Claudia Klüppelberg · Daniel Straub
Isabell M. Welpe *Editors*

Risk

A Multidisciplinary Introduction

Springer

Risk – A Multidisciplinary Introduction

Claudia Klüppelberg • Daniel Straub •
Isabell M. Welpe
Editors

Risk – A Multidisciplinary Introduction

Springer

Editors

Claudia Klüppelberg
Department of Mathematics
Munich University of Technology
Munich, Germany

Isabell M. Welpe
TUM School of Management
Munich University of Technology
Munich, Germany

Daniel Straub
Faculty of Civil, Geo
and Environmental Engineering
Munich University of Technology
Munich, Germany

ISBN 978-3-319-04485-9 ISBN 978-3-319-04486-6 (eBook)
DOI 10.1007/978-3-319-04486-6
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014934154

Cover illustration: Peter Bühlmann

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Introduction

Risk is a multi-faceted and complex phenomenon; one that defies pure disciplinary treatment and whose analysis and evaluation requires interdisciplinary competencies. Recent events like the 2008 financial crisis, large scale black-outs in energy supply systems, the Deepwater Horizon oil spill in the Gulf of Mexico and the earthquake/tsunami event triggering the Fukushima nuclear accident highlight the fact that risks are increasingly caused by the complex and interconnected nature of today's societies and technologies. One of the major conclusions drawn from these and many other events is the strong need for improving the interdisciplinary analysis, evaluation, management and communication of risk.

Risk and security issues have always been important in modern societies; but they re-emerge and change shape, involving new issues due to rapid and unprecedented technological and climatic changes or political developments. These developments cause major challenges to identification, understanding and management of risk. While traditional tasks (e.g. reliability, availability of technical systems) need to be reframed, and approaches and methods need to be further advanced, they still allow for primarily disciplinary treatment. In contrast, modern risks tend to be systemic in nature and clearly demand an interdisciplinary and trans-sectional approach.

Present-day research and training with its sectorial approach cannot meet the challenges posed by multiple and interlinked events and systemic risk. In the classical areas of risk and reliability analysis, such as transportation systems, pharmaceuticals and structures, undesired triggering events and event sequences, their frequencies and consequences are evaluated within clear sectorial limitations in space and time. However, todays challenges pose new demands on risk analysis and risk evaluation:

- technical, economic and social systems become more integrated, e.g. by digital ICT, and extend to large scale-networked systems;
- risks are interconnected and coupled to varying degrees;
- risk events once triggered, cascade, cross borders and are becoming systemic;
- consequences of events extend to long-lasting, trans-boundary problems as well as loss of products and services systems.

New and interlinked risk and security problems in a modern information society—with its highly computerized trading centres, banks, the Internet and the large-scope interactions between humans and the environment—add an unknown complexity to the classical areas of technical risk analysis. Their management requires special socio-technical knowledge and methods and should also include behavioural components as convincingly demonstrated throughout the 2008 financial crisis or the Fukushima event. The miscalculation, misunderstanding or miscommunication of risk by politicians, scientists, and executives poses an additional challenge. Today's interconnectedness of economic, social and political changes results in complex and simultaneous reactions in various areas, and affects scientific development and technological assessment. Furthermore, climate risks, environmental risks and biotechnical food as well as medical risks have increased immensely over the last decades. Global warming, endangered balance in soil-water systems and the decay of biodiversity are long-term risks that are highly interdependent and will influence each other and our society in unfamiliar and complex ways in the future. The scope of their influence is hard to predict by today's disciplinary risk management approaches and methodologies.

One book can hardly address all challenges in risk outlined above. Nevertheless, we hope that this book inspires multi-disciplinary learning, stimulates systemic thinking, sharpens multi-methodological competencies, and brings risk and security issues closer to readers with various backgrounds. The goal of this book is to integrate risk and security issues into core domains of natural sciences, engineering, life sciences, management, and medicine. It should encourage readers to work with methods and subjects of various disciplines and on specific cross-sectorial risk issues. With this book we want to promote a common language and thus contribute to risk communication across disciplines and between theory and practice. It can serve as a training guide when dealing with complex risk decisions, which typically have direct and indirect effects in economic, technical and social areas.

Understanding and assessing systemic risk quantitatively is currently the main challenge of risk research, and it is therefore one goal of this book to foster systemic risk understanding in an interdisciplinary way. With the foundation of the Munich Center for Technology in Society (MCTS), the Technische Universität München (TUM) commits itself to explore how society impacts research and vice-versa, which ethical factors should be taken into consideration when developing new technologies, and how science and the general public can communicate with each other. Projects like "sociotechnical systems, robotics and demographic change", "water management", "mistakes, ignorance, contingency, and error in science and technology", or "from prognosis to predictive medicine" have always a risk component, and such projects have influenced the writing of this book. Leading experts from TUM present novel exciting fields, surveys of recent developments, or focus on some of the most challenging applications in future risk research. Despite the wide range of topics, each chapter is written in an expository style, with two bibliographies at the end. A selected bibliography gives fundamental publications to the topic at hand. Additional references aim at those readers eager to dive deeper into the topic. In this way, each article makes an invaluable comprehensive reference text. The intended readers of this book are researchers from all fields, PhD students and postdoctoral researchers, who look for an introduction to risk from different angles, and want to get an overview of old and new applications in different fields and, not least, are looking for inspiration from other areas of research.

At TUM, the book will be used at the MCTS to establish a unique research training program that will create a new generation of researchers, who will understand risk and security in an integrated and interdisciplinary way. Young researchers of the MCTS will receive multi-methodological training that draws from state-of-the-art methodologies of all disciplines involved. These methodologies contain a variety of risk-relevant concepts and will have a toolbox of quantitative methods at their disposal. Our training develops communication skills and enables graduates to work in an interdisciplinary team in the context of risk and security-related issues.

The book is organized in three sections.

Part I Risk in History, Society and Science: four chapters provide the context in which risk is to be seen. The first chapter gives a historical introduction into the change of the perspective on risk during the centuries. Risk and business ethics is the topic of Chaps. 2, and 3 explains the often difficult decision making process from different perspectives. Chapter 4 by the Director of the MCTS, Prof. Klaus Mainzer, makes the transition to the methodological section.

Part II Quantitative Risk Methodology: six chapters provide insight into quantitative methods for risk assessment. The first Chap. 5 introduces into the modern theory of risk measures, the second Chap. 6 reviews extreme value theory and statistics as a basis for extreme risk assessment. Statistical prediction by linear models and Bayesian modelling is presented in Chaps. 7 and 8. Some of the sometimes disastrous consequences of dependence for risk modelling can be seen in Chap. 9. Finally, Chap. 10 deals with model risk, one of the big issues of quantitative risk assessment. All chapters contain a fair bit of theory, but all theoretical concepts are illustrated with applications from various fields.

Part III Risk Treatment in Various Applications: in contrast to Part II, the focus of these chapters is the application. Methodology is presented, because of its relevance to the application at hand. Such applications range from management problems, classical engineering risk problems via information systems to medical cancer risk research.

We editors and authors take pleasure in thanking our home institutions for providing excellent working conditions. As most of us are members of the scientific TUM community, this is a possibility to pay tribute to its scientific environment. The TUM Institute for Advanced Study, founded with the support of the German Excellence Initiative, provided since 2007 an interdisciplinary atmosphere, where the idea of this interdisciplinary introduction to risk was born. The editors and several authors acknowledge financial support from the Institute, and they also profited immensely from the various interdisciplinary meetings at the TUM-IAS.

We also take pleasure in thanking a number of people, who supported the book at various stages. We are particularly grateful to Prof. Wolfgang Kröger, Managing Director of the ETH Risk Centre for various discussions on risk related research issues. Also other members of the ETH Risk Centre were most supportive. Furthermore, Prof. Orthwin Renn, Chair of the Environmental Sociology and Technology

Assessment, University of Stuttgart, deserves our gratitude for constant support. On the technical side Dr. Thomas Klein helped to set up the design of the chapters and started the communication process with the authors. Dr. Anita Behme took over as Managing Editor and handled the refereeing procedure efficiently, which was a major task with so many authors from different fields and varying background. Her careful inspection of the book chapters and her patient communication with authors and referees have contributed immensely to the quality of the book. She also handled the book production with the Springer staff. The negotiations with Springer concerning the contract and the following printing process went smoothly, and we thank the staff involved. Finally, we would like to thank Prof. Charles Goldie, Emeritus from Sussex University, who carefully read several book chapters and helped them to remain within certain bounds of proper English. Of course, the responsibility for all final contents lies by the authors.

Munich, Germany
October 30, 2013

Claudia Klüppelberg
Daniel Straub
Isabell M. Welpe

Contents

Part III Risk Treatment in Various Applications

Part I
Risk in History, Society and Science

Chapter 1
Risk in Historical Perspective: Concepts, Contexts, and Conjunctions

Karin Zachmann

Although the etymological roots of the term risk can be traced back as far as the late Middle Ages, the modern concept of risk appeared only gradually, with the transition from traditional to modern society. The modern understanding of risk presupposes subjects or institutions, accountable for their actions, that make decisions under conditions of apparent uncertainty. Some apparent uncertainties, however, can be measured or quantified probabilistically and are, therefore, more precisely called "risks". Situations of "risk" in human society can thus be "managed". Relying on probability calculation, which emerged during the 17th and the 18th centuries but became truly prevalent only in the 20th century, risk became a theoretical focus designed to bolster a scientific, mathematically-based approach toward uncertainty. Insurance companies led in demanding and developing a concretely applicable concept of risk, since calculating the probability of premature death or material hazards related to either humans or material things, such as ships, buildings, and their contents, was essential for their core business and success. However, by the middle of the 20th century—an Age of Extremes, as it has been aptly characterized—nuclear weapons and their use in Japan and subsequent further development early in the Cold War dramatically increased awareness of potential hazards derived from these and other achievements in science, engineering and warfare. Therefore, the Age of Extremes stimulated more and new research on risk. With new tools, such as operations research, digital computers, systems analysis, and systems management, all of which had been introduced in the military and aerospace sectors in the course of World War II, the intellectual resources necessary to estimate the extent and the probability of failures and accidents in nuclear warfare and beyond increased dramatically. Out of the Cold War effort to create the "Peaceful Atom", nuclear-power reactor safety studies became landmarks in risk analysis, and this type of study later achieved relevance in many more areas. This chapter seeks to explore the evolution of risk research and risk management in its social and political contexts in order to understand the underlying concepts of risk and safety as social constructs. The

K. Zachmann (✉)
History of Technology, Munich Center for Technology in Society, Technische Universität München, c/o Deutsches Museum, 80306 Munich, Germany
e-mail: Karin.Zachmann@mzwtg.mwn.de

C. Klüppelberg et al. (eds.), *Risk – A Multidisciplinary Introduction*,
DOI 10.1007/978-3-319-04486-6_1,

historical survey focuses mainly on the last two centuries. It starts with the advent of the modern era when with spreading bourgeois virtues it became common to plan for the future but not to bet on it. This involved an increasing need to calculate future uncertainties in order to manage them as risks. The study stops at the end of the Cold War, when the collapse of the socialist bloc settled the risky confrontation between the two opposing societal camps. By no means did the termination of the Cold War end the story about risk. On the contrary, as late modern societies accumulate more and more knowledge they simultaneously increase the amount of ignorance that is the cause of newly emerging risk. How these risks are tackled is the topic of the other chapters in this book. This historical survey does not aim at completeness but rather at understanding the major transformations in the evolution of risk. Thus, not all areas in the history of risk are covered here; for instance, the important field of financial risk is treated by other Chap. 4.

Keywords Food safety regulation · Probabilistic health risk research · Quality and reliability engineering · Reactor safety studies · Steam boiler safety

The Facts

- While mathematicians in the era of the scientific revolution and the enlightenment began to approach uncertainty as probability, the early-modern passion for gambling shaped notions on risk as genuine uncertainty and, therefore, precluded the early application of the nascent tools of probability.
- Both the quality of uncertainties and attitudes toward them changed in conjunction with the great political, technological and social transformation of societies in the Western world since the beginning of the 19th century.
- Human-made dangers and threatening uncertainties resulted from the introduction of new technologies, from urbanization and from the industrialization of food; these induced Western societies to commence framing and managing uncertainties as risks.
- The burgeoning insurance industry, which, since the 19th century, sold its customers a new degree of control over uncertainty, evolved as an important promoter of research as to the causes and prevention of risk, and became an important contributor to the quantitative understanding of risk.
- The adoption of state compulsory accident insurance especially gave rise to the emergence of industrial medicine, which also furthered probabilistic approaches in medical research and industrial hygiene.
- The development of quantitative approaches to system safety and reliability in the Bell Telephone System in the 1920s, as well as the German beginnings of “Großzahlforschung” (see Sect. 3.5), constituted an important building block for the emergence of quality and reliability engineering and Probabilistic Risk Assessment in various fields of complex engineering systems.
- Safety engineering in the aerospace and defense sector gave rise to pioneering quantitative as well as qualitative methods of risk assessment.

- The so-called Rasmussen Reactor Safety Study, issued October 1975, was a contested and yet celebrated breakthrough of Probabilistic Risk Assessment, and its method spread to other branches as well as countries beyond the US.
- Risk research and risk management have become an increasingly professionalized endeavor since the 1970s, when late modern societies began to pay more attention to the swelling uncertainties that accompanied the experience of increasing ignorance as an unavoidable side effect of the production of more and more knowledge and unbounded Promethean technological and industrial development.

1 Introduction

Risk gained the popularity of a keyword in the latter part of the last century (cf. Williams [99]). Politicians, civil organizations of various kinds, researchers, experts, doctors, generals, publishers, and many more people and institutions felt the need to tackle problems of risk in a more systematic fashion (Renn [34, 35]). In 1980 the international risk research community established its own professional society—the Society of Risk Analysis (SRA)—which has published Risk Analysis: An International Journal since 1981 (Thompson, Deisler, and Schwing [38]). When German sociologist Ulrich Beck produced his analysis of late modern society under the thrilling title Risk Society shortly after the Chernobyl reactor catastrophe focused people's attention on the enormous dangers of nuclear power plants, the book immediately became a big success (Beck [2]). According to Luhmann, this phenomenon of sustained focus on risk reveals a remarkable characteristic of late modern society; as he argues, risk became the main approach to addressing the problems of uncertainty (Luhmann [24]).[1] Uncertainty, however, is a fundamental anthropological experience. People in all societies have had to deal with uncertainty in one way or another. Thus, if we want to understand the significance of risk in our present society, we need to explore the following questions: when did the attitude toward future uncertainties change so that the understanding of uncertainties became narrowed down to risk? How did the modern concept of risk determine people's ways to deal with uncertainties? How widely accepted has modern risk analysis become, in what ways has such analysis proved to be particularly problematic, and in what manner has risk analysis become professionalized?

2 Pre-modern Ways of Coping with Uncertainty and the Emergence of Proto-Modern Notions on Risk

Members of pre-modern societies experienced uncertainties in manifold ways as their success of everyday action was highly vulnerable to a great variety of unex-

[1] On the classical differentiation between uncertainty and risk see Knight [22]. As a well informed and yet popular story of risk see Bernstein [43].

pected or inalterable events, such as premature death, famines, natural disasters, wars, epidemics such as pestilence and the plague, violent politics, and so on and so forth. Most of all, religious belief systems and magical as well as divinatory practices provided methods for coping with these uncertainties. Confidence in the wisdom of gods helped humans to accept uncertainties as one's fate, and collectively practised magical rituals did so as well (cf. Luhmann, 16–17 [24] and Douglas and Wildavsky [12]).

Fateful resignation, though the main method, was just one way to cope with uncertainty. Already in the 12th and 13th centuries a new attitude toward uncertainty emerged in the Italian cities and city states. Merchants and seafarers started to take uncertainties as a chance to improve their welfare. Speculating on a fortunate course of events, they ventured out beyond known places and thus risked long sea journeys. Here uncertainty was no longer seen only as danger and passively endured as fate, but taken as a challenge that could pay off if their calculations worked out. Calculations, however, meant nothing but informed guesses at that time when available information remained exceedingly less than sparse. It is important to note that in this very context the term "risk" came to be used (cf. Bonß, 49–50 [4] and Luhmann, 17–18 [24]). While risk expressed a new, active, and positively connoted stance on uncertainty, it also gave rise to a new need. In order to get the calculations right, risk takers wished to learn new methods of forecasting the future course of events beyond traditional practices of divination, the belief in the wisdom of gods, and resignation to an unknowable fate.

The emerging new attitude toward uncertainty spread throughout Europe, and this boosted the desire to gain control over an unknown future. This development signifies a remarkable shift from "traditional" to "modern" perspectives, as the risk seekers hoped to determine their own future. Thus they increasingly gained confidence that nature could be conquered and the world improved by human action (Bonß, 52 [4]).

In the mid-16th century risk-taking even advanced to become a new business as the new legal category of aleatory contracts revealed. According to Daston these contracts subsumed "all agreements involving an element of chance, any trade of here-and-present certain goods for uncertain future goods: annuities, gambling, expectation of an estate, purchase of a future harvest or the next catch of a fisherman's net..." (Daston, 238 [10]). In the late 17th and early 18th centuries, England's bustling capital London provided the most fertile breeding ground for the business of risk-taking, as is evident from the quickly expanding insurance market. Maritime insurance multiplied on the initiative of individual brokers who gathered in places like Lloyd's Coffee House. In addition, new branches emerged such as fire and life insurance, not to mention the many adventurous schemes that promised protection against any and every contingency of life. It was, however, not yet prudent foresight but a reckless spirit of gambling that fueled this early boom of insurance (Daston, 165 [11]).

As for the calculation of risk, however, contractors relied on rules of thumb and all forms of experience rather than statistical approaches. The fact that past experience took manifold forms and obscured any regularity prevented early entrepreneurs of risk from attempting calculations based on systematic empirical data (Daston, 240 [10]). The practitioners' non-statistical stance notwithstanding, aleatory contracts paved the way toward mathematical probability because they put new problems and questions before mathematicians. The latter, however, remained caught in the mindset of the jurists who posed the problem when they sought to determine the fair price of an annuity or a life insurance premium. Thus, the mathematicians began to tackle the new field in terms of mathematical expectations, i.e. the product of a probability of an event and its outcome value or "payoff" (Daston, 240 [10]). Their approach to quantifying uncertainty as probability, however, worked against the application of mathematics in this early modern business of risk as the aleatory contracts defined risk as "genuine uncertainties". Quantification could have diluted the genuine uncertainty and thus would have worked against the playful rationality of aleatory contracts (Daston, 247–248 [10]). Therefore, deploying the mathematicians' new achievements of probability as a way to control uncertainty required a new attitude toward risk. The latter had to be redefined from something to be desired into something to be avoided. A favorable context for this redefinition evolved as soon as bourgeois values of familial responsibility, control, and predictability began to determine the norms of society (Daston, 182 [11]).

3 Industrialization, Urbanization and Competitive Markets: New Qualities of Uncertainty and the Beginnings of Risk Management

Within the great political, technological and social transformation of Western societies that was pioneered by the British Industrial Revolution and the French Bourgeois Revolution, the meaning of uncertainty changed substantially. In contrast to the gambler as well as the venturesome man of action in the Ancien Régime who had appreciated uncertainty as a chance to make a fortune and as a way to escape the fate of the natural as well as the religious order, the capitalist entrepreneur as well as the male breadwinner who was entitled to vote did not want to bet on the future but to plan for it. Thus, they strove to enlist knowledge in order either to reduce or to circumvent uncertainty. Gaining control on the unknown worked as a strong motive. That was, for example, the case for the French revolutionaries and the German bourgeois reformers who wanted to determine the state of society. It was also true for the agriculturalists, engineers, entrepreneurs, architects, members of the academic elite, and many others in Britain, France, Germany, and elsewhere as they all together aimed at extending human control over nature. And indeed, people who had been living in the Western world since the mid-19th century experienced a higher degree of predictability during the course of their lives when more children than ever before survived past infancy, when dwellings withstood fires for generations, when famines no longer constituted the rule but became exceptional events in the experiences of Western men and women, to name just a few most fundamental improvements in human existence.

More stability and predictability, however, did not free the urban middle class or capitalist entrepreneurs and farmers from fear. At the same time as people accumulated more knowledge and competencies to put an end to uncertainties, they increasingly felt ignorant about many things that were coming into their lives. Railroad accidents, steam boiler explosions, collapsing bridges, adulterated food, and several waves of cholera epidemics in rapidly expanding cities, among other perils, marked a new class of human-made dangers and threatening uncertainties.

How did men and women in mid-19th century Europe and North America cope with such new dangers? They developed a whole range of strategies and institutions to gain control of uncertainties and to decrease the probability as well as the extent of these misfortunes. This was the context out of which the modern politics of risk management gradually emerged, notwithstanding the fact that the term risk was only seldomly used and if so in a much narrower sense.[2] Thus, by exploring this emerging new field of politics we can learn a great deal about how the current concept of risk evolved and changed over time. We will see how, following the efforts of industrializing societies to develop approaches and institutions for regulating dangerous activities, uncertainties became framed and managed as risks and thus necessarily also gave rise to new notions of security.

3.1 Controlling Technical Risk: From Steam Boiler Associations to Safety Standard Authorities

The steam engine is often seen as a paradigmatic invention of the so-called British Industrial Revolution. Its widespread use in powering factories and river and rail transportation also decisively triggered the transformational process of introducing new perils into society because it was prone to explode, leading to deaths, serious injuries, and destruction of valuable property. Steam boiler explosions constituted a completely new form of threat because they exposed people for the first time to the destructive potential of modern technology. Therefore, steam boiler explosions mobilized a concerned public, led to pioneering scientific and engineering investigations of such "failures", and required governments to institutionalize construction and operation standards and regular safety inspections. Hence, the state felt obliged to diminish the risk of explosions and thus to establish a new concept of technological safety.

In France, as well as in Prussia and some other German territorial states, the state set up steam boiler legislation and introduced rules and institutions for inspection (for France cf. Fressoz [15] and for the German states cf. Wiesenack, 5–18 [40]). In Great Britain the owners of steam boilers established boiler insurance and introduced private inspections. In the United States, public outrage about increasingly

[2] During the 19th century the term risk remained confined to the economic sphere and was used with the meaning of venture or hazard of loss (cf. Schulz and Basler, 452 [88]).

numerous and deadly explosions of steamboats led the US government to commission the Franklin Institute to investigate the causes of steam boiler explosions and to recommend means by which they could be prevented. The institute's investigation resulted in the first form of federal regulation of technology in the US, but the regulations and Federal power were so weak that boiler failures remained a common occurrence well into the 20th century, when the American Society of Mechanical Engineers promulgated its Steam Boiler Code in 1916 based on what became known as a consensus standards-making process (Burke [5], Sinclair [89, 90]). In the German states, at first the state conducted inspection, but this system was gradually replaced by privately founded steam boiler associations wherein boiler owners and manufacturers set up a self-organized inspection process. The associations claimed autonomy based on technological expertise that the states did not possess. But the real problem at stake here was this: who would more successfully ensure the workers' and citizens' safety with regard to technology, the authoritarian state or private entrepreneurs and engineers in a liberal market? In the years from 1866 and 1911, in all German states, 36 steam boiler associations came into being (Wiesenack, 19–21 [40]). The federal law of 1872 assigned the privately organized associations the task of inspections, and in subsequent years, until the outbreak of World War I, the German states extended the associations' responsibility of regularly conducted revisions onto newly emerging fields of potentially dangerous technological installations and artifacts such as steam vessels, elevators, motor vehicles, vessels for pressurized or liquidized gases, mineral water apparatus, acetylene-generating and -storing units, and electrical installations (Wiesenack, 38–74 [40]). The steam boiler associations took up these new fields of activity with hesitation because the new tasks had to be carried out on behalf of the state for nonmembers of the associations in technical areas beyond the specific expertise of steam boiler engineers (Wiesenack, 42–46 [40]). Such resistance notwithstanding, especially in the interwar period, the new areas and technologies—in particular, the inspection of motor vehicles—gained increasing importance; thus the steam boiler associations changed into safety standards authorities. About one year before the Nazi regime triggered World War II and thus began to deploy the destructive forces of technology in new and unknown dimensions that put millions of people at risk and to death, the federal minister of economic affairs reorganized the technical safety inspection system when he transferred the powers of regulation from the states to the Reich and officially transformed the steam boiler associations into state-regulated but self-governed safety standards authorities (Wiesenack, 77–92 [40]).

Thus, if we focus only on the German case, we can see that in the nearly 70 years from the unification of Germany in 1871 to the advent of World War II (WWII) the danger from accidents of technological artifacts and installations that were prone to explode, to cause fire, or to go out of control gave rise to the establishment of a still-important field of risk management. To be sure, the participants in this development hardly used the term risk prior to the second half of the last century. Nevertheless, they developed regulations, strategies, and routines for coping with a new class of human-made dangers: technical risks. One way to accomplish this task was to broaden the field of technical knowledge. Therefore, the associations collaborated with technical universities (or they even established their own research laboratories, as the Bavarian steam boiler association did in 1904 under the direction of the eminent inventor, engineer, and industrialist Carl von Linde—The Steam Research Laboratory, cf. Wiesenack, 22–23 [40]). Furthermore, the associations not only conducted inspections, but also worked as consultants. They participated in developing norms of technological safety and pushed for safety improvements (Wiesenack, 73–74 [40]). Whereas the steam boiler inspectors' notion of risk was confined to the likelihood of failure of technological equipment, this notion became broader after World War II, when the safety standards authorities in Germany and elsewhere extended their domains, as they included dangers that resulted not from failure but from "normal operation" of technology. This new awareness of dangers emerged with the spread of large technological systems (Perrow [30]). Safety standards authorities, however, were politically unsuccessful in establishing legally binding safety norms for the design and use of technology. These legal constraints worked as a strong impetus toward the development of risk analysis because the assessment of risks was to supersede legally inadequate regulations via safety norms (Lukes [25]). But until today it is an open question in engineering, whether probabilistic calculations are superior to safety margins or not (Doorn and Hansson [54]).

3.2 Managing Health Risk: City Sanitation and the Coalition of Experts and Stake Holders Against the Cholera Threat

The introduction of new technologies was not the only source of new perils to industrial society. Industrialization itself led to rapidly growing cities, which in turn exposed people to more danger, as the likelihood of epidemics spreading from crowded quarters with poor living conditions, lack of adequate public sanitation (i.e., human waste management), insufficient water supply, and high pollution in even remote and wealthier waterways and neighborhoods grew (as a pioneering study see Simson [37] and more literature in Labisch and Vögele [72]). In the time span from 1831 to 1892, the northwest of Europe was struck by four waves of cholera epidemics with a death toll of 50 percent of all men and women who fell ill. (Because of increasing "globalization" of commerce and emigration, the United States experienced an equal number of cholera epidemics over the same seven decades. For the US see the eminent book of Rosenberg [85] and for Hamburg see Evans [57].) In fighting this danger, European city authorities, in collaboration with technical and medical experts (i.e. engineers and doctors), developed increasingly successful strategies of risk management. In local politics, engaged hygienists—a new, interdisciplinary oriented group of experts—took up the issue of city pollution as a health problem and established coalitions of local politicians, businessmen, engineers, doctors, and other experts. These coalitions mobilized knowledge, experience, and competencies from various fields in order to advise municipal authorities on appropriate solutions for their city's sanitation and improved public health. In Germany, the Frankfurt doctor and local politician Georg Varrentrapp (1809–1886) decisively shaped the coalition of experts when he established the German Association for Public Health in 1873 (Hardy [64]). Among the first 230 members, there were 20 mayors of big cities (Berlin, Frankfurt, Munich, Danzig, ...) besides other municipal authorities, 112 physicians, and a wide range of architects, engineers, entrepreneurs, chemists, pharmacists, journalists, as well as famous hygienists from abroad. Meetings of the association provided a forum to negotiate core problems of city hygiene and public health among the interdisciplinary group of experts. Participants gave lectures that were extensively discussed by all members. The aim was to find common ground between the medical, technical, and financial arguments. Via majority vote the association settled its negotiations and thus established a base of knowledge for enabling municipal authorities to take decisions on appropriate sanitation systems. With such mobilization of experts from different fields, as well as engaged and concerned citizens, local authorities and stakeholders of various kinds accumulated and disseminated knowledge and evaluated alternative strategies for reducing the risk of an epidemic's outbreak. Thus, the protagonists of the 19th-century hygiene movement invented a pattern of risk management that enabled the hygienist activists to push decision-making in favor of sanitation systems, although the question as to the causes of infectious diseases was not yet settled (Hardy, 108 [64]).

3.3 Regulating Food Risk: The Introduction of Science-Based Food Control

In the mid-19th century complaints about food adulteration and consumer fraud began to make headlines in the press of industrial countries. The range of new food products on the markets stemming either from imports or from innovations of industrially processed food challenged the experience-based knowledge not just of consumers but also of food merchants to make judgments on food quality (Zachmann and Østby [41]). A remarkable percentage of these product innovations and product changes were initially perceived as adulteration, and this caused heightened uncertainty at the food market. Because inadequate food supply can easily result in political unrest—many German cities experienced bread riots on the eve of the 1848 Revolution—national legislators strove to establish an infrastructure for food control through enforcing nation-wide food laws that were to supersede local regulations.

Great Britain pioneered the development. In 1860 Parliament enacted a landmark food law aimed at preventing adulteration of all food and drink. (For more details see Clow and Clow [49], Wohl [100], Smith and Phillips [91].) The German empire followed in 1879, and between 1890 and 1906 national food laws were enacted in Belgium, Austria, Switzerland, France and the United States. These laws, however, provided just the framework of food controls, and had to be supplemented with food standards as benchmarks for proving food quality. But who was to define food standards? Practitioners of the food business claimed to have the last word on how to secure food quality and food safety, and they for the most part showed limited interest in collaborating with experts such as chemists, hygienists, or doctors. The chemists developed more and more interest in food chemistry as the chemical analysis of food promised to become a rewarding field for exploiting professional expertise. Thus, chemists pushed chemical analysis and employed the nutrients paradigm for determining food standards and subsequently food quality (Spiekermann [93], Dessaux [51], Hierholzer [65]). National legislators again faced the task of reconciling the interest of the food industry in liberal markets with consumers' demand for safe food and the states' interest in public health and political stability. Thus, national food legislation at the turn of the 20th century gave rise to nationally slightly different systems of food control in order to manage food risk (Spiekermann [93]). At the same time, however, hygienists and chemical experts pushed for an international approach toward food regulation (Dessaux [51]). In September 1907, La Croix Blanche de Genéve was created as an international association, based in Paris, specifically in order to fight food fraud and adulteration. The association organized two congresses, the first in Geneva in 1908 and the second in Paris a year later. Then it petered out. In spite of its short life and the fact that it took the Codex Alimentarus, its successor, almost half a century to get established, the association had a great impact on food safety regulation. It strengthened the authority of chemical expertise in the food market, as the association's organizers had managed to reach agreement on a broad catalog of food definitions. These definitions provided the fundamentals of food evaluation based on chemical analysis. Thus, at the turn of the 20th century, food risk management was established as food regulation, and subsequent food regulation based on food standards became established in a tense collaboration of chemical experts and food industry representatives. The institutions established in the late 19th and early 20th century have continued to be the primary institutions dealing with food safety, even as the globalization of food supply has raised many questions about food safety.

3.4 Capitalizing Risk and Enhancing Social Security: The Emergence of Insurance as Catalyst of Modern Strategies Toward Risk and Security

Whereas the aforementioned strategies of risk management aimed at preventing individual and societal harm from technologically produced hazardous products and environments ranging from steam engines to crowded cities and food adulteration, the advancing insurance system of the nineteenth century promised to compensate persons harmed, the survivors of deceased victims, and the owners of damaged property. Modern insurers, who had severed their business practices from gambling, now capitalized on risk, as they sold their customers no longer chance but a new degree of control over uncertainty through empirically established probabilities. Hence, risk became a new commodity.

Insurance companies that, in the modern sense, offered contracts with mathematically calculated premiums and a legal claim on the indemnification payment were founded at first in London. The Amicable Society (est. 1706) pioneered the advance, as the world's first life insurance company, but operated at first more as a friendly society than as a business. The Amicable, however, induced a rejected applicant who was a mathematician to establish the Equitable Society in 1762. As the world's oldest mutual insurer the Equitable owed its success, as we learn from Daston (175 [11]), to "its exploitation of the regularity of the mortality statistics and the mathematics of probability to fix premiums [. . .], but also. . . [to] its creation of an image of life insurance diametrically opposed to that of gambling".

From the early 19th century a whole range of new insurance branches emerged that signaled where witnesses of industrialization and urbanization perceived new, potential threats to their bodies, businesses, and property and thus felt compelled to make provision for such contingencies. In Germany, for instance, private entrepreneurs insured against the risk of transport damages on the Rhine river traffic (1818), the risk of harm by railroad accidents (1853), injury by broken glass (1864), damage from broken taps (1886), and losses caused by mechanical breakdown (1900). Furthermore, in 1829 the first reinsurance business was established, and in 1875 personal liability insurance was set up (Koch [71]). While in all these cases private entrepreneurs developed a need for more safety as a chance to earn money, nation states also detected the potential advantages of the insurance trade. In contrast to the fund-seeking politics of early modern states that sold annuities for getting the sovereign money, nation states sought to utilize modern insurance in order to provide for political stability via social security systems. The founder of the German empire, Bismarck, pioneered the institutionalization of state compulsory insurance, i.e. social security, as well as health and accident insurance (Ritter [84]). As soon as the states enacted compulsory forms of insurance, provisions for mitigating risks became a pillar of the welfare state (Ewald [13]).

The enhancement of risk policies, together with the enormous extension of the insurance system throughout the long 19th century, necessitated the accumulation of knowledge and experience on how to assess and to manage risks. For insurers this was of critical importance, as the success of their business stemmed in large measure from such knowledge. The first insurance branch to develop and apply theoretical knowledge was life insurance. Insurers could build upon the well-developed classical probability theory and upon mortality statistics. Therefore, it came at little surprise that in Great Britain in 1848 the Institute of Actuaries was founded (Pabst, 26 [29]). In Germany, however, insurers had been much more reluctant to develop an interest in scientific knowledge. Only at the turn of the 20th century did some German universities, such as Göttingen, Leipzig, Frankfurt am Main, and Cologne, and Technische Hochschulen, such as Dresden and Aachen, set up study courses related to the insurance business. Göttingen was the first to establish a "seminar on insurance science" in 1895 (Pabst, 26–29 [29]). Insurance science, however, was not a coherent field of knowledge but a conglomerate of many special fields. The theoretically most advanced and exacting field was actuarial mathematics, which is first and foremost probability theory. Actuaries, however, were employed only in the life insurance area until well after 1950, as Reinhard Pabst has shown in his dissertation (Pabst, 116–118 [29]).

Except for life insurers, practitioners in the insurance business proved to be quite averse to theoretical approaches to risk calculation. One main reason was the lack of appropriate statistical data. Another reason was economic success based on more traditional methods. Empirical knowledge and experience remained very important for estimating risks and insurance premiums. For example, even as late as the mid-20th century, maritime insurers would gauge, as their predecessors in 16th century Venice had done, "the integrity of the ship-owner, the skill of the ship's officers, [and] the quality of the crew" (Pfeffer, 69 [31], Gigerenzer, 257 [19]).

Pabst's study on machine insurance reveals that insurers did not put much emphasis on more elaborate risk assessment for improving premium calculations but preferred to make provisions for damage prevention by increasing the availability of new technological knowledge. Allianz, the largest supplier in this field, published a journal called "The Mechanical Breakdown" to teach strategies of how to avoid breakdowns. Furthermore, the insurer organized company inspections, better turbine control procedures, and manager training classes, and set up its own materials-testing institute and museum. With such measures, insurers of technological risks developed a new domain of employment for engineers (Pabst, 52–79 [29]).

The increasing availability of technological knowledge notwithstanding, experts in the property insurance business began to articulate a need for more theoretical knowledge by the end of the 1920s. Founded in 1935, the German Association of Actuaries put the development of mathematics for the property insurance business on the agendas of its congresses in subsequent years. Thus, expectations grew that probability theory would begin to be applied beyond life insurance for the analysis of uncertainties and the identification of risk in property and indemnity insurance (Pabst, 80–97 [29]). A first mathematical model for non-life insurance, however, had been presented by the Swedish actuary Filip Lundberg in 1909. It was largely ignored until the Swedish professor Harald Cramér from Stockholm University built his insurance risk theory based on Lundberg's approach. Even Cramér's risk theory was slow to be used; only well after World War II did the insurance industry widely adopt it, albeit the first publication dated from 1930 (Pabst, 52–53 [29]). This delay reveals that practitioners paid little attention to the ambitions of actuaries, and with the outbreak of World War II all priorities changed anyhow. General diffusion of actuarially based risk theory in non-life insurance was delayed until the international community of actuaries established the Actuarial Studies in Non-Life Insurance (A.S.T.I.N.) organization in 1957. Establishment of this organization proved to be an important step for the diffusion of probability theory in non-life insurance, even if the transition from actuarial theory to practice took longer and diffused at different rates in the various branches of property and indemnity insurance (Pabst, 126–130, 165–193 [29]).

The extension of the insurance business increased risk awareness, and at the same time promoted research as to the causes and the prevention of risks. This was true not just for the aforementioned property risks due to technological breakdowns, but also for health risks caused by industrial accidents. When national governments in many countries, following Bismarck's pioneering example, began to insure workers against industrial accidents, research in industrial medicine received a tremendous boost. Physicians who worked for the state in compulsory health and accident insurance developed industrial medicine. The subject area of the newly emerging field was the detection and prevention of health risks and risks of accidents in industrial work places (Lengwiler, 146–148 [23]). The physician's task to provide insurers with medical certificates as to the causes of damage to insured workers' health boosted research on medical causalities. Up to the interwar era of the 20th century, medical causality was discussed most in bacteriology. Here Robert Koch's and Louis Pasteur's explorations of tuberculosis and anthrax as bacteriologically caused diseases gave rise to a mono-causal, deterministic concept of disease that replaced manifold etiologies (Schlich, 8 [87]). But with the more frequent appearance of particular diseases in specific industrial environments, such as silicosis or various kinds of cancer, mono-factorial chains of causes did not work. Therefore, in the interwar era, industrial medicine gradually began to abandon strictly deterministic concepts of causality in favor of probabilistic health risk research.

As Martin Lengwiler has shown in his study on the development of accident insurance in Switzerland from 1870 to 1970, probabilistic concepts gained ground particularly in the emerging field of toxicology (Lengwiler, 149–158 [23]). An important figure in this field was the director of the forensic institute at the University of Zurich, Heinrich Zangger (1874–1957). Poison gas attacks in World War I, as well as high incidence of poisoning from wartime-promoted chemical substitutes, inspired him to deal with military as well as industrial poisoning. With improved measurement methods based on new instruments, he began to use a statistical approach to evaluating the effects of poisons on human bodies. Thus he paved the way toward probabilistic diagnoses. Zangger defined industrial medicine as a "science of danger" aimed at control and prevention by describing potential dangers of industrial and technological environments. Zangger's concept of a science of danger stands for an early approach toward an independent and theoretically ambitious discipline of medical risk research (Lengwiler, 152 [23]). Toxicology as pioneering medical risk research was to determine the risk of poisoning emanating from human exposure to dangerous materials (Hounshell and Smith [67]). During the 1930s toxicologists introduced threshold value definitions under the heading of "maximum acceptable/allowable concentration" (MAC) of hazardous materials in workplaces (e.g., exposure of workers to a range of organic chemicals used in the manufacture of synthetic dyes). In 1933, industrial physicians within the Soviet public health system had been the first to succeed in getting MAC-values enacted into law. US industrial medicine changed to MAC values in 1937. Other countries followed after WWII. The West German Association for Industrial Safety set up a MAC committee in 1954 (Bächi, 421 [42]). Just one year later the senate of the German Research Council established a commission on materials with adverse health effects in workplaces as an advisory body for government authorities (Bächi, 422 [42]). The enactment of MAC-values as litigable criteria in accordance with insurance law signified a shift toward risk assessment based on probabilistic concepts with a statistical understanding of causality in industrial medicine (Lengwiler, 155 [23]). The statistical understanding and probabilistic assessment of health risk in industrial medicine proved to be a useful and enduring point of departure for the development in social and preventative medicine that began in the interwar era but gained momentum only in the post World War II era (Lengwiler, 155–158 [23]).

3.5 *Controlling Quality via Statistics: Quantitative Approaches to System Safety and Reliability in the Bell Telephone System*

Whereas the insurance trade pioneered the quantitative understanding of risk, problems of electrical engineering gave rise to quantitative approaches to system safety and reliability that were to constitute an important building block for the emergence of Probability Risk Assessment (PRA) in various fields of complex engineering systems, and thus they contributed decisively to the evolving intellectual core of scientific risk research. Both, the increasing scale of mass production and the growing size, complexity, and interdependencies of large technical systems challenged the hitherto common ways of assuring the safety and reliability of those systems. Because the reliability of a technical system depended on the manufactured quality of each part, it soon became clear that quality control for the millions of components in these rapidly expanding systems would become a bottleneck for warranting the safety and reliability of those systems. American Telephone and Telegraphy (AT&T, owner of what was simply called "the Bell system" until 1984) was the first company to tackle this new challenge. In the second decade of the 20th century the company concluded that future growth depended on the geographical extension of telephone service (Miranti, 51 [78]). To meet this challenge the company had to improve transmission quality. One way for quality improvements led through innovations in the quality inspection regime. George A. Campbell, a MIT and Harvard trained electrical engineer who also studied advanced mathematics under Felix Klein in Göttingen and electricity and magnetism under Ludwig Boltzmann in Vienna, pioneered the introduction of probability-based techniques in the Bell system for positioning loading coils on transcontinental telephone lines. Around 1924 he strongly encouraged his colleagues to also use probability theory in confronting uncertainties related to management problems (Miranti, 55–56 [78]). A pioneer in industrially applied probability theory, Campbell called for developing a common knowledge base—industrial mathematics (Campbell [47]). As early as 1925 Bell Telephone Laboratories did indeed follow this advice: they established a Mathematical Research Department, headed by applied mathematician Thornton C. Fry, who in 1928 published his widely received text, Probability and its Engineering Uses [16].

Bell Labs' research statistician, W.A. Shewhart, recognized the usability of statistics as a scientific approach toward improving the quality control regime of the company's equipment manufacturing operations. He suggested analyzing product-defect distributions with the help of the properties of the bell-shaped normal (i.e., Gaussian) curve. According to Miranti, Shewhart "defined manufacturing control in terms of acceptable levels of variance, measured in standard deviations, from the mean number of deviations in a product lot" (Miranti, 60–61 [78]). This proved to be the decisive point of departure for the subsequent development and introduction of Statistical Quality Control (SQC) in the Bell system—and eventually beyond it (Shewhart [36]). With the advent of the Great Depression, when AT&T's labor force shrank and manufacturing inspection teams dwindled, the company recruited more graduates with strong mathematical backgrounds. This boosted the full exploitation of SQC in the Bell system's factories and elsewhere in the company's operations (on the history of SQC see also Juran [70]).

Not only in the US but also in Germany SQC came into being in the inter-war-period. Here, Karl Daeves, the head of the research laboratories of the Rhenish Steelworks, developed the method of SQC to control variations in steel production. Daeves called his method "Großzahlforschung" (large number research) and praised it as a way to replace the "doubtfully intuitive information that is based on subjective experience, by statistical values of objectified experience" (Daeves [8]). Via an analysis of frequency distributions with the help of probability graph papers that Karl Daeves developed together with the food chemist August Beckel in the early 1930s, these industrial researchers laid much of the groundwork for the use of probability theory in industry (Daeves and Beckel [9]). In Germany and the US alike the method of Großzahlforschung received the most attention in the electrical industry. Industrial researchers of the German electric light bulb producer Osram and the giant of the electrical industry Siemens collaborated with well-known professors from the Technische Hochschule Berlin in a lecture series on SQC during the winter term of 1928–1929 and again at the beginning of 1936. The Nazis hampered these fruitful beginnings when they forced leading practitioners and promoters of industrial mathematics, and mathematical statistics especially, to flee from the anti-Semitic regime (Tobies, 190 [39]). In contrast to Germany, the US state encouraged industrial mathematics when the National Defense Research Committee established the Applied Mathematics Panel (AMP) at Columbia University in 1942. As an appointed member of AMP, the Romanian-Austrian mathematician Abraham Wald (1902–1950) developed the statistical technique of sequential analysis in 1943. An important development in SQC theory and method, sequential analysis allowed reduction in the number of random samples necessary to maintain quality control in armaments production, thereby increasing manufacturing productivity and saving the US state a lot of money (Morgenstern, 183–192 [26]).

The multiple efforts to develop SQC paved the way for the new profession of quality engineering. During World War II, Bell engineers transferred knowledge of SQC to war industries, and the US Department of Education and the War Production Board set up training courses. By 1946 the number of newly trained quality engineers had reached a critical mass, which resulted in professionalization; that is, the American Society for Quality Control was founded in 1946 and more than 2000 professionals attended the organization's first technical conference in 1947 (Miranti, 67 [78]). In postwar Europe quality control was pushed via the Marshall Plan and subsequent recovery programs. The largely US-funded European Productivity Agency initiated the establishment of the European Organization for Quality Control in 1956, which was allied with the American Society for Quality Control. In the same year the German journal "Qualitätskontrolle" appeared for the first time. It changed its title into “Qualität und Zuverlässigkeit” (quality and reliability) in 1970 (Masing, 411–415 [74]).

Growing imperatives for reliability in weapons systems during the Cold War arms race led to further extensions of probabilistic quality control and gave rise to Reliability Engineering and quantitative reliability analysis. For example, in the early 1950s the US Department of Defense commissioned a study on how to increase the reliability of one of the most ubiquitous but also most failure-prone components of military electronics—the vacuum tube (Stott et al. [94]). Issued in 1957, this so-called AGREE (Advisory Group on Reliability of Electronic Equipment) Report furthered the development of quantitative reliability analysis and constituted an important building block for the emergence of Probabilistic Risk Assessment (PRA). It will come as no surprise that electrical engineers, who were well-grounded in probability theory, contributed significantly to this development.

4 Hot and Cold War, Large Technological Systems and Safety Concerns: Tackling Uncertainties via New Knowledge and Methods of Assessing Risks

The World War II experience changed people’s attitudes toward risk and uncertainty in quite contradictory ways. Having survived the Second World War and the deadly Nazi regime, some people emerged with confidence that contingencies could be controlled and the world changed for the better. Economists claimed to apply the right instruments to stabilize the equilibrium of markets. Keynesianism promised full employment. Bretton Woods re-established the stability of the gold standard of the 19th century. The International Monetary Fund and the World Bank promised economic advancement for the developing world. The United Nations was set up to secure peace and progress around the world. Engineers lined up not just to do away with the enormous destruction and rubble of the war but also to improve the safety of technology. With the development of more and more large and complex technological systems, the tasks of improving systems’ reliability—and thus of increasing safety—triggered new approaches to risk management. Governments strove toward political stability based on improved welfare systems and the transition toward mass consumption. This also included the responsibility that was felt on the part of the governing parties and administrations to protect populations from environmental, health and technological risks. Since the 1950s national legislation enacted new regulations, e.g., Food Additive Amendments to improve food safety, regulations for radiation safety, and new laws to increase highway and motor vehicle safety.

Thus, we find an ambivalent situation in the first two decades after the war. There was, on the one hand, great confidence that uncertainties could be controlled and risks assessed. This confidence was based on the assumption that everybody would behave rationally, an assumption that proved to be fertile ground for the spread of new concepts and methods to deal with future uncertainty. One of those methods was game theory, developed by the eminent mathematician John von Neumann and economist Oskar Morgenstern prior to and during World War II (von Neumann and Morgenstern [27]). Game theory became a highly used analytical tool during the Cold War, and major developments proceeded as its use spread, the Nash equilibrium being perhaps the most important. By the end of the Cold War, game theory had come to dominate scholarship in economics and had spread to many areas where analysis of present and future decisions in contexts of uncertainty must be made.[3] On the other hand, increased confidence went together with an increased awareness of and greater attention to potential dangers and perils. And there was good reason for increased concern, as the war had brought into being technologies with hitherto unknown potential dangers. One case in point was nuclear technology.

4.1 Nuclear Technology as New Challenge to Deal with Problems of Safety and Risk

When US President Dwight D. Eisenhower announced the decision of his administration to promote peaceful uses of atomic science and technology on an international scale in his famous Atoms for Peace speech in front of the United Nations' General Assembly on December 8, 1953, nuclear-fuelled power plants ranked high on the agenda of desirable peaceful applications of the atom (see Eisenhower's "Atoms for Peace" Speech [80]). Consequently, the Atoms for Peace initiative prompted national governments of many countries as well as international institutions under the aegis of the United Nations and the Organisation for European Economic Co-Operation to establish programs for the use of atomic energy in many domains.[4] However, the paradoxically overheated Cold War expectations about the seemingly unlimited potential of nuclear technologies could not erase the fear—fuelled by the atomic bombs dropped on Hiroshima and Nagasaki—that the power of the atom would have lethal effects when chain reactions ran out of control and when humans were exposed to ionizing radiation from fissionable materials. Imagining nuclear accidents and estimating potential damage became a major issue as the question arose as to who would assume liability for private nuclear power plants in case of an accident. With no historical knowledge of reactor safety and with seemingly unlimited liability should a reactor blow up or "melt down", the US insurance industry was unwilling to underwrite insurance risks for private nuclear energy. This

[3] Even a short history of game theory is beyond the scope of this chapter, but the interested reader should consult the following work: Poundstone [79]. On game theory in the Cold War think tank RAND see Hounshell, 253–255 [66].

[4] On programs to put the peaceful atom in service of food and agriculture see e.g. Zachmann [101].

refusal threatened to delay the development of Eisenhower's "peaceful atom". Thus, in 1957 the US Congress passed the Price-Anderson Act by which "the federal government provided insurance to cover losses above the $60 million private insurers were willing to cover (under considerable federal pressure), up to a total of $560 million" (Carlisle, 931 [6]). The government intended the law to be in force for only ten years, as it assumed that major safety improvements would occur and sufficient nuclear power plant operating data would be accumulated so that the state could withdraw and leave the field to private insurers. That, however, did not happen. Instead, the act was reinstated several times. Still in 2005 the Bush administration and Congress renewed the Price-Anderson Act as part of the Energy Policy Act of 2005 and extended it for the hitherto longest period of 20 years till 2025 (Price-Anderson Amendment Act [81]).

But how did engineers think about risk? Here two approaches had been prevalent, a deterministic and a probabilistic approach. The difference resulted from different engineering cultures. Chemical engineers from Du Pont who designed and built the first three plutonium production reactors at Hanford, Washington, took the deterministic approach. They explored potential component failure step by step and sought to determine what precautions needed to be taken to prevent such failure. In this approach, any effort to pre-calculate the mathematical probability of a component failure was completely absent. But as soon as the electrical engineers entered the nuclear power field in bigger numbers—US Admiral Hyman Rickover's Naval Reactor Program had opened the door—the probabilistic approach toward reactor risk gained ground. It was based in the electrical engineers' culture, as they saw the reactor as a product that "they fully thought out and put on paper before construction began" (Carlisle, 928 [6]). In this process, they calculated the probability of failure of crucial components. Increasingly available digital computer power increased the feasibility of such calculations. Thus, PRA in engineering emerged out of the professional culture of electrical engineers.[5] The two ways of thinking about risk set different priorities. Whereas deterministic engineering put physical problems and their remedies center stage but did not pursue any quantification, probabilism evaluated the reliability of entire complex systems as it calculated or estimated the likelihood of failure of crucial system's components.

The electrical engineers who introduced probabilistic methods into reactor design were able to build on an early tradition of probabilistic approaches that had already found fertile ground in the Bell system in the first half of the 20th century. Teachers such as Ernst Frankel also guided the electrical engineers. He taught at the Massachusetts Institute of Technology and wrote a textbook for a course on systems reliability that applied probabilistic thinking to complex systems (Carlisle, 926 [6]). He did what engineers and mathematicians at Bell system and elsewhere had envisaged since the 1920s when they explored the possibilities of applying probability theory to practical engineering problems (see e.g. Fry [16]).

[5] See the paragraph on SQC above.

4.2 Safety Engineering in the Aerospace and Defense Sector: Pioneering New Methods of Risk Assessment

Besides reactor design it was the aerospace and defense sector that fostered the application of probabilistic methods in safety engineering (Rip, 4 [83]). Beginning in the early 1960s fault trees became a commonly used technique that was applied for the first time in safety evaluations of the Launch Control System of the US Minuteman ICBM (Ericson [55]). Fault Tree Analysis is grounded in reliability theory, Boolean algebra and probability theory. The framework of FTA for analyzing very complex systems and complex relationships between hardware, software, and humans is comprised of a basic set of rules and symbols. FTA's initial development is ascribed to Bell Labs' researcher Hugh A. Watson who graduated with a PhD in nuclear physics from MIT in 1949 and worked at Bell Labs afterwards. In 1961 Watson conceived of FTA in connection with a US Air Force contract to perform the above-mentioned study of the Minuteman Launch Control System (Ericson, 1 [55] and Haasl, 1 [63]). Boeing Aircraft Company engineer David Haasl recognized the value of Watson's new method and organized the application of FTA to the entire Minuteman Missile System. Other departments of Boeing got interested as well, and Boeing began to use FTA in the design of commercial aircraft. In assigning probabilities to the events or component failures involved, the aerospace engineers aimed at calculating the overall probability of system failure in advance of use. In 1965 Boeing collaborated with the University of Washington in holding the first System Safety Conference. The rapid spread of FTA, however, stemmed mostly from the fact that it emerged in the very heated Cold War context of nuclear weapons systems development. Relying upon a policy called Mutually Assured Destruction (MAD) from the ever-growing number of atomic and thermonuclear weapons, the US believed the new Cold War imperative was to control systems safety of its increasingly potent weapons delivery systems. Already in 1950 the Air Force had established a Directorate of Flight Safety Research that was to be followed by a safety center of the Navy in 1955 and the Army in 1957 (Ericson [56]). In the late 1950s system safety began to be perceived as a new engineering discipline. That the military was its midwife became obvious with the publication of a document entitled "System Safety Engineering for the Development of United States Air Force Ballistic Missiles" in 1962 (Dhillon, 265 [52]). FTA's primary contribution to this development was its probability-based quantitative technique for analyzing system safety and reliability of space and defense systems. Improved FTA methods were developed, thanks to advances in both statistics and digital computer applications (Ericson [55]). The Department of Defense soon built FTA into specifications for all its weapons systems development contracts.

In the midst of the first wave of FTA-hype, however, the National Aeronautics and Space Administration (NASA) refrained from quantitative approaches to risk and safety analysis. John Garrick, a pioneer in nuclear risk assessment and a leading figure of the US risk analysis community, has retold the events as follows: "The time is remembered as about 1960, and the event was a bad experience with a probability calculation on the likelihood of successfully getting a man to the moon and back. The calculation was very pessimistic and embarrassing to NASA officials and soured them on the utility of probability calculations. From that point forward, NASA chose not to do probability, that is, quantitative risk and safety analysis, on their space systems. Rather, they adopted a qualitative approach utilizing Failure Mode and Effects Analysis (FMEA) as the principal building block for their risk analysis program" (Garrick, 1 [60]. For information on Garrick see Profile [82]). Only after the Challenger accident on January 28, 1986, did NASA re-visit its earlier decision and integrate quantitative risk assessment into its systems safety management processes (Garrick, 3–7 [60] and [18]). Meanwhile, however, NASA's preference for the qualitative FMEA concept pushed its development and made the qualitative approach toward systems safety attractive to other circles. The automobile industry took it up in the late 1970s when the Ford Motor Company adapted the method after the Ford Pinto debacle in which the company's hitherto unremarkable small car had to be recalled because of safety concerns related to the location and integrity of its gas tank (Tietjen and Müller [95]). From the automobile industry FMEA spread to other branches, became more diversified methodologically, and eventually developed as a risk-mitigating tool that became a standard element of prevention strategies.[6] The food industry developed its own version of FMEA even before the automobile makers when, during the Apollo moon program, NASA established new safety requirements for the astronauts' diet. The food company Pillsbury was the prime contractor for the space food program and adapted military experiences of critical control point (CCP) identification and FMEA into what became known as "Hazard Analysis and Critical Control Point System" (HACCP) for food safety in the early 1970s (Sperber and Stier [92]).

By no means did NASA's initial rejection of probabilistic risk assessment in favor of more qualitative approaches to safety result in any serious setback for Probabilistic Risk Analysis. By the late 1960s and early 1970s PRA was moving swiftly toward broad acceptance, thanks especially to developments in the nuclear sector. In turn, the perceived success of PRA boosted the professionalization of risk research and risk communication.

4.3 The Rasmussen Reactor Safety Study as Contested and Yet Celebrated Breakthrough of Probabilistic Risk Assessment

A decisive event in this process was the Rasmussen report, a reactor safety study that made extensive use of fault tree analysis and probabilistic techniques for estimating and quantifying risks (Rasmussen [33]). In 1972 the US Atomic Energy Commission (AEC) set up a new panel, headed by MIT engineering professor Norman R. Rasmussen, to evaluate the safety of nuclear reactors. The new head of the embattled AEC, James Schlesinger, aimed at presenting the AEC as a referee between the

[6] On FMEA and FTA as methods to increase dependability in engineering systems today see Vogel-Heuser and Straub in this book.

nuclear industry and an increasingly concerned public; therefore he strove to mitigate heightened safety concerns (Walker, 41–41 [96]). The latter had been voiced, e.g., by the recently founded Union of Concerned Scientists, which criticized how the AEC had dealt with unsettled questions about deficiencies in emergency core cooling systems in the AEC's licensing procedures (Walker, 33 [96]). More safety concerns arose as a result of the growing environmental movement, especially concerning thermal pollution, the effects of low-level radiation from routine operation of nuclear power plants, and the risks posed by high-level radioactive waste storage and disposal. Thus, the Rasmussen panel's task to assess accident risks in US commercial nuclear power was bound up with high expectations on the part of the AEC. The study was to demarcate the field the AEC felt responsible for—reactor safety—in advance of the pending renewal of the Price-Anderson Act (Carlisle, 931 [6]). When in October 1975 the US Nuclear Regulatory Commission (the AEC's successor regulatory agency) presented the final Rasmussen report to the public, the report immediately won a lot of attention. This was largely due to its scale and political significance, but also to its extensive use of probabilistic techniques. It must be stressed, though, that the Rasmussen report was by no means the first study to apply probabilistic approaches in the assessment of technical risks. As already noted, physicists, electrical engineers, and aerospace engineers had done so earlier to varying degrees and in various contexts (Carlisle, 933 [6]). Nevertheless, the Rasmussen report made a pioneering contribution because it introduced a general public of non-specialists to the application of probabilistic techniques in reactor safety studies based on fault trees and other forms of probabilistic risk analyses. Furthermore, the Reactor Safety Study made use of Monte Carlo simulations that had come into being in the context of the development of thermonuclear and enhanced fission weaponry as a kind of lingua Franca among physicists, nuclear theorists, chemists, electrical engineers, mathematicians, statisticians and others for dealing with problems of mutual interest: nuclear atomic structure, molecular structure, equilibrium calculations, reaction rates, resonance energy calculations, shielding calculations, and the fitting of decay curves (Lee, Grosh, Tillman, and Lie, 198 [73]).[7] Monte Carlo simulation has become standard fare across a wide number of science, engineering, and social science disciplines and also in industries and the finance and insurance business.

Despite its achievements the Rasmussen report also received serious criticism. The Union of Concerned Scientists pointed to the fact that fault tree analyses had been developed in order to compare risks and to make decisions within the design process. Fault trees, it argued, were not suited for determining exact numerical probability data of accidents (Ford, 23 [14] and Öko-Institut, 18 [28]). Serious criticism was also uttered over the report's way of presenting risk. In order to guide the risk perceptions of the public, the Rasmussen report developed numerical measures to compare accident risks of reactors to more socially familiar risks, such as traffic accidents, dam breaks, and catastrophic fires. In doing so the Rasmussen panel introduced the criterion of acceptable risk, as it assumed that risks of nuclear reactors

[7] For an excellent historical interpretation of Monte Carlo simulations (see Galison, 689–780 [17]).

which lay within the range of risks of other technical systems—to which people had grown accustomed already—would be as easily accepted (Carlisle, 934–935 [6]). Not just the public but also internal staff from the Nuclear Regulatory Commission voiced serious doubts about the results of the Rasmussen report. In January 1979 the NRC went so far to issue a statement withdrawing its full endorsement of the report's executive summary (Walker, 49 [96]).

When on 28 March 1979 a serious accident at the Three Mile Island II nuclear power plant near Harrisburg, Pennsylvania, occurred that the Nuclear Regulatory Commission had not thought to be possible, the USA encountered a severe setback for the public acceptance of nuclear energy. The nuclear establishment responded to the TMI accident with a series of measures, such as, e.g., the setting-up of a database and reporting system for accidents and the introduction of PRA as part of the documentation in pending plant applications for licenses. Thus, PRA gained more ground, despite the initially harsh criticism of the Rasmussen report and even though the occurrence at TMI had proved that heavy reliance on fault tree analysis was inadequate for the assessment of nuclear accident risks. The NRC, e.g., subsequently required PRA as part of the licensing procedure for nuclear power plants (Walker, 51 [96]).

The Rasmussen report worked as catalyst of Probabilistic Risk Assessment not just in the USA but also abroad. The Federal Minister of Research and Technology in Germany, e.g., issued the first German reactor safety study in 1976, only a year after the publication of the Rasmussen report. In the midst of the first wave of anti-nuclear power protests, the minister felt obliged no longer to rely on American nuclear safety research but to entrust the newly founded Gesellschaft für Reaktorsicherheit GRS (Society for Reactor Safety) with conducting the first German risk study on nuclear power plants that would pay attention to German characteristics, such as specific German design and safety features and especially their location in far more densely populated areas compared to US plant sites (Der Bundesminister, 1–2 [45]).[8] The first German risk study, however, closely followed the methodology of the Rasmussen report. In their Festschrift for the 30th-anniversary of the GRS, the authors praised the risk study as the first probabilistic safety analysis that inaugurated the new instrument of probabilistic safety assessment in Germany (GSR, 9 [61]). Only a few years earlier, however, probabilistic approaches had still met with resistance in many parts of Germany. In 1966, the head of the laboratory of nuclear power control and plant safety at the Technical University Munich, Professor Adolf Birkhofer, who was to become the managing director of GRS in 1977 and would keep that position till 2002, belittled probabilistic safety research as passing fashion (Radkau, 361 [32]). The mentor of Birkhofer's Habilitation, Ludwig Merz, who was an expert on measurement and control engineering and responsible for

[8] According to Radkau, the first German research program on reactor safety was instituted by the Minister of Research and Technology only in 1971. It was triggered by the project of BASF to establish a nuclear power plant in Ludwigshafen and thus near big cities. This project was abandoned in 1972 (Radkau, 381–382 [32]).

the instrumentation of the first German-designed research reactor (FR-2 in Karlsruhe), repeatedly insisted on deterministic approaches as more appropriate or at least equally important in reactor safety research (Merz [76, 77]). As head of GRS and thus responsible for the first German nuclear power plant risk study, however, Birkhofer changed his mind and subscribed to Probabilistic Risk Assessment.

The timing for publication of the German risk study coincided with the accident at the Three Mile Island nuclear power plant. The GRS managed this situation by adding an analysis of the nuclear accident in Harrisburg, PA, as an appendix to the main study. Here the authors concluded that the events in TMI did not undermine but rather confirmed the results of the risk study (Der Bundesminister, 265–257 [45]). At the same time, however, the authors already envisioned a "phase B" of the risk study that would reveal internal safety-relevant weak points, whereas phase A had analyzed accident-caused damage outside of nuclear power plants and especially the dimension and frequency of health damage to the population (Der Bundesminister, 245–247 [45] and 6–7 [46]). Phase B was published in 1989, the same year the last two German nuclear power plants were connected to the nation's electric grid. The risk studies had not mitigated the public's safety concerns about nuclear power, and after the turn of the millennium the German government decided to abandon nuclear energy altogether.

By the mid-1980s in the US and elsewhere PRA had become, as Carlisle framed it, "part of the safety orthodoxy" and an object of Gierynian "boundary work", leading to the formation of professional risk research organizations (Carlisle, 938 [6] and Gieryn [62]). This was true not just for the nuclear sector. As we have mentioned above, after 1986 NASA returned to PRA. Also the chemical and petroleum industry developed an increased interest in PRA after major accidents at Flixborough in England, Seveso in Italy, and Bhopal in India. The Bhopal accident, especially, triggered greater activities in risk and safety research and its applications in the chemical industry (Garrick, 197 [18]). Thus, since the mid-1970s and especially during the 1980s PRA emerged as a new business. Private firms performed PRA on nuclear power plants, chemical plants, transportation systems, space systems, and defense systems (Profile, 936 [82]. The practitioners of quantitative risk assessment developed new ways of thinking about risk and safety. PRA became the intellectual core for the emerging community of risk research that began to organize itself in the late 1970s.

4.4 *Swelling Uncertainties in the "Epoch of Landslide" and the Mobilization of Professionalized Research to Deal with New Risks*

That reactor safety studies and PRA made headlines in the media and fired public controversies in the 1970s signalled changing attitudes toward uncertainties and risks. The post-World War II optimism that uncertainties can be controlled and transformed into calculable risks that would allow humans to make wise decisions was superseded by new concerns because of newly emerging uncertainties. Increasing environmental concerns spread as indicated by the growing amount of readers of Rachel Carson's book Silent Spring [48] and the publication of the Meadows et al. report, Limits to Growth [75]. Growing fears of a deteriorating state of the earth stemming from industrial activities and economic growth, however, were not the only cause of concern. Wars and political unrest, uprisings and scandals in all parts of the world, reaching from the war in Vietnam via the increasingly violent conflicts in the Arab peninsula and the crushed Prague Spring up to the Watergate scandal revealed a fragile political state and the weakness of the United Nations in fulfilling its task of securing peace and progress across the community of peoples. Other forces were also unleashed. Economies crumbled when oil prices skyrocketed and the Bretton Woods Agreements broke down. In the wake of these economic storms, structural changes gathered speed, putting an end to full employment and undermining faith in Keynesianism. This was a period that Eric Hobsbawm called the years of landslide. These years historians only recently began to define as an epochal threshold, leading to an era "after the boom" (Hobsbawm, 502–720 [20] and Doering-Manteuffel and Raphael [53]). In this context, late-modern societies developed a heightened awareness of uncertainties and a changing attitude toward risks, notwithstanding the fact that fundamental anthropometric data, such as longevity and body height, and world population counts, indicated fundamentally improved living conditions in many parts of the world (on improved living conditions see Fogel [58]).

Sociologists identified the risks in late modern societies as having a new character. According to Ulrich Beck new risks result from such sources as nuclear power plants, genetic engineering, and volatile capital markets (Beck, 11 [3] and Bonß [44]). These new risks are no longer completely known nor are they fully verifiable. To a certain extent, these and other new risks remain hypothetical. Managing these risks may produce unintended side effects. In temporal, material, and social respects, the risks of the late-modern world reveal a new dimension: potential damages can no longer be compensated with money. The nuclear reactor catastrophes of Chernobyl (1986) and Fukushima (2011) may be cited as proof. Thus, new risks are no longer considered as chances that can be taken based on confidence in a basic certainty but rather as threats that should be avoided based on a fundamental awareness of uncertainty. To be sure, Beck's diagnosis of the characteristics of late-modern risks is widely known, but other authors take different, less normative, and more analytical positions (Luhmann, 13–14 [24]). The success of Beck's book, however, supports his diagnosis as a relevant description of swelling uncertainty.

Swelling uncertainties triggered a tremendous boost in risk regulation and risk research. From the end of 1960s and the early 1970s, first in the US and shortly thereafter elsewhere, there were dramatic increases in the number of agencies implementing risk-related legislation that dealt with health, safety, and environmental concerns (Covello and Mumpower, 116–117 [50]; Jasanoff, 2–3 [21]; Thompson, Deisler, and Schwing, 1334–1336 [38]). Legislative mandates to protect the environment and public health and to ensure safety furthered new federal research centers and research programs in the US and elsewhere. As more researchers than ever before in a broader array of fields began to analyze risks, they developed a need for greater communication and interaction. Historical reports on the developing field of risk analysis underscore the importance of the 1975 multidisciplinary conference at Asilomar, CA, on the risks resulting from research on recombinant DNA molecules as one of the first meetings with risk as the main subject. The Asilomar Conference resulted in an interdisciplinary Recombinant DNA Advisory Committee that was to review all proposals for conducting rDNA research in order to prevent possible harm to human health and the environment through the unchecked spread of undesired genes (Jasanoff, 47 [69]). In 1979 another early, interdisciplinary, and explicitly risk-related meeting was organized by two General Motors Laboratory researchers as part of the General Motors symposia series under the title: "How Safe is Safe Enough?" (on the conference see Thompson, Deisler, and Schwing, 1335–1336 [38]). The conference gathered together experts from many disciplines—as diverse as anthropology and nuclear physics—and it was opened by Chauncey Starr, whose 1969 article, "Social Benefits versus Technological Risk: What is Our Society Willing to Pay for Safety", was considered by many as a landmark in risk research.[9] Thus, by the late-1970s, risk had become a subject of research that—as Sheila Jasanoff highlighted—connected disciplines as different as "mathematics, biostatistics, toxicology, and engineering on the one hand and law, psychology, sociology and economics on the other hand" (Jasanoff, 123 [68]). In their preference for either quantitative, model- and measurement-oriented approaches, or qualitative investigations as to the ethical, legal, political, and cultural aspects of risk, the researchers remained confined to the two cultures of science.[10] Jasanoff, however, did not stress the differences but the complementarity of the two cultures of risk analyses (Jasanoff, 124 [68]).

Common problems encountered across many disciplines requiring probabilistic calculation led a range of researchers to contemplate developing risk analysis as an academic discipline that would hasten the professionalization of risk research. In 1980 they founded the Society for Risk Analysis (SRA) and began publishing its journal, Risk Analysis, in 1981, which provided a forum for both debate about professionalization and new research on risk analysis. Robert B. Cumming is reputed to have been the "spiritus rector" for establishing the new society and its journal (Thompson, Deisler, and Schwing, 1336 [38]). As member of the Environmental Mutagens Society and a genetic toxicologist in the Biology Division of Oak Ridge National Laboratory in Tennessee, Cumming had been one of the participants at the Asimolar Conference and other meetings on risk research, and thus he knew the emerging community of risk analysts quite well. In the first issue of Risk Analysis, Cumming included an editorial posing the question: "Is Risk Assessment a Science?" (Cumming [7]). Cumming answered "no". Instead, he warned explicitly

[9] The article was published in Science 165, 1232–1238. Thompson, Deisler, and Schwing, 1334 [32] praised it as providing "the basis for approaching risk issues systematically and quantitatively and (introducing) the concept of tradeoffs between risks and benefits for a wide range of risks".

[10] For an extensive and knowledge-able overview on the disciplinary perspectives on risk see Althaus, 567–588 [1].

against "dangers of professionalism" because these aspirations would serve only special interest groups but not the community of risk researchers as a whole. He envisaged the main purpose of the new society and its journal as "providing better communication among the diverse elements involved in risk management", i.e. the whole range of contributing scientific disciplines as well as political and social institutions (Cumming, 2 [7]). The author of the second article in the same issue, Alvin Weinberg, the distinguished nuclear and bio-physicist with research and policy experiences going back to the Manhattan Project, spoke on "the art of risk assessment" in order to distinguish it from science (Weinberg [98]). He pointed to strong trans-scientific elements in risk assessment, and thus referred to an idea of thinking on science and ignorance that he had developed a decade before. In 1972 he had introduced the term trans-scientific for "questions which can be asked of science and yet which cannot be answered by science" (Weinberg [97]). As examples of trans-scientific questions he named among others the biological effect of low-level radiation exposure or the probability of extremely improbable events such as catastrophic reactor accidents. Risk analysis was fundamentally important in addressing trans-scientific questions, but its practitioners could by no means claim absolute authority in offering answers.

Notwithstanding the hesitant stance of its founders, SRA both fostered and tracked many activities toward developing risk analysis into a coherent academic discipline with well-defined educational programs from the undergraduate up to the postgraduate level (Thompson, Deisler, and Schwing, 1380–1381 [38]). But the desired coherence was hard to achieve. This becomes clear with regard to the unsuccessful strivings to find a common definition of risk on which all members of the risk community could agree. In the mid-1980s, SRA tried to tackle this problem by setting up an Ad Hoc Definitions Committee that, about a decade later, finally settled the question by providing a list of definitions on the society's website without officially endorsing any one of them (Thompson, Deisler, and Schwing, 1380 [38]). Another indicator of the great diversity of the risk research community is the emergence of other, more specialized societies that are focusing on risk, such as, e.g., the Society of Environmental Toxicology and Chemistry (1979), the International Society of Regulatory Toxicology and Pharmacology (1984), the Association of Environmental Health Sciences (early 1980s), the International Society of Exposure Analysis (1989), the International Association for Probabilistic Safety Assessment and Management (1991), and the Risk Assessment & Policy Association (1994) (cf. Thompson, Deisler, and Schwing, 1347 [38]). Thus, risk research blossomed much more as an interdisciplinary rather than a disciplinary endeavor.

5 Food for Thought

The great societal transformation of the 19th century involved changing attitudes toward risk. As soon as the urban middle class of professionals and tradesmen became entitled to vote and acquired more social responsibilities, both in the public and the private realms, they subscribed to an ethos of control and predictability and began seeking ways to avoid risks. The burgeoning economic life of the industrial revolution, however, required the entrepreneurial men of the middle class to take risks because setting up businesses involved calculating on an uncertain future. How were these contradictory attitudes toward risk reconciled in Western societies of the 19th century?

This chapter has been concerned only with developments in the Western world and has shed light on events and processes that signified shifts in concepts of risks in Great Britain, Germany, and the United States mainly. From anthropological studies, however, we have learned that culture matters in determining approaches toward risk. How do non-Western cultures experience risk, and how do these differences affect economic, financial, technological, political, and military endeavors in an increasingly globalized world?

Life insurers were the first within the broader insurance business to develop and apply theoretical knowledge to underwriting insurance policies. Practitioners in the non-life insurance trades, however, preferred empirical knowledge and experience for estimating risks and rating insurance premiums well into the second half of the 20th century, even though the Swedish actuary Filip Lundberg had published a theory of risk in 1909. Why did it take so long for probability theory to be applied in non-life insurance?

Societies developed multiple ways to deal with risks, such as developing new knowledge on dangers, imposing legally binding safety norms, developing quality and reliability engineering, requiring risk analysis of safety-critical ventures, demanding compulsory insurance for activities in danger zones, and many other measures. How did societies decide on the most appropriate means of risk management, and to what extent did professional cultures and intellectual fashions influence such decisions?

Probabilistic Risk Assessment got a boost via the Rasmussen Report, the reactor safety study that was presented to the US public in 1975. Paradoxically, the study gained acceptance only after the severe accident at the Three Mile Island nuclear power plant in 1979, although the study's calculated probability of such an accident occurring was far too low to seem plausible to policy makers and regulatory authorities. Why and how did Probabilistic Risk Assessment become an object of "boundary work" that was soon applied to many areas beyond the nuclear power sector and helped to propel forward the formation of the risk research community?

6 Summary

In this chapter we have investigated the changing concepts of and attitudes toward risk. As a point of departure, we used Luhmann's reflection on risk as future uncertainty that is caused by human-made decisions, a notion based on the economist Frank Knight's more well known concept of risk as calculable uncertainty. The questions we sought to answer were framed as follows: when did the attitude toward future uncertainties change so that our understanding of uncertainties became narrowed down to risk? How did the modern concept of risk determine people's ways of dealing with uncertainties? How widely accepted has modern risk analysis become, in what ways has such analysis proved to be particularly problematic, and in what manner has risk analysis become professionalized?

Proto-modern notions on risk emerged in the context of the early modern rage for gambling and other forms of aleatory contracts that treated future uncertainties as a chance to make a fortune and therefore worth a wager. These aleatory contracts inspired mathematicians to develop calculations on the probable outcome of future events; thus, they began to quantify uncertainty as probability. The mathematicians' achievement, however, was incompatible with the gamblers' proto-modern understanding of risk as genuine uncertainty that precluded quantification. Therefore, the understanding of risk had to be changed before probability calculations could find acceptance as a way to manage uncertainties. The great political, technological and social transformation of Western societies ushered in this development as the attitudes toward risks now changed from something to be sought into something to be avoided—or at least managed. When bourgeois values of familial responsibility, control, and predictability began to determine the norms of society, its citizens strove to gain control over uncertainties. In this context, however, they developed a heightened awareness of a new class of human-wrought dangers and threatening uncertainties. In this chapter we have explored how steam boiler explosions, food adulteration, and cholera epidemics were not endured in fateful resignation but gave rise to modern modes of risk management. The agents of this development established regulations based on newly produced technical knowledge, formed coalitions of experts among a broad range of fields, and introduced standards of safety for technologies as well as for food. These strategies of risk management aimed at preventing individual and societal harm from human-made hazardous products and environments. At the same time, the advancing insurance system of the 19th century promised to compensate victims for their harmed bodies and damaged properties. Thus, insurers capitalized on risk as they sold their customers a new degree of control over uncertainty. As the success of the insurers' business largely depended on knowledge of how to assess and to manage risks, insurers were important promoters of research as to the causes and prevention of risks. Except for life insurers, however, practitioners in most of the other fields of the insurance trade proved to be quite reluctant to employ theoretical approaches to risk calculation.

Only after World War II did the need for a more systematic and mathematically rigorous risk analysis encourage the statistical understanding and probabilistic assessment of risks in fields beyond life insurance. Whereas the insurance trade led in developing a quantitative understanding of risk, problems of electrical engineering gave rise to quantitative approaches toward system safety and reliability that were to constitute an important building block for the emergence of Probability Risk Assessment (PRA) in the engineering fields. These developments occurred at the intersection of increasingly complex, large-scale technological systems and the establishment of formal organizations in which advanced mathematically-based science and engineering knowledge was produced and applied to those systems. We traced the quantitative approaches of system safety and reliability back to statistical quality control in the Bell Telephone System, where it was developed beginning in the 1920s.

Formal methods of PRA emerged out of the need to determine the safety of nuclear reactors. Reactor safety became a hot-button issue when the success of Eisenhower's Atoms for Peace Cold War initiative hinged upon the acceptance and subsequent spread of nuclear power. In addition to reactor design and operation, the aerospace and defense sectors also fostered the application of probabilistic methods in safety engineering. Here Fault Tree Analysis was developed and introduced for safety evaluations of the Launch Control System of the US Minuteman ICBM; avoiding an accidentally initiated thermonuclear World War III thus served as what economist Nathan Rosenberg has termed a "focusing device" for innovation in risk research and analysis (Rosenberg [86]). As a probability-based quantitative technique for analyzing system safety and the reliability of space and defense systems, the Department of Defense built FTA into specifications for weapons systems development contracts. In order not to jeopardize its Apollo moon program, however, the civilian National Aeronautics and Space Administration decided to refrain from quantitative risk and safety analysis, adopting instead a qualitative approach using Failure Mode and Effects Analysis as the principal building block for the agency's risk analysis program. This qualitative approach toward systems safety also became attractive to other circles, such as the automobile industry and the food industry. However, despite NASA's initial rejection, PRA gained further ground and boosted the application and further development of risk research and risk communication.

As we have seen, the 1974 Rasmussen report, a reactor safety study that made extensive use of fault tree analysis and probabilistic techniques for estimating and quantifying risks, proved to be decisive for the spread and acceptance of PRA. This was the case not just in the USA but also abroad, e.g., in the Federal Republic of Germany. By the mid-1980s PRA had become an object of "boundary work", furthering professional risk research communities and spreading across the nuclear sector to a whole range of problems and applications. With more sophisticated approaches toward assessing risk, however, the awareness of new risks also increased. Since the 1980s ever more disciplines are contributing to this truly interdisciplinary endeavor, and thus are expanding and deepening the approaches to analyzing, communicating, and managing risks.

References

Selected Bibliography

1. C. Althaus, A disciplinary perspective on the epistemological status of risk. Risk Anal. **25**, 567–588 (2005)
2. U. Beck, *Risikogesellschaft. Auf dem Weg in eine andere Moderne* (Suhrkamp, Frankfurt, 1986)
3. U. Beck, *Weltrisikogesellschaft. Auf der Suche nach der verlorenen Sicherheit* (Suhrkamp, Frankfurt, 2007)

4. W. Bonß, *Vom Risiko. Unsicherheit und Ungewissheit in der Moderne* (Hamburger Edition, Hamburg, 1995)
5. J.G. Burke, Bursting boilers and the federal power. Technol. Cult. **7**, 1–23 (1966)
6. R. Carlisle, Probabilistic risk assessment in nuclear reactors: engineering success, public relations failure. Technol. Cult. **38**, 920–941 (1997)
7. R.B. Cumming, Is risk assessment a science? Risk Anal. **1**, 1–3 (1981)
8. K. Daeves, *Großzahlforschung. Grundlagen und Anwendungen eines neuen Arbeitsverfahrens für die Industrieforschung mit zahlreichen praktischen Beispielen* (Stahleisen m.b.H., Düsseldorf, 1924)
9. K. Daeves, A. Beckel, *Großzahl-Forschung und Häufigkeits-Analyse. Ein Leitfaden* (Verlag Chemie, Weinheim, 1948)
10. L. Daston, The domestication of risk: mathematical probability and insurance 1650–1830, in *The Probabilistic Revolution*, ed. by L. Krüger, L.J. Daston, M. Heidelberger. Ideas in History, vol. 1 (MIT Press, Cambridge, 1987), pp. 237–260
11. L. Daston, *Classical Probability in the Enlightenment* (Princeton University Press, Princeton, 1988)
12. M. Douglas, A. Wildavsky, *Risk and Culture. An Essay on the Selection of Technical and Environmental Dangers* (University of California Press, Berkeley, 1982)
13. F. Ewald, *Der Vorsorgestaat* (Suhrkamp, Frankfurt, 1993)
14. D. Ford, *A History of Federal Nuclear Safety Assessments: From WASH-740 Through the Reactor Safety Study* (Union of Concerned Scientists, Cambridge, 1977)
15. J.-B. Fressoz, Beck back in the 19th century: towards a genealogy of risk society. Hist. Technol. **23**, 333–350 (2007)
16. T.C. Fry, *Probability and Its Engineering Uses* (Van Nostrand, New York, 1928)
17. P. Galison, *Image and Logic* (University of Chicago Press, Chicago, 1997)
18. J. Garrick, The approach to risk analysis in three industries: nuclear power, space systems, and chemical process. Reliab. Eng. Syst. Saf. **23**, 195–205 (1988)
19. G. Gigerenzer et al., *The Empire of Chance. How Probability Changed Science and Everyday Life* (Cambridge University Press, Cambridge, 1989)
20. E. Hobsbawm, *Das Zeitalter der Extreme. Weltgeschichte des 20. Jahrhunderts* (Deutscher Taschenbuch Verlag, München, 1999)
21. S. Jasanoff, *The Fifth Branch. Science Advisers as Policymakers* (Harvard University Press, Cambridge, 1990)
22. F. Knight, *Risk, Uncertainty and Profit*, 1st edn. (Houghton Mifflin, Boston, 1921)
23. M. Lengwiler, *Risikopolitik im Sozialstaat. Die schweizerische Unfallversicherung* (Böhlau, Köln, 2006)
24. N. Luhmann, *Soziologie des Risikos* (de Gruyter, Berlin, 1991)
25. R. Lukes, 150 Jahre Recht der technischen Sicherheit in Deutschland—Geschichtliche Entwicklung und Durchsetzungsmöglichkeiten, in *Risiko—Schnittstelle zwischen Recht und Technik*, ed. by VDE (VDE-Verlag, Berlin, 1982), pp. 11–43
26. O. Morgenstern, *Spieltheorie und Wirtschaftswissenschaft* (Oldenburg, Wien, 1963)
27. J.v. Neumann, O. Morgenstern, *Theory of Games and Economic Behavior* (Princeton University Press, Princeton, 1944)
28. Öko-Institut Freiburg (ed.), *Die Risiken der Atomkraftwerke. Der Anti-Rasmussen-Report der Union of Concerned Scientists* (Adolf Bonz, Fellbach, 1980)
29. R. Pabst, *Theorie und Methodenentwicklung bei der Versicherung technischer Risiken am Beispiel der Maschinenversicherung in Deutschland* (Diss., Fakultät Wirtschaftswissenschaft, TU München, 2011)
30. C. Perrow, *Normale Katastrophen. Die unvermeidbaren Risiken der Großtechnik* (Campus, Frankfurt, 1992)
31. I. Pfeffer, *Insurance and Economic Theory* (Richard D. Irwin, Boston, 1956)
32. J. Radkau, *Aufstieg und Krise der deutschen Atomwirtschaft 1945–1975. Verdrängte Alternativen in der Kerntechnik und der Ursprung der nuklearen Kontroverse* (Rowohlt, Reinbek bei Hamburg, 1983)

33. N.C. Rasmussen, Reactor safety study. An assessment of accident risk in U.S. Commercial Nuclear Power Plants (WASH-1400, NUREG 75/014). U.S. Nuclear Regulatory Commission, Washington, 1975
34. O. Renn, Three decades of risk research: accomplishments and new challenges. J. Risk Res. **1**, 49–71 (1998)
35. O. Renn, *Risk Governance. Coping with Uncertainty in a Complex World* (Earthscan, London, 2008)
36. W.A. Shewhart, *Economic Control of Quality of Manufactured Product*, 1st edn. (Van Nostrand, New York, 1931)
37. J.V. Simson, *Kanalisation und Stadthygiene im 19. Jahrhundert* (VDI, Düsseldorf, 1983)
38. K.M. Thompson, P.F. Deisler Jr., R.C. Schwing, Interdisciplinary vision: the first 25 years of the society for risk analysis (SRA), 1980–2005. Risk Anal. **25**, 1333–1386 (2005)
39. R. Tobies, *Morgen möchte ich wieder 100 herrliche Sachen ausrechnen*. Iris Runge bei Osram und Telefunken (Franz Steiner, Stuttgart, 2010)
40. G. Wiesenack, *Wesen und Geschichte der Technischen Überwachungsvereine* (Carl Heymanns, Köln, 1971)
41. K. Zachmann, P. Østby, Food, technology, and trust: an introduction. Hist. Technol. **27**, 1–10 (2001)

Additional Literature and Sources

42. B. Bächi, Zur Krise der westdeutschen Grenzwertpolitik in den 1970er Jahren: Die Verwandlung des Berufskrebses von einem toxikologischen in ein sozioökonomisches Problem. Ber. Wiss.gesch. **33**, 419–435 (2010)
43. P. Bernstein, *Against the Gods. The Remarkable Story of Risk* (Wiley, New York, 1996)
44. W. Bonß, (Un-)Sicherheit als Problem der Moderne, in *Handeln unter Risiko. Gestaltungsansätze zwischen Wagnis und Vorsorge*, ed. by H. Münkler, M. Bohlender, S. Meurer (transcript, Bielefeld, 2010), pp. 33–53
45. Der Bundesminister für Forschung und Technologie (ed.), *Deutsche Risikostudie Kernkraftwerke. Eine Studie zu dem durch Störfälle in Kernkraftwerken verursachten Risiko* (TÜV Rheinland, Bonn, 1980). Hauptband
46. Der Bundesminister für Forschung und Technologie (ed.), *Deutsche Risikostudie Kernkraftwerke Phase B* (TÜV Rheinland, Bonn, 1989). Available at http://www.grs.de/sites/default/files/pdf/Dt._Risikostudie_Kernkraftwerke_Phase_B.pdf
47. G.A. Campbell, Mathematics in industrial research: ‘selling’ mathematics to the industries. Bell Syst. Tech. J. **3**, 550–557 (1925)
48. R. Carson, *Silent Spring* (Houghton Mifflin, Boston, 1962)
49. A. Clow, N.L. Clow, *The Chemical Revolution: A Contribution to Social Technology* (Batchworth Press, London, 1952)
50. V. Covello, J. Mumpower, Risk analysis and risk management: an historical perspective. Risk Anal. **5**, 103–120 (1986)
51. P.-A. Dessaux, Chemical expertise and food market regulation in Belle-Epoque France. Hist. Technol. **23**, 351–368 (2007)
52. B.S. Dhillon, Systems safety: a survey. Microelectron. Reliab. **22**, 265–275 (1982)
53. A. Doering-Manteuffel, L. Raphael, *Nach dem Boom. Perspektiven auf die Zeitgeschichte seit 1970* (Vandenhoeck & Ruprecht, Göttingen, 2008)
54. N. Doorn, S.O. Hansson, Should probabilistic design replace safety factors? Philos. Technol. **24**, 151–168 (2011). Available at http://www.springerlink.com/content/818781xk76376m40/fulltext.pdf
55. C. Ericson, Fault tree analysis—a history, in *Proceedings of the 17th International System Safety Conference* (1999). Available at http://www.fault-tree.net/papers/ericson-fta-history.pdf

56. C. Ericson, A short history of system safety. J. Syst. Saf. **42**, 3 (2006). Available at http://www.system-safety.org/ejss/past/novdec2006ejss/clifs.php
57. R. Evans, *Tod in Hamburg. Stadt, Gesellschaft und Politik in den Cholera-Jahren 1830–1910* (Rowohlt, Reinbek bei Hamburg, 1990)
58. R.B. Fogel, *The Escape from Hunger and Premature Death, 1700–2100: Europe, America, and the Third World* (Cambridge University Press, Cambridge, 2004)
59. T.C. Fry, Industrial mathematics, in *Research—A National Ressource—II, Section VI, Part 4*, Washington, D.C. (1940), pp. 268–288
60. J. Garrick, Risk assessment practices in the space industry: the move toward quantification. Risk Anal. **9**, 1–7 (1989)
61. Gesellschaft für Anlagen- und Reaktorsicherheit (GRS), (ed.), *30 Jahre Forschungs- und Sachverständigentätigkeit* (GRS, Köln, 2007). Available at http://www.grs.de/sites/default/files/kum/festschrift_30_jahre.pdf
62. T. Gieryn, Boundary-work and the demarcation of science from non-science: strains and interests in professional ideologies of scientists. Am. Sociol. Rev. **48**, 781–795 (1983)
63. D. Haasl, Advanced concepts in fault tree analysis. Presented at system safety symposium sponsored by University of Washington and the Boeing Company, June 8–9, 1965, 1. Available at http://www.fault-tree.net/papers/haasl-advanced-concepts-in-fta.pdf
64. A.I. Hardy, Der Arzt, die Ingenieure und die Städteassanierung. Georg Varrentrapps Visionen zur Kanalisation, Trinkwasserversorgung und Bauhygiene in deutschen Städten (1860–1880). Technikgeschichte **72**, 91–126 (2005)
65. V. Hierholzer, *Nahrung nach Norm. Regulierung von Nahrungsmittelqualität in der Industrialisierung 1871–1914* (Vandenhoeck & Ruprecht, Göttingen, 2010)
66. D. Hounshell, The cold war, RAND, and the generation of knowledge, 1946–1962. Hist. Stud. Phys. Sci. **27**, 237–267 (1997)
67. D.A. Hounshell, J.K. Smith, *Science and Corporate Strategy. Du Pont R&D, 1902–1980* (Cambridge University Press, Cambridge, 1988), pp. 555–572
68. S. Jasanoff, Bridging the two cultures of risk analysis. Risk Anal. **13**, 123–129 (1993)
69. S. Jasanoff, *Designs of Nature. Science and Democracy in Europe and the United States* (Princeton University Press, Princeton, 2005)
70. J.M. Juran, Early SQC: a historical supplement. Qual. Prog. **30**, 73–81 (1997)
71. P. Koch, *Versicherungsgeschichte in Stichworten*. Schriftenreihe des Vereins zur Förderung der Versicherungswissenschaft in München e.V., vol. 32 (1988), pp. 1–16
72. A. Labisch, J. Vögele, Stadt und Gesundheit. Anmerkungen zur neueren sozial- und medizinhistorischen Diskussion in Deutschland. Arch. Soz.gesch. **37**, 396–424 (1997)
73. W.S. Lee, D.L. Grosh, F.A. Tillman, C.H. Lie, Fault tree analysis, methods, and applications—a review. IEEE Trans. Reliab. **34**, 198 (1985)
74. W. Masing, Von TESTA zur Protagonistin der Business Excellence—Geschichte der Deutschen Gesellschaft für Qualität e.V, in *Qualitätsmanagement—Tradition und Zukunft. Festschrift zum 50-jährigen Bestehen der Deutschen Gesellschaft für Qualität e.V.*, ed. by W. Masing et al. (Hanser, München, 2003), pp. 389–418
75. D.H. Meadows, J. Randers, D.L. Meadows, *The Limits to Growth* (Universe Books, New York, 1972)
76. L. Merz, Philosophie des Reaktorschutzes. atw, 118–126 (1970)
77. L. Merz, Restrisiko. Das Doppelgesicht der Reaktorsicherheit. atw, 294–298 (1981)
78. P. Miranti, Corporate learning and quality control at the bell system, 1877–1929. Bus. Hist. Rev. **79**, 39–72 (2005)
79. W. Poundstone, *Prisoner's Dilemma* (Anchor Book/Doubleday, New York, 1993)
80. President Eisenhower's, "Atoms for Peace" Speech, 1953. Available at http://www.atomicarchive.com/Docs/Deterrence/Atomsforpeace.shtml
81. Price-Anderson Amendments Act of 2005. Available at http://www.govtrack.us/congress/billtext.xpd?bill=s109-865
82. Profile—John Garrick: nuclear risk assessment pioneer. Risk Anal. **29**, 935–939 (2009)

83. A. Rip, The mutual dependence of risk research and political context. Sci. Technol. Stud. **4**, 3–15 (2001)
84. G.A. Ritter, *Der Sozialstaat: Entstehung und Entwicklung im internationalen Vergleich* (Oldenbourg, München, 1991)
85. C.A. Rosenberg, *The Cholera Years: The United States in 1832, 1849, and 1866* (University of Chicago Press, Chicago, 1962)
86. N. Rosenberg, The direction of technological change: inducement. Mechanisms and focusing devices. Econ. Dev. Cult. Change **18**, 1–24 (1969)
87. T. Schlich, Einführung. Die Kontrolle notwendiger Krankheitsursachen als Strategie der Krankheitsbeherrschung im 19. und 20. Jahrhundert, in *Strategien der Kausalität: Konzepte der Krankheitsverursachung im 19. und 20. Jahrhundert*, ed. by C. Gradmann, T. Schlich (Centaurus, Pfaffenweiler, 1999), pp. 3–28
88. H. Schulz, O. Basler (eds.), *Deutsches Fremdwörterbuch* (Walter de Gruyter, Berlin, 1977), p. 452
89. B. Sinclair, *Philadelphia's Philosopher Mechanics: A History of the Franklin Institute, 1824–1865* (Johns Hopkins University Press, Baltimore, 1974)
90. B. Sinclair, *A Centennial History of the American Society of Mechanical Engineers* (University of Toronto Press, Toronto, 1980)
91. D.F. Smith, J. Phillips (eds.), *Food, Science, Policy and Regulation in the Twentieth Century: International and Comparative Perspectives* (Routledge, New York, 2000)
92. W.H. Sperber, R.F. Stier, Happy 50th birthday to HACCP: retrospective and prospective. Food Saf. Mag. (December 2009/January 2010). Available at http://www.foodsafetymagazine.com/article.asp?id=3481
93. U. Spiekermann, Redefining food: the standardization of products and the production in Europe and the United States, 1880–1914. Hist. Technol. **27**, 11–36 (2001)
94. J.E. Stott, P.T. Britton, R.W. Ring, F. Hark, G.S. Hatfield, Common cause failure modeling: aerospace vs nuclear (2010). Available at http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20100025991_2010028311.pdf
95. T. Tietjen, D.H. Müller, *FMEA Praxis. Das Komplettpaket für Training und Anwendung* (Hanser, München, 2003), pp. 4–5
96. S. Walker, US Nuclear Regulatory Commission, *A Short History of Nuclear Regulation, 1946–1999* (US Nuclear Regulatory Commission, Washington, 2000), pp. 41–42
97. A.M. Weinberg, Science and trans-science. Minerva **10**, 209–222 (1972)
98. A.M. Weinberg, Reflections on risk assessment. Risk Anal. **1**, 5–7 (1981)
99. R. Williams, *Keywords: A Vocabulary of Culture and Society* (Oxford University Press, New York, 1983)
100. A.S. Wohl, *Endangered Lives: Public Health in Victorian Britain* (Harvard University Press, Cambridge, 1983)
101. K. Zachmann, Grenzenlose Machbarkeit und unbegrenzte Haltbarkeit? Das „friedliche Atom“ im Dienst der Land- und Ernährungswirtschaft. Technikgeschichte **78**, 231–253 (2003)

Chapter 2
Risk Management and Business Ethics: Integrating the Human Factor

Christoph Luetge, Eberhard Schnebel, and Nadine Westphal

Risk is defined differently in different disciplines, its meaning sharing the uncertainty that is its essence. This chapter explores "risk" as an active verb, that is, what it means to take risk or to risk, and what it means to manage the risk taking of individuals within the modern corporation. Rather than the usual focus on the well-worn material of the processes and uses of risk measurement and assessment, we explore its ethical aspects, specifically, the conscious decision to take on risk, and its management through incentives that shape organizations' and individuals' focus on risk.

There are ethical issues in risky business activities and risky aspects of business ethics. As business ethics is also a dimension of theories of leadership and human resources management, this article focuses on the ethical aspects of risk management, as outlined by business ethics. After a short introduction, there is a general overview of the different ethically relevant dimensions of risk in business (Sect. 2). Section 3 focuses specifically on management risks and risk assessment. Section 4 outlines the role of normative loopholes in risk management to frame our ethical perspective: order ethics. In Sect. 5 we discuss its relation to risk, specifically the way it deals with the human factor in organizations. Section 6 concludes by sketching a theoretical framework for what corporations can do to effectively manage ethical risks and fulfill their social responsibilities at the same time.

Keywords Business ethics · Ethical risk management · Human factor · Prisoner's dilemma · Corporate social responsibility

C. Luetge (✉)
Peter Löscher Chair of Business Ethics, Technische Universität München, Arcisstr. 21,
80333 Munich, Germany
e-mail: luetge@tum.de

E. Schnebel
Business Ethics, TUM School of Education, Technische Universität München, Arcisstr. 21,
80333 Munich, Germany

N. Westphal
Munich Business School, Munich, Germany

C. Klüppelberg et al. (eds.), *Risk – A Multidisciplinary Introduction*,
DOI 10.1007/978-3-319-04486-6_2,

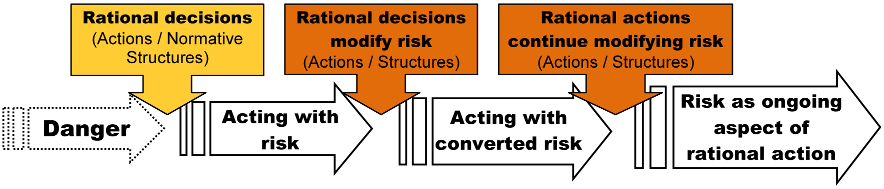

Rational decisions
(Actions / Normative Structures)
Rational decisions modify risk
(Actions / Structures)
Rational actions continue modifying risk
(Actions / Structures)
Danger
Acting with risk
Acting with converted risk
Risk as ongoing aspect of rational action

Fig. 1 Perpetuation of actions and decisions as risk and risk management

The Facts

- Risk can be defined as an active verb requiring both consciousness and courage, which are grounded in responsibility.
- The classical approach of risk management can be explained through the formation of expectations based on observation and measurement.
- Risk plays an important role in human resources management and the corporate role of cultural expectations in approaches to risk.
- There are different personal risk attitudes in motivating behavior and shaping corporate culture.
- Order ethics in a modern, competitive economy sets incentives that encourage cooperative or constructive competition, and thus close the normative loopholes opened by competition.
- The “prisoner’s dilemma” models the ethical dimension of taking risk.
- Ethical risk must be added to other specific business risks such as country risk, settlement risk, market risk, credit risk or operational risk.

1 Introduction: Risk, Rational Choice, and Risk in Business

Wherever people face failure, they may recognize the threat of danger and experience fear. Regardless of the chances of success, when they decide to take action to avert danger they become conscious managers of their circumstances. The decision to act decisively and avoid being at the mercy of danger is when people start calculating or expecting concrete hazard: they assume risk.

In this sense, risk is a process of rational choice, reliant on consciousness and courage. Risk exists wherever people shape their future by rational arguments or classified observations, wherever they act consciously and calculate ongoing actions. It is the attempt to arrange the future and gain power by dealing with an unknown or at least unexpected progression of events (Fig. 1). Despite the fact that we cannot know the scope of our activity at all times, risk empowers future actions. Starting with the courage of the actor, risk rationalizes unknown future events. It is the enlightened counterpart to the rational decision, based on the Kantian “courage to use your own understanding”![1] Risk is the rational equivalent to rational choice,

[1] Kant 1784.

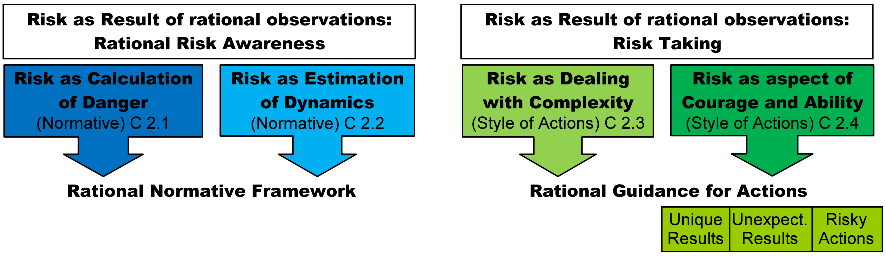

Risk as Result of rational observations: Rational Risk Awareness
Risk as Calculation of Danger (Normative) C 2.1
Risk as Estimation of Dynamics (Normative) C 2.2
Rational Normative Framework
Risk as Result of rational observations: Risk Taking
Risk as Dealing with Complexity (Style of Actions) C 2.3
Risk as aspect of Courage and Ability (Style of Actions) C 2.4
Rational Guidance for Actions
Unique Results
Unexpect. Results
Risky Actions

Fig. 2 Categories of risk awareness and risk taking

but deals with uncertainty, making the human capable of future action in the classical meaning. Risk management focuses on the conceivable scope of action, and in the end it helps to establish normative structures as standards of risk selection.

2 Rational Expectations as the Center of Risk

At its origin, the theory of rational choice refers to the fact of conscious awareness of different choices and to the possibility of taking different actions. It refers to anticipated and unanticipated influences affecting an individual's life, intentions, or targets. The asymmetric difference between known risk and unknown risk, favoring the known and conceptualized in risk aversion, focuses on risk consciousness as well as on risk management. The rise of rational risk awareness is part of the transition to modern society: "perhaps, this was simply a loss of plausibility of the old rhetoric of Fortuna as an allegorical figure of religious content and of prudentia as a noble virtue in the emerging commercial society".[2] There are two rational concepts to classify risk and risk management with two opposite directions of analysis (Fig. 2): risk management as avoidance of harm and danger, and risk management as establishment of new opportunities. We characterize four elements of risk as: rational awareness of (unspecified) danger; rational observation of complexity and possible escalations; estimation and calculation of developments, dynamics, and volatilities for management; and finally risk related to courageous and powerful action.

2.1 Risk and Danger: Results of Rational Observations

Nature is the framework of human planning and social life, but nature is dangerous: earthquakes, storms, and tsunamis affect life on earth in particularly huge and universal ways. But life itself also contains natural dangers, like epidemic diseases. These natural dangers create a mode of rational observations among rational beings.

[2] Luhmann [23].

People draw conclusions from these observations and adjust their behavior. They do not live in areas of explosive volcanoes, or in flood plains, etc. The observation of natural dangers entails appreciating their complexity. There are irregularities in natural phenomenon and catastrophes and these irregularities follow the natural law of complexity and contingency. It is the management of these circumstances that bares the rationality in risk: human beings settle in places where natural dangers are calculable, infrequent incidents.

2.2 Risk as Estimation and Calculation of Developments, Dynamics, and Volatilities

Rational observation leads to assessment of the development of business and its dynamics. Certain estimations characterize risk as a measurement to adjust the probability of success for single decisions: the appreciation of business cycles and the position of a single business within a cycle, and the development of cash flows as part of the dynamic development of that business in the global or local economy. These same assessments are often used in portfolio management. A third application estimates the interrelationship between two ecological systems, if they are coupled strictly. For example, the extension of industrial fishing is a direct risk for the global ecological system if the dynamics and volume of global fishing is continuously increasing.

2.3 Risk as Dealing with Complexity and Contingency: The Awareness of Irritation and Escalation

Continued rational observation of complexity and contingency in order to defeat danger is not limited to natural or unknown situations. In complex natural or social systems it leads to a realization of the phenomenon of contingency and escalation: in particular, unexpected results accumulate unpredictably at single points and lead to escalating systems in a complex set of interrelationships. The possibility of failure rises as the number of combinations of things that can go wrong increases. The complexity of large systems like communications networks means that even tiny glitches can cascade into catastrophic events. In fact, catastrophic events are almost guaranteed to occur in many complex systems, much like big earthquakes are bound to happen.

Without the benefit of perfect foresight, businesses can uncover the fatal flaws and forestall the nascent disasters lurking within their organizations by: (1) assessing risk for informed decisions, such as purchasing an insurance policy; (2) spotting vulnerabilities and addressing them before catastrophic events occur; and (3) designing for resilience. “These ideas have been around for years, but researchers have recently had to reinvent them in the context of extremely complex, interconnected cascade-prone systems".[3] These techniques are helpful even in other complex systems like pandemic viruses in modern societies or local epidemics where a huge number of complex weaknesses converge and escalate unpredictably.

2.4 Risk as Courageous and Powerful Action

The last level of rational observation of risk manages actions and improves results. There are three types of actions in risky situations. One tries to achieve better-than-average results, another is forced by undefined or unexpected developments, and the third is undertaken without knowing the returns.

2.4.1 Risk Management for Achieving Outstanding and Unique Results

First, taking and managing risks is a way of achieving outstanding or unique results. In this sense, an engineer's trip into space or a musician's performance of very difficult pieces are that kind of risk. In both cases, the actors will prepare themselves very well, and will exclude all known disruptive factors or source irritations. They plan how to act with courage, knowing where the main risks are and how dangerous the action is.

Also in this category are other business or scientific innovations in fields where no common experiences exist, such as in clinical studies to develop new active pharmaceutical ingredients, where participants have to take new steps to achieve unique results. Risk management has to calculate everything it knows and to consider even the negative outcomes if the actors fail.

2.4.2 Awareness of Undefined and Unexpected Results or Developments

Risk and risk management are a way of dealing with situations where the circumstances are changing and unpredictable. Managers in these cases know about possible natural, social, or human variations and are prepared to be responsive to changes. For example, in aviation, risk is managed not only by avoidance but also by establishing alternatives in case of changing circumstances. Pilots can change routes, use different airfields, and alternate the altitude or speed. It is not a hierarchic security that affects risk in aviation but a dynamic handling of possible solutions.

[3] Bonabeau [5].

2.4.3 Risky Actions

Risk and risk management are essential whenever people put themselves in jeopardy. Risk in these circumstances is the rational and feasible handling of uncertainty or of known risk, for example, fire fighters who put themselves in risky situations. Our focus is less on the danger of forest fire itself but on how fire fighters act professionally to expose themselves to danger and to return safely.[4] In the same professional style, other people act in dangerous jobs, like frogmen or parachutists.

3 Dimensions of Risk and Risk Management in Modern Business

In business and in business organizations, risk is a term covering the uncertainty of business opportunities. This leads to the need to succeed by developing competitive or organizational advantages. It also leads to acquiring or supporting resources and therefore generating stable payments and sound cash flows. In this way, risks in business and of business opportunities are always measured as financial changes. Accordingly, risk in business can be separated into two parts. There is external risk, tangible as the financial risk of managing current payments, future payments, and financial credits, all essential for acquiring resources. There is also internal risk, dealing with organizational structures, human cooperation, and individual action. Both kinds of risks require guidance from economic rules as well as from ethical insights when facing future challenges and social circumstances. For both, there is a tremendous need to distinguish clearly between risk awareness and risk estimations in order to get a precise understanding of risk management and its ethical aspects (Fig. 3). Risk awareness is the acceptance of what actions we are going to undertake and why they can fail. This helps us to recognize how we can change our actions. Risk estimations are the rational assessments of how risks may develop. They illuminate the urgency of additional efforts or special diligence. It is the rational calculation and planning for unexpected occurrences that separates risk from bare uncertainty as the core characteristic of risk management.

3.1 Risk and Uncertainty: The Knightian Distinction

The distinction between uncertainty and risk as captured uncertainty was first proposed by Frank Knight in his work "Risk, Uncertainty, and Profit".[5] Knight mentioned that the probability or threat of damage, injury, liability, loss, or other negative occurrence, caused by external or internal vulnerabilities, may be neutralized

[4] A very detailed study on this issue and on intrinsic social and organizational problems is introduced by Weick [39] with "The Mann Gulch Disaster".

[5] Knight [17].

Dimensions of risk management	**Aspects of risk in business administration**		
	External risk (eg. country risk, market risk, credit risk)	**Internal Risk** (eg. operational risk, compliance risk, loyalty risk qualification risk)	
Separation from uncertainty	External Risks cover unexpected development of external partners, society, or environment	Risks related to organizational structures entail institutional flexibility	Risks related to individual action cover misbehavior, missing skills and missing motivation
Risk awareness	Observations and judgments of external situations or actions differ from own expectations	Organizational Standards respond always invariably on different and shifting challenges	Individuals take wrong decisions and act without solving problems reduced productivity
Risk estimation	Calculation and assessment of the probability of unexpected situations	Determination of changing recommendations	Calculation of rational behavior and from individual deviation
Rational reason for risk management	Observations of environment and markets presuppose modifications of decisions	Standardized tools of business administration do not cope with substantial changes	Inadequate individual action due to rational dilemmas or nonrational influences lead to inefficient social results

Fig. 3 Classification of different types of risk in business

through premeditated action. His approach combines the measurable likelihood of a hazardous event and the severity of injury that can be caused by the event. Uncertainty, however, persists immeasurably, impossible to calculate. In his words:

> … Uncertainty must be taken in a sense radically distinct from the familiar notion of risk, from which it has never been properly separated. The term "risk", as loosely used in everyday speech and in economic discussion, really covers two things which, functionally at least, in their causal relations to the phenomena of economic organization, are categorically different… The essential fact is that "risk" means in some cases a quantity susceptible of measurement, while at other times it is something distinctly not of this character; and there are far-reaching and crucial differences in the bearings of the phenomenon depending on which of the two is really present and operating.… It will appear that a measurable uncertainty, or risk proper, as we shall use the term, is so far different from an immeasurable one that it is not in effect an uncertainty at all.[6]

3.2 Risk Awareness Leads to Risk Estimation

This distinction between uncertainty and risk rests on the contingency of achieved results relative to the goals of the action and the circumstances.[7] Measurement of uncertainty is a set of probabilities assigned to a set of possibilities. Risk, again, is a state of uncertainty where some of the possibilities involve a loss or other undesirable outcome. The measurement of risk is a set of possibilities, each with quantified probabilities and quantified losses. One may have uncertainty without risk but not

[6] Knight [17].

[7] Hubbard [13].

risk without uncertainty. We can be uncertain about the winner of a contest, but unless we have some personal stake in it, we have no risk. If we bet money on the outcome of the contest, then we have a risk. In both cases, there is more than one outcome. But the measure of risk requires both probabilities for outcomes and for losses (quantified for the outcomes).

The results of the impact of the likelihood (probability) of a hazardous event or phenomenon and the impact of the severity (consequence) constitute what we measure as risk according to what we perceive as risky. For example, for the carcinogen effect, risk is estimated as the incremental probability of an individual developing cancer over a lifetime (70 years) as a result of exposure to a potential carcinogen. For the non-carcinogen effect, it is evaluated by comparing an exposure level over a period to a reference dose derived from experiments on animals. The two impacts engender insights in how to cope: whether we need standardized models of behavior, or whether we only need the awareness and sensibility of scarce effects.

3.3 Risk Assessment: Event Statistics and Bayesian Methods

The many formal methods used to "assess" or to "measure" risk are a critical factor in human decision-making. Some of these quantitative definitions of risk are well grounded in sound statistical theory. However, these measurements of risk rely on failure occurrence data that may be sparse. This makes risk assessment difficult in hazardous industries such as nuclear energy, where the frequency of failures is rare and the harmful consequences are astronomical. The dangerous consequences often necessitate actions to reduce the probability of failure to infinitesimally small values that are both hard to measure and hard to confirm empirically. Often, the probability of a negative event is estimated by using the frequency of equivalent past similar events or by event-tree or fault-tree methods.[8] But probabilities for rare failures may be difficult to estimate if an event tree is not well defined.

In game theory and other models of complex systemic interrelations, more subjective judgments are used for the assessment of risk, related to good judgment or common sense, like the Bayesian probability.[9] Bayesian probability focuses on evidential probabilities as an extension of logic that enables reasoning with propositions whose truth is uncertain. Bayesian probability is an abstract concept to represent a state of individual knowledge in contrast to interpreting probability as a frequency or "propensity" of some phenomenon.[10] Bayesian probability is a mathematical treatment of a non-trivial problem of inference[11] offering two views to

[8] Hanley and Hiromitsu [11].

[9] Stigler [36]: 131.

[10] Jaynes [15].

[11] Stigler [36].

interpret the probability concept: the objectivist view as an extension of logic justified by requirements of rationality and consistency, and the subjectivist view of probability as "personal belief".

3.4 Business Ethics in Risk Management

For risk management issues, a more action-oriented approach is desirable, reflecting ongoing adjustments, interactions, and the ongoing need to reach decisions. Therefore, risk management distinguishes between risks as "future issues" that can be avoided or mitigated and "present problems" that must be immediately addressed.[12] For management, risk is a probability issue. While possibility, as a binary condition, describes something as either possible or not, probability reflects the smooth transition between absolute certainty and impossibility. The term risk only describes "the probable frequency and probable magnitude of future loss".[13] But establishing probabilities is quite different from foretelling the future; therefore action-oriented risk assessment is relevant mainly for individual or subjective judgments.

Additional definitions refer to risk as the effect of the probability of a hazard resulting in an adverse event, combined with the severity of the event.[14] The term "hazard" is used to mean an event that could cause harm, while the term "risk" is used to mean simply the probability of something happening. One of the first major uses of this concept was at the planning of the Delta Works in 1953, a flood protection program in the Netherlands. With the aid of mathematical calculations, the probability of the occurrence of a storm surge was combined with the average cost of damages.[15] This kind of risk analysis is implemented in fields like nuclear power, aerospace, the chemical industry, health, and, increasingly in the last ten years, in the financial industry.[16] Ethical coordination allows assessment of whether the harm is too much or not (e.g. nuclear power) and whether its probability is acceptable or not (e.g. transportation systems).

4 Shaping Risk: Order Ethics and the Human Factor

Motivation is a significant factor in realizing economic goals, as well as a beneficent resource within an appropriate structure for optimal operation. At best, companies'

[12] E.g. "Risk is the unwanted subset of a set of uncertain outcomes" (Cornelius Keating, acr.).

[13] This definition was accepted by The Open Group, see: The Open Group [26].

[14] "Risk is a combination of the likelihood of an occurrence of a hazardous event or exposure(s) and the severity of injury or ill health that can be caused by the event or exposure(s)" (Occupational Health & Safety Advisory Service [25]).

[15] Wolman [40].

[16] Cf. the other articles in this volume.

aims should be precisely outlined by management and communicated clearly to the staff. Ideally, the companies' goals or ends correspond to the staff members' goals or ends. The question is how to organize the communication of these ideas, since there are different human characters, risk strategies, and levels of risk tolerance. Some actors try to avoid risks at any time, while others take risks at any opportunity. Moreover, risk avoidance is not helpful for new actions, as we have seen in Sects. 2 and 3. Thus, the risk tolerance of the decision makers is discussed as a human factor.[17] The main issue is to find fitting settings where the required risk level of the position or task matches people's risk tolerance in order to achieve success. This empowers people to take advantage of existing various levels, attributes and types of behavior and conduct.

4.1 Fundamentals of Order Ethics

This section focuses on what order ethics can contribute to risk management. We will take a closer look at social dilemmas and their relation to issues in risk management.

Most types of ethics are still based on the circumstances of pre-modern societies: the words of the successful 15th century Florentine merchant Giovanni Rucellai: "by being rich, I make others (who I might not even know) poor"[18] illustrate the zero-sum games played in pre-modern societies. In these situations, it is not rational to focus on win-win-situations, but instead to oblige people to be moderate, to share, and to sacrifice.

Order ethics, by contrast, is based on the concept of modern societies, characterized by sustained growth over long periods of time, where economic or financial crises interrupt or deflate economic development for several years but don't seriously affect long-term growth. This development has mainly been made possible by the modern competitive market economy, which enables individuals to follow their own interests and to make independent acquisitions of resources within a carefully built institutional system. In this system of ideal competition, positive sum games are played which create situations where the position of every individual can be improved at the same time. Competition is seen here as a fundamental social condition.

Ideal competitive situations, however, also bring about critical situations: in dilemma situations, possible mutual gains are often left unattained. These dilemmas lie at the core of ethical questions in modern societies, and to avoid or curtail them by designing adequate institutional structures is the challenge of business ethics.[19] The most important dilemma situation is the prisoner's dilemma, which models a fundamental structure of economic action in a globalized world full of interdependence.

[17] Tversky and Kahneman [37].

[18] Rucellai [32] (written about 1450).

[19] Cf. Homann et al. [12], Luetge [19, 20].

The prisoner's dilemma is a fundamental problem in game theory that demonstrates why actors might not cooperate even if it is in their best interests to do so: two suspects of a major crime get caught by the police and are interrogated separately. Although their involvement in that crime cannot be proven, there are some evident misdemeanors that justify detaining them. Each prisoner now has the option of being exempted from punishment by testifying against his accessory (defect) or by remaining silent (cooperate). However, the total exemption (i.e. no sentence at all) works only if just one of them defects while the other remains silent, and applies only to the defector. If both choose to cooperate, they will both be prosecuted for the minor offence and receive a light sentence. If both defect, however, each will receive a higher sentence, which leaves both worse off. In this case, each prisoner's rational choice is to defect and hope that the other does not. Consequently, both actors defect and receive the higher sentence for the major crime, even though they both would have been better off if they had both kept silent.

In this model, the incentives of the social situation force actors to ignore the common fruits of a possible cooperation. The prisoners cannot be expected to cooperate, because the conditions of the situation (the "rules of the game") lead to the other player's defection. In other words: in the prisoner's dilemma, all actors are faced with the possibility of being exploited by others if they behave cooperatively. Therefore they preemptively stop cooperating. This leads to a situation where rational, self-interested actors end up with a result that leaves all worse off and no one better off: morality gets displaced. The situation can be remedied by making conditions equal for all participating social actors: the rules of the game must be changed in order to rule out exploitation.

Members of a cartel are often involved in a prisoner's dilemma, because defecting, that is, selling below an agreed minimum price level, means taking business and profits away from the other cartel members. Inside the cartel this leads to lost profits and therefore is the classical prisoner's dilemma. From a societal or consumer's point of view, defection in a cartel leads to lower consumer prices and therefore to an increase in social welfare.

In a vein similar to the German model of "Ordnungspolitik",[20] the focus of the concept of order ethics is a regulatory framework. This concept emphasizes the importance of rules, too, and of a scope for moral actions that erodes under competitive conditions without institutional rules. Note that the prisoner's dilemma is not always in need of being dissolved. In some situations it is desirable to establish the dilemma in a productive manner, especially to keep actors (firms) in a competitive situation.

Thus, order ethics uses economics as a key theoretical resource and focuses on institutions for implementing moral norms. Individual actors should not be forced to act against their own interest in competitive situations. People cannot, as a general rule, systematically be expected to accept being exploited by others. Especially in the field of risk management, the fear of being exploited may lead to instability, which is not favorable to the entire organizational process. Well-intended moral appeals without sanctions are systematically ineffective and inevitably lead to failure.

[20] Eucken [7].

Order ethics aims at changing the order framework of a society rather than appealing to moral behavior. Further, the principle of morality must not be applied in opposition to but in compliance with economic reality. For that reason, an adequate institutional regulatory framework should provide stable conditions and offer incentives for individual actors based on individual motivations—and for mutual benefit. Institutions should be arranged in such a way that they can contribute to overcoming the previously described dilemma situations.

Being aware of which important role human actors play in complex settings within modern, work-sharing societies, order ethics identifies risk as an object of human resources management and offers mental models to take care of this issue. Embedding into human resources management risk attributes such as "avoids new actions" (risk aversion) or "prefers new actions" (risk courage) may help to improve business organizations, projects, and products, shaping risk aims toward reducing operational risk by taking care of the human factor. An adequate conceptual framework enhances the possible output by expanding the involved actors' capacities to act. As nobody can be expected to accept being intentionally exploited, the rules must be set in an appropriate way, targeting the actors' self-interest: it is better to tell businesspeople how much money they can make by offering us their goods, instead of appealing to their good will.

We can sum up the main four points of the concept of order ethics:

1. Moral actors are exploitable. This is the basic problem of business ethics: people find that they are exploitable as social actors when acting morally, and so do corporations and organizations. Corporations acting in an ethically desirable way may find themselves at a competitive disadvantage compared to others who act less morally.
2. Adam Smith, the founder of economic ethics, was the first to systematically bring together the difference between actions and conditions of actions in order to link competition and morality: morality (incorporated in the idea of the solidarity for all, for example) can be found on the level of the conditions, the rules. Only by making the individuals' moves amoral in principle can competition be made productive. With the aid of rules, of adequate conditions for actions, competition is directed at realizing advantages for all people involved. In this way, others cannot exploit moral behavior, since the rules are the same for everybody. Therefore Smith's approach is based on the distinction between action and conditions of action, between individual actions and the rules of the game. Individual actions are subject to rules.[21] This distinction between the two different levels is often overlooked in ethics.
3. Competition within appropriate rules generates solidarity. Competition should not be restrained or abolished but should be channeled by suitable rules.
4. Within an adequate framework of rules, solidarity as a basic ethical ideal calls primarily for intense competition by self-interested actors, not for sacrificing or sharing.

[21] Cf. Smith [35].

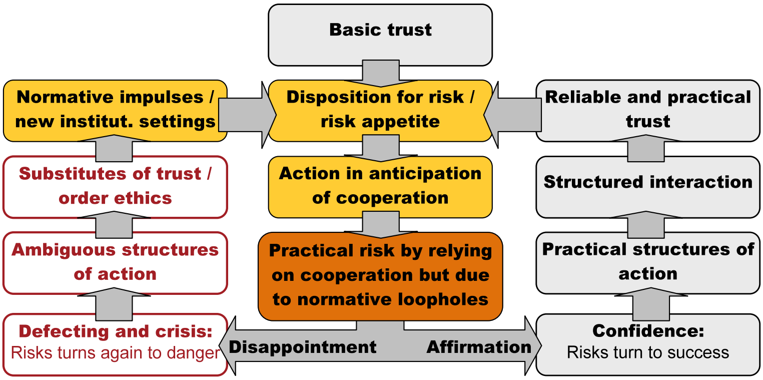

Basic trust
Normative impulses / new institut. settings
Disposition for risk / risk appetite
Reliable and practical trust
Substitutes of trust / order ethics
Action in anticipation of cooperation
Structured interaction
Ambiguous structures of action
Practical risk by relying on cooperation but due to normative loopholes
Practical structures of action
Defecting and crisis: Risks turns again to danger
Disappointment
Affirmation
Confidence: Risks turn to success

Fig. 4 Normative structures for risk management as substitution of missing trust

Human beings act differently when conditions change, either in a positive or a negative way, especially when they find themselves in dilemma situations. This factor should also be taken into consideration once staff is entrusted with decisions where risk behavior matters. The following section offers a mental model for progressively shaping individual risk appetite.

4.2 Risk, Trust, and Normative Loopholes: Ethical Analysis to Improve Risk Management

In regular business settings, wherever business participants fail to cooperate and wherever they are in need of constraints and conditions to avoid defecting, normative loopholes occur. In regular settings, managers learn by making decisions that lead to practical patterns for actions. Their disposition for risk gets modified too: making new decisions leads to new risks and varied experiences that eventually lead to gradually building up more reliable trust—which in turn leads to new and "risky" decisions and actions. But while actions and decisions can either succeed or fail, trust is always brittle. Normative structures can help actors and corporations regain trust.[22] These categories help in identifying areas where normative structures improve the stability of social systems or even where trust is missing (Fig. 4).

Wherever ethical norms are affected to re-establish trust or to avoid defection, it is a matter of order ethics. Order ethics, as mentioned in the introduction, identifies dilemmas where the rational risk orientation of interconnected actors avoids cooperation. It may be rational for business companies to obviate investments in underdeveloped countries where clear legal systems are missing. The economic solution would be to establish transnational institutions or guarantees to minimize the

[22] This model of action refers to Baier [1, 2], Hume [14] and Luhmann [22].

business risks for the companies and to facilitate future investments. The solution of order ethics would be to specify the aspects by which ethical norms may empower managers to take even more risk in special situations without the perspective of increasing profits.[23] It looks for social solutions and new institutional arrangements that fill these gaps or loopholes. In this context, trust is an "institutional" solution for avoiding losses resulting from lack of cooperation—both outside and inside an organization. The human factor emphasizes opportunities inside the company, outlined as part of human resource management.

4.3 Risk Appetite as an Object of Human Resources Management

The active management of risk is not restricted to classic economics, but also involves aspects of human capital. It considers the individual motivation to cooperate as an opportunity to manage risk, but this has to be spelled out in detail to be applicable to managerial tools. The avoidance of defecting business partners or hazardous behavior is genuine risk management. In institutional economics, this is related to the distinction between coordination and motivation as the two basic aspects of managing an organization.[24] Coordination is the skillful distribution of resources and workforce based on work-sharing principles, while motivation concerns the reasons people have for doing what they should in relation to their principles and preferences.

This describes risk aversion as missing a preference for new actions and risk courage (a key attribute for dealing, negotiation, and proceeding) as a willingness to focus on ongoing factors of motivation. Both concepts structure our understanding of risk as organizational knowledge: what kind of risk attribute, corresponding behavior, and conduct is adequate for which project, which new steps, and which level of hierarchy? Typically, the desire to commit to more or less risk is called "risk appetite".[25] Risk appetite as a organization's or individual's attitude towards risk taking is the effort to search for corresponding elements of risk, like profits, sustainability, or a social set of values by organizing businesses, projects, and products. Risk aversion and risk courage are corresponding attributes in a complex world for shaping risk. It is the power to shape operational risk that grows by the knowledge of the role human actors play in complex settings.

To tackle this social problem, we recommend identifying risk appetite as an important factor in business. Within an inappropriate framework of rules and institutions, risk appetite may lead to failures and losses, which have economic as well as ethical aspects. Consequently, we will first discuss ethical aspects of risks that directly involve human beings. Second, we will relate actions of involved social actors to ethical issues such as trust. We will employ the concept of order ethics: order

[23] About the function of normative values in decisions see Schnebel [33].

[24] Cf. Picot et al. [27].

[25] Cf. the Institute of Risk Management, Risk Appetite and Tolerance, Crowe Horwath.

ethics offers, first, “mental models” to take care of these ethical aspects, and second (as a philosophical approach) it stresses the semantic and conceptual aspects of ethics in organizations: a tangible conceptual framework increases the capacity to act, which in turn may have a positive impact on the economic output.

4.4 The Human Factor in Terms of Risk

In modern corporations, social dilemmas[26] are increasingly responsible for reduced productivity due to ignored potential in human resources. This type of risk takes the form of dilemma situations and might be handled better by effective human resources management. Motivation or coordination problems are organizational problems. An efficient process should enhance a well-balanced mix of coordination and motivation in a “human factor-oriented decision making”. Being aware that corporations are affected not only by economic, but also by political and ethical decisions, risks occur in a social context.

Three autonomous European companies (A, B, and C) merge. The headquarters will be based in company A, and all central administrative processes will be organized from there. The headquarters now have to live up to diverse expectations. The new operations must be cost-saving and effective, while at the same time the solution must represent all three companies equally. Moreover, there is the important question of authority. Who has the power of control over the system and who is in charge of risk responsibility? Unfortunately, deficiencies concerning the distribution of power and their consequences are often not discovered until the stage of actual usage. At first, the applicant as well as the users will have to live with this dilemma. Mainly, failed projects carry with them the risk of significant delays as well as extra expenses charged after the project has started. This often results from inefficient decisions and inadequate action, such as withholding relevant information or mishandling access authorization, consciously or not.[27]

Therefore, it may be worth considering the human or ethical factor involved in organizing businesses, projects, and products. The ethical potential of risks should be exposed and more transparency brought into the team-building process. Once every social actor is optimally integrated and effectively deployed by considering risk attributes, team processes might develop more effectively.

Michael Power[28] stated that categories for decisions on actions in relation to individual know-how are possible risk dimensions. They show how managers make a decision on the risky matter itself and do not avoid risk by communicating. Further, he detects that secondary risk, like reputation risk, cannot be managed on the level of

[26] Many of the most challenging problems in modern societies, from interpersonal to intergroup issues, are at their core social dilemmas (cf. Liebrand et al. [18] and Beckenkamp [4]).

[27] Cf. Westphal [38].

[28] Power [29].

communication only. Risks have to be decided on and consistently be adapted to the specific operation. Accounting systems demand an enormous amount of time and energy, while certificates and reports are often too vague and complex to really be useful. Categories that are the basis for risk management decisions are the only items that are important in this regard. According to Power, these are all issues difficult to evaluate in a quantifiable and reliable way: the management culture, its strategy, or its trust in the staff. And even if they are somehow quantified, they are very difficult to manage or administrate. For this reason, the know-how of the participating social actors should be integrated and not ignored.[29]

The model of order ethics does not evaluate the management culture, strategy, or trust in the staff, but can contribute to finding out more about intrinsic incentive structures that go hand in hand with the factors of motivation and may give information about individual risk appetite. Thanks to a smart organization by effective rules, which respect social integration, this factor may be raised and has the potential to become an important economic factor. As soon as we are dealing with situations of risk, the factor of responsibility is significant. Responsibility can only be assigned to people or collective entities, not to categories. As we learned above, games help to reconstruct interaction situations for investigation. The following section elaborates how order ethics integrates the analysis of social-economic phenomena within dilemma structures for improvement.

4.5 *Avoiding Wasted Potential Through Professional Risk Management of Human Resources*

Coordination and motivation as the basis of good organization are linked to individuals, so it is worth considering the principles and preferences of social actors in a process of planning or integrating these aspects in risk decisions, e.g., in the field of mergers and acquisitions. With regard to what effects different levels of managers' risk appetite may have on successful business, it is rational to manage operational risk by integrating the human factor wherever employees are less integrated.[30] The human resources department could deliver valuable input on risk shaping by identifying idle potential based on a concept of order ethics. Idle potential, in this case, denotes the lack of ideal integration of a willing workforce. By using intrinsic incentive structures properly, motivation can be raised, which in turn may lead to a better economic output. Further, trust also plays an important role in this context. An employee who trusts her employer and co-workers may act in a different way in a risk situation than an employee in precarious working conditions.[31]

[29] Cf. Power [29, 30].

[30] Cf. Rücker [34].

[31] Cf. Matthes [24], 56.

Nowadays, corporate decisions are not only concerned with economic, but also with political and ethical responsibility. Consequently, mental models suggest the integration of ethics into risk analysis. This could be organized in five steps:

1. The concept of order ethics may help to implement ethical standards and managerial frameworks in risk management.
2. The primary focus is on rules, not on motives. The imposition of sanctions should support compliance with rules.
3. Self-interest of social actors is assumed to be beneficial within an appropriate regulatory framework.
4. Incentive structures should reward ethical behavior, not punish it.
5. Rules should be beneficial for all involved parties, at least in the longer run.

This way of thinking and acting brings new positive aspects to all stakeholders. For example, team processes have the potential to develop more effectively once every social actor is integrated by considering risk attributes. Moreover, there are serious risks associated with paying for damages, such as fines for corruption or for ecological damage. The advantages of integrating ethical values into risk management may help to organize processes in a more transparent way. In addition, decision makers may be motivated to improve ethical standards and values such as solidarity and fairness in daily work life, as this implies improvement of efficiency at the same time. That is to say, synergies can be used, as each actor brings and/or extracts expert knowledge and empirical values. Next, we focus on examples of how order ethics can help to improve risk management as well as risk taking and decision making.

5 Risk in Resource Optimization

“Risk comes from not knowing what you are doing” (Warren Buffett). Buffett’s statement emphasizes the action-related aspect of risk: take your present decisions and strive to be capable for making new decisions to continue your business. This requires management differentiation between those factors that have to be sensitive to external settings, and other factors that are created and help create the settings of our organization.

5.1 The Current Angle of Risk in Business

In modern business, risk also covers the probability that an actual return on an investment will be lower than the expected return. Therefore, risk is the observation and classification of uncertainties to define areas of decision and of functional structures. Risk covers the fact that the consumption of resources is higher than the production of new resources or services taking into account the effects of losses.

Following the Basel Committee,[32] we will therefore divide risk in business into the following five general categories:[33]

1. *Country risk* covers the uncertainties of political systems and the internal dynamics of societies.
2. *Settlement risk* covers the external uncertainties of running financial processes as well as of operating sequences.
3. *Market risk* covers the uncertainties of factors inside markets in relation to price mechanisms and valuation.
4. *Credit risk* covers the uncertainties of external economic influences on resources needed from the economic environment. This relates to external circumstances other than those of the respective actor, e.g., the repayment of a credit.
5. *Operational risk* extends to parameters and structures of internal organizational processes in relation to individual misbehavior and individual failure, covering therefore the efficiency of adequate integration of an organization's members.

The first two risk categories are observations of things with direct influence on our own activities, but which actually detract from our actions. These are alterations that happen and that we have to cope with. The third and fourth categories of risk are observations of circumstances we accept in order to gain various advantages in terms of additional resources: credit risk covers the uncertainty of expected results due to previous inputs, whereas market risk covers the change of social evaluation. Both risks are accepted if the profits (not only the monetary ones) are appropriate. The fifth risk category, operational risk, belongs to appropriate action of all involved individuals and to the question of an adequate fit of human behavior and organizational structures. The following sections outline the details of these risk categories.

5.2 Country Risk and Settlement Risk

Country risk refers to a country's political system and its economic reliability. It acknowledges that economic and political changes in a foreign country will affect loan repayments. As a result, a buyer or seller of a financial instrument, of other economic obligations, or of foreign currency will not be able to meet associated delivery obligations at maturity. In a way, this risk is related to exchange rate risk, the appreciation or depreciation of a currency resulting in a loss or a "naked position" with regard to the exchange rate. In the end, the state as an actor is referred to in political risk, covering concerns that political changes in a debtor's country will jeopardize debt-service payments. The state is also concerned with sovereign risk: the risk that a local or foreign debtor-government will refuse to honor its debt obligations.

[32] www.bis.org, publications and papers.

[33] Beyond this categorization is a huge number of different other categorizations offered, e.g., for project management, IT, running nuclear power plants, health and cancer.

Settlement risk is a term for all issues that could prevent the fulfillment of a business contract. This could be, for example, the failure of a major bank, resulting in a chain-reaction that reduces other banks' ability to honor commitments. It also includes underwriting risk, i.e. the risk that a new issue of securities will not be sold or that its market price will drop. Settlement risk also covers payment system risk, where payment systems of a major bank will malfunction and will hinder its payments.

5.3 Credit Risk and Market Risk

Market factors challenge the expected availability of resources outside the organization or the financial system and also the liquidity of resources inside respective markets (e.g., financial markets). Generally, they are separated into default risks and capital risk. Default risk denotes a situation where a business partner or borrower might not be able to repay principal and interest from delivered resources. Capital risk refers to losses that accrue from unrecovered loans or from contracts with business partners. They can affect the organization's capital base and may necessitate new capital. Economic risk, in addition, designates changes in the state of the economy that will impair the debtors' ability to pay or the potential borrower's ability to borrow.

Interest rate risk indicates possible declines in net interest income that will result from changes in the relationship between interest income and interest expense. Liquidity risk denotes a deficiency of resources, cash, or cash-equivalents to meet the needs of principals, depositors, and borrowers. This risk is related to reinvestment risk, the lack of opportunities for reinvesting interest-earning assets (loans) at current market rates. Finally there is a refinancing risk, the lack of opportunities to refinance maturing liabilities (deposits) at economic cost and terms.

5.4 Operational Risk

Operational and reputational risk indicate possible and real failure that will prevent an organization from maintaining its critical operations, or meeting the expectations of customers and business partners, especially in core values of business behavior. Compliance risk specifically identifies the failure of fulfilling the intrinsic meaning of a rule or of all organizational guidelines due to misunderstandings and misbehavior of individuals.

Focusing more on the individual aspects than on processes of human resource management, we face motivation risk and loyalty risk as a lack of employee integration into the organization or into the personal requirements of the processes. Professional alignment of organizational gains and ideas among the employees solves these problems. This leads to qualification risk as the lack of employee skills with which to fulfill their job description and to meet all required spontaneous actions, and finally to the risk of poor skill adjustment. The latter constitutes the lack of integration of people into complex project requirements, which impairs expected or required human cooperation.

6 Ethical Risk Management and Responsibilities of Corporations

According to Milton Friedman's famous dictum, "the social responsibility of business is to increase its profits",[34] corporations would have—at most—a responsibility for the order framework of the market. However, we observe corporations doing much more: providing social welfare, or engaging in environmental protection or in cultural and scientific affairs. Therefore, Friedman's picture must be expanded.

6.1 Three Corporate Responsibilities in Order Ethics

Order ethics suggests that the responsibilities of corporations can be differentiated into the following three dimensions:

1. Corporations are responsible for their actions and the immediate consequences that result. This can be defined as their action responsibility. Corporations must comply with laws, and they are responsible for their products, marketing methods, employment policy, corporate culture, and philanthropic activities.

   In an extended sense, action responsibility also encompasses activities that go beyond the traditional, rather passive meaning, such as investing in educational programs, directly fighting corruption and discrimination, or founding trusts. These are important activities in the globalized world. However, they typically have rather local or regional character, and they are mostly uncoordinated, because corporations hesitate to cooperate in this field with others who are normally their competitors. Thus, larger structural problems like hunger, poverty, terrorism, and environmental destruction are not dealt with systematically.
2. In a second step, corporations are responsible for the social and political order framework. In the national setting, this framework is easily identified. But in the global setting, it does not (yet) exist and there is not much reason to suggest that it will come into existence in the near future. Thus, there is room for corporate order responsibility, which can have much greater impact than action responsibility. The main task is to help in establishing basic human rights, a trustworthy judicial system, property rights, and so on. This in turn improves the conditions for future, long-term company benefits.

[34] Friedman [9].

3. This leads directly to the third and most important element: certain mental models can block necessary reforms and create vehement opposition. Many people even regard it as their moral duty to oppose globalization, “neoliberalism”, and the market. These people, however, are usually not convinced by “economic” benefits, narrowly understood, such as improving factors like GNP, but only by engaging in a discourse about the social and economic structures and factors that shape the world. From the perspective of order ethics, it can be shown that many traditional moral ideals are better served by intensifying, not by slowing down, competition within an adequate institutional framework. This is because strong and fitting traditional morals will support organizational success while weak morals will naturally take a back seat. If traditional morals become widely accepted in intensive competition they are more convincing than if they are enforced in an authoritarian way. What is called for is the discourse responsibility of corporations. Corporations must engage in (public) discourse about the social and political order of the global society. People who cannot reconcile this social and political order with their own normative self-image, with their moral or ethical views, will stand in the way of much mutually fruitful and productive cooperation—and endanger the long-run well-being of corporations. In this way, engaging in discourse responsibility is a way of long-term risk management.

6.2 Dangers of Corporate Social Responsibilities

In some cases, people are indeed reinforced in their opinions by bad arguments in favor of the market: for example, if the market is justified by calling it an expression of human freedom—the classic Milton Friedman [8] view—this creates immediate opposition by many people who daily experience otherwise. Many people in Germany, for example, see a growing danger in globalization and in the activities of corporations. Unemployed people—and those afraid of losing their jobs—experience pressure mainly from competition, not freedom. It is therefore vital to stress that freedom and pressure always go hand in hand in the market economy: pressure on suppliers creates freedom of choice for consumers.

As an example, the German system of the “Social Market Economy” is quite often justified or equally criticized by others saying that the role of the “social” is to correct the “anti-social” consequences of the market. In this picture, the market in itself is regarded as morally dubious, to say the least. A better view, and one that the discourse responsibility of corporations should find worthwhile, would be that the word “social” can only mean to create a better, more competitive market that fulfills more of the expectations and goals of its participants. This market can be called an ethically more desirable market. This argument would proceed by showing that people can take more risks as market competitors if they know that the social system will support them. If the concept of a social market economy is to make sense at all in the globalized world, then this strategy of argumentation should be followed. However, two major criticisms are regularly raised against the political activities of corporations:

1. Corporations are "only" maximizing their profits and are therefore only following their own interests. In the political sphere, this is supposed to amount only to lobbying. Certainly, no corporation can control the global social order on its own. Corporations have to justify their actions in public, and that is not the only means of controlling companies. This leads to the second criticism.
2. It is often alleged that corporations lack democratic legitimacy, as CEOs and managers are not elected democratically. This argument presupposes that democracy can be reduced to elections and to the vote of the majority in a Lockean sense. However, following authors like K. Popper,[35] the main function of democracy is not majority vote, but control. In a democracy, control is exercised through many mechanisms, not only by voting. Other methods include competition through markets and public discourse, but also through control of politics by corporations that must reckon with the possibility of being punished in the capital markets. Likewise, these control mechanisms exist in a global setting, with the addition of NGOs, who are of course no more "democratically" elected (in the traditional sense) than corporations. Democratic legitimacy of corporations depends on these control mechanisms being in place. By making their activities more transparent, corporations can enhance their acceptance and thus their democratic legitimacy. This is in their own interest and not simply the moral duty of the "good corporate citizen". It is another method of long-term ethical risk management.

6.3 Conclusion/Food for Thought

Ethical risks play a large role in business, particularly for corporations. The ethical concept of order ethics, which draws some of its main theoretical resources from economics, puts these ethical risks within an adequate theoretical framework, pointing especially to the role of implementation. First, to effectively manage ethical risks within a corporation, the human factor should be taken seriously and dealt with according to the lines sketched here. Second, corporations should engage in risk management in a much wider sense: by actively taking on their political role, corporations fulfill their discourse responsibility, which calls for caring about the ethically relevant arguments used in public discourse. We have given an example of how a bad argument can be detrimental to an adequate understanding of business and ethics in the globalized world, and also to companies themselves, which increases risks. Certainly, corporations cannot fulfill their discourse responsibility entirely on their own. Here, business ethics can help in developing, shaping, and promoting ethical ideas about business ethics and risk management.

[35] Popper famously wrote that the main advantage of democracy is to be able to get rid of its governments "without bloodshed—for example, by way of general elections" (Popper [28], vol. 1, 124, our italics). Note the wording "for example".

7 Food for Thought

- Why should risk management limit itself to classical economic risks? What is the economic rationale for actively dealing with ethical risks?
- What is the role of self-interest for ethics? Should ethics require us to be moderate, to share and to sacrifice?
- What corporate cultures could be instrumental in rationally dealing with ethical risks?

8 Summary

Risk is the consciousness of danger necessitated by uncertainty and the decision to act to lessen its impacts. Corporate culture may be designed with incentives structured to manage risk appetite. Corporate responsibility should use order ethics to close the normative loopholes inherent in competitive markets.

References

Selected Bibliography

1. E. Bonabeau, Understanding and Managing Complexity Risk. MIT Sloan Manag. Rev. **48**(4), 62–68 (2007)
2. A. Crane, D. Matten, *Business Ethics. Managing Corporate Citizenship and Sustainability in the Age of Globalization*, 2nd edn. (Oxford University Press, Oxford, 2007)
3. C. Luetge, Economic Ethics, in *Encyclopedia of Applied Ethics. 4 Vols* (Elsevier, Oxford, 2012)
4. C. Luetge (ed.), *Handbook of the Philosophical Foundations of Business Ethics* (Springer, Heidelberg, 2013)
5. M. Power, *Organized Uncertainty: Designing a World of Risk Management* (Oxford University Press, Oxford, 2007)

Additional Literature

6. K. Arrow, *Social Choice and Individual Values* (Yale University Press, New Haven, 1951)
7. A. Baier, Trust and antitrust. Ethics **96**, 231–260 (1986)
8. A. Baier, Moralism and cruelty: reflections on Hume and Kant. Ethics **103**, 436–457 (1993)
9. M. Beckenkamp, *Sanktionen im Gemeingutdilemma: eine spieltheoretische und psychologische Analyse* (Beltz, Weinheim, 2002)
10. A. Crane, D. Matten, *Business Ethics. Managing Corporate Citizenship and Sustainability in the Age of Globalization*, 2nd edn. (Oxford University Press, Oxford, 2007)
11. W. Eucken, *Ordnungspolitik*. Hrsg. Walter Oswald (LIT, Münster, 1999)
12. M. Friedman, *Capitalism and Freedom* (Chicago University Press, Chicago, 1962)

13. M. Friedman, The social responsibility of business is to increase its profits. New York Times Mag. **32**, 122–126 (1970)
14. D. Green, L. Shapiro, *Pathologies of Rational Choice Theory. A Critique of Applications in Political Science* (Yale University Press, New York, 1994)
15. E. Hanley, K. Hiromitsu, *Reliability Engineering and Risk Assessment* (Prentice Hall, Englewood Cliffs, 1981)
16. K. Homann, P. Koslowski, C. Luetge (eds.), *Globalisation and Business Ethics* (Ashgate, Aldershot, 2007)
17. D. Hubbard, *The Failure of Risk Management: Why It's Broken and How to Fix It* (Wiley, Hoboken, 2009)
18. D. Hume, *An Enquiry Concerning the Principles of Morals* (Oxford University Press, Oxford, 1998). Originally published in 1751
19. E.T. Jaynes, Bayesian methods: general background, in *Maximum-Entropy and Bayesian Methods in Applied Statistics*, ed. by J.H. Justice (Cambridge Univ. Press, Cambridge, 1986)
20. I. Kant, *An Answer to the Question: What Is Enlightenment?* (1784). Originally published
21. F. Knight, *Risk, Uncertainty, and Profit* (Houghton Mifflin, Chicago, 1921)
22. W. Liebrand, P. van Lange, D. Messick, Social dilemmas, in *The Blackwell Encyclopedia of Social Psychology*, ed. by A.S.R. Manstead, M. Hewstone (Blackwell, Oxford, 1996), pp. 546–551
23. C. Luetge, Economic ethics, business ethics, and the idea of mutual advantages. Bus. Ethics, Eur. Rev. **14**, 108–118 (2005)
24. C. Luetge, An economic rationale for a work and savings ethic? J. Buchanan's late works and business ethics. J. Bus. Ethics **66**, 43–51 (2006)
25. N. Luhmann, Die Programmierung von Entscheidung und das Problem der Flexibilität, in *Bürokratische Organisation*, ed. by R. Mayntz (Kiepenheuer & Witsch, Köln, 1968), pp. 324–339
26. N. Luhmann, Familiarity, confidence, trust: problems and alternatives, in *Trust: Making and Breaking Cooperative Relations*, ed. by D. Gambetta (Blackwell, Oxford, 1988), pp. 94–107
27. N. Luhmann, Modern society shocked by its risks. University of Hong Kong, Department of Sociology Occasional Papers 17, Hong Kong (1996)
28. A. Matthes, *Die Wirkung von Vertrauen auf die Ex-Post-Transaktionskosten in Kooperation und Hierarchie* (Deutscher Universitäts-Verlag, Wiesbaden, 2007)
29. Occupational Health & Safety Advisory Services. Document 18001, 2007
30. The Open Group, Technical standard risk taxonomy: document number: C081. The Open Group, 2009
31. A. Picot, R. Reichwald, R. Wigand, *Information, Organization and Management* (Springer, Berlin, 2008)
32. K. Popper, *The Open Society and Its Enemies, 2 Vols.*, 5th edn. (Routledge/Kegan Paul, London, 1966). Originally published in 1945
33. M. Power, *The Audit Society: Rituals of Verification* (Oxford University Press, Oxford, 1997)
34. M. Power, *Organized Uncertainty: Designing a World of Risk Management* (Oxford University Press, Oxford, 2007)
35. R. Rorty, *Contingency, Irony, and Solidarity* (Cambridge University Press, Cambridge, 1992)
36. G. Rucellai, *Ricordanze*. Padua, 1772
37. S. Rücker, *Le risque moral par rapport aux fusions & acquisitions*. Schriftenreihe des ESB Research Institute, vol. 45 (Ibidem Verlag, Stuttgart, 2007)
38. E. Schnebel, Values in decision-making processes: systematic structures of J. Habermas and N. Luhmann for the appreciation of responsibility in leadership. J. Bus. Ethics **27**, 79–88 (2000)
39. A. Smith, *An Inquiry into the Nature and Causes of the Wealth of Nations* (Penguin, Harmondsworth, 1982). Originally published in 1776
40. S. Stigler, *The History of Statistics* (Harvard University Press, Cambridge, 1986), p. 131
41. A. Tversky, D. Kahneman, Rational choice and the framing of decisions. The Journal of Business **59**, S251–S278 (1981)

42. K. Weick, The collapse of sensemaking in organisations: the Mann gulch disaster. Adm. Sci. Q. **38**, 628–652 (1993)
43. N. Westphal, *Ethik als Wettbewerbsfaktor. Wirtschaftsethische Potenziale im Unternehmen* (Lit, Münster, 2009)
44. D. Wolman, Before the levees break: a plan to save the Netherlands. Wired Mag. 17.01. 22.12.2008

Chapter 3
Decision-Making Under Risk: A Normative and Behavioral Perspective

Daniel Straub and Isabell Welpe

This chapter introduces the theories of decision-making under uncertainty and risk of socio-technical systems. Following the historic development of the main conceptions of rationality, we start with expected utility theories and explain the rational choice (or normative) perspective. We explain how decisions under risk can be optimized consistently within the framework of the theory, and under which conditions such analyses are particularly applicable and when they are reduced to an economic cost-benefit analysis. It is then discussed why the classic theories are sometimes misused and why the normative perspective is not suitable to describe or predict actual human behavior, perception or evaluation of decisions and their outcomes under uncertainty and risk. We then outline alternative theories of decision-making, including descriptive approaches from behavioral economics (e.g. cognitive biases) as well as ecological rationality and heuristic decision making. As is discussed in this article, the normative approach is suited for optimizing decisions in a consistent manner for relatively well defined (often technical) problems, whereas the alternative theories are more suitable to predict actual human and social evaluations and behavior and can provide improved decision making in complex situations where socio-technical system parameters as well as the decision maker's preferences are not well defined.

Keywords Expected utility · Risk · Decision theory · Behavoral economics

Mathematics Subject Classification (2010) 62C05 · 62C10 · 91B06 · 62P30 · 62P20

D. Straub
Engineering Risk Analysis Group, Faculty of Civil, Geo and Environmental Engineering,
Technische Universität München, Theresienstr. 90, 80333 Munich, Germany

I. Welpe (✉)
Chair for Strategy and Organization, TUM School of Management, Technische Universität
München, Arcisstr. 21, 80798 Munich, Germany
e-mail: welpe@tum.de

C. Klüppelberg et al. (eds.), *Risk – A Multidisciplinary Introduction*,
DOI 10.1007/978-3-319-04486-6_3,

The Facts

- In theories of judgment and decision making one has to distinguish between how people should make decisions (idealistic, normative approaches) and how people actually make decisions (realistic, descriptive approaches).
- Normative decision theory assumes that under certain circumstances decision makers (should) follow a certain set of rules that ensures consistency among decisions as well as optimal decision outcomes. Descriptive decision theory accounts for the fact that people do not follow these rules and for such situations in which optimal set of rules cannot be given.
- Normative decision theory is applicable to well defined and contained (often technical) problems, and can be used to optimize risk levels. A number of tools, including decision trees and graphs, exist. It can also be used to optimize the amount of information that should be collected to reduce uncertainty before making the decision.
- The utility function describes decision maker's preferences. It is an empirical function that can differ between individuals and is influenced by subjective perceptions. No mathematical form of the utility function is justified by some "universal law".
- Different from what the classical normative theory would propose, the subjective, observer-dependent perception of "objective" values and probabilities has a strong impact on human perceptions, evaluations and decisions. The normative theory therefore generally fails to accurately recognize, describe or predict actual decision making under risk and uncertainty.
- When optimization is not possible, people often make good decisions through the use of heuristics and "gut feelings".
- Most risks are embedded into socio-technical systems, thus is it advisable to be familiar with and use both normative and descriptive risk decision theories.
- There is no "fixed formula" for ideal decision making under risk and uncertainty.

1 Introduction

Decision making under conditions of uncertainty and risk is an every-day task. When deciding whether or not to take the umbrella upon leaving the house, when deciding on whether or not to wear a helmet for bicycling or when deciding whether to take the train or the airplane, you are making a decision that involves outcomes that are uncertain (Will it rain? Will you be hit by a car? Will the train or the plane be safer?) and that are associated with risks (of catching a cold; of sustaining injuries). In our every-day life, we often use intuition (also called heuristics or gut feeling—see Sect. 3) to make such decisions, which often works well. As professionals dealing with risk and uncertainty we often have to make complex and far-reaching decisions or advise the ones that make those decisions, e.g. a committee of experts in health risk that must make a recommendation on acceptable levels of air pollution, a team of engineers that must determine the optimal flood protection strategy for a city or a team of corporate manager that must weigh the economic risks against the technical risks in the introduction of new products and technologies. Even as individuals we frequently must decide between decision alternatives involving uncertainty on which we have little experience and intuition, for example as a patient between different treatment options, as we save for retirement, between different investment strategies or in private life when deciding for or against a life partner. Decision theory has been developed to describe and model the process of making such decisions and ideally supports us in identifying the best options.

Decision theory started out by assuming that the outcomes of decisions can be assessed following a set of consistent decision rules (often—and somewhat misleadingly—referred to as "rational decision making"). Based on these rules, it is then possible to mathematically identify optimal decisions under conditions of uncertainty. Today, this theory is called the *normative* decision theory, because it is useful in describing how decisions should ideally be made under some idealistic, objective and observer-independent assumptions (compare Sect. 3.2), which will be discussed in this article. When studying the behavior of decision makers, it is observed that people's assumptions and resulting actions are not consistent with the assumptions and rules of the normative decision theory. Instead, decisions made by people are influenced by a number of cognitive, motivational, affective and a number of other factors that are not addressed by the classical normative theory. Decisions associated with risk and uncertainty are often concerned with socio-technical systems of some sort, in which human, social and technical dimensions continuously interact. In order to understand, model and reduce risk in these anthropogenic systems, it is necessary to understand how people involved in the process actually perceive, evaluate and decide about risk, which is the aim of *descriptive* decision theory that concerns itself with the empirical reality of how people think and decide.

Examples for the application of the normative theory in risk management include the optimization of decisions on the optimal level of flood protection for a city based on probabilistic models of future flood events and infrastructure performance, or decisions on optimal levels of insurance and reinsurance coverage. Examples for the application of the descriptive theory arise when dealing with processes whose outcomes substantially depend on the perceptions, evaluations, decisions and interventions of humans. For example, consumers decide if genetically modified food is safe for them to buy and eat, or if nuclear energy is an acceptable form of energy technology.

As described in the above paragraphs, in this chapter we distinguish between the normative and the descriptive decision theory. Normative decision analysis uses a mathematical modeling approach based on the expected utility theory (sometimes also called normative, prescriptive, rational or economical decision analysis) and provides a framework for analyzing the optimality of decisions when knowledge of the probability and consequences involved in the decision is available or can be approximated. Descriptive or behavioral decision analysis supports risk-related decisions in complex, socio-technical systems that involve uncertainties with regard to probability and outcomes that make exact quantification difficult. Using either normative or descriptive decision theory in isolation gives an incomplete assessment of the realities of the risk situation. Risk management in socio-technical systems and situations should always consider both normative and descriptive aspects of decision analysis. Risk managers and decision makers need thus be familiar with different risk theories and perceptions.

Section 2 of this chapter presents an introduction to the normative theory while Sect. 3 introduces the descriptive theory. Finally, Sect. 4 concludes with a comparison of the main theories with regard to their assumptions, approach, decision criteria and applicability.

2 Normative Decision Making: Optimal Decision Making Based on the Expected Utility Criterion

2.1 Mathematical, Technical and Economical Perspective: The Rational Approach

In many professional situations it is desirable to select the right decision following a set of logical and reproducible rules and criteria.[1] This holds true in particular when making decisions in groups, where different verbal arguments have to be "translated" into numbers and outcomes, when probabilities and outcome can be sufficiently quantified, and when decisions affect others, as is the case in risk management of anthropogenic systems (e.g. technical systems, environmental systems or companies). When authorities prescribe an acceptable level of air pollution, society expects that the decision on the value of this level is made on a rational and consistent basis (i.e. that the decisions are perceived as legitimate), taking into account all costs and benefits; on the one hand the potential health and environmental effects and on the other hand the economic costs and benefits of setting stringent criteria. A main difficulty in making such decisions is that many of the influencing factors and future outcomes are not and cannot be known with certainty. Neither the health impact of the pollutants nor the cost of reducing them or the value derived thereof for people can be precisely quantified.

To identify optimal decisions in situations when outcomes are uncertain is the goal of classical decision analysis, which has its foundation as a scientific discipline in the publication of the book by Von Neumann and Morgenstern [49] on utility. It is worthwhile noting that although their work is entitled "Theory of games and economical behaviour", it is written by mathematicians and not by empirical scientists. Classical decision analysis is based on the premise that outcomes are uncertain

[1] We note that at least two reasons for this preference can be distinguished: (1) Rules and numbers allow for an "objective" and "true" assessment of risks, probabilities and outcomes. (2) In social interactions, the legitimacy and acceptability of decisions is increased by justifying them through the use of (sometimes just seemingly) objective and true assessment of risks, probabilities and outcomes.

but that it is possible to quantify their probabilities of occurrence. It furthermore assumes that the preferences of the decision makers follow certain rules that are considered rational, as described by utility theory introduced in Sect. 2.3. According to these rules, decisions should not be influenced by any factors that are considered irrelevant for the outcome, in particular not by the context in which gains or losses occur. Despite of (or even because of) these idealized assumptions, classical decision theory provides a useful framework for analyzing decisions involving risk in and quantifying outcomes and probabilities and in describing how decisions should be made in an ideal world. This theory makes it possible to set up consistent (i.e. reproducible and comparable) criteria for making decisions, which is often relevant when decisions need to be justified in social contexts and affect a larger group, as is commonly the case in a socioeconomically or technical context.

In short, the classical decision theory provides a rationale for identifying the decisions and actions that *should* be taken under conditions of uncertainty and risk. For this reason it is often termed the *normative* or *prescriptive* approach. Because it also forms the basis for classical economic theory, it is also often referred to as *economic* decision theory. Hereafter, we will generally use the term normative decision theory.

2.2 System Model, Decisions and Utility

Normative decision analysis requires a model of the relevant system and time frame, the identification of possible decision alternatives and the probabilities and outcomes as well as a measure for evaluating the optimality of the decision alternatives. For engineering problems, the relevant system is typically represented by physical, chemical and/or logical models with input and output variables, some of which are uncertain. In deference to the literature on decision analysis, we will represent the system by a vector of random variables $\mathbf{\Theta}$. Often, $\mathbf{\Theta}$ is referred to as “state of nature”. As an example, consider the problem of determining the optimal flood protection for a city. Here, $\mathbf{\Theta}$ might represent the future maximum water height and discharge of the river, as well as the future land use in the areas at risk.

The decision alternatives can be separated into decisions on actions and decisions on gathering further information. The former, which we will denote by $\mathbf{a}$, actively change the state of the system as represented by $\mathbf{\Theta}$. As an example, the decision on building a dam upstream will change the probability of a flooding of the city or the decision on allowing no building close to the river will alter the damage in the case of a flood. On the other hand, decisions on gathering further information, denoted by $\mathbf{e}$, will not change the state of the system. Upon obtaining the information, our estimate of the system state may change, however. If, for example, one decides to perform an extended hydrological study, one will reduce the uncertainty on the estimate of the intensity of future flood events and obtain a more accurate estimate of maximum floods. In the following we will focus on decisions on actions $\mathbf{a}$; decisions on collecting information $\mathbf{e}$ are considered in pre-posterior decision analysis as introduced in Sect. 2.5.

Finally, we must identify the attributes of the system upon which to assess the optimality of a decision alternative. In the decision on flood protection, these attributes include safety, monetary cost of measures and damages as well as societal and environmental consequences. For optimization purposes, we translate these attributes into a unique metric that allows comparing the alternatives in a quantitative manner. This metric is termed utility u and the associated utility theory, outlined in Sect. 2.3, forms the basis of normative decision analysis

2.3 Utility Theory

The quality of an outcome of a set of decisions on an anthropogenic system is judged on the basis of a number of attributes. As an example, in a decision analysis on the management of contaminated sediments, the following attributes were identified, Kiker et al. [25]:

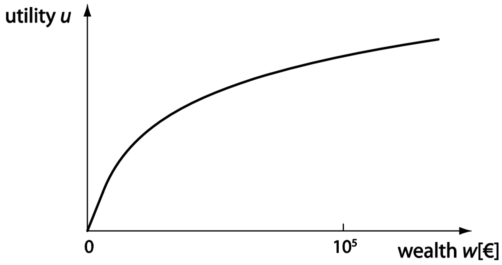
- monetary cost;
- size of the affected area;
- impact on human health (safety);
- impact on ecological health.

In finding an optimal decision, all attributes must be taken into account. Typically, a situation arises where one decision alternative is more optimal with respect to one attribute while another decision alternative is more optimal with respect to another attribute. Cost and safety are common attributes in risk-related problems, and in general a trade-off between the two must be made. If safety was the only attribute, then a system should be designed as safe as possible (consider the pyramids as an example of such a safe structural system). However, it is the art of engineering to design structures that are not only safe but also economical (as well as functional and aesthetical).

The motivation for utility theory is the need for a formalism that allows assessing the optimality of decision alternatives such that the preferences of the decision maker are consistently reflected. Such a formalism enables us to extrapolate from past behavior to new decision situations, both with respect to the trade-off between different attributes and the trade-off among different values of the same attribute. To this end, we define a single metric for measuring the optimality of a decision. This metric is called *utility*. Then, all attributes are transformed into utility by a suitable transformation that consistently reflects the preferences of the decision maker. It is assumed that this transformation, i.e. the weighing assigned to different attributes, is constant with time. To introduce the concept, we study the transformation of the attribute *money* into utility in the following.

First, we note that the utility function, which transforms attributes into utility, is a property of the decision maker. Different decision makers will have different utility functions. In Fig. 1, an exemplarily utility function for an individual is shown. This utility function is continuously increasing, which appears logical, since almost

Fig. 1 Utility function for an individual decision maker, transforming monetary values into utility



everybody would prefer more over less money. However, this is not a necessary condition for the theory; in principle, the utility function can have any arbitrary shape.

Second, we note that the utility is not linear with money over the entire domain. The increase in utility associated with a small increase in wealth, i.e. $\mathrm{d}u(w)/\mathrm{d}w$, is called *marginal utility*. Most decision makers have a marginal utility that decreases with increasing wealth w. (In economics, this is sometimes referred to as the law of diminishing marginal utility.) In simple words: obtaining two million Euros is not simply two times more preferable than obtaining one million Euros.

To understand how the exact form of the utility function is derived, we consider the basic principle of utility theory developed by Von Neumann and Morgenstern [49]. This principle is that:[2]

> *Utility is assigned to the attributes in such a way that a decision* (*on which action to take*) *is preferred over another if, and only if, the expected utility of the former is larger than the expected utility of the latter.*

That is, the utility function is derived to ensure that among different set of decision alternatives, the preferable one will always result in the higher expected utility, $\mathrm{E}[U]$. Expectation is a mathematical operation, which for the case that the utility depends only on the single random variable θ, is defined as

$$\mathrm{E}[U] = \int_{-\infty}^{\infty} u(\theta) f(\theta) \mathrm{d}\theta \quad \text{or} \quad \mathrm{E}[U] = \sum_{\text{all } \theta} u(\theta) p(\theta) \tag{1}$$

where $u(\theta)$ is the utility as a function of the system state θ and $f(\theta)$ is the probability density function (PDF) of Θ if it is continuous and $p(\theta)$ is the probability mass function (PMF) of Θ if it is discrete.

A common way of determining the utility function $u(\theta)$ for monetary values is to consider a series of decisions on whether or not to accept a bet. In each bet, there

[2] For this to hold, a number of consistency requirements must be fulfilled, i.e. the preferences of the decision maker must fulfill a set of axioms, which, however, are in agreement with what is commonly considered to be consistent behaviour. As an example, one of the axioms states that the ordering of the preferences among different outcome events E_i is transitive. Formally, if $\succ$ means "preferred to" then transitivity demands that if $E_j \succ E_k$ and $E_k \succ E_l$ then it must also be $E_j \succ E_l$. For a more formal introduction and the full set of necessary axioms, consult e.g. (Luce and Raiffa [5], Sect. 2.5).

is a probability of p to win a monetary prize of x_1 and a probability of $(1-p)$ to loose x_0. For this bet, the expected utility of the two decision alternatives are as follows:

$$\begin{aligned} &\text{Decision not to bet, } a_0\text{:} && \mathrm{E}[U \mid a_0] = u(0), \\ &\text{Decision to bet, } a_1\text{:} && \mathrm{E}[U \mid a_1] = (1-p)u(-x_0) + p \cdot u(x_1). \end{aligned}$$

$\mathrm{E}[U \mid a_0]$ stands for: the expected value of U for given decision a_0. If, for particular values of x_0, x_1 and p, the decision maker prefers the decision a_0 over the decision a_1, it must hold that $u(0) > (p-1)u(-x_0) + p \cdot u(x_1)$. If she prefers a_1 over a_0 the opposite must hold, and if she is indifferent it is $u(0) = (p-1)u(-x_0) + p \cdot u(x_1)$. By varying the values of x_0, x_1 and p, it is now possible to determine the value of the utility function for different monetary values, so that it is *consistent with the actual decisions made by the decision maker*. You may try to establish your own utility function by playing such an imaginary game.

(We note that a linear transformation of the utility function does not alter the ordering of preferences, i.e. with $u_1(X) = c + b \cdot u(X)$ and b and c being constants, if $\mathrm{E}[u(X) \mid a_0] > \mathrm{E}[u(X) \mid a_1]$ it must also hold that $\mathrm{E}[u_1(X) \mid a_0] > \mathrm{E}[u_1(X) \mid a_1]$. For this reason, any linear transformation of the utility function is allowed, which implies that two points of the utility function can be freely selected.)

2.3.1 Probability

Decision making based on the expected utility theory requires one to assess the probability of all relevant system outcomes. In practice, these probabilities must often be estimated by the decision maker on the basis of limited or no data. The probabilities represent the knowledge of the decision maker at the time of making the decision, and are therefore subjective values. The problem of assessing these probabilities in real situation is further addressed in Sect. 3.1 and in Chap. 12, [42].

2.3.2 Risk

In the context of utility theory and normative decision analysis, we will use the following definition of risk:

> Risk is the expected change in utility associated with uncertain, undesirable outcomes.

Following utility theory, decisions are not made based on risk, but on the basis of the expected utility (of which risk is a part). The optimal decision is the one that leads to the highest expected utility. It follows that the risk that should optimally be taken is the risk associated with this decision.

2.3.3 Risk-Aversion

Utility functions are often concave, like the one of Fig. 1, corresponding to diminishing marginal utility. When considering losses, this can be explained by the fact

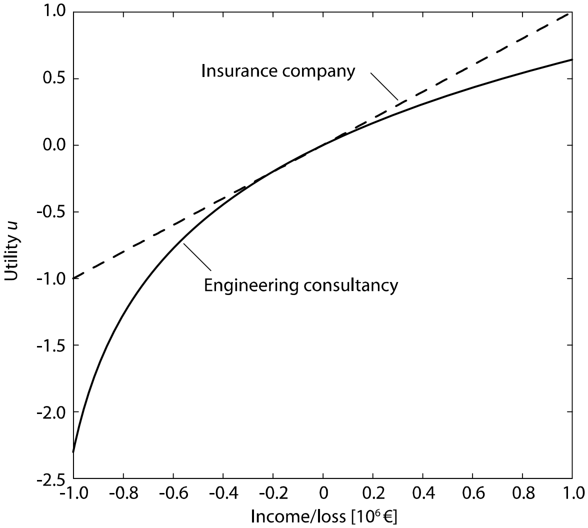


Fig. 2 The utility function for a small engineering consultancy versus the utility function for a large insurance company

that substantial losses can have consequences that go beyond the direct losses, and which therefore cannot be compensated by gains elsewhere. As an example, for a company the loss of 10,000€ is likely to be twice as bad as the loss of 5,000€, but the loss of 2 Million € can be disproportionally worse than the loss of 1 Million € if such a loss threatens the liquidity of the company.

Typically, the utility function is linear (or almost linear) within a range that is small compared to the working capital of the decision maker. This "size effect" is illustrated in Fig. 2, showing the difference in the utility function of a small versus a large company. In the considered range, the utility function is linear for the large company (these sums are "peanuts" for the insurance company), whereas it is concave for the small company where the loss of one million is a critical event.

A consequence of the concave shape of the utility function is that decision makers tend to avoid risks. Consider an event A, causing a loss of 10^5€, and an event B, with associated loss 10^6€. Assume that the probabilities of these events are $p_A = 0.1$ and $p_B = 0.01$. The expected monetary loss of both events is $p \cdot Loss = -10^4$€. Assume that the decision maker is the engineering consultancy whose utility function is shown in Fig. 2. The utility associated with the losses are $u(-10^5€) = -0.09$ and $u(-10^6€) = -2.3$, respectively. The expected utility associated with events A and B (the risks) are $\mathrm{E}[U_A] = 0.1 \cdot (-0.09) = -0.009$ and $\mathrm{E}[U_B] = 0.01 \cdot (-2.3) = -0.023$. Therefore, although the expected monetary loss is the same, the risks associated with event B are higher. This effect is commonly referred to as *risk aversion*.

Illustration 1 (Why Risk Aversion Motivates Insurance) This illustration is taken from Straub [6]. Consider the engineering consultancy whose preference is represented by the utility function in Fig. 2:

$$u(x)=\ln\left(\frac{0.9}{10^6}x+1\right),\quad [x \text{ in } €].$$

This company is managing a project that involves considerable risk because of a penalty in case of a delay. It is estimated that the probability of the event "project delayed" is $p=5\ \%$, and the penalty associated with that event is 800,000€. The company is now offered an insurance that, in the event of a delay, covers the penalty minus a deductible of 80,000€. The premium is 50,000€.

For the engineering consultancy, the expected utility of action a_0, not to buy insurance, is

$$\mathrm{E}[U \mid a_0]=p\cdot u(-800{,}000\,€)=0.05\cdot\ln\left[\frac{0.9}{10^6}(-800{,}000\,€)+1\right]=-0.064.$$

The expected utility of action a_1, to buy insurance, is

$$\begin{aligned}\mathrm{E}[U \mid a_1]&=p\cdot u(-130{,}000€)+(1-p)\cdot u(-50{,}000€)\\&=0.05\cdot\ln\left[\frac{0.9}{10^6}(-130{,}000\,€)+1\right]+0.95\cdot\left[\frac{0.9}{10^6}(-50{,}000\,€)+1\right]\\&=-0.050.\end{aligned}$$

Since it is $\mathrm{E}[U \mid a_1]>\mathrm{E}[U \mid a_0]$, the optimal decision for the consultancy is to buy the insurance.

On the other hand, for the insurance company (whose utility function is $u_1(x)=x/10^6$) the optimal action is to sell the insurance, since $\mathrm{E}[U_1 \mid a_0]=0$ and $\mathrm{E}[U_1 \mid a_1]=p\cdot u_1(-670{,}000€)+(1-p)\cdot u_1(50{,}000€)=0.008$.

It is important to realize that insurance only makes sense if the insured party has a different utility function than the insurer. If the engineering company had a linear utility function, it should not buy the insurance, since the expected utility associated with that decision would be lower. (It corresponds to computing expected monetary values.) This linearity holds approximately when losses are small. (You can verify this yourself by repeating the above calculations for the case where all costs are reduced by a factor of 10, i.e. when the penalty cost is 80,000€, the premium is 5,000€, and the deductible is 8,000€. You will find that in this case, insurance is not an optimal strategy for the consultancy.)

The above example illustrates the effect of *risk-averse* behaviour. A decision maker is said to be risk-averse whenever his utility function is concave; mathematically this corresponds the utility function having a negative second derivative: $\mathrm{d}^2u(w)/\mathrm{d}w^2<0$. This decision maker tries to avert risks, even though this reduces his expected monetary gains, because it maximizes his expected utility.

Measures for risk aversion have been proposed by economists. The most well known measure is the coefficient of *absolute risk aversion* (*ARA*), introduced by Arrow and Pratt [32], defined as

$$ARA(w)=-\frac{u''(w)}{u'(w)} \tag{2}$$

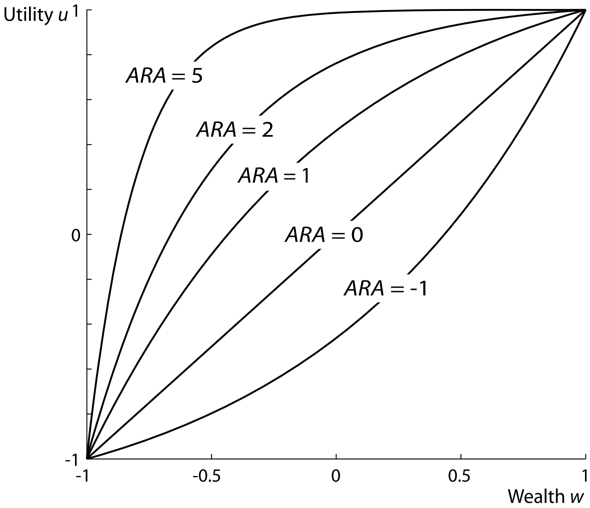


Fig. 3 Utility functions with different absolute risk aversion (*ARA*). All utility functions have been scaled to give $u(-1) = -1$ and $u(1) = 1$

where $u'(w) = \mathrm{d}u(w)/\mathrm{d}w$ is the first derivative and $u''(w) = \mathrm{d}^2u(w)/\mathrm{d}w^2$ the second derivative of the utility function with respect to wealth w. Figure 3 shows several utility functions with varying *ARA*. These are of the form

$$u(w) = 1 - \exp(-cw). \tag{3}$$

This utility function results in an $ARA(w) = c$ that is constant for all values of w (you can verify this claim by inserting the utility function in Eq. (2)). For a negative *ARA*, the decision maker is said to be risk seeking. This corresponds to a convex utility function, as exemplified in Fig. 3 by the utility function with $ARA = -1$.

Alternative measures of risk aversion exist, e.g. the Arrow-Pratt coefficient of relative risk aversion (*RRA*):

$$RRA(w) = -w\frac{u''(w)}{u'(w)}. \tag{4}$$

There is a vast body of literature available investigating these and other measures of risk aversion (e.g. Menezes and Hanson [30]; Binswanger [11]), most of which is rather technical. It is, however, important to realize that the utility function is an empirical function and there is no mathematical form of the utility function that is justified by some "universal law". In fact, Rabin [33] shows that already relatively weak assumptions on the form of the utility function, namely the assumption of diminishing marginal utility for all levels of wealth w, can lead to absurd predictions when extrapolating from decisions involving small sums to decisions with large consequences. The reason behind this is that people do not generally behave consistently according to the expected utility theory, as discussed later in Sect. 3. This observation does not invalidate the use of expected utility theory, but it points to the fact that extrapolation of the utility function assuming some underlying mathematical form (like the one of Eq. (3)) should not be performed. If this is taken into consideration, then utility theory (and the measures of risk aversion) provides rules for optimizing decisions under uncertainty and risk.

2.3.4 Expected Utility Theory vs. Economic Cost-Benefit Analysis

Many decisions involve events with consequences that are small compared to the "working capital" of the decision maker. This is particularly true if the decision maker is society or a representative of society, e.g. a governmental body such as the federal transportation administration. In this case, the utility function will be linear with respect to monetary values. As we have seen earlier, the ordering of the expected utility of different decision alternatives is not altered by a linear transformation of the utility function; we can thus set the utility function equal to monetary values when all consequences are in the linear range of the utility function. In this case, the decision problem can be reduced to an economic cost-benefit analysis (Chap. 11, [36]).

Because monetary values are commonly used in society and economics for exchanging and comparing the value of different goods and units, decisions are often assessed based on expected monetary values. However, it is important to be aware that such an approach is only valid under the conditions stated above (i.e., a linear utility function in the relevant range of consequences). For example, if the engineering consultancy in the example above would make its decision based on expected monetary values, it would decide not to buy the insurance, which would not be optimal according to the company's preferences expressed by the non-linear utility function.

2.4 Multi-attribute Decision Making

So far we have seen utility functions of a single attribute (wealth), yet in most real-life problems involving risks, consequences are associated with several attributes (e.g. economical cost and safety). When multiple attributes are relevant, it becomes necessary to define joint utility functions of the different attributes. Multi-attribute utility theory (MAUT) as presented in Keeney and Raiffa [3] is concerned with decision problems involving multiple attributes.

As an example, consider a decision problem with two attributes X_1 and X_2. A possible joint utility function is constructed from the marginal utility functions $u_1(X_1)$ and $u_2(X_2)$ by

$$u(X_1, X_2) = c_1 u_1(X_1) + c_2 u_2(X_2) + c_{12} u_1(X_1) u_2(X_2). \tag{5}$$

In this case, the two attributes X_1 and X_2 are said to be utility independent. Often, it is $c_{12} = 0$ and the joint utility function reduces to

$$u(X_1, X_2) = c_1 u_1(X_1) + c_2 u_2(X_2). \tag{6}$$

In this case, the two attributes X_1 and X_2 are said to be additive utility independent.

Once the joint utility function u is established, decision analysis proceeds as in the case of the single attribute: the optimal decision is identified as the one that leads to the highest value of the expected utility.

We do not go further into the details of MAUT, but we note that whenever multiple attributes are present (and they are so in most decision problems), a joint utility function is necessary to make consistent decisions. It is important to be aware of this, because it is sometimes argued that it is unethical to assess attributes such as the health of humans or ecological values by the same metric as monetary values (in particular if that metric happens to be the monetary value itself). These arguments are generally misleading, however. In the end, a decision is made, which always implies a trade-off between individual attributes. If two designs for a new roadway are possible, one with lower costs and one with lower environmental impacts, then the final decision made will imply a preference that weights these two attributes, if only implicitly. In fact, it is possible to deduce an implicit trade-off from past decisions. Viscusi and Aldy [48] present an overview on research aimed at estimating the "value of a statistical life" based on societal decisions and choices, and Lentz [28] demonstrates how such deduced trade-offs can be used to assess the acceptability of engineering decisions. The problem with not making these trade-offs explicit is the possibility for making decisions that reflect an inconstant assessment of society's preferences and which lead to an inefficient use of resources. An example of such inconsistent decision making is given by Tengs [44], who compares 185 potential life-saving measures that are or could be implemented in the United States. She finds that with current policies, around 600,000 life years are saved by these measures at a cost of 21 Billion US$ (the numbers are valid for the 1990s). By optimizing the implemented measures using cost-effectiveness criteria, she concludes that with the same amount around 1,200,000 life years could be saved. It follows that the inefficient use of resources here leads to a loss of around 600,000 life years (corresponding to around 15,000 pre-mature deaths each year that could be avoided at no additional cost).[3]

The above argument does not discard the benefits of communicating the values of individual attributes for different decision alternatives. In particular for important and complex decisions it is strongly advocated that decision makers and stakeholders should be given the information on the effect of their decisions on all the relevant attributes.

[3] We note that, in principle, such a cost-effectiveness analysis does not require us to assign our preferences, i.e. it is not necessary to make the trade-off between money and safety explicit. Theoretically it would be sufficient to list the measures according to their effectiveness, as done by Tengs [44], and then starting from the top of the list select all measures that are affordable. In practice, however, such an approach is not possible, because these measures are implemented by different governmental agencies and other actors, who do not make a joint planning. By assigning an explicit trade-off between safety and cost (i.e. by putting a monetary value to human life), however, it can be ensured that money is spent optimally even without performing a joint optimization. Each decision can be tested individually against the criteria set by decision analysis, based on the joint utility function of life-savings and money (see also Lentz [28]).

2.5 Modeling and Optimizing Decisions with Decision Trees and Influence Diagrams

Utility theory prescribes that the optimal set of decisions is the one maximizing the expected utility. Therefore, normative decision analysis essentially corresponds to computing the expected utility for a given set of decisions $\mathbf{a}$, $\mathrm{E}[u(\mathbf{a},\boldsymbol{\Theta}) \mid \mathbf{a}]$, and then solving the optimization problem:

$$\mathbf{a}_{opt} = \arg\max_{\mathbf{a}} \mathrm{E}\big[u(\mathbf{a},\boldsymbol{\Theta}) \mid \mathbf{a}\big]. \tag{7}$$

The operator $\arg\max_{\mathbf{a}}$ reads: the value of the argument that maximizes the expression on the right hand side. The expectation E[ ] is with respect to the random variables describing the uncertain system state $\boldsymbol{\Theta} = [\Theta_1; \ldots; \Theta_n]$. It is defined as

$$\mathrm{E}\big[u(\mathbf{a},\boldsymbol{\Theta}) \mid \mathbf{a}\big] = \int_{\Theta 1} \cdots \int_{\Theta n} u(\mathbf{a},\boldsymbol{\theta}) f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) \mathrm{d}\theta_1 \cdots \mathrm{d}\theta_n. \tag{8}$$

This is a generalization of Eq. (1) to the case of multiple random variables. Equation (8) applies to the case where all uncertain quantities $\boldsymbol{\Theta} = [\Theta_1; \ldots; \Theta_n]$ are described by random variables with joint probability density function $f_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$. If all or some of the random variables are discrete, the corresponding integration operations in Eq. (8) must be replaced with summation operations.

To represent and model the decisions $\mathbf{a}$ and their effect on (expected) utility, *decision trees* and *influence diagrams* have emerged as useful tools. The presentation in this section is limited to decision problems with given information, i.e. for problems in which all uncertain quantities are described by known probability distributions and it is not possible to gather further information. The possibility to collect further information will be introduced in Sect. 2.6.

2.5.1 Decision Trees

In a decision tree, all decisions $\mathbf{a}$ as well as random vectors $\boldsymbol{\Theta}$ describing the states of the system are modeled sequentially from left to right. Each decision alternative is shown as a branch in the tree, as is each possible outcome of the random variables. A generic decision tree is shown in Fig. 4, with only one random variable Θ with m outcome states $\theta_1, \ldots, \theta_m$. The tree is characterized by the different decision alternatives a, the system outcomes Θ described by a probability distribution conditional on a, and the utility u as a function of a and Θ. The decision alternatives as well as the system outcomes can be defined either in a discrete space, a continuous space or a combination thereof.

The analysis proceeds from left to right: for each decision alternative a_i, the expected value of the utility is computed following Eq. (8) and the optimal decision is found according to Eq. (7).

Illustration 2 (Pile Selection) This example, which involves only discrete random variables and decision alternatives, is due to Benjamin and Cornell [10]. A construction engineer has to select the length of steel piles at a site where the depth to the

Table 1 Utility function

State of nature	Actions	
	a_1: Drive 15 m piles	a_2: Drive 20 m piles
θ_1: Depth to bedrock is 15 m	No loss	5 m of the pile must be cut off, 100 unit loss
θ_2: Depth to bedrock is 20 m	Piles must be spliced and welded and construction is delayed, 400 unit loss	No loss

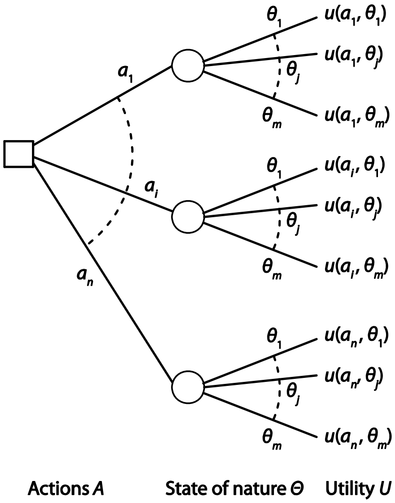


Fig. 4 Generic decision tree for the analysis with given information

bed-rock is uncertain. The engineer has the choice between 15 m and 20 m piles and the possible states of nature are a 15 m or 20 m depth to the bedrock. The consequences (utility) associated with each combination of decision and system state is summarized in Table 1.

The probabilities of the different outcomes are $p(\theta_1) = 0.7$ and $p(\theta_2) = 0.3$. The full decision tree for this problem is shown in Fig. 5. The expected utilities for decisions a_1 and a_2 are obtained as $\mathrm{E}[U \mid a_1] = 0.7 \cdot 0 + 0.3 \cdot (-400) = -120$ and $\mathrm{E}[U \mid a_2] = 0.7 \cdot (-100) + 0.3 \cdot 0 = -70$. Obviously, the optimal decision is to order the larger piles.

The decision tree grows exponentially with the number of decisions and random variables considered, due to the necessary ordering of decisions and random variables (each decision must be made conditional on the decisions and random variables to its left, and each random variable is described by a probability distribution conditional on the decisions and random variables to its left). The decision tree is thus not convenient for representing decision problems involving more than just a few parameters. A more efficient and flexible alternative are influence diagrams, introduced in the following section.

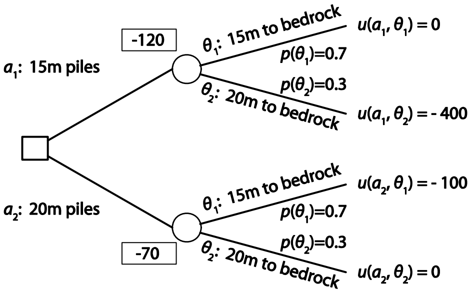

-120
a_1: 15m piles
θ_1: 15m to bedrock
$p(\theta_1)$=0.7
$p(\theta_2)$=0.3
θ_2: 20m to bedrock
$u(a_1, \theta_1) = 0$
$u(a_1, \theta_2) = -400$
a_2: 20m piles
-70
θ_1: 15m to bedrock
$p(\theta_1)$=0.7
$p(\theta_2)$=0.3
θ_2: 20m to bedrock
$u(a_2, \theta_1) = -100$
$u(a_2, \theta_2) = 0$

Fig. 5 Decision tree for the pile selection problem with given information

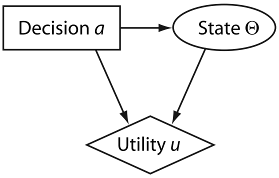


Fig. 6 Influence diagram for a basic decision problem corresponding to the decision tree in Fig. 4

2.5.2 Influence Diagrams

As an alternative to decision trees, decision problems can be represented by influence diagrams. These are more concise representations of the problem, and they are particularly useful in problems where several decisions have to be considered. They were first proposed by Howard and Matheson [21].

Influence diagrams are acyclic directed graphs, whose nodes represent random variables (round nodes), decisions (squared nodes) and utility functions (diamond-shaped nodes). Directed arrows among the nodes represent the dependence structure of the problem. Figure 6 shows a generic influence diagram with one decision a and one random variable Θ. Here it is assumed that Θ depends on the decision a and the utility is a function of both a and Θ.

To understand the semantics of the influence diagrams, it is useful to interpret them as extensions of Bayesian networks (BN) (Jensen and Nielsen [2]). The rules for dependence among the variables follow directly from the BN, with only a few additions: in influence diagrams, links have the additional meaning of representing the flow of information. When making a decision a, the state of the variables that have links going to the node a are known, as are all the ancestors of those variables. Consider the example of Fig. 7, which is different from the one in Fig. 6 only in the direction of the link between a and Θ. This graph implies a completely different decision problem: because the state of Θ is known at the time of making the decision, this represents a decision problem under certainty. A second important rule in influence diagrams is that for the case of several utility nodes, it is assumed that the utility functions are additive independent, Eq. (6).

We do not go further into the details of the influence diagrams here, but note that they can often be constructed from intuition. However, care is needed in ensuring that the relations among the nodes are consistent with causality and with

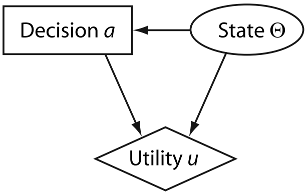


Fig. 7 Alternative influence diagram for a basic decision problem. Here, the uncertain state of the system Θ is known at the time of making the decision a: this is a decision problem under certainty

the assumptions regarding independence among variables. Examples for the construction of such models are given e.g. in Jensen and Nielsen [2], Straub [6]. Free software that allows the construction and computation of influence diagrams (and Bayesian networks) is available, e.g. the Genie/Smile code that can be downloaded from http://genie.sis.pitt.edu/.

2.6 Preposterior Decision Analysis (How to Optimize Decisions on Collecting Information?)

Previously, we have assumed that all information is available at the time of making the decision and that it is not possible to obtain additional information on the uncertain state of nature $\mathbf{\Theta}$. However, in most cases when decisions must be made under conditions of uncertainty, it is possible to gather additional information to reduce the uncertainty prior to making the decisions $\mathbf{a}$. As an example, in the decision on flood protection, it might be possible to perform additional detailed studies to reduce the uncertainty in estimating damages for given levels of flood. The question that must be answered is: is it efficient to collect additional information before deciding $\mathbf{a}$? Or in other words: is the value of the information higher than the cost of obtaining it?

Preposterior decision analysis aims at optimizing decisions on gathering additional information $\mathbf{e}$, together with decisions on actions $\mathbf{a}$ (the letter $\mathbf{e}$ is derived from the word experiment). Typical applications of preposterior decision analysis are:

- Optimization of monitoring systems and inspection schedules
- Decision on the appropriate level of detailing in an engineering model
- Development of quality control procedures
- Design of experiments

It is important to realize that collecting and analyzing information does not alter the system. (Exceptions are destructive tests, which sometimes worsen the state of the system.) For this reason, decisions on gathering information $\mathbf{e}$ do not directly lead to a change in the risk, unlike decisions on actions $\mathbf{a}$. The benefit of $\mathbf{e}$ is the reduction in uncertainty on the system state $\mathbf{\Theta}$, which in turn facilitates the selection of optimal actions $\mathbf{a}$. Preposterior decision analysis allows quantifying this benefit, the so-called *value of information*. (The word preposterior derives from the fact that we calculate in advance (pre-) the effect of information on the model, i.e. the updating of the prior model with the information to the posterior model.)

The quality of the information obtained by performing $\mathbf{e}$ is described by a likelihood function $L(\boldsymbol{\theta} \mid \mathbf{z}) \propto \Pr(\mathbf{Z} = \mathbf{z} \mid \boldsymbol{\theta})$, which is well known from classical statistics. The change in the probability distribution of the system state $\boldsymbol{\Theta}$ with information $\mathbf{z}$ is obtained via Bayes' rule as

$$f_{\boldsymbol{\Theta}|\mathbf{Z}}(\boldsymbol{\theta} \mid \mathbf{z}) \propto L(\boldsymbol{\theta} \mid \mathbf{z}) f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}). \tag{9}$$

Once the information $\mathbf{z}$ is obtained (posterior case), the optimal decisions $\mathbf{a}_{opt}$ are found according to the procedure described in the previous section, whereby $f_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$ is replaced with $f_{\boldsymbol{\Theta}|Z}(\boldsymbol{\theta} \mid \mathbf{z})$. Prior to obtaining the information, however, it is necessary to consider all possible outcomes $\mathbf{Z}$ to assess the benefit of collecting the information in the first place.

In preposterior analysis, we jointly optimize the decisions $\mathbf{e}$ and $\mathbf{a}$. If additional information is obtained through $\mathbf{e}$, then the decision on $\mathbf{a}$ will be based on that information. Therefore, it is not reasonable to determine the optimal action $\mathbf{a}$ a-priori. In contrast, it is possible to optimize so-called *decision rules* d, which determine which actions $\mathbf{a}$ to take based on the type of experiment performed $\mathbf{e}$ and the outcomes of the experiment $\mathbf{Z}$, i.e., $\mathbf{a} = d(\mathbf{e}, \mathbf{z})$. For example, a decision rule in the case of a medical test would be to subscribe a treatment if the test results in a positive indication and do nothing if the test result is negative. The optimization problem in preposterior analysis can thus be written as

$$[\mathbf{e}_{opt}, d_{opt}] = \arg\max_{e,d} \mathrm{E}\big[u\big(\mathbf{e}, \mathbf{Z}, d(\mathbf{e}, \mathbf{Z}), \boldsymbol{\Theta}\big) \mid \mathbf{e}, d\big] \tag{10}$$

where the utility is now a function of the selected experiments $\mathbf{e}$, the outcome of the experiments $\mathbf{Z}$, the state of the system $\boldsymbol{\Theta}$ and the final actions $\mathbf{a}$, $u(\mathbf{e}, \mathbf{z}, \mathbf{a}, \boldsymbol{\theta})$, and the expectation is with respect to the system state $\boldsymbol{\Theta}$ and the experiment outcomes $\mathbf{Z}$.

Details on how to compute the above expectations, as well as on modeling the information, can be found in the literature, in particular in the classical reference of Raiffa and Schlaifer [35] and in Straub [43]. Here, we restrict ourselves to presenting the computations by means of an illustrative example in the following.

Illustration (Pile Selection) We reconsider the pile selection problem introduced earlier. The engineer is now considering whether or not she should use a simple sonic test to obtain a better estimate of the depth to the bedrock. A sound wave created at the surface is reflected at the bedrock and the time between the hammer blow and reception at the surface is utilized to estimate the depth. The test has three possible outcomes, namely estimates of 15 m depth, 17.5 m depth and 20 m depth. The corresponding test likelihoods $L(\theta_i \mid z_i) = \Pr(Z = z_i \mid \Theta = \theta_i)$ are summarized in Table 2.

The sonic test e_1 comes at a cost, corresponding to the deployment of the test equipment and the analysis of the test results. This cost is 20 utility units, i.e. $u_e(e_1, z) = -20$. (The utility associated with different combinations of bedrock depth and pile lengths are given in Table 1.)

To determine whether the sonic test should be carried out or not, the engineer carries out a preposterior decision analysis. She summarizes the problem in the form of a influence diagram, Fig. 8.

Table 2 *Test likelihoods* $L(\theta_j \mid z_i) = \Pr(Z = z_i \mid \Theta = \theta_j)$

Test result	State of nature	
	θ_1: Depth is 15 m	θ_2: Depth is 20 m
z_1: 15 m indication	0.6	0.1
z_2: 17.5 m indication	0.3	0.2
z_3: 20 m indication	0.1	0.7

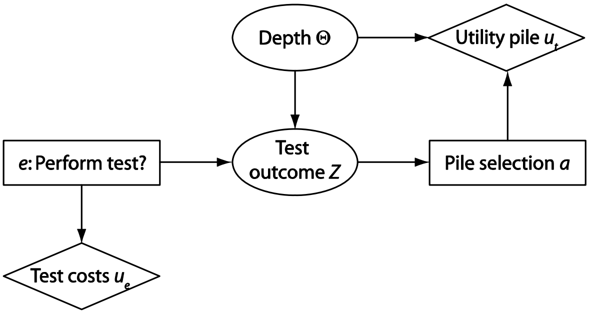


Fig. 8 Influence diagram for the pile selection preposterior analysis

The influence diagram can be implemented in software, since all the relevant information is provided earlier in the text. For this small example, calculations can also be performed manually, as illustrated in Straub [6]. The decision not to inspect leads to an expected utility of −70, as was calculated earlier. The decision to inspect leads to an expected utility of −60, and is therefore optimal. The reason for this higher utility is that the test might indicate a lower depth and the smaller pile can be chosen in this case. Even though this indication is not completely reliable (there is a probability $\Pr(\Theta = \theta_2 \mid Z = z_1) = 0.07$ that the depth is 20 m despite an indication of 15 m), it is sufficiently accurate to provide a higher expected utility.

The value of information of the test can be computed by comparing the expected utility with and without the test and subtracting the cost of the test itself. For the considered sonic test, the value of information is −60 − (−70) − (−20) = 30.

3 Descriptive Decision Making: Decision Making Based on Empirical Observation

3.1 Challenges and Limitations of Normative Decision Theory

> “When the map and the territory don’t agree,
> always believe the territory”
> Gause and Weinberg [17]—describing Swedish Army Training

Normative decision theory is widely used in economics, mathematics and engineering, and in many other decision-related sciences. Its strength lies in the quantification of probabilities and outcomes, and thus of translating verbal arguments into a common (mathematical) language making different risks directly comparable. Yet, this strength of the theory is also the source of its weaknesses. Normative decision theory struggles when quantification cannot be easily accurately achieved, which is particularly the case when dealing with many of the more complex challenges and problems involving risk. In particular those, that involve human and social systems and their the interaction with technical systems. Moreover, empirical research has repeatedly demonstrated that by using normative decision theory one cannot accurately predict how people will decide in a given situation.

The following anecdote reported by Gigerenzer [20, p. 62] illustrates how these two points of criticism often limit the practical usefulness of normative decision theory in guiding our decision-making. He describes how

> *A decision theorist from Columbia University struggled with the decision on whether to accept an alternative offer from another university or whether he should stay at his current university. His colleague allegedly gave him the following advice: "Just maximize your expected utility—you always write about doing this". To which the decision theorist replied. "Come on, this is serious".*

It sheds a light on the dispute between the different branches of decision theory that the decision theorist in question, Howard Raiffa, never actually said this, but on the contrary did decide to move to Harvard using a formal decision analysis to guide his decision, as he recalls in [34].

Broadly, the limitations of normative decision theory can be divided into the following two categories:

People Decide Based on Their Subjective and Observer-Dependent Perceptions and Observations A main assumption of normative decision theory is that peoples' evaluations and decisions are guided by "objective" and "observer-independent" criteria. However, empirical research has repeatedly shown us that the same objective characteristics of a situation can be assessed completely differently by different people (cf. Welpe et al. [50]). Someone might, for example, think that the probability of 80 % of failing with their entrepreneurial start-up is too high a risk for them to take, whereas someone else in the same situation might find a 10 % probability of success to be "a good chance" and "well worth the risk". In other words, normative decision theory does not take into account that economic and social evaluations and decision are subjectively perceived and thus observer-dependent. Thus, different utility functions can lead to different "best or optimized decisions" by different individuals in the same situation or with the same information. Whenever people are part of the decision-making, there is no universal objective reality that can be quantified and calculated. What does this mean for the empirical study of risk and uncertainty?

Probabilities and Outcomes Often Cannot be Quantified in Risk Decisions Economists have in the past studied risk by looking at rather simple economic risk games ("gambles"), such as the centipede game. This enhances our understanding of decision-making in situations where probabilities and outcomes are well-known in advance. It does, however, help us little in understanding the real-life decisions of entrepreneurs or politicians, as they are typically not faced with decision situations in which all different outcomes along with their probabilities are known in advance. In many situations, decision-makers (regardless of which decision theory is used) are unable to rigorously determine probabilities and outcome values of all risk-related events in advance (Sect. 2.3.1). Risk managers, entrepreneurs, decision-makers typically encounter situations that are not entirely mathematically resolvable, unlike when betting on a number in the Roulette game, where the probabilities of winning and losing as well as the potential pay offs are known in advance to all players (i.e. decision-makers). This is rarely the case in complex sociotechnical risk problems. This might call into question the usefulness of economic risk experiments that use gambles to understand risk decision making (Stanton and Welpe [41]).

Whenever accurate predictions are necessary (e.g. when important issues are at stake) but impossible, it is advisable better to realize and accept these limitations instead of falsely relying on alleged and delusive certainty. For some problems, the issues can be addressed by making a decision analysis and forecast based on the best available estimates followed by sensitivity analyses. For all problems that are not sufficiently well understood and the interrelations of the parameters are not well known, in particular with social and economic systems that are inherently complex, self-emergent and variable, it is often impossible to accurately predict the future of such systems. It is advisable to employ several alternative approaches for risk assessment and risk decisions in order to harvest the strengths of multiple approaches and compensate for their respective limitations.

3.2 Examining the Underlying Assumptions of Normative Decision Theory

The assumptions of normative decision theory closely resemble and are based on the well-known (some people think: infamous) “Homo Oeconomicus”. Homo Oeconomicus is an artificial model of human perception and decision-making, who is self-oriented, has preferences that are stable over time and is able to process information fully and rationally. Following Kirchgässner [26], “Homo Oeconomicus” lives in an unrealistic world in which all information including probabilities and outcome values of all choice options are known and freely available without any transaction costs, which also include the time and energy necessary to search, evaluate, contract, and control information and information providers (e.g. Kirchgässner [26]). The model of “Homo Oeconomicus” makes a number of additional assumptions among which are *optimality*, *universality* and *omniscience* (Kurz-Milcke and Gigerenzer [27]). Here, *optimality* means that individuals strive for the best possible solution instead of a solution which is good-enough. *Omniscience* implies that individuals have complete information about positive and negative consequences of a decision. Kurz-Milcke and Gigerenzer [27] further argue: (1) that *universality* is an expression of the idea that a common currency or calculus exists which underlies all decisions, (2) that normative decision theory assumes that humans are always both willing and cognitively capable of identifying the optimal decision, which would be one that maximizes according to a certain criterion (e.g. money, happiness), (3) that individuals (as well as organizations) are fully aware of all existing decision possibilities and their associated costs, benefits and probabilities in the present and future. Of course, these assumptions are a "mathematical idealization" of reality, and are not adequate to completely describe the current evaluations, decisions and behaviors of people, let alone predict their future utilities and actions. The question to ask is whether a completely accurate description is necessary, use- or helpful for any given risk management problem.

Previous research has repeatedly shown that the formal conceptualization of rational decision-makers and the empirically observed human behavior differ substantially (e.g. Tversky and Kahneman [47]; Kahneman and Tversky [23]). Since 1970, Akerlof [8] has argued that information is typically unevenly shared between any two transaction partners, resulting in ubiquitous "information asymmetry" as the rule not the exception. Having full information during a decision process is in reality impossible. Furthermore, transaction costs exist in virtually all transactions (Coase [13]). Even if such a world ever existed in which all information is known and freely available, Simon [38] was one of the first scholars to point out that the limited cognitive ability of individuals limits the identification of any best option from several alternatives. People are simply unable to process and evaluate every alternative in an acceptable time frame.

Ford et al. [16] review 45 studies that investigate the outcomes of decision-making and shows that humans often use heuristics instead of weighing pros and cons as normative decision theory would predict. They conclude with the statement that "*the results conclusively demonstrate that non-compensatory*[4] *strategies were the dominant mode used by decision makers. Compensatory strategies* (i.e. trading off good and bad aspects of two competing alternatives—parentheses added by Straub and Welpe) *were typically used only when the number of alternatives and dimensions were small or after a number of alternatives have been eliminated from consideration*".

3.3 Behavioral Decision-Making Theories

The following section introduces two theories in decision making that address the limitations of the classical theory for descriptive decision analysis, namely, (a) prospect theory that emphasizes the limitations, cognitive and affective biases of human decision making and (b) the approach of ecological rationality that emphasizes the human ability to make correct decisions under limited time and information through the use of heuristics and "gut feeling". The goal of this section is

[4] Heuristics are an example of a non-compensatory strategy.

to illustrate that human decision-making is inevitably influenced by a great number of biases, emotional and cognitive influencing factors which are difficult to foresee and quantify and which sometimes benefit and sometimes deteriorate the outcomes of human decisions.

3.3.1 Prospect Theory

Prospect theory was introduced by Kahneman and Tversky [23]. They were awarded the Nobel Prize in Economic Sciences in 2002 "*for having integrated insights from psychological research into economic science, especially concerning human judgment and decision-making under uncertainty*" (Royal Swedish Academy of Sciences [46, p. 1]). Their work integrates normative decision theory with insights from behavioral sciences and cognitive psychology. Furthermore, they introduced experiments as an innovative methodology for economics in their research. These developments have laid the foundation for a new field of research called behavioral economics, which has been the starting point of a paradigmatic shift in the study of human decision-making under risk. In contrast to expected utility theory, which is considered to be a *prescriptive and normative* theory, prospect theory is a *descriptive* theory of human behavior in decision making under risk constituting an extension of the normative expected utility theory.

One of the main contributions of prospect theory is in its explicit consideration and inclusion of the observer-dependent *perceptions* of utility and in the subjective weighting of outcome probabilities. An important aspect economists have previously overlooked (some continue to overlook it) is that human preferences with regard to seemingly "objective facts" are highly context-dependent and can consequently show a great deal of inter-individual differences. To illustrate this further: a glass of water can be worth a few pennies if you are sitting at home and are not thirsty and it can be worth a million dollars if you are alone in the desert, close to dying of thirst. This seemingly trivial example illustrates a central point. Standard economic theory uses normative decision theory, which has not found a way yet to incorporate how individuals perceive, evaluate, weigh and judge objective probabilities, risks, outcomes, costs and benefits depending on the context and their subjective mental states. Even though these observations and deliberations are hardly surprising to social scientists, especially psychologists, and probably also to the average lay person, they had a great impact on economists and economic theory, for reasons outlines in Sects. 3.1 and 3.2.

Kahneman, Tversky and colleagues empirically investigate the value function of individuals, in which the loss curve has a steeper decline than the gain curve based on the person's respective reference point or status quo. A main finding of prospect theory (e.g. [23]) shows that people react more sensitively to any losses (i.e. changes below their individual status quo on the value function) than to gains (i.e. changes above the individual status quo), even when the resulting value of the outcome is the same (so that normative decision theory predicts the same utility).

Furthermore, this line of research has consistently shown that the subjective perceptions of objectively equal risk alternatives can vary because of different wording and phrasing of the decision alternatives. The most basic example of this kind would be to describe a glass, which only contains half of its content as 50 % empty versus 50 % filled. Scholars (e.g. Tversky and Kahneman [7]; Levin and Chapman [29]) have repeatedly demonstrated in numerous experiments that individuals' preferences change simply due to a different wording, the so called "framing" alone.

Illustration (The Framing Effect in an Example by Messick and Bazerman [31, p. 13])

Situation 1: *A large car manufacturer has recently been hit with a number of economic difficulties. It appears that it needs to close three plants and lay off 6,000 employees. The vice president of production, who has been exploring alternative ways to avoid the crisis, has developed two plans:*

Plan A: *Will save one of three plants and 2,000 jobs*
Plan B: *Has a one-third probability of saving all three plants and all 6,000 jobs, but has a two-thirds probability of saving no plants and no jobs*

Situation 2: *Same situation as in situation 1, but two different plans*

Plan C: *Will result in the loss of two plants and 4,000 jobs*
Plan D: *Has a two-thirds probability of resulting in the loss of all three plants and all 6,000 jobs, but has a one-third probability of losing no plants and no jobs*

Empirical studies show that most executives choose plan A in situation 1, but plan D in situation 2, despite of Plan A and C being equivalent and Plan B and D being equivalent. This example shows: when the glass is described as half-full it is more attractive than when it is described as half-empty. Messick and Bazerman [31] explain this result by the fact that the reference point of the decision-makers is a different one: in the first case, the reference point is the good situation where all plants are OK; in the second case, the reference point is the bad situation, namely the one where all plants must be shut-down. The typical pattern of responses is consistent with the general tendency to be risk averse with gains and risk seeking with losses. If the problem is framed in terms of saving jobs and plants (plans A and B) executives tend to avoid the risk and take the sure plan. If the problem is framed in terms of losing jobs and plants (plans C and D) executives tend to seek the risk and not to take the sure plan.

Kahneman, Knetsch, and Thaler [24] argue that loss aversion described in prospect theory influences decision processes in that humans are generally more negative about potential losses (risks) than they are positive about possible gains (opportunities). Related to prospect theory, Kahneman et al. [24] have identified a number of additional cognitive biases and so-called irrational "anomalies" with regard to human decision-making. For instance, the *status quo bias* or the *endowment bias* (Samuelson and Zeckhauser [37], Kahneman et al. [24]).

The *endowment bias* effect is closely associated with loss aversion (Thaler [45]) and is salient when a loss of any asset weighs much higher in the decision-making than a win of an asset with the same size and value would. The decisive aspect with the endowment effect stems from the ownership of an object. Research on the endowment effect shows that assets are valued more highly when they are in the possession of the decision-maker than when they are not. Again, this finding confirms that subjective perceptions of seemingly objective characteristics are more important when describing and predicting human decision-making. In a similar way, the *status quo bias* describes the tendency of individuals to prefer the status quo over taking chances and risks in decision making (Samuelson and Zeckhauser [37], Fernandez and Rodrik [14]).

According to *status quo bias theory*, consumer choices depend on which option is framed as the default (i.e. status quo) option. Kahneman et al. [24] have suggested that the status quo bias is the result of a combination of loss aversion and the endowment effect. For politicians, management executives and anyone managing risk-related challenges, the status quo bias means that thinking about what will constitute the "default" in the organization or decision processes will greatly influence which decisions will be taken. An example for a risk-related default would be an organizational rule such as "safety first—when in doubt do what is best for the safety of our products and not what is best from an economic perceptive".

3.4 Ecological Rationality and Heuristic Decision Making

The previous sections have dealt with the abilities and inabilities of humans to optimize decisions and make full use of all information available. More often than not, individuals have to make decisions under limited time and information, which rules out the application of any analytic decision making procedure to determine an "optimal" decision. How do people decide in situations like this? To illustrate this, we first consider an example.

Gigerenzer [20] gives an example that mirrors the different theories and approaches of decision making humans can use: the problem of catching a ball flying in the air in baseball. One could approach this problem by calculating all the probabilities and utilities or one could use a simple heuristic to catch the ball. It is impossible for humans to know all necessary parameters of the flight of the ball to correctly calculate the "parabolic trajectories", i.e. the "ball's initial distance, velocity, wind strength and projection angle" necessary to catch the ball. All of these parameters would need to be assessed and calculated in the short time while the ball is in the air. As the calculation of these parameters is impossible, Gigerenzer [20] suggests, the use of so-called "heuristics", in this case the gaze heuristics to accomplish the task of catching the ball. The gaze heuristic works in the following way: a player fixates the ball and starts running and adjusts his or her speed of running in an extent that allows him or her to keep the angle of his or her gaze constant. The player will probably be unable to know or "calculate" where exactly the ball will touch the ground, but more importantly, keeping the angle between his or her eyes and the ball constant the player will be at the spot where the ball lands. The gaze heuristic is a well-known example of a fast and frugal heuristic. It is called fast, because the heuristic can address problems within matters of seconds, and it is called frugal because it requires little information to work accurately.

Descriptive (and behavioral) decision theory generally agrees that the human information processing capacity is limited, for example through cognitive and affective biases, which make human decision making in general—including heuristic decision making—sub-optimal. In contract, the heuristics approach as pointed out by Gigerenzer and colleagues takes an evolutionary perspective—and argues that such "fast and frugal heuristics" have emerged as a result of human evolution in order to facilitate good decision-making under limited information and time.

Gigerenzer [18, 19] and colleagues are also critical of behavioral economics for a number of points. First, with regard to biases (see Sect. 3.3) they argue that these are "first-best solutions" and "environmental adjustments" of human decision making resulting from long evolutionary processes. In contrast to behavioral economists, he does not categorize heuristic decision making or so called "irrationalities" in decision making in any negative way as "errors" or "second best solutions". They argue that calculating probabilities is much more difficult to accomplish for humans than understanding frequencies (Gigerenzer [18]). Their basic argument is that bounded rationality as introduced by Herbert Simon and what he calls effective "ecological rationality" (i.e. heuristic decision making) do not contradict each other and in fact often co-exist together closely (Gigerenzer and Goldstein [1], Gigerenzer [20]). The original thinking behind this idea is that heuristic decision making, i.e. decision-making that is not based on an exact number or their calculations, is more efficient than decision making based on classic utility maximization. In other words, heuristics are particularly efficient in situations with limited information and time for decision making were mathematical optimization is impossible, which is regularly the case for decisions in managerial or political (and also personal) decisions. Heuristics, nevertheless, need to constantly be adapted to fit the contexts in which they are applied in as no heuristic is effective or useful in all decision situations.

In the following, we present examples for heuristics, namely the *representativeness heuristic*, the *availability heuristic* and the *affective heuristic*.

The *representativeness heuristics* refers to judgments of probabilities of a future event or the representativeness of a sample. In other words, it describes individuals' subjective assessment of probabilities based on the comparison of previous experiences with events or individuals that represent a current event or sample. Particularly important is the subjectively perceived similarity, which can lead to misjudgments because the more individuals perceive events to be similar the more they are likely to ignore important information and previous probabilities about a current situation or sample.

Another important heuristic is the *availability heuristic*, which refers to the evaluation of the probability of events based on one's own previous experiences and memories, which can be easily recalled. The more easily they are recalled, the higher individuals evaluate the likelihood of similar current events (Kahneman and Tversky [22]).

A number of recent approaches focus on the role of *affect in risk perception* (Loewenstein et al. [4]). The "risk-as-feeling" hypothesis (Slovic et al. [40], Slovic and Peters [39]) implies that affects are important determinants for risk perception and evaluation. Loewenstein et al. [4] argue that individuals perceive risk depending on their emotions. Researchers have repeatedly shown that emotions have the potential to influence human decisions through human information processing of the perceived risk. For example, Finucane et al. [15] and colleagues showed that people use affective cues in decision situations under risk. A potential implication of the risk-as-feeling hypothesis would be that positive affect could lead to a biased estimation of risk perception and evaluation.

4 Discussion

All models are wrong, but some are useful.[5]

This chapter has outlined a number of different decision theories, all of which have their merits and their limitations. The choice of the theoretical approach must thus be problem dependent, as emphasized throughout the text. Table 3 summarizes the three main decision theories presented in this chapter.

The classical decision theory is relatively far refined and current research in this area focuses mostly on computational aspects of the optimization problem in various fields of application. There are, however, some alternative novel developments, which address the difficulty in realistically assessing probabilities in real decision situations. One example is the info-gap theory, which is developed to provide robust decisions on a non-probabilistic basis (Ben-Haim [9]). The descriptive and the heuristic theories, due to their empirical nature and shorter history, seem wide open for development and adaptation. In addition, there is ample potential for research on the application of both lines of decision theory to practical problems involving risk. Real decisions (be it in business, technology, politics or other fields) are seldom based on rigorous applications of decision theory, be it normative, descriptive or heuristic. One reason for this lack is the gap between researchers living in an "idealized world" and the practitioners dealing with the "dirty reality".

Concerning the different lines of decision theory, researchers should aim to link the formalism of classical utility analysis with the empirical appropriateness of descriptive and behavioral models. In order to understand and improve decision making on systemic and complex risks, an integrative perspective of normative, descriptive and heuristic decision making may offer many benefits. Another promising area for future research would be to study the normative and behavioral perspectives looking at group decisions as opposed to individual decisions. Furthermore, scholars may want to examine which institutions (rules, regulations, etc.) can be successfully implemented in order to enhance the effectiveness and efficiency of individual and group decisions (e.g. debiasing strategies).

[5] Quoted from the statisticians Box and Draper [12].

Table 3 Overview on the three decision theories presented in this chapter

Decision-theory	Approach	Decision criterion	Suitable for/applicable to	Tools
Classical (normative) decision theory, expected utility theory	Normative (how decisions should be made), mathematical, axiomatic theory	Expected utility (reflecting attributes such as money, safety, happiness); objective/observer-independent; consistent rules; sometimes reduced to cost-benefit analysis	Optimizing decision-making when problems are well-defined, i.e. when probability and consequences can be reasonably quantified. Sufficient time for calculations is available. Important to reduce risks related to technological and environmental hazards	Expected utility maximization. Decision trees and influence diagram, Mathematical optimization, Advanced probabilistic models
Descriptive (descriptive, behavioral) approaches of human decision making	Descriptive (how decisions are made) Empirical, i.e. fitted to observed human behavior (behavioral economics, e.g. prospect theory)	The aim of descriptive decision theory is to describe what people will actually do, not necessarily what they should do. According to prospect theory, individuals compare decision criteria (objective and subjective) against a reference point	Describing (and predicting) actual human behavior Understanding how people actually make decisions (important to reduce risks associated with human and organizational behavior)	Empirical analyses (e.g. experiments or questionnaire studies) to describe actual decision behavior
Heuristic decision making	Descriptive (how decisions are made) Empirical Normative elements (decision heuristics in certain situations) Assumption: Decision makers have intuition on the problem	Subjective/observer-dependent cost-benefit analysis Utility (money, safety happiness)	Optimizing decision making under certain conditions (little time and limited information) and within complex systems	Use of decision heuristics (e.g. representativeness heuristic; cause and result; availability heuristic; heuristic)

5 Food for Thought

- What is the value of economics and classical utility theory given that they make a number of often unrealistic assumptions? Where can they and where can they not create value added by applying them?

- It has been said, that all models are wrong to some degree—is there a point however, where a model becomes "too wrong" or "right enough"—if so, how would one know?
- How can economic theory account for the role of subjective perception of "objective" values and probabilities in human decision making?
- What is the value of information and how can it be assessed?
- How does one (theoretically) construct a utility function for a decision maker, following the classical utility theory?
- From two engineering designs for a tunnel construction, which only differ in safety and cost, one is selected. How can the implicit trade-off between safety and risk be deduced from this solution?
- It has been argued that by not following the expected utility principle when making decisions involving life safety, "we are in effect killing people". Discuss this statement.
- A popular "economics joke": what do economists mean when they write in the conclusion of their paper: "The evidence for our hypotheses is mixed?" It means that economic theory supports the hypotheses but the empirical data does not. Discuss.

6 Summary

Classical normative decision analysis, which is based on the expected utility theory developed by mathematicians, provides an axiomatic framework for optimizing decisions under uncertainties. It is well suited for identifying optimal decisions when copying with risks if probabilities and consequences of adverse events can be reasonably well quantified. Descriptive decision analysis is a generalization of the expected utility theory, accounting for the influence of psychological factors on the decisions made. It is better suited than the classical theory to describe the behavior of humans under uncertainty and risk. Finally, the chapter outlines newer attempts to formalize heuristic decision making, which is based on relatively simple rules, and which assume that these heuristics have developed in an evolutionary process. These theories are particularly well suited to describe (and sometimes optimize) decision making under uncertainty and limited time and information.

References

Selected Bibliography

1. G. Gigerenzer, D.G. Goldstein, Reasoning the fast and frugal way: models of bounded rationality. Psychol. Rev. **103**, 650–669 (1996)
2. F.V. Jensen, T.D. Nielsen, *Bayesian Networks and Decision Graphs*. Information Science and Statistics (Springer, New York, 2007)

3. R.L. Keeney, H. Raiffa, *Decisions with Multiple Objectives* (Wiley, New York, 1976). Reprinted by Cambridge University Press, 1993
4. G.F. Loewenstein, E.U. Weber, C.K. Hsee, N. Welch, Risk as feelings. Psychol. Bull. **127**, 267–286 (2001)
5. R.D. Luce, H. Raiffa, *Games and Decisions: Introduction and Critical Survey* (Wiley, New York, 1957)
6. D. Straub, Lecture notes in engineering risk analysis. TU München (2011)
7. A. Tversky, D. Kahneman, The framing of decisions and the psychology of choice. Science **211**, 453–458 (1981)

Additional Literature and Sources

8. G. Akerlof, The market for 'lemons': quality uncertainty and the market mechanism. Q. J. Econ. **84**, 488–500 (1970)
9. Y. Ben-Haim, *Info-Gap Decision Theory: Decisions Under Severe Uncertainty* (Academic Press, San Diego, 2006)
10. J.R. Benjamin, C.A. Cornell, *Probability, Statistics, and Decision for Civil Engineers* (McGraw-Hill, New York, 1970)
11. H.P. Binswanger, Attitudes toward risk: experimental measurement in rural India. Am. J. Agric. Econ. **62**, 395–407 (1980)
12. G.E. Box, N.R. Draper, *Empirical Model-Building and Response Surfaces*. Wiley Series in Probability and Statistics (1987)
13. R. Coase, The nature of the firm. Economica **4**, 386–405 (1937)
14. R. Fernandez, D. Rodrik, Resistance to reform: status quo bias in the presence of individual-specific uncertainty. Am. Econ. Rev. **81**, 1146–1155 (1991)
15. M. Finucane, A. Alhakami, P. Slovic, S.M. Johnson, The affect heuristic in judgments of risks and benefits. J. Behav. Decis. Mak. **13**, 1–17 (2000)
16. J.K. Ford, N. Schmitt, S.L. Schechtman, B.M. Hults, M.L. Doherty, Process tracing methods: contributions, problems, and neglected research questions. Org. Behav. Hum. Decis. **43**, 75–117 (1989)
17. D.C. Gause, G.M. Weinberg, *Exploring Requirements: Quality Before Design* (Dorset House, New York, 1989)
18. G. Gigerenzer, From tools to theories: a heuristic of discovery in cognitive psychology. Psychol. Rev. **98**, 254–267 (1991)
19. G. Gigerenzer, On narrow norms and vague heuristics: a reply to Kahneman and Tversky. Psychol. Rev. **103**, 592–596 (1996)
20. G. Gigerenzer, Fast and frugal heuristics: the tools of bounded rationality, in *Blackwell Handbook of Judgment and Decision Making*, ed. by D. Koehler, N. Harvey (Blackwell, Malden, 2006), pp. 62–88
21. R. Howard, J. Matheson, Influence diagrams, in *The Principles and Applications of Decision Analysis*, Vol. II. (Strategic Decisions Group, Menlo Park, 1981). Published again in: Decis. Anal. **2**, 127–143 (2005)
22. D. Kahneman, A. Tversky, On the psychology of prediction. Psychol. Rev. **80**, 237–251 (1973)
23. D. Kahneman, A. Tversky, Prospect theory: an analysis of decision under risk. Econometrica **47**, 263–292 (1979)
24. D. Kahneman, J.L. Knetsch, R.H. Thaler, Anomalies: the endowment effect, loss aversion, and status quo bias. J. Econ. Perspect. **5**, 193–206 (1991)
25. G.A. Kiker et al., Application of multicriteria decision analysis in environmental decision making. Integr. Environ. Assess. Manag. **1**, 95–108 (2005)

26. G. Kirchgässner, *Homo Oeconomicus: The Economic Model of Behaviour and Its Applications in Economics and Other Social Sciences* (Springer, Berlin, 2008)
27. E. Kurz-Milcke, G. Gigerenzer, Heuristic decision making. Mark. J. Res. Manag. **3**, 48–56 (2007)
28. A. Lentz, Acceptability of civil engineering decisions involving human consequences. PhD thesis, TU München, Germany (2007)
29. I.P. Levin, D.P. Chapman, Risk taking, frame of reference, and characterization of victim groups in AIDS treatment decisions. J. Exp. Soc. Psychol. **26**, 421–434 (1990)
30. C.F. Menezes, D.L. Hanson, On the theory of risk aversion. Int. Econ. Rev. **11**, 481–487 (1970)
31. D.M. Messick, M.H. Bazerman, Ethical leadership and the psychology of decision making. MIT Sloan Manag. Rev. **37**, 9–22 (1996)
32. J.W. Pratt, Risk aversion in the small and in the large. Econometrica **32**, 122–136 (1964)
33. M. Rabin, Risk aversion and expected-utility theory: a calibration theorem. Econometrica **68**, 1281–1292 (2000)
34. H. Raiffa, Decision analysis: a personal account of how it got started and evolved. Oper. Res. **50**, 179–185 (2002)
35. H. Raiffa, R. Schlaifer, *Applied Statistical Decision Theory* (Cambridge University Press, Cambridge, 1961)
36. J. Roosen, Cost-benefit analysis, in *Risk – A Multidisciplinary Introduction*, ed. by C. Klüppelberg, D. Straub, I. Welpe (2014)
37. W. Samuelson, R. Zeckhauser, Status quo bias in decision making. J. Risk Uncertain. **1**, 7–59 (1988)
38. H. Simon (ed.), *Models of Man: Social and Rational* (Wiley, New York, 1957)
39. P. Slovic, E. Peters, Risk perception and affect. Curr. Dir. Psychol. Sci. **15**, 322–325 (2006)
40. P. Slovic, M. Finucane, E. Peters, D.G. MacGregor, Risk as analysis and risk as feelings: some thoughts about affect, reason, risk, and rationality. Risk Anal. **24**, 1–12 (2004)
41. A.A. Stanton, I.M. Welpe, Risk and ambiguity: entrepreneurial research from the perspective of economics, in *Neuroeconomics and the Firm*, ed. by A.A. Stanton, M. Day, I.M. Welpe (Edward Elgar, Cheltenham, 2010), pp. 29–49
42. D. Straub, Engineering risk assessment, in *Risk – A Multidisciplinary Introduction*, ed. by C. Klüppelberg, D. Straub, I. Welpe (2014)
43. D. Straub, Value of information analysis with structural reliability methods. Struct. Saf. (2014). doi:10.1016/j.strusafe.2013.08.006
44. T.O. Tengs, Dying too soon: how cost-effectiveness analysis can save lifes. NCPA Policy Report #204, National Center for Policy Analysis, Dallas (1997)
45. R. Thaler, Toward a positive theory of consumer choice. J. Econ. Behav. Organ. **1**, 39–60 (1980)
46. The Royal Swedish Academy of Sciences. Press release, advanced information on the prize in economic sciences 2002, 17 December 2002 (retrieved 28 August 2011). http://www.nobelprize.org/nobel_prizes/economics/laureates/2002/ecoadv02.pdf
47. A. Tversky, D. Kahneman, Judgment under uncertainty: heuristics and biases. Science **185**, 1124–1131 (1974)
48. W.K. Viscusi, J.E. Aldy, The value of a statistical life: a critical review of market estimates throughout the world. J. Risk Uncertain. **27**, 5–76 (2003)
49. J. von Neumann, O. Morgenstern, *Theory of Games and Economical Behaviour* (Princeton University Press, Princeton, 1944)
50. I.M. Welpe, M. Spörrle, D. Grichnik, T. Michl, D. Audretsch, Emotions and opportunities: the interplay of opportunity evaluation, fear, joy, and anger as antecedent of entrepreneurial exploitation. Entrep. Theory Pract. **36**, 1–28 (2012)

Chapter 4
The New Role of Mathematical Risk Modeling and Its Importance for Society

Klaus Mainzer

This book on risk and security is an example for the new role of mathematical modeling in science. In Newtonian times, mathematical models were mainly applied to physics and astronomy (e.g., planetary systems) as definitive mappings of reality. They aimed at explanations of past events and predictions of future events. Models and theories were empirically corroborated or falsified by observations, measurements and lab experiments. Mathematical predictions were reduced to uniquely determined solutions of equations and the strong belief in one model as mapping of reality. In probabilistic models, extreme events were underestimated as improbable risks according to normal distribution. The adjective "normal" indicates the problematic assumption that the Gaussian curve indicates a kind of "natural" distribution of risks ignoring the fat tails of extreme events. The remaining risks are trivialized. The last financial crisis as well as the nuclear disaster in Japan are examples of extreme events which need new approaches of modeling.

Mathematical models are interdisciplinary tools used in natural and engineering sciences as well as in financial, economic and social sciences. Is there a universal methodology for turbulence and the emergence of risks in nature and financial markets? Risks which cannot be reduced to single causes, but emerge from complex interactions in the whole system, are called *systemic risk*. They play a dominant role in a globalized world. What is the difference between microscopic interactions of molecules and microeconomic behavior of people? Obviously, we cannot do experiments with people and markets in labs. Here, the new role of computer simulations and data mining comes in.

These models are mainly stochastic and probabilistic and can no longer be considered as definitive mappings of reality. The reason is that, for example, a financial crisis cannot be predicted like a planetary position. With this methodic misunderstanding, the political public blamed financial mathematics for failing anticipations. Actually, probabilistic models should serve as stress tests. *Model ambiguity* does not allow to distinguish a single model as definitive mapping of reality. We have to consider a whole class of possible stochastic models with different weights. In this way,

K. Mainzer (✉)
Chair for Philosophy and Theory of Science, München Center for Technology in Society,
Technische Universität München, Arcisstr. 21, 80333 Munich, Germany
e-mail: mainzer@cvl-a.tum.de

C. Klüppelberg et al. (eds.), *Risk – A Multidisciplinary Introduction*,
DOI 10.1007/978-3-319-04486-6_4,

we can overcome the old philosophical skepticism against mathematical predictions from David Hume to Nassim Taleb. They are right in their skepticism against classical axiomatization of human rationality. But they forget the extreme usefulness of robust stochastic tools if they are used with sensibility for the permanent model ambiguity. It is the task of philosophy of science to *evaluate risk modeling* and to consider their *interdisciplinary possibilities and limits*.

Keywords Risk modeling · Systemic risk · Model scepticism · Risk measuring · Rational behavior

The Facts

- Mathematical modeling is crucial for understanding the dynamics of natural and societal systems.
- The emergence of systemic risks can be explained in nonlinear models of systems science.
- Philosophy of science delivers criteria of good models and their application in risk modeling.
- In risky situations, model skepticism is a challenge of research.
- Risk modeling has its historical origin in financial and insurance mathematics.
- Securitized credit models, their increasing networks of risks, and the crisis of risk modeling lead to a new paradigm of risk measuring and rational behavior.
- We can no longer trust in a single risk model, but we must consider a class of more or less appropriate models, supplemented by experimental behavioral case studies.

1 Introduction

Mathematical models are mathematical descriptions of systems in different sciences. They refer in particular to natural systems in astronomy (e.g., planetary systems), physics (e.g., atomic systems), chemistry (e.g., molecular bonds), and biology (e.g., cellular networks), but also to social systems in economics (e.g., financial markets), sociology (e.g., social networks) and political science (e.g., administrative organizations). When engineers analyze a technical system to be controlled or optimized, they also use a mathematical model. In mathematical analysis, engineers can build a model of the system as a hypothesis of how the system should work, or try to estimate how an unforeseeable event could affect the system. Examples are extreme events and risks emerging in complex systems. Similarly, in control of a system, engineers can try out different control approaches in simulations. Simulations are often represented by computer programs and tested on computers (Bungartz et al. [1]). In the natural sciences, the validity of models is tested by derived explanations or predictions which are confirmed or falsified by observations, measurements and experiments. A hypothetical model is a more or less appropriate mapping of reality.

A *mathematical model* usually describes a system by a set of variables and a set of equations that establish relationships between the variables (cf. Gershenfeld [5], Weidlich [13], Yang [14]). A *dynamical system* is characterized by its elements and the time-depending development of their states. The *states* can refer to moving planets, molecules in a gas, gene expressions of proteins in cells, excitation of neurons in a neural net, nutrition of populations in an ecological system, or products in a market system. The *dynamics* of a system, i.e. the change of system states depending on time, can mathematically be described by, e.g., time-depending differential equations. In physics, a *conservative system*, e.g. an ideal pendulum, is determined by the reversibility of time direction and conservation of energy. *Dissipative systems*, e.g., a real pendulum with friction, are irreversible. In a more intuitive way, a conservative system is "closed" with respect to external influences and only determined by its intrinsic dynamics. A dissipative system can be considered to be "open" to external influences, e.g., air or other material friction forces of the pendulum. Models of conservative and dissipative systems can also be applied in ecology and economics.

Case Study (Conservative and Dissipative Systems in Ecology) At the beginning of the 20th century, fishermen in the Adriatic Sea observed a periodic change of numbers in fish populations. These oscillations are caused by the interaction between predator and prey fish. If the predators eat too many prey fish, the number of prey fish and then the number of predators decreases. The result is that the number of prey fish increases, which then leads to an increase in the number of predators. Thus, a cyclic change of both populations occurs. In 1925, the Italian mathematicians Lotka [36] and Volterra suggested a dynamical model to describe the prey and predator system. Each *state* of the model is determined by the numbers of prey fish and the number of predator fish. So the *state space* of the model is represented by a two-dimensional Euclidean plane with a coordinate for prey fish and a coordinate for predator fish. The observations, over time, of the two populations describe a dotted line in the plane. Births and deaths change the coordinates by integers, a few at a time. To apply continuous dynamics, the dotted lines must be idealized into continuous curves. Obviously, the *Lotka-Volterra model* is closed to other external influences of, e.g., temperature or pollution of the sea. If these external forces of "ecological friction" were added to the model, its dynamics would change the cyclic behaviour.

Case Study (Conservative and Dissipative Systems in Economy) In 1967, the economist Goodwin proposed a conservative dynamical model to make the 19th-century idea of class struggle in a society mathematically precise (cf. Goodwin [26], Mainzer [7]). He considered an economy consisting of workers and capitalists. Workers spend all their income on consumption, while capitalists save all their income. Goodwin used a somewhat modified predator-prey model of Lotka and

Volterra. This *conservative model* supports the idea that a capitalist economy is permanently oscillating. Obviously it is superficial, because it does not refer directly to the functional income shares of capitalists and workers or to their population size. But it is mainly its conservative character that makes Goodwin's model seem economically unrealistic. Thus, the model has been made more realistic by the assumption of "*economic friction*". In reality, an economic system cannot be considered as isolated from other dynamical systems. An economic model of coupled oscillatory systems is provided by international trade. In other cases, economic systems are influenced by political interventions. We will come back to these examples later on.

Mathematical models can be classified in several ways (Mainzer [8, 9]). In classical physics, dynamics of a system is considered a *continuous process*. But, continuity is only a mathematical idealization. Actually, a scientist has single observations or measurements at discrete-time points which are chosen equidistant or defined by other measurement devices. In discrete processes, there are finite differences between the measured states and no infinitely small differences (differentials) which are assumed in a continuous process. Thus, *discrete processes* are mathematically described by difference equations.

Random events (e.g., Brownian motion in a fluid, mutation in evolution, innovations in economy) are represented by additional fluctuation terms. *Classical stochastic processes*, e.g. the billions of unknown molecular states in a fluid, are defined by time-depending differential equations with distribution functions of probabilistic states. In *quantum systems* of elementary particles, the dynamics of quantum states is defined by Schrödinger's equation with observables (e.g., position and momentum of a particle) depending on Heisenberg's principle of uncertainty which only allows probabilistic forecasts of future states.

2 Emerging Risks in Complex Dynamical Systems

2.1 Linear and Nonlinear Models

Historically, during the centuries of classical physics, the universe was considered a deterministic and conservative system. We say that a system is *deterministic* when future events are causally set by past events. A finite-difference equation like $x_{t+1} = f(x_t)$ is deterministic as long as $f(x_t)$ has only one value for each possible value of x_t. Given the past value x_t, the function f determines the future value x_{t+1}. The astronomer and mathematician P.S. Laplace (1814) assumed the total computability and predictability of nature if all natural laws and initial states of celestial bodies are well known. The Laplacean spirit expressed the belief of philosophers in determinism and computability of the world during the 18th and 19th century.

Laplace was right about linear and conservative dynamical systems. In general, a *linear relatio*n means that the rate of change in a system is proportional to its cause:

Small changes cause small effects while large changes cause large effects. Changes of a dynamical system can be modeled in one dimension by time series with changing values of a time-depending quantity along the time axis. Mathematically, linear equations are completely solvable. This is the deeper reason for Laplace's philosophical assumption to be right for linear and conservative systems.

In systems theory (Mainzer [8, 9, 39]), the complete information about a dynamical system at a certain time is determined by its state at that time. The *state of a complex system* is determined by more than two quantities. Then, a *higher dimensional state space* is needed to study the dynamics of a system. From a methodological point of view, time series and phase spaces are important instruments to study systems dynamics. The state space of a system contains the complete information of its past, present and future behavior.

Case Study (State Space in Ecology) Let us consider the state space of a Lotka-Volterra system of predator and prey fishes. The vector field on the *two-dimensional state space* can roughly be described in terms of four regions (Fig. 1a). In region A, both populations are relatively low. When both populations are low, predator fish decreases for lack of prey fish while prey fish increase because of less predation. The interpretation of this habitual tendency as a bound velocity vector is drawn as an arrow. In region B, there are many prey fish, but relatively few predators. But when there are many prey fish and few predator fish, both populations increase. This is interpreted by the vector in region B. In region C, both populations are relatively large. The predator fish are well fed and multiply, while the prey fish population declines. This tendency is shown by the vector in region C. In region D, there are few prey fish but many predator fish. Both populations decline. This tendency is shown by the vector in region D. The *phase portrait* of this system can be visualized by a closed trajectory, because the flow tends to circulate.

In Fig. 1b, the *phase portrait* is a nest of closed trajectories, around a central equilibrium point. As dynamical systems theory tells what to expect in the long run, the phase portrait enables the ecologist to know what happens to the two populations in the long run. Each initial population of predator and prey fish will recur periodically.

If some kind of ecological friction were added to the model, the center would become a point attractor. This would be a model for an ecological system in static equilibrium (Fig. 1c). A different but perhaps more realistic modification of the model results in a *phase portrait* like Fig. 1d, with only one periodic trajectory.

At the end of the 19th century, H. Poincaré (1892) discovered that celestial mechanics is not a completely computable clockwork, even if it is considered a deterministic and conservative system. The mutual gravitational interactions of more than two celestial bodies ('*Many-bodies-problem*') can be illustrated by causal feedback loops analytically represented by nonlinear and non-integrable equations with instabilities and irregularities. In a strict dynamical sense, the degree of complexity depends on the degree of nonlinearity of a dynamical system. According to the Laplacean view, similar causes effectively determine similar effects. Thus, in the

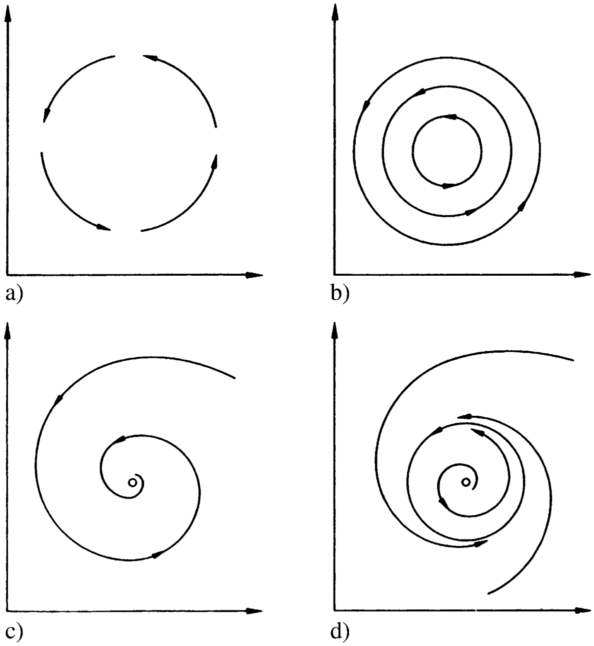


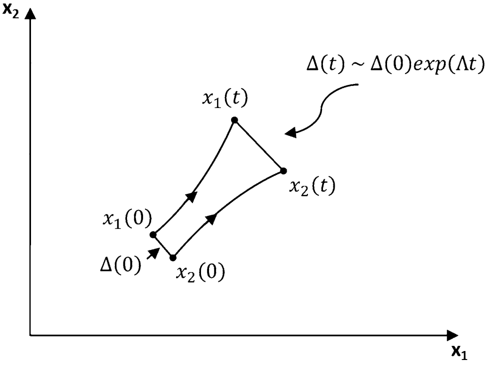
Fig. 1 Phase portraits of an ecological system with a prey and predator population (Lotka-Volterra): (**a**) a closed trajectory, (**b**) a nest of closed trajectories, (**c**) a point attractor, (**d**) a periodic trajectory [7, p. 114]

state space, trajectories that start close to each other also remain close to each other during time evolution. Dynamical systems with deterministic chaos exhibit an exponential dependence on initial conditions for bounded orbits: the separation of trajectories with close initial states increases exponentially.

Important Consequence for Risk Analysis (Butterfly Effect of Chaotic Dynamics) Consider two trajectories starting from nearly the same initial data. In chaotic dynamics only a tiny difference in the initial conditions can result in the two trajectories diverging exponentially quickly in the state space after a short period of time (Fig. 2). In this case, it is difficult to calculate long-term forecasts, because the initial data can only be determined with a finite degree of precision. Tiny deviations in digits behind the decimal point of measurement data may lead to completely different forecasts. This is the reason why attempts to forecast weather fail in an unstable and chaotic situation. In principle, the wing of a butterfly may cause a global change of development. This "*butterfly effect*" can be measured by the so-called *Lyapunov exponent*. A trajectory $\mathbf{x}(t)$ starts with an initial state $\mathbf{x}(0)$. If it develops exponentially fast, then it is approximately given by $|\mathbf{x}(t)| \sim |\mathbf{x}(0)|e^{\Lambda t}$. The exponent Λ is smaller than zero if the trajectory is attracted by attractors, such as stable points or orbits. It is larger than zero if it is divergent and sensitive to very small perturbations of the initial state.

Thus, tiny deviations of initial data lead to exponentially increasing computational efforts for future data limiting long-term predictions, although the dynamics is in principle uniquely determined. According to the famous *KAM-Theorem* of A.N. Kolmogorov (1954), V.I. Arnold (1963), and J.K. Moser (1967), trajectories in the

Fig. 2 Exponential dependence on initial conditions measured by Lyapunov exponent Λ [7, p. 83]



phase space of classical mechanics are neither completely regular, nor completely irregular, but depend sensitively on the chosen initial conditions.

Models of dynamical systems can be classified on the basis of the effects of the dynamics on a region of the state space (Weidlich [13]). A conservative system is defined by the fact that, during time evolution, the volume of a region remains constant, although its shape may be transformed. In a dissipative system, dynamics causes a volume contraction.

An *attractor* is a region of a state space into which all trajectories departing from an adjacent region, the so-called basin of attraction, tend to converge. There are different kinds of attractors (Lorenz [35]). The simplest class of attractors contains the *fixed points*. In this case, all trajectories of adjacent regions converge to a point. An example is a dissipative harmonic oscillator with friction: the oscillating system is gradually slowed down by frictional forces and finally come to a rest in an equilibrium point.

Conservative harmonic oscillators without friction belong to the second class of attractors with *limit cycles*, which can be classified as being periodic or quasi-periodic. A periodic orbit is a closed trajectory into which all trajectories departing from an adjacent region converge. For a simple dynamical system with only two degrees of freedom and continuous time, the only possible attractors are fixed points or periodic limit cycles. An example is a Van der Pol oscillator modeling a simple vacuum-tube oscillator circuit.

In continuous systems with a state space of dimension $n > 2$, more complex attractors are possible. Dynamical systems with *quasi-periodic limit cycles* show a time evolution which can be decomposed into different periodic parts without a unique periodic regime. The corresponding time series consist of periodic parts of oscillation without a common structure. Nevertheless, closely starting trajectories remain close to each other during time evolution. The third class contains dynamical systems with *chaotic attractors* which are non-periodic, with an exponential dependence on initial conditions for bounded orbits. A famous example is the chaotic attractor of a Lorenz system simulating the chaotic development of weather caused by local events, which cannot be forecast in the long run (*butterfly effect*).

2.2 Linear and Nonlinear Time Series Analysis

In the previous chapter we have analyzed dynamical systems and their types of behavior with fixed points, limit cycles, and chaos. Modeling means that these mathematical systems are applied to physical, biological or social systems of interest. The Lotka-Volterra equations, for example, constitute a mathematical system modeling the interaction of prey and predators in zoology. Modeling in this way is a top down procedure from mathematical equations to applications by appropriate interpretations of variables. In a bottom up approach, we start with a sequence of measurements and ask what the data themselves can tell us about the laws of dynamics. Sequences of data are called *times series*. Time series analysis is used to find types of appropriate equations fitting the data, or to compare the predictions of mathematical models to measurements made in the field of research.

In an ideal case, *time-series analysis* delivers a computer program providing a mathematical model fitting the measured data. But these data-generated models have a severe shortcoming, because they work without any understanding of the physical system. In practice, model building is combined with times-series analysis. Model building is based on knowledge of a physical system, while time-series analysis can be used to detect features of a system, inspiring model building.

Dynamical systems are governed by difference equations of the form $x_{t+1} = f(x_t)$ or differential equations of the form $dx/dt = g(x, y)$ and $dy/dt = h(x, y)$ with time-depending variables $x(t)$ and $y(t)$. In a *top-down approach* of modeling, the functions f, g, and h are given and the dynamical behavior with, e.g., fixed points, limit cycles, and chaos attractor is derived by mathematical analysis. In a *bottom-up approach*, we can only measure a limited set of quantities with limited precision. In our example of prey and predator dynamics, we might be able to measure the population of the predator only, although predator and prey are correlated and important for the dynamics of the whole prey and predator system.

For a mathematical model of observed data, we need an equation relating the measurements to the corresponding dynamical variables. The measurements approximate the dynamical variables with a difference which is called the *measurement error*. The measurement error depends on several factors like systematic bias, measurement noise, and dynamical noise. Systematic bias means a deficiency in the measurement process. Measurement noise results from random fluctuations in measurements. Dynamical noise is affected by outside influence, because dynamical systems are not isolated. A prey and predator system, for example, does not only depend on the two variables of prey and predator, but also on the environment with climate, nutrition, temperature et al.

Case Study (Linear Model of Dynamics) The dynamics of a finite-difference equation $x_{t+1} = A + \rho x_t$ has a steady state at $x_t = A/(1 - \rho) = M$ which is stable if $|\rho| < 1$. The solution to the finite-difference equation is exponential decay to the steady state. After the transient passes, there is steady-state behavior $x_t = M$. A direct measurement of the dynamical variable x_t is assumed. But, with respect to *measurement noise*, the measurement data at time t is $D_t = x_t + W_t$, where W_t is a

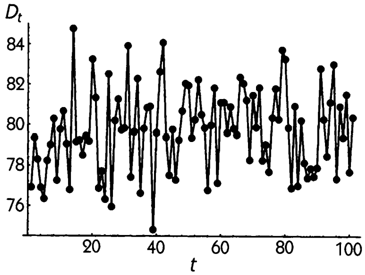


Fig. 3 Data dynamics of a linear model [6, p. 286]

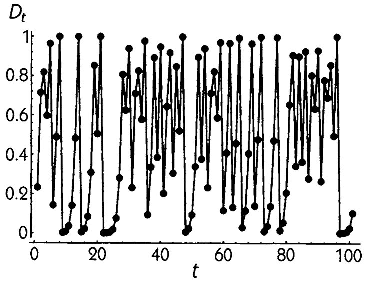


Fig. 4 Data dynamics of a nonlinear model [6, p. 302]

random number independently at each t in a Gaussian probability distribution with a mean of zero and standard deviation σ. Figure 3 shows data D_t generated by this model with $A = 4$, $\rho = 0.95$, and $M = 80$. P is Gaussian white measurement noise with a standard deviation of $\sigma = 2$.

The model describes a system maintained at a steady level (e.g., a population level or amount of prices at a market) without outside perturbations. For the interpretation of measured data, the model leads to following questions:

- What is the value of the steady state in the data?
- What is the level of measurement noise in the data?
- Is there evidence that there really is a steady state?
- Is there evidence that there is only measurement noise and no outside perturbations to the state x_t?

Case Study (Nonlinear Model of Dynamics) The previous model has *linear dynamics* and the stable fixed point is approached asymptotically in the absence of dynamical noise. *Nonlinear models* can have non fixed asymptotic behavior. For example, the quadratic map $x_{t+1} = \mu x_t(1 - x_t)$ can show a variety of behavior from stable fixed points to stable periodic cycles and chaos. The equation indicates no dynamical noise. Further on, there is no measurement noise, $D_t = x_t$ (Fig. 4). Thus, the model is completely deterministic. In this case all future data can be calculated for given initial conditions. In the case of chaos, there are practical limitations with respect to the sensitive dependence of the chaotic dynamics on initial data.

For a nonlinear model, the following questions may arise:

- What evidence is there that the data are generated by a deterministic process?

- What evidence is there for a nonlinear process?
- How large is the sensitive dependence on initial data in the case of chaos?

The mean of the data in Fig. 4 is $M_{est} = 0.471$. The fluctuations about the mean $V_t = D_t - M_{est}$ can be used to calculate the correlation coefficient between V_{t+1} and V_t. This is $\rho_{est} = 0.054$, close to zero. In fact, the autocorrelation function for the data of the nonlinear model is very similar to that for the data of the linear model. This suggests that the data from the nonlinear model are white noise, apparently contradicting the fact that the data are from a deterministic model. This paradox is solved by the fact that the correlation coefficient and the autocorrelation function measure linear correlations in the data. A scatter plot of V_{t+1} and V_t shows a very strong relationship, but actually the relationship is nonlinear and hence not accurately represented by the correlation coefficient and autocorrelation function.

Obviously, statistics of correlation coefficient and autocorrelation function cannot distinguish between the data in linear and nonlinear models. *Nonlinear time series analysis* helps to reconstruct nonlinear dynamics of a system from measured data. The idea of using a scatter plot to display the relationship between successive measurements is fundamental to the analysis of data from nonlinear systems. They are also called *return plot*, *Poincaré map*, or *return map*. In many cases, data are collected from a continuous-time dynamical system defined by differential equations rather than finite-difference equations. In these cases, it is appropriate to use the phase-plane or embedding reconstruction procedure to find the laws of dynamics from measured data.

Case Study (Harmonic Oscillator and Nonlinear Time Series Analysis) As an example, we consider a second-order differential equation describing a harmonic oscillator which is often used to model natural or economic systems with oscillating behavior (cf. Kaplan [6] p. 306): $d^2x/dt^2 = -bx$. In order to illustrate the flow of dynamics in a harmonic oscillator, this equation is rewritten with two first-order differential equations $dx/dt = y$ and $dy/dt = -bx$ for the variables x and y as coordinates of the phase plane of the system. In a *bottom-up approach*, we start with measuring a *time-series* $D(t) = x(t)$. In a next step, we must reconstruct the *state plane* and the flow on it from the measured data. At any instant, the position on the state plane is given by the coordinates (x, y) representing the state of the dynamical system at that instant. We can also measure $y(t)$ from $D(t)$ by noticing that $y = dx/dt = dD/dt$. If we plot dD/dt versus D, the trajectory in the state plane describes the flow based on the measured data.

But the harmonic oscillator is only a special case because dx/dt provides y. In general, dynamics on the state plane are given by a pair of coupled differential equations $dx/dt = f(x, y)$ and $dy/dt = g(x, y)$. Again, the question arises how to calculate the values of y if only $x(t)$ is measured. Measuring $x(t)$ and calculating dx/dt provide a direct measurement of x and a calculated value of $f(x, y)$. Some information about y is contained in the value of $f(x, y)$, and sometimes this information helps to an idea of the *whole dynamics* of the system.

Example (Chaotic Behavior and Weather Forecasting) Two-dimensional dynamics in a state plane cannot represent *chaotic behavior*. A continuous-time system generating chaos must consist of, at least, three equations. As an example, the Lorenz system of (simplified) weather forecasting is modeled by the three equations $dx/dt = 10(y - x)$, $dy/dt = 28x - y - xy$, and $dz/dt = 28xy - 8z/3$. If the values of $x(t)$, $y(t)$, and $z(t)$ can be measured simultaneously, it is easy to reconstruct the dynamics in a three-dimensional phase space. But if only one of the variables, e.g., $D(t) = x(t)$, can be measured, one must use heuristic procedures to reconstruct a model from measured data faithful to the geometry of the original.

2.3 Deterministic and Stochastic Models

Measurements are often contaminated by unwanted noise which must be separated from the signals of specific interest. Further on, in order to forecast the behavior of a system, the development of its future states must be reconstructed in a corresponding state space from a finite sequence of measurements. Thus, *time-series analysis* is an immense challenge in different fields of research from, e.g., climatic data in meteorology, ECG-signals in cardiology, and EEG-data in brain research to economic data of economics and finance. Beyond the patterns of dynamical attractors, randomness of data must be classified by statistical distribution functions.

Typical phenomena of our world, such as weather, climate, the economy and daily life, are much too complex for a simple deterministic description to exist. Even if there is no doubt about the deterministic evolution of, e.g., the atmosphere, the current state whose knowledge would be needed for a deterministic prediction contains too many variables in order to be measurable with sufficient accuracy. Hence, our knowledge does not usually suffice for a deterministic model. Instead, very often a stochastic approach is more situated. Ignoring the unobservable details of a complex system, we accept a *lack of knowledge*. Depending on the unobserved details, the observable part may evolve in different ways. However, if we assume a given probability distribution for the unobserved details, then the different evolutions of the observables also appear with specific probabilities. Thus, the lack of knowledge about the system prevents us from deterministic predictions, but allows us to assign probabilities to the different possible future states. It is the task of a time series analysis to extract the necessary information from past data.

Complex models contain nonlinear feedback, and the solutions to these are usually obtained by numerical methods (Bungartz et al. [1]). Statistical complex models are data driven and try to fit a given set of data using various distribution functions. There are also hybrids, coupling dynamic and statistical aspects, including deterministic and stochastic elements. Simulations are often based on computer programs, connecting input and output in nonlinear ways. In this case, models are calibrated by training the programs, in order to minimize the error between output and given test data.

Example (Power Laws and Risks) In the simplest case of statistical distribution functions, a *Gaussian distribution* has exponential tails situated symmetrically to the far left and right of the peak value. Extreme events (e.g., disasters, tsunamis, pandemics, worst case of nuclear power plants) occur in the tails of the probability distributions (Embrechts et al. [2]). Contrary to the Gaussian distribution, probabilistic functions $p(x)$ of heavy tails with extreme fluctuations are mathematically characterized by *power laws*, e.g., $p(x) \sim x^{-\alpha}$ with $\alpha > 0$. Power laws possess scale invariance corresponding to the (at least statistical) self-similarity of their time series of data. Mathematically, this property can be expressed as $p(bx) = b^{-\alpha} p(x)$ meaning that the change of variable x to bx results in a scaling factor independent of x while the shape of distribution p is conserved. So, power laws represent *scale-free* complex systems. The Gutenberg-Richter size distribution of earthquakes is a typical example of natural sciences. Historically, Pareto's distribution law of wealth was the first power law in the social sciences with a fraction of people presumably several times wealthier than the mass of a nation (Mainzer [8]).

3 Criteria of Risk Modeling in Philosophy of Science

3.1 What is a Good Model?

Mathematical modeling problems are often classified into black-box or white-box models, according to how much a priori information is available of the system. A *black-box model* is a system of which there is no a priori information available. A *white-box model* is a system where all necessary information is available. Practically all systems are somewhere between the black-box and white-box models, so this concept only works as an intuitive guide for approach.

Usually it is preferable to use as much a priori information as possible to make a model more accurate (cf. Gershenfeld [5]). Therefore the white-box models are usually considered easier, because if one has used the information correctly, then the model will behave correctly. Often the a priori information comes in forms of knowing the type of functions relating different variables. For example, if we make a model of how a climate model works in an ecological environment, we know that usually the amount of data is a varying function. Thus we are still left with several *unknown parameters*: how rapidly does pollution increase, and what is the initial state of the system? This example is therefore not a completely white-box model. These parameters have to be estimated through some means before one can use the model.

In black-box models one tries to estimate both the functional form of relations between variables and the numerical parameters in those functions. Using a priori information we could end up, for example, with a set of functions that probably could describe the system adequately. If there is no a priori information we would try to use functions as general as possible to cover all different models. The problem with using a large set of functions to describe a system is that estimating the parameters becomes increasingly difficult when the amount of parameters (and different types of functions) increases.

Another basic issue is the *complexity of a model*. If we were, for example, modeling the route of a railway plane, we could embed each mechanical part of the train into our model and would thus acquire an almost white-box model of the system. However, the computational cost of adding such a huge amount of detail would effectively inhibit the usage of such a model. Additionally, the uncertainty would increase due to a complex system, because each separate part induces some amount of variance into the model. It is therefore usually appropriate to make some approximations to reduce the model to a sensible size. Engineers often can accept some approximations in order to get a more robust and simple model. For example Newton's classical mechanics is an approximated model of the real world. Still, Newton's model is quite sufficient for most ordinary-life situations, that is, as long as particle speeds are well below the speed of light, and we study macro-particles only with respect to Einstein's theory of relativity and to quantum physics.

An important part of the modeling process is the *evaluation of an acquired model*. How do we know whether a mathematical model describes the system well? This is not an easy question to answer. Usually the engineer has a set of measurements from the system which are used in creating the model. Then, if the model was built well, the model will adequately show the relations between system variables for the measurements at hand. The question then becomes: how do we know that the measurement data is a representative set of possible values? Does the model describe well the properties of the system between the measurement data (interpolation)? Does the model describe well events outside the measurement data (extrapolation)?

Extrapolations are a challenge with increasing complexity of models. How well does this model describe events outside the measured data? Is it an adequate mapping of reality? Let us consider Newtonian classical mechanics-model, again. Newton made his measurements without advanced equipment, so he could not measure properties of particles travelling at speeds close to the speed of light. Likewise, he did not measure the movements of molecules and other small particles, but macro particles only. It is then not surprising that his model does not extrapolate well into these domains, even though his model is quite sufficient for ordinary life physics.

3.2 Model Skepticism—From David Hume to Nassim Taleb

Since Newton's century, there have been deep doubts in causality and the reliability of model-based predictions. An important progress of this criticism was the British philosopher David Hume (1711–1776) who was—like Adam Smith—one of the most important figures of Scottish Enlightenment. From a methodological point of view, Hume's critical analysis of human reason was a milestone in the history of philosophy. Kant mentioned that it was Hume waking him up from his "dogmatic slumbers". The problem concerns the question of how we are able to make inductive inferences. *Inductive inference* is reasoning from the observed behavior of objects to their behavior when unobserved. As Hume said, it is a question of how things behave when they go beyond the present test by our senses, and the records of our memory. He noticed that we tend to believe that things behave in a regular manner, i.e., that patterns in the behavior of objects will persist into the future, and throughout the unobserved present.

Hume's argument is that we cannot rationally justify the claim that nature will continue to be uniform, as justification only allows two arguments, and both of these are inadequate. According to Hume, the two sorts are: (1) *demonstrative reasoning*, and (2) *probable reasoning*. With regard to (1), Hume argues that the regularity of nature cannot be demonstrated, as, without logical contradiction, we can assume that nature might stop being regular. Considering (2), Hume argues that we cannot hold that nature will continue to be uniform because it has been in the past, as this is using the very sort of reasoning (induction) that is under question: it would be circular reasoning. Thus no form of justification will rationally warrant our inductive inferences.

Hume's solution to this skeptical problem is to argue that, rather than reason, it is natural *instinct* that explains our ability to make inductive inferences. He asserts that "All inferences from experience, therefore, are effects of custom, not of reasoning". (Hume [31]).

On the same line, the Lebanese philosophical essayist and practitioner of finance Nassim Taleb has argued in front of the recent financial crisis (Taleb [12]). His argument centers on the idea that predictive models are based on axiomatic "*Platonism*" (cf. Popper [47]), gravitating towards mathematical purity and failing to take some key ideas into account, such as: complete information is impossible, small unknown variations in the data could have a huge impact, and flawed models are based on empirical data without considering events that have not taken place but could have taken place. These rare and risky events are symbolized as "*black swans*" against the general belief that all swans are white. From a methodological point of view, Taleb follows Sir Karl Popper's philosophy of falsification (Popper [11]).

Logical Excursion (Falsification and Black Swans) In more details, Popper argues in the following way. A *general hypothesis* like "All swans are white" has the logical form "For all objects x is assumed: if x is a swan, then x is white". This general statement is especially true for a special object x_o, i.e. "If x_o is a swan, then x_o is white". Let the condition of this conclusion be true for a special object x_o, i.e. "x_o is a swan" is true. Then, our hypothesis *predicts* for the special swan x_o that it is white. This prediction follows by a logical direct conclusion (modus ponens): let A and B be propositions which can be either true or false. The *direct conclusion* (*modus ponens*) claims if A is true and conclusion $A \rightarrow B$ ("if A, then B") is true, then B is true. If we observe that the prediction is true, i.e. the observed swan x_o is actually white, then the general hypothesis is only *corroborated* by the example x_o, but not *verified* for *all* possible cases. In general, it is not possible to verify a general statement of empirical sciences for all possible objects, locations, and points of time. Only in mathematics, we can verify a general proposition on all natural numbers by a proof of complete induction. Therefore, according to Popper, a general hypothesis in empirical sciences can logically only be *falsified*, but not verified: Again, let the condition A ("x_o is a swan") be true. By observation, the swan is not white, but black, i.e. B is false. Then, the conclusion $A \rightarrow B$ must be false by logical reasons. In this case, the general hypothesis "All swans are white" is said to be falsified by the example x_o of a black swan.

The occurrence of black swans may be rare, but we must take black swan events into account. Therefore, according to Taleb, the foundations of quantitative economics are faulty and highly self-referential. He states that statistics is fundamentally incomplete as a field, as it cannot predict the risk of rare events, a problem that is acute in proportion to the rarity of these events. Taleb sees his main challenge as mapping his ideas of "*robustification*" and "*anti-fragility*", that is, how to live and act in a world we do not understand, and build robustness to black swan events. He advocates what he calls a "black swan robust" society, meaning a society that can withstand difficult-to-predict events. Like Hume he argues that, rather than mathematical modeling, it is natural instinct that explains our ability to make inductive inferences. He favors "*stochastic tinkering*" as a method of scientific discovery, by which he means experimentation and fact-collecting instead of top-down directed modeling.

3.3 Human Instinct, Probabilistic Thinking, and the Brain

Most of Taleb's critique could only be detected by sophisticated mathematical analysis. Thus, the question arises how Hume's and Taleb's confidence in *human instinct* can be sufficient in front of a world with increasing complexity. Traditionally, philosophy of science defended the belief in human rationality and the possibility of logical reasoning. Therefore, in the 20th century, logical empirism argued for scientific rules of inductive reasoning.

Logical Excursion (Inductive Logic) Since Isaac Newton, induction was proclaimed a fundamental method to derive a general natural law or hypothesis from observational data and measurements. Although there is no logical justification to derive a general proposition for all cases of a domain from some confirmed examples, logicians and philosophers of science suggested formal rules to handle the problem of induction. Rudolf Carnap (1891–1970) suggested a probabilistic calculus of hypotheses. The probability of a hypothesis h is defined as degree of belief in h with respect to given data of experience e. The task of inductive logic is the definition of a function of confirmation $c(h, e) = r$, which correlates an inductive resp. a priori-probability r to the proposition h. Carnap's c-function was defined on elementary propositions, complex propositions of logically connected elementary propositions, and general propositions for infinite many cases (e.g., all space-time points). But his axioms were too weak for practical applications. Thus, he did not longer rely in one unique inductive method, but suggested a class of different confirmation functions. Anyway, in modern philosophy of science, probabilistic arguments and the meaning of probability play a crucial role. To evaluate the probability of a hypothesis, the concept of *Bayesianism* assumes some prior probability, which is then updated with respect to new data (Hacking [28]).

Carnap also initiated a logical theory of rational decisions under risk. The degree of belief of a person at time T is defined by a belief function *Cr* which is interpreted as betting quotient. Obviously, this approach makes no sense in the natural sciences, because natural laws do not depend on betting. In social sciences, decisions under risks depend on personal degrees of belief which Carnap assumed to be measurable. But, in modern brain research and cognitive science, gut feeling is no longer only a source of irrationality. New insights in human intuition and unconscious experience lead to behavioral skills which are even useful in management. The philosopher of science Michael Polyani (1891–1976) introduced the term "*tacit knowledge*", in order to describe these unconscious abilities. Polyani argued that we sometimes cannot only more than we can express by language, but that all kind of knowledge is based on tacit knowledge (Polyani [46]). Daily activities like car driving or the routines of our jobs are rooted in unconscious abilities which were trained and learnt in earlier time. These schemes of behavior let us react under stress and risk. Without trust in these abilities, we would not be able to act under risk. Modern brain research and cognitive science are extremely interested to understand these mechanisms. Therefore, *experimental* and *behavior-oriented economics* as well as *neuroeconomics* provide important tools to complement mathematical risk modeling (Fehr [23]).

4 Classical Risk Modeling in Financial and Insurance Mathematics

In economics as well as in financial theory uncertainty and information incompleteness prevent exact predictions. A widely accepted belief in financial theory is that time series of asset prices are unpredictable. Chaos theory has shown that unpredictable time series can arise from deterministic nonlinear systems. The results obtained in the study of physical, chemical, and biological systems raise the question whether the time evolution of asset prices in financial markets might be due to underlying nonlinear deterministic dynamics of a finite number of variables. If we analyze financial markets with the tools of nonlinear dynamics, we may be interested in the reconstruction of an attractor. In time series analysis, it is rather difficult to reconstruct an underlying attractor and its dimension. For chaotic systems, it is a challenge to distinguish between a chaotic time evolution and a random process, especially if the underlying deterministic dynamics are unknown. From an empirical point of view, the discrimination between *randomness* and *chaos* is often impossible. Time evolution of an asset price depends on all the information affecting the investigated asset. It seems unlikely that all this information can easily be described by a limited number of nonlinear deterministic equations.

4.1 Beginning of Insurance Mathematics: Poisson Distribution of Risks

Mathematical modeling in finance and insurance can be traced back for centuries. Insurance of risks against the chances of life is an old topic of mankind (cf. Mainzer [8]). *Commercial insurance* dates back to Renaissance, when great cities of trading introduced bets on safe routes of ships. In the 17th century, the great British insurance company Lloyd arose from this system of bookmakers. The philosopher and mathematician Gottfried Wilhelm Leibniz (1646–1716) already suggested a health insurance in which people should pay with respect to their income. In Germany, the ingenious idea of Leibniz was realized not earlier than in the 19th century by Bismarck. In the time of Leibniz, life insurances were first applications of probability calculations.

Historical Excursion (Huygens and Insurances in the 17th Century) The Dutch physicist Christiaan Huygens (1629–1695) applied the law of large numbers to calculations of insurance rates. In his approach, an insurance is considered as a game between the insurer and clients. The insurer diminishes his risk by adapting the premium payed by a client. Let $c_1, \ldots, c_n$ be the costs of the insurer and $p_1, \ldots, p_n$ the probabilities that the damages happen. The expected damage of the insurer is assumed to be $p_1c_1 + \cdots + p_nc_n$. The average gain is equal to the premium Q paid by the clients. His *risk* is zero for a premium $Q = p_1c_1 + \cdots + p_nc_n$. The risk of clients is also zero, their loss Q and the expected gain $p_1c_1 + \cdots + p_nc_n$. In this case, Q is called a fair premium to be paid by clients. It is assumed that the probabilities $p_1, \ldots, p_n$ can be estimated according to the law of large numbers. But this assumption was the flaw of Huygens' approach. The law of large numbers cannot be applied in cases of rare damages with extreme costs.

In 1898 the Russian economist and statistician Ladislaus Josephovich Bortkiewicz (1868–1931) published a book about the Poisson distribution, titled *The Law of Small Numbers*. In this book he first noted that events with low frequency in a large population follow a Poisson distribution even when the probabilities of the events varied. Modern insurance mathematics started with the thesis of the Swedish mathematician Filip Lundberg (1876–1965). He introduced the *collective risk m*odel for insurance claim data. Lundberg showed that the homogeneous Poisson process, after a suitable time transformation, is the key model for insurance liability data. Risk theory deals with the modeling of claims that arrive in an insurance business and which gives advice on how much premium has to be charged in order to avoid ruin of the insurance company. Lundberg started with a simple model describing the basic dynamics of a *homogeneous insurance portfolio*.

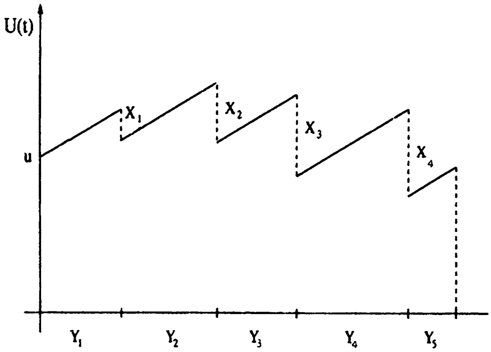


Fig. 5 A realization of Lundberg's risk process [2, p. 9]

Lundberg's Model of a Homogeneous Insurance Portfolio This means a portfolio of contracts for similar risks (e.g., car or household insurance) under three assumptions:

- Claims happen at time T_i satisfying $0 \leq T_1 \leq T_2 \leq T_3 \leq \cdots$ which are called claim arrivals.
- The ith claim arriving at time T_i causes the claim size. The latencies between the claim arrivals T_i are iid (exponential) distributed.
- The claim size process (X_i) and the claim arrival process (T_i) are mutually independent.

According to Lundberg's model, the *risk process* $U(t)$ of an insurance company is determined by the *initial capital* u, the *loaded premium rate* c and the *total claim amount* $S(t)$ of claims X_i with $U(t) = u + ct - S(t)$ and $S(t) = \sum_{i=1}^{N(t)} X_i (t \geq 0)$. $N(t)$ is the *number of the claims* that occur until time t. Lundberg assumed that $N(t)$ is a *homogeneous Poisson process*, independent of (X_i). Figure 5 illustrate a realization of the risk process $U(t)$.

Lundberg's model is fine for *small claims*. But the question arises how the global behaviour of $U(t)$ is influenced by individual extreme events with *large claims*. Under Lundberg's condition of small claims, Harald Cramér estimated bounds for the *ruin probability* of an insurance company which are exponential in the initial capital u. Actually, claims are mostly modeled by *heavy-tailed distributions* like, e.g., Pareto which are much heavier than exponential.

4.2 *Beginning of Financial Mathematics: Gaussian Distribution of Risks*

With the up-coming stock markets during the period of industrialization, people became more and more interested in their risky dynamics. Asserts price dynamics are assumed to be stochastic processes. An early key-concept to understand stochastic processes was the random walk. The first theoretical description of a *random walk* in the natural sciences was performed in 1905 by Einstein's analysis of molecular interactions. But the first mathematization of a random walk was not realized in physics, but in social sciences by the French mathematician Louis Jean Bachelier (1870–1946). In 1900 he published his doctoral thesis with the title "*Théorie de la Spéculation*" [17]. During that time, most market analysis looked at stock and bond prices in a causal way: something happens as cause and prices react as effect. In complex markets with thousands of actions and reactions, a causal analysis is even difficult to work out afterwards, but impossible to forecast beforehand. One can never know everything. Instead, Bachelier tried to estimate the odds that prices will move. He was inspired by an analogy between the diffusion of heat through a substance and how a bond price wanders up and down. In his view, both are processes that cannot be forecast precisely. At the level of particles in matter or of individuals in markets, the details are too complicated. One can never analyze exactly how every relevant factor interrelate to spread energy or to energize spreads. But in both fields, the broad pattern of probability describing the whole system can be seen.

Bachelier introduced a stochastic model by looking at the bond market as a *fair game*. In tossing a coin, each time one tosses the coin the odds of heads or tails remain 1:2, regardless of what happened on the prior toss. In that sense, tossing coins is said to have no memory. Even during long runs of heads or tails, at each toss the run is as likely to end as to continue. In the thick of the trading, price changes can certainly look that way. Bachelier assumed that the market had already taken account of all relevant information, and that prices were in equilibrium with supply matched to demand, and seller paired with buyer. Unless some new information came along to change that balance, one would have no reason to expect any change in price. The next move would as likely be up as down.

Actually, prices follow a *random walk*. Imagine a blind drunk staggering across an open field. How far will he have gotten after some time? He could go one step left, two steps right, three backwards, and so on in an aimless path. On average, just as in tossing coins, he gets nowhere. On the average, his random walk will be forever stuck at his starting point. In the same way, the prices on markets can go up or down, by big increments or small. With no new information to push a price in one direction or another, a price on average will fluctuate around its starting point. In that case, the best forecast is the price today. Each variation in price is unrelated to the last. In a stochastic model, the price-changes form a sequence of *independent* and *identically distributed random variables*. In that case, a chart of changes in price from moment to moment illustrates a more or less *uniform distribution over time*. The size of most price changes varies within a narrow range. There are also bigger fluctuations. But they barely stand up from the bulk of changes, as some outliers of grass rise above the average height of an unmown lawn, in that most of the blades of grass fall within a narrow range of heights, while a minority rise above this range (Mainzer [8], Mandelbrot and Hudson [10]).

In order to illustrate this smooth distribution, Bachelier plotted all of a bond's price-changes over a month or year onto a graph. In the case of independent and identically distributed price-changes, they spread out in the well-known bell-curve shape of a *normal* ("*Gaussian*") *distribution*: the many small changes clustered in the center of the bell, and the few big changes at the edges. Bachelier assumed that price changes behave like the random walk of molecules in a Brownian motion. Long before Bachelier and Einstein, the Scottish botanist Robert Brown had studied the way that tiny pollen grains jiggled about in a sample of water. Einstein explained it by molecular interactions and developed equations very similar to Bachelier's equation of bond-price probability, although Einstein never knew that. In 1923 (*Journal of Mathematical Physics* 2, 131–174), Norbert Wiener proved the existence of Brownian motion and considered advanced related mathematical theories. Therefore, Brownian motion is also called a Wiener process. It is a remarkable *interdisciplinary coincidence* that the *movement of security prices*, the *motion of molecules*, and the *diffusion of heat* are described by mathematically analogous models.

Bachelier's Hypotheses of Price Changes In short, Bachelier's model depends on the three hypotheses of (1) *statistic independence* (i.e., each change in price appears independently from the last), (2) *statistic stationarity* of price changes, and (3) *normal distribution* (i.e., price changes follow the proportions of the Gaussian bell curve).

4.3 Models of Efficient Markets and Computable Risks

But it took a long time that economists recognized the practical virtues of describing markets by the laws of chance and Brownian motion (Mainzer [8], Mandelbrot and Hudson [10]). In 1956, Bachelier's idea of a fair game was used by Paul A. Samuelson and his school to formulate the *Efficient Markets Hypothesis*. They argued that in an ideal market, security prices fully reflect all relevant information. A financial market is a fair game in which buyer balances seller. By reading price charts, analyzing public information, and acting on inside information, the market quickly discounts the new information that results. Prices rise or fall to reach a new equilibrium of buyer and seller. The next price change is, once again, as likely to be up as down. So, one can expect to win half the time and loose half the time. If one has special insights into a stock, one could profit from being the first in the market to act on it. But one cannot be sure to be right or first, because there are many clever people in a market as intelligent as oneself.

Since Samuelson Bachelier's theory was not only elaborated into a mature theory of how prices vary and how markets work. It was more important for the financial world that the theory has been translated into practical tools of finance. In the 1950s, Markowitz [43] was inspired by Bachelier to introduce *Modern Portfolio Theory* (MPT) as a method for selecting investments. In the early 1960s, Sharpe [51] devised a method of valuing an asset, called Capital Asset Pricing Method (CAPM). A third tool is the *Black-Scholes formula* for valuing options contracts and assessing risk. Its inventors were Black and Scholes [18] in the early 1970s. These three innovations, CAPM, MPT, and Black-Scholes, are still the fundamental tools of classical financial theory until today, resting on Bachelier's hypotheses of financial markets.

Black-Scholes Conditions of Financial Markets The Black-Scholes formula tries to implement risk-free portfolios. Black and Scholes assumed several conditions of financial markets: (1) The change of price $Y(t)$ at each step t can be described by the stochastic differential equation of a *geometric Brownian motion*. This assumption implies that the changes in the (logarithm of) price are Gaussian distributed. (2) *Security trading* is continuous. (3) *Selling of securities* is possible at any time. (4) There are no *transaction costs*. (5) The *market interest rate* r is constant. (6) There are no *dividends* between $t = 0$ and $t = T$ (maturity).

(7) There are no *arbitrage opportunities*. Arbitrage is a key concept for the understanding of markets. It means the purchase and sale of the same or equivalent security in order to profit from price discrepancies. A stock may be traded in two different stock exchanges in two different countries with different currencies. By buying several shares of the stock in New York and selling them in Frankfurt, the arbitrager makes a profit apart from the transaction costs. Traders looking for arbitrage opportunities contribute to a market's ability to evolve the most rational price for a good. The reason is obvious: if someone has discovered an arbitrage opportunity and succeeded in making a profit, he will repeat the same action. After carrying out this action repeatedly and systemically for several opportunities, the prices will be adapted and no longer provide arbitrage opportunities. In short: New arbitrage opportunities continually appear in markets. But as soon as they are discovered, the market moves in a direction to eliminate them gradually (Mandelbrot and Hudson [10]).

Now, in the absence of arbitrage opportunities, the change in the value of a portfolio must equal the (expected) gain obtained by investing the same amount of money in a riskless security providing a return per unit of time. The assumed dynamics of prices allows to derive the *Black-Scholes partial differential equation* which is valid for both call and put European options. Under some boundary conditions and substitutions the Black-Scholes partial differential equation becomes formally equivalent to the heat-transfer equation of physics which is analytically solvable.

Assumptions of Classical Economic Models These financial tools are deeply rooted in assumptions of classical economic models, but refuted by observables of real human behavior (Mandelbrot and Hudson [10]):

1. Assumption: People are rational in the sense of Adam Smith's *homo oeconomicus*. Consequently, when presented with all the relevant information about a stock or bond, investors will make the obvious rational choice leading to the greatest possible wealth and happiness. Their preferences can be expressed in mathematical formulas of utility functions which can be maximized. By that, rational investors make a rational model of an efficient market. Actually, people do not

only think in terms of mathematical utility functions, and are not always rational and self-interested. They are driven by emotions distorting their decisions. Sometimes, they miscalculate probabilities and feel differently about loss than gain.

2. Assumption: *All investors are alike*. Consequently, people have the same investment goals and react and behave in the same manner. In short: they are like the molecules in an idealized gas of physics. An equation that describes one such molecule or investor can be replaced to describe all of them. Actually, people are not alike. If one drops the assumption of homogeneity, one gets a more complex model of the market. For example, there are at least two different types of investors: a fundamentalist believes that each stock has its own value and will eventually sell for that value. On the other side, a chartist ignores the fundamentals and only watchs the price trends in order to jump on or off band waggons. Their interactions can lead to price bubbles and spontaneously arising crashes. The market switches from a well-balance linear system in which one factor adds predictably to the next, to a chaotic nonlinear system in which factors interact with the emergence of synergetic and unanticipated effects.
3. Assumption: *Price change is practically continuous*. Consequently, stock quotes or exchange rates do not jump up or down, but move smoothly from one value to the next. In this way, continuity has been assumed in classical physics, according to the motto of Leibniz “natura non facit saltum” (nature does not make leaps) which was repeated by Alfred Marshall in his text book “Principles of Economics” (1890) for economic systems. From a methodological point of view, the belief in a continuous behavior of nature and economy opens the possibility to apply continuous functions and differential equations, in order to solve physical or economic problems analytically. But actually, prices in economy and quantum states in quantum physics do jump, and discontinuity, far from being an anomily, characterizes the reality. Contrary to Einstein’s famous objection against quantum physics: god plays with dice—in nature and society.
4. Assumption: *Price changes follow* a *Brownian motion*. The Brownian motion is also a famous model of physics applied to financial markets by Bachelier. In more details, it implies three assumptions: first, each change in price is believed to appear independently from the last (statistical independence). Second, the process generating price changes stays the same over time (statistical stationarity). Third, price changes follow the proportions of the Gaussian bell curve (normal distribution). Financial data clearly contradict to a smooth normal distribution of changing prices. The analysis of the real distribution patterns is a challenge of stochastic mathematics and systems theory and opens new avenues to the complexity of modern society.

4.4 Securitized Credit Model and Increasing Networks of Risks

Nevertheless, the demand for profit and security has initiated a wave of financial innovation, based on these classical assumptions. They are focused on the origination, packaging, trading and distribution of *securitised credit instruments*. Simple forms of securitised credit have existed for almost as long as modern banking. But from the mid-1990s the system entered explosive growth in both scale and complexity. We observe a huge growth in the value of the total stock of credit securities, an explosion in the complexity of the securities sold, with the growth of structured credit products, and with the related explosion of the volume of credit derivatives, enabling investors and traders to hedge underlying credit exposures, or to create synthetic credit exposures.

This financial innovation sought to satisfy the demand for yield uplift. It was predicated on the belief that by slicing, structuring and hedging, it was possible to create value, offering investors combinations of risk, return, and liquidity which were more attractive than those available from the direct purchase of the underlying credit exposures. It resulted not only in massive growth in the importance of securitised credit, but also in a profound change in the nature of the securitised credit model. As securitisation grew in importance from the 1980s on, its development was praised as a means to reduce banking system risks and to cut the total costs of credit intermediation, with credit risk passed through to end investors, reducing the need for unnecessary and expensive bank capital. Credit losses would be less likely to produce banking system failure (Turner [54]).

But there is no "*free lunch*" or financial "*perpetuum mobile*". When the crisis broke, it became apparent that this diversification of risk holding had not actually been achieved. Instead most of the holdings of the securitised credit, and the vast majority of the losses which arose, were not in the books of end investors intending to hold the assets to maturity, but on the books of highly leveraged banks and bank-like institutions. This reflected an evolution of the securitised credit model away from the initial descriptions. To an increasing extent, credit securitised and taken off one bank's balance sheet, was not simply sold through to an end investor, but bought by the propriety trading desk of another bank, sold by the first bank but with part of the risk retained via the use of credit derivatives, resecuritised into increasingly complex instruments (e.g. CDOs and CDO squareds) or used as collateral to raise short-term liquidity (International Monetary Fund [32]).

The financial innovations of structured credit resulted in the creation of products, e.g. the lower credit tranches of CDOs or even more so of CDO-squareds, which had very high and imperfectly understood embedded leverage, creating positions in the trading books of banks which were hugely vulnerable to shifts in confidence and liquidity. This process created a complex chain of multiple relationships between multiple institutions, each performing a different small slice of the credit intermediation and maturity transformation process, and each with a leveraged balance sheet requiring a small slice of capital to support that function (Sinn [52]). A *complex network of dependences* has emerged in a hidden and intransparent world of financial shadows. The new model left most of the risk still somewhere on the balance sheets of banks and bank-like institutions but in a much more complex and less transparent way.

The evolution of the *securitised credit model* was accompanied by a growth in the relative size of financial services within economy, with activities internal to the banking system growing far more rapidly than end services to the real economy. The growing size of the financial sector was accompanied by an increase in total system leverage. But this process also drived the boom and created vulnerabilities of the whole financial network that have increased the severity of the crisis. According to the Turner Report [54], from about 2003 onwards, there were significant increases in the measured on-balance sheet leverage of many commercial and investment banks, driven in some cases by dramatic increases in gross assets and derivative positions. This was despite the fact that measures of leverage (e.g. Value at Risk (VaR) relative to equity) showed no such rise. This divergence reflected the fact that VaR measures of the risk involved in taking propriety trading positions, in general suggested that risk relative to the gross market value of positions had declined. It is clear in retrospect that the VaR measures of risk were faulty (Stutz [53]).

4.5 The Risk of Value at Risk (VaR)

The increasing complexity of the securitised credit market was obvious to some participants, regulators and academic observers (Greenspan [27]). But the predominant assumption was that increased complexity had been matched by the evolution of mathematically sophisticated and effective techniques for measuring and managing the resulting risks (Colander et al. [19]). Central to many of the techniques was the concept of *Value-at-Risk* (VaR), enabling inferences about forward-looking risk to be drawn from the observation of past patterns of price movement. The *risk-forecasting models* of value-at-risk (VaR) are based on the assumption that forecasting credit risk is an activity not unlike that of forecasting weather. It is assumed that one’s own action, based on past volatility, does not affect future volatility itself just like forecasting weather does not influence future weather.

This technique, developed in the early 1990s, was not only accepted as standard across the industry, but adopted by regulators as the basis for calculating trading risk and required capital. Therefore, VaR was incorporated within the European Capital Adequacy Directive (Danielsson et al. [21]). In financial mathematics and financial risk management, Value at Risk (VaR) is a widely used risk measure of the risk of loss on a specific portfolio of financial assets. For a given portfolio, probability and time horizon, VaR is defined as a *threshold value* such that the probability that the mark-to-market loss on the portfolio over the given time horizon exceeds this value in the given probability level. VaR has five main uses in finance: risk management, risk measurement, financial control, financial reporting and computing regulatory capital (Kleeberg and Schlenger [33]). VaR is sometimes used in non-financial applications as well. Important related ideas are economic capital, backtesting, stress testing and expected shortfall.

Mathematical Definition of VaR Mathematically (Föllmer and Schied [3, 24]; compare also Chap. 5 of Biagini et al.), the uncertainty in the future of a portfolio is usually described by a function $X : \Omega \to R$, where Ω is a fixed set of scenarios.

For example, X can be the *value of a portfolio*. The goal is to determine a number $\rho(X)$ that quantifies the risk and can serve as a capital requirement or the minimal amount of capital which, if added to the position and invested in a risk-free manner, makes the position acceptable. Given some *confidence level* $\alpha \in (0, 1)$, the *Value at Risk* (VaR) of the portfolio value X at the confidence level α is given by the smallest number $m \in R$ such that the probability of a loss is not larger than the confidence level α:

$$VaR_\alpha(X) = \inf\{m \in R \mid P(X + m < 0) \leq \alpha\}.$$

Obviously, value at risk (VaR) only pays attention that the boundary of the confidence level is not exceeded. But, it does not consider the degree of loss. Further on, it assumes that the probability distribution of losses is well-known because of historical data. Only in this case value at risk (VaR) can forecast credit risk like weather, which means that future volatility can be derived from past volatility.

There are, however, fundamental questions about the validity of VaR as a measure of risk. The use of VaR measures based on relatively short periods of historical observation (e.g. 12 months) introduced dangerous procyclicality into the assessment of trading book risk (Turner [51]). Short-term observation periods and the assumption of normal distribution can lead to large *underestimation of probability of extreme loss events*. Interconnected market events in complex networks can produce *self-reinforcing cycles* which models do not capture. *Systemic risk* may be highest when measured risk is lowest, since low measured risk encourages behavior which creates increased systemic risks.

This kind of mathematics, used to measure and manage risk by VaR, was not very well understood with all its conditions and restrictions by top management and boards to assess and exercise judgement over the risks being taken. Mathematical sophistication ended up not containing risk, but providing false assurance that other indicators of increasing risk (e.g. rapid credit extension and balance sheet growth) could be safely ignored.

The global financial system, combining with macroeconomic imbalances, created an *unsustainable credit boom* and *asset price inflation*. Those consequences of the financial crisis transmitted financial system problems into real economy effects. The shock to the banking system has been so great that its impaired ability to extend credit to the real economy has played a major role in enforcing the economic downturn, which in turn undermines banking system strength in a self-reinforcing feedback loop.

From a historical point of view, it is remarkable that the academic professionals were well aware of the methodological weakness of VaR measures. In an "Academic Response to Basel II" [21], the methodology of value-at-risk (VaR) was criticized to be insufficient: (1) VaR risk models treat risk as a fixed *exogenous* process, but its *endogeneity* may matter enormously in times of crisis. (2) VaR is a misleading risk measure when the returns are not *normally distributed*, as in the case with credit, market, and operational risk. It does not measure the distribution of risk in the tail, but only provides an estimate of a particular region in the distribution. Thus, VaR models generate imprecise and widely fluctuating forecasts.

5 New Paradigm of Risk Modeling and Rational Behavior

The development of an expanded financial sector and the rapid growth and increased complexity of the securitised model of credit intermediation was accompanied by the development of increasingly sophisticated mathematical techniques for the measurement and management of position taking risks. The techniques entailed numerous variants to cope with, for instance, different categories of option. Their application required significant computing power to capture relationships between different market prices, the complex nature of structured credit instruments, and the effects of diversification across correlated markets. But the underlying methodological assumption was the old idea that analysis of past price movement patterns could deliver statistically robust inferences relating to the probability of price movements in future.

5.1 Crisis of Risk Modeling

The financial crisis has revealed, however, severe problems with these techniques. They suggest the need for significant changes in the way that VaR-based methodologies have been applied. But, the most fundamental question concerns our ability in principle to infer *future risk* from *past observed patterns*. Can financial models still be considered true mappings of an external world in order to derive predictions of future events like in the natural sciences? (Lux and Westerhoff [37].)

Models in the tradition of Bachelier assume that the distribution of possible events, from which the observed price movements are assumed to be a random sample, is normal with the shape of a Gaussian bell curve. But there is no clearly robust justification for this assumption. Actually, the financial market movements are inherently characterized by *fat-tail distributions*. This implies that any use of VaR models needs to be analyzed by the application of stress test techniques which consider the impact of extreme movements beyond those which the model suggests are at all probable.

One explanation of fat-tail distributions may lie in the complex networks of financial dependences. VaR models implicitly assume that the actions of the individual firm, reacting to market price movements, are both sufficiently small in scale as not themselves to affect the market equilibriums, and independent of the actions of other firms. But this is a deeply misleading assumption if it is possible that developments in markets will induce similar and simultaneous behavior by numerous players. If this is the case, which it certainly was in the financial crisis, VaR measures of risk may not only fail adequately to warn of rising risk, but may convey the message that risk is low and falling at the precise time when systemic risk is high and rising.

For example, according to VaR measures, risk was low in spring 2007. Actually, the system was overwhelmed with huge *systemic risk*. This suggests that *stress tests* are needed to consider the impact of second order effects, for example, the impact on one bank of another bank's likely reaction to the common systemic stress.

5.2 A New Paradigm of Risk Modeling

The most fundamental insight is, however, philosophical: it is important to realize that the assumption that past distribution patterns carry robust inferences for the probability of future patterns is methodologically insecure. It involves applying to the world of social and economic relationships a technique drawn from the world of physics, in which a random sample of a definitively existing universe of possible events is used to determine the probability characteristics which govern future random samples. But it is doubtful when applied to economic and social decisions with inherent uncertainty. Economists sometimes refer to it as "*Knightian*" uncertainty which is a reference to the classical distinction between risk and uncertainty in Frank Knights' Ph.D. "Risk, Uncertainty, and Profit" [34] from 1921. But it would also suggest that no system of regulation could ever guard against all risks and uncertainties.

Analysis of the causes of the crisis suggests that there is a limit to the extent to which risks can be identified and offset at the level of the individual firm. We explained how the origins of the crisis lay in systemic developments: the crucial shift required in regulatory philosophy is towards one which focuses on macro-analysis, *systemic risks* and judgements about business model sustainability, and away from the assumption that all risks can be identified and managed at a firm specific level. As a result most of the changes we propose relate to the redesign of global regulation combined with a major shift in methodology (Colander et al. [19]).

But improvements in the effectiveness of *internal risk management* and *firm governance* are also essential. While some of the problems could not be identified at firm specific level, and while some well run banks were affected by systemic developments over which they had no influence, there were also many cases where internal risk management was ineffective and where boards failed adequately to identify and constrain excessive risk taking. Achieving high standards of risk management and governance in all banks is therefore essential. Detailed proposals are necessary to support an FSA (Financial Service Authority) in all countries.

The origins of the past crisis entailed the development of a complex, highly leveraged and therefore risky variant of the securitised model of credit intermediation. Large losses on structured credit and credit derivatives, arising in the trading books of banks and investment banks, directly impaired the capital position of individual banks, and because of uncertainty over the scale of the losses, created a *crisis of confidence* which produced *severe liquidity strains* across the entire system. As a result, a wide range of banking institutions suffered from an impaired ability to extend credit to the real economy, and have been recapitalized with large injections of taxpayer money.

The mathematical rigor and numerical precision of risk management and asset pricing tools has a tendency to conceal the weakness of models and their assumptions to those who have not developed them and do not know the potential weakness of the assumptions. *Models* are only approximations to the real world dynamics and partially built upon *idealized assumptions*. A typical example is the belief in normal distribution of asset price changes completely neglecting the importance of extreme events. Considerable progress has been made by moving to more sensitive models with *fat-tailed Lévy processes* (Mandelbrot [41]). Of course, such models better capture the intrinsic volatility of markets. But they might again contribute to enhancing the control illusion of the naïve user.

Therefore, market participants and regulators have to become more sensitive towards the potential weakness of risk management models. Since there is not only one true model, robustness should be a key concern. *Model uncertainty* should be taken into account by applying more than a single model. For example, one could rely on probabilistic procedures that cover a whole class of specific models. The theory of robust control provides a toolbox of techniques that could be applied for this purpose.

5.3 Convex Models of Risk

In the field of *financial economics* there are a number of ways that risk can be defined (Marrison [40]). To clarify the concept mathematicians have axiomatically described a number of properties that *a risk measure* might or might not have (Föllmer and Schied [24], York [55]).

Mathematical Definition of Coherent Risk Measure A *coherent risk measure* (Artzner et al. [16]) is a risk measure ρ that satisfies properties of *monotonicity*, *subadditivity*, *homogeneity*, and *translational invariance*. Consider a random outcome X viewed as an element of a linear space L of measurable functions, defined on an appropriate probability space. A functional $\rho : L \rightarrow R$ is said to be a coherent risk measure for L if it satisfies the following properties:

Monotonicity: If $X_1, X_2 \in L$ and $X_1 \leq X_2$, then $\rho(X_1) \leq \rho(X_2)$.

That is, if portfolio X_2 always has better values than portfolio X_1 under all scenarios then the risk of X_2 should be less than the risk of X_1.

Sub-additivity: If $X_1, X_2 \in L$, then $\rho(X_1 + X_2) \leq \rho(X_1) + \rho(X_2)$.

Indeed, the risk of two portfolios together cannot get any worse than adding the two risks separately. This is the diversification principle.

Positive homogeneity: If $\alpha \geq 0$ and $X \in L$, then $\rho(\alpha X) = \alpha\rho(X)$.

Loosely speaking, if you double your portfolio then you double your risk.

Translation invariance: If $m \in R$ and $X \in L$, then $\rho(X + m) = \rho(X) - m$.

The value m is just adding cash to the portfolio X, which acts like an insurance. The risk of $X + m$ is less than the risk of X, and the difference is exactly the added cash m. Therefore, translational invariance is also called cash invariance. In particular, if $m = \rho(X)$ then $\rho(X + \rho(X)) = 0$.

The notion of coherence has been subsequently relaxed. Indeed, the notions of sub-additivity and positive homogeneity can be replaced by the notion of convexity:

Convexity: If $X_1, X_2 \in L$ and $0 \leq \lambda \leq 1$, then $\rho(\lambda X_1 + (1-\lambda)X_2) \leq \lambda\rho(X_1) + (1-\lambda)\rho(X_2)$.

Consider the collection of possible future outcomes that can be generated with the resources available to an investor. One investment strategy leads to X_1, while a second strategy leads to X_2. If one diversifies, spending only the fraction λ of the resources on the first possibility and using the remaining part for the second alternative, one obtains $\lambda X_1 + (1-\lambda)X_2$. Thus, the axiom of convexity gives a precise meaning to the idea that diversification should not increase the risk.

It is well known that *value at risk* (VaR) is positively homogeneous, but it is not in general a coherent risk measure as it does not respect the sub-additivity property. Hence, it is *not convex*. An immediate consequence is that *value at risk* might discourage diversification. *Value at risk* is, however, *coherent*, under the assumption of *normally distributed* losses when the portfolio value is a linear function of the asset prices. However, in this case the value at risk becomes equivalent to a mean-variance approach where the risk of a portfolio is measured by the variance of the portfolio's return. *Average value at risk at level* $\lambda \in (0, 1]$,

$$AVaR_\lambda = \frac{1}{\lambda}\int_0^\lambda VaR_\alpha(X)d\alpha$$

also called *conditional value at risk*, expected shortfall, or tail value at risk, is a coherent risk measure (Detlefsen and Scandolo [22], Riedel [48]).

We previously underlined that model uncertainty should be taken into account, since we do not know the distinguished true model of financial reality. Therefore, we should consider a whole class of possible probabilistic models with different penalty. In the dual representation theory of convex risk measures one aims at deriving their representation in a systematic manner. The class M contains possible *probabilistic models* Q which are taken more or less seriously according to the size of a *penalty function* $\pi(Q)$. In this way, we take the message of praxis seriously that we should not rely on one single model, but flexibly vary the models with respect to different contextual applications under special attention to the worst case.

Mathematical Definition of Convex Risk Measure A dual representation of a *convex risk measure* computes the *worst case expectation* taken over all *models* Q and *penalized* by $\pi(Q)$. The class M of possible probabilistic models is a set of probability measures such that the *expectation* $E_Q(X)$ is well defined for all models Q and *portfolios* X. According to Föllmer and Schied [21], the dual representation of a convex risk measure ρ has the form

$$\rho(X) = \sup_{Q \in M}\big(E_Q(-X) - \pi(Q)\big).$$

These models are no longer considered definitive mappings of reality. But they serve as stress tests. One does not rely on a fixed model, but chooses the sure side for every position and focuses on the corresponding worst case model. Thus, the model ambiguity is explicitly considered during the procedure.

5.4 Model Ambiguity and Rational Behavior

Model ambiguity is linked to the economic theory of rational behavior under uncertainty (Cont [20], Maccheroni [38]). Classical economic models are mainly built upon the two assumptions of rational expectations with well-known probabilities of utilities and a representative agent ("*homo oeconomicus*"). They imply a complete understanding of the economic laws governing the world. These models leave no place for *imperfect knowledge* discovered in empirical psychological studies of real humans (Frydman and Goldberg [4, 25]). Their behavior in financial markets is even strongly influenced by emotional and hormonal reactions. Thus, economic modeling has to take *bounded rationality* seriously. But, model ambiguity does not mean the collapse of mathematical modeling. Mathematically, a fixed probability measure of expected utilities should be replaced by a convex risk measure which simultaneously considers a whole class of possible stochastic models with different penalties. Financial praxis warned us not to rely on a fixed model, but to vary possible models in a flexible way and to pay attention to the worst case. This is also the mathematical meaning of a convex risk measure.

The differences between the overall system and its parts, macro- and microeconomics, remain incomprehensible from the viewpoint of classical rationality which assumes a *representative agent*. Since interaction depends on differences in information, motives, knowledge and capabilities, this implies heterogeneity of agents (Hayek [29, 30]). Only a sufficiently rich structure of connections between firms, households and a dispersed banking sector will allow insights in *systemic risks* and synergetic effects in the financial sector. The reductionism of the representative agent or "homo oeconomicus" has prevented economists from modeling these phenomena.

For natural scientists, the distinction between micro-level phenomena and those originating on a macro originated from the interaction of microscopic units is well-known. In those models, the current crisis would be seen as an *emergent phenomenon* of the macroeconomic activity (Aoki and Yoshikawa [15], Mainzer [40]). The reductionist paradigm blocks any understanding of the interplay between micro and macro level.

Models with interacting heterogeneous agents would also open the door to interdisciplinary research from different sciences. Complex networks of different agents or statistical physics of interacting agents can model dynamic economic systems (Mantegna and Stanley [38], McCauley [45]). *Self-organized criticality* is another area that seems to explain boom-and-bust cycles of the economic non-equilibrium dynamics (Scheinkman [49]).

6 Food for Thought

In macroeconomics, data mining is often driven by the pre-analytic belief in the validity of certain models which should justify *political* or *ideological opinions*. The political belief in deregulation of the 1990 years is a typical example. Rather than misusing statistics as a means to illustrate these beliefs, the goal should be to put theoretical models to scientific tests like in the natural sciences. We should follow the line of a more data-driven methodology.

A chain of specification tests and estimated statistical models for simultaneous systems would provide a benchmark for the tests of models based on economic behavior. Significant and robust relations within a simultaneous system would provide empirical regularities that one would attempt to explain, while the quality of fit of the statistical benchmark would offer a confidence for more ambitious models. Models that do not reproduce (even) approximately the quality of the fit of statistical models would have to be rejected. This methodological criterion also has an aspect of *ethical responsibility* of researchers: economic policy models should be theoretically and empirically sound. Economists should avoid giving policy recommendations on the base of models with a weak empirical grounding and should, to the extent possible, make clear to the public how strong the support of the data is for their models and the conclusions drawn from them.

A neglected area of methodology is the degree of connectivity and its interplay with the *stability of the complex system*. It will be necessary for supervision to analyze the network aspects of the financial system, collect appropriate data, define measures of connectivity and perform macro stress testing at the system level. In this way, new measures of financial fragility would be obtained. This would also require a new area of accompanying academic research that looks at agent-based models of the financial system, performs scenario analyses and develops aggregate risk measures. Network theory and the theory of self-organized criticality of highly connected systems would be appropriate starting points (Scheinkman and Woodford [50], Mainzer [7]).

Such scientific analysis must be supported by more practical consequences. The hedge fund market is still widely unregulated. The interplay between *connectivity*, *leverage* and *system risks* needs to be investigated at the whole level. It is highly likely that extreme leverage levels of interconnected institutions impose dangerous social risks on the public.

On the macroeconomic level, it would be desirable to develop *early warning schemes* that indicate the formation of bubbles. Combinations of indicators with time series techniques could be helpful in detecting deviations of financial or other prices from their long-run averages. Indication of structural change would be a sign of changes of the behavior of market participants of a bubble-type nature (McCauley [45]).

Obviously, there is no single causal model as definitive mapping of reality. In this sense, David Hume and his followers were right in their skepticism against classical axiomatization of rationality in the world. But that does not mean a complete deny of mathematical tools and models. We have to consider whole *classes of possible* *stochastic models* with different weights. They must be combined with a *data-driven methodology* and insights in the factual human behavior and its diversity. Therefore, *psychological* and *sociological case studies* of human behavior under risk conditions (e.g., stakeholders at stock markets) are necessary. In experimental economics, decision behavior is already simulated under laboratorial conditions. Even *philosophical ethics* can no longer only argue with arm-chaired considerations and a priori principles, but must relate to empirical observations of factual decision behavior. That is done in the new approaches of experimental ethics. We argue for this kind of *interdisciplinary methodology* which opens new avenues for mathematical modeling in science. In this case, robust stochastic tools are useful, because they are used under restricted conditions and with sensibility for the permanent model ambiguity.

7 Summary

In a globalized world, risks are mainly *systemic* and cannot be reduced to single causes. They emerge from *complex interactions* in natural, technical, economic, and social systems. Examples are complex information and communication networks, power ("smart") grids as well as cellular interactions in organisms or transactions in financial markets. Therefore, *systems theory* with *linear* and *nonlinear dynamics*, *stochastic* and *statistic modeling*, and *computer models* are important methodologies in RISE. We must consider their explanatory power as well as their limitations. Then, they can supplement themselves mutually.

But, formal models are not sufficient. *Risk-awareness* even of experts is often *subjective* and depends on individual experience, societal and cultural contexts. Remember the extremely different reactions of the public to the Fukushima disaster in Japan and Germany. Therefore, formal risk-models must be complemented by sociological and cultural studies. Psychic behavior in decision situations must also be taken into account. Therefore, experimental economics and ethics relate to observations of factual behavior of people, e.g., at stock markets. Behavioral studies under experimental lab conditions are even useful for social philosophy and ethics.

The past crises might be characterized as example of final stages of well-known boom-and-bust patterns that have been repeated so many times in the course of economic history. But, there are several new aspects leading to a shift of methodological paradigm: the preceding boom had its origin in the development of new financial products with increasing complexity which seemed to promise diminishing risks. The financial market detaches itself from the real market. Profit seems to be possible by clever financial innovations loosing their connection to real economy. But, like in nature, there is no "*free lunch*" or "*perpetuum mobile*" of profit in finance. Further on, the past crises were due to the increasing complexity of interconnected financial networks. These aspects have been largely ignored by traditional economic models.

Therefore, we cannot trust in a single risk model, but must consider a *class of more or less appropriate models*, supplemented by *experimental behavioral case* *studies*. The lack of methodological understanding of models and the lack of ethical responsibility to warn the public against the limitations of models were the main reasons of the past economic crises. It is the task of *philosophy of science* to evaluate scientific modeling and the ethical responsibility of scientists. During booming periods we should better prepare the next crisis in a countercyclical manner.

References

Selected Bibliography

1. H.-J. Bungartz, S. Zimmer, M. Buchholz, D. Pflüger, *Modellbildung und Simulation. Eine anwendungsorientierte Einführung* (Springer, Berlin, 2009)
2. P. Embrechts, C. Klüppelberg, T. Mikosch, *Modeling Extremal Events for Insurance and Finance*, 4th edn. (Springer, Berlin, 2003)
3. H. Föllmer, A. Schied, *Stochastic Finance. An Introduction into Discrete Time*, 2nd edn. (De Gruyter, Berlin, 2004)
4. R. Frydman, M.D. Goldberg, *Imperfect Knowledge Economics* (Princeton University Press, Princeton, 2007)
5. N. Gershenfeld, *The Nature of Mathematical Modeling* (Cambridge University Press, Cambridge, 1998)
6. D. Kaplan, L. Glass, *Understanding Nonlinear Dynamics* (Springer, New York, 1995)
7. K. Mainzer, *Thinking in Complexity. The Computational Dynamics of Matter, Mind, and Mankind*, 5th edn. (Springer, Berlin, 2007)
8. K. Mainzer, *Der kreative Zufall. Wie das Neue in die Welt kommt* (C.H. Beck Verlag, München, 2007)
9. K. Mainzer, *Komplexität* (UTB-Profile, Paderborn, 2008)
10. B.B. Mandelbrot, R.L. Hudson, *The (mis) Behavior of Markets. A Fractal View of Risk, Ruin, and Reward* (Basic Books, New York, 2004)
11. K.R. Popper, *The Logic of Scientific Discovery* (Routledge, London, 1959)
12. N.N. Taleb, *The Black Swan—The Impact of the Highly Improbable* (Random House, New York, 2007)
13. W. Weidlich, *Sociodynamics. A Systematic Approach to Mathematical Modeling in the Social Sciences* (Taylor and Francis, London, 2002)
14. X.-S. Yang, *Mathematical Modeling for Earth Sciences* (Dudedin Academic, Edinburg, 2008)

Additional Literature

15. M. Aoki, H. Yoshikawa, *Reconstructing Macroeconomics—A Perspective from Statistical Physics and Combinatorical Stochastic Processes* (Cambridge University Press, Cambridge, 2007)
16. P. Artzner, F. Delbaen, J.-M. Eber, D. Heath, Coherent measures of risk. Math. Finance **9**(3), 203–228 (1999)
17. L. Bachelier, Théorie de la spéculation. Dissertation. Ann. Sci. Ec. Norm. Super. **17**, 21–86 (1900)
18. F. Black, M. Scholes, The pricing of options and corporate liabilities. J. Polit. Econ. **81**, 637–654 (1973)

19. D. Colander, H. Föllmer, A. Haas, M. Goldberg, K. Juselius, A. Kirman, T. Lux, B. Sloth, The financial crisis and the systemic failure of academic economics. Discussion Papers 09-03, Department of Economics, University of Copenhagen, 2008
20. R. Cont, Model uncertainty and its impact on the pricing of derivative instruments. Math. Finance **16**, 519–542 (2006)
21. J. Danielsson, P. Embrects, C. Goodhart, C. Keating, F. Muennich, O. Renault, H. Song Shin, An academic response to Basel II. Special paper series No. 130, LSE Financial Markets Group, ESRC Research Centre, May 2001
22. K. Detlefsen, G. Scandolo, Conditional and dynamic convex risk measures. Finance Stoch. **9**(4), 539–561 (2005)
23. E. Fehr, *Neuroeconomics. Decision Making and the Brain* (Academic Press, Waltham, 2008)
24. H. Föllmer, A. Schied, Convex and coherent risk measures. Working paper, Institute for Mathematics, Humboldt-University Berlin, October 2008
25. R. Frydman, M.D. Goldberg, Macroeconomic theory for a world of imperfect knowledge. Capital. Soc. **3**(3), 1–76 (2008)
26. R.M. Goodwin, *Chaotic Economic Dynamics* (Clarendon Press, Oxford, 1990)
27. A. Greenspan, We will never have a perfect model of risk. Financial Time (17 March 2008)
28. I. Hacking, *An Introduction to Induction and Probability* (Cambridge University Press, Cambridge, 2001)
29. F.A. Hayek, *Individualism and Economic Order* (The University of Chicago Press, Chicago, 1948)
30. F.A. Hayek, The pretence of knowledge, Nobel Lecture 1974, in *New Studies in Philosophy, Politics, Economics, and History of Ideas* (The University of Chicago Press, Chicago, 1978)
31. D. Hume, *An Enquiry Concerning Human Understanding*. Havard Classics, vol. 37 (Collier, New York, 1910)
32. International Monetary Fund, Global financial stability report: responding to the financial crisis and measuring systemic risk. Washington, April 2009
33. J.M. Kleeberg, C. Schlenger, Value-at-risk im asset management, in *Handbuch Risikomanagement*, ed. by L. Johannig, B. Rudolph (Uhlenbruch-Verlag, Bad Soden, 2000), pp. 973–1014
34. F. Knights, Risk, uncertainty, and profit. Ph.D., Yale, 1921
35. H.-W. Lorenz, *Nonlinear Dynamical Economics and Chaotic Motion* (Springer, Berlin, 1989)
36. A.J. Lotka, *Elements of Mathematical Biology* (Dover, New York, 1956). Reprint of the first publication 1924
37. T. Lux, F. Westerhoff, Economics crisis. Nat. Phys. **5**, 2–3 (2009)
38. F. Maccheroni, M. Marinaci, A. Rustichini, Ambiguity aversion, robustness, and the variational representation of preferences. Econometrica **74**, 1447–1498 (2006)
39. K. Mainzer (ed.), *Complexity*. European Review, vol. 17 (Cambridge University Press, Cambridge, 2009)
40. K. Mainzer, Challenges of complexity in economics. Evol. Inst. Econ. Rev., Jpn. Assoc. Evol. Econ. **6**(1), 1–22 (2009)
41. B.B. Mandelbrot, *Multifractals and 1/f Noise: Wild-Self-Affinity in Physics* (Springer, New York, 1999)
42. R.N. Mantegna, H.E. Stanley, *An Introduction to Econophysics. Correlations and Complexity in Finance* (Cambridge University Press, Cambridge, 2000)
43. H.H. Markowitz, *Portfolio Selection: Efficient Diversification of Investments* (Yale University Press, New Haven, 1959)
44. C. Marrison, *The Fundamentals of Risk Measurement* (McGraw Hill, New York, 2002)
45. J.L. McCauley, *Dynamics of Markets. Econophysics and Finance* (Cambridge University Press, Cambridge, 2004)
46. M. Polanyi, *Personal Knowledge. Towards a Post Critical Philosophy* (Routledge, London, 1998)
47. K. Popper, *The Poverty of Historicism* (Routledge, London, 1957)
48. F. Riedel, Dynamic convex risk measures: time consistency, robustness, and the variational representation of preferences. Econometrica **74**, 1447–1498 (2006)

49. J.A. Scheinkman, Nonlinearities in economic dynamics. Econ. J., Suppl. **100**(400), 33–47 (1990)
50. J.A. Scheinkman, M. Woodford, Self-organized criticality and economic fluctuations. Am. Econ. Rev. 417–421 (2001)
51. W.F. Sharpe, Capital asset prices: a theory of market equilibrium under conditions of risk. J. Finance **19**, 425–442 (1964)
52. H.W. Sinn, *Kasino-Kapitalismus* (Ullstein-Verlag, Berlin, 2010)
53. R.M. Stutz, Was Risikomanager falsch machen. Harvard Bus. Manag. **April**, 67–75 (2009)
54. L. Turner, The turner review. A regulatory response to the global banking crisis. The Financial Services Authority, 25 The North Colonnade, Canary Wharf, London E14 5HS, March 2009
55. M. York (ed.), *Aspects of Mathematical Finance* (Springer, Berlin, 2008)

Part II
Quantitative Risk Methodology

Chapter 5
The Mathematical Concept of Measuring Risk

Francesca Biagini, Thilo Meyer-Brandis, and Gregor Svindland

One of the key tasks in risk management is the quantification of risk implied by uncertain future scenarios which then has to be interpreted with respect to certain risk management decisions. Mathematically, the usual tool for doing so is a quantitative risk measure. The financial industry standard risk measure Value-at-Risk exhibits some serious deficiencies and a vital research activity has been ongoing to search for better alternatives. In this chapter we give an introduction to the general theory of monetary, convex, and coherent risk measures and present illustrating and motivating examples.

Keywords Risk measures · Acceptance sets · Robust representation · Risk sharing

Mathematics Subject Classification (2010) 91B30 · 91G99

The Facts

- Quantitative risk measures are key tools in financial risk management. The most prominent examples are the Value-at-Risk and the Average Value-at-Risk risk measures, see Sect. 2.
- Due to their importance in modern risk assessment, there is a vivid research activity, both in practice as well as in academia, on the topic of classifying suitable risk measures. This has amongst others led to the development of the theory of convex monetary risk measures.
- We present three basic approaches to defining such risk measures, one purely axiomatic, and two more constructive ones. The axiomatic approach classifies

F. Biagini
Chair of Financial Mathematics, Department of Mathematics, University of München,
Theresienstr 39, 80333 Munich, Germany

T. Meyer-Brandis · G. Svindland (✉)
Financial Mathematics, Department of Mathematics, University of München, Theresienstr 39,
80333 Munich, Germany
e-mail: svindla@mathematik.uni-muenchen.de

C. Klüppelberg et al. (eds.), *Risk – A Multidisciplinary Introduction*,
DOI 10.1007/978-3-319-04486-6_5,

the monetary risk measures as functions with certain properties, see Sect. 3.1. In the two other approaches convex monetary risk measures are constructed by specifying either sets of acceptable portfolios (Sect. 3.2) or by considering the worst case given a set of probability models (Sect. 3.3).

- The exposition is completed by illustrating examples throughout the text.

1 Introduction

Mathematically, the possibility of different, uncertain outcomes of future states of the world can be modeled by a function $X : \Omega \to \mathbb{R}$ where Ω denotes a fixed set of scenarios; i.e. each possible scenario $\omega \in \Omega$ is represented by a real number $X(\omega)$. For example, X could represent the uncertain (discounted) values of a portfolio of financial assets at a future point in time. If there exists an idea about how likely the realizations of the possible future scenarios are, then this is typically modeled by further assuming X to be a random variable; that is one considers a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ where the σ-algebra $\mathcal{F}$ denotes the set of all events and $\mathbb{P}$ is a probability measure that determines the probability of each event. Probability theory, which is a concise and axiomatic translation of our intuition about randomness into a mathematical theory, was initiated and developed in the ground breaking work of the Russian mathematician Kolmogorov in 1933 [16].

We note that in the above model approach there exists risk or uncertainty at two levels. When X is assumed to be a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, which is the usual approach in quantitative risk management, then uncertainty about the future realization of X is described by the assumed probabilistic structure implied by the probability model $\mathbb{P}$. This type of 'measurable' uncertainty is often referred to as *risk*. However, the choice of a specific probability model $\mathbb{P}$ is a disputed approach since in general there is not enough knowledge to make a reliable choice of one specific probability structure $\mathbb{P}$. This controversial discussion has been nurtured again by the recent financial crisis. There is thus a second level of (model) uncertainty which concerns the 'immeasurable' risk of not knowing the correct probability model and which is referred to as *Knightian uncertainty*. The distinction between measurable risk as opposed to immeasurable uncertainty was established first by Frank Knight in his work 'Risk, Uncertainty, and Profit' [15]. It is an important challenge to extend risk management approaches by mathematical tools that deal with Knightian uncertainty, and, as we will see, the theory of (convex) risk measures is contributing to this research objective.

A *risk measure* helps the risk manager to measure and quantify the risk implied by the uncertainty about the future realization of X. Such quantitative risk measures are usually obtained by applying a certain functional ρ to X which yields a real number $\rho(X)$ that indicates the risk level which then has to be interpreted in terms of risk management decisions. The definition and theory of quantitative risk measures has been initiated and closely influenced by the need for quantitative risk management in the financial and insurance industry. In these areas, the outcome $\rho(X)$ of a risk measure may be interpreted as required capital reserve to hedge against the risk of future losses, as management tool for limiting the amount of risk a unit within an institution may take (for example to constrain a single trader's portfolio), or as insurance premium required to compensate the insurance company for bearing the risk of the insured claims.

Historically, the first one to systematically consider the return of an investment portfolio in relation to its risk was Markowitz in 1952 [18]. He modeled the value of a portfolio by a random variable X and used the *standard deviation* (or *variance*) of the distribution as risk measure. By determining the so-called efficient frontier the portfolio manager could then optimize the return for a given risk level. In 1973, Black, Scholes, and Merton developed the famous *Black-Scholes-Merton formula* to determine the price of a European call option which had enormous impact on the development of financial derivatives markets [8, 19]. This price can be interpreted as a risk measure to hedge against the risk of selling such an option, where the risk measure is the expectation under the so called risk-neutral probability measure. In the early 1990s, the financial industry, public sector, and academia alike started to recognize the need for systematic risk management of the enormous increase in so-called off-balance-sheet products like derivatives. At the investment bank JP Morgan, for instance, the introduction of the so-called Weatherstone 4.15 report asked for the daily assessment of the firms market risk measured in terms of Value-at-Risk (VaR). VaR, which quickly has been established as the industry standard and most prominent risk measure, is a certain quantile of the loss distribution of a portfolio (see Sect. 2.1 for more details). In response to a sequence of disasters in particular over the last two decades, like the ruin of the Barings Bank caused by the single trader Nick Leason in 1995, the fall of the hedge fund Long-Term Capital Management in 1998, or the most serious recent financial crisis that started in 2007, a road to systematic regulation of the banking and insurance industry has evolved. The currently applicable regulation guidelines, which in Germany are implemented by the BAFIN (Bundesanstalt für Finanzdienstleistungsaufsicht), are formulated in the so-called Basel II Accord for the banking sector and Solvency II Accord for the insurance sector. In these guidelines one of the main regulation principles is to require sufficient capital reserves of a firm as to hedge against the risk exposure of future losses using the risk measure VaR.

Despite the status as the industry standard, VaR is often criticized mainly by academics for some fundamental deficiencies. In particular, in certain scenarios VaR is punishing the pooling (or diversification) of risk and is encouraging the accumulation of shortfall risk, which is the opposite of what our intuition about good properties of risk measures would be. Also, while VaR considers the probability that a loss occurs it is not concerned about the size of possible losses. This criticism about VaR has initiated a vital research activity aiming at specifying desirable axioms for risk measures. The outcome of this research has been the axiomatic definitions of the families of *monetary*, *convex*, and *coherent* risk measures; see Artzner et al. [1], Föllmer and Schied [2], and Frittelli and Rosazza-Gianin [5]. Further, the characterization of convex risk measures presented in Sect. 3.3 will reveal that the concept of convex risk measures takes Knightian uncertainty into account. Despite the academic advances and warnings concerning VaR, more appropriate risk measures like Average Value-at-Risk (AVaR), also referred to as Expected Shortfall ($\mathbb{ES}$) (see Sect. 2.2 for more details), have not yet been incorporated into the regulation guidelines. However, the experiences gained during the financial crisis, which revealed the deficiencies in the use of VaR as proposed in the regulation guidelines of Basel II, have initiated a further reformation of regulation mechanisms that takes place under the notion Basel III.

The objective of this chapter is to present the general mathematical concept of monetary risk measures. As described above, the theory of monetary risk measures has been developed in the environment of financial and insurance markets. We will remain in that framework and assume in the following that X models the (discounted) value of a portfolio and a risk measure $\rho(X)$ measures the risk in terms of capital. However, the fundamental concept of (financial) risk management has potential to also be applied to other fields where it is necessary to quantify risk exposure in a concise way. We start by describing in more detail the two most prominent monetary risk measures VaR and AVaR in Sect. 2. In Sect. 3 we then provide the general axiomatic definitions of monetary, convex, and coherent risk measures as well as presenting two alternative methods of constructing convex monetary risk measure. The theory is illustrated by several concrete examples. In Sect. 4 we discuss the optimal risk sharing problem as an example of a typical research question in this field before we conclude in Sect. 5 with some food for thoughts.

2 Two Prominent Examples: VaR and AVaR

2.1 Value-at-Risk

The most common quantitative risk measure in use is the so-called *Value-at-Risk* (VaR_α) at level $\alpha \in (0, 1)$. We model the risk of a discounted portfolio at a future point in time by a random variable X on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and denote by $F(x) := \mathbb{P}(X \leq x)$, $x \in \mathbb{R}$, the distribution function of the risk X. Note that the risk manager is concerned about the *downside risk* of X (i.e. small values of X which imply losses). The risk measure VaR_α measures the minimal amount of cash $m \in \mathbb{R}$ that has to be added to the portfolio X such that the probability of a loss of $X + m$ is less than α. In other words,

$$\mathrm{VaR}_\alpha(X) = \inf\big\{m \in \mathbb{R} \mid \mathbb{P}(X + m < 0) \leq \alpha\big\}. \tag{2.1}$$

Depending on the risk management situation, typical values for α are 0.05 (5 %), 0.01 (1 %), or 0.001 (0.1 %). VaR_α can thus be interpreted as the required capital reserve such that the probability of losses at a given future point in time is less than α.

Example 2.1 As mentioned in the introduction, the risk measure stipulated by the Basel II regulations to compute a bank's required capital reserve is VaR_α. More precisely, in Basel II the probability distribution of X is assumed to be normal, that is

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(y-\mu)^2}{\sigma^2}} dy \quad \text{for } x \in \mathbb{R}. \tag{2.2}$$

The mean parameter μ and variance parameter σ^2 have to be estimated from historical data. The computation of VaR_α for a general normal distribution with mean μ and variance σ^2 can be reduced to the computation of the inverse of the standard normal distribution function:

$$\mathrm{VaR}_\alpha(X) = -\mu - \sigma\Phi^{-1}(\alpha),$$

where Φ denotes the standard normal distribution function (i.e. F as in (2.2) with $\mu = 0$ and $\sigma^2 = 1$). Indeed, since the normal distribution function in (2.2) is continuous and strictly increasing it is sufficient by the definition of VaR_α in (2.1) to observe that

$$\begin{aligned} \mathbb{P}\big(X + \big(-\mu - \sigma\Phi^{-1}(\alpha)\big) < 0\big) &= \mathbb{P}\big(X \le \mu + \sigma\Phi^{-1}(\alpha)\big) \\ &= \mathbb{P}\left(\frac{X-\mu}{\sigma} \le \Phi^{-1}(\alpha)\right) \\ &= \Phi\big(\Phi^{-1}(\alpha)\big) = \alpha, \end{aligned}$$

where the last equality comes from the fact that $\frac{X-\mu}{\sigma}$ is standard normal distributed. VaR_α for normally distributed random variables is thus easily computed. However, the assumption that the risk X is normally distributed is very critical since empirical data analysis indicates that the normal distribution strongly underestimates the probability of extreme events in most situations (see also [10], Chap. 6).

Despite its popularity, the use of VaR_α exhibits some serious deficiencies. In particular, it has been fundamentally criticized because of the following two major problems:

(i) VaR_α only considers the probability of encountering losses but not the size of the potential loss in case a loss scenario occurs. Hence, optimizing portfolios under constraints on the risk given by VaR_α may result in portfolios which indeed have a low probability of loss, i.e. less than the specified level α, but which may produce (extremely) high losses if a loss scenario occurs. These effects have been observed, for instance, during the last financial crisis, and it seems evident that in such situations a sound measuring of risk should not only take into account the likeliness of a loss scenarios, but also depend on the quantity that might be lost.

(ii) As we will see in Example 3.3, VaR_α is a positively homogeneous, monetary but not a convex risk measure (see definitions in Sect. 3.1). In particular, one can construct realistic examples of two risks X and Y such that:

$$\mathrm{VaR}_\alpha\left(\frac{1}{2}(X+Y)\right) > \frac{1}{2}\mathrm{VaR}_\alpha(X) + \frac{1}{2}\mathrm{VaR}_\alpha(Y).$$

So $\mathrm{VaR}_\alpha(X+Y)$ of a merged portfolio is not necessarily bounded above by the sum of the VaR_α's of the individual portfolios. But this means that measuring risk with VaR_α may penalize diversification instead of encouraging it. Further, decentralization of risk management might be difficult using VaR_α because one cannot be sure that by aggregating the VaR_α levels of different portfolios (or different units) one will obtain a bound for the overall risk.

2.2 Average Value-at-Risk

The fundamental criticism about VaR_α has led to the search for better alternatives. Before we present the general axiomatic approach which has been the outcome of these efforts in the next section, we will introduce another popular risk measure which attempts to solve problems (i) and (ii) described above. Instead of just considering the quantile corresponding to some specified level $\alpha \in (0,1)$, one averages over all quantiles less than the given level α. One thus takes into account not only the likeliness of a loss scenario but also the quantity that might be lost, which addresses problem (i) above. The corresponding risk measure is called the *Average Value-at-Risk* (AVaR_α) and is given by

$$\mathrm{AVaR}_\alpha(X) = \frac{1}{\alpha}\int_0^\alpha \mathrm{VaR}_\lambda(X)d\lambda, \tag{2.3}$$

where we assume that the expectation is finite, i.e. $\mathbb{E}(|X|) < \infty$, in order to have the right hand side of (2.3) well-defined. Obviously, like VaR_α, also AVaR_α only depends on the probability distribution of X and we have $\mathrm{AVaR}_\alpha \geq \mathrm{VaR}_\alpha$. In Fig. 1 $\mathrm{VaR}_\alpha(X)$ and $\mathrm{AVaR}_\alpha(X)$ of a normal distributed random variable X are plotted against the respective normal density.

But contrary to VaR_α, we will see in Example 3.4 that AVaR_α is convex and even *coherent* (see the definition in Sect. 3.1), which addresses problem (ii) above. One can actually show that AVaR_α is the best approximation of VaR_α in the class of convex risk measures which only depend on the distribution of the portfolio (see [4]). Sometimes, AVaR_α is also referred to as Expected Shortfall ES_α. This is motivated by the following alternative representation which is valid when X has a continuous distribution function:

$$\mathrm{AVaR}_\alpha(X) = \mathbb{E}\big[-X \mid -X \geq \mathrm{VaR}_\alpha(X)\big]. \tag{2.4}$$

For general distribution functions the equality in (2.4) turns into a greater-or-equal inequality.

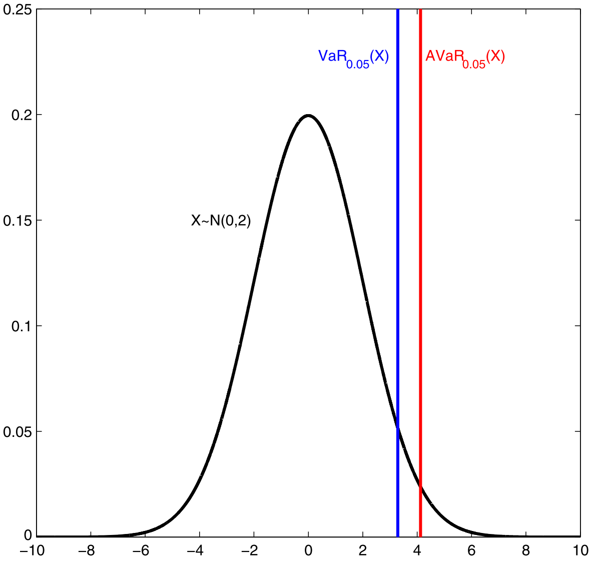


Fig. 1 $\mathrm{VaR}_\alpha(X)$ vs. $\mathrm{AVaR}_\alpha(X)$ at level $\alpha = 5$ % of a normally distributed r.v. X. Indeed $\mathrm{AVaR}_\alpha(X) > \mathrm{VaR}_\alpha(X)$

Example 2.2 Consider again the example of a normally distributed X with mean μ and variance σ^2. Then the distribution function is continuous and we can use representation (2.4) to compute AVaR_α for a given level $\alpha \in (0, 1)$ as

$$\mathrm{AVaR}_\alpha(X) = -\mu + \sigma \frac{\phi(\Phi^{-1}(\alpha))}{\alpha},$$

where ϕ is the density and Φ is the distribution function of a standard normal distribution. Indeed, using representation (2.4) we observe that

$$\mathrm{AVaR}_\alpha(X) = -\mu + \sigma \mathbb{E}\left[-\frac{X-\mu}{\sigma} \,\middle|\, -\frac{X-\mu}{\sigma} \geq \mathrm{VaR}_\alpha\left(\frac{X-\mu}{\sigma}\right)\right],$$

which reduces the problem to the computation of AVaR_α for the standard normal random variable $\frac{X-\mu}{\sigma}$. Again by (2.4) we get

$$\begin{aligned}\mathrm{AVaR}_\alpha\left(\frac{X-\mu}{\sigma}\right) &= -\frac{1}{\alpha}\int_{-\infty}^{\Phi^{-1}(\alpha)} x\phi(x)dx \\ &= \frac{1}{\alpha}\big[\phi(x)\big]_{-\infty}^{\Phi^{-1}(\alpha)} = \frac{\phi(\Phi^{-1}(\alpha))}{\alpha}.\end{aligned}$$

3 Monetary, Convex, and Coherent Risk Measures

Motivated by the examples above, we now introduce the general mathematical theory and characterization of monetary risk measures. More precisely, we give three alternative approaches to monetary risk measures, and we will observe that they are basically equivalent. Namely, in Sect. 3.1 we will undertake an axiomatic approach, in Sect. 3.2 we will construct monetary risk measures by means of a set of acceptable portfolios, whereas in Sect. 3.3 the construction incorporates ideas on how to deal with the uncertainty about the right probabilistic model—the Knightian uncertainty—mentioned in the introduction. The axiomatic approach to (coherent) monetary risk measures goes back to [1] and was further extended by [2, 5, 12]. For an exhaustive treatment of the subject, and in particular for the mathematical details, we refer to [4], for a survey to [3] and [17], Chap. 4.

3.1 What Properties Should a Risk Measure Have?

Consider a set of portfolios $\mathcal{X}$ which we assume to be of sufficiently nice mathematical structure in order to allow for the mathematical analysis which follows; see [4] for the details. We assume that the portfolios are discounted which allows us to compare values in the future with cash amounts today. In the following we are concerned with depicting basic properties that any monetary risk measure ρ should satisfy in order to be suited to measure the risk in terms of cash amounts needed to secure a given portfolio $X \in \mathcal{X}$. One undoubted feature of any risk evaluation is that more is better than less, that is if the payoff of a portfolio X is higher than the payoff of another portfolio Y, then the measured risk $\rho(X)$ of X should be lower than the measured risk $\rho(Y)$ of Y. This property is referred to as monotonicity of the risk measure ρ. Another property of monetary risk measures is based on the observation that cash amounts have no intrinsic risk in the sense that facing a sure loss m, we know that we have to have a corresponding security of $-m$ in order to be able to meet future payments. Therefore, it seems natural to assess the risk of a certain amount m as $-m$ or, more generally, if we add a certain amount m to a given portfolio X then the risk $\rho(X+m)$ of $X+m$ as compared to the risk $\rho(X)$ of X should be increased or decreased by m—depending on whether m is a loss or a gain—and thus corresponds to $\rho(X)-m$. Mathematically these basic features of a monetary risk measure are expressed in the following way:

Definition 3.1 $\rho : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ is called a *monetary risk measure* if $\rho(0) < \infty$ and ρ satisfies the following conditions for all $X, Y \in \mathcal{X}$:

- *Monotonicity*: If $X \leq Y$, then $\rho(X) \geq \rho(Y)$;
- *Cash invariance*: If $m \in \mathbb{R}$, then $\rho(X+m) = \rho(X) - m$.

Since $\rho(X + \rho(X)) = 0$, we may interpret $\rho(X)$ as a capital requirement, i.e. as the minimal amount of capital that must be added to or can be withdrawn from X in order to obtain zero risk and thus make X acceptable from the point of view of a supervising agency.

Example 3.2 Let X be a square integrable random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The *standard deviation* (or *volatility*) σ_X of X

$$\sigma_X := \sqrt{\mathbb{E}\big[\big(X - \mathbb{E}[X]\big)^2\big]},$$

which first was systematically introduced by Markowitz in the risk analysis [18] of portfolio choices and still is a widely used risk indicator, is neither cash invariant nor monotone and thus not a monetary risk measure. The lack of monotonicity is mainly due to the fact that the standard deviation considers risk symmetric in the sense that the risk of gains is assessed in the same way as the risk of losses, whereas a risk manager is usually only concerned with the risk of losses. For example, let $X > 0$ be a random variable with positive support. Then $Y := aX > X$ for any constant $a > 1$, and also $\sigma_Y = a\sigma_X > \sigma_X$. Thus monotonicity is not fulfilled. Augmenting σ_X to

$$\mathrm{mv}(X) := \mathbb{E}[-X] + \sigma_X,$$

we obtain the so-called *mean-variance risk measure* which is cash invariant, but still does not satisfy monotonicity. The lack of monotonicity is the reason why the very popular asset pricing based on optimizing a portfolio under σ_X or mv has drawbacks such as producing negative prices in some cases.

It is often required that the risk measure ρ should favor diversification: Consider two portfolios X, Y and the possibilities to either invest in X or in Y or in a fraction $\lambda X + (1-\lambda)Y$, $\lambda \in [0, 1]$, of both. Favoring diversification means that the risk of the diversified investment $\lambda X + (1-\lambda)Y$ should not exceed the risks of both X and Y, thereby accounting for the fact that the downside risk, in particular the risk of default, is lower in the diversified investment $\lambda X + (1-\lambda)Y$ as compared to the most risky of X and Y. Formally this property is know as quasi-convexity of the risk measure ρ:

- *Quasi Convexity*: $\rho(\lambda X + (1-\lambda)Y) \le \max(\rho(X), \rho(Y))$, for $0 \le \lambda \le 1$.

If the risk measure ρ satisfies cash invariance then it can indeed be shown that quasi-convexity is equivalent to convexity, i.e.

- *Convexity*: $\rho(\lambda X + (1-\lambda)Y) \le \lambda\rho(X) + (1-\lambda)\rho(Y)$, for $0 \le \lambda \le 1$.

The latter property is very desirable from an analytic point of view since it allows for an analysis of convex monetary risk measures by means of tools from the field of convex analysis and optimization. This field provides comprehensive toolboxes for dealing with optimization problems that naturally occur in the financial risk context, like e.g. portfolio optimization under constraints on the portfolio risk given by some convex monetary risk measure.

Recall the Value-at-Risk discussed in Sect. 2.1. It follows immediately that VaR_α is a monetary risk measure in the above sense. However, as we will show in the following example and as was already mentioned in Sect. 2.1, VaR_α is not convex (and thus not quasi convex either).

Example 3.3 Consider two portfolios X, Y which are independent and identically distributed. Let X and Y take the value 100 with probability 0.99 and -100 with probability 0.01. Then the convex combination $\frac{1}{2}(X + Y)$ assumes the values 100, 0, and -100 with probabilities 0.9801, 0.0198, and 0.0001 respectively. Hence, $\mathrm{VaR}_{0.01}(X) = \mathrm{VaR}_{0.01}(Y) = -100$, whereas $\mathrm{VaR}_{0.01}(\frac{1}{2}(X + Y)) = 0$.

As mentioned in Sect. 2.2 the Average Value-at-Risk indeed satisfies convexity and thus favors diversification. Moreover, it inherits a scaling invariance property from the Value-at-Risk which is known as positive homogeneity:

- *Positive Homogeneity*: $\rho(\lambda X) = \lambda\rho(X)$, for $\lambda \geq 0$.

It might be debated whether this latter property is reasonable in every setting as it implies a linear dependence of risk with respect to the amount invested into a portfolio. Especially for large multipliers $\lambda > 0$ this might not be very realistic, and one should instead have $\rho(\lambda X) > \lambda\rho(X)$, and thus pure convexity, to penalize a concentration of risk. The idea here is that if the maximal loss of a portfolio is for instance 1 Euro, then this might not be seen as very risky, whereas the risk of a million times the very same portfolio, and thus a possible loss of a million Euros, may be viewed as providing much more risk than simply a million times the very low risk of the initial portfolio.

Nevertheless, many risk measures exhibit the positive homogeneity property and the positively homogeneous convex monetary risk measures form an important subclass called *coherent risk measures*.

Example 3.4 The Average Value-at-Risk AVaR_α as presented in Sect. 2.2 is a coherent risk measure. Indeed, cash invariance, monotonicity, and positive homogeneity follow immediately from the properties of VaR_α. The crucial property of convexity can easily be deduced from the following representation of AVaR_α:

$$\mathrm{AVaR}_\alpha(X) = \lim_{n\to\infty} \frac{\sum_{i=1}^{[n\alpha]}(-X_{i,n})}{[n\alpha]},$$

where $[n\alpha]$ is the integer part of $n\alpha$, $X_1, \ldots, X_n$ is a sequence of independent random variables which have the same distribution as X, and $X_{1,n} \geq \cdots \geq X_{n,n}$ is the order statistics of $(X_1, \ldots, X_n)$.

3.2 Constructing Risk Measures via Acceptance Sets

As an alternative to the axiomatic approach, one could define a monetary risk measure on $\mathcal{X}$ by fixing a class of portfolios $\mathcal{A} \subset \mathcal{X}$ which are acceptable in the sense that they do not require additional capital in order to secure their risk. Now the risk of any portfolio $X \in \mathcal{X}$ is evaluated as the minimal amount of cash m that has to be added to X such that $X+m$ is acceptable, that means $X+m \in \mathcal{A}$. The risk measure $\rho_{\mathcal{A}}$ induced by $\mathcal{A}$ in the described way has a formal representation as follows

$$\rho_{\mathcal{A}}(X) = \inf\{m \in \mathbb{R} \mid m + X \in \mathcal{A}\}. \tag{3.1}$$

Assuming that $\mathcal{A}$ satisfies certain properties such as convexity and a monotonicity property ($X \in \mathcal{A}_\rho$, $Y \in \mathcal{X}$, $Y \geq X$, then $Y \in \mathcal{A}_\rho$), one can prove that $\rho_{\mathcal{A}}$ is a convex monetary risk measure as defined in Sect. 3.1; see [4], Proposition 4.7.

Furthermore, one can show that the converse is also true: every convex risk measure in the sense of Sect. 3.1 is induced by an acceptance set $\mathcal{A}$ as in (3.1), and thus the two approaches to defining a risk measure are equivalent. To see this, consider any monetary risk measure ρ and let

$$\mathcal{A}_\rho = \big\{X \in \mathcal{X} \mid \rho(X) \leq 0\big\} \tag{3.2}$$

be the set of portfolios that are acceptable under ρ, the so called acceptance set of ρ. Then, $\rho = \rho_{\mathcal{A}_\rho}$.

In the following we provide some prominent examples of this approach.

Example 3.5 (Monetary Risk Measure Induced by Expected Utility) In economic theory the preferences of some agent are often modeled by a utility function that is a strictly concave and strictly increasing function $u : \mathbb{R} \to \mathbb{R}$, which quantifies how utile the agent considers payoffs compared to each other. The property of being increasing encodes the fact that more is better. The concavity of u is due to the fact that a typical agent is very sensitive to losses, fairly sensitive to gains, but not that sensitive to very large gains as compared to slightly smaller gains, simply because above a certain level she has reached an amount of wealth above which her consumption cannot be significantly improved. Such agents are usually assumed to assess the utility of a portfolio $X \in \mathcal{X}$ by taking the expected utility value of X, i.e. $\mathbb{E}[u(X)]$. Hence, a natural way to obtain a reasonable acceptance set is to call acceptable the set of portfolios $X \in \mathcal{X}$ such that the expected utility exceeds a certain threshold c, that is $X \in \mathcal{A}$ if and only if

$$\mathbb{E}\big[u(X)\big] \geq c.$$

The corresponding acceptance set defines a convex monetary risk measure via (3.1).

Example 3.6 (Shortfall Risk) Recall Example 3.5 If the focus is more on the losses, instead of considering a utility function u and a lower bound c on the expected utility as in the previous example, it is more natural to replace u by a loss function $l : \mathbb{R} \to \mathbb{R}$ which is assumed to be convex and increasing and non-constant. Then the corresponding acceptance set is given by

$$\mathcal{A} := \big\{X \in \mathcal{X} \mid \mathbb{E}\big[l(-X)\big] \leq \tilde{c}\big\} \tag{3.3}$$

where $\tilde{c} \in \mathbb{R}$ is an upper bound on the *shortfall risk* $\mathbb{E}[l(-X)]$ of a portfolio X. This acceptance set defines a convex monetary risk measure $\rho_{\mathcal{A}}$ as in (3.1). If

$l(x) = -u(-x)$, then the acceptance set $\mathcal{A}$ equals the acceptance set in Example 3.5, and the associated risk measures coincide. Note, however, that the loss function l is only required to be increasing, so it may indeed be flat on some interval $(-\infty, a]$. In particular l may vanish on $(-\infty, 0]$, which means that the corresponding acceptability criterion in (3.3) only depends on the possible losses of a portfolio X.

3.3 A Robust Approach to Measuring Risk

Consider any portfolio $X \in \mathcal{X}$. A very natural way to assess its risk is looking at the expected value $\mathbb{E}_{\mathbb{P}}[X]$ of X under some probability measure $\mathbb{P}$. It can be easily verified that $\mathbb{E}_{\mathbb{P}}[-X]$ is indeed a coherent risk measure as introduced in Sect. 3.1. But computing the expected value means that we need full information about the probability distribution of X, that is we need to know the right probability measure $\mathbb{P}$ describing the real world. However, in many cases we don't know which probability model is appropriate to describe the true distribution of X. In other words, there is ambiguity on the right probability measure under which we should take the expectation. We may attempt to overcome this problem by specifying not just one but a class $\mathcal{Q}$ of probability measures that one considers as possible descriptions of reality and then to taking the expectation under each probability measure and to looking at the worst case. The corresponding risk measure is a coherent risk measure and is given by the following expression:

$$\rho(X) = \sup_{\mathbb{Q} \in \mathcal{Q}} \mathbb{E}_{\mathbb{Q}}[-X], \quad X \in \mathcal{X}. \tag{3.4}$$

One can even go one step further and penalize each probability measure $\mathbb{Q} \in \mathcal{Q}$ according to how likely the corresponding probability model appears to be. This is achieved by introducing a penalizing function $\alpha : \mathcal{Q} \to \mathbb{R}$ which assigns to each $\mathbb{Q} \in \mathcal{Q}$ a certain penalization $\alpha(\mathbb{Q})$. The corresponding risk measure is

$$\rho(X) = \sup_{\mathbb{Q} \in \mathcal{Q}} \big(\mathbb{E}_{\mathbb{Q}}[-X] - \alpha(\mathbb{Q})\big), \quad X \in \mathcal{X}, \tag{3.5}$$

which is a convex monetary risk measure. Clearly, letting $\alpha(\mathbb{Q}) = 0$ for all $\mathbb{Q} \in \mathcal{Q}$ we obtain (3.4).

Conversely, one can prove by means of tools from convex analysis that basically every convex monetary risk measure can be represented as in (3.5) where the penalizing function α is known as the dual function of the risk measure ρ. Hence, we observe that the three approaches to defining a risk measure presented throughout Sects. 3.1, 3.2, and 3.3 in principle are equivalent and lead to the same class of risk measures, these are the convex monetary risk measures. Moreover, we note that representation (3.5) reveals a robust structure with respect to model uncertainty and thus the capability of monetary convex risk measures to deal with Knightian uncertainty as discussed in the introduction. For further details, we refer to [4, 9, 14].

Example 3.7 (Entropic Risk Measure) Consider some reference probability measure $\mathbb{P}$ which we believe is the best description of the likeliness of any future events. In presence of ambiguity about the true probability measure, we decide to take into account all probability measures $\mathbb{Q}$ which are consistent with $\mathbb{P}$ in the sense that there are no events that are likely under $\mathbb{Q}$ but have zero probability under our reference probability measure $\mathbb{P}$. We say that such a probability measure $\mathbb{Q}$ is *absolutely continuous* with respect to $\mathbb{P}$, written as $\mathbb{Q} \ll \mathbb{P}$. If $\mathbb{Q}$ strongly deviates from $\mathbb{P}$, it should not play the same role in our risk analysis as probability measures which are just slight modifications of $\mathbb{P}$. This is realized by penalizing each $\mathbb{Q}$ by a kind of distance to $\mathbb{P}$ which is known as the (relative) entropy $H(\mathbb{Q} \mid \mathbb{P})$ of $\mathbb{Q}$ with respect to $\mathbb{P}$. The formal definition is

$$H(\mathbb{Q} \mid \mathbb{P}) := \mathbb{E}_{\mathbb{Q}}\left[\log \frac{d\mathbb{Q}}{d\mathbb{P}}\right]$$

where $\frac{d\mathbb{Q}}{d\mathbb{P}}$ is the density of $\mathbb{Q}$ with respect to $\mathbb{P}$. Indeed we have that $H(\mathbb{Q} \mid \mathbb{P}) \geq 0$ and that $H(\mathbb{Q} \mid \mathbb{P}) = 0$ if and only if $\mathbb{Q} = \mathbb{P}$. Taking the worst case over all probability models penalized with the relative entropy as in (3.5) yields the following convex monetary risk measure

$$e_\beta(X) = \sup_{\mathbb{Q} \ll \mathbb{P}} \left(\mathbb{E}_{\mathbb{Q}}[-X] - \frac{1}{\beta} H(\mathbb{Q} \mid \mathbb{P}) \right), \quad X \in \mathcal{X}, \tag{3.6}$$

where we allow for a parameter $\beta > 0$ determining the impact of the weighting. Solving the variational problem appearing on the right hand side of (3.6) we obtain that

$$e_\beta(X) = \frac{1}{\beta} \log \mathbb{E}\big[e^{-\beta X}\big], \quad X \in \mathcal{X}. \tag{3.7}$$

This is the so called *entropic risk measure*.

Example 3.8 (Acceptability Floor) Consider a set $\mathcal{Q}$ of probability measures, and let $\gamma : \mathcal{Q} \to \mathbb{R}$ be such that $\sup_{\mathbb{Q} \in \mathcal{Q}} \gamma(\mathbb{Q}) < \infty$. The function γ specifies an acceptability floor in the sense that a portfolio X is considered to be acceptable if and only if

$$\mathbb{E}_{\mathbb{Q}}[X] \geq \gamma(\mathbb{Q}) \quad \text{for all } \mathbb{Q} \in \mathcal{Q}.$$

The corresponding acceptance set $\mathcal{A}$ defines a convex monetary risk measure $\rho_{\mathcal{A}}$ as in (3.1) which also has the following representation

$$\rho_{\mathcal{A}}(X) = \sup_{\mathbb{Q} \in \mathcal{Q}} \big(\mathbb{E}_{\mathbb{Q}}[-X] + \gamma(\mathbb{Q}) \big), \quad X \in \mathcal{X}.$$

4 A Case Study: Optimal Risk Sharing

So far we have been concerned with specifying a class of risk measures—the (convex) monetary risk measures—which satisfy certain conditions that make them apt for risk management of financial portfolios. However, the choice of an appropriate risk measure within the vast class of convex monetary risk measures, where being *appropriate* depends on factors such as the business structure or stability under optimization, is a non-trivial problem. And even after solving that problem, this is by far not the end of the story. There are a lot of issues arising beyond the level of specifying an appropriate risk measure and simply applying it to quantify the risk of some portfolios. In what follows we present a typical problem arising in risk management when more than one agents are involved. In that case it is very natural to look for cooperation opportunities from which all agents benefit in the sense that the individual risk of each agent is reduced by mutual protection. In other words the agents seek an optimal risk sharing:

Consider n agents with initial portfolios W_i, $i = 1, \dots, n$, who assess the risk of any portfolio offered to them by individual convex monetary risk measures $\rho_i : \mathcal{X} \to \mathbb{R}$, $i = 1, \dots, n$. The aggregate portfolio is $W = W_1 + \dots + W_n$. An (re-)allocation of W is any $(X_1, \dots, X_n) \in \mathcal{X}^n$ such that $\sum_{i=1}^n X_i = W$. Denote by $\mathbb{A}(W)$ the set of all reallocations of W. We assume that the agents are allowed to exchange risks without changing the aggregate portfolio, that is the agents may agree on exchanging the initial allocation $(W_1, \dots, W_n)$ for some other allocation $(X_1, \dots, X_n)$ of W. The optimal risk sharing problem is to find a reallocation of W amongst the n agents such that the total risk is minimized, that is find $(\bar{X}_1, \dots, \bar{X}_n) \in \mathbb{A}(W)$ such that

$$\sum_{i=1}^{n} \rho_i(\bar{X}_i) = \inf\left\{ \sum_{i=1}^{n} \rho_i(Y_i) \,\middle|\, (Y_1, \dots, Y_n) \in \mathbb{A}(W) \right\}. \tag{4.1}$$

Note that, due to cash-invariance, if $(\bar{X}_1, \dots, \bar{X}_n)$ is a solution to (4.1), then we have that for every $(c_1, \dots, c_n) \in \mathbb{R}^n$ such that $\sum_{i=1}^n c_i = 0$ the allocation $(\bar{X}_1 + c_1, \dots, \bar{X}_n + c_n)$ is a solution to (4.1) too. Hence, provided that (4.1) allows for a solution, it is always possible to find a solution which respects the individual rationality constraints of the agents, that is a solution $(\bar{X}_1, \dots, \bar{X}_n)$ to (4.1) such that all agents are better off: $\rho_i(\bar{X}_i) \le \rho_i(W_i)$, $i = 1, \dots, n$.

Consider the following function

$$\Box \rho_i(W) := \inf\left\{ \sum_{i=1}^{n} \rho_i(Y_i) \,\middle|\, (Y_1, \dots, Y_n) \in \mathbb{A}(W) \right\}, \quad W \in \mathcal{X}. \tag{4.2}$$

Problem (4.1) is equivalent to finding $(\bar{X}_1, \dots, \bar{X}_n) \in \mathbb{A}(W)$ such that

$$\rho_1(\bar{X}_1) + \dots + \rho_n(\bar{X}_n) = \Box \rho_i(W). \tag{4.3}$$

Provided that $\Box_i \rho_i > -\infty$ it can be shown that $\Box_i \rho_i$ is again a convex monetary risk measure which is interpreted as the risk measure of the market or the representative agent. This has a particularly nice interpretation in the case of a company, e.g. an insurer, with different business units/entities possibly being exposed to different kinds of risks. Suppose that the risk of each unit is measured by a convex monetary risk measure, which depends on the business structure of that particular unit. Thus we may view each unit as an agent in the above sense. Then the stand alone risk of unit i with business W_i is $\rho_i(W_i)$. In this situation there are two major questions to be answered. First of all, given the structure of risk measurement for the units, what is a sound monetary risk measure for the whole company as such? Secondly, keeping in mind the typical situation that the measured risks correspond to e.g. solvency capital requirements, what is the advantage of each unit of being member of a group in the sense of being part of the company? One should expect that the risk profile of the unit should profit from the fact that the company may to some extent cover potential losses in that unit with gains from another. If this is the case, this obviously implies a competitive advantage, at least over competitors with similar business plans, but without comparable backup. Otherwise, if this *diversification effect* is not observed, that is if the risk of the unit would simply remain its stand alone risk $\rho_i(W_i)$, then from a risk perspective there is no reason to stay within the company. In that case the shareholders might for instance be tempted to sell off that unit. However, according to the results above, assuming that problem (4.1) admits a solution $(\bar{X}_1, \dots, \bar{X}_n)$, it is very natural to consider the convex monetary risk measure (4.2) as the risk measure of the company and the optimal allocation $(\bar{X}_1, \dots, \bar{X}_n)$ as the businesses of the units after an optimal mutual reinsurance. Since we may assume that $(\bar{X}_1, \dots, \bar{X}_n)$ respects the individual rationality constraints we have that

$$d_i := \rho_i(W_i) - \rho_i(\bar{X}_i) \geq 0 \quad \text{and} \quad D := \sum_{i=1}^{n} d_i = \sum_{i=1}^{n} \big(\rho_i(W_i) - \rho_i(\bar{X}_i)\big) \geq 0 \tag{4.4}$$

where d_i is the diversification effect for unit i, and D is the diversification effect of the company. This gives an elegant answer to the posed questions.

Apparently, the assumption that the risk sharing in (4.1) is over all possible allocations of the aggregate risk W may seem far from reality. Hence, it appears that in a next step one should allow for constraints on the set of allocations $\mathbb{A}(W)$. However, in many cases the optimal allocation coming from solving the unconstrained problem (4.1) indeed exhibits structures which are very often traded, such as linear sharing of the aggregate portfolio or reinsurance by means of stop-loss contracts; see [6, 7, 11, 13]. Moreover, when developing new kinds of (reinsurance) contracts, knowledge of the structure of solutions to the unconstrained problem (4.1) might be of advantage. For a detailed discussion of the optimal risk sharing problem when agents apply convex monetary risk measures we refer to [6, 7, 11, 13].

5 Food for Thoughts

In this section we give an overview of possible research topics in the field of monetary risk measures that are directly related to our presented examples and case studies.

5.1 Appropriate Risk Measures

The class of monetary risk measures is quite broad. Even though monetary risk measures, and in particular convex monetary risk measures, exhibit basic properties that from a certain point of view any risk measure should satisfy, this class of risk measures still includes functions that are not reasonable in application. For instance the worst case risk measure

$$\rho_{worst}(X) := -\inf_{\omega\in\Omega} X(\omega)$$

is a coherent risk measure. But measuring risks by means of ρ_{worst} implies taking no risk, and thus not at all taking an active part in the economy, at least if the markets are assumed to be arbitrage-free. Hence, an important task is, given some specific setting, e.g. a specific field of business or risk profile, to find a suitable convex monetary risk measure for that setting. This involves understanding what suitable in some given setting means, depicting additional requirements that a convex monetary risk measure for that setting should satisfy, and studying and testing the corresponding class of monetary risk measures. In that context, apart from describing certain risk averseness or matching observed structures like e.g. the behavior of risks of large, highly diversified portfolios in certain markets, also numerical issues like stability in optimization play an important role.

5.2 Risk Sharing

The optimal risk sharing problem was outlined in Sect. 4 above. The tasks are to prove the existence of solutions to (4.3), to explicitly characterize these solutions given certain classes of convex monetary risk measures, and to study the problem under additional constraints on the set of feasible allocations.

5.3 Optimization Under Convex Monetary Risk Measures

Many applications of convex monetary risk measures lead to a convex optimization problem which is very often not easily solved analytically. Hence numerical methods have to be applied. However, since convex monetary risk measures are highly non-linear and non-smooth structures, they very often behave poorly in optimization. The field of convex optimization provides a lot of tools to deal with even these kind of problems. The challenge here is to spot the right methods, and maybe to develop new ones which are particularly suited in case of optimization under some convex monetary risk measure, taking advantage of the properties like cash invariance and monotonicity.

6 Summary

The purpose of any risk modeling is, in a first step, to understand what is risk associated to some random outcome and what are the major sources of risk for that outcome. Then, in a second step, the aim is to apply the gained knowledge in order to quantify the risk, thereby opening for the possibility to compare risks of different outcomes and to seek to some extent protection against risk. In case of a financial portfolio two major sources of risk are depicted as being the uncertainty of the exact outcome given multiple scenarios, and the ambiguity about the right probabilistic model for the likeliness of the different scenarios. These sources of risk are accounted for in the theory of convex monetary risk measures which quantify the risk in terms of a cash amount that has to be added to the analyzed portfolio in order to make it acceptable. As, with the increasing complexity of financial products and in the aftermath of the financial crisis, risk analysis and quantification rapidly gains importance, there is a vivid ongoing research activity in the field of (convex) monetary risk measures. Apparently, the developed risk measuring machinery may also be adopted to other than merely financial risks.

References

Selected Bibliography

1. P. Artzner, F. Delbaen, J.M. Eber, D. Heath, Coherent measures of risks. Math. Finance **9**, 203–228 (1999)
2. H. Föllmer, A. Schied, Convex measures of risk and trading constraints. Finance Stoch. **6**, 429–447 (2002)
3. H. Föllmer, A. Schied, Convex and coherent risk measures, in *Encyclopedia of Quantitative Finance* (2010), pp. 355–363
4. H. Föllmer, A. Schied, *Stochastic Finance—An Introduction in Discrete Time*, 3rd edn. (De Gruyter, Berlin, 2011)
5. M. Frittelli, E. Rosazza Giannini, Putting order in risk measures. J. Bank. Finance **26**, 1473–1486 (2005)

Additional Literature

6. B. Acciaio, Optimal risk sharing with non-monotone monetary functionals. Finance Stoch. **11**, 267–289 (2007)
7. P. Barrieu, N. El Karoui, Inf-convolution of risk measures and optimal risk transfer. Finance Stoch. **5**, 269–298 (2005)
8. F. Black, M. Scholes, The pricing of options and corporate liabilities. J. Polit. Econ. **81**, 637–654 (1973)
9. P. Carr, H. Geman, D. Madan, Pricing and hedging in incomplete markets. J. Financ. Econ. **62**, 131–167 (2001)
10. V. Fasen, C. Klüppelberg, A. Menzel, Quantifying extreme risk, in *Risk – A Multidisciplinary Introduction*, ed. by C. Klüppelberg, D. Straub, L. Welpe (2014)

11. D. Filipović, G. Svindland, Optimal capital and risk allocations for law- and cash-invariant convex functions. Finance Stoch. **12**, 423–439 (2008)
12. D. Heath, Back to the future, in *Plenary Lecture*, First World Congress of the Bachelier Finance Society, Paris (2001)
13. E. Jouini, W. Schachermayer, N. Touzi, Optimal risk sharing for law invariant monetary utility functions. Math. Finance **18**, 269–292 (2008)
14. K. Larsen, T. Pirvu, S. Shreve, R. Tütüncü, Satisfying convex risk limits by trading. Finance Stoch. **9**, 177–195 (2005)
15. F. Knight, *Risk, Uncertainty, and Profit* (1921). Originally published
16. A.N. Kolmogorov, Grundbegriffe der Wahrscheinlichkeitsrechnung. Ergeb. Math. (1933)
17. K. Mainzer, The new role of mathematical risk modeling and its importance for society, in *Risk – A Multidisciplinary Introduction*, ed. by C. Klüppelberg, D. Straub, L. Welpe (2014)
18. H.M. Markowitz, Portfolio selection. J. Finance **7**, 77–91 (1952)
19. R.C. Merton, The theory of rational option pricing. Bell J. Econ. Manag. Sci. **7**, 141–183 (1973)

Chapter 6
Quantifying Extreme Risks

Vicky Fasen, Claudia Klüppelberg, and Annette Menzel

Understanding and managing risks caused by extreme events is one of the most demanding problems of our society. We consider this topic from a statistical point of view and present some of the probabilistic and statistical theory, which was developed to model and quantify extreme events. By the very nature of an extreme event there will never be enough data to predict a future risk in the classical statistical sense. However, a rather clever probabilistic theory provides us with model classes relevant for the assessment of extreme events. Moreover, specific statistical methods allow for the prediction of rare events, even outside the range of previous observations. We will present the basic theory and relevant examples from climatology (climate change), insurance (return periods of large claims) and finance (portfolio losses and Value-at-Risk estimation).

Keywords Climate risk · Extreme risk analysis · Extreme value distribution · Financial risk · Peaks over thresholds

Mathematics Subject Classification (2010) 60G70 · 62G32

The Facts

- Modern risk measures like Value-at-Risk and Expected Shortfall are defined by high quantiles, such that the probability of a large loss is small.

V. Fasen
Institute of Stochastics, Department of Mathematics, Karlsruhe Institute of Technology, Kaiserstr. 89, 76133 Karlsruhe, Germany

C. Klüppelberg (✉)
Chair of Mathematical Statistics, Center for Mathematical Sciences, Technische Universität München, Boltzmannstr. 3, 85748 Garching bei München, Germany
e-mail: cklu@tum.de

A. Menzel
Chair of Ecoclimatology, Center of Life and Food Sciences Weihenstephan, Technische Universität München, Hans-Carl-von-Carlowitz-Platz 2, 85354 Freising-Weihenstephan, Germany

C. Klüppelberg et al. (eds.), *Risk – A Multidisciplinary Introduction*,
DOI 10.1007/978-3-319-04486-6_6,

- Poisson's classic theorem on rare events (also called the law of small numbers) is the basis for extreme value statistics, because it says that the Poisson distribution is the limit of binomial distributions with very small success probabilities.
- The distribution of maxima of large samples can only be a Generalized Extreme Value (GEV) distribution. This is one of the most fundamental results of extreme value theory. On this basis methods to estimate far out tails and high quantiles were developed.
- Another method to estimate far out tails and high quantiles is the Peaks-Over-Threshold (POT) method using the fact that exceedances over high thresholds for large samples follow a Generalized Pareto distribution (GPD).
- We quantify extreme events for three data examples:
  - yearly temperature maxima from 1879–2008;
  - claim sizes of a Danish fire insurance;
  - daily returns of the Standard and Poors 500 Index.

1 Introduction

Extreme risks accompany our lives. Although every single person hopes that she does not suffer any losses, some lose a fortune in a financial crises, some others lose their property in a hurricane, or they have to leave their homes because of a nuclear accident, another person may even lose her life in a car accident or because of a terrorist attack. Whereas our ancestors took dangers and risks as God-given, nowadays we trace the occurrence of most types of risk back to the actions of men. This implies that risk is precisely calculable (an assumption that is mostly wrong), and that somebody has to be responsible. This applies to technical risk, where safety measures are implemented in order to prevent disasters, which still happen occasionally. We even try to adapt to risk of natural catastrophes, when we develop strategies like, for instance, building dikes or simply sign an insurance contract.

In a society guided by such believes it is natural to require formulas from Mathematics and Statistics for risk assessment. It is within this framework that *extreme value theory* and *extreme value statistics* find their natural place. However, the modeling and the assessment of extreme events is not so simple and cannot be gained with standard methods.

We illustrate the problem with a classical example.

Illustration 1.1 (Determine the Height of a Dike) In the Netherlands, where substantial parts of the country are below sealevel, dikes of appropriate height are of vital importance as protection against floods. The dikes have to be built higher than a wave height, which happens at most every 10,000 years. How high has the dike at least to be? Or formulated otherwise, how does one estimate the height of the highest wave in 10,000 years, if one has only measurements of some hundred years available? The problem is to estimate the probability of an event which is more extreme than any recorded to date. This requires a special method, which is provided by statistical methods based on extreme value theory.

Extreme value theory is a fundamental mathematical theory, which can be transferred to statistical methods. It was developed during the last 50 years and is not undebated. Extreme value theory allows (under appropriate conditions) to predict rare events, which are not included in the previous observations because of their rareness. Based on extreme data (later they will be yearly temperature maxima, large insurance claims and large changes in a financial time series) it is possible to extrapolate the data for the prediction of events, which cause higher temperatures, insurance claims or financial losses than have ever been observed before. Naturally it is easy to criticize this extrapolation out of the sample data and it is clear that extrapolation is unreliable by nature. However, extreme value theory provides a solid mathematical basis, and no other reliable alternative has been suggested. We cite the following assessment of Professor Richard Smith (http://www.unc.edu/~rls/), who has substantially contributed to the development of extreme value statistics: "There is always going to be an element of doubt, as one is extrapolating into areas one doesn't know about. But what extreme value theory is doing is making the best use of whatever you have about extreme phenomena".

We emphasize that the statistical treatment of rare events as the far-out tail behavior can only succeed with specific methods, which implement probabilistic results of extreme value theory into the estimation procedure and, hence, compensate for the insufficient amount of data. This will be the topic of Sects. 3 and 4. Parts of this chapter have corresponding parts in Fasen and Klüppelberg [29].

2 Extreme Risks

2.1 Climate Risk

Fire, water, air—these three basic elements cause climate or weather-related natural disasters. They comprise meteorological hazards (such as storm, hail, lightning), hydrological (flooding, mass movement), and climatological ones (such as extreme temperatures, heat waves, drought, forest fire). Apart from devastating earthquakes in Chile, Haiti (2010) and Japan, New Zealand (2011), making 2011 the costliest year ever, the natural catastrophe losses in the last few years were dominated by weather-related catastrophes, such as devastating floods in Pakistan (2010) and Thailand (2011), the Winter Storm Xynthia in western Europe (2010), Hurricane Sandy in the US (2012), wildfires in Russia (2010) and the summer drought in the US (2012) (see also Sect. 2.3 Insurance Risks). According to Munich Re data, there is an increasing trend of these natural disasters in respect to intensities, frequencies, damages and losses. The Intergovernmental Panel on Climate Change (IPCC) concluded in its last report in 2007 (see [6]) that in past records the dominant signal was significantly increased in the values of exposure at risk. However climate change has likely altered and will virtually certainly alter also the occurrence of extreme events dramatically: frequency and magnitude of extreme events are strongly linked to anthropogenic induced climate change.

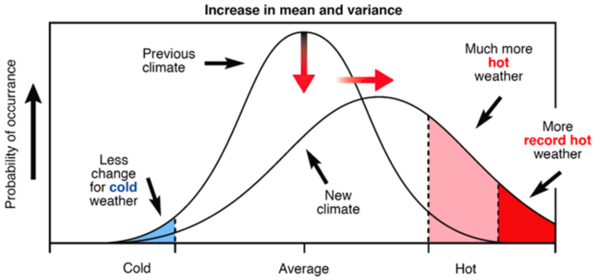


Fig. 1 Illustration of the consequences of an increase of temperature in mean and variance

The latest IPCC report confirmed a 100-year linear trend (1906–2005) of 0.74 °C, more precisely, eleven of the last twelve years (1995–2006) ranked among the 12 warmest years in the instrumental record of global surface temperature since 1850. Most of the observed warming since the mid-20th century is very likely due to the observed increase in anthropogenic greenhouse gas concentrations. Linked to this climate change are marked observed changes in extreme events, much more intense and longer droughts since the 1970s, particularly in the tropics and subtropics, higher frequency of heavy precipitation events, or widespread changes in extreme temperatures. For the latter one, a human contribution to the observed trends is likely. Also future trends have been assessed by simulation of different scenarios with strong impacts on extreme events, e.g., increase in intense tropical cyclone activity or incidence of extreme high sea level are likely at the end of the 21th century. Due to the importance of extreme events the IPCC published a Special Report Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation (SREX) in 2012.

Many important research questions are linked to this increase in weather related extreme events. First of all, is climate becoming more extreme under climate change conditions? This question has traditionally been answered by fitting Gaussian distributions to temperatures. Figure 1 displays how an increase in mean and variance of temperature causes more hot and more record hot weather. However, Gaussian distributions do not provide a good fit for the distribution tails of high temperature measurements.

Second, if there are changes in extremes, which vulnerability of humans is to be expected? Not all extreme events end in disasters. The most recent World Risk Report of 2012, published by the BündnisEntwicklungshilfe in cooperation with the United Nations University (UNU-EHS) (http://www.weltrisikobericht.de), summarizes the risk by natural hazards to nations with different vulnerability, starting with

(1) the likelihood of extremes to occur (exposition),
(2) the vulnerability of societies with respect to infrastructure, housing, food, poverty, economy,
(3) the coping capacity based on governance, catastrophe precautions, medical situation, social networks, insurances, and
(4) the adaptation capacity linked to education, environmental protection, projects and investments.

Similarly, it is a question of tremendous importance how the occurrence of physical extreme events translates to extreme biological impacts or hazards which threaten the fitness and survival of ecosystems more than any change in mean conditions (cf. Hegerl, Hanlon, and Beierkuhnlein [5], Menzel, Seifert, and Estrella [8]). Not all rare climatological events translate into extreme impacts: the responses in nature may be non-linear, the species may be resilient, resistant, recover fast, or are well adapted by management. Due to this variation in response, always more and more data on impacts of extreme events are needed. The goal is to bridge the gap between extreme events and extreme impacts, especially for climatological hazards, such as temperature extremes, heat waves, cold spells, frost events, drought or fire. They impact primarily agricultural and forest ecosystems, however, as combined, longer lasting events their proper statistical modeling and assessment is a scientific challenge.

2.2 Financial Risks

The Basel Committee for Banking Supervision (http://www.bis.org/bcbs/) recommends for insurance companies and financial institutions the building of capital reserves to hedge against unpredictable risks. This is in Germany explicitly required by the regulatory authorities, the BAFIN (Bundesanstalt für Finanzdienstleistungsaufsicht, http://www.bafin.de/) in the framework of "Basel II" for banks (http://www.bis.org/publ/) and in the framework of "Solvency II" for insurance companies (http://ec.europa.eu/internal_market/insurance/). The risk management department of every company is responsible for the respective calculations of the required capital reserves and their administration, which requires a mathematical-statistical training.

The focus of Basel II, which was initially published in June 2004, was to manage and measure *credit risks*, *operational risks* and *market risks*. In this chapter we will only pay attention to market risk, the risk that a value of a portfolio will change due to movements in the market risk factors as, e.g., interest rates, foreign exchange rates, equity prices and commodity prices.

In the Basel framework the capital requirement for market risk is based on the so-called *Value-at-Risk*, which is the p-quantile of the portfolio risk, and is defined as follows.

Let X be the financial risk in terms of the *daily losses*, defined as the negative profit/loss of the market portfolio. To be precise, if Z_t for $t = 1, 2, \dots$ denote the daily market prices of the portfolio, then the losses X_t represent the daily negative log-returns defined as $X_t = -(\log Z_t - \log Z_{t-1}) \approx -(Z_t - Z_{t-1})/Z_{t-1}$, approximating the negative relative price changes for each day.

The distribution function of the daily portfolio loss X is given by $F(x) = \mathbb{P}(X \leq x)$ for $x \in \mathbb{R}$. We define the *quantile function* of F or *Value-at-Risk* as

$$\mathrm{VaR}_p(X) = F^{-1}(p) = \inf\{x : F(x) \geq p\}, \quad p \in (0, 1). \tag{2.1}$$

(Note that for strictly increasing F this is simply the analytic inverse.) Hence, $\mathrm{VaR}_p(X)$ is the smallest number such that the probability of a loss larger than $\mathrm{VaR}_p(X)$ does not exceed $1-p$. Then for a large value of p (usually $p=0.95$ or larger) $\mathrm{VaR}_p(X)$ is a prominent risk measure.

Depending on the specific risk, choices are $p=95\ \%$ (0.95) or $p=99\ \%$ (0.99) or even $p=99.9\ \%$ (0.999). In the case of market risks $p=99\ \%$.

By the perception and experiences gained through the financial crises, which started in 2007, the Basel Committee on Banking Supervision decided a reformation of Basel II to strengthen the regulation, supervision and risk management of the banking sector in September 2010. This revision had to be implemented until 31 December 2011 [16] and introduced—as a response to the crises—a *stressed Value-at-Risk* requirement taking into account a historic one-year observation period relating to significant losses, which must be estimated in addition to the classical Value-at-Risk based on the recent one-year observation period. Basel III [15] now aims at raising the resilience of the banking sector by strengthening the risk coverage of the capital reserves. It suggests reforms of capital requirements for counterparty credit risk using stressed inputs, addresses the systemic risk arising from the interconnectedness of banks and other financial institutions, and supplements the risk-based capital requirement to constrain too high leverage (details to the changes in market risk can be found in http://www.bis.org/publ/bcbs193.htm). The implementation of Basel III will start in 2013.

Typical methods to estimate the Value-at-Risk in practice are *historical simulations*, the *variance-covariance method* and *Monte Carlo simulation*.

The "historical simulation method" simply estimates $\mathrm{VaR}_p(X)$ by the corresponding empirical quantile based on the required one year of data. For instance, $\mathrm{VaR}_{0.99}(X)$ is estimated as the largest 1 % of daily losses. Alternatively, a weighted estimation scheme is used, which gives higher weights to those data near to the current date and lower to the more distant data. Criticism of this method is obvious: reliable estimation of high quantiles like $\mathrm{VaR}_{0.99}(X)$ requires a large amount of high losses, but 1 % of the required one year of data provides no reliable estimator. Consequently, the estimated $\mathrm{VaR}_{0.99}(X)$ depends very much on the present market situation and estimates can differ substantially almost from day to day. We shall analyse the Standard and Poors 500 Index data during 1990–2004, abbreviated as S&P500. Moreover, $\mathrm{VaR}_{0.99}(X)$ is supposed to predict future high losses, which may be substantially higher than losses of the previous year and requires extrapolation outside the observations.

For the "variance-covariance method" the risk factors are assumed to be multivariate normal distributed. Then the distribution function of the portfolio X is a one-dimensional normal distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma > 0$ determined by the portfolio weights, the means and variances of the components and the pairwise correlations of the components. The loss distribution F of X is given by

$$F(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} \mathrm{e}^{-\frac{(y-\mu)^2}{\sigma^2}}\,\mathrm{d}y \quad \text{for } x \in \mathbb{R}. \tag{2.2}$$

Then $\mathrm{VaR}_{0.99}(X) = \mu + \sigma z_{0.99}$, where $z_{0.99}$ is the 0.99-quantile of the standard normal distribution. It is particularly easy to estimate and to update, when the estimates

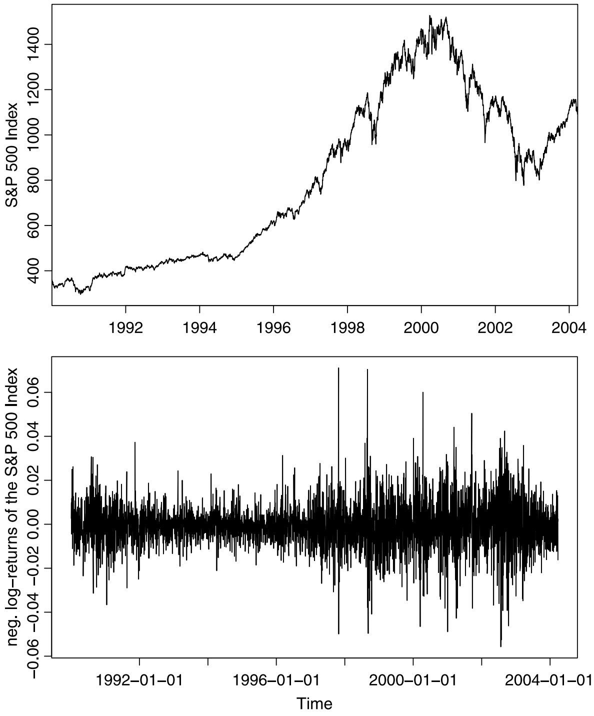


Fig. 2 The S&P500 (*top*) and the corresponding losses (*bottom*) during 1990–2004

for μ and σ change in time. In Fig. 2 we see the S&P500 (left) and its losses (right) during 1990–2004.

From this example we see that the normal model is completely inadequate: The histogram (empirical density) of the daily losses of the S&P500 and the normal density with mean and standard deviation estimated from the data are depicted in Fig. 3. The histogram clearly shows that the daily losses of the S&P500 have more mass in the tails than the normal distribution; i.e. for ±0.03 and larger/smaller the histogram exhibits more large/small values than is likely for the normal distribution. This mismatch leads to an underestimation of the required capital reserve. The fact that the empirical distribution and the normal distribution differ around 0 is for risk management based on high quantiles irrelevant. Moreover, financial loss data are usually negatively skewed and leptokurtic, again properties which can not be captured by a Gaussian distribution.

The third VaR estimation method is the "Monte Carlo simulation". Here a more sophisticated parametric distributional model is fitted to the daily losses, its parameters are estimated, and then large numbers of random samples of arbitrary length are simulated, its VaR estimated for each sample, and then the average VaR is taken as an estimate. This method can be made more efficient by variance reduction methods (Glasserman [33], Korn [38]), and estimates VaR for a given model with arbitrary precision. However, the estimate depends on the chosen model (as it does for the

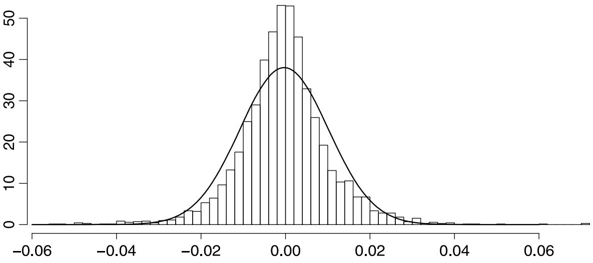


Fig. 3 Histogram of the daily losses of the S&P500 in comparison to the density of the normal distribution. The mean μ and the variance σ^2 have been estimated by their empirical versions

normal model in the variance-covariance method), so model risk can be considerable; cf. Chap. 10, Bannör and Scherer [14].

Remark 2.1 (i) In the Basel II market risk framework the calculation of the capital reserves requires as risk measure the Value-at-Risk for a holding period of 10 days at a confidence level 0.99 %. A standard method in practice to calculate the Value-at-Risk for a holding period of 10 days is to calculate the Value-at-Risk for a holding period of one day and scale it by $\sqrt{10}$. This scaling factor is based on the scaling property of the normal distribution and can be completely wrong.

(ii) In the amendments to the Basel II accord, which have been incorporated into Basel III ([15]), the $\mathrm{VaR}_{0.99}$ has been extended to incorporate so-called stressed periods like the financial crises during 2007/2008. Let X denote the loss of a market risk portfolio (over the next 10 days) and $\mathrm{VaR}_{0.99,avg}(X)$ the average of the estimated VaR values of the preceding 60 business days. Then the new capital requirement has to be calculated according to

$$\begin{aligned}&\max\bigl\{\mathrm{VaR}_{0.99}(X), m_c\mathrm{VaR}_{0.99,avg}(X)\bigr\}\\&\quad+\max\bigl\{\mathrm{SVaR}_{0.99}(X), m_s\mathrm{SVaR}_{0.99,avg}(X)\bigr\}\end{aligned} \tag{2.3}$$

where m_c and m_s are multiplication factors, which are not smaller than 3 (and are related to the ex-post performance of the bank's model). The quantity SVaR is the Value-at-Risk of the loss portfolio estimated from historical data of a 12-month period of significant financial stress; e.g the financial crises 2007/2008.

(iii) Finally, we argue that the Value-at-Risk is not an appropriate risk measure. It is appropriate for the dike height of Illustration 1.1, for financial risk however, the situation is different. If a flood with waves higher than the dike happens, the dike usually breaks and nothing can be done for salvation. The land behind the dike disappears under water. For financial risks, however, it is extremely relevant to know also the amount of resulting losses. This quantity is taken into account, when using the *Average Value-at-Risk* as an alternative risk measure, which describes the expected losses given a loss larger than the Value-at-Risk happens. It is given as

$$\mathrm{AVar}_p(X) = \frac{1}{1-p}\int_p^1 \mathrm{VaR}_\gamma \, d\gamma$$

(cf. Chap. 5, Biagini, Meyer-Brandis, and Svindland [19] for a detailed introduction into risk measures). If X has continuous distribution function F, then $\mathrm{AVar}_p(X) =$

$\mathbb{E}(X \mid X > \mathrm{VaR}_p(X))$, which represents exactly the expected losses, given an extreme loss occurs. A second drawback of the Value-at-Risk is that it is in general not subadditive, i.e. $\mathrm{VaR}_p(X+Y) \leq \mathrm{VaR}_p(X) + \mathrm{VaR}_p(Y)$ may not hold for risks X, Y. Subadditivity reflects the diversification effect. It is better to have a portfolio of risks than several individual risks. However, if for example X and Y are independent with distribution $F(x) = 1 - \frac{1}{1+x}$ for $x \geq 0$, then $\mathrm{VaR}_p(X+Y) > \mathrm{VaR}_p(X) + \mathrm{VaR}_p(Y)$ and there is no chance for risk diversification. In contrast, the Average Value-at-Risk is a subadditive risk measure. Although there were serious attempts to communicate to regulators that the Average Value-at-Risk may be a more appropriate risk measure (cf. Danielsson et al. [24]), this academic initiative was not successful. The lobby work of the banks has prevented this: the capital reserves calculated on the basis of Expected Shortfall would be substantially larger than the Value-at-Risk.

2.3 Insurance Risks

Insurance companies take over the risks of their customers. Typical insurance risks are health problems, death, accidents, burglary, floods and fire. With the acquisition of an insurance contract customers transfer their risk to an insurance company, which is then financially liable to insurance claims. Also the insurance company does not know the risk for a claim to happen to a customer, but by selling a large number of policies, it subsumes customers with similar risk in a portfolio and takes advantage of the fact that in a large portfolio with similar and independent risks the total claim amount is constant in mean. In probability theory this fact is proved and is called the *Law of Large Numbers*. For the insurance company this makes the risk of a portfolio of similar and independent risks calculable. Random fluctuations in the portfolio are hedged by reserves. In this context insurance companies have to evaluate the frequency as well as the severity of risks. To do this they have to suggest appropriate risk models and estimate the model parameters, they have to analyze the model statistically and test it under extreme conditions. But they also have to calculate the premiums and reserves. As capital reserves of insurance companies are substantial, it is also subject to capital regulations like Basel II. Taking the total insurance business into account, new regulations are being implemented under Solvency II, following the very same ideas as the Basel framework. We do not want to explain these ideas in detail, but instead want to present the very traditional concept of the *return period*, which is used universally to describe extreme events and serves as a risk measure, in particular, for abnormally large insurance claims.

Large claims are rare events with very high costs for an insurance company. They include natural catastrophes like earth quakes, fire, storms or floods, which are typical events where large claims occur (cf. Fig. 4), but also so-called *man-made claims* from large industrial structures. In 2010 the earth quake in Chile and the sinking of the drilling rig "Deepwater Horizon" were large claims, in 2011 the event in Fukushima, which combined natural catastrophe with man-made disaster, and the hurricane Sandy was a major catastrophe in 2012. It is common practice that an insurance company insures itself against large claims by a contract with a

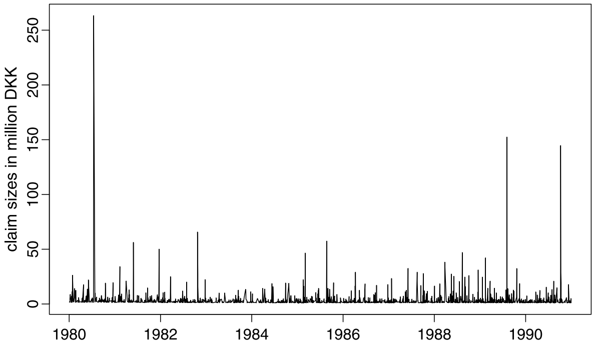


Fig. 4 Claim sizes of a Danish fire insurance during 1980–1990 in million Danish Krone (DKK)

reinsurance company. To-date the hurricane Katrina in 2005 is the most expensive insurance claim in history with about 76.25 billion US-Dollar, followed by the earth quake and the tsunami in Japan by 35.7 billion US-Dollar, hurricane Sandy in 2012 with about 35 billion US-Dollar, hurricane Andrew in 1992 with about 26.1 billion US-Dollar and the terror attack to the World Trade Center in 2001 with about 24.3 billion US-Dollar (the data are going back to http://de.statista.com/).

It is a common feature of large claims that they happen rarely, and hence little data are available to allow for reliable statistical prediction. But obviously, an insurance company and, even more so, a reinsurance company has to prepare for extreme events. Certain quantities can help to assess the frequency and severity of large claims. In the following we denote by $X_1, X_2, \ldots$ the accumulated claims per year of an insurance or reinsurance company (X_k is the total claim amount in year k) and we assume that these yearly claim amounts are independently and identically distributed (shortly i.i.d.) with distribution function F. We further assume that $F(0) = 0$ (a claim can only be positive) and that $F(x) < 1$ for all $x \in \mathbb{R}$ (claims can be arbitrarily large, which has been proved over and over by reality). We denote by $\overline{F}(x) = 1 - F(x)$ for $x \geq 0$ the so-called *tail of* F. We want to determine now the distribution of the first year in the future, where the yearly total claim exceeds a fixed yearly reserve u for the first time. This year is determined by

$$Z(u) = \min\{k \in \mathbb{N} : X_k > u\}.$$

Setting

$$q := \mathbb{P}(X > u) = \overline{F}(u), \tag{2.4}$$

the random variable $Z(u)$ is geometrically distributed with parameter q, i.e. the probability that $Z(u)$ takes the value k is given by

$$\mathbb{P}\big(Z(u) = k\big) = (1-q)^{k-1}q \quad \text{for } k \in \mathbb{N}$$

(in $k-1$ years we experience no excess, but then in year k there is an excess). The *return period* is now the mean waiting time until a yearly total claim amount exceeds the threshold u (denoted by $\mathbb{E}(Z(u))$), where $\mathbb{E}$ is the mathematical symbol for expectation or mean. The expectation is then

$$\begin{aligned}\mathbb{E}\big(Z(u)\big) &= \sum_{k=1}^{\infty} k\mathbb{P}\big(Z(u)=k\big) = q\sum_{k=1}^{\infty} k(1-q)^{k-1}\\ &= \frac{1}{q} = \frac{1}{\mathbb{P}(X>u)} = \frac{1}{\overline{F}(u)}.\end{aligned} \tag{2.5}$$

This provides now a trick to estimate the expectation. The standard way to estimate the expectation is by the arithmetic mean (the sum of all observation values divided by the number of all observations). Note however that, in order to do this, one would need many years, where exceedances have happened. Since the events we are interested in are rare, this classical statistical method can not be applied simply by lack of data. However, estimation via the right hand side of (2.5) is also not straightforward: the problem has been shifted now to the estimation of the tail $\overline{F}(u)$. Also for this tail estimation only few data are available. However, we can now compensate the lack of data by using clever methods from extreme value theory. We will explain this in detail in Sects. 3 and 4.

But also the inverse problem is of great interest. The insurance company wants to calculate premiums and reserves such that a yearly total claim amount larger than u should happen with a probability 0.1 at most every 50 years, which means that $\mathbb{P}(Z(u) \le 50) \le 0.1$. Since

$$\mathbb{P}\big(Z(u)\le 50\big) = q\sum_{i=1}^{50}(1-q)^{i-1} = 1-(1-q)^{50},$$

we have $1-(1-q)^{50} = 0.1$. This implies that $q = 0.002105$. Hence the return period in this example is $1/q = 475$ years. For the calculation of premiums and reserves we need now also the threshold u, and this requires the estimation of the quantile of the distribution function F. With the definition of the p-quantile in (2.1) we conclude with (2.4) that $u = x_{1-q}$ holds. We come back to this in Sect. 4.

3 Basic Extreme Value Theory

In the following we present the most important concepts for realistic modeling and quantification of rare events. The precise mathematical background as well as many application examples can be found in Beirlant et al. [1], Coles [3], Embrechts, Klüppelberg, and Mikosch [4], McNeil, Frey, and Embrechts [7], Reiss and Thomas [9], Stephenson [43] gives an excellent overview on extreme events in climatology.

Figure 3 presents a rather typical figure in many statistical applications areas. The normal distribution is often wrongly applied to extreme risk problems. This can only be explained by the fact that everybody with a basic statistical education has learnt about the normal distribution. Moreover, the sum of normally distributed random variables is again normally distributed, and the mean and the standard deviation of this sum are easy to calculate.

There is no doubt that the normal distribution is a very important distribution in probability theory and statistics: it is the limit distribution for sums. For a sequence of i.i.d. random variables $X_1, X_2, \dots$ (under the weak condition of a finite variance), we have

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{n} \big(X_k - \mathbb{E}(X_k)\big) \xrightarrow{d} \mathcal{N}\big(0, \sigma^2\big) \quad \text{as } n \to \infty,$$

where the random variable on the right hand side is normally distributed with distribution function as in (2.2). The symbol $\xrightarrow{d}$ stands for convergence in distribution; i.e. the distribution functions of the random variables on the left hand side converge to the normal distribution function with mean 0 and variance σ^2. This is the so-called *Central Limit Theorem*. Because of this very basic result the normal distribution is an excellent model for random variables, which can be approximated by a sum of many small random effects. The great German mathematician Carl Friedrich Gauß (1777–1855) has derived it in his book [32].

It has long been known that the normal distribution is unrealistic for risk considerations. But which model is a good model for extreme events? The answer to this question has been given by the great French mathematician Siméon [40] (1781–1840), which we formulate nowadays as follows.

Theorem 3.1 (Poisson Theorem, [40]) *A statistical experiment with possible outcome E_n is repeated independently n times. The probability that the event E_n happens in one of the n trials is $\mathbb{P}(E_n) = p_n$. If $\lim_{n\to\infty} np_n = \tau$ holds for some $0 < \tau < \infty$, then*

$$\begin{aligned} &\lim_{n\to\infty} \mathbb{P}\,(\textit{in exactly } m \textit{ of the } n \textit{ trials we have outcome } E_n) \\ &\quad = \lim_{n\to\infty} \binom{n}{m} p_n^m (1-p_n)^{n-m} = \mathrm{e}^{-\tau} \frac{\tau^m}{m!} \quad \textit{for } m = 0, 1, 2, \dots, \end{aligned} \tag{3.1}$$

where $\binom{n}{m} = \frac{n!}{m!(n-m)!}$ with $0! = 1$ and $m! = 1 \cdot 2 \cdots m$.

In honor of Poisson, the distribution on the right hand side of (3.1) is called Poisson distribution with parameter τ, abbreviated by $\mathrm{Poi}(\tau)$. The distribution on the left hand side of (3.1) (before the limit is taken) is the binomial distribution $\mathrm{Bin}(n, p_n)$, which for large n and small p_n approximates the Poisson distribution (cf. Fig. 5). Note that $\lim_{n\to\infty} np_n = \tau > 0$ implies obviously that $\lim_{n\to\infty} p_n = 0$. Hence the events E_n happen with vanishing probability, when the number of trials n is getting large. For this reason the Poisson distribution is also called the distribution of rare events. We want to present some ideas concerning the applicability of the

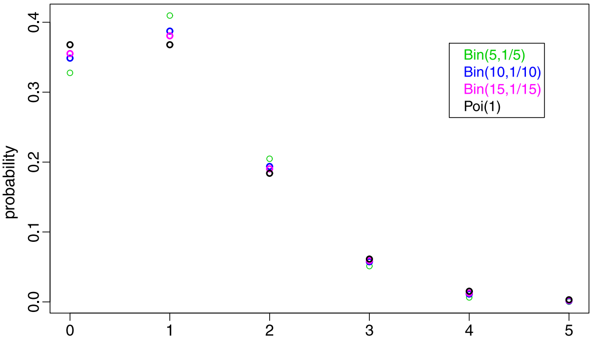


Fig. 5 Counting density of Bin(5, 1/5)-, Bin(10, 1/10)-, Bin(15, 1/15)-distribution and the Poi(1)-distribution. Note that for all parameters of the binomial distributions presented $np = 1$ holds

Poisson distribution, which leads to the two essential statistical concepts of extreme value theory. The first statistical method is called the *blocks method*, and the second one the *Peaks-Over-Thresholds (POT) method*. Which method to use depends on the question posed and on the data at hand. We will come back to both statistical methods in Sect. 4.

In the following we present the necessary mathematical results to understand the concepts. Let $X_1, \ldots, X_n$ be a sample of random variables; think for instance of yearly total claim amounts of an insurance company or losses of a financial asset. We assume that $X_1, \ldots, X_n$ are i.i.d. having the same distribution function as the random variable X; we denote it again by $F(x) = \mathbb{P}(X \le x)$ for $x \in \mathbb{R}$.

We show first how to use the Poisson Theorem 3.1 for the description of the behavior of the maximum of a sample and investigate in a first step the so-called *partial maxima*

$$M_n = \max(X_1, \ldots, X_n) \quad \text{for } n \in \mathbb{N}.$$

As in real life we assume that risks larger than any we have observed before can continue to occur. This is formulated mathematically by investigating $\mathbb{P}(M_n \le u_n)$, where the sequence u_n increases with n (and hence with M_n). Then the following fundamental result holds (which one can prove by means of the Poisson Theorem 3.1):

$$\lim_{n\to\infty} n\mathbb{P}(X_1 > u_n) = \tau \quad \Longleftrightarrow \quad \lim_{n\to\infty} \mathbb{P}(M_n \le u_n) = \mathrm{e}^{-\tau}. \tag{3.2}$$

We want to motivate the implication from the left side to the right side:

Consider a rare event E, for example the event that the loss of a financial asset at a day is larger than a threshold u for large u. The daily losses of an asset constitute again a sample $X_1, \ldots, X_n$. Then

$$p = \mathbb{P}(E) = \mathbb{P}(X > u).$$

Invoking the same argument as Poisson, we find that the probability that the event E within the sample occurs m times is given by

$$\binom{n}{m} p^m (1-p)^{n-m} \quad \text{for } m = 0, \ldots, n;$$

i.e. it is $\mathrm{Bin}(n, p)$-distributed. Now we let u depend on n in the sense that u_n increases with the sample size n. Then p becomes p_n, which converges to 0, and E becomes $E_n = \{X > u_n\}$. When u_n is chosen such that

$$\lim_{n\to\infty} np_n = \lim_{n\to\infty} n\mathbb{P}(X > u_n) = \tau \in (0, \infty),$$

then the Poisson Theorem 3.1 implies

$$\lim_{n\to\infty} \binom{n}{m} p_n^m (1-p_n)^{n-m} = \mathrm{e}^{-\tau} \frac{\tau^m}{m!} \quad \text{for } m = 0, 1, 2, \ldots.$$

In particular,

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}(M_n \le u_n) &= \lim_{n\to\infty} \mathbb{P}(E_n \text{ never occurs in the } n \text{ trials}) \\ &= \lim_{n\to\infty} \binom{n}{0} p_n^0 (1-p_n)^n = \mathrm{e}^{-\tau}.\end{aligned}$$

Consequently, we have shown how by the Poisson Theorem 3.1 the right hand side follows from the left hand side of (3.2). We shall resist to prove the reverse here.

The following result by [31] dating back to 1928 complements the above result; it describes precisely the possible limit distributions of partial maxima and provides the relevant tools for the estimation of tails and quantiles. For extreme value theory the Theorem of Fisher and Tippett is of equal fundamental importance as the Central Limit Theorem. The English statistician Ronald A. Fisher (1890–1962) has been one of the creators of modern statistics, working in many diverse areas.

Theorem 3.2 (Fisher-Tippett Theorem, [31]) *Let $X_1, X_2, \ldots$ be i.i.d. random variables, and $a_n > 0$ and $b_n \in \mathbb{R}$ appropriate constants. Moreover we assume that*

$$\lim_{n\to\infty} \mathbb{P}\big(\max(X_1, \ldots, X_n) \le a_n x + b_n\big) = G(x) \quad \textit{for } x \in \mathbb{R} \tag{3.3}$$

holds for a distribution function G. Then G belongs to the class $\{G_{\gamma,\sigma,\mu} : \gamma, \mu \in \mathbb{R}, \sigma > 0\}$, where

$$G_{\gamma,\sigma,\mu}(x) = \left\{ \begin{array}{ll} \mathrm{e}^{-(1+\gamma\frac{x-\mu}{\sigma})^{-\frac{1}{\gamma}}}, & \textit{if } \gamma \in \mathbb{R}\backslash\{0\}, \\ \mathrm{e}^{-\mathrm{e}^{-\frac{x-\mu}{\sigma}}}, & \textit{if } \gamma = 0, \end{array} \right\} \quad \textit{for} \left\{ \begin{array}{ll} 1+\gamma\frac{x-\mu}{\sigma} > 0, & \textit{if } \gamma \neq 0, \\ x \in \mathbb{R}, & \textit{if } \gamma = 0. \end{array} \right.$$

The class of distributions $\{G_{\gamma,\sigma,\mu} : \gamma, \mu \in \mathbb{R}, \sigma > 0\}$ is called *generalized extreme value distribution (GEV)*. We recall that the *support* of a distribution function is the set of all $x \in \mathbb{R}$, where $0 < F(x) < 1$. Since $G_{\gamma,\sigma,\mu}(x) = G_{\gamma,1,0}(\frac{x-\mu}{\sigma})$, μ is called *location parameter* and σ is called *scale parameter*. The parameter γ is known as *shape parameter* and defines the type of distribution: if $\gamma > 0$ the distribution $G_{\gamma,\sigma,\mu}$ is a Fréchet distribution with support on $[\mu - \sigma/\gamma, \infty)$; if $\gamma = 0$ the distribution $G_{0,\sigma,\mu}$ is a Gumbel distribution with support on $\mathbb{R}$; if $\gamma < 0$ the distribution is a Weibull distribution with support on $(-\infty, \mu - \sigma/\gamma]$. The Fisher-Tippett

Theorem 3.2 thus states that the limit distribution of maxima are necessarily generalized extreme value distributions (and the normal distribution does obviously not belong to this class).

We want to explain the modelling and statistical consequences of the Fisher-Tippett theorem leading to the so-called *blocks method*. Recall the classical central limit theorem, which ensures that the distributions of sums and means of random variables converge to a normal distribution (for i.i.d. and even weakly dependent variables under the assumption of a finite variance). This motivates the modelling of random variables, which can be regarded as sums or means of random quantities by a normal distribution. Similarly, random variables which represent extreme quantities can be modelled by an extreme value distribution; Sect. 4.1 discusses the typical example of yearly maxima. Underlying this example the measurements consist of daily temperature values, and the maximum over every year is considered. So an extreme value distribution is an appropriate model for these yearly maxima. Moreover, the assumption of independence between the different maxima is also realistic as the time between two of such maxima is several months. We will discuss in Sect. 4.1, if the assumption of those maxima being identically distributed is realistic.

Under the conditions of the Fisher-Tippett Theorem 3.2 much more holds. We denote the class $\{H_{\gamma,\sigma} : \gamma \in \mathbb{R}, \sigma > 0\}$ of distribution functions *Generalized Pareto Distribution functions (GPD)*, which are defined as

$$H_{\gamma,\sigma}(x) = \left\{ \begin{array}{ll} 1-(1+\gamma\frac{x}{\sigma})^{-\frac{1}{\gamma}}, & \text{if } \gamma \in \mathbb{R}\backslash\{0\} \\ 1-\mathrm{e}^{-\frac{x}{\sigma}}, & \text{if } \gamma = 0 \end{array} \right\} \quad \text{for} \quad \begin{cases} x \geq 0, & \text{if } \gamma \geq 0, \\ 0 \leq x < -\sigma/\gamma, & \text{if } \gamma < 0. \end{cases}$$

Again γ denotes the shape parameter and σ the scale parameter. Indeed the parameter γ here is the same as in the Fisher and Tippett Theorem 3.2. Then the following theorem holds, which was proved independently by Pickands [39] and by Balkema and de Haan [13].

Theorem 3.3 (Pickands-Balkema-de Haan Theorem) *Assume that the conditions of the Fisher-Tippett Theorem* 3.2 *hold and that F is the distribution function of X. Then there exists a function* $\sigma : (0,\infty) \to (0,\infty)$ *and some* $\gamma \in \mathbb{R}$ *such that*

$$\lim_{u\to\infty} \mathbb{P}\big(X > u + \sigma(u)x \mid X > u\big) = \lim_{u\to\infty} \frac{\overline{F}(u+\sigma(u)x)}{\overline{F}(u)} = \overline{H}_{\gamma,1}(x)$$

for x in the support of $H_{\gamma,1}$.

It is now important for the *Peaks-Over-Threshold (POT)* method that for a large threshold u the following approximation holds by Theorem 3.3, where we set $y = \sigma(u)x$ and use that $\overline{H}_{\gamma,1}(y/\sigma(u)) = \overline{H}_{\gamma,\sigma(u)}(y)$:

$$\mathbb{P}(X > u + y \mid X > u) = \frac{\overline{F}(u+y)}{\overline{F}(u)} \approx \overline{H}_{\gamma,\sigma(u)}(y) \quad \text{for } y \geq 0. \tag{3.4}$$

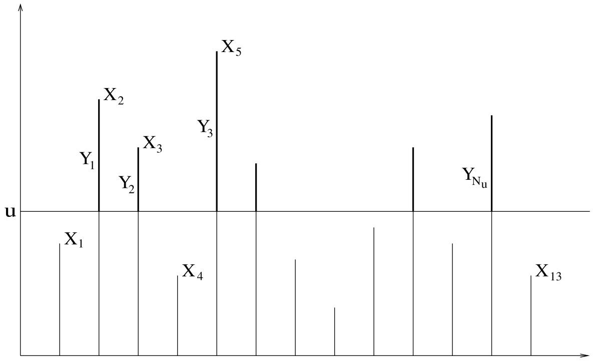


Fig. 6 Data $X_1, \ldots, X_{13}$ with corresponding excesses $Y_1, \ldots, Y_{N_u}$

Note first that an observation larger than $u + y$ is only possible, if the observation is larger than u; this means one needs a so-called *exceedance* of u. Such an observation has then necessarily a so-called *excess* over the threshold u, which is larger than y; cf. Fig. 6. If we investigate the special case that X has distribution $H_{\gamma,\sigma}$, we already have after some calculations that

$$\mathbb{P}(X > u + y \mid X > u) = \frac{\overline{H}_{\gamma,\sigma}(u + y)}{\overline{H}_{\gamma,\sigma}(u)} = \overline{H}_{\gamma,\sigma+\gamma u}(u) \tag{3.5}$$

and $\sigma(u) = \sigma + \gamma u$.

Let now $X_1, X_2, \ldots$ (as illustrated in Fig. 6) be i.i.d. with distribution $H_{\gamma,\sigma}$, then (3.5) means that $Y_1, Y_2, \ldots$, the exceedances of u, namely, $(X - u \mid X > u)$, are $H_{\gamma,\sigma+\gamma u}$ distributed. In the case $\gamma = 0$, where $H_{0,\sigma}$ is the exponential distribution with parameter σ^{-1}, $Y_1, Y_2, \ldots$ are again exponentially distributed with parameter σ^{-1}. This phenomena is well known as *loss-of-memory* property. In the general context of Theorem 3.3 with $X_1, X_2, \ldots$ i.i.d. with distribution function F, (3.4) says that $Y_1, Y_2, \ldots$ are asymptotically generalized Pareto distributed.

In contrast to the Fisher-Tippett Theorem 3.2, which models extreme observations directly, the Pickands-Balkema-de Haan Theorem 3.3 models all large values of a sample, more precisely, all those which exceed a high threshold. This is, where the acronym "Peaks-Over-Thresholds" (POT) originates. Compared to the modelling of yearly extremes (the so-called blocks method) the POT method has a positive and a negative property: on the one hand, taking all exceedances of a sample usually gives more observations, on the other hand, such exceedances can occur in clusters, so that the independence property can be violated. We will apply the POT method in Sect. 4.3.

4 Fundamental Results from Extreme Value Statistics

The books of Beirlant et al. [1], Coles [3], McNeil, Frey, and Embrechts [7], Reiss and Thomas [9] mentioned at the beginning of Sect. 3 provide also their own software package for analyzing extremal events. An extensive overview on quite a number of R-packages and other extreme statistics software is given in [11]; cf. http://www.ral.ucar.edu/~ericg/softextreme.php and http://www.isse.ucar.edu/extremevalues/extreme.html. In particular, we want to mention the Extremes Toolkit (extRemes) developed in R by Eric Gilleland, which provides a user friendly graphical interface.

4.1 Fitting the GEV to a Sample of Extreme Data (the Blocks Method)

The GEV family can be applied as any other parametric family of distributions, whenever the model is justified by the data. Consequently, the GEV has been used for a sample of i.i.d. random variables, which result from some experiment and justify such a model.

Assume we have given yearly maxima $Y_1, \dots, Y_n$, which can be assumed to be i.i.d. GEV distributed with distribution function $G_{\gamma,\sigma,\mu}$ and density $g_{\gamma,\sigma,\mu}$ with realizations $y_1, \dots, y_n$. This means that data are block maxima and every year is a block. Then the maximum likelihood estimator of the parameters is given as

$$(\widehat{\gamma}, \widehat{\sigma}, \widehat{\mu}) = \underset{\gamma,\sigma,\mu}{\operatorname{argmin}} \prod_{t=1}^{n} g_{\gamma,\sigma,\mu}(y_t). \tag{4.1}$$

We will use and slightly extend this concept to assess a possible trend in the location or scale parameter of the data over time.

The next example is classic in this respect: we will fit a GEV to a sample of yearly temperature maxima.

Illustration 4.1 (Climate Risk) Hot days are one of the prominent climatological phenomenon changing. According to IPCC 2007 (cf. [6]), it is very likely that warmer and more frequent hot days over most land areas have occurred in the late 20th century, a human contribution to this trend is likely, and it is—following their likelihood classification—virtually certain that this trend will continue for the 21th century. Daily maximum temperatures for example influence the well-being of humans putting additional stress to the thermal regulation and thus the cardiovascular system. Temperature maxima are very closely linked to average summer temperatures, each degree of warming increasing the maximum temperatures by 1.2 °C in Basel (Switzerland); see Beniston and Diaz [17]. Other projected impacts of more hot days comprise decreasing agricultural and forest yields in warmer environments, reduced energy demand for heating, increased demand for cooling, or declining air quality in cities.

We study long-term changes in daily maximum temperatures recorded at the oldest mountain climate station in the world, the observatory Hohenpeißenberg (977 m above sealevel, south-west of Munich), where regular meteorological observations

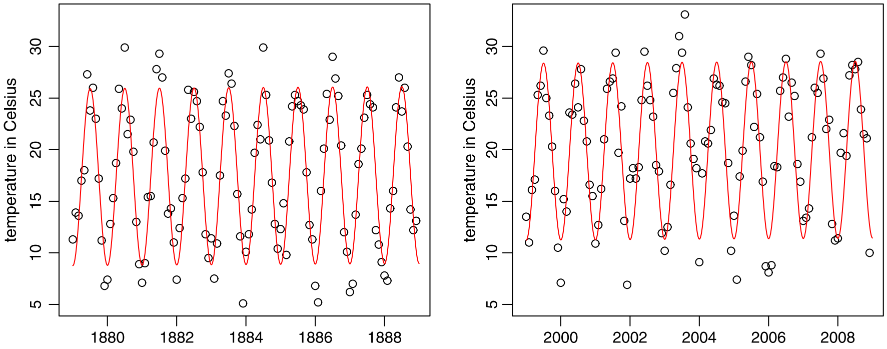


Fig. 7 Two decades of monthly temperature maxima: 1879–1888 and 1999–2008. The *red line* shows the estimated seasonality and trend

started beginning of 1781. We restrict our analysis to the period of 1879–2008, because in 1879 observations started being measured with new instruments under the guidance of the Munich Meteorological Central Station and thus the time series is homogenous. Due to its location on top of a mountain, summer temperatures are 2 °C to 3 °C lower than in the surrounding lowlands, whereas winter inversion layers lead to higher temperatures than in the valleys. The absolute maximum so far was recorded on July 29th in 1947 with 33.8 °C. Figure 7 displays the first (1879–1888) and last decade (1999–2008) of monthly temperature maxima.

It is one of the most demanding problems in environmental statistics to deal with trend and seasonality in data. When we are interested in the development of extreme events, we have to specify the event we want to study. In environmental statistics a usual measure of extremes is the return period as defined in (2.5). We could investigate the return periods of extremes in each month, January to December. Then we could answer, for instance, whether extreme temperatures in winter or summer have changed. Alternatively we could investigate the difference to a long-term mean or some other quantities, which describe extreme events.

In the present paper we will concentrate on a possible long-term trend in high temperatures at the station Hohenpeißenberg. Consequently, our analysis will be based on yearly maxima (see Fig. 8), which we assume to be GEV distributed (in Fig. 9 we shall see that this assumption is justified). Recall, however, that based on the IPCC 2007 report a 130 year temperature time series cannot be regarded as stationary. Thus, we want to incorporate some time-dependence into our model, i.e. a linear warming trend, although we know that there was not a uniform increase in mean temperature, but two periods with particular warming during approximately 1900–1945 and 1975–today.

We will investigate two possibilities to introduce non-stationarity into the model. Recall that classical time series theory (e.g. Brockwell and Davis [2]) suggests for a time series $Y_1, Y_2, \ldots$ either an arithmetic model of the form $Y_t = \Lambda_t + X_t$ or a multiplicative model $Y_t = \Lambda_t X_t$ for $t = 1, 2, \ldots$, where $\Lambda_1, \Lambda_2, \ldots$ models a non-stationary deterministic effect like drift and seasonality, and $X_1, X_2, \ldots$ is a stationary process. If $X_1, X_2, \ldots$ are identically GEV distributed, then we see immediately

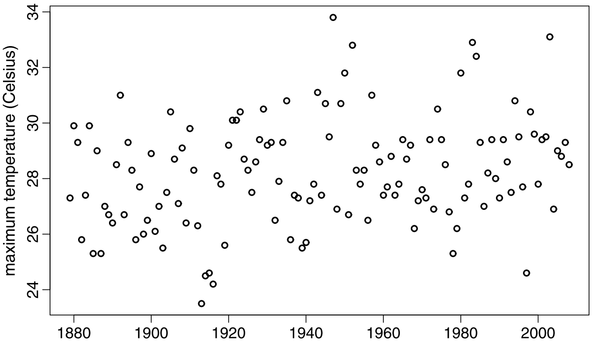


Fig. 8 Maximum yearly temperature over 130 years of data. The highest temperature has been measured in 1947

that $\Lambda_1, \Lambda_2, \ldots$ affect either the location parameter μ (for the arithmetic model) or the scaling parameter σ (for the multiplicative model) of the GEV distribution of $Y_1, Y_2, \ldots$. But the shape parameter γ remains the same under these deterministic location and scale changes. For simplicity, we introduce a linear trend into the location and scale parameter of the yearly maximal temperatures; i.e. we assume that the yearly maximal temperature $Y_1, \ldots, Y_{130}$ are an independent sequence with

$$Y_t \sim G_{\gamma,\sigma(t),\mu(t)} \quad \text{for } t = 1, \ldots, 130,$$

where $\mu(t) = \mu + at$ and $\sigma(t) = \sigma + bt$. Consequently, we will estimate by maximum likelihood estimation and compare the following models:

(1) Model 1: $\mu(t) = \mu$ and $\sigma(t) = \sigma$,
(2) Model 2: $\mu(t) = \mu + at$ and $\sigma(t) = \sigma$,
(3) Model 3: $\mu(t) = \mu$ and $\sigma(t) = \sigma + bt$,
(4) Model 4: $\mu(t) = \mu + at$ and $\sigma(t) = \sigma + bt$.

The estimation results are presented in Table 1.

For a comparison of the four different models, we notice that the negative log-likelihoods indicate already that Models 2 and 4 are better than Models 1 and 3, respectively. Although Model 1 is a special case of Model 3, the likelihood of Model 1 is nearly the same as the likelihood of Model 3. We guess already that the trend in the scale parameter may not be statistically significant, which is indeed true; the fluctuations do not significantly change over time. We have applied *likelihood ratio tests* to all nested pairs of models. Our model pairs are nested, when some of our parameters (a or b) may be zero or not. For details we refer to Coles [3], Sect. 2.6.6.

The tests compare (as we have already done informally) the likelihoods of two models. Rejection is now determined by asymptotic theory. More precisely, assume two (nested) models, say (I) and (II), with parameter $\theta^{(1)} \in \mathbb{R}^{d-k}$ for $k < d$, in model (I) and $\theta^{(2)} = (\theta_1^{(2)}, \theta_2^{(2)}) \in \mathbb{R}^d$ (where $\theta_1^{(2)} \in \mathbb{R}^k, \theta_2^{(2)} \in \mathbb{R}^{d-k}$) in model ($II$) with maximum likelihood estimators $\hat{\theta}^{(1)}$ and $\hat{\theta}^{(2)}$. Then, under some regularity conditions for the maximum likelihood functions $L_1(\hat{\theta}^{(1)})$ and $L_2(\hat{\theta}^{(2)})$, it can be shown that the quantity $-2(\log L_1(\hat{\theta}^{(1)}) - \log L_2(\hat{\theta}^{(2)}))$ is asymptotically χ_k^2-distributed. We present the results of 3 of our tests:

Table 1 Maximum likelihood estimators for μ, a, σ, b, and γ with standard errors in brackets below. The negative log-likelihood corresponding to the estimated models is given in the right-hand column

Parameters	μ	a	σ	b	γ	$-\log L$
Model 1	27.49671 (0.17721)	–	1.84122 (0.12203)	–	−0.20125 (0.05070)	268.9776
Model 2	26.65174 (0.32672)	0.01320 (0.00426)	1.76802 (0.11814)	–	−0.19624 (0.05253)	264.3865
Model 3	27.21659 (0.18851)	–	1.70919 (0.23720)	0.00199 (0.00377)	−0.18065 (0.06075)	268.9581
Model 4	26.65110 (0.32730)	0.01321 (0.00426)	1.77117 (0.22692)	−0.00005 (0.00301)	−0.19605 (0.05332)	264.3863

- Model 1 against Model 2: $H_0 : a = 0$ versus $H_1 : a \neq 0$

$$\begin{aligned}&-2\big(\log L_1\big(\widehat{\mu}^{(1)}, \widehat{\sigma}^{(1)}, \widehat{\gamma}^{(1)}\big) - \log L_2\big(\widehat{\mu}^{(2)}, \widehat{a}^{(2)}, \widehat{\sigma}^{(2)}, \widehat{\gamma}^{(2)}\big)\big)\\ &\quad = 9.1823 > 3.8415 = \chi_1^2(0.95),\end{aligned}$$

  i.e. we reject H_0 (p-value $= 0.002444$).
- Model 1 against Model 4: $H_0 : a = b = 0$ versus $H_1 : a \neq 0$ or $b \neq 0$

$$\begin{aligned}&-2\big(\log L_1\big(\widehat{\mu}^{(1)}, \widehat{\sigma}^{(1)}, \widehat{\gamma}^{(1)}\big) - \log L_4\big(\widehat{\mu}^{(4)}, \widehat{a}^{(4)}, \widehat{\sigma}^{(4)}, \widehat{b}^{(4)}, \widehat{\gamma}^{(4)}\big)\big)\\ &\quad = 9.1826 > 5.9915 = \chi_2^2(0.95),\end{aligned}$$

  i.e. we reject H_0 (p-value $= 0.0104$).
- Model 2 against Model 4: $H_0 : b = 0$ versus $H_1 : b \neq 0$

$$\begin{aligned}&-2\big(\log L_2\big(\widehat{\mu}^{(2)}, \widehat{a}^{(2)}, \widehat{\sigma}^{(2)}, \widehat{\gamma}^{(2)}\big) - \log L_4\big(\widehat{\mu}^{(4)}, \widehat{a}^{(4)}, \widehat{\sigma}^{(4)}, \widehat{b}^{(4)}, \widehat{\gamma}^{(4)}\big)\big)\\ &\quad = 3 \times 10^{-4} < 3.8415 = \chi_1^2(0.95),\end{aligned}$$

  i.e. we do not reject H_0 (p-value $= 0.986983$).

The p-value is an indicator of significance: the p-value of 0.002444 as calculated in the first test ensures that we can reject H_0 for all significance levels larger than this value. So the smaller the p-value, the more justified is a rejection of H_0. The comparison shows that a trend in the location parameter of the GEV model is significant but not the trend in the scale parameter. Model 4 gives no improvement to Model 2. Hence, again with support by statistical theory we conclude that the best model is Model 2, and there is no significant difference between Models 2 and 4, justifying the choice for Model 2.

In order to assess the model fit graphically, we will use a Gumbel probability plot (based on the GEV $G_{0,1,0}$) (PP-plot) and a Gumbel quantile-quantile plot (QQ-plot) for our transformed data set. Therefore, we show that any $G_{\gamma,\sigma(t),\mu(t)}$ distributed random variable Y_t with $\gamma < 0$ (the relevant regime for the temperature example is

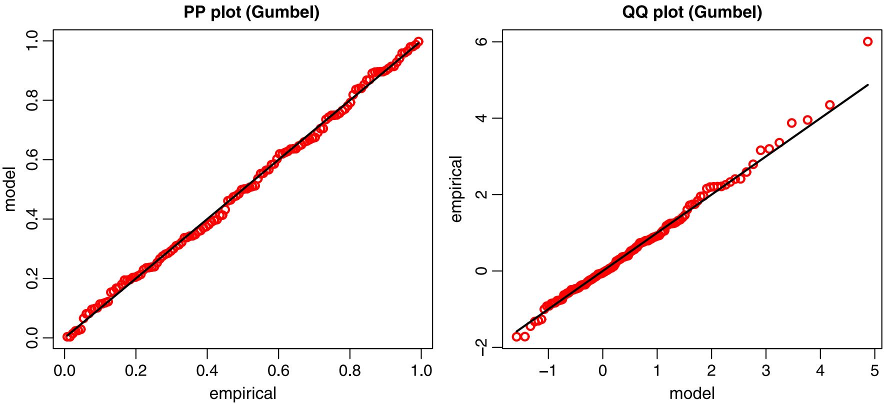


Fig. 9 The linear location Model 2 transformed to standard Gumbel: *PP*-plot and *QQ*-plot

a Weibull GEV distribution) can be transformed to a Gumbel random variable as follows. Afterwards we can use standard software for the plots.

We define

$$Z_t = \frac{1}{\gamma} \ln\left(1 + \gamma \frac{(Y_t - \mu(t))}{\sigma(t)} \right),$$

and prove below that indeed Z_t is standard Gumbel distributed. Note first that the Gumbel distribution has support on the whole of $\mathbb{R}$, whereas the Weibull distribution $G_{\gamma,\sigma(t),\mu(t)}$ has support $(-\infty, \mu(t) - \sigma(t)/\gamma]$; i.e. $G_{\gamma,\sigma(t),\mu(t)}(x) = 1$ for all $x > \mu(t) - \sigma(t)/\gamma$. Then $1 + \gamma(Y_t - \mu(t))/\sigma(t) > 0$ and, hence, Z_t has full support $\mathbb{R}$. Now we calculate

$$\begin{aligned}\mathbb{P}(Z_t \le x) &= \mathbb{P}\left(\frac{1}{\gamma} \ln\left(1 + \gamma \frac{(Y_t - \mu(t))}{\sigma(t)} \right) \le x \right) \\ &= \mathbb{P}\left(Y_t \le \frac{\sigma(t)}{\gamma} \left(e^{\gamma x} - 1 \right) + \mu(t) \right) = e^{-e^{-x}} \quad \text{for } x \in \mathbb{R}.\end{aligned}$$

This means that, provided $Y_1, Y_2, \ldots$ are independent Weibull distributed random variables, then $Z_1, Z_2, \ldots$ are independent Gumbel distributed random variables. Consequently, once we have estimated $\mu(t)$, $\sigma(t)$ and γ, we transform our data Y_t to

$$\widehat{Z}_t := \frac{1}{\widehat{\gamma}} \ln\left(1 + \widehat{\gamma} \frac{(Y_t - \widehat{\mu}(t))}{\widehat{\sigma}(t)} \right),$$

which should be close to a Gumbel distribution, provided the data are indeed Weibull GEV distributed with the estimated parameters. Figure 9 assesses the distribution fit by a *PP*-plot and a *QQ*-plot for the estimated parameters of Model 2 with linear location parameter. In the first plot, the *PP*-plot, the empirical distribution of $\widehat{Z}_1, \ldots, \widehat{Z}_{130}$ is plotted against the Gumbel distribution. In the second plot, the

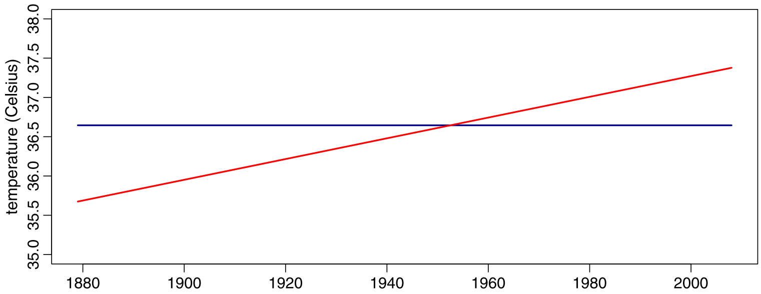


Fig. 10 Estimated right endpoint of the GEV distribution of Model 1 (*blue line*) and the linear trend Model 2 (*red line*). For Model 1 we estimate the constant right endpoint of 36.645 °C. For Model 2 the right endpoint starts at 35.674 °C and ends at 37.377 °C

QQ-plot, the quantiles of the Gumbel distribution are plotted against the empirical quantiles of $\widehat{Z}_1, \ldots, \widehat{Z}_{130}$. Both look very convincing, since they follow a 45° line confirming again Model 2.

For Model 2 we estimated the asymptotic 95 % confidence interval for γ. Let $z_{1-\alpha/2}$ be the $1-\alpha/2$-quantile of the normal distribution and $\widehat{s}_\gamma$ be the estimated standard deviation of $\widehat{\gamma}$. Then by classical likelihood theory (see Smith [10]), at least for $\gamma < 1/2$,

$$(\hat{\gamma} - z_{1-\alpha/2}\,\widehat{s}_\gamma, \hat{\gamma} + z_{1-\alpha/2}\,\widehat{s}_\gamma)$$

denotes the asymptotic $(1-\alpha) \times 100$ % confidence interval for γ. In Model 2 this results in the 95 % confidence interval $(-0.29972, -0.06158)$ for γ. As mentioned after the Fisher-Tippet Theorem 3.2 a negative γ indicates a Weibull distribution with finite right endpoint, meaning that there should be a limit of extreme maximum temperatures, which is not exceeded. Similarly, we obtain for a the 95 % confidence interval

$$(\widehat{a} - z_{1-\alpha/2}\,\widehat{s}_a, \widehat{a} + z_{1-\alpha/2}\,\widehat{s}_a) = (0.0048504, 0.0215496),$$

which reflects that a is positive; we have a statistically significant increase in the location parameter and a trend in the extremal temperatures.

The right endpoint of the Weibull distribution is given by $\mu(t) - \sigma(t)/\gamma$ (representing the maximum yearly temperature), which we can also estimate after having estimated the parameters. Figure 10 visualizes the constant endpoints of Model 1, where we have assumed fixed parameters over the whole time period, and the increase of the endpoint for the linear trend Model 2 caused by the linearity in the location parameter.

From this analysis presented in Fig. 11 we see that the return levels of high temperatures have increased considerably over the last 130 years. This increase is due to an increase of the location parameter of the extreme temperatures, the levels of the return periods have increased. The estimated parameters suggest an increase of $at = 0.01320t = 0.01320 \times 130 = 1.716$ °C over 130 years, corresponding to an

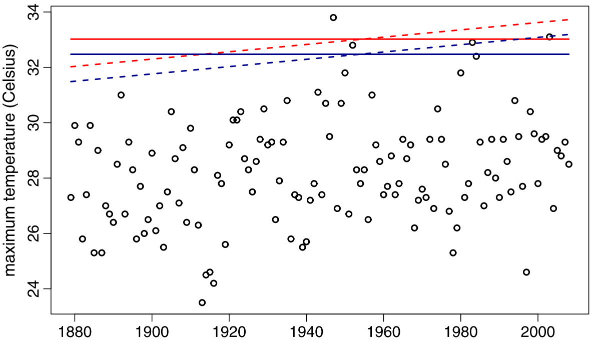


Fig. 11 The *red lines* show the estimated 100-year return level (which is the 99 % quantile), where the *straight line* is based on Model 1 and the *dashed line* on Model 2. Similarly the *blue lines* show the estimated 50-year return level based on Model 1 and Model 2, respectively

increase of 1.32 °C over a century. In contrast, simple least square linear regressions reveal increases in daily mean temperature of 1.472 °C and in daily maximum temperature of 1.515 °C over 130 years at the climate station Hohenpeißenberg, corresponding to an increase of 1.13 °C for the mean and of 1.17 °C for the daily maximum temperature over a century. Compared to these naive estimators the more realistic assessment by EVT methods yields a considerably higher prediction for the daily maximum temperatures in the future.

Prediction could now be based on this analysis. If we believe that the linear trend remains the same over the next 10 years, then we would estimate the value of $37.377 + 0.132 = 37.5090$ for the maximal yearly temperature in 2018. Note however, that such a fixed number is very unlikely. A confidence interval would be needed to give some idea about the statistical variability. By our estimation method we have been able to calculate confidence intervals for every single parameter estimate. However, for a confidence interval of the prediction we would need the whole distribution, which involves all three parameters, and their estimates are dependent. So besides standard errors (based on the estimated variance of the maximum likelihood estimators) also the asymptotic correlations between parameter estimates enter. Such theory, however, goes beyond this introductory paper, and gives rather food for thought.

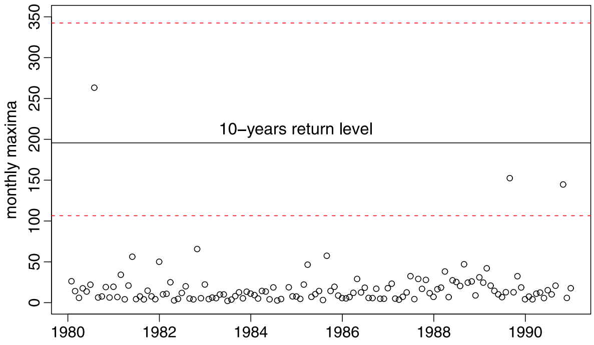
Apart from this statistical discussion, there is also some doubt on the assumption that future maximum temperatures increase with the same linear drift as the past ones. This also depends on political measures being taken against the threatening climate change.

4.2 The Blocks Method from Scratch

In the previous section we have simply started with maximum yearly temperatures over 130 years, and fitted an extreme value distribution to these data. This model choice was first based on the Fisher-Tippet Theorem 3.2, and later justified by a *PP*-plot and a *QQ*-plot depicted in Fig. 9.

As the name *blocks method* suggests the idea behind it is to divide the data $X_1, X_2, \dots, X_{nm}$ into m blocks of roughly the same length n and consider the block

Fig. 12 The largest claims of a Danish fire insurance per month. The *black line* is the 10-year return level, and the *dashed red lines* indicate the 95 %-confidence interval



maxima, i.e. we define $M_{n,j} = \max(X_{(j-1)n+1}, \dots, X_{jn})$ for $j = 1, \dots, m$. Recall that on the one hand we want to choose the blocks so small that we get as many block maxima as possible, on the other hand we have to choose them large enough so that we can assume that block maxima follow an extreme value distribution and also that they are independent.

Illustration 4.2 (10-Year Return Period for Danish Fire Data) For the daily losses of the fire insurance portfolio over m months $X_1, X_2, \dots, X_{nm}$ (i.e., X_k is the loss at the kth day), we determine the maximum losses within a month, respectively. These monthly block sizes are roughly equal, more precisely, n is between 28 and 31 days, and $M_{n,j}$ is the maximum loss during the jth month. As a first ansatz, according to the Fisher-Tippett Theorem 3.2, we exploit the fact that the distribution of $M_{n,j}$ can be approximated by a GEV distribution, so that

$$\mathbb{P}(M_{n,j} \le u) \approx G_{\gamma,\sigma,\mu}(u),$$

where γ, σ, μ are parameters, which have to be estimated, and the constants a_n and b_n are integrated in σ and μ. We denote by $\widehat{\gamma}, \widehat{\sigma}, \widehat{\mu}$ the respective estimators. Then we approximate

$$\mathbb{P}(M_{n,j} \le u) \approx G_{\widehat{\gamma},\widehat{\sigma},\widehat{\mu}}(u).$$

The level of the 10-year return period of the largest monthly claim, which happens in mean only once in 10 years can be estimated by means of (2.5). Since also $q = 1/(10 \times 12)$ holds, we obtain

$$\widehat{u} = \widehat{x}_{1-q} = G^{-1}_{\widehat{\gamma},\widehat{\sigma},\widehat{\mu}}\big(1 - (10 \times 12)^{-1}\big). \tag{4.2}$$

For the Danish fire data as depicted in Fig. 4 we estimate 195.7 million Danish Krone as level for extreme monthly claims, which happen in mean every 10 years (see Fig. 12).

4.3 The POT Method

It has been argued that applying the blocks method to data has the drawback of disregarding data, which may contribute information to the statistics of extreme values. Moreover, the blocks method can easily be applied to yearly, monthly or to other blocks-structured data, but what to do, if this is not the case. The *Peaks-Over-Threshold* (POT) method presents a valuable alternative.

The following section is dedicated to the POT method for a sample $X_1, \dots, X_n$, where we assume for the distribution function F that $F(x) = \mathbb{P}(X \le x) < 1$ for $x > 0$. We define further for a high threshold u

$$\overline{F}_u(y) := \mathbb{P}(X - u > y \mid X > u) = \frac{\overline{F}(u+y)}{\overline{F}(u)} \quad \text{for } y \ge 0.$$

Consequently, we obtain

$$\overline{F}(u+y) = \overline{F}(u)\overline{F}_u(y) \quad \text{for } y \ge 0. \tag{4.3}$$

How can we use these identities now to estimate tails and quantiles?

If now N_u denotes the number of all $k \in \{1, \dots, n\}$ satisfying $X_k > u$ given by

$$N_u = \# \left\{k \in \{1, \dots, n\} : X_k > u\right\},$$

then we denote by $Y_1, \dots, Y_{N_u}$ the excesses of $X_1, \dots, X_n$, i.e. the heights of the exceedances of u (cf. Fig. 6). We obtain an estimator for the tail (for values larger than u) by estimating both tails on the right hand side of (4.3). We estimate $\overline{F}(u)$ by the relative frequency

$$\widehat{\overline{F}(u)} = \frac{N_u}{n} \tag{4.4}$$

and approximate $\overline{F}_u(y)$ by the Generalized Pareto Distribution (GPD) of (3.4), where the scale parameter $\sigma(u)$ has to be considered. It is integrated as parameter $\sigma(u)$ into the limit distribution such that

$$\overline{F}_u(y) \approx \left(1 + \gamma \frac{y}{\sigma(u)}\right)^{-1/\gamma} \quad \text{for } y \ge 0, \tag{4.5}$$

where γ and $\sigma(u)$ have to be estimated by some estimators denoted by $\widehat{\gamma}$ and $\widehat{\sigma}(u)$. From (4.3)–(4.5) we obtain a tail estimator of the form

$$\widehat{\overline{F}(u+y)} = \frac{N_u}{n}\left(1 + \widehat{\gamma} \frac{y}{\widehat{\sigma}(u)}\right)^{-1/\widehat{\gamma}} \quad \text{for } y \ge 0. \tag{4.6}$$

Then for given $p \in (0, 1)$ we obtain an estimator $\widehat{x}_p$ for the p-quantile x_p taken from (2.1) by solving the equation

$$1 - p = \frac{N_u}{n}\left(1 + \widehat{\gamma} \frac{\widehat{x}_p - u}{\widehat{\sigma}(u)}\right)^{-1/\widehat{\gamma}}.$$

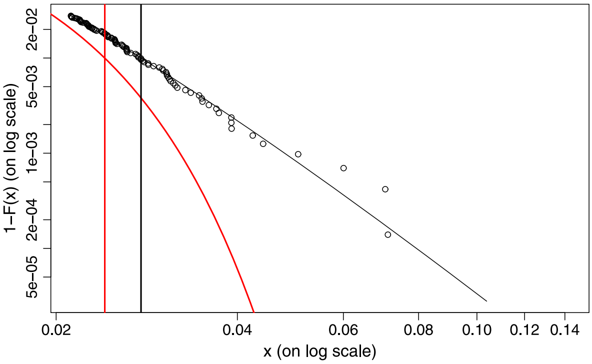


Fig. 13 Estimated tail of the daily losses of the S&P500. The *black curve* shows the tail estimated by the POT method with threshold $u = 0.0212$, $\widehat{\gamma} = 0.193$, $\widehat{\sigma} = 0.00575$ and the *red line* shows the distribution tail estimated under the assumption of a normal distribution for the daily losses. The *vertical black line* indicates the logarithmic $\mathrm{VaR}_{0.99}^{\mathrm{POT}}(X) = 0.028$ estimated by the POT method and the *vertical red line* shows the logarithmic $\mathrm{VaR}_{0.99}^{\mathrm{norm}}(X) = 0.024$ estimated from a normal distribution

This gives

$$\widehat{x}_p = u + \frac{\widehat{\sigma}(u)}{\widehat{\gamma}}\left(\left(\frac{n}{N_u}(1-p)\right)^{-\widehat{\gamma}} - 1\right). \tag{4.7}$$

Illustration 4.3 (Tail and Quantile Estimation) We apply the POT method to the S&P500 loss data using the tail estimate from (4.6) and, for comparison, we also fitted a normal distribution to the data by estimating mean and variance by their empirical versions. Figure 13 depicts both tail estimates in logarithmic scale for a threshold $u = 0.0212$ and $y > 1$. Moreover, $\mathrm{VaR}_{0.99}^{\mathrm{POT}}(X)$ was estimated for the daily losses using the POT estimator (4.7) as well as the normal estimator $\mathrm{VaR}_{0.99}^{\mathrm{norm}}(X) = \widehat{\mu} + \widehat{\sigma} z_{0.99}$, where $z_{0.99}$ is the 0.99-quantile of the normal distribution. Plotted are again the logarithmic quantities; i.e. $\log \mathrm{VaR}_{0.99}^{\mathrm{POT}}(X) = 0.028$ and $\log \mathrm{VaR}_{0.99}^{\mathrm{norm}}(X) = 0.024$, which correspond to $\mathrm{VaR}_{0.99}^{\mathrm{POT}}(X) = 2.795$ and $\mathrm{VaR}_{0.99}^{\mathrm{norm}}(X) = 2.784$; the difference of 0.011 does not look too substantial, but recall that our data are relative losses (i.e. percentage points). Moreover, the standardized S&P500 portfolio value compares only to a standardized bank portfolio, so has to be multiplied by millions to obtain a realistic value.

We clearly see that the normal distribution tail is completely inadequate to estimate the tail of the daily losses of the S&P500. The data are far above its normal tail estimate. Usage of the normal distribution underestimates the risk considerably and yields a completely inadequate risk capital.

In Illustration 4.3 we have estimated the tail and the $\mathrm{VaR}_{0.99}(X)$ for the S&P500 losses and depicted in Fig. 13. The estimation was based on the assumption that the losses (or at least the excesses) are i.i.d. However, modelling of financial data goes

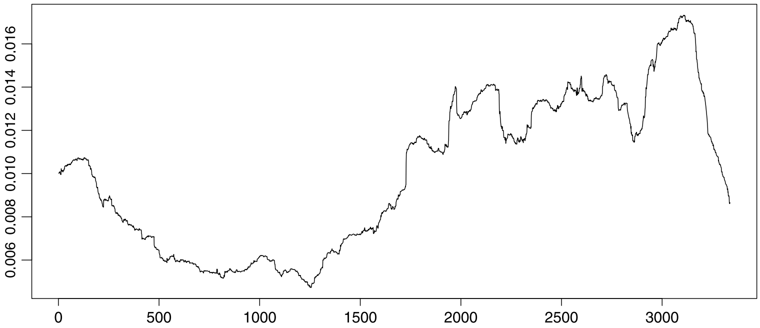


Fig. 14 The empirical standard deviations of the daily losses of the S&P500 during 1991–2004 with estimators based on the previous 250 days, respectively

far beyond marginal distributions. It has been a relevant research area for decades, and we conclude with some facts and references.

Remark 4.4 (i) Dependence between portfolio components are in the normal model given by correlations, which only model linear dependence. Market risk portfolios, however, consist of such different assets as shares, options, and more complex derivatives, which are known to be non-linearly dependent. It is of high importance to have a comprehensive understanding of the influence of the portfolio components to the portfolio loss. Dependence modeling and different dependence measures are discussed in Chap. 9, [37].

(ii) Already from the daily losses depicted in the right plot of Fig. 2 it is clear that the data vary considerably in their structure. We see immediately that a period of low volatility is followed by a period of high volatility (the standard deviation is called volatility in banking jargon). It is certainly not obvious that all observations can be modelled with the same distribution. Recall that (2.3) requires daily estimates based on past year's observations. Figure 14 shows the running empirical estimates of the volatility σ of the daily losses of the S&P500 based on observations of the past one year, respectively. This simple window estimate shows clearly the time-varying volatility, which is typical for most financial time series.

(iii) Until now we have not touched the important questions of time dependence within the time series of daily returns. Financial data show an interesting dependence structure; although most daily returns are uncorrelated, the data do not originate from independent observations. As seen in Fig. 15 the sample autocorrelation function of the daily losses of the S&P500 is almost 0 for all lags, whereas the sample autocorrelation function of the squared returns is substantial, contradicting the independence assumption. The most prominent financial time series model is the GARCH (Generalized AutoRegressive Conditional Heteroskedasticity) model. Volatility is modelled as a stochastic process and can capture a dependence structure as seen in Fig. 15. An excellent overview on discrete-time and continuous-time

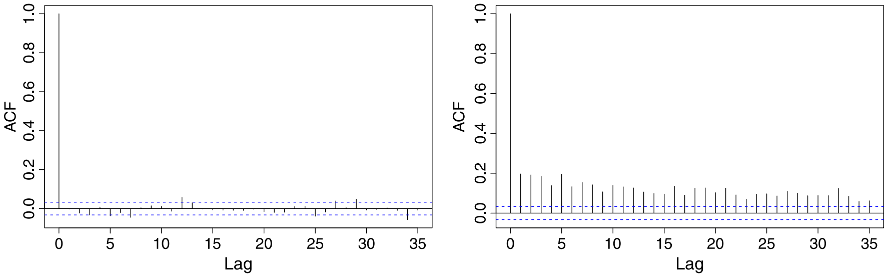


Fig. 15 Sample autocorrelation function of the daily losses (*left*) and the squared daily losses (*right*) of the S&P500

stochastic volatility models in one and multivariate dimensions is the book edited by [12].

5 Food for Thought

Extreme value theory has gone a long way since its beginnings with Fisher and Tippett in 1928. New applications, unheard-of in the 1920-ies have emerged. Climate change, large insurance claims and extreme financial risk are just three of them. Extreme value theory has found its way also into the areas of technical safety and reliability theory, as well as the statistical assessment of environmental quantities like temperatures, floods, droughts, earthquakes and storms.

Concerning the statistical methods we have presented, we want to emphasize the following. In our statistical analyses we have assumed that data are independent and have the same distribution (perhaps enriched by a linear trend, which can easily be implemented). This assumption is often unrealistic. As reported in Remark 4.4 financial time series exhibit in general a very complex dependence structure; for the S&P500 see Figs. 14 and 15. Many data, also insurance claims, are affected by seasonal effects or exhibit some clusters of claim events. Such effects can influence estimation and prediction procedures considerably. Moreover, the one-dimensional case treated above is rather unrealistic. Portfolios of market risks are composed of many components (often several hundreds), and it may be interesting to understand the dependence in the combination of extreme risks. Moreover, risks are often influenced by some latent variables, whose influence would have to be assessed as well. Such problems are hot research topics at the moment and require still a considerable amount of theoretical and practical work.

Extreme value theory has been extended to multivariate data, which is rather demanding, since there exists no finite parameterizations as in the one-dimensional case as seen in the Fisher-Tippett Theorem 3.2. The dependence between different components of a vector is modeled by an integral with respect to some measure and Poisson random measures provide a very powerful tool to deal with such problems; cf. Resnick [41].

Moreover, extreme value theory for time series with marginals ranging from Gaussian to heavy-tailed ones is still a lively research area. The usual picture is that for light-tailed time series models one can more or less ignore the dependence structure, whereas for heavy-tailed models the dependence creeps into the extreme tails (events) by leading to clusters of extremes; cf. Fasen [28], Fasen, Klüppelberg, and Schlather [30].

More recently, also spatial and space-time extreme value models have come into focus in particular for environmental data like heavy rainfall or storms requiring special statistical methods; cf. Davis, Klüppelberg, and Steinkohl [25, 26] for details and further references.

For those interested in the state of the art of extreme value theory research, we recommend to consider the journal "Extremes" (http://www.springer.com/statistics/journal/10687), which is solely devoted to theory and applications of extreme values.

6 Summary

We hope that we have convinced our readers that extreme value theory and extreme value statistics offer an important theory and statistical estimation procedures to assess extreme risks in different applications areas.

We have presented the basic theory and also three estimation procedures to find the distribution and other quantities describing extreme events. The first one was to fit a GEV to extreme data, where we also took care of non-stationarity of the data either in the location parameter (linear trend) or in the scaling parameter (higher fluctuations). The second one was to use the block-maxima method for a sample where only the blocks maxima were distributed according to a GEV distribution. And finally, we introduced the POT method, which models high threshold exceedances.

As a result we obtained for our three examples:

- The climate change data exhibit a higher trend in the yearly maxima over the last century than the mean trend at the corresponding station. The Weibull distribution is the appropriate extreme value distribution, which shows that high temperature is bounded, although the maxima increase.
- Danish insurance claims, which are from a fire insurance portfolio, are very heavy-tailed data, and the model suggests that with a (non-negligible) positive probability the insurance company may experience a claim, which is easily twice as high as they have ever seen before.
- The daily losses of the S&P500 have a 99 % Value-at-Risk of 0.028 % when estimated by the POT method, while based on the normal distribution, it is only 0.024 %. While these numbers look small, in banking business one has to multiply them by millions of Euros, so that the difference becomes substantial. Since capital reserves have to be calculated built on such numbers, the banks are much happier about the smaller numbers coming from the Gaussian distribution.

References

Selected Bibliography

1. J. Beirlant, Y. Goegebeur, J. Segers, J. Teugels, D. De Waal, C. Ferro, *Statistics of Extremes: Theory and Applications* (Wiley, New York, 2004)
2. P. Brockwell, R.A. Davis, *Introduction to Time Series and Forecasting* (Springer, New York, 1996)
3. S.G. Coles, *An Introduction to Statistical Modeling of Extreme Values* (Springer, Berlin, 2001)
4. P. Embrechts, C. Klüppelberg, T. Mikosch, *Modelling Extremal Events for Insurance and Finance* (Springer, Berlin, 1997)
5. G.C. Hegerl, H. Hanlon, C. Beierkuhnlein, Elusive extremes. Nat. Geosci. **4**, 143–144 (2011)
6. IPCC 2007, Climate change 2007: the physical science basis, in *Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, ed. by S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor, H.L. Miller (Cambridge University Press, Cambridge, 2007)
7. A.J. McNeil, R. Frey, P. Embrechts, *Quantitative Risk Management: Concepts, Techniques, and Tools* (Princeton University Press, Princeton, 2005)
8. A. Menzel, H. Seifert, N. Estrella, Effects of recent warm and cold spells on European plant phenology. Int. J. Biometeorol. **55**, 921–932 (2011)
9. R.-D. Reiss, M. Thomas, *Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields*, 2nd edn. (Birkhäuser, Basel, 2001)
10. R.L. Smith, Estimating tails of probability distributions. Ann. Stat. **15**, 1174–1207 (1987)
11. A. Stephenson, E. Gilleland, Software for the analysis of extreme events: the current state and future directions. Extremes **8**, 87–109 (2005)

Additional Literature

12. T.G. Andersen, R.A. Davis, J.-P. Kreiss, T. Mikosch, *Handbook of Financial Time Series* (Springer, New York, 2009)
13. A.A. Balkema, L. de Haan, Residual life time at great age. Ann. Probab. **2**, 792–804 (1974)
14. K.F. Bannör, M. Scherer, Model risk and uncertainty—illustrated with examples from mathematical finance, in *Risk – A Multidisciplinary Introduction*, ed. by C. Klüppelberg, D. Straub, I. Welpe (2014)
15. Basel Committee on Banking Supervision, Basel III: a global regulatory framework for more resilient banks and banking systems (2011). Revised June 2011
16. Basel Committee on Banking Supervision, Revisions to the Basel II market risk framework (2011). Updated as of 31 December 2010
17. M. Beniston, H.F. Diaz, The 2003 heat wave as an example of summers in a greenhouse climate? Observations and climate model simulations for Basel, Switzerland. Glob. Planet. Change **44**, 73–81 (2004)
18. C. Bernhardt, C. Klüppelberg, T. Meyer-Brandis, Estimating high quantiles for electricity prices by stable linear models. J. Energy Mark. **1**, 3–19 (2008)
19. F. Biagini, T. Meyer-Brandis, G. Svindland, The mathematical concept of measuring risk, in *Risk – A Multidisciplinary Introduction*, ed. by C. Klüppelberg, D. Straub, I. Welpe (2014)
20. K. Böcker, C. Klüppelberg, Operational VaR: a closed-form approximation. Risk **12**, 90–93 (2005)
21. E. Brodin, C. Klüppelberg, Extreme value theory in finance, in *Encyclopedia of Quantitative Risk Assessment*, ed. by B. Everitt, E. Melnick (Wiley, Chichester, 2007)

22. S.J. Brown, J. Caesar, C.A.T. Ferro, Global change in extreme daily temperature since 1950. J. Geophys. Res. **113**, D05115 (2008)
23. R. Cont, Empirical properties of asset returns: stylized facts and statistical issues. Quant. Finance **1**, 223–236 (2001)
24. J. Danielsson, P. Embrechts, C. Goodhart, C. Keating, F. Muennich, O. Renault, H.S. Shin, An Academic Response to Basel II. Financial Markets Group, London School of Economics (2001)
25. R.A. Davis, C. Klüppelberg, C. Steinkohl, Max-stable processes for modelling extremes observed in space and time. J. Korean Stat. Soc. **42**(3), 399–414 (2013)
26. R.A. Davis, C. Klüppelberg, C. Steinkohl, Statistical inference for max-stable processes in space and time. J. R. Stat. Soc., Ser. B **75**(5), 791–819 (2013)
27. S. Emmer, C. Klüppelberg, M. Trüstedt, VaR – ein Mass für das extreme Risiko. Solutions **2**, 53–63 (1998)
28. V. Fasen, Extremes of continuous-time processes, in *Handbook of Financial Time Series*, ed. by T.G. Andersen, R.A. Davis, J.-P. Kreiss, T. Mikosch (Springer, Berlin, 2009), pp. 653–667
29. V. Fasen, C. Klüppelberg, Modellieren und Quantifizieren von extremen Risiken, in *Facettenreiche Mathematik*, ed. by K. Wendland, A. Werner (Vieweg/Teubner, Wiesbaden, 2011)
30. V. Fasen, C. Klüppelberg, M. Schlather, High-level dependence in time series models. Extremes **13**, 1–33 (2010)
31. R.A. Fisher, L.H.C. Tippett, Limiting forms of the frequency distribution of the largest or smallest member of a sample. Proc. Camb. Philos. Soc. **24**, 180–190 (1928)
32. C.F. Gauß, *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solum Ambientium* (Hamburgi Sumptibus F. Perthes et I.H. Besser, Hamburg, 1809)
33. P. Glasserman, *Monte Carlo Methods in Financial Engineering* (Springer, New York, 2004)
34. S. Haug, C. Klüppelberg, L. Peng, Statistical models and methods for dependent insurance data; invited discussion paper. J. Korean Stat. Soc. **40**, 125–139 (2011)
35. IPCC 1990, Climate change: the IPCC scientific assessment, in *Report Prepared for IPCC by Working Group 1*, ed. by J.T. Houghton, G.J. Jenkins, J.J. Ephraums (Cambridge University Press, Cambridge, 1990)
36. R.W. Katz, Statistics of extremes in climate change. Clim. Change **71–76**, 100 (2010)
37. C. Klüppelberg, R. Stelzer, Dealing with dependent risks, in *Risk – A Multidisciplinary Introduction*, ed. by C. Klüppelberg, D. Straub, I. Welpe (2014)
38. R. Korn, E. Korn, G. Kroisandt, *Monte Carlo Methods and Models in Finance and Insurance*. Financial Mathematics Series (Chapman & Hall/CRC Press, London/Boca Raton, 2010)
39. J. Pickands, Statistical inference using extreme order statistics. Ann. Stat. **3**, 119–131 (1975)
40. S. Poisson, *Recherches sur la Probabilité des Jugements en Matière Criminelle et en Matière Civile Précédées des Règles Générales du Calcul des Probabilités* (Bachelier, Imprimeur-Libraire, Paris, 1837)
41. S.I. Resnick, *Extreme Values, Regular Variation, and Point Processes* (Springer, New York, 1987)
42. R.L. Smith, Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone (with discussion). Stat. Sci. **4**, 367–393 (1989)
43. D.B. Stephenson, Definition, diagnosis, and origin of extreme weather and climate events, in *Climate Extremes and Society*, ed. by R. Murnane, H. Diaz (Cambridge University Press, Cambridge, 2008), pp. 11–23
44. G.-R. Walther, E. Post, P. Convey, A. Menzel, C. Parmesani, T.J. Beebee, J.-M. Fromentin, Ecological responses to recent climate change. Nature **416**, 389–395 (2002)

Chapter 7
Statistical Models for the Prediction of Genetic Values

Chris-Carolin Schön and Valentin Wimmer

Agricultural and medical genetics are currently revolutionized by the technological developments in genomic research. The genetic analysis of quantitatively inherited traits and the prediction of the genetic predisposition of individuals based on molecular data are rapidly evolving fields of research. We ask how phenotypic variation for a quantitative trait can be linked to genetic variation at the DNA level. Advances in high-throughput genotyping technologies return data on thousands of loci per individual. We present linear models to identify molecular markers significantly associated with quantitative traits. We discuss the drawbacks arising from a large number of predictor variables and a high degree of collinearity between them. We illustrate how linear mixed models can overcome the limitations through shrinkage and allow the prediction of genetic values inferred from genome-wide marker data. With a small example from maize breeding, we present how the models can be applied to predict the risk of genetically diverse individuals to be damaged by insects and why predictions based on whole-genome marker profiles are likely to be more accurate than those based on pedigree information. The choice of appropriate methods for quantitative genetic analyses based on high-throughput genomic data for medical and agricultural genetics is discussed.

Keywords Quantitative genetics · Genome-based prediction · Linear mixed models · Disease risk · Genetic value

Mathematics Subject Classification (2010) 62J05 · 62J07 · 62P10

C.-C. Schön (✉)
Chair of Plant Breeding, Center of Life and Food Sciences Weihenstephan, Technische Universität München, Liesel-Beckmann-Str. 2, 85354 Freising-Weihenstephan, Germany
e-mail: chris.schoen@tum.de

V. Wimmer
Plant Breeding, Center of Life and Food Sciences Weihenstephan, Technische Universität München, Liesel-Beckmann-Str. 2, 85354 Freising-Weihenstephan, Germany

C. Klüppelberg et al. (eds.), *Risk – A Multidisciplinary Introduction*,
DOI 10.1007/978-3-319-04486-6_7,

The Facts

- Agricultural and medical genetics are currently revolutionized by the technological developments in genomic research.
- Many quantitative traits are complex, i.e. they are controlled by a large number of genes. The genetic predisposition of individuals can be predicted based on their DNA marker profile.
- The major challenge in predicting the phenotype from the genotype is that the number of predictor variables p often exceeds the number of observations n.
- With $n < p$ and model selection the estimated marker effects are biased in multiple marker regression models.
- Linear mixed models provide a framework for simultaneously estimating marker effects even if $n \ll p$.

1 Introduction

Crops and livestock species have been genetically improved by breeders for more than 10,000 years. In the selection process, breeding populations are evaluated for traits such as productivity, quality or resistance against diseases and environmental stress, and only those individuals that meet specific requirements form the parents of the next generation. Because most traits of importance follow a continuous distribution, understanding and predicting quantitative genetic variation is crucial in agricultural genetics. The trait value or *phenotypic value* (P) of an individual is determined by the joint action of many genes and the environment. Thus, the phenotypic value of an individual can be expressed as the sum of its *genotypic value* (G) and an environmental deviation (E)

$$P = G + E.$$

This decomposition also holds when measuring the amount of variation in a population of individuals. Assuming independence of genotypic and environmental effects, the total phenotypic variance (σ_P^2) is given as the sum of the genotypic (σ_G^2) and the environmental (σ_E^2) variance components, thus

$$\sigma_P^2 = \sigma_G^2 + \sigma_E^2.$$

An important question in quantitative genetics is the relative contribution of the genotype and the environment to trait variation. For a given population, the heritable portion of trait variation can be quantified with the trait *heritability*, which is defined as

$$h^2 = \frac{\sigma_G^2}{\sigma_P^2} = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_E^2}. \tag{1.1}$$

The genetic variance can be further subdivided into the additive genetic variance component σ_A^2 resulting from the effects of individual alleles, the dominance genetic variance component σ_D^2 resulting from the interaction of two alleles of the same gene and the epistatic genetic variance component σ_I^2 resulting from the interaction of alleles of different genes [1]. It is the additive genetic variance that is most important in breeding because it is the major determinant for the resemblance between parent and offspring. Thus, for simplicity we will focus on the additive genetic variation and assume absence of dominance and epistasis throughout this chapter. For a more detailed description of the decomposition of the genotypic variance the reader is referred to textbooks on quantitative genetics [1, 5].

Linking phenotypic variation with genetic variation at the DNA level is crucial for understanding the inheritance of quantitative traits. The most abundant and also the simplest form of genetic variation at the DNA level is the exchange of a single nucleotide at a given position in the genome which is called single nucleotide polymorphism (SNP). Some SNPs are functional, i.e. their variation translates directly into phenotypic variation. Many SNPs are silent, i.e. they do not have a direct effect on the phenotype but they are still useful as genetic markers. If a silent and a functional SNP are both located on the same chromosome, the probability that a specific allele at a silent SNP is associated with a specific allele at the functional SNP increases with increasing proximity on the chromosome. This non-random association of alleles between SNPs is called linkage disequilibrium (LD). In a given population LD arises from a shared history of mutation and recombination. Given the non-random association of alleles between SNP markers and genes affecting a quantitative trait, variation at the SNPs will track the genetic variation at the unobservable quantitative trait loci (QTL). The extent and pattern of LD across the genome varies depending on the studied species and its breeding history. For many species, SNP arrays have been developed that can determine the genotype of an individual at tens or hundreds of thousands of nucleotide sites. With these tools at hand, it is possible to achieve good genome coverage and visualize variation at many loci with a high probability of the SNP marker being in high LD with one or more QTL affecting the trait of interest. A very good introduction to genetic analysis can be found in [2].

The choice of experimental design and statistical model employed in the genetic analysis of complex traits depends on the genetic architecture of the trait under study. Let our trait of interest be the risk of a plant to be susceptible to a disease or to insect damage. We assume this risk to be controlled by a few QTL with sizeable effects. First of all, we need to identify a population of individuals that genetically differ with respect to our trait of interest. For every individual of our population we analyze the SNP genotypes at many marker loci in the lab and collect phenotypic data on the risk of disease in the field or in a resistance test in the greenhouse. The aim of our study will be to identify significant associations between SNP markers and QTL, localize the QTL on the genome, characterize their molecular properties and gain insight into how the different alleles at the QTL contribute to phenotypic trait variation. With this knowledge we can predict the disease risk of another set of individuals as long as we know their SNP genotypes and we will select those individuals for which the risk of being affected by the disease is minimized. It will no longer be necessary to test the selection candidates in the field or in the greenhouse for their phenotypes. The prediction accuracy and also our selection success will depend on the statistical power of the experiment, in which we identified the marker-trait associations. If we were able to identify all or most of the QTL contributing to trait variation our prediction accuracy will be high. If many QTL remain undetected by the SNP markers prediction accuracy will be low. The statistical models for the identification of significant associations between SNP markers and QTL are introduced in Sect. 2.

Our assumption was that disease risk is regulated by a few genes with sizeable effects. However, we have ample evidence from marker-based studies that this might be rather the exception than the rule. In crop plants, many genes with small effects contribute to the genetic variation of traits like grain yield or flowering time [9, 18]. Similar results have been shown for livestock species and in human genetic studies [17, 20]. Thus, for these traits it may be more appropriate to predict the genetic value of an individual using many markers randomly dispersed across the genome instead of selected SNPs. This concept is based on the hypothesis that with a sufficiently high density of marker data all of the genetic polymorphisms contributing to trait variation are in high LD with the markers segregating in the studied population. Statistical methods for genome-wide prediction of the genetic merit of individuals from high-density, genome-wide marker data were suggested in a seminal paper [16]. Prediction models are developed based on large training populations for which genotypic and phenotypic data are available. However, with high-density SNP assays the number p of predictor variables, i.e. SNP markers, often exceeds the number n of observations (small n, large p). Furthermore, with SNP markers in LD multicollinearity is prevalent among predictor variables leading to an inflation of variance of their estimated effects [3]. In Sect. 3 we describe statistical models that can cope with this situation.

The remainder of this chapter is structured as follows. In Sect. 2, we introduce linear models. In Sect. 3 we give an overview of linear mixed models, a model class which is suitable for problems with $n \ll p$. In Sect. 4, we present how the genetic value of an individual can be predicted based on information from its relatives. Section 5 compares the different models with respect to their assumptions and pitfalls. All models are illustrated by a simulated data example. We also give an outlook how genome-wide marker information can be used to predict the genetic predisposition in human genetics.

2 Linear Models

In quantitative genetic analyses we use *linear models* to investigate the relationship of an observed response or dependent variable Y (phenotype) and a predictor or independent variable X (SNP marker) that is observed along with Y. Note, that we use the term linear, because the model is linear in the model parameters and not necessarily in X. We can fit a simple linear regression model to find out, if a given SNP marker is significantly associated with variation in the response variable and how much of that variation can be explained by the predictor variable. The SNP markers are generally biallelic and we can observe two alleles (A and a) in the population. The rare allele is called the *minor allele* as opposed to the *major allele*. In a diploid species each individual carries two alleles which yields three different genotypes for a given SNP marker (AA, Aa, aa). If both alleles are identical by state, the individual is called *homozygous* and *heterozygous* otherwise. See [2] for more details on the genetic nomenclature. The marker genotype of an individual is coded by the number of copies of the minor allele, i.e. 0, 1, and 2. Thus, the slope of the regression line estimates the additive effect of one additional copy of the minor allele.

Let us assume we observe data on X and Y for n independent individuals, i.e. tuples $(x_1, y_1), \dots, (x_i, y_i), \dots, (x_n, y_n)$. With the simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, \dots, n, \tag{2.1}$$

we estimate two unknown constants, the intercept β_0 and the slope β_1 of the regression line. The residual term e_i represents the part of the data which is not explained by the model, i.e. the random deviation of each (x_i, y_i) from the regression line. Inference in a linear model is based on the assumption of normally distributed, uncorrelated residuals with mean 0 and variance σ^2, i.e. $e_i \sim \mathrm{N}(0, \sigma^2)$. Estimates of the regression coefficients β_0 and β_1 can be obtained using the least squares estimation procedure. This method minimizes the sum of squares of the estimated residuals

$$\mathrm{SQR} = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right]^2$$

with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$. The least squares estimates for the intercept and the slope are

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}, \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}. \end{aligned}$$

To decide if the regressor variable X has a statistically significant influence on the response variable, hypothesis testing of the slope of the regression β_1 is necessary. We formulate the following hypotheses

$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_1 : \beta_1 \neq 0. \tag{2.2}$$

In an analysis of variance (ANOVA) the total sum of squares (SQT) of the response variable Y is partitioned into a component explained by the model (SQE) and the residual sum of squares (SQR) which is not explained by the model

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{\text{SQT}}=\underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{\text{SQE}}+\underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{\text{SQR}}. \tag{2.3}$$

Under the normal distribution assumption on the e_i an F-statistic can be calculated from the ratio of the corresponding mean squares, i.e. $F = \text{MQE}/\text{MQR}$ with $\text{MQE} = \text{SQE}/1$ and $\text{MQR} = \text{SQR}/(n-2)$. If the F-statistic exceeds the tabulated value at the given significance level we reject H_0 and conclude that the SNP marker is significantly associated with variation in the response variable. Consequently we can use this marker for prediction of the genetic merit of our selection candidates. Prediction of a new individual's genetic merit based on the SNP marker genotype x_{n+1} is given by $\hat{y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 x_{n+1}$. The fit of the model can be judged from the coefficient of determination which is defined as $R^2 = \text{SQE}/\text{SQT}$. It quantifies the proportion of variability in the response variable Y that is explained by the statistical model. In the genetic literature this type of analysis is called *single marker regression* (SMR) or single marker ANOVA.

Illustration 2.1 (Resistance against the European corn borer) Let us now look at an example that illustrates the method of SMR. We want to investigate the susceptibility risk of eight maize individuals (I1–I8) to an insect called European corn borer (*Ostrinia nubilalis*) which is a major pest of maize, see Figs. 1A and 1B. We start with a simulation study using eight maize plants that are derived from different selection cycles of the breeding process. Some of them are related through common ancestors. We will have a closer look at their relatedness in Sect. 4. Our response variable $Y \in \mathbb{R}$ is the average tunnel length [cm] in the stalk of each individual caused by feeding of the larvae, see Fig. 1C.

We assume that the trait tunnel length is normally distributed and that the genetic variation for tunnel length is purely additive. Each of the eight plants is assigned a genotype (0, 1 or 2) at each of four biallelic SNP markers. SNP1 is assumed not to be associated with tunnel length. SNP2, SNP3 and SNP4 each have an effect on tunnel length. SNP2 has an effect of 1 cm, SNP3 and SNP4 of -4 cm and 4 cm. The sign indicates whether the trait increasing allele is associated with the minor or the major allele. We obtain phenotypic values for individuals I1–I6 by adding random environmental noise to the genotypic value from a normal distribution with mean 0 cm and variance $\sigma^2 = 4\ \text{cm}^2$. To avoid negative values, we add a constant to each simulated trait value ($c = 10$ cm), hence $e_i \sim \text{N}(10, 4)$. Let's assume that we have no phenotypic data for individuals I7 and I8. The resulting data comprising pedigrees (parent P1 × parent P2), phenotypic values, and marker genotypes for the eight individuals are given in Table 1. Of course, this is a simplifying example because in real life the true model will be much more complex, the number of markers will be much larger and will most often exceed the number of observations.

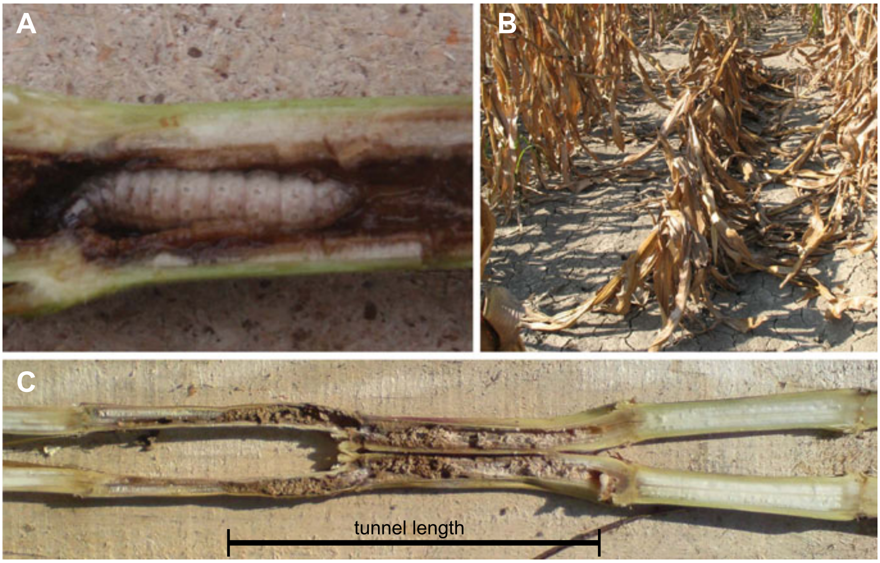


Fig. 1 (**A**) Larvae of the European corn borer in the stalk; (**B**) Maize plants damaged by stalk breakage through feeding of corn borer larvae; (**C**) Tunnel in stalk caused by feeding of the larvae

Table 1 Pedigree, phenotypic values, and marker genotypes for eight simulated maize individuals

Cycle	Individual	Pedigree	Tunnel length [cm]	SNP			
				1 (0)[a]	2 (1)	3 (−4)	4 (4)
1	I1	P1 × P2	13	2	2	0	1
1	I2	P3 × P4	17	0	0	0	1
1	I3	–	1	0	1	2	0
2	I4	I1 × I2	17	1	1	0	2
2	I5	I1 × I2	11	1	1	0	1
2	I6	I2 × I3	6	0	1	1	0
2	I7	I1 × I2	–	1	1	0	1
2	I8	I1 × I2	–	1	1	0	0

[a] Simulated SNP effects

Let us now estimate the genetic effects of the four SNPs with SMR based on data from individuals I1–I6. For each SNP ($j = 1, \ldots, 4$) we fit the following model

$$y_i = \beta_0 + \beta_1 x_{ij} + e_i, \quad i = 1, \ldots, 6,$$

where y_i denotes the observed phenotype (tunnel length in cm) of the ith individual and x_{ij} the marker genotype of the ith individual for the jth SNP. When fitting the model for SNP4 we obtain the least squares estimates $\hat{\beta}_0 = 4.7$ cm and $\hat{\beta}_1 = 7.4$ cm. The data and the regression line are visualized in Fig. 2.

Table 2 Summary of a single marker regression for each SNP

	β_0	β_1	SQE	SQR	R^2	F-value	Pr(> F)
SNP1	8.7	3.2	34.13	166.70	0.17	0.82	0.42
SNP2	12.8	−2.0	8.00	192.83	0.04	0.17	0.71
SNP3	14.3	−7.0	171.50	29.33	0.85	23.39	0.01
SNP4	4.7	7.4	153.19	47.65	0.76	12.86	0.02

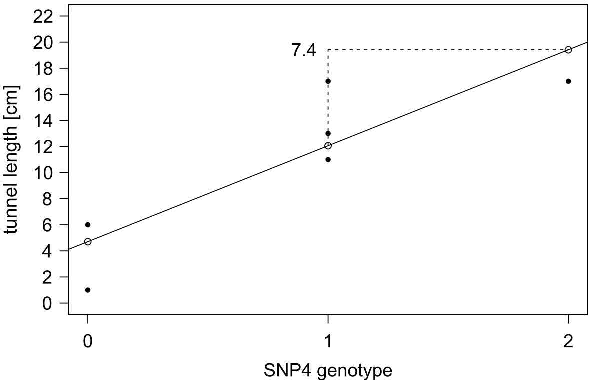


Fig. 2 Regression of tunnel length (cm) on marker genotypes at SNP4. *Solid circles* represent phenotypic values, *open circles* represent the genotypic values fitted for the 3 marker genotypes by the regression model

The fitted value of all individuals with a given genotype at SNP4 can be calculated by $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i4}$. The regression line indicates an increase in the average tunnel length for the minor allele. Thus, the mean tunnel length of genotypes homozygous for the major allele is 4.7 cm, 19.5 cm for genotypes homozygous for the minor allele and 12.1 cm for heterozygous genotypes. Note, that the coding of the three genotypes is arbitrary and that estimates of the intercept and the slope might change in magnitude or sign when genotypes are coded differently.

Repeating the SMR for the three remaining SNPs results in different estimates for the intercept and the slope for each marker as can be seen in Table 2.

To decide which of the four SNPs is significantly associated with genetic variation for tunnel length we perform an ANOVA for each SNP. At a 5 % error rate the null hypothesis $H_0 : \beta_1 = 0$ can only be rejected for SNP3 and SNP4, see Table 2. Note, for simplicity we ignore the fact that we perform statistical tests on more than one SNP and that we should control our experiment wise error rate accordingly. From the ANOVA we conclude, that SNP3 and SNP4 are in LD with a QTL affecting trait variation for tunnel length. For SNP1 and SNP2 with simulated true effects of 0 and 1 cm we cannot reject H_0. The true non-zero effect of SNP2 could not be detected because of the small simulated genetic effect and because of the limited size of our data set ($n = 6$).

To make inferences on the genetic risk of insect damage for individuals I7 and I8 we can use the estimates obtained with SMR. We know that the genotypes of the two individuals differ at SNP4. Remember that the minor allele increases tunnel length. From Table 1 we know I7 to carry one copy of the minor allele and I8 none. Thus, I7 has a higher risk of being susceptible to the European corn borer than I8.

With SMR we tested the effect of each SNP marker individually and we identified two out of four SNPs to be significantly associated with insect damage. We will now see how we can estimate the effects of the two SNP markers simultaneously by formulating a model that can account for the effects of more than one predictor variable. A natural extension of the simple marker regression model (2.1) is the *multiple marker regression model* (MMR). The data are given by $(p+1)$-tuples $(y_1, x_{11}, \ldots, x_{1j}, \ldots, x_{1p}), \ldots, (y_n, x_{n1}, \ldots, x_{nj}, \ldots, x_{np})$ and the model for p SNP effects is

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \cdots + \beta_j x_{ij} + \cdots + \beta_p x_{ip} + e_i, \\ i &= 1, \ldots, n, \; j = 1, \ldots, p, \; n \geq p+1. \end{aligned} \tag{2.4}$$

The solutions of MMR models are generally given in matrix notation and, therefore, we express the model (2.4) as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{2.5}$$

with

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1p} \\ \vdots & & & \\ 1 & & x_{ij} & \\ \vdots & & & \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_j \\ \vdots \\ \beta_p \end{bmatrix}, \quad \mathbf{e} = \begin{bmatrix} e_1 \\ \vdots \\ e_i \\ \vdots \\ e_n \end{bmatrix},$$

where the n-dimensional vector of phenotypic records $\mathbf{y}$ is described by the $(p+1)$-dimensional vector of regression coefficients $\boldsymbol{\beta}$. The $n \times (p+1)$ matrix $\mathbf{X}$ allocates observations in $\mathbf{y}$ to regression coefficients and contains n rows (one for each individual) and $p+1$ columns. As in the case of simple regression we assume the residuals e_i to be independent with equal variance (homoscedastic error terms) and to follow a normal distribution $\mathbf{e} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$ with $\mathbf{I}$ being the $n \times n$ identity matrix. The first column in $\mathbf{X}$ is a vector of ones, the following columns contain the genotype readings for p SNP markers.

Assuming $\mathbf{X}$ to be of full column rank (i.e. the unique inverse matrix $\mathbf{X}^{-1}$ exists), we obtain the ordinary least squares solution for $\boldsymbol{\beta}$ as

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}. \tag{2.6}$$

Illustration 2.2 (Continuation of Illustration 2.1) In the SMR model we identified SNP3 and SNP4 to be significantly associated with the response variable tunnel length. So let us now build the MMR model with the two SNP markers as predictors. The multiple marker regression model is

$$y_i = \beta_0 + \beta_3 x_{i3} + \beta_4 x_{i4} + e_i, \quad i = 1, \ldots, 6.$$

Table 3 Summary of results from multiple marker regression for the European corn borer data

	Estimate of $\boldsymbol{\beta}$	Partial R^2	F-value	Pr($>F$)
$\hat{\beta}_0$	10.55			
$\hat{\beta}_3$	−4.7	0.14	4.63	0.12
$\hat{\beta}_4$	3.2	0.05	1.70	0.28

Table 3 gives least squares estimates for the intercept β_0 and the partial regression coefficients β_3 and β_4 obtained with the MMR model. The regression coefficients we obtain with the MMR model deviate from those obtained with the SMR model (2.1) due to dependencies allowed between the predictor variables.

The partial regression coefficient $\hat{\beta}_4 = 3.2$ tells us that an additional copy of the minor allele at SNP4 increases tunnel length by 3.2 cm, if we hold the other variables in the model constant. The importance of a single predictor variable in a MMR model can be inferred from the partial F-statistic, which indicates whether the respective predictor variable significantly increases the proportion variance explained by a model involving all other predictors. At a 5 % error rate partial F-values for both SNP markers are not significant. The same can be shown for SNP2 and we leave it to the interested reader to verify this statement. However, from simulation of our data we know that the three SNPs have an effect on tunnel length. So from which SNP markers should we build our model? The process of variable selection in model building is not trivial and from our small example we can already see that the choice of the correct model is not unambiguous. It would be beyond the scope of this text to give an overview of the different methods employed in model selection. However, the reader should be aware that a compromise needs to be found between a model that is too simple and a model that includes too many variables. While underfitting can lead to severely biased regression coefficients and prediction, overfitting leads to large variances in both, the coefficients and prediction. The latter is especially true if dependencies exist between predictor variables. With genetic data this is often the case because high marker densities generate high LD. For a more detailed description of the choice of best model the interested reader is referred to [3, 6].

Analogously as for the SMR model we partition the total sum of squares of the response variable into a component explained by the model SQE and the residual sum of squares SQR which is not explained by the model. From this decomposition we obtain the coefficient of determination for our MMR model with two SNP markers to be $R^2 = \text{SQE}/\text{SQT} = 182.1/200.8 = 0.91$. Thus, in our example 91 % of the phenotypic variation for tunnel length can be explained by fitting the two SNP markers. This is quite a high proportion and higher than in the individual SMRs, see Table 2. In experimental studies, the proportion of the phenotypic variance explained by the model strongly depends on the genetic architecture and heritability of the trait under study, the LD in the population, and the statistical power of the experiment. Even in well powered studies, the proportion phenotypic variance explained by the model can be quite low ($\ll$50 %) if the genetic architecture of the target trait is complex [18].

3 Linear Mixed Models

In the preceding section we assumed the coefficients in the linear model to be fixed constants or *fixed effects*. If the coefficients in our model are assumed to be realizations of random variables we model them as *random effects* possibly with a variance-covariance structure. A linear model with fixed effects that is extended by effects modeled as random is called a *linear mixed model* (LMM). The theoretical foundation of LMMs was laid by Henderson in the context of livestock improvement [4].

In matrix notation the LMM is given by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}, \tag{3.1}$$

where the n-dimensional vector of phenotypic records $\mathbf{y}$ is described by the r-dimensional vector of fixed effects $\boldsymbol{\beta}$ and the p-dimensional vector of random effects $\mathbf{u}$. The $n \times r$ matrix $\mathbf{X}$ allocates observations in $\mathbf{y}$ to fixed effects and the $n \times p$ matrix $\mathbf{Z}$ allocates the observations to the random effects, and $\mathbf{e}$ is the n-dimensional vector of residuals. For the random effects and the error terms we assume normal distributions with the following expectation and variance-covariance structure

$$\mathrm{E}\begin{bmatrix}\mathbf{u}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix} \quad \text{and} \quad \mathrm{Var}\begin{bmatrix}\mathbf{u}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{bmatrix}. \tag{3.2}$$

The $p \times p$ matrix $\mathbf{G}$ and the $n \times n$ matrix $\mathbf{R}$ are the variance-covariance matrices of the random effects and residuals, respectively. By using (3.2) and the multivariate normal assumption for $\mathbf{u}$ and $\mathbf{e}$ we obtain

$$\mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} \quad \text{and} \quad \mathrm{Var}(\mathbf{y}) = \mathbf{V} = \big(\mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}\big).$$

In linear mixed models different concepts of inference apply to fixed and random effects. For the fixed effects, inference is based on estimators that are best in the sense that they have minimum sampling variance, they are linear functions of the observations in $\mathbf{y}$ and they are unbiased in the sense that $\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$. Thus, we call them best linear unbiased estimators (BLUEs). If $\mathbf{G}$ and $\mathbf{R}$ are known and the inverse of $\mathbf{V}$ exists, we obtain the BLUE of $\boldsymbol{\beta}$ with

$$\hat{\boldsymbol{\beta}} = \big(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\big)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}. \tag{3.3}$$

Note that the difference of (3.3) and (2.6) is the inclusion of the variance-covariance structure $\mathbf{V}$. If $\mathbf{V}$ has the form $\sigma^2\mathbf{I}$, Eq. (3.3) reduces to the ordinary least squares solution in (2.6).

Now let's look at the random variables involved. Our goal is to *predict* the realized values of the random variables in our model. According to [4] the best linear unbiased predictor (BLUP) of $\mathbf{u}$ is

$$\hat{\mathbf{u}} = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}). \tag{3.4}$$

The vector $\hat{\mathbf{u}}$ is an estimator of the conditional mean of $\mathbf{u}$ given $\mathbf{y}$. The effects of the vector $\hat{\mathbf{u}}$ are unbiased in the sense that $\mathrm{E}(\hat{\mathbf{u}}) = \mathrm{E}(\mathbf{u})$.

If the number of observations in $\mathbf{y}$ is large, calculations in Eqs. (3.3) and (3.4) can become computationally quite demanding because they involve $\mathbf{V}^{-1}$. Henderson [4] showed that a set of equations not involving $\mathbf{V}^{-1}$ can lead to the same solutions for $\boldsymbol{\beta}$ and $\mathbf{u}$ as (3.3) and (3.4). This set of equations

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix} \tag{3.5}$$

is generally denoted as the *mixed model equations*. If $\mathbf{G}$ and $\mathbf{R}$ are diagonal matrices, computational demands are substantially reduced compared to Eqs. (3.3) and (3.4). Let us make some even more simplifying assumptions. We assume the residuals $\mathbf{e}$ to be independent and homoscedastic with variance-covariance matrix $\mathbf{R} = \sigma^2\mathbf{I}$. If our data are generated in balanced and well designed experiments as is often the case in plant breeding this can be a realistic assumption. Let us also assume that covariances between random factors are zero and that the effects of each random factor are independent with a common, unknown variance σ_g^2, i.e. $\mathbf{G} = \sigma_g^2\mathbf{I}$ for the variance-covariance matrix of the random effects.

The mixed model equations (3.5) now reduce to

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \frac{\sigma^2}{\sigma_g^2}\mathbf{I} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}. \tag{3.6}$$

Now let us take a closer look at these equations. If our model comprised no fixed and only one random factor, the solution vector for our random effects would be

$$\hat{\mathbf{u}} = \left(\mathbf{Z}'\mathbf{Z} + \frac{\sigma^2}{\sigma_g^2}\mathbf{I}\right)^{-1} \mathbf{Z}'\mathbf{y}. \tag{3.7}$$

The difference of this solution for random effects to the ordinary least square solution for fixed effects (2.6) lies in the addition of the term $\lambda = \sigma^2/\sigma_g^2$ to the diagonal elements of $\mathbf{Z}'\mathbf{Z}$. In the prediction of the random effects λ is called the *shrinkage* parameter [3]. The predictors of the random effects are shrunken based on prior knowledge because they are assumed to have been sampled from a normal distribution with mean zero (3.2). Shrinkage induces an estimation bias for the random effects but in turn reduces the prediction variance. This can help to enhance the predictive ability of a linear regression model [3]. The amount of shrinkage is determined by the "noise to signal ratio" λ. High values of λ lead to strong shrinkage. By adding the term $\lambda\mathbf{I}$ to the $\mathbf{Z}'\mathbf{Z}$ matrix we can obtain a unique solution for $\hat{\mathbf{u}}$ even if the number of random effects p exceeds the number of observations ($n < p$). This would not be possible if we assumed the effects to be fixed, because the matrix $\mathbf{Z}'\mathbf{Z}$ does not have full column rank for $n < p$.

Remember that we assumed $\mathbf{G}$ and $\mathbf{R}$ to be known, which means that we also know σ^2 and σ_g^2. However, in real life these variance components are often unknown and must be estimated from the data. If λ cannot be assumed to be known without error the statistical properties of the estimators and predictors of the LMM change. Another problem that arises from not knowing the variance components entering into the model is how to perform hypothesis tests of the fixed effects. Remember that the BLUEs of the fixed effects are a function of $\mathbf{V}$ and consequently of the trait heritability. If σ^2 and σ_g^2 are not known but estimated from the data the resulting hypothesis tests are only approximate. For the random effects we do not perform hypothesis tests because in our model we assumed the variation due to random factors to be different from zero. In plant and animal breeding an iterative algorithm such as restricted maximum likelihood estimation (REML) [4, 5] is often used for estimation of the variance components. It is beyond the scope of this text to outline the procedures and the specific properties of the resulting estimators and predictors. The interested reader is referred to textbooks such as [5, 7].

Illustration 3.1 (Continuation of Illustration 2.1) In our example on the resistance against European corn borer we now model the SNP effects as random and the model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{m} + \mathbf{e} \tag{3.8}$$

where $\boldsymbol{\beta}$ is the vector of fixed effects including only the mean and $\mathbf{m}$ is the vector of random SNP marker effects. The distribution of the SNP effects is assumed to be $\mathbf{m} \sim \mathrm{N}(\mathbf{0}, \sigma_m^2\mathbf{I})$ with σ_m^2 being the genetic variance pertaining to a single SNP and $\mathbf{I}$ being the 4×4 identity matrix. By choosing $\mathbf{G} = \sigma_m^2\mathbf{I}$ we assume the same unknown variance σ_m^2 for the effects of all SNP markers. Note, that this does not mean that the realizations of the random variables are equal but that we assume that all effects are sampled from the same normal distribution. However, the genetic variance contributed by individual SNPs is a function of their effects and their allele frequency in the population under study. Thus, assuming a common variance for all SNP markers is a simplifying assumption. The residual vector $\mathbf{e}$ is assumed to follow a normal distribution with $\mathbf{e} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$. The matrix $\mathbf{X}$ allocates the fixed effects to observations in $\mathbf{y}$. We replace the matrix $\mathbf{Z}$ in model (3.1) by the $n \times p$ matrix $\mathbf{W}$ assigning the p SNP marker genotypes to observations in $\mathbf{y}$ and obtain the model

$$\begin{bmatrix} 13 \\ 17 \\ 1 \\ 17 \\ 11 \\ 6 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \left[\beta_0\right] + \begin{bmatrix} 2 & 2 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 2 & 0 \\ 1 & 1 & 0 & 2 \\ 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \\ m_3 \\ m_4 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \end{bmatrix}.$$

The small sample size of our example would not give us meaningful estimates of the required variance components. We use prior knowledge and assume $\lambda = \sigma^2/\sigma_m^2 = 2$.

With Eq. (3.6) we obtain solutions for the mean and the random effects by

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\mathbf{m}} \end{bmatrix} = \left[\begin{array}{c|cccc} 6 & 4 & 6 & 3 & 5 \\ \hline 4 & 8 & 6 & 0 & 5 \\ 6 & 6 & 10 & 3 & 5 \\ 3 & 0 & 3 & 7 & 0 \\ 5 & 5 & 5 & 0 & 9 \end{array}\right]^{-1} \left[\begin{array}{c} 65 \\ \hline 54 \\ 61 \\ 8 \\ 75 \end{array}\right]$$

$$\hat{\beta}_0 = 11.2 \quad \text{and} \quad \hat{\mathbf{m}} = \begin{bmatrix} 0.5 \\ -1.3 \\ -3.1 \\ 2.5 \end{bmatrix}.$$

Resulting SNP effects are smaller in absolute value compared to SMR and MMR models due to the shrinkage with $\lambda > 0$. With a value of $\lambda = 5$ resulting SNP effects are shrunken heavily with $\hat{\mathbf{m}}' = (0.6, -0.7, -2.2, 1.8)$. A value of $\lambda = 0.5$ gives us $\hat{\mathbf{m}}' = (0.3, -1.8, -4.1, 3.1)$. For $\lambda \to 0$ the effects would approach those obtained from the MMR, for $\lambda \to \infty$ the entries of $\hat{\mathbf{m}}$ would be shrunken to zero. In practice, we could use cross-validation to select a value for λ which maximizes the prediction performance of the model.

In this section we have shown, how linear mixed models are used in the analysis of genetic data. They allow us to account for covariance structures of the residuals and the random effects. The limitation of multiple linear regression to cases with $n > p$ can be addressed in LMMs by assuming SNP effects as random. Within this framework we can arrive at predictions for the marker effects inferred from high-density genome-wide marker data with $n \ll p$. In the next sections we will look at the application of LMMs for predicting the genetic value of individuals based on information from relatives.

4 Prediction of Genetic Values

The prediction of genetic values based on information from relatives has a long tradition in breeding. We will now learn how to predict the genetic value of individuals by exploiting information from relatives. We formulate a linear mixed model which is based on pedigree data and denoted by *modA*

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a} + \mathbf{e}, \tag{4.1}$$

with the vector of phenotypes $\mathbf{y}$, the vector of fixed effects $\boldsymbol{\beta}$, the vector of genetic values $\mathbf{a}$ being sampled from a distribution $\mathbf{a} \sim \mathrm{N}(\mathbf{0}, \mathbf{G_A} = \sigma_A^2\mathbf{A})$, and the vector of residual effects $\mathbf{e}$ with $\mathbf{e} \sim \mathrm{N}(\mathbf{0}, \mathbf{R} = \sigma^2\mathbf{I})$. Matrices $\mathbf{X}$ and $\mathbf{Z}$ allocate the fixed and random effects to observations $\mathbf{y}$, respectively. The matrix $\mathbf{G_A}$ defines the variance-covariance structure for the random effects. In the following we will show how $\mathbf{G_A}$ can be constructed based on pedigree information.

The genetic covariance between relatives increases with their degree of relatedness. If we measure a quantitative trait such as body height on ten pairs of mothers and daughters, we expect a positive genetic covariance for this trait, because we know that body height is highly heritable. We expect tall mothers to have daughters which are taller than average. If we measure the covariance in body height for pairs of first cousins we expect the covariance to be smaller than for pairs of mothers and daughters, because the genetic covariance between two relatives is proportional to the probability that they share ancestral alleles. This probability is called the *kinship coefficient or coefficient of coancestry* [1]. Let's assume that individual I carries alleles A_k and A_l at a given gene and individual I′ carries alleles A_r and A_s at the same gene. We have four pairwise possibilities that two randomly selected alleles of individuals I and I′ are *identical by descent* (here denoted with the symbol ≡). The kinship coefficient f for I and I′ is given by

$$\begin{aligned} f_{\mathrm{II'}} &= \mathrm{P}\big(\mathrm{I} \equiv \mathrm{I'}\big) \\ &= \frac{1}{4}\big[\mathrm{P}(A_k \equiv A_r) + \mathrm{P}(A_k \equiv A_s) + \mathrm{P}(A_l \equiv A_r) + \mathrm{P}(A_l \equiv A_s)\big]. \end{aligned} \tag{4.2}$$

Because the genetic covariance originates from alleles shared between two individuals, it is intuitive that it must be a function of the genetic variance pertaining to these alleles. Using quantitative genetic theory [1] it can be shown that the genetic covariance between two relatives I and I′ is given by $2f_{\mathrm{II'}}\sigma_A^2$, i.e. twice the kinship coefficient times the additive genetic variance. Recall from the introduction that we assume the genetic variance to be purely additive, and dominance and epistasis to be absent. Thus, the genetic variance-covariance matrix based on pedigree information becomes $\mathbf{G_A} = \sigma_A^2\mathbf{A}$ with the elements of $\mathbf{A}$ given by $2f_{\mathrm{II'}}$. The matrix $\mathbf{A}$ is called the numerator relationship matrix, because it provides us with the coefficients of the additive genetic covariance for all pairwise combinations of individuals.

Illustration 4.1 (Continuation of Illustration 2.1) In the European corn borer example we do not have phenotypic values for individuals I7 and I8, but we still want to know their genetic value. Both individuals are derived from the cross I1 × I2. Because we have phenotypic data from relatives of I7 and I8 we can predict their genetic values with model (4.1)

$$\begin{bmatrix} 13 \\ 17 \\ 1 \\ 17 \\ 11 \\ 6 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \mu + \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} a_{\mathrm{I1}} \\ a_{\mathrm{I2}} \\ a_{\mathrm{I3}} \\ a_{\mathrm{I4}} \\ a_{\mathrm{I5}} \\ a_{\mathrm{I6}} \\ a_{\mathrm{I7}} \\ a_{\mathrm{I8}} \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \end{bmatrix}.$$

The matrix $\mathbf{Z}$ has eight columns, the last two containing only zeros because we have no phenotypic records for the corresponding individuals. Since we know the pedigree of the eight individuals we can construct the symmetric 8 × 8 matrix $\mathbf{A}$

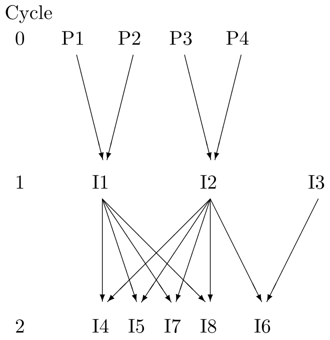


Fig. 3 Pedigree for eight maize individuals and their parents from three breeding cycles

that provides the coefficients of the variance-covariance matrix $\mathbf{G_A}$. In Fig. 3 the pedigree structure of the eight individuals in our example data set is depicted (for pedigrees see also Table 1).

We assume the individuals from breeding cycle 0 to be unrelated and non-inbred. Individuals I4 and I5 are both progeny of I1 and I2, i.e. they are full siblings, while individuals I5 and I6 are half siblings with only one common parent. Individuals I5 and I6 can inherit alleles identical by descent only through their common parent I2. Let the allele descended from I2 be A_k in I5 and A_r in I6. Thus we obtain $\mathrm{P}(A_k \equiv A_r) = 2 \cdot (\frac{1}{2} \cdot \frac{1}{2})$ and $\mathrm{P}(A_k \equiv A_s) = \mathrm{P}(A_l \equiv A_r) = \mathrm{P}(A_l \equiv A_s) = 0$. Using (4.2) we get $f_{\mathrm{I5I6}} = 0.125$ for these half siblings. With individuals I4 and I5 that have two parents in common we get $f_{\mathrm{I4I5}} = 0.25$ the same value that we also get for a parent and its offspring. The off-diagonal elements of $\mathbf{A}$ are given by $2f_{\mathrm{II'}}$, but how can we calculate the kinship coefficient of an individual with itself? It is the same as the kinship coefficient of monozygotic twins. We follow the rules of (4.2) and get $f_{\mathrm{II}} = 0.5$. The coefficients for more complex pedigrees may be obtained using the formulas presented in [5]. For the corn borer example the genetic variance-covariance matrix becomes

$$\mathbf{G_A} = \sigma_A^2 \mathbf{A} = \sigma_A^2 \begin{bmatrix} 1 & 0 & 0 & 0.50 & 0.50 & 0 & 0.50 & 0.50 \\ & 1 & 0 & 0.50 & 0.50 & 0.50 & 0.50 & 0.50 \\ & & 1 & 0 & 0 & 0.50 & 0 & 0 \\ & & & 1 & 0.50 & 0.25 & 0.50 & 0.50 \\ & & & & 1 & 0.25 & 0.50 & 0.50 \\ & & & & & 1 & 0.25 & 0.25 \\ & & & & & & 1 & 0.50 \\ & & & & & & & 1 \end{bmatrix}.$$

Note, that the matrix $\mathbf{A}$ is symmetric and that elements below the diagonal are not given for the sake of readability. Coefficients of zero indicate unrelated individuals, values of 0.25 half siblings and values of 0.50 full siblings or parents and offspring. To solve the mixed model equations in (3.5), we need to choose the shrinkage parameter λ. From prior knowledge on the trait heritability we assume that $h^2 = \sigma_A^2/(\sigma_A^2 + \sigma^2) = 0.5$. Thus we infer $\lambda = \sigma^2/\sigma_A^2 = 1$ and solve the mixed model equations for $\mathbf{a}$. The estimated fixed and predicted genetic values for I1–I8 are

$$\hat{\beta}_0 = 10.2 \quad \text{and} \quad \hat{\mathbf{a}}' = \left[1.8, 3.3, -5.1, 4.0, 2.0, -2.0, 2.6, 2.6\right].$$

Without having measured their phenotype we can predict genetic values of I7 and I8. We get the same value for both individuals (2.6 cm), because they are full siblings and have the same degree of relatedness with all individuals for which we have phenotypes (see columns 7 and 8 of matrix **A** for confirmation). Because most of their close relatives show high values for tunnel length, their predicted genetic value is also quite high. Thus, we conclude that individuals I7 and I8 both carry a high genetic risk of being damaged by the European corn borer, but we cannot differentiate between them.

We have seen how the linear mixed model *modA* can be used to obtain the vector of predicted genetic values $\hat{\mathbf{g}}_{\mathbf{A}} = \mathbf{I}\hat{\mathbf{a}}$. A similar approach can be taken for prediction of genetic values from genomic data. In Sect. 3 we fitted the LMM with SNP markers as random factors. Let us call this model *modRR* because we perform *random regression* on the SNP markers. From *modRR* we obtain a vector of predicted SNP effects $\hat{\mathbf{m}}$. The predicted genetic value of an individual based on marker information can be obtained by multiplying the genotype score of an individual at a given SNP marker with the corresponding SNP effect and summation over all m SNP markers. Thus the predicted genetic value of individual i is given by

$$\hat{g}_i = \sum_{j=1}^{m} w_{ij}\hat{m}_j, \quad i = 1, \ldots, n;$$

or in matrix notation as $\hat{\mathbf{g}}_{\mathbf{RR}} = \mathbf{W}\hat{\mathbf{m}}$ with $\mathbf{W}$ comprising the genotype scores for all individuals including those without phenotypes.

Which of the two models should we use to infer the genetic value of individuals? To answer this question let us take a closer look at the major differences between the two models. Model *modA* is based on a linear mixed model where the genetic value for each individual is the random effect. Pedigree information is used to exploit the resemblance between relatives and to construct the variance-covariance structure of the genetic values. Model *modRR* is based on a linear mixed model where the genetic value for each individual is given by the sum of the random SNP effects. The resemblance between relatives in *modRR* is modeled by the dependencies of marker genotypes between pairs of individuals. So how do these two sources of information differ?

Let us assume a family consisting of mother, father and four direct progeny. We can arrange the four progeny in six different pairs. On average, the kinship between individuals of a given pair is expected to be 0.25. However, genetic recombination and sampling of the parental gametes will result in some pairs that share more and some that share less than 25 % of their ancestral alleles. The elements of matrix **A** give us the same *expected* coefficient for all six pairs because they are all full siblings. However, the SNP marker data can quantify the deviation from the expected

Table 4 True and predicted genetic values for risk of insect damage of eight maize individuals obtained with different prediction procedures. The prediction is evaluated with the mean squared error (MSE) between true and predicted genetic values and the coefficient of determination R^2 from a regression of the predicted on the true genetic values

	I1	I2	I3	I4	I5	I6	I7	I8		
$\tilde{g}$	6	4	−7	9	5	−3	5	1	MSE	R^2
modA	1.8	3.3	−5.1	4.0	2.0	−2.0	2.6	2.6	8.1	0.85
modRR	1.0	2.5	−7.5	4.3	1.8	−4.4	1.8	−0.7	9.4	0.95

value, because we can infer the *realized* proportion of shared alleles between two individuals from their marker profile. Thus, predictions based on marker information should be superior to predictions based on model *modA* because the marker information can model the realized genetic relatedness between two individuals whereas the matrix $\mathbf{A}$ can only account for the expected relatedness. The interested reader who wants to learn more about the statistical dependencies of the two models is referred to [8, 15].

Illustration 4.2 (Continuation of Illustration 2.1) We will take the corn borer example to illustrate the differences between *modA* and *modRR*. Table 4 summarizes true and predicted genetic values for the eight maize individuals from Table 1. The true genetic values were calculated from simulation parameters as $\tilde{\mathbf{g}} = \mathbf{W}\tilde{\mathbf{m}}$ with $\tilde{\mathbf{m}}' = (0, 1, -4, 4)$. Predictions from models *modA* and *modRR* were obtained as described above with $\hat{\mathbf{g}}_{\mathbf{A}} = \mathbf{I}\hat{\mathbf{a}}$ ($\lambda = 1$) and $\hat{\mathbf{g}}_{\mathbf{RR}} = \mathbf{W}\hat{\mathbf{m}}$ ($\lambda = 2$), respectively.

Due to the small sample size of our example our predictors cannot be expected to be very precise. However, the data still give evidence for differences between the two models. It is important to note, that the ranking of the eight individuals can change when modeling the genetic relationship between individuals based on marker data as compared to pedigree data. In addition, model *modRR* provides distinct values for individuals I7 and I8 thus improving our prediction on their respective risk of being damaged by insects compared to *modA*. On the other hand we can see that with model *modRR* the predicted genetic values for individuals I5 and I7 are identical. Due to the small number of markers simulated they had identical marker genotypes at all SNP loci. Nowadays we assess the marker profile of each individual with thousands of markers, so the probability of having identical marker genotypes is extremely small for real life data. However, if the number of markers is very large, solving the mixed model equations in (3.5) with model *modRR* can be computationally demanding. In order to avoid this computational burden, an $n \times n$ genomic relationship matrix can be constructed from high-dimensional marker data. If this genomic relationship matrix is used in (4.1) to replace the matrix $\mathbf{A}$, we obtain genetic predictions equivalent to those of *modRR*. In the statistical literature this approach is also known as “kernel trick”. For details the interested reader is referred to [3] and [8, 15].

5 Choice of Models

In the preceding sections we have learnt how fixed linear regression models and linear mixed models can be used for estimation of marker effects and prediction of genetic values. Major differences between the models lie in the assumption of treating marker effects as random or fixed effects. In many real life experiments the number of assayed SNP markers exceeds by far the number of observations. For example in maize genetics we currently use technologies that return data on more than 50,000 SNP markers per individual but population sizes are within the range of a few hundred to a few thousand. So which model do we choose to obtain meaningful results from the given data? First of all, we need to keep in mind what we already know about our quantitative trait of interest. For many traits such as disease or insect resistance for example it is legitimate to assume that they are regulated by few genes with sizeable effects. Thus, our aim will be to identify the SNP markers located next to these genes with SMR. To account for dependencies between the selected predictors, we formulate a MMR model and estimate the effects of all selected variables simultaneously. Finally, the coefficient of determination from the MMR model will give us an indication how much of the variation of the response variable can be explained by the model. We need to keep in mind though that a high coefficient of determination does not necessarily mean that our model has high predictive ability. With thousands of SNP markers tested for significance we run into the problem of multiple testing and we are likely to identify false positive signals due to inflated error rates. Hence more SNPs than actually required would be included in the model. At the same time estimated SNP effects can be severely biased due to model selection or hidden population structure. This will lead to an overestimation of their importance in the regulation of our quantitative trait. Consequently we need to assess whether the SNP markers we have selected will predict the genetic merit of our selection candidates with sufficient accuracy. For the data at hand, statistical methods such as cross-validation will give a first indication of the bias associated with SNP effects [19]. However, before entering into fine mapping, cloning or marker-based selection with the selected SNPs it is highly advisable to confirm the significance and the strength of the statistical association with the trait of interest in independent validation studies.

An alternative to selecting a few "winner" SNPs is to formulate models that use the entire set of SNP markers simultaneously. This appears to be a logical approach for traits for which we have conclusive evidence from quantitative genetics that they are regulated by many genes distributed over the entire genome. If many genes act on the trait of interest, the effect of individual markers will be small and with SMR it will be difficult to identify SNPs that contribute significantly to phenotypic trait variation. Thus, for truly quantitative traits it is more appropriate to take a genome-wide approach and predict the genetic value of an individual from the accumulated effects of a large number of SNPs distributed evenly across the genome. However, when interpreting the magnitude and distribution of the individual SNP effects caution is advised, because the large number of predictors in the model and the high degree of multicollinearity between them lead to heavily shrunken and hence biased individual marker effects.

A combination of the presented methods is frequently adopted in genome-wide association studies. Many genetic studies in plant, animal or human genetics aim at the identification of individual SNP markers in strong LD with a QTL. Because the limitations of SMR are well known, a linear mixed model combining some of the properties of single marker regression and *modA* is chosen. Each SNP marker is modeled as a fixed factor but the dependencies between the individuals under study are modeled by including a random factor with a variance-covariance structure that accounts for their genetic relatedness based on marker or pedigree data. The interpretation of results from this model with respect to multiple testing and ascertainment bias of the estimated effects is similar to what has been discussed for single marker regression.

6 Food for Thought

This article presented a survey of models used for the prediction of complex phenotypes. However, throughout this chapter we have made simplifying assumptions. For example we ignored interactions between alleles at the same gene and between alleles at different genes and our predictions were based on a purely additive model. Remember that in a linear model the observations must be linear functions of the model parameters, but we can still formulate models accommodating many types of interactions between factors. However, with 50,000 SNP markers the number of predictors in the model and the number of possible models becomes inconceivably large. In the literature it has been discussed that non parametric methods such as machine learning or neural networks might improve predictive abilities if dominance and epistasis are present [12]. However, so far only little evidence from experimental data has been available to test the validity of this hypothesis.

In addition to neglecting genetic interactions we also assumed that we only have one observation per individual thus ignoring heteroscedasticity of the residual effects in our models and interactions between genotypes and the environment. In plant breeding, we measure phenotypes in many environments because we know that genotypes react differently to changing environmental conditions. Thus, an additional complication in model building arises from the need to accommodate replicated data. It would be beyond the scope of this chapter to describe the many methods offered by the literature to deal with genotype × environment interactions, but the reader should be aware that the analysis of replicated data is not trivial especially when data are highly unbalanced.

Theoretical and applied research on the above mentioned topics is bustling and the methods for genetic data analysis are constantly improved. The aim of this chapter was to introduce some of the concepts and methodologies related to the analysis of genomic data for prediction of complex phenotypes in agricultural genetics. For the sake of brevity we could only give an introduction to a small selection of topics. Many more statistical procedures have been suggested for prediction of genetic values of individuals for which phenotypic data is not available. One important class of methods are Bayesian approaches. The Bayesian methods employed in prediction of genetic values differ from *modRR* mainly with respect to assumptions on the distributions of the variances of marker effects [11]. Some of the Bayesian methods assume a subset of markers to be sampled from a distribution with zero variance introducing a variable selection component into the model. What has become apparent from the analysis of experimental data is that the predictive abilities of the Bayesian variable selection methods and *modRR* do not differ to a great extent unless genes with large effects are segregating in the population under study. This leads us back to the discussion on the genetic architecture of our quantitative trait of interest.

We have seen that there are many methods to analyze and interpret genetic data. A model that gives satisfactory answers to one question might not be appropriate for answering another question. Different research questions often may require different statistical models. What is important for the researcher is to realize which assumptions have to be met to apply a specific model and the potential drawbacks that can be associated with a specific procedure.

Agricultural and medical genetics are currently revolutionized by the technological developments in genomic research. High-throughput genotyping has become reality for many species and sequencing of whole genomes at reasonable costs is within striking distance. Thus, the genetic analysis of quantitatively inherited traits and the prediction of the genetic predisposition of individuals based on molecular data are rapidly evolving fields of research. The statistical methods introduced in this chapter have been developed in agricultural genetics for the analysis of high-dimensional genomic and phenotypic data. First reports in the literature have also suggested their use for prediction of the genetic predisposition in humans [10, 13], because many important traits studied in human genetics such as high blood pressure, diabetes or psychiatric disorders also follow a quantitative pattern of inheritance. The validity of the models for predicting disease risk in humans is yet to be shown. Human cohorts employed in genetic analyses have very different population structures compared to highly selected and often inbred agricultural populations. In addition, the requirements with respect to prediction accuracies are certainly more stringent in human than in agricultural genetics. Nevertheless, the proposed methods certainly are valuable tools in the analysis of human genetic data. They can be adapted to the analysis of already existing data sets and provide valuable insights into the genetic architecture of quantitative traits. How human genetics will cope with the increasing possibilities of genetic prediction and counseling will need to be addressed by medical ethics research. A debate about the consequences of our steadily growing knowledge on the inheritance of quantitative traits has started in human genetics [14].

7 Summary

We hope it has become apparent from reading this chapter that the analysis of genetic data in agriculture and medicine is a highly interdisciplinary task. Technological developments in sequence, protein and metabolite analyses are moving at an enormous pace, challenging quantitative genetics and biostatistics to cope with the exponentially growing amount of data. High-throughput, high-quality genomic data in combination with efficient statistical and computational tools are currently advancing our knowledge on the inheritance of quantitative traits at mind boggling speed. These developments make research in the field of quantitative genetics groundbreaking and supremely exciting. In addition, the concepts for data mining to arrive at accurate predictions based on high-dimensional data presented here are applicable to a plethora of research fields and applications.

Acknowledgements We thank Theresa Albrecht for help with the European corn borer example and Peter Westermeier for providing the corn borer photos. Valentin Wimmer acknowledges financial support by the German Federal Ministry of Education and Research (BMBF) within the AgroClustEr *Synbreed—Synergistic plant and animal breeding* (*FKZ 0315528A*).

References

Selected Bibliography

1. D.S. Falconer, T.F.C. Mackay, *Introduction to Quantitative Genetics* (Longman Technical, Harlow, 1996)
2. A.J.F. Griffiths, S.R. Wessler, S.B. Carroll, J. Doebley, *Introduction to Genetic Analysis*, 10th edn. (Palgrave Macmillan, Basingstoke, 2012)
3. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics (Springer, Berlin, 2009)
4. C.R. Henderson, *Applications of Linear Models in Animal Breeding* (University of Guelph, Guelph, 1984)
5. M. Lynch, B. Walsh, *Genetics and Analysis of Quantitative Traits* (Sinauer, Sunderland, 1998)
6. R.H. Myers, *Classical and Modern Regression with Applications* (Duxbury, Belmont, 1994)
7. S.R. Searle, G. Casella, C.E. McCulloch, *Variance Components*. Wiley Series in Probability and Statistics (Wiley-Interscience, Hoboken, 2006)

Additional Literature

8. T. Albrecht, V. Wimmer, H.-J. Auinger, M. Erbe, C. Knaak, M. Ouzunova, H. Simianer, C.-C. Schön, Genome-based prediction of testcross values in maize. Theor. Appl. Genet. **123**(2), 339–350 (2011)
9. E.S. Buckler, J.B. Holland, P.J. Bradbury, C.B. Acharya, P.J. Brown, et al., The genetic architecture of maize flowering time. Science **325**(5941), 714–718 (2009)
10. G. de los Campos, D. Gianola, D.B. Allison, Predicting genetic predisposition in humans: the promise of whole-genome markers. Nat. Rev. Genet. **11**(12), 880–886 (2010)
11. D. Gianola, G. de los Campos, W.G. Hill, E. Manfredi, R.L. Fernando, Additive genetic variability and the Bayesian alphabet. Genetics **183**(1), 347–363 (2009)
12. D. Gianola, H. Okut, K.A. Weigel, G.J.J. Rosa, Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. BMC Genet. (2011). doi:10.1186/1471-2156-12-87

13. M.E. Goddard, N.R. Wray, K. Verbyla, P.M. Visscher, Estimating effects and making predictions from genome-wide marker data. Stat. Sci. **24**(4), 517–529 (2009)
14. D.B. Goldstein, Growth of genome screening needs debate. Nature **476**(7358), 27–28 (2011)
15. D. Habier, R.L. Fernando, J.C.M. Dekkers, The impact of genetic relationship information on genome-assisted breeding values. Genetics **177**(1), 2389–2397 (2007)
16. T.H.E. Meuwissen, B.J. Hayes, M.E. Goddard, Prediction of total genetic value using genome-wide dense marker maps. Genetics **157**(4), 1819–1829 (2001)
17. E.C.G. Pimentel, M. Erbe, S. König, H. Simianer, Genome partitioning of genetic variation for milk production and composition traits in Holstein cattle. Front. Livest. Genomics **2**, 19 (2011)
18. C.-C. Schön, H.F. Utz, S. Groh, B. Truberg, S. Openshaw, A.E. Melchinger, Quantitative trait locus mapping based on resampling in a vast maize testcross experiment and its relevance to quantitative genetics for complex traits. Genetics **167**(1), 485–498 (2004)
19. H.F. Utz, A.E. Melchinger, C.-C. Schön, Bias and sampling error of the estimated proportion of genotypic variance explained by quantitative trait loci determined from experimental data in maize using cross validation and validation with independent samples. Genetics **154**(4), 1839–1849 (2000)
20. P.M. Visscher, Sizing up human height variation. Nat. Genet. **40**(5), 489–490 (2008)

Chapter 8
Bayesian Risk Analysis

Claudia Czado and Eike Christian Brechmann

Uncertainty in the behavior of quantities of interest causes risk. Therefore statistics is used to estimate these quantities and assess their variability. Classical statistical inference does not allow to incorporate expert knowledge or to assess the influence of modeling assumptions on the resulting estimates. This is however possible when following a Bayesian approach which therefore has gained increasing attention in recent years. The advantage over a classical approach is that the uncertainty in quantities of interest can be quantified through the posterior distribution. We first introduce the Bayesian approach and illustrate its use in simple examples, including linear regression models. For more complex statistical models Markov Chain Monte Carlo methods are needed to obtain an approximate sample from the posterior distribution. Due to the increase in computing power over the last years such methods become more and more attractive for solving complex problems which are intractable using classical statistics, for instance spam e-mail filtering or the analysis of gene expression data. We illustrate why these methods work and introduce two most commonly used algorithms: the Gibbs sampler and Metropolis Hastings algorithms. Both methods are derived and applied to statistical models useful in risk analysis. In particular a Gibbs sampler is developed for a change point detection in yearly counts of events and for a regression model with time dependence, while a Metropolis Hastings algorithm is derived for modeling claim frequencies in an insurance context.

Keywords Bayesian inference · Markov Chain Monte Carlo samplers · Bayesian risk · Prior · Posterior

Mathematics Subject Classification (2010) 62F15 · 62F86 · 62J05

C. Czado (✉) · E.C. Brechmann
Applied Statistics, Center for Mathematical Sciences, Technische Universität München,
Boltzmannstr. 3, 85748 Garching bei München, Germany
e-mail: cczado@ma.tum.de

C. Klüppelberg et al. (eds.), *Risk – A Multidisciplinary Introduction*,
DOI 10.1007/978-3-319-04486-6_8,

The Facts

- Risk is regarded as induced by the uncertainty in the behavior of quantities of interest. Therefore this random behavior has to be modeled using probability models and characteristics such as expected value and variance to be estimated.
- An introduction to Bayesian statistics is given, which—in contrast to classical statistics—can accommodate prior knowledge about the risk parameters under consideration, in particular using Bayes' famous theorem. Especially expert knowledge can be incorporated.
- Bayesian inference is based on the posterior distribution of the risk parameters which summarizes the knowledge about the risk quantity after the data is observed. Common Markov Chain Monte Carlo methods for deriving the Bayesian posterior distribution are discussed, namely the Gibbs sampler and Metropolis Hastings algorithms.
- Concepts are illustrated by examples from insurance, health care, mining and agriculture involving the risk quantities number of claims, complication rate of new medical treatment, number of coal-mining disasters and crop yield, respectively.

1 The Bayesian Approach

In this chapter we are interested in the study of quantities which are subject to uncertainty. In this context we understand risk as a process which is induced by uncertainty or randomness in the behavior of these quantities. To be more precise we will consider among other the following risk quantities: yearly crop rates, number of complications following a new medical treatment and the annual number of claims for a car insurance company. For the statistical risk analyst these quantities are random variables for which a probability distribution has to be chosen which depends on unknown population parameters and fits the observed data well. These population parameters determine the expectation and variance of the risk quantity. Classical—usually called *frequentist*—statistics uses solely the observed data to estimate the unknown population parameters. This is a sensible approach, however, the randomness in the observations and the limited number of observations available can lead to errors in subsequent inference. We assume that the reader has basic knowledge in probability and statistics; for convenience a glossary is provided in Appendix. Three illustrative examples are presented after this first short introduction.

In the simplest possible setting, we assume that observations come from a population whose members follow a specific probability distribution which depends on a single parameter θ. Given that we know this particular underlying distribution, we are interested in estimating θ based on the observed data. We denote such an estimate by $\hat{\theta}$. For example, if θ is the expectation of the distribution, we can estimate it by the average of all observations, that is

$$\hat{\theta} = \bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{1.1}$$

where n is the number of observations with values $x_1, \ldots, x_n$.

In practice, the estimate $\hat{\theta}$ will however pretty much never equal the true parameter θ, that is, in general $\hat{\theta} \neq \theta$. Moreover, we might obtain an estimated value $\hat{\theta}$ which is unbelievable because it maybe lies outside a range where we expected the parameter to be in. If we however still believe that our probability model for the observed data is correct, we are in the dilemma that we have to decide between our belief in the data model and our prior belief in the parameter.

Bayesian statistics solves this problem by combining prior expert knowledge with information obtained from the observations. From now on, let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)' \in \Theta$ be the unknown parameter of interest belonging to the parameter space Θ, where usually $\Theta \subset \mathbb{R}^k$. Then we a priori assign a probability to each parameter value $\boldsymbol{\theta}$ according to the prior expert knowledge available, that is, we treat the population parameter as random variable and not as a fixed unknown quantity. Statistically speaking, this means that we choose an appropriate *prior distribution* with density or probability function $p(\boldsymbol{\theta})$, which summarizes the knowledge about the parameter of interest. We now observe a random sample $\boldsymbol{x} = (x_1, \ldots, x_n)'$, which are realizations of random variables $\boldsymbol{X} = (X_1, \ldots, X_n)'$ with true probability density $f(\cdot|\boldsymbol{\theta})$. For example x_i is the observed crop yield in plot i of the random crop yield X_i. Considering $f(\boldsymbol{x}|\boldsymbol{\theta})$ as a function of the parameter $\boldsymbol{\theta}$ for given observations $\boldsymbol{x}$ yields the *likelihood* denoted as

$$\ell(\boldsymbol{\theta}|\boldsymbol{x}) := f(\boldsymbol{x}|\boldsymbol{\theta}), \tag{1.2}$$

which summarizes the available information in the data about the parameter.

Note that in frequentist statistics, parameters are often estimated by so-called *maximum likelihood estimation* which means finding the parameter values $\hat{\boldsymbol{\theta}}$ that maximize (1.2), that is, finding the value of $\boldsymbol{\theta}$ which makes the observations "most likely". For example, the quantity in (1.1) is the maximum likelihood estimate of the expectation μ of a normal distribution (see Illustration 1.1 below).

In Bayesian statistics, we however would like to incorporate prior knowledge about the parameter $\boldsymbol{\theta}$, that is the prior distribution, into the estimation procedure. Since the observations $\boldsymbol{x}$ contain information about $\boldsymbol{\theta}$, we update our knowledge about $\boldsymbol{\theta}$ by considering the conditional distribution of $\boldsymbol{\theta}$ given observations $\boldsymbol{x}$. This distribution is called the *posterior distribution* and can be calculated by *Bayes' theorem* as

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{f(\boldsymbol{x})}, \tag{1.3}$$

where

$$f(\boldsymbol{x}) = \int_{\Theta} f(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{1.4}$$

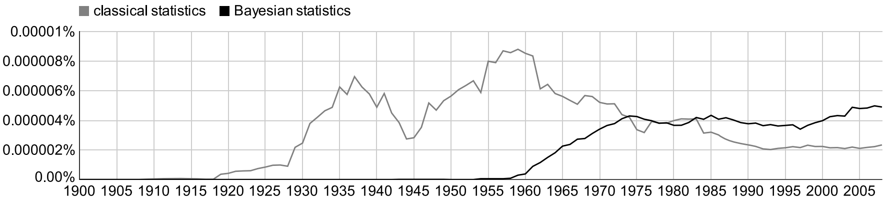


Fig. 1 Ngram of "classical statistics" (*gray*) and "Bayesian statistics" (*black*) created using Google Books Ngram Viewer available at http://books.google.com/ngrams

is the unconditional density function of the observations $\boldsymbol{x}$, called the *marginal distribution*. It does not depend on $\boldsymbol{\theta}$, in other words, it is only a normalizing constant with respect to $\boldsymbol{\theta}$ that ensures that the posterior distribution is a proper density expression integrating to 1. Hence it holds that

$$p(\boldsymbol{\theta}|\boldsymbol{x}) \propto \ell(\boldsymbol{\theta}|\boldsymbol{x})p(\boldsymbol{\theta}), \tag{1.5}$$

that is, the posterior is proportional to the product of the likelihood and the prior. The computation of the posterior distribution however often is rather intricate so that so-called Markov Chain Monte Carlo methods are needed as discussed in Sect. 2.

A standard reference on Bayesian inference is the book by Berger [8], more recent references are Lee [5], Gelman et al. [15], Bolstad [1] and Hoff [20]. To illustrate the increasing importance of Bayesian methods in statistics, Fig. 1 shows how often the terms "classical statistics" and "Bayesian statistics" have occurred in books since 1900.

Three illustrative examples for different types of data (continuous, binary, count) are given below. These represent common types of risk quantities.

Illustration 1.1 (Crop Yields) Too small crop yields constitute a major risk to farmers. A reliable estimate of the expected crop yield and its variability therefore is needed for careful business planning. For this purpose, an agronomist studies the behavior of the random annual crop yields $X_1, \dots, X_n$ of n acres of the same size and with similar soil and growth conditions. From her experience and discussions with farmers she assumes that the crop yields are normally distributed with common mean θ and (known) variance σ^2 and independent of each other, that is $X_i \sim N(\theta, \sigma^2)$, $i = 1, \dots, n$. Then the likelihood (1.2) is

$$\ell(\theta|\boldsymbol{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x_i - \theta)^2\right] \propto \exp\left[-\frac{n}{2\sigma^2}(\overline{x} - \theta)^2\right],$$

where $\overline{x}$ is the empirical mean as defined in (1.1). It is a unimodal function in θ with mode given by $\overline{x}$.

From previous years the agronomist has some prior knowledge about the likely values of the expected crop yield θ and therefore specifies a prior distribution as normal with known mean μ and known variance τ^2. Having observed the crop yields

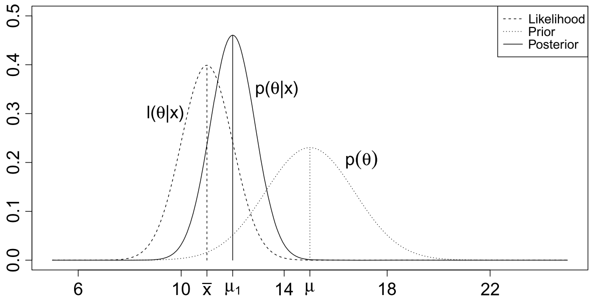


Fig. 2 Likelihood, prior and posterior densities for $n = 5$, observation variance $\sigma^2 = 5$, prior mean $\mu = 15$, prior variance $\tau^2 = 3$ and observed mean $\overline{x} = 11$

$x_1, \ldots, x_n$, she therefore calculates the posterior density (1.3) using (1.5) as

$$p(\theta|\boldsymbol{x}) \propto \exp\left[-\frac{n}{2\sigma^2}(\overline{x}-\theta)^2\right]\exp\left[-\frac{1}{2\tau^2}(\theta-\mu)^2\right] \propto \exp\left[-\frac{1}{2}\frac{(\theta-\mu_1)^2}{\tau_1^2}\right], \tag{1.6}$$

where

$$\tau_1^2 = \frac{1}{n\sigma^{-2}+\tau^{-2}} \quad \text{and} \quad \mu_1 = \tau_1^2\left(\frac{\overline{x}}{n^{-1}\sigma^2}+\frac{\mu}{\tau^2}\right). \tag{1.7}$$

From (1.6) it follows that the posterior distribution is again normal but now with mean μ_1 and variance τ_1^2. To illustrate these concepts further, let us assume the agronomist expects an average yield of 15 per acre, that is, she sets the prior mean $\mu = 15$. She is however uncertain about her guess and therefore allows for a large uncertainty by choosing the prior variance τ^2 to be 3. After harvesting $n = 5$ acres, the observed average yield was $\overline{x} = 11$ per acre. The seed manufacturer claims that the variability under normal growing conditions is $\sigma^2 = 5$ per acre. Therefore the posterior distribution has posterior moments $\tau_1^2 = 0.75$ and $\mu_1 = 2$. This is illustrated in Fig. 2.

The expression of the posterior expectation μ_1 in (1.7) can conveniently be rewritten as

$$\mu_1 = w\overline{x} + (1-w)\mu, \tag{1.8}$$

where $w := w(\sigma^2, \tau^2, n) := \frac{\tau^2}{\tau^2+\sigma^2/n}$ is a weight varying from 0 to 1. Expression (1.8) shows that the posterior mean is the weighted average of the empirical mean $\overline{x}$ and the prior mean μ. As the uncertainty in the prior knowledge, reflected by the prior variance τ^2, increases, the weight $(1-w)$ for the prior mean decreases and the posterior mean is more heavily pulled towards the empirical mean. Moreover, the belief in the observed data as measured by the weight w also increases when the number of observations n, the number of acres under consideration, is increased. In the example it is $w = 0.75$. This means that there is already a quite strong belief in the data.

Illustration 1.2 (Complication Rate in Medical Studies) In a medical study, the researcher is interested in the rate of complications θ of n subjects. Clearly, the risk of the researcher is that this rate θ is higher than a small but admissible limit rate. At the end of the study, for each subject it is known whether he or she developed a complication or not. The event of complication occurrence can be modeled by a binary random variable X_i which is either 1 if the patient $i \in \{1, \ldots, n\}$ develops a complication or 0 otherwise. Because the researcher developed a completely new treatment, no prior knowledge about the success probability θ of the Bernoulli distribution representing the complication probability is available. Hence, she simply assumes equal likelihood for each parameter value θ, in other words, a prior density $p(\theta) = 1$ corresponding to the uniform distribution. For observations $x_1, \ldots, x_n$ the posterior distribution (1.3) for θ therefore simplifies to the likelihood (1.2):

$$p(\theta|\boldsymbol{x}) \propto \ell(\theta|\boldsymbol{x}) p(\theta) = \ell(\theta|\boldsymbol{x}) = \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i}.$$

If, however, prior information based on studies of similar treatments is available, the researcher can specify a more informative prior distribution. For a parameter in the range of 0 to 1, the Beta distribution with parameters $\alpha > 0$ and $\beta > 0$ is a reasonable and quite flexible choice. Its density is given by

$$p(\theta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}, \tag{1.9}$$

with normalizing constant $B(\alpha, \beta) = \int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta$. Furthermore, its mean and variance are $E(\theta) = \alpha/(\alpha+\beta)$ and $\text{Var}(\theta) = \alpha\beta/((\alpha+\beta)^2(\alpha+\beta+1))$, respectively, and for $\alpha = \beta = 1$ the Beta distribution corresponds to the uniform distribution on [0, 1]. For example, if the researcher expects a 20 % complication rate with 0.1 standard error, then she solves $E(\theta) = 0.2$ and $\text{Var}(\theta) = 0.1^2$ for α and β and obtains $\alpha = 3$ and $\beta = 12$.

It can be shown that the posterior distribution is again Beta with parameters $\alpha_1 = \sum_{i=1}^n x_i + \alpha$ and $\beta_1 = n - \sum_{i=1}^n x_i + \beta$. The posterior mean then can be written similarly to (1.8) as a weighted average of the sample mean and the prior mean:

$$\frac{\alpha_1}{\alpha_1 + \beta_1} = \frac{\sum_{i=1}^n x_i + \alpha}{n + \alpha + \beta} = w\overline{x} + (1-w)\frac{\alpha}{\alpha+\beta},$$

where $w := w(\alpha, \beta, n) := \frac{n}{n+\alpha+\beta}$. As before, belief in the observed data increases as the number of subjects n increases.

Illustration 1.3 (Claim Numbers in Car Insurance) In car insurance, a good estimate of the expected number of claims is essential for adequate policy pricing. An insurance company here faces a two-way risk. Overestimation of the expected number of claims means too high premiums and therefore a loss of clients. Expecting too few claims however poses the risk of large losses in the portfolio. Assuming that an insurance company has a portfolio of n homogeneous policy holders, a common choice for the distribution of the number of claims $X_i, i = 1, \ldots, n$, is the Poisson distribution with mean and variance parameter θ and probability mass function

$$f(x_i|\theta) = \frac{\theta^{x_i}}{x_i!}e^{-\theta} \quad \text{for } x_i \in \{0, 1, 2, \ldots\}. \tag{1.10}$$

Even if the portfolio consists of rather homogeneous policy holders, there is significant uncertainty regarding the expected number of claims θ because it also depends on unobservable quantities such as risk affinity or exogenous risks like extreme weather events.

The insurance company decides to choose a Gamma prior distribution with parameters $\alpha > 0$ and $\beta > 0$, mean α/β, and density

$$p(\theta) = \frac{1}{\Gamma(\alpha)}\beta^{\alpha}\theta^{\alpha-1}e^{-\beta\theta}, \tag{1.11}$$

where $\Gamma(\alpha)$ is the Gamma function $\Gamma(\alpha) = \int_0^\infty \theta^{\alpha-1}e^{-\theta}d\theta$.

The posterior distribution (1.3) based on observations $x_1, \ldots, x_n$ from a previous year for example, is then obtained as follows:

$$p(\theta|\boldsymbol{x}) \propto \left[\prod_{i=1}^{n} \frac{\theta^{x_i}}{x_i!}e^{-\theta}\right]\frac{1}{\Gamma(\alpha)}\beta^{\alpha}\theta^{\alpha-1}e^{-\beta\theta} \propto \theta^{\alpha_1-1}e^{-\beta_1\theta},$$

which is again a Gamma distribution with parameters $\alpha_1 = \sum_{i=1}^n x_i + \alpha$ and $\beta_1 = n + \beta$. As before, the posterior mean can be decomposed into a weighted average of the empirical and prior mean. Such a convenient decomposition is however not always possible.

This mixture of Poisson and Gamma densities has another interesting interpretation: if the insurance company is interested in the claim number probabilities given an unknown parameter θ, Bayes' theorem can be "inverted" to compute the marginal density as $f(x_i) = f(x_i|\theta)p(\theta)/p(\theta|x_i)$ which results in a negative binomial distribution with the same mean as the Poisson distribution but with a higher variance due to the uncertainty in the unknown parameter.

1.1 From Non-informativeness to Conjugacy

Illustrations 1.1 and 1.2 also demonstrate a general problem of Bayesian statistics, namely the question: how do we choose an appropriate prior distribution? In certain applications, this choice might be evident but in general this is a non-trivial question and should be as objective as possible in order to not influence the results in an unwanted way. If for example in Illustration 1.1 the uncertainty in the prior knowledge τ^2 is very large, that is, the prior knowledge is rather vague, the prior will be close to $p(\boldsymbol{\theta}) \propto 1$ like the first prior choice in Illustration 1.2. Such a prior is called *non-informative* because it assigns equal likelihood to each possible parameter value.

One however has to be careful if the parameter space Θ is unbounded. In that case we have $\int_{\Theta} p(\boldsymbol{\theta})d\boldsymbol{\theta} = \infty$, and $p(\boldsymbol{\theta})$ is an improper prior.

Hence, such non-informative priors has to be dealt with care to ensure that the resulting posterior is proper. In Illustration 1.1, as $\tau^2 \to \infty$ corresponding to a non-informative prior, the posterior density is a normal density with mean $\overline{x}$ and variance $\frac{\sigma^2}{n}$, which is a proper distribution.

Another issue of non-informative priors is that they are not invariant under reparametrization of the model. For example a uniform prior on the success probability $\theta \in (0,1)$ (see Illustration 1.2) does not result in a uniform prior on the so-called *odds* parameter given by $\theta/(1-\theta)$. An alternative approach for defining non-informative priors which has this invariance property was developed by Jeffreys [21]. Jeffreys prior is given as

$$p(\boldsymbol{\theta}) \propto \left|I(\boldsymbol{\theta})\right|^{\frac{1}{2}},$$

where

$$I(\boldsymbol{\theta}) = E\left[-\frac{\partial^2 \ln f(\boldsymbol{X}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\bigg|\boldsymbol{\theta}\right] \tag{1.12}$$

is the expected *Fisher information matrix* about $\boldsymbol{\theta}$, which is a measure for the information about the parameter contained in the sample. In general, Jeffrey's approach leads to prior densities in the form of $p(\boldsymbol{\theta}) \propto 1$ for *location parameters* $\boldsymbol{\theta}$ and $p(\sigma) \propto \sigma^{-1}$ for *scale parameters* σ. For example the mean μ of a normal distribution is a location parameter and the standard error σ is a scale parameter.

On the other hand, the choice of an informative prior is always preferable if there is some kind of a priori knowledge about the parameter of interest. However, it will not be possible to get an analytically closed form expression of the posterior in complex situations, since the normalizing constant $f(\boldsymbol{x})$ defined in (1.4) of the posterior distribution requires a possibly high-dimensional integration. Posterior calculations are however simple if one considers *conjugate prior distributions*. A class of prior distributions $\mathcal{P}$ is conjugate to a class of observational models $\mathcal{F}$ if for every prior p out of $\mathcal{P}$ and for any observational distribution f from $\mathcal{F}$, the posterior distribution $p(\cdot|\boldsymbol{x})$ remains in the class of the prior distribution $\mathcal{P}$.

Example 1.4 (Conjugate Prior Distributions) The class of normal priors for the mean (Illustration 1.1) is conjugate for the observational model of normal distributions with known variance, while the class of Beta priors (Illustration 1.2) is conjugate for the observational model of Bernoulli distributions. Finally Illustration 1.3 also shows that the class of Gamma priors is conjugate for Poisson distributions.

1.2 Bayesian Inference

In Bayesian statistics all information about the parameter $\boldsymbol{\theta}$ is contained in the posterior distribution, while in classical statistics the information about $\boldsymbol{\theta}$ is captured by point and interval estimates. However, for the Bayesian, these quantities can be straightforwardly derived as well.

The main location measures are the *posterior mean*, as discussed in Illustrations 1.1–1.3, the *posterior median* and the *posterior mode*, where the last quantity is closest to the maximum likelihood principle from frequentist statistics, that is, the parameter $\boldsymbol{\theta}$ is most likely to be observed as judging from the available information contained in the observations. In maximum likelihood (ML) estimation we choose $\hat{\boldsymbol{\theta}}_{ML} = \text{argmax}_{\boldsymbol{\theta}\in\Theta}\, \ell(\boldsymbol{x}|\boldsymbol{\theta})$, while the posterior mode (PM) is augmented by the prior and given by $\hat{\boldsymbol{\theta}}_{PM} = \text{argmax}_{\boldsymbol{\theta}\in\Theta}\, \ell(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. Note that, for example, for normal distributions the mean, mode and median coincide, while this is in general not the case, such as for the Gamma distribution.

The main dispersion measures are the *variance*, *standard deviation* (square root of the variance), *precision* (inverse of the variance) and *interquartile range* (difference between 75 %- and 25 %-quantiles) of the posterior distribution. Corresponding to the Fisher information defined in (1.12), one also often considers the *posterior curvature at the mode* which is the matrix of second derivatives of the posterior density in log form at the mode. If $\boldsymbol{\theta}$ is a vector, marginal densities can also be assessed.

In addition to these Bayesian point estimates $100(1-\alpha)$ % *credible intervals* provide interval estimates for $\boldsymbol{\theta}$ and are given for a scalar parameter θ by an interval $I(\boldsymbol{x})$, depending on the observations $\boldsymbol{x}$, such that

$$\int_{I(\boldsymbol{x})} p(\theta|\boldsymbol{x})d\theta = 1 - \alpha.$$

In contrast to the confidence interval in classical statistics, the credible interval allows the interpretation that the parameter θ is contained with probability $1-\alpha$ in $I(\boldsymbol{x})$, since θ is now modeled as a random quantity.

Example 1.5 (Inference of the Normal Distribution) In Illustration 1.1 we have seen that the posterior distribution is given by the normal distribution with mean μ_1 and variance τ_1^2. Therefore the posterior mean, mode and median are μ_1, while the posterior variance is τ_1^2 and the posterior precision is τ_1^{-2}, which is also the posterior curvature at the mode.

A $100(1-\alpha)$ % credible interval $[\theta_l(\boldsymbol{x}), \theta_u(\boldsymbol{x})]$ for θ is given by appropriate quantiles of the posterior distribution: $\theta_l(\boldsymbol{x}) = \mu_1 - \tau_1\Phi^{-1}(1-\frac{\alpha}{2})$ and $\theta_u(\boldsymbol{x}) = \mu_1 + \tau_1\Phi^{-1}(1-\frac{\alpha}{2})$, where Φ^{-1} is the inverse of the standard normal distribution function. This is also the shortest possible credible interval. Note that the corresponding classical $100(1-\alpha)$ % and confidence interval for θ is given by $\bar{x} \pm \frac{s}{\sqrt{n}}\Phi^{-1}(1-\frac{\alpha}{2})$ where $s^2 := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is the sample variance.

Returning to the specific example of Illustration 1.1, a corresponding 95 % credible interval for the mean yield is [10.303, 13.697], while a 95 % confidence interval is [9.040, 12.960] when assuming a sample variance of $s^2 = 5$. From the Bayesian theory the agronomist can say that the mean yield is between 10.303 and 13.697 with 95 % probability. The frequentist approach gives that the random interval $\bar{x} \pm \frac{s}{\sqrt{n}}\Phi^{-1}(1-\frac{\alpha}{2})$ covers the mean in 95 % of times. For the specific observations this interval is given by [9.040, 12.960].

1.3 Conjugacy and Regression Models

Before closing this section we consider the problem of modeling the influence of potential *explanatory variables* on a risk quantity called response. The simplest such model is the *linear regression model* for the *response* vector $\boldsymbol{Y} = (Y_1, \dots, Y_n)'$:

$$Y_i \sim N\big(x_{i1}\beta_1 + \cdots + x_{id}\beta_d, \sigma^2\big) \quad \text{independent for } i = 1, \dots, n, \tag{1.13}$$

where $x_{i1}, \dots, x_{id}$ are known values of d explanatory variables for the ith observation and $\beta_1, \dots, \beta_d$ are unknown *regression coefficients*. We can rewrite this model in matrix form as follows:

$$\boldsymbol{Y} \sim N_n\big(X\boldsymbol{\beta}, \sigma^2 I_n\big), \tag{1.14}$$

where $N_n(\boldsymbol{\mu}, \Sigma)$ denotes the n-dimensional normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. Further we define

$$X := \begin{pmatrix} x_{11} & \dots & x_{1d} \\ \vdots & & \vdots \\ x_{n1} & \dots & x_{nd} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\beta} := \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_d \end{pmatrix}.$$

The matrix X is called the *design matrix* and we assume that its columns are not linearly dependent.

Applications of such models can be found in virtually all areas of scientific research. For example, in Illustration 1.1 the agronomist may also try to model the crop yields with respect to a set of explanatory variables such as rainfall or sunshine duration. An experienced agronomist may have some prior expert knowledge about the effect of these variables and therefore can choose appropriate prior distributions for the regression coefficients. Similarly, based on her experience she may also be able to specify a prior for the variance parameter of the model parameters.

In model (1.14) it is more convenient to formulate priors in terms of $\boldsymbol{\beta}$ and the precision $\phi := \sigma^{-2}$. A typical choice is the Normal-Gamma, NG$(\boldsymbol{b}_0, B_0, n_0, S_0)$, prior, which is, for known constants n_0 and S_0, known vector $\boldsymbol{b}_0$ and known matrix B_0, defined in a hierarchical way as

$$\boldsymbol{\beta}|\phi \sim N_d\left(\boldsymbol{b}_0, \frac{B_0}{\phi}\right) \quad \text{and} \quad \phi \sim \text{Gamma}\left(\frac{n_0}{2}, \frac{n_0 S_0}{2}\right). \tag{1.15}$$

Equivalently we can assume $\boldsymbol{\beta}|\sigma^2 \sim N_d(\boldsymbol{b}_0, \sigma^2 B_0)$ and $\sigma^2 \sim$ Inverse Gamma$(\frac{n_0}{2}, \frac{n_0 S_0}{2})$. Here the Inverse Gamma distribution is derived as follows: if $X \sim$ Gamma(α, β) then $1/X \sim$ Inverse Gamma(α, β). Under this setup the following theorem holds:

Theorem 1.6 (Conjugacy in Regression) *For the linear model given in* (1.14) *with observed response* $\mathbf{y}$ *and prior distribution given by* (1.15) *the posterior distribution* *of* $(\boldsymbol{\beta}, \phi)$ *is given by an* $NG(\boldsymbol{b}_1, B_1, n_1, S_1)$ *distribution with*

$$\boldsymbol{b}_1 = B_1\big(B_0^{-1}\boldsymbol{b}_0 + X'\boldsymbol{y}\big), \qquad B_1 = \big(B_0^{-1} + X'X\big)^{-1},$$

$$n_1 = n_0 + n, \qquad S_1 = \frac{1}{n_1}\big[n_0 S_0 + (\boldsymbol{y} - X\boldsymbol{b}_1)'y + (\boldsymbol{b}_0 - \boldsymbol{b}_1)'B_0^{-1}\boldsymbol{b}_0\big].$$

See Gamerman and Lopes [4, Sect. 2.3.2] for a proof.

2 MCMC—Markov Chain Monte Carlo

In Sect. 1.1 we studied the choice of prior distributions. In particular, we discussed non-informative priors and conjugate families which allow for an easy derivation of the posterior distribution (1.3). This is however not the case in general. *Markov Chain Monte Carlo* (MCMC) methods are used to approximate the posterior in more complex situations. Although being very computer intensive, the increasing availability of computing power nowadays makes the use of MCMC methods increasingly attractive. In particular, MCMC methods may be used to solve complex problems which cannot be treated using classical statistics. Examples of such problems are spam e-mail filtering and the analysis of gene expression data, just to name a few.

MCMC methods are based on the two well-known concepts of *Markov Chains* and *Monte Carlo* techniques. Both concepts will be explained first, before we then introduce the two most commonly used algorithms, namely the Gibbs sampler and Metropolis Hastings algorithms. Recent comprehensive references on MCMC methods include Gamerman and Lopes [4] and Marin and Robert [22].

2.1 ∗∗MC—Monte Carlo

To understand MCMC methods, we begin with the second "MC" which refers to "Monte Carlo" and which is due to the often used *Monte Carlo integration* techniques. In general Monte Carlo methods repeatedly sample from a probability distribution to determine analytically difficult quantities. For example, let us assume that $t(\cdot)$ is a function and we are interested in computing the integral

$$I = \int_0^1 t(\theta)d\theta, \tag{2.1}$$

of which no closed form solution is known. This is, for example, often the case for the marginal density function f defined in (1.4) which is part of the posterior distribution defined in (1.3). For such problems we use the following numerical approximation. First let $\theta \in (0,1)$ be a random variable with density p. Then the expectation of the random variable $t(\theta)$ is $E(t(\theta)) = \int_0^1 t(\theta)p(\theta)d\theta$. If we can sample from p, an estimate of $E(t(\theta))$ is the sample mean. In particular, let θ be uniform on $(0, 1)$ and $\theta_1, \ldots, \theta_n$ a corresponding independent and identically distributed (i.i.d.) random sample. Then (2.1) can be estimated by

$$\hat{I} := \frac{1}{n} \sum_{i=1}^{n} t(\theta_i). \tag{2.2}$$

By the strong law of large numbers (see Durrett [12]) $\hat{I}$ converges to $I = E(t(\theta))$ with probability 1, since $p(\theta) = 1$ for all $\theta \in (0, 1)$.

In Bayesian statistics the posterior expectation $E(t(\boldsymbol{\theta})|\boldsymbol{x})$ can be estimated by the sample mean (2.2) when $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ is a sample from the posterior distribution $p(\cdot|\boldsymbol{x})$. As long as the posterior distribution and sampling algorithms are available, there are no problems and the first "MC" referring to "Markov chain" is not needed.

As mentioned above, it is unfortunately not the case that an analytical form of the posterior density $p(\cdot|\boldsymbol{x})$ is always available. The idea of MCMC methods therefore is to construct a Markov chain with limiting distribution $p(\cdot|\boldsymbol{x})$. If the Markov chain is run for a sufficiently long time, it can be assumed that the stationary state is reached and therefore the realizations of the chain represent a sample from $p(\cdot|\boldsymbol{x})$. In the following section we therefore give a brief overview of Markov chain theory. Readers familiar with it can skip Sect. 2.2 and continue reading with Sects. 2.3 and 2.4 which discuss the two most common MCMC methods.

2.2 MC∗∗—Markov Chains

We give a short introduction to Markov chains and state major results. A more detailed treatment can be found in Meyn and Tweedie [24], Nummelin [25], Resnick [26] and Guttorp [18]. The set of random variables $\{\boldsymbol{\theta}^{(t)} : t \in T\}$ is said to be a stochastic process taking values in the state space S for time points t in the index set T. In our discussion we will only consider discrete time stochastic processes with T being the set of natural numbers $\mathbb{N} = \{1, 2, \ldots\}$. The state space S can generally be a subset of the d-dimensional set of real numbers, $\mathbb{R}^d$, but in the following we will concentrate on a discrete state space S. Details on continuous state space Markov chains can be found in Meyn and Tweedie [24].

A *Markov chain* is a process, such that given the present state, past and future states are independent:

$$\begin{aligned} P\big(\boldsymbol{\theta}^{(n+1)} = \boldsymbol{x}_{n+1} | \boldsymbol{\theta}^{(n)} = \boldsymbol{x}_n, \boldsymbol{\theta}^{(n-1)} = \boldsymbol{x}_{n-1}, \ldots, \boldsymbol{\theta}^{(0)} = \boldsymbol{x}_0\big) \\ = P\big(\boldsymbol{\theta}^{(n+1)} = \boldsymbol{x}_{n+1} | \boldsymbol{\theta}^{(n)} = \boldsymbol{x}_n\big) \end{aligned} \tag{2.3}$$

for all $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{n+1} \in S$. If the probabilities in (2.3) do not depend on n, we say that the Markov chain is *homogenous*. In this case we define the *transition probability* $P(\boldsymbol{x}, \boldsymbol{y})$ of moving from state $\boldsymbol{x}$ to state $\boldsymbol{y}$ as:

$$P(\boldsymbol{x}, \boldsymbol{y}) := P\big(\boldsymbol{\theta}^{(n+1)} = \boldsymbol{y} | \boldsymbol{\theta}^{(n)} = \boldsymbol{x}\big).$$

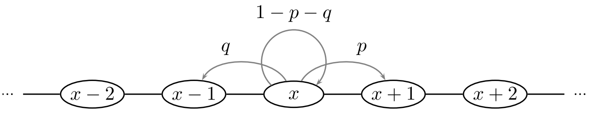


Fig. 3 Probabilities of molecule movement

In general, for $A \subset S$, $P(\boldsymbol{x}, A) := \sum_{\boldsymbol{y}\in A} P(\boldsymbol{x}, \boldsymbol{y})$ is called the *transition kernel.*

Illustration 2.1 (Molecule Movement) Consider a molecule traveling in a liquid or a gas which moves independently left and right with successive displacements from its current position governed by a probability function f over the integers, that is $S = \mathbb{Z}$. Such a process is called a *random walk*. Let $\theta^{(n)}$ represent the position of the molecule at time n. Therefore we have

$$\theta^{(n)} = \theta^{(n-1)} + w_n = \theta^{(0)} + w_1 + \cdots + w_n,$$

where $w_i \sim f$ independently and for all $i \geq 1$. For the initial position $\theta^{(0)}$ we assume an initial distribution $\pi^{(0)}$.

The case where the probabilities of right, left or stay move are given by p, q and $1 - p - q$, respectively, is represented by assuming $f(1) = p$, $f(-1) = q$ and $f(0) = 1 - p - q$. This implies that

$$P(x, y) = P\big(\theta^{(n+1)} = y|\theta^{(n)} = x\big) = \begin{cases} p, & \text{if } y = x + 1, \\ q, & \text{if } y = x - 1, \\ 1 - p - q, & \text{if } y = x, \\ 0, & \text{if } y \neq x - 1, x, x + 1, \end{cases}$$

which is illustrated in Fig. 3.

If the state space $S \subset \mathbb{R}^d$ is not only discrete but also finite, that is $S = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_r\}$, we can consider the *transition matrix* P defined by

$$P := \begin{pmatrix} P(\boldsymbol{x}_1, \boldsymbol{x}_1) & \ldots & P(\boldsymbol{x}_1, \boldsymbol{x}_r) \\ \vdots & & \vdots \\ P(\boldsymbol{x}_r, \boldsymbol{x}_1) & \ldots & P(\boldsymbol{x}_r, \boldsymbol{x}_r) \end{pmatrix}.$$

Higher order transition probabilities P^m for $m \geq 2$ can be obtained as follows

$$\begin{aligned} P^m(\boldsymbol{x}, \boldsymbol{y}) &:= P\big(\boldsymbol{\theta}^{(m)} = \boldsymbol{y}|\boldsymbol{\theta}^{(0)} = \boldsymbol{x}\big) \\ &= \sum_{\boldsymbol{x}_1\in S} \cdots \sum_{\boldsymbol{x}_{m-1}\in S} P\big(\boldsymbol{\theta}^{(m)} = \boldsymbol{y}, \boldsymbol{\theta}^{(m-1)} = \boldsymbol{x}_{m-1}, \ldots, \boldsymbol{\theta}^{(1)} = \boldsymbol{x}_1|\boldsymbol{\theta}^{(0)} = \boldsymbol{x}\big) \\ &= \sum_{\boldsymbol{x}_1\in S} \cdots \sum_{\boldsymbol{x}_{m-1}\in S} P\big(\boldsymbol{\theta}^{(m)} = \boldsymbol{y}|\boldsymbol{\theta}^{(m-1)} = \boldsymbol{x}_{m-1}\big) \cdots P\big(\boldsymbol{\theta}^{(1)} = \boldsymbol{x}_1|\boldsymbol{\theta}^{(0)} = \boldsymbol{x}\big) \\ &= \sum_{\boldsymbol{x}_1\in S} \cdots \sum_{\boldsymbol{x}_{m-1}\in S} P(\boldsymbol{x}, \boldsymbol{x}_1)P(\boldsymbol{x}_1, \boldsymbol{x}_2) \cdots P(\boldsymbol{x}_{m-1}, \boldsymbol{y}), \end{aligned}$$

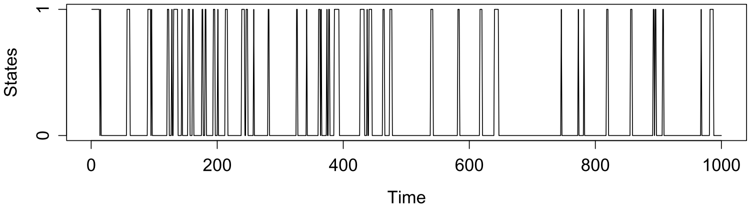


Fig. 4 Example of health states (0 = healthy, 1 = sick) of a policy holder over time ($p = 0.05$, $q = 0.3$, $\pi^{(0)}(0) = 0.8$, $\pi^{(0)}(1) = 0.2$)

where the second equality is due to the Markov property (2.3). In matrix notation we have $P^m = P \cdots P$ meaning matrix multiplication m times of the matrix P.

Further, let $\pi^{(0)}$ be the initial distribution of the chain, $\pi^{(0)}(\boldsymbol{x}) := P(\boldsymbol{\theta}^{(0)} = \boldsymbol{x})$. The marginal distribution after n time steps is given by

$$\begin{aligned}\pi^{(n)}(\boldsymbol{y}) := P\big(\boldsymbol{\theta}^{(n)} = \boldsymbol{y}\big) &= \sum_{\boldsymbol{x}\in S} P\big(\boldsymbol{\theta}^{(n)} = \boldsymbol{y} | \boldsymbol{\theta}^{(0)} = \boldsymbol{x}\big) P\big(\boldsymbol{\theta}^{(0)} = \boldsymbol{x}\big)\\ &= \sum_{\boldsymbol{x}\in S} P^n(\boldsymbol{x}, \boldsymbol{y})\pi^{(0)}(\boldsymbol{x}), \qquad (2.4)\end{aligned}$$

which can also be written as $\pi^{(n)} = \pi^{(0)}P^n = \pi^{(0)}P^{n-1}P = \pi^{(n-1)}P$.

Before we move on to discuss some major results which are the basis of MCMC methods, we consider an illustrative example.

Illustration 2.2 (Daily Allowance in Health Insurance) A health insurance company sells policies which pay a daily allowance to sick policy holders. In order to price the policies, the company sets up the following simplifying model. The health state of a person is modeled as a Markov chain $\{\theta^{(n)} : n \geq 0\}$ with states $S = \{\mathit{healthy}, \mathit{sick}\}$, denoted as $S = \{0, 1\}$, respectively. The initial distribution (the proportions of healthy and sick policy holders when the policy is sold) is denoted by $\pi^{(0)} = (\pi^{(0)}(0), \pi^{(0)}(1))'$ and the transition matrix P by

$$P = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix} = \begin{pmatrix} P(0,0) & P(0,1) \\ P(1,0) & P(1,1) \end{pmatrix}.$$

That is, a healthy policy holder today is assumed to fall ill tomorrow with a probability of p versus staying healthy with a probability of $1-p$. Similarly, a sick policy holder becomes healthy with a probability of q and stays sick with a probability of $1-q$. An exemplary realization of this Markov chain is shown in Fig. 4

The probability that a person is healthy after n days (independent of whether or not he or she was sick in the meantime) is given by

$$\begin{aligned}P\big(\theta^{(n)} = 0\big) = P\big(\theta^{(n)} = 0 | \theta^{(n-1)} = 0\big) P\big(\theta^{(n-1)} = 0\big)\\ + P\big(\theta^{(n)} = 0 | \theta^{(n-1)} = 1\big) P\big(\theta^{(n-1)} = 1\big)\end{aligned}$$

$$\begin{aligned}
&= (1-p)P\big(\theta^{(n-1)}=0\big) + qP\big(\theta^{(n-1)}=1\big)\\
&= (1-p-q)P\big(\theta^{(n-1)}=0\big) + q\\
&= (1-p-q)\big[(1-p-q)P\big(\theta^{(n-2)}=0\big) + q\big] + q\\
&\vdots\\
&= (1-p-q)^n\pi^{(0)}(0) + q\sum_{k=0}^{n-1}(1-p-q)^k.
\end{aligned}$$

If $p = q = 0$, that is, healthy (sick) persons always stay healthy (sick), then $P(\theta^{(n)} = 0) = \pi^{(0)}(0)$ and $P(\theta^{(n)} = 1) = \pi^{(0)}(1)$. If $p + q > 0$, using results for the finite geometric series gives

$$\begin{aligned}
P\big(\theta^{(n)}=0\big) &= (1-p-q)^n\pi^{(0)}(0) + q\frac{1-[(1-p-q)^n]}{1-(1-p-q)}\\
&= (1-p-q)^n\left[\pi^{(0)}(0) - \frac{q}{p+q}\right] + \frac{q}{p+q}. \qquad (2.5)
\end{aligned}$$

If the initial distribution is given by $\pi^{(0)} = (\frac{q}{p+q}, \frac{p}{p+q})'$, then the marginal probability $P(\theta^{(n)} = 0) = \frac{q}{p+q}$ is the same for all time points n.

If $p + q < 2$, then $(1 - p - q)^n$ converges to zero as n goes to infinity and therefore

$$\lim_{n\to\infty} P\big(\theta^{(n)}=0\big) = \frac{q}{p+q} \quad \text{and} \quad \lim_{n\to\infty} P\big(\theta^{(n)}=1\big) = \frac{p}{p+q},$$

which shows that the initial distribution is obtained as the limiting distribution of the Markov chain. For the realizations of the Markov chain shown in Fig. 4 the convergence is illustrated in Table 1.

To obtain the probability that an initially healthy policy holder is also healthy after n days, denoted by $P^n(0, 0)$, we assume that we always start in the healthy state, that is $\pi^{(0)}(0) = 1$. Using (2.4) with $\pi^{(0)}(0) = 1$ this gives

$$\begin{aligned}
P^n(0,0) = P\big(\theta^{(n)}=0\big) &= (1-p-q)^n\left(1-\frac{q}{p+q}\right) + \frac{q}{p+q}\\
&= (1-p-q)^n\frac{p}{p+q} + \frac{q}{p+q}.
\end{aligned}$$

Similarly, we compute $P^n(1, 0)$, $P^n(0, 1)$ and $P^n(1, 1)$ to determine the nth order transition matrix P^n as

$$P^n = \frac{(1-p-q)^n}{p+q}\begin{pmatrix} p & -q \\ -q & q \end{pmatrix} + \frac{1}{p+q}\begin{pmatrix} q & p \\ q & p \end{pmatrix}. \qquad (2.6)$$

Table 1 Empirical marginal probabilities after different time points of the Markov chain shown in Fig. 4

Health states	Time										Limit prob.
	100	200	300	400	500	600	700	800	900	1000	
0	0.75	0.76	0.78	0.79	0.79	0.82	0.83	0.84	0.85	0.86	0.86
1	0.25	0.24	0.22	0.21	0.21	0.18	0.17	0.16	0.15	0.14	0.14

Finally, we denote by T_0 the first time that a person becomes healthy again. Given that he or she was healthy when taking out the policy, we have

$$\begin{aligned} P\bigl(T_0 = n|\theta^{(0)} = 0\bigr) &= P_0\bigl(\theta^{(n)} = 0, \theta^{(j)} \neq 0, 1 \leq j \leq n-1\bigr) \\ &= P(0,1)P(1,1)^{n-2}P(1,0) = p(1-q)^{n-2}q. \end{aligned}$$

Similarly let T_1 be the first time that a person falls ill. Then it holds that $P(T_1 = n|\theta^{(0)} = 0) = P(0,0)^{n-1}P(0,1) = p(1-p)^{n-1}$.

A fundamental problem for Markov chains in the context of simulation is the study of the asymptotic behavior of the chain as the number of steps or iterations n goes to infinity. A key concept for this is the *stationary distribution* π, which satisfies

$$\sum_{\boldsymbol{x} \in S} \pi(\boldsymbol{x}) P(\boldsymbol{x}, \boldsymbol{y}) = \pi(\boldsymbol{y}) \quad \forall \boldsymbol{y} \in S, \tag{2.7}$$

and can be written in matrix notation as $\pi = \pi P$. The reason for the name is clear from the above equation. If the marginal distribution at any step n is π, then the distribution of the next step is πP. Once the chain reaches a stage where π is the distribution of the chain, the chain retains this distribution for all subsequent stages.

Illustration 2.3 (Illustration 2.2 Continued) Since the policies sold by the health insurance company are valid for the full lifetime of a policy holder, the company would like to investigate the long term expected proportions of healthy and sick persons. Since $S = \{0, 1\}$, in this case condition (2.7) is equivalent to

$$\pi(0)P(0,y) + \pi(1)P(1,y) = \pi(y), \quad y = 0, 1.$$

The solution is $\pi = (\frac{q}{p+q}, \frac{p}{p+q})$. Also for $p + q < 2$ it follows from (2.6) that

$$\lim_{n\to\infty} P^n = \frac{1}{p+q}\begin{pmatrix} q & p \\ q & p \end{pmatrix} = \begin{pmatrix} \pi(0) & \pi(1) \\ \pi(0) & \pi(1) \end{pmatrix}$$

and the distribution of $\theta^{(n)}$ converges to π at an exponential rate. This shows that for $p + q < 2$ the proportion of healthy and sick policy holders is asymptotically given by the stationary distribution π.

The case $p+q=2$ still produces a stationary distribution π but this does not provide a unique limiting distribution since from (2.5) it follows that

$$P\big(\theta^{(n)}=0\big)=(-1)^n\left(\pi^{(0)}(0)-\frac{q}{2}\right)+\frac{q}{2} \quad \forall n\geq 1.$$

This case is somewhat different, since the states are always alternating over time corresponding to the case that persons are healthy one day and always fall ill the next day which is evidently rather unrealistic. The chain has a periodic nature that will be addressed below.

Having established some basic properties of Markov chains, we are interested in characterizing the limiting behavior. For this a classification of the states of the Markov chain is necessary. For a more complete treatment see for example Chap. 2 of Resnick [26]. We define the first visit time to $\boldsymbol{y}$ as $T_{\boldsymbol{y}}=\inf\{n\geq 1:\boldsymbol{\theta}^{(n)}=\boldsymbol{y}\}$ and the probability of visiting $\boldsymbol{y}$ after starting in $\boldsymbol{x}$ in finite time by $\rho_{\boldsymbol{xy}}:=P(T_{\boldsymbol{y}}<\infty|\boldsymbol{\theta}^{(0)}=\boldsymbol{x})$. Then a state $\boldsymbol{y}\in S$ is *recurrent* if and only if $\rho_{\boldsymbol{yy}}=1$, and—more strongly—*positive recurrent* if and only if $\boldsymbol{y}$ is recurrent and $E(T_{\boldsymbol{y}}|\boldsymbol{\theta}^{(0)}=\boldsymbol{y})<\infty$. Further, the state $\boldsymbol{x}$ is said *to hit* $\boldsymbol{y}$ or $\boldsymbol{y}$ is *accessible* from $\boldsymbol{x}$, denoted by $\boldsymbol{x}\rightarrow\boldsymbol{y}$ if and only if $\rho_{\boldsymbol{xy}}>0$. One can show that $\boldsymbol{x}\rightarrow\boldsymbol{y}$ if and only if there exist an $n\geq 0$ such that $P^n(\boldsymbol{x},\boldsymbol{y})>0$ (see Resnick [26, p. 78]). Let $\boldsymbol{x}\leftrightarrow\boldsymbol{y}$ if and only if $\boldsymbol{x}\rightarrow\boldsymbol{y}$ and $\boldsymbol{y}\rightarrow\boldsymbol{x}$. This is an equivalence relationship. The Markov chain is called *irreducible* if $\boldsymbol{x}\rightarrow\boldsymbol{y}$ for every pair $\boldsymbol{x},\boldsymbol{y}\in S$.

Finally, to establish limit distributions one also needs to introduce the notation of periodicity. The *period* of state $\boldsymbol{x}$ is given by

$$d_{\boldsymbol{x}}=\text{largest common divisor of }\big\{n\geq 1:P^n(\boldsymbol{x},\boldsymbol{x})>0\big\}.$$

It follows that the condition $P(\boldsymbol{x},\boldsymbol{x})>0$ implies $d_{\boldsymbol{x}}=1$. Such a state is called *aperiodic*. Thus the states 0 and 1 in Illustration 2.3 are aperiodic if $p+q<2$. On the other hand, if $p+q=2$, it holds that $d_0=d_1=2$, in other words, the states 0 and 1 are periodic with period 2.

A state $\boldsymbol{x}$ is called *ergodic* if it is aperiodic and positive recurrent. Similarly, a Markov chain is called *ergodic* if all states are aperiodic and positive recurrent. These concepts are sufficient to characterize the limiting distribution.

Theorem 2.4 (Limiting Distribution) *Let* $\{\boldsymbol{\theta}^{(n)},n\geq 0\}$ *be an irreducible and ergodic Markov chain with stationary distribution* π, *then*

$$\lim_{n\to\infty}P^n(\boldsymbol{x},\boldsymbol{y})=\pi(\boldsymbol{y}) \quad \forall\boldsymbol{x},\boldsymbol{y}\in S.$$

A proof can be found in Guttorp [18, Theorem 2.9]. This shows that the stationary distribution is also the limiting distribution under the assumptions of Theorem 2.4.

While the empirical mean converges to the population mean as the sample size increases for i.i.d. samples by the strong law of large numbers, a Markov chain equivalent will now be given.

Theorem 2.5 (Ergodic Theorem) *If the chain is ergodic and* $E_\pi(t(\boldsymbol{\theta})) < \infty$ *for the unique limiting distribution* π *then*

$$\bar{t}_n := \frac{1}{n}\sum_{i=1}^{n} t\left(\boldsymbol{\theta}^{(i)}\right) \xrightarrow{n\to\infty} E_\pi\left(t(\boldsymbol{\theta})\right) \quad \textit{with probability } 1.$$

A proof can be found on page 49 of Guttorp [18]. This theorem can be used as justification for using $\bar{t}_n$ as an estimate for $E_\pi(t(\boldsymbol{\theta}))$, see also the discussion in Sect. 2.1. A central limit theorem for Markov chains can also be formulated and is found for example in Gilks, Richardson, and Spiegelhalter [17]. It can be used for constructing asymptotic confidence intervals.

Having established the asymptotic theory of Markov chains, the final, and crucial, step is simulation. For this, consider an ergodic Markov chain $\{\boldsymbol{\theta}^{(n)}, n \geq 0\}$ with state space $S \subset \mathbb{R}^d$, transition probabilities $P(x, y)$ and initial distribution $\pi^{(0)}$. To generate values from this Markov chain the following algorithm can be used.

- Sample a starting value $\boldsymbol{\theta}^{(0)}$ from the initial distribution $\pi^{(0)}$.
- For $i = 1, \ldots, n$, sample value $\boldsymbol{\theta}^{(i)}$ from the probability mass function $f(\cdot) := P(\boldsymbol{\theta}^{(i-1)}, \cdot)$.

As n gets large the sampled values will have a distribution close to the limiting distribution π and can therefore be considered as an approximate sample from π. Note that all samples drawn after convergence are also samples from π since it is the stationary distribution. Here, convergence of a Markov chain means that the stationary distribution is approximated sufficiently accurately, which is difficult to assess. Relevant references will be given below. The values before convergence are called the *burn-in period* and will be deleted when considering the ergodic averages such as $\bar{t}_n$. The sampled values are dependent, since they arise from a Markov chain, however so-called thinning and batching methods can be applied to achieve an approximately i.i.d. sample. This general method of approximate sampling from the stationary distribution is called the *Markov Chain Monte Carlo* (MCMC) approach.

We can now use this approach to draw approximate samples from a complex posterior distribution $p(\cdot|\boldsymbol{x})$, which is analytically not tractable, by assuming that $p(\cdot|\boldsymbol{x})$ is the stationary distribution π of a Markov chain. The next two sections will study two famous MCMC algorithms in detail.

2.3 Gibbs Sampler

This chapter introduces and discusses the first widely used sampling scheme for constructing a Markov chain with prespecified limiting distribution π. It was first developed for approximately sampling from the *Gibbs distribution* used in image analysis. Geman and Geman [16] discussed this problem for several sampling schemes. Gelfand and Smith [14] were the first to point out to the statistical community at large that this sampling scheme could be used for other distributions than the Gibbs distribution. Before stating the sampling algorithm, we consider a small illustrative example.

Illustration 2.6 (Health States of a Couple) (Casella and George [2]) Let $S = \{(0,0)', (1,0)', (0,1)', (1,1)'\}$ be a two-dimensional state space with probability distribution π for the random vector $\boldsymbol{\theta} = (\theta_1, \theta_2)'$ given by

$$\begin{aligned} P(\theta_1 = 0, \theta_2 = 0) &= \pi_{00}, \qquad & P(\theta_1 = 0, \theta_2 = 1) &= \pi_{01}, \\ P(\theta_1 = 1, \theta_2 = 0) &= \pi_{10}, \qquad & P(\theta_1 = 1, \theta_2 = 1) &= \pi_{11}. \end{aligned} \tag{2.8}$$

In view of Illustration 2.2 this can be interpreted as the healthy and sick states of a married couple. For example, if the first component corresponds to the health state of the husband and the second to that of his wife, then $\theta_1 = 1$ and $\theta_2 = 0$ indicates that the husband is sick, while his wife is healthy.

The Markov chain now consists of a bivariate vector $\boldsymbol{\theta}^{(n)} = (\theta_1^{(n)}, \theta_2^{(n)})'$ and the following transition probabilities are assumed.

- For $\theta_1^{(n)}$ the probability of moving from $\theta_2^{(n-1)} = j$ to $\theta_1^{(n)} = 0$ and $\theta_1^{(n)} = 1$, respectively, is given by

$$\pi_1(0|j) = \frac{\pi_{0j}}{\pi_{0j} + \pi_{1j}} \quad \text{and} \quad \pi_1(1|j) = \frac{\pi_{1j}}{\pi_{0j} + \pi_{1j}}. \tag{2.9}$$

  Note that $\pi_1(\cdot|j)$ is the conditional probability function of θ_1 given $\theta_2 = j$, $j = 0, 1$.

- For $\theta_2^{(n)}$ the probability of moving from $\theta_1^{(n)} = i$ to $\theta_2^{(n)} = 0$ and $\theta_2^{(n)} = 1$, respectively, is given by

$$\pi_2(0|i) = \frac{\pi_{i0}}{\pi_{i0} + \pi_{i1}} \quad \text{and} \quad \pi_2(1|i) = \frac{\pi_{i1}}{\pi_{i0} + \pi_{i1}}. \tag{2.10}$$

  Note that $\pi_2(\cdot|i)$ is the conditional probability function of θ_2 given $\theta_1 = i$, $i = 0, 1$.

This means that the husband's health state depends on his wife's yesterday's state and today's health state of the wife depends on today's health state of the husband. For a transition from state (i, j) yesterday to state (k, l) today we have

$$\theta_2^{(n-1)} = j \quad \xrightarrow{\frac{\pi_{kj}}{\pi_{0j}+\pi_{1j}}} \quad \theta_1^{(n)} = k \quad \xrightarrow{\frac{\pi_{kl}}{\pi_{k0}+\pi_{k1}}} \quad \theta_2^{(n)} = l.$$

Therefore the overall transition probability is given by

$$\begin{aligned} P\big((i,j),(k,l)\big) &= P\big(\boldsymbol{\theta}^{(n)} = (k,l) | \boldsymbol{\theta}^{(n-1)} = (i,j)\big) \\ &= P\big(\theta_2^{(n)} = l | \theta_1^{(n)} = k\big) P\big(\theta_1^{(n)} = k | \theta_2^{(n-1)} = j\big) \\ &= \frac{\pi_{kl}}{\pi_{k0} + \pi_{k1}} \frac{\pi_{kj}}{\pi_{0j} + \pi_{1j}}, \end{aligned}$$

for $(i, j), (k, l) \in S$. Thus a 4×4 transition matrix P can be formed.

One can further show that $\{\boldsymbol{\theta}^{(n)} = (\theta_1^{(n)}, \theta_2^{(n)})', n \geq 0\}$ forms a Markov chain and that π defined in (2.8) is the stationary distribution of the chain. If all elements of π are positive, it is also a limiting distribution. In particular, chains formed by the superposition of the conditional distributions have a stationary distribution given by the joint distribution.

Illustration 2.6 can easily be extended to the case where $\boldsymbol{\theta}$ consists of d components with $m_1, \dots, m_d$ values.

In general, *Gibbs sampling* is an MCMC scheme where the transition probabilities are formed by the full conditional distributions. Assume as before that the distribution of interest is $\pi(\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_d)'$. Each of the d components can be a scalar, vector or matrix. Further assume that for each $i \in \{1, \dots, n\}$ the *full conditional distribution* for $\boldsymbol{\theta}_i$

$$\pi_i^{FC}(\boldsymbol{\theta}_i) := \pi(\boldsymbol{\theta}_i | \boldsymbol{\theta}_{-i}) \quad \text{where } \boldsymbol{\theta}_{-i} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{i-1}, \boldsymbol{\theta}_{i+1}, \dots, \boldsymbol{\theta}_d)'$$

is known and can be sampled, for example, using Eqs. (2.9) and (2.10) in the above example. The Gibbs sampling algorithm can now be described as follows.

1. Set the iteration counter to $j = 1$ and set initial values $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\theta}_1^{(0)}, \dots, \boldsymbol{\theta}_d^{(0)})'$.
2. Obtain a new value $\boldsymbol{\theta}^{(j)} = (\boldsymbol{\theta}_1^{(j)}, \dots, \boldsymbol{\theta}_d^{(j)})'$ through successive generation of values

$$\begin{aligned} \boldsymbol{\theta}_1^{(j)} &\sim \pi\big(\boldsymbol{\theta}_1 \big| \boldsymbol{\theta}_2^{(j-1)}, \dots, \boldsymbol{\theta}_d^{(j-1)}\big), \\ \boldsymbol{\theta}_2^{(j)} &\sim \pi\big(\boldsymbol{\theta}_2 \big| \boldsymbol{\theta}_1^{(j)}, \boldsymbol{\theta}_3^{(j-1)}, \dots, \boldsymbol{\theta}_d^{(j-1)}\big), \\ &\vdots \\ \boldsymbol{\theta}_d^{(j)} &\sim \pi\big(\boldsymbol{\theta}_d \big| \boldsymbol{\theta}_1^{(j)}, \dots, \boldsymbol{\theta}_{d-1}^{(j)}\big). \end{aligned}$$

3. Change counter j to $j + 1$ and return to step 2 until convergence is reached.

When convergence is reached the resulting value $\boldsymbol{\theta}^{(j)}$ is a draw from π. Often convergence is assessed by choosing an error bound $\varepsilon > 0$ and assuming convergence when the distance between $\boldsymbol{\theta}^{(n+1)}$ and $\boldsymbol{\theta}^{(n)}$ is less than ε. A further example is given in the following.

Illustration 2.7 (Coal Mining Disasters) Carlin, Gelfand, and Smith [10] discuss the following problem: yearly numbers $Y_1, \dots, Y_M$ of British coal-mining disasters as measured over more than a century are unlikely to have stayed at a similar level due to better technology and increased safety requirements. It is therefore reasonable to assume the presence of a change point $m \in \{1, \dots, M\}$ at which the general level of disasters significantly changed. Therefore Carlin et al. [10] assume the number of coal-mining disasters before that (unknown) change point to be Poisson distributed with another intensity parameter than after. They consider the following hierarchical model:

$$\begin{aligned} Y_i|\lambda, m &\sim \text{Poisson}(\lambda) \quad \text{for } i = 1, \dots, m \text{ (independent)}, \\ Y_i|\phi, m &\sim \text{Poisson}(\phi) \quad \text{for } i = m+1, \dots, M \text{ (independent)}, \\ \lambda &\sim \text{Gamma}(\alpha, \beta), \\ \phi &\sim \text{Gamma}(\gamma, \delta), \\ m &\sim \text{uniform over } \{1, \dots, M\}, \end{aligned} \tag{2.11}$$

where α, β, γ and δ are known constants and the model is termed "hierarchical", since the parameters of the Poisson distributions are modeled as random themselves. That is, m is the year where there is a significant change in the number of disasters as modeled by $Y_1, \dots, Y_M$ with different (random) intensities λ and ϕ depending on whether Y_i is measured before or after the change point m, respectively. Due to missing prior knowledge about the change point m its distribution is modeled as uniform.

The joint posterior density of λ, ϕ and m given data $\boldsymbol{y} = (y_1, \dots, y_M)'$ satisfies

$$\begin{aligned} &\pi(\lambda, \phi, m|\boldsymbol{y}) \\ &\quad\propto f(y_1, \dots, y_M|\lambda, \phi, m)p(\lambda, \phi, m) \\ &\quad= \left[\prod_{i=1}^{m} f_P(y_i; \lambda)\right]\left[\prod_{i=m+1}^{M} f_P(y_i; \phi)\right] f_G(\lambda; \alpha, \beta) f_G(\phi; \gamma, \delta) 1_{\{1,\dots,M\}}(m) \\ &\quad\propto \left[\prod_{i=1}^{m} e^{-\lambda}\lambda^{y_i}\right]\left[\prod_{i=m+1}^{M} e^{-\phi}\phi^{y_i}\right] \lambda^{\alpha-1}e^{-\beta\lambda}\phi^{\gamma-1}e^{-\delta\phi}1_{\{1,\dots,M\}}(m) \\ &\quad\propto \lambda^{\alpha+(\sum_{i=1}^{m} y_i)-1}e^{-(\beta+m)\lambda}\phi^{\gamma+(\sum_{i=m+1}^{M} y_i)-1}e^{-(\delta+M-m)\phi}1_{\{1,\dots,M\}}(m), \end{aligned}$$

where 1_A is the indicator function satisfying $1_A(m) = 1$ if $m \in A$ and $1_A(m) = 0$ otherwise. Further f_P and f_G denote the Poisson and Gamma density functions, respectively (see Glossary A.2).

Therefore the full conditionals can be calculated as

$$\begin{aligned} \pi_\lambda^{FC}(\lambda) := p(\lambda|\phi, m, \boldsymbol{y}) &= \frac{p(\lambda, \phi, m, \boldsymbol{y})}{p(\phi, m, \boldsymbol{y})} = \frac{f(\boldsymbol{y}|\lambda, \phi, m)p(\lambda, \phi, m)}{p(\phi, m, \boldsymbol{y})} \\ &\propto \pi(\lambda, \phi, m|\boldsymbol{y}) \quad \text{as function of } \lambda \\ &\propto \lambda^{\alpha+(\sum_{i=1}^{m} y_i)-1}e^{-(\beta+m)\lambda} \propto \text{Gamma}\left(\alpha + \sum_{i=1}^{m} y_i, \beta + m\right) \end{aligned}$$

and similarly, $\pi_\phi^{FC}(\phi) \propto \text{Gamma}(\gamma + \sum_{i=m+1}^{M} y_i, \delta + M - m)$, and for the discrete random parameter m for $m = 1, \dots, M$ as

$$\pi_m^{FC}(m) = \frac{\lambda^{\alpha+\sum_{i=1}^{m} y_i - 1} e^{-(\beta+m)\lambda} \phi^{\gamma+\sum_{i=m+1}^{M} y_i - 1} e^{-(\delta+M-m)\phi}}{\sum_{l=1}^{M} \lambda^{\alpha+\sum_{i=1}^{l} y_i - 1} e^{-(\beta+l)\lambda} \phi^{\gamma+\sum_{i=l+1}^{M} y_i - 1} e^{-(\delta+M-l)\phi}}.$$

Therefore the Gibbs sampler for (λ, ϕ, m) draws $\lambda^{(n+1)}$ from Gamma$(\alpha + \sum_{i=1}^{m^{(n)}} y_i, \beta + m^{(n)})$, $\phi^{(n+1)}$ from Gamma$(\gamma + \sum_{i=m^{(n)}+1}^{n} y_i, \delta + M - m^{(n)})$ and chooses $m^{(n+1)} = m$ with probability $\pi_m^{FC}(m)$. Here $\pi_m^{FC}(m)$ depends on $\lambda^{(n+1)}$ and $\phi^{(n+1)}$.

To get a first impression on the behavior of this Gibbs sampler, we simulated data from the model (2.11) with $M = 50, \alpha = 5, \beta = 1, \gamma = 1$ and $\delta = 1$ (left panel of Fig. 5) and implemented the Gibbs sampler for 100 iterations. Note that the Gamma priors for λ and ϕ are quite informative, since the signal-to-noise ratio (mean divided by standard deviation) is 1. For illustration we used the true values as starting values. In the left panel of Fig. 5 the data is presented and the time plots of the MCMC iterations and posterior density estimates for each parameter are shown in the right panel of the same figure. The true values are indicated by a vertical dotted line.

The time plots (first column of right panel) indicate that the sampler is converged, which we expect since we used the true values as starting values. The true values of λ and ϕ are reasonably in the center of the sampled posterior distribution. The sampler has no difficulty finding the true break point. In general, the assessment of convergence is difficult especially for higher dimensions and convergence diagnostics have to be considered.

We now establish a few basic facts for the Gibbs sampler. First of all the Gibbs sampler defines a Markov chain, since the update step at iteration j involves only values of the chain at $j - 1$. Also the chain is homogeneous, since transitions are only affected by the iteration through the chain values. The transition kernel from $\boldsymbol{\phi} = (\phi_1, \dots, \phi_d)'$ to $\boldsymbol{\theta} = (\theta_1, \dots, \theta_d)'$ is given by

$$p(\boldsymbol{\theta}, \boldsymbol{\phi}) = \prod_{i=1}^{d} \pi(\boldsymbol{\phi}_i | \boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_{i-1}, \boldsymbol{\theta}_{i+1}, \dots, \boldsymbol{\theta}_d). \tag{2.12}$$

The limiting distribution of a Markov chain with transition kernel (2.12) is π, which we established for $d = 2$ and the discrete case in Illustration 2.6. For the continuous case the exact conditions under which the Markov chain resulting from the Gibbs sampler has limiting distribution π are given in Roberts and Smith [27]. For the continuous case π-irreducibility and aperiodicity are sufficient conditions (see Nummelin [25]). However, there are Markov chains derived from the Gibbs sampler which are not irreducible, see, for example, Gilks et al. [17]. Finally it can also be shown that π is stationary.

Even though theoretical results assure the convergence of the Gibbs sampler, they are difficult to validate theoretically for many complex statistical problems. In these

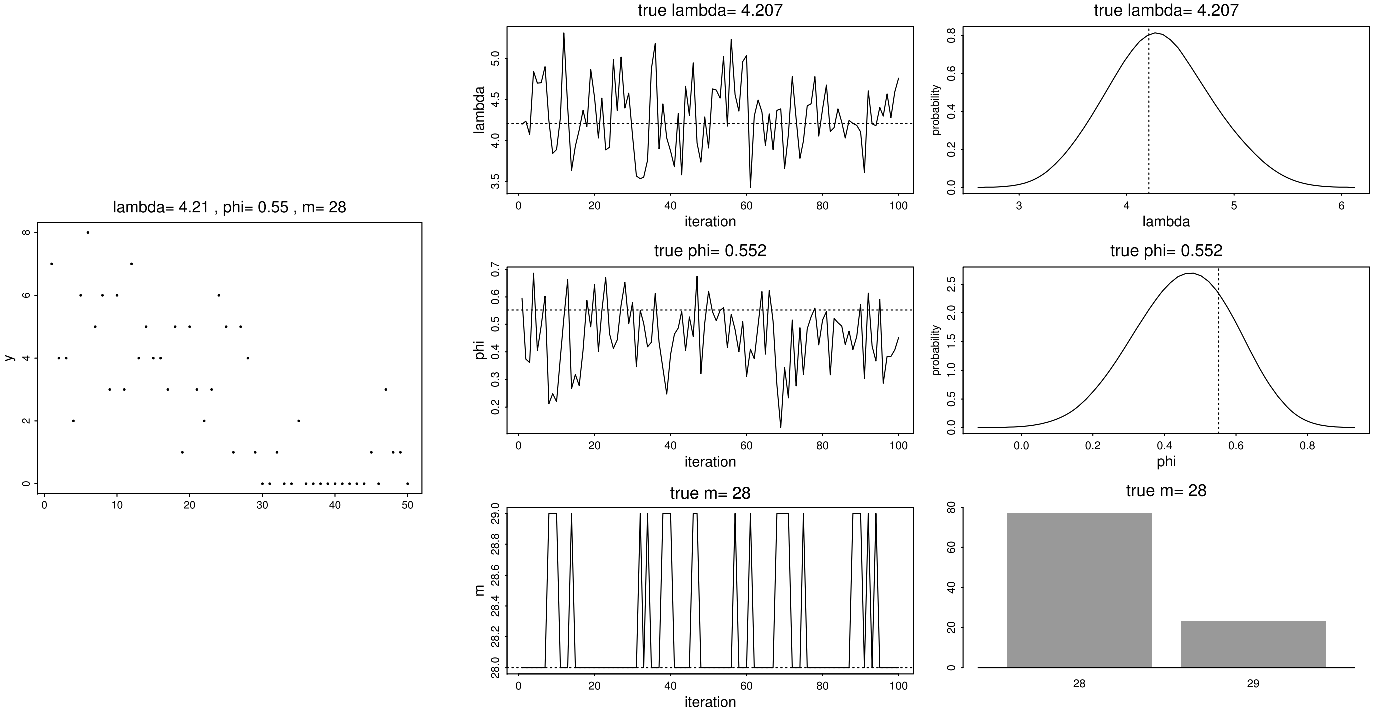


Fig. 5 *Left panel*: simulated data from model (2.11) with $M = 50$, $\alpha = 5$, $\beta = 1$, $\gamma = 1$ and $\delta = 1$. *Right panel*: time plot of MCMC iterations and posterior density estimates based on 100 iterations from the Gibbs sampler

cases a more practical approach is to assess the convergence by plotting n versus $\boldsymbol{\theta}^{(n)}$. If the variability of $\boldsymbol{\theta}^{(n)}$ for $n \geq n_0$ is approximately constant, then a burn-in of n_0 iterations is sufficient. Further MCMC sample based convergence assessments and comparison of several samplers with regard to burn-in iterations and required arithmetic operations are considered in Gilks et al. [17] and Marin and Robert [22] and the references therein.

Next, we draw attention to the use of the sample. For this, assume that we have a sample $\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(n)}$ from the posterior distribution π now available as generated by the Gibbs sampler, after some burn-in period and possibly thinning or batching to reduce autocorrelation of the sampled MCMC iterates. Suppose we are interested in the posterior distribution of the statistics $\boldsymbol{\psi} = t(\boldsymbol{\theta})$. The standard estimator

$$\hat{\boldsymbol{\psi}} := \hat{E}_{\pi(\boldsymbol{\theta}|\boldsymbol{x})}(\boldsymbol{\psi}) := \frac{1}{n}\sum_{j=1}^{n} t\big(\boldsymbol{\theta}^{(j)}\big)$$

estimates the posterior mean $E_{\pi(\boldsymbol{\theta}|\boldsymbol{x})}(\boldsymbol{\psi})$ of $\boldsymbol{\psi}$, while the posterior variance $\sigma^2_{\boldsymbol{\psi}} := \mathrm{Var}_{\pi(\boldsymbol{\theta}|\boldsymbol{x})}(\boldsymbol{\psi}) = E_{\pi(\boldsymbol{\theta}|\boldsymbol{x})}(\boldsymbol{\psi}^2) - [E_{\pi(\boldsymbol{\theta}|\boldsymbol{x})}(\boldsymbol{\psi})^2]$ is estimated by

$$\hat{\sigma}^2_{\boldsymbol{\psi}} := \hat{E}_{\pi(\boldsymbol{\theta}|\boldsymbol{x})}\big(\boldsymbol{\psi}^2\big) - \big[\hat{E}_{\pi(\boldsymbol{\theta}|\boldsymbol{x})}(\boldsymbol{\psi})\big]^2 = \frac{1}{n}\sum_{j=1}^{n}\big[t\big(\boldsymbol{\theta}^{(j)}\big) - \hat{\boldsymbol{\psi}}\big]^2.$$

Moreover, *posterior credibility intervals* for $\boldsymbol{\psi}$ can be estimated by using sample quantiles as the estimates of the interval limits. For example if one is interested in estimating a 95 % credible interval for $\boldsymbol{\psi}$ and $n = 1000$, then the estimated credible interval is given as the interval between the 25th and 975th largest sampled value for $\boldsymbol{\psi}$. This section concludes with a continuation of the example on linear regression models.

Illustration 2.8 (Linear Regression with Ar(1) Disturbances) Sometimes the observed risk quantities are not independent, but might depend on previous observations. For example if we consider monthly plant growth rates, then the growth rate might depend on the variety but also on the previous month growth rate. Therefore we extend the linear regression model of Sect. 1.3 to include autoregressive lag 1 (AR(1)) disturbances, that is, the response variables are no longer assumed independent but dependent upon the previous response. We change indices from i to t to acknowledge the time dependencies. Similar to (1.13) the model is then given by

$$Y_t = x_{t1}\beta_1 + \cdots + x_{td}\beta_d + u_t \quad \text{where } u_t = \rho u_{t-1} + \varepsilon_t$$

for a time series of responses Y_t with possibly time dependent covariates $\boldsymbol{x}_t = (x_{t1}, \dots, x_{td})' \in \mathbb{R}^d$ for $t = 1, \dots, T$. Further we assume $|\rho| < 1$ and $\epsilon_t \sim N(0, \sigma^2)$ are i.i.d. As an initial condition we use $u_0 \sim N(0, \frac{\sigma^2}{1-\rho^2})$. The following informative priors can be used:

- $\boldsymbol{\beta}|\sigma^2 \sim N_d(\boldsymbol{\beta}_0, \sigma^2 A_0^{-1})$

- $\sigma^2 \sim$ Inverse Gamma$(\frac{n_0}{2}, \frac{\delta_0}{2})$
- $\rho \sim N(\rho_0, R_0^{-1})$ truncated to $(-1, 1)$, where a truncated normal distribution is a normal distribution whose values are bounded below, above or both. Thus the usual normal density is multiplied with an indicator function $1_{(a,b)}$ for an interval with endpoints $a < b$ and rescaled appropriately to ensure that it integrates to 1.

In the following we determine the full conditional distributions of the parameters, which can be used in a corresponding Gibbs sampling scheme.

1. *Regression parameter:* To update the vector of regression parameters $\boldsymbol{\beta}$ consider the following transformations

$$\boldsymbol{Y}^* := \begin{pmatrix} \sqrt{1-\rho^2}Y_1 \\ Y_2 - \rho Y_1 \\ Y_3 - \rho Y_2 \\ \vdots \\ Y_T - \rho Y_{T-1} \end{pmatrix} \quad \text{and} \quad X^* := \begin{pmatrix} \sqrt{1-\rho^2}\boldsymbol{x}_1' \\ \boldsymbol{x}_2' - \rho \boldsymbol{x}_1' \\ \boldsymbol{x}_3' - \rho \boldsymbol{x}_2' \\ \vdots \\ \boldsymbol{x}_T' - \rho \boldsymbol{x}_{T-1}' \end{pmatrix}.$$

   Therefore $\boldsymbol{Y}^*$ follows a standard linear model with

$$\boldsymbol{Y}^* = X^*\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \text{where } \boldsymbol{\varepsilon} \sim N_T\big(0, \sigma^2 I_T\big).$$

   Since the full conditional for $\boldsymbol{\beta}$ given $\boldsymbol{Y}, X, \rho$ and σ^2 is the same as the full conditional for $\boldsymbol{\beta}$ given $\boldsymbol{Y}^*$, X^*, ρ and σ^2, we can use Theorem 1.6 to show that

$$\boldsymbol{\beta}|\boldsymbol{Y}, X, \rho, \sigma^2 \sim N_p\big(\boldsymbol{\beta}_1, \sigma^2 B_1^{-1}\big),$$

   with $B_1 = (A_0 + X^{*\prime}X^*)^{-1}$ and $\boldsymbol{\beta}_1 = B_1(A_0\boldsymbol{\beta}_0 + X^{*\prime}\boldsymbol{Y}^*)$.
2. AR(1) *error variance:* By again considering the precision $\phi := \frac{1}{\sigma^2}$ and using the equality of the following conditional distributions $\phi|\boldsymbol{Y}, X, \boldsymbol{\beta}, \rho = \phi|\boldsymbol{Y}^*, X^*, \boldsymbol{\beta}, \rho$, it can be shown that

$$\sigma^2|\boldsymbol{Y}, X, \boldsymbol{\beta}, \rho \sim \text{Inverse Gamma}\left(\frac{n_1}{2}, \frac{\delta_1}{2}\right),$$

   with $n_1 = T + n_0 + d$ and $\delta_1 = \delta_0 + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'X^{*\prime}X^*(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + (\boldsymbol{Y}^* - X^*\hat{\boldsymbol{\beta}})'(\boldsymbol{Y}^* - X^*\hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \boldsymbol{\beta}_0)'A_0(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$, where $\hat{\boldsymbol{\beta}} = (X^{*\prime}X^*)^{-1}X^{*\prime}\boldsymbol{Y}^*$.
3. *Correlation parameter:* Finally for updating the parameter ρ we can use Bayes' theorem to show that

$$\rho|\boldsymbol{Y}, X, \boldsymbol{\beta}, \sigma^2 \sim N(\tilde{\rho}, \tilde{R}) \text{ truncated to } (-1, 1),$$

   where $\tilde{R} := \sigma^{-2}(\sum_{t=1}^{T} u_{t-1}^2 + R_0)$ and $\tilde{\rho} := \tilde{R}^{-1}(\sigma^{-2}\sum_{t=1}^{T} u_t u_{t-1} + R_0\rho_0)$.

2.4 Metropolis Hastings Algorithms

The final MCMC algorithms presented here are the *Metropolis Hastings algorithms* (Metropolis et al. [23]; Hastings [19]). A nice introduction to the Metropolis Hastings algorithms is given in Chib and Greenberg [3]. As before, we are interested in constructing a Markov Chain with given stationary distribution π. First we consider a small example to motivate the discussion below.

Illustration 2.9 (Metropolis Hastings Algorithms) Consider a distribution π for $\boldsymbol{x} \in S$, where $S \subset \mathbb{R}^d$, $d \geq 1$. For a possible application recall Illustration 2.6, where we investigated the health states of a couple as modeled by the two-dimensional state space $S = \{0, 1\}^2$ and the probability distribution π.

Our aim is to construct a Markov chain with stationary and limiting distribution π. For this, let Q be any four-dimensional irreducible transition matrix on S satisfying the symmetry condition $Q(\boldsymbol{x}, \boldsymbol{y}) = Q(\boldsymbol{y}, \boldsymbol{x})\ \forall \boldsymbol{x}, \boldsymbol{y} \in S$ and define a Markov chain $\{\boldsymbol{\theta}^{(n)}, n \geq 0\}$ as having transitions from $\boldsymbol{x}$ to $\boldsymbol{y}$ proposed according to the probabilities $Q(\boldsymbol{x}, \boldsymbol{y})$. This proposed value for $\boldsymbol{\theta}^{(n+1)}$ is accepted with probability $\min\{1, \frac{\pi(\boldsymbol{y})}{\pi(\boldsymbol{x})}\}$ and rejected otherwise, leaving the chain in $\boldsymbol{x}$. This implies that for $\boldsymbol{x} \neq \boldsymbol{y}$

$$\begin{aligned} P(\boldsymbol{x}, \boldsymbol{y}) &= P\big(\boldsymbol{\theta}^{(n+1)} = \boldsymbol{y},\ \text{transition accepted}|\boldsymbol{\theta}^{(n)} = \boldsymbol{x}\big) \\ &= P\big(\boldsymbol{\theta}^{(n+1)} = \boldsymbol{y}|\boldsymbol{\theta}^{(n)} = \boldsymbol{x}\big) P(\text{transition accepted}) \\ &= Q(\boldsymbol{x}, \boldsymbol{y}) \min\left\{1, \frac{\pi(\boldsymbol{y})}{\pi(\boldsymbol{x})}\right\} \end{aligned}$$

and for $\boldsymbol{x} = \boldsymbol{y}$

$$\begin{aligned} P(\boldsymbol{x}, \boldsymbol{x}) = {} & P\big(\boldsymbol{\theta}^{(n+1)} = \boldsymbol{x},\ \text{accepted}|\boldsymbol{\theta}^{(n)} = \boldsymbol{x}\big) \\ & + P\big(\boldsymbol{\theta}^{(n+1)} \neq \boldsymbol{x},\ \text{not accepted}|\boldsymbol{\theta}^{(n)} = \boldsymbol{x}\big) \\ = {} & P\big(\boldsymbol{\theta}^{(n+1)} = \boldsymbol{x}|\boldsymbol{\theta}^{(n)} = \boldsymbol{x}\big) P(\text{accep.}) \\ & + \sum_{\boldsymbol{y} \neq \boldsymbol{x}} P\big(\boldsymbol{\theta}^{(n+1)} = \boldsymbol{y}|\boldsymbol{\theta}^{(n)} = \boldsymbol{x}\big) P(\text{not accep.}) \\ = {} & Q(\boldsymbol{x}, \boldsymbol{x}) \min\left\{1, \frac{\pi(\boldsymbol{x})}{\pi(\boldsymbol{x})}\right\} + \sum_{\boldsymbol{y} \neq \boldsymbol{x}} Q(\boldsymbol{x}, \boldsymbol{y}) \left[1 - \min\left\{1, \frac{\pi(\boldsymbol{y})}{\pi(\boldsymbol{x})}\right\}\right]. \end{aligned}$$

Further observe that if we assume that $\pi(\boldsymbol{y}) > \pi(\boldsymbol{x})$ for $\boldsymbol{x} \neq \boldsymbol{y}$, then

$$\begin{aligned} \pi(\boldsymbol{x}) P(\boldsymbol{x}, \boldsymbol{y}) = \pi(\boldsymbol{x}) Q(\boldsymbol{x}, \boldsymbol{y}) \min\left\{1, \frac{\pi(\boldsymbol{y})}{\pi(\boldsymbol{x})}\right\} &= \pi(\boldsymbol{x}) Q(\boldsymbol{x}, \boldsymbol{y}) \\ = \pi(\boldsymbol{y}) \min\left\{1, \frac{\pi(\boldsymbol{x})}{\pi(\boldsymbol{y})}\right\} Q(\boldsymbol{y}, \boldsymbol{x}) &= \pi(\boldsymbol{y}) P(\boldsymbol{y}, \boldsymbol{x}), \end{aligned}$$

and similarly if $\pi(\boldsymbol{y}) < \pi(\boldsymbol{x})$. This result is referred to as *reversibility* of a Markov chain and ensures that π constitutes the stationary distribution of the chain. If Q is aperiodic, so will be P and the stationary distribution is also the limiting distribution.

In general, Metropolis Hastings algorithms also exploit the concept of reversibility as in Illustration 2.9. That is, in order to construct a Markov chain with stationary distribution π we require the following *reversibility condition* for the transition kernel $P(\boldsymbol{\theta}, \boldsymbol{\phi})$:

$$\pi(\boldsymbol{\theta}) P(\boldsymbol{\theta}, \boldsymbol{\phi}) = \pi(\boldsymbol{\phi}) P(\boldsymbol{\phi}, \boldsymbol{\theta}) \quad \forall \boldsymbol{\theta}, \boldsymbol{\phi}.$$

Hastings [19] proposes to define the acceptance probability in such a way that when combined with an arbitrary transition probability, it defines a reversible chain. Such an acceptance probability is given by

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\phi}) = \begin{cases} \min\{1, \frac{\pi(\boldsymbol{\phi}) Q(\boldsymbol{\phi}, \boldsymbol{\theta})}{\pi(\boldsymbol{\theta}) Q(\boldsymbol{\theta}, \boldsymbol{\phi})}\}, & \text{if } \pi(\boldsymbol{\theta}) Q(\boldsymbol{\theta}, \boldsymbol{\phi}) > 0, \\ 1, & \text{otherwise.} \end{cases} \tag{2.13}$$

Algorithms based on (2.13) are called *Metropolis Hastings* (*MH*) *algorithms*. MH algorithms define reversible chains with stationary distribution π if $P(\boldsymbol{\theta}, \boldsymbol{\phi}) > 0$. Roberts and Smith [27] show that if Q is irreducible and aperiodic and $\alpha(\boldsymbol{\theta}, \boldsymbol{\phi}) > 0$ for all $(\boldsymbol{\theta}, \boldsymbol{\phi})$, then the algorithm defines an irreducible and aperiodic Markov chain with limiting distribution π. The MH algorithm can now be described as follows:

1. Set iteration counter $j = 1$ and arbitrary initial value $\boldsymbol{\theta}^{(0)}$.
2. Move the chain to a new value $\boldsymbol{\phi}$ generated from the density $Q(\boldsymbol{\theta}^{(j-1)}, \cdot)$.
3. Evaluate the acceptance probability of the move given by $\alpha(\boldsymbol{\theta}^{(j-1)}, \boldsymbol{\phi})$ in (2.13). If the move is accepted, then $\boldsymbol{\theta}^{(j)} = \boldsymbol{\phi}$. If the move is not accepted, then $\boldsymbol{\theta}^{(j)} = \boldsymbol{\theta}^{(j-1)}$ and the chain does not move.
4. Change the counter from j to $j + 1$ and return to Step 2 until convergence is reached.

Step 3 can easily be performed by generating an independent uniform quantity u. If $u \leq \alpha$, then the move is accepted and else it is not.

Note that you do not need to know the often complicated normalizing constant of the stationary distribution π to perform the MH algorithm. Further, when using a symmetric proposal probability as in Illustration 2.9, (2.13) simplifies to $\alpha(\boldsymbol{\theta}, \boldsymbol{\phi}) = \min\{1, \frac{\pi(\boldsymbol{\phi})}{\pi(\boldsymbol{\theta})}\}$ if $\pi(\boldsymbol{\theta}) > 0$ and $\alpha(\boldsymbol{\theta}, \boldsymbol{\phi}) = 1$ otherwise. Other common choices for Q lead to a *random walk* (new value = old value + disturbance; Illustration 2.1), *independence* (new value chosen independently of old value) or *hybrid chains* (Metropolis within Gibbs algorithm).

We close our discussion of MCMC methods with an example resuming and extending the Poisson model for claim frequencies of Illustration 1.3.

Illustration 2.10 (Claim Frequencies) Scollnik [29] considered the following model for modeling claim frequency data for group insurance policies: let X_{ij} be the number of claims for the ith group of policy holders in the jth policy year and P_{ij} the payroll count for the ith group of company employees in the jth policy year for $i=1,\dots,I$, $j=1,\dots,J$. The payroll counts give the number of employees which are at risk to incur a claim. The dependency among the claim counts over different years for the same policy i is modeled by introducing an unobserved random unit rate θ_i which has a common distribution for all policies. In particular Scollnik [29] assumed that X_{ij} given θ_i are independent with

$$X_{ij}|P_{ij},\theta_i \sim \text{Poisson}(P_{ij}\theta_i), \qquad \alpha \sim \text{Gamma}(5,5),$$
$$\theta_i|\alpha,\beta \sim \text{Gamma}(\alpha,\beta), \qquad \beta \sim \text{Gamma}(25,1).$$

The prior specification for α and β are rather arbitrary, but they imply that each θ_i has a prior mean and standard deviation approximately equal to 0.041 and 0.048, which might not be unreasonable in this context according to Scollnik [29]. Denote by $\boldsymbol{X}_i=(X_{i1},\dots,X_{iJ})'$ the number of claims vector of policy group i over all years and $\boldsymbol{X}=(\boldsymbol{X}_1',\dots,\boldsymbol{X}_I')'$ the total number of claims vector. Further, let $\boldsymbol{\theta}=(\theta_1,\dots,\theta_I)'$. Then the joint distribution of $(\boldsymbol{X},\boldsymbol{\theta},\alpha,\beta)$ can be written as follows:

$$p(\boldsymbol{X},\boldsymbol{\theta},\alpha,\beta)=\left[\prod_{j=1}^{J}\prod_{i=1}^{I} f_P(X_{ij}|P_{ij},\theta_i)\right]\left[\prod_{i=1}^{I} f_G(\theta_i|\alpha,\beta)\right]p(\alpha)p(\beta).$$

To update the unobserved latent rates θ_i we have as full conditional

$$\begin{aligned}p(\theta_i|\boldsymbol{X},\boldsymbol{\theta}_{-i},\alpha,\beta) &\propto \left[\prod_{j=1}^{J} f_P(X_{ij}|P_{ij},\theta_i)\right] f_G(\theta_i|\alpha,\beta)\\ &\propto \theta_i^{\alpha+\sum_{j=1}^{J}X_{ij}-1}\exp\left[-\left[\beta+\sum_{j=1}^{J}P_{ij}\right]\theta_i\right],\end{aligned}$$

which is a Gamma distribution with parameters $\alpha+\sum_{j=1}^{J}X_{ij}$ and $\beta+\sum_{j=1}^{J}P_{ij}$ and where $\boldsymbol{\theta}_{-i}=(\theta_1,\dots,\theta_{i-1},\theta_{i+1},\dots,\theta_I)'$. We see that these conditionals are actually independent of $\boldsymbol{\theta}_{-i}$. For updating α note that

$$p(\alpha|\boldsymbol{X},\boldsymbol{\theta},\beta)\propto\prod_{i=1}^{I} f_G(\theta_i|\alpha,\beta)p(\alpha)\propto\left[\frac{\beta^{\alpha}}{\Gamma(\alpha)}\right]^{I}\left[\prod_{i=1}^{I}\theta_i\right]^{\alpha}\alpha^4\exp(-5\alpha).$$

This is not a standard distribution and an MH step is needed.

Finally, to update β, we obtain for $\beta|\boldsymbol{X},\boldsymbol{\theta},\alpha$ again a Gamma distribution with parameters $I\alpha+25$ and $\sum_{i=1}^{I}\theta_i+1$.

According to Scollnik [29] we implemented a hybrid chain for the small data set with $I=3$ and $J=5$ shown in Table 2 using `WinBUGS` (**B**ayesian inference **U**sing **G**ibbs **S**ampling; http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml), which can be called directly from the statistical computing environment `R` (see

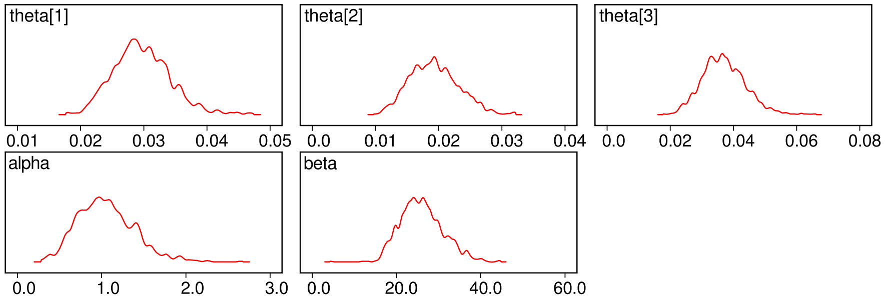


Fig. 6 Estimated posterior densities of 1000 iterations for $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)'$, α and β

Table 2 Data set of claim numbers and payroll counts for groups of policy holders and policy years

Year	Group 1		Group 2		Group 3	
	Payroll	Claims	Payroll	Claims	Payroll	Claims
1	280	9	260	6	267	6
2	320	7	275	4	145	8
3	265	6	240	2	120	3
4	340	13	265	8	105	4
5	325	10	285	5	115	7

Ntzoufras [7] for more information). The estimated posterior densities of 1000 iterations are shown in Fig. 6.

3 Food for Thought

There is software for Bayesian inference based on MCMC methods available in specialized problems. To the interested reader we particularly recommend to have a look at the above mentioned software `WinBUGS` and the illustrative book by Ntzoufras [7]. The recent book by Lunn et al. [6] also covers software for Bayesian statistical methods.

Another important issue of MCMC methods which could not be treated here appropriately are burn-in diagnostics which were briefly mentioned in Sect. 2.2 and provide tools for determining when we consider the values of the sampler as realizations from the posterior distribution. Further information can be found for example in Cowles and Carlin [11] and in Brooks and Roberts [9]. Related to this is the theoretical study of convergence questions.

Other areas of interest are, on the one hand, so-called ABC (Approximate Bayesian computation) methods which were developed for computationally very complex problems such as large-scale applications. Roberts et al. [28] and Frühwirth-Schnatter and Sögner [13], on the other hand, use MCMC methods for estimating stochastic volatility models commonly used in financial applications.

4 Summary

In this chapter, we gave a brief introduction to the main concepts of Bayesian statistics. After discussing the fundamental Bayes' theorem and three illustrating examples, we examined the problem of an appropriate prior choice in more detail and introduced Bayesian inference techniques. The first section closed with the commonly used linear regression model.

In the second section, we introduced the important class of MCMC methods, which are increasingly becoming popular for estimating parameters in complex statistical models. They are based on Monte Carlo techniques and properties of Markov chains, which were discussed before turning to the two most common MCMC algorithms, namely the Gibbs sampler and the Metropolis Hastings algorithms. There were discussed and illustrated using relevant examples involving risk quantities on different scales and with different contexts.

Appendix: Glossary

A.1 Foundations

Symbol	Explanation
X	random variable (r.v.)
$X = x$	realization or observed value of r.v. X
X continuous	r.v. X takes on any value in an interval (e.g., $X =$ annual crop yield $\in [0, \infty)$)
X discrete	r.v. X takes on only finite or countable many values (e.g., $X =$ number of mining disasters $\in \{0, 1, 2, \ldots\}$)
i.i.d.	independent and identically distributed
θ	unknown parameter of a distribution (e.g., $\theta =$ probability of occurrence of a complication after a medical treatment)
$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$	unknown parameters of a distribution (e.g., $\boldsymbol{\theta} = (\mu, \sigma^2)$, μ mean, σ^2 variance of a normal distribution)
$P_{\boldsymbol{\theta}}(A)$	probability that event A occurs when parameters $\boldsymbol{\theta}$ are true
$F(x\|\boldsymbol{\theta})$	cumulative distribution function (cdf) of r.v. X, i.e., $F(x\|\boldsymbol{\theta}) = P_{\boldsymbol{\theta}}(X \leq x)$
$f(x\|\boldsymbol{\theta})$	probability density function (pdf), when X continuous, i.e., $f(x\|\boldsymbol{\theta}) \geq 0$, $\int_{-\infty}^{\infty} f(x\|\boldsymbol{\theta})dx = 1$, $P_{\boldsymbol{\theta}}(X \leq x) = \int_{-\infty}^{x} f(x\|\boldsymbol{\theta})dx$
$f(x\|\boldsymbol{\theta})$	probability mass function (pmf), when X discrete, i.e., $f(x\|\boldsymbol{\theta}) = P_{\boldsymbol{\theta}}(X = x)$
$\mu = E(X)$	mean or expectation of r.v. X ($E(X) = \int_{-\infty}^{\infty} x f(x\|\boldsymbol{\theta})dx$ for X continuous)
$\sigma^2 = \text{Var}(X)$	variance of r.v. X ($\text{Var}(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x\|\boldsymbol{\theta})dx$ for X continuous)
$\phi = \frac{1}{\sigma^2}$	precision of r.v. X
$X \sim F(\cdot\|\boldsymbol{\theta})$	X has cdf $F(\cdot\|\boldsymbol{\theta})$

Symbol	Explanation
$X \sim f(\cdot\|\boldsymbol{\theta})$	X has pdf/pmf $f(\cdot\|\boldsymbol{\theta})$
$(X, Y) \sim f(\cdot, \cdot\|\boldsymbol{\theta})$	r.v.s X and Y have joint pdf/pmf $f(\cdot, \cdot\|\boldsymbol{\theta})$
$f_X(x\|\boldsymbol{\theta})$ $(f_Y(y\|\boldsymbol{\theta}))$	marginal pdf for X (Y): $f_X(x\|\boldsymbol{\theta}) = \int_{-\infty}^{\infty} f(x, y\|\boldsymbol{\theta})dy$ $(f_Y(y\|\boldsymbol{\theta}) = \int_{-\infty}^{\infty} f(x, y\|\boldsymbol{\theta})dx)$
$f_X(x\|\boldsymbol{\theta})$ $(f_Y(y\|\boldsymbol{\theta}))$	marginal pmf for X (Y): $f_X(x\|\boldsymbol{\theta}) = \sum_{i=1}^{\infty} f(x, y_i\|\boldsymbol{\theta})$ $(f_Y(y\|\boldsymbol{\theta}) = \sum_{i=1}^{\infty} f(x_i, y\|\boldsymbol{\theta}))$
$P_{\boldsymbol{\theta}}(A\|B)$	conditional probability of A given B: $P_{\boldsymbol{\theta}}(A\|B) = \frac{P_{\boldsymbol{\theta}}(A \cap B)}{P_{\boldsymbol{\theta}}(B)}$ if $P_{\boldsymbol{\theta}}(B) > 0$
x_α	α-quantile of continuous r.v. X: $P_{\boldsymbol{\theta}}(X \le x_\alpha) = \alpha$
$x_{0.5}$	median of continuous r.v. X
x_{mode}	mode of continuous r.v. X, that is the value which maximizes $f(x\|\boldsymbol{\theta})$ over x
$\boldsymbol{X} = (X_1, \ldots, X_n)'$	$\boldsymbol{X}$ random vector, where $X_1, \ldots, X_n$ r.v.s
$F(\boldsymbol{x}\|\boldsymbol{\theta})$ $(f(\boldsymbol{x}\|\boldsymbol{\theta}))$	cdf (pdf/pmf) of $\boldsymbol{X}$
$E(\boldsymbol{X}) = (E(X_1), \ldots, E(X_n))$	mean vector of random vector $\boldsymbol{X}$
$\Sigma = (\Sigma_{ij})_{i,j=1,\ldots,n}$	covariance matrix of random vector $\boldsymbol{X}$ with $\Sigma_{ij} = \text{Cov}(X_i, X_j) = E((X_i - \mu_i)(X_j - \mu_j))$
Σ^{-1}	precision matrix of random vector $\boldsymbol{X}$
$I(\boldsymbol{\theta}) = (I(\boldsymbol{\theta})_{ij})_{i,j=1,\ldots,n}$	Fisher information matrix with $I(\boldsymbol{\theta})_{ij} = E(\frac{\partial^2 \ln f(\boldsymbol{X}\|\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j})$

A.2 Distributions

Symbol	Explanation
$X \sim N(\mu, \sigma^2)$	X is normally distributed with mean μ, variance σ^2 and pdf $f(x\|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(x - \mu)^2\}$, $x \in \mathbb{R}$
$X \sim \text{Bernoulli}(\theta)$	X is Bernoulli distributed with success probability $\theta \in (0, 1)$ and pmf $f(x\|\theta) = \theta^x(1 - \theta)^{1-x}$, $x = 0, 1$, $E(X) = \theta$, $\text{Var}(X) = \theta(1 - \theta)$
$X \sim \text{Beta}(\alpha, \beta)$	X is Beta distributed with parameters $\alpha > 0$, $\beta > 0$ and pdf $f(x\|\alpha, \beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1 - x)^{\beta-1}$, $x \in (0, 1)$, $B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1 - x)^{\beta-1}dx$, $E(X) = \frac{\alpha}{\alpha+\beta}$, $\text{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
$X \sim \text{Poisson}(\theta)$	X is Poisson distributed with parameter $\theta > 0$ and pmf $f(x\|\theta) = \frac{\theta^x}{x!} e^{-x}$, $x \in \{0, 1, 2, \ldots\}$, $E(X) = \text{Var}(X) = \theta$
$X \sim \text{Gamma}(\alpha, \beta)$	X is Gamma distributed with parameters $\alpha > 0$, $\beta > 0$ and pdf $f(x\|\alpha, \beta) = \frac{1}{\Gamma(\alpha)} \beta^\alpha x^{\alpha-1} e^{-\beta x}$, $x > 0$, $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$, $E(X) = \frac{\alpha}{\beta}$, $\text{Var}(X) = \frac{\alpha}{\beta^2}$
$X \sim N(0, 1)$	X is standard normal with pdf $\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}x^2\}$, and cdf $\Phi(x) = \int_{-\infty}^x \varphi(u)du$, $E(X) = 0$, $\text{Var}(X) = 1$
$\boldsymbol{X} \sim N_n(\boldsymbol{\mu}, \Sigma)$	$\boldsymbol{X}$ is multivariate normally distributed with mean vector $\boldsymbol{\mu}$, covariance matrix Σ and pdf $f(\boldsymbol{x}\|\boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{n/2}} \|\Sigma\|^{-1/2} \exp\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})' \times \Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\}$, $\boldsymbol{x} \in \mathbb{R}^n$, $E(\boldsymbol{X}) = \boldsymbol{\mu}$, $\text{Var}(\boldsymbol{X}) = \Sigma$

A.3 Classical Statistics

Symbol	Explanation
θ ($\boldsymbol{\theta}$)	unknown fixed parameter to be estimated
$(x_1, \ldots, x_n)'$	i.i.d. sample (realizations) from r.v. X
$\hat{\theta}$ ($\hat{\boldsymbol{\theta}}$)	estimate of θ ($\boldsymbol{\theta}$) based on data $\boldsymbol{x} = (x_1, \ldots, x_n)$
$\ell(\boldsymbol{\theta}\|\boldsymbol{x})$	likelihood for $\boldsymbol{\theta}$ based on data $\boldsymbol{x}$ from $X \sim f(\cdot\|\boldsymbol{\theta})$ given as $\ell(\boldsymbol{\theta}\|\boldsymbol{x}) = f(\boldsymbol{x}\|\boldsymbol{\theta})$
$\hat{\boldsymbol{\theta}}_{ML}$	maximum likelihood estimator of $\boldsymbol{\theta}$: maximizes the likelihood $\ell(\boldsymbol{x}\|\boldsymbol{\theta})$ over $\boldsymbol{\theta}$
$I^{-1}(\boldsymbol{\theta})$	inverse Fisher information matrix, corresponds to asymptotic covariance matrix of the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_{ML}$
$\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i$	sample or empirical mean for the i.i.d. sample $(x_1, \ldots, x_n)$
$s^2 := \frac{1}{n-1} \times \sum_{i=1}^{n}(x_i - \bar{x})^2$	sample variance for the i.i.d. sample $(x_1, \ldots, x_n)$
$Y_i \sim N(x_{i1}\beta_1 + \cdots + x_{id}\beta_d, \sigma^2)$ independent for $i = 1, \ldots, d$	linear regression model for response Y_i, covariates $x_{i1}, \ldots, x_{id}$ and unknown regression coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)$
$\hat{\boldsymbol{\beta}}_{LS}$	least square estimator of $\boldsymbol{\beta}$, given by minimizing $Q(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - x_{i1}\beta_1 - \cdots - x_{id}\beta_d)^2$ for observed responses $y_1, \ldots, y_n$
$[l(\boldsymbol{x}), u(\boldsymbol{x})]$	$100(1-\alpha)$ % confidence interval for θ if $P_\theta(l(\boldsymbol{x}) \le \theta \le u(\boldsymbol{x})) \ge 1 - \alpha$, that is, the random interval $[l(\boldsymbol{x}), u(\boldsymbol{x})]$ covers the true parameter θ in $100(1-\alpha)$ % of times

A.4 Bayesian Statistics

Symbol	Explanation
θ ($\boldsymbol{\theta}$)	unknown random parameter
$p(\boldsymbol{\theta})$	prior pdf/pmf for $\boldsymbol{\theta}$
$p(\boldsymbol{\theta}\|\boldsymbol{x})$	posterior pdf/pmf of $\boldsymbol{\theta}$ given the observed sample $\boldsymbol{x}$ from $\boldsymbol{X} \sim f(\cdot\|\boldsymbol{\theta})$ Bayes' theorem: $p(\boldsymbol{\theta}\|\boldsymbol{x}) = \frac{\ell(\boldsymbol{\theta}\|\boldsymbol{x})p(\boldsymbol{\theta})}{\int_{-\infty}^{\infty}\ell(\boldsymbol{\theta}\|\boldsymbol{x})p(\boldsymbol{\theta})d\boldsymbol{\theta}}$
$\boldsymbol{\theta}_{\text{mode}}(\boldsymbol{x})$	posterior mode = mode of posterior distribution
$\boldsymbol{\theta}_{\text{mean}}(\boldsymbol{x})$	posterior mean = mean of posterior distribution
$I(\boldsymbol{x})$	$100(1-\alpha)$ % credible interval for θ if $\int_{I(\boldsymbol{x})} p(\theta\|\boldsymbol{x})d\theta = 1 - \alpha$
$\theta_\alpha(\boldsymbol{x})$ ($\hat{\theta}_\alpha(\boldsymbol{x})$)	(empirical) α-quantile of posterior distribution
$[\hat{\theta}_{\alpha/2}(\boldsymbol{x}), \hat{\theta}_{1-\alpha/2}(\boldsymbol{x})]$	$100(1-\alpha)$ % credible interval based on empirical quantiles
$f(y\|\boldsymbol{x})$	predictive density of future observation y given the observations $\boldsymbol{x}$: $f(y\|\boldsymbol{x}) = \int f(y\|\boldsymbol{\theta})p(\boldsymbol{\theta}\|\boldsymbol{x})d\boldsymbol{\theta}$ if Y is independent of $\boldsymbol{X}$ given $\boldsymbol{\theta}$

A.5 MCMC Methods

Symbol	Explanation
$\{\boldsymbol{\theta}^{(t)} : t \in T\}$	stochastic process with random vectors $\boldsymbol{\theta}^{(t)}$ taking values in the state space S for each t out of the index set T
$\{\boldsymbol{\theta}^{(n)} : n = 1, 2, \ldots\}$	Markov chain (MC) if (2.3) holds
$\boldsymbol{\theta}^{(n)}$ homogeneous	if (2.3) does not depend on n
$P(\boldsymbol{x}, \boldsymbol{y}) := P(\boldsymbol{\theta}^{(n+1)} = \boldsymbol{y} \mid \boldsymbol{\theta}^{(n)} = \boldsymbol{x})$	transition probability of homogeneous MC $\boldsymbol{\theta}^{(n)}$ with discrete state space S
$P = (P(\boldsymbol{x}_i, \boldsymbol{x}_j))_{i,j=1,\ldots,r}$	transition matrix for a homogeneous MC $\boldsymbol{\theta}^{(n)}$ with finite state space $S = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_r\}$: $P(\boldsymbol{x}_i, \boldsymbol{x}_j) = P(\boldsymbol{\theta}^{(n+1)} = \boldsymbol{x}_j \mid \boldsymbol{\theta}^{(n)} = \boldsymbol{x}_i)$
$P^m(\boldsymbol{x}, \boldsymbol{y}) := P(\boldsymbol{\theta}^{(n+m)} = \boldsymbol{y} \mid \boldsymbol{\theta}^{(n)} = \boldsymbol{x})$	mth order transition probability for $m > n$
$\pi^{(0)}(\boldsymbol{x}) = P(\boldsymbol{\theta}^{(0)} = \boldsymbol{x})$	initial distribution of MC $\boldsymbol{\theta}^{(n)}$
$\pi^{(n)}(\boldsymbol{x}) = P(\boldsymbol{\theta}^{(n)} = \boldsymbol{x})$	nth step marginal distribution of MC $\boldsymbol{\theta}^{(n)}$
π stationary	if (2.7) holds
$T_{\boldsymbol{y}}$	first visit of MC $\boldsymbol{\theta}^{(n)}$ to $\boldsymbol{y}$
$\rho_{\boldsymbol{x}\boldsymbol{y}}$	probability of visiting $\boldsymbol{y}$ after starting in $\boldsymbol{x}$
$\boldsymbol{y} \in S$ (positive) recurrent	$\rho_{\boldsymbol{y}\boldsymbol{y}} = 1$ ($\rho_{\boldsymbol{y}\boldsymbol{y}} = 1$ and $E(T_{\boldsymbol{y}} \mid \boldsymbol{\theta}^{(0)} = \boldsymbol{y}) < \infty$)
$\boldsymbol{\theta}^{(n)}$ irreducible	$\rho_{\boldsymbol{x}\boldsymbol{y}} > 0$, $\rho_{\boldsymbol{y}\boldsymbol{x}} > 0$ $\forall \boldsymbol{x}, \boldsymbol{y} \in S$
$\boldsymbol{\theta}^{(n)}$ aperiodic	if largest common divisor of $\{n \geq 1 : P^n(\boldsymbol{x}, \boldsymbol{x}) > 0\} = 1$ $\forall \boldsymbol{x} \in S$
$\boldsymbol{\theta}^{(n)}$ ergodic	if $\boldsymbol{\theta}^{(n)}$ aperiodic and irreducible
Full conditionals of random parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)'$	conditional distributions of θ_i, $i = 1, \ldots, d$, given all other components different from i
Autocorrelation of lag k	correlation $Cor(\boldsymbol{\theta}^{(n)}, \boldsymbol{\theta}^{(n+k)})$ in homogeneous MC $\boldsymbol{\theta}^{(n)}$

References

Selected Bibliography

1. W.M. Bolstad, *Introduction to Bayesian Statistics* (Wiley, Hoboken, 2004)
2. G. Casella, E.I. George, Explaining the Gibbs sampler. Am. Stat. **46**, 167–174 (1992)
3. S. Chib, E. Greenberg, Understanding the Metropolis-Hastings algorithm. Am. Stat. **49**, 327–335 (1995)
4. D. Gamerman, H.F. Lopes, *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference* (Taylor & Francis, Boca Raton, 2006)
5. P.M. Lee, *Bayesian Statistics: An Introduction*, 4th edn. (Wiley, Hoboken, 2012)
6. D. Lunn, C. Jackson, N. Best, A. Thomas, D. Spiegelhalter, *The BUGS Book—A Practical Introduction to Bayesian Analysis* (Chapman & Hall/CRC, London, 2012)
7. I. Ntzoufras, *Bayesian Modeling Using WinBUGS* (Wiley, Hoboken, 2009)

Additional Literature

8. J.O. Berger, *Statistical Decision Theory and Bayesian Analysis*, 2nd edn. (Springer, Berlin, 1985)
9. S.P. Brooks, G.O. Roberts, Assessing convergence of Markov chain Monte Carlo algorithms. Stat. Comput. **8**, 319–335 (1998)
10. B.P. Carlin, A.E. Gelfand, A.F.M. Smith, Hierarchical Bayesian analysis of changepoint problems. Appl. Stat. **41**, 389–405 (1992)
11. M.K. Cowles, B.P. Carlin, Markov chain Monte Carlo convergence diagnostics: a comparative review. J. Am. Stat. Assoc. **91**, 883–904 (1996)
12. R. Durrett, *Probability: Theory and Examples*, 4th edn. (Cambridge University Press, Cambridge, 2010)
13. S. Frühwirth-Schnatter, L. Sögner, Bayesian estimation of stochastic volatility models based on OU processes with marginal Gamma law. Ann. Inst. Stat. Math. **61**(1), 159–179 (2009)
14. A.E. Gelfand, A.F.M. Smith, Sampling-based approaches to calculating marginal densities. J. Am. Stat. Assoc. **85**, 398–409 (1990)
15. A. Gelman, J.B. Carlin, H.S. Stern, D.B.R. Rubin, *Bayesian Data Analysis*, 2nd edn. (Chapman & Hall, London, 2003)
16. S. Geman, D. Geman, Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. **6**, 721–741 (1984)
17. W.R. Gilks, S. Richardson, D.J. Spiegelhalter, *Markov Chain Monte Carlo in Practice* (Chapman & Hall, London, 1996)
18. P. Guttorp, *Stochastic Modeling of Scientific Data* (Chapman & Hall/CRC, London, 1995)
19. W.K. Hastings, Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57**, 97–109 (1970)
20. P.D. Hoff, *A First Course in Bayesian Statistical Methods*. Springer Texts in Statistics (Springer, New York, 2009)
21. A. Jeffreys, *The Theory of Probability* (Cambridge University Press, Cambridge, 1961)
22. J.-M. Marin, C.P. Robert, *Bayesian Core: A Practical Approach to Computational Bayesian Statistics* (Springer, New York, 2007)
23. N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, Equation of state calculations by fast computing machines. J. Chem. Phys. **21**, 1087–1092 (1953)
24. S.P. Meyn, R.L. Tweedie, *Markov Chains and Stochastic Stability*, 2nd edn. (Cambridge University Press, Cambridge, 2009)
25. E. Nummelin, *General Irreducible Markov Chains and Non-negative Operators* (Cambridge University Press, Cambridge, 1984)
26. S.I. Resnick, *Adventures in Stochastic Processes* (Birkhäuser, Boston, 1992)
27. G.O. Roberts, A.F.M. Smith, Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. Stoch. Process. Appl. **49**, 207–216 (1994)
28. G.O. Roberts, O. Papaspiliopoulos, P. Dellaportas, Bayesian inference for non-Gaussian Ornstein-Uhlenbeck stochastic volatility processes. J. R. Stat. Soc., Ser. B **66**, 369–393 (2004)
29. D.P.M. Scollnik, Actuarial modeling with MCMC and BUGS. N. Am. Actuar. J. **5**, 96–124 (2001)

Chapter 9
Dealing with Dependent Risks

Claudia Klüppelberg and Robert Stelzer

In most real life situations we are confronted not only with one single source of risk or one single risk, but with several sources of risk or combinations of risks. An important question is whether individual risks influence each other or not. This may involve the time of their occurrence and/or their severity. In other words, we need to understand how to model and describe the dependence structure of risks. Clearly, if risks influence each other in such a way that they tend to occur together and increase the severity of the overall risk, then the situation may be much more dangerous than otherwise.

We illustrate this with a concrete example. Consider a building which could be hit by an earthquake and a flood. If the building is situated on the Japanese coast, an earthquake may damage the building and cause a tsunami, which in turn floods the building. Hence, it is quite likely that by these two combined sources of risk a particularly disastrous event occurs. In other words, there is a strong positive dependence between these two risks (high damage from an earthquake will often come along with high damage from a flood). This does not mean that they always occur together, since an earthquake does not necessarily cause a tsunami, and there may be a flood caused only by heavy rain.

Keywords Copula · Correlation · Extreme dependence · Rank correlation · Risk measure

Mathematics Subject Classification (2010) 60E99 · 62E10 · 62H20 · 60-01 · 62-01

C. Klüppelberg
Chair of Mathematical Statistics, Center for Mathematical Sciences, Technische Universität München, Boltzmannstr. 3, 85748 Garching bei München, Germany

R. Stelzer (✉)
Institute of Mathematical Finance, Faculty of Mathematics and Economics, Ulm University, Helmholtzstr. 18, 89081 Ulm, Germany
e-mail: robert.stelzer@uni-ulm.de

C. Klüppelberg et al. (eds.), *Risk – A Multidisciplinary Introduction*,
DOI 10.1007/978-3-319-04486-6_9,

The Facts

- Dependence between risks and/or sources of risks is crucial for risk assessment, quantification and management.
- Adequate mathematical measures for the dependence between risks (or more generally between random variables) are needed.
- Correlation measures linear dependence, but characterises the full dependence structure only in special parametric models (the multivariate normal distribution is *the* typical example).
- Correlation is also useful in spherical and elliptical distributions.
- Rank correlations are appropriate dependence measures in certain situations.
- Copulae provide a way to characterise the dependence structure completely, but are rather complex objects.
- For risk assessment it is mainly the dependence structure of extreme events that matters. Thus, measures for dependence in extreme observations provide useful dependence measures for combined risks.

1 Introduction

In most situations (both in our professional and our daily life) risks are present. Often there exist various sources of risk which, in the end, determine the overall risk of a more-or-less complex system. This is a common situation in the financial world (i.e., for any bank and insurance company), in any engineering system, when working as a physician or when dealing with environmental consequences. It is then necessary to assess and deal with combinations of risks in an appropriate way.

There is a huge difference between two risks possibly occurring together and risks happening at different times. In one situation you need to be prepared to deal with both risks at the same time, whereas in the other situation it suffices to cope with one risk at a time. For example, if you consider the people needed on stand-by for the emergency services, you will need many more people in the first case. However, in almost all situations life is not even that easy; risks do not have to occur at the same time; instead they may or tend to occur at the same time. Then we need to understand and quantify this tendency. This is exactly what this paper is about, to understand how to model the statistical dependence between different risks.

There are two classical approaches. The first assesses the single risk factors by some monetary risk measure, and simply adds the different values of the single risk measures together. The second combines the monetary risks with a multivariate normal model, and assesses the dependence via the pairwise correlations.

Both approaches capture only part of the truth, and in this chapter we discuss their appropriateness, other approaches and the pros and cons of different approaches to model and measure risks of complex systems.

We are concerned with risk under dependence and thus we briefly have to make precise what we mean by this. In the end we want to use risk measures (see

Chap. 5, [15] for a detailed introduction) to quantify risks, as well as to assess the effects of risk management strategies. Essentially, we want to understand the effects the dependence structure has on these risk measures. In models it is of utmost importance to have an appropriate dependence structure capturing all effects relevant for the risk measures. So we want to discuss both how to model dependence and the effects of different ways of modelling on the final risk assessment.

Therefore let us briefly introduce two risk measures and note that we identify risk with a random variable; i.e., the outcome of a risky event.

Definition 1.1 (Examples of "Risk Measures") For a random variable X with distribution function $F(x) = P(X \leq x)$ for $x \in \mathbb{R}$ we define the following risk measures:

(a) *Variance:* $\operatorname{var}(X) = E((X - E[X])^2) = E(X^2) - (E(X))^2$ is the mean squared deviation from the mean or expected value of X.

(b) *Value-at-Risk:* Define the *quantile function* of F as

$$F^{-1}(\alpha) = \inf\bigl\{x : F(x) \geq \alpha\bigr\}, \quad \alpha \in (0, 1). \tag{1.1}$$

Note that for strictly increasing F this is simply the analytic inverse.
Then for a large value of α (usually $\alpha = 0.95$ or larger) $\mathrm{VaR}_\alpha = F^{-1}(\alpha)$ is called the Value-at-Risk (for the level α).

The first risk measure, i.e. the variance, gives the average squared difference between a random variable (the realisation of a risk) and its mean outcome. It measures how widely spread various outcomes are. Clearly, it is a very simplistic risk measure, since e.g. it does not differentiate between values higher and values lower than the mean, as it looks only at the squared distance. Normally, only one direction really matters when considering a particular risk. For instance, if we consider the level of a river in a German city and the flood risk, then it is irrelevant when the level is far smaller than the mean (of course, the "downside" direction may well matter for other risks, e.g. that water becomes scarce).

The Value-at-Risk or VaR is a very popular risk measure, in particular in the financial world. Above it has been assumed that the high realizations of X are "risky", but this is only a convention and can be changed to low realizations being risky. Intuitively, the value at risk gives the level which is not exceeded in $100 \cdot \alpha$ % of all cases (e.g. if the VaR at the level 0.95 is 500, then the relevant variable, "the risk", is above 500 in 5 % of all cases and in 95 % of all cases it is below 500). Moreover, the VaR has been incorporated into the Basel II regulations (the international rules governing how much capital banks must set aside to cover future losses from their business) and Solvency II (similar international rules for insurance companies), and the national legislation which enforces these international standards. VaR is the standard risk measure in use there (cf. Chap. 6, [20] for estimation methods).

We will see later, in particular in Illustration 2.3, that changing the dependence structure usually has major effects on the VaR. But note that VaR has been rightly criticized for various reasons:

(a) VaR takes only the event of large losses into account, but not the size of losses. In this sense the so called Tail-VaR is preferable, which measures the average of all losses exceeding VaR. So if a bank sets aside capital equal to its VaR it certainly goes bankrupt (or needs to be "rescued"), as soon as a loss occurs which is higher than VaR. In contrast to this, if it used the Tail-VaR to determine its risk capital, it has set aside enough capital to withstand such an event on average. So there should be a realistic chance that the capital is sufficient to cover the loss.
(b) VaR is not always a coherent risk measure. For a risk measure to be coherent (cf. Chap. 5, [15]) it is necessary that the risk measure of the sum of two risks is always below the sum of the two risk measures. Since e.g. banks estimate the VaR for each unit and add all resulting VaRs up to estimate the risk of the whole bank, the use of VaR may underestimate the true bank's VaR considerably.

As a very readable paper on dependence measures and their properties and pitfalls, which goes far beyond the present chapter, we recommend [2].

This paper is structured as follows. In Sect. 2 we introduce the mathematical definitions of (in)dependence of random variables and illustrate the effects of different dependence structures. In Sect. 3 we recall the multivariate normal distribution and discuss which kind of dependence it is able to model. We continue this in Sect. 4 where we consider the correlation as a popular dependence measure, discussing in detail its properties, problems, limitations and popular misconceptions. As the next natural step we present spherical and elliptical distributions in Sect. 5. Thereafter, we turn our focus onto alternative dependence measures starting with rank correlations in Sect. 6. Then in Sect. 7 we consider a concept—copulae—at length. In principle it is able to encode completely all possible dependence structures. As typically extreme events are the really dangerous risks, we indicate in Sect. 8 how to quantify and model the dependence of extreme events. Finally, we give you as our readers some Food for Thought in Sect. 9 and provide a brief summary in Sect. 10.

2 Independence and Dependence

The first simple question to answer is, when exactly do we have dependence between risks? The best answer seems to be a negative one, viz. risks are dependent whenever they are not independent.

Clearly, this means that we have to give a mathematical definition of independence. We do this for two random variables X and Y (which represent the risks we are interested in). Think for instance of our example of an earthquake and a flood at the beginning. Intuitively, independence should mean that whatever happens in one random variable, say X (the earthquake), should in no way affect what happens in Y

(the flood). If we know the value of X, this should not change our knowledge of what might happen and with what probability to Y. In proper mathematical terms one says that two random variables are independent if their joint distribution is the product of the two marginal distributions; i.e., $P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y)$ for all $x, y \in \mathbb{R}$ (note: $P(A)$ means the probability that some event A occurs). This implies that the probability distribution of Y conditional on X does not depend on X, but is simply equal to the distribution of Y; i.e., $P(Y \leq y \mid X \leq x) := P(X \leq x, Y \leq y)/ (P(X \leq x)) = P(Y \leq y)$ for all $x, y \in \mathbb{R}$. Obviously this is in line with the intuition given above.

Note that the necessity of a negative definition of dependence tells us that there are (too) many ways in which risks can be dependent. Hence, any mathematical object completely describing the dependence of arbitrary random variables has to be a very complex object. Turned the other way around any simple quantification of dependence—such as one real number obtained from the joint distribution of two random variables—will necessarily reflect only a very special aspect of dependence, or describe the dependence completely only in very special situations/set-ups. This should be kept in mind throughout the rest of this chapter and whenever trying to quantify dependence in applications.

In truly realistic situations we are interested in the (in)dependence of more than two random variables. We give the general definition and discuss and illustrate it afterwards.

Definition 2.1 (Independence) Let $X_1, X_2, \ldots, X_n$ for $n \in \mathbb{N}$ be random variables. Then $X_1, X_2, \ldots, X_n$ are called *independent* if

$$P(X_1 \leq x_1, \ldots, X_d \leq x_d) = P(X_1 \leq x_1) \cdots P(X_d \leq x_d) \tag{2.1}$$

holds for all $x_1, \ldots, x_d \in \mathbb{R}$.

Let us consider two special cases that are particularly relevant in applications.

(a) Assume the random variables $X_1, X_2, \ldots, X_n$ are discrete; i.e., they can only assume countably many values (e.g. all random variables take only values 0 or 1, or all possible outcomes are natural numbers). Then $X_1, X_2, \ldots, X_n$ are independent if and only if

$$P(X_1 = x_1, \ldots, X_d = x_d) = P(X_1 = x_1) \cdots P(X_d = x_d)$$

for all possible values of $x_1, \ldots, x_d$.

(b) Assume that the random variables $X_1, X_2, \ldots, X_n$ have densities (non-negative functions f_i such that $P(X_i \leq x) = \int_{-\infty}^{x} f_i(t)dt$ for all $i \in \{1, \ldots, n\}$ and $x \in \mathbb{R}$). Provided they have also a joint density; i.e., a non-negative function f such that $P(X_1 \leq x_1, \ldots, X_d \leq x_d) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \cdots \int_{-\infty}^{x_d} f(t_1, t_2, \ldots, t_d)dt_1 dt_2 \cdots dt_d$, then they are independent if and only if

$$f(x_1, x_2, \ldots, x_d) = f_1(x_1) f_2(x_2) \cdots f_d(x_d)$$

for all $x_1, \ldots, x_d \in \mathbb{R}$.

2.1 Misconceptions of the Independence Concept

Unfortunately, there are several popular misunderstandings regarding independence, which we shall discuss now.

Misconception 1: "Pairwise Independence Entails Independence" One may be tempted to believe that instead of checking the definition of independence, which involves all random variables, one could check whether all possible pairs of two variables are independent. Unfortunately, such pairwise independence does not imply independence in the sense of Definition 2.1 above. This is illustrated by the following example. Simple random variables (indicator variables) are defined via events A, B, C by 1_A, 1_B and 1_C, where 1_A is equal to one if the event A occurs and equal to zero else; analogously for B and C. Then our Definition 2.1 is consistent with the usual definition of independent events, which says that events A, B, C are independent, if $P(A \cap B \cap C) = P(A)P(B)P(C)$, $P(A \cap B) = P(A)P(B)$, $P(A \cap C) = P(A)P(C)$ and $P(B \cap C) = P(B)P(C)$ all hold. The following examples shows that independence of all pairs of indicator variables (or events) does not imply independence of all three indicator variables (or events).

Illustration 2.2 Think of two thunderstorms which we assume to be independent. We care only whether a thunderstorm comes accompanied by hail or not. The probability for a single thunderstorm to come with hail shall be $1/2$. Let A be the event that it hails during the first thunderstorm and B the event that it hails during the second thunderstorm. Finally, let C be the event that it either hails or does not hail in both the first and the second thunderstorm. One easily calculates $P(A) = P(B) = P(C) = 1/2$. However, $P(A \cap B \cap C) = P(A \cap B) = 1/4 \neq 1/8 = P(A)P(B)P(C)$, because if it hails both in the first and the second thunderstorm, the event C given by no hail in both thunderstorms can no longer occur. So clearly A, B, C are not independent. However, A, B are independent by construction and $P(A \cap C) = P(A \cap B) = P(B \cap C) = 1/4$; and thus we have pairwise independence.

Misconception 2: "Total Risk is Smallest/Largest for Independent Events" One cannot conclude in general that the situation of independent risks is particularly (un)favourable from the point of view of the total risk. The reason is that dependence can act both in a risk-reducing and risk-enhancing way, since typically risk measures are non-linear. We will present some real life examples and discuss the variance and the Value-at-Risk as risk measures.

Illustration 2.3 Assume that we are confronted with two different risks modelled by two random variables X, Y. Both random variables are either 0 or 1 (in some monetary unit like 1 million Euros), corresponding to no loss or loss of one monetary unit, each with probability $1/2$. For example, in an insurance company X, Y may describe whether or not damages have been reported for two different insurance contracts and the claim had to be paid (then the corresponding variable is 1, else it is zero). The insurer regards $X+Y$ as the random variable describing the total risk of both contracts.

When using the variance as risk measure, we simply have to apply the formula

$$\mathrm{var}(X+Y)=\mathrm{var}(X)+\mathrm{var}(Y)+2\,\mathrm{cov}(X,Y).$$

Consequently, the risk (in terms of the variance) is equal to the sum of the risks, if X and Y are uncorrelated (see Sect. 4 for the use of the correlation as a dependence measure). The risk of the sum is larger than the sum of risks, if X and Y are positively correlated, and likewise the risk of the sum is smaller than the sum of the risks if they are negatively correlated. For the VaR (as well as other more advanced risk measures) the situation is not quite as simple.

Situation 1: X and Y are independent (e.g. X models a life insurance contract and Y a personal liability insurance for the same person). Then the loss $X+Y$ is 0 with probability $1/4$, 1 with probability $1/2$, or 2 with probability $1/4$. When using the Value-at-Risk at the 90 % or 70 % level as risk measures, one obtains $\mathrm{VaR}_{0.9}(X+Y)=2$ and $\mathrm{VaR}_{0.7}(X+Y)=1$.

Situation 2: X,Y are "completely positive dependent" (e.g. X,Y are insurances against hurricanes for two neighbouring houses of same value; i.e., $X=Y$). Then the loss $X+Y$ is 0 with probability $1/2$, or 2 with probability $1/2$. It can never be 1 and one obtains $\mathrm{VaR}_{0.9}(X+Y)=\mathrm{VaR}_{0.7}(X+Y)=2$.

Situation 3: X,Y are "completely negative dependent" (e.g. X is an insurance cover for a farmer against too little rain measured by the annual amount of rain being below a level c, and Y is an insurance cover for a holiday resort at the same place against bad weather which pays 1 if the amount of rain is above the same level c; i.e., $X=1-Y$). Then the loss $X+Y$ is 1 with probability 1. It can never be 0 or 2 and one obtains $\mathrm{VaR}_{0.9}(X+Y)=\mathrm{VaR}_{0.7}(X+Y)=1$.

Comparing the values of the VaR for the two different levels in the three examples shows that the risk in the independent situation is neither an upper nor a lower bound on the risk in dependent situations. Note that in the last situation there is actually no risk at all in the sense of an uncertain outcome, because $X+Y$ is always equal to 1.

Note here also that typical risk measures are non-linear. This is in contrast to the expected value, which for $X+Y$ is in all situations equal to 1. Hence, our examples illustrate also that the expected value does not at all care about the dependence structure.

3 Normal Distribution

The normal (or Gaussian) distribution is the most widely used probability distribution in applications. Its popularity is due to the facts that it is rather easy to handle, that many properties are known completely explicitly, and often there are arguments that it is a natural distribution to use. By a classical result called the central limit theorem, one can argue that whenever a variable of interest is generated by the averaged results of many different small random effects, this random variable should be approximately normally distributed. However, this argument has to be used with care and one should always check in detail whether data at hand may reasonably come from a normal distribution.

Definition 3.1 A random variable X is said to be *normally distributed with mean* $\mu \in \mathbb{R}$ *and variance* $\sigma^2 > 0$, if it has a probability density given by

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}. \tag{3.1}$$

If $\mu = 0$ and $\sigma^2 = 1$, we speak of a *standard normal random variable*.

Dependence issues make sense only for at least two random variables, hence we now turn our focus to multivariate normal distributions. We summarize all risks in a (column) vector $\mathbf{X} = (X_1, \dots, X_d)^\top$. We also need the notion of a *positive definite* $d \times d$ matrix Σ; that is, a matrix which is symmetric (i.e., the transposed $\Sigma^\top = \Sigma$) and satisfies $\mathbf{x}^\top \Sigma \mathbf{x} > 0$ for all $\mathbf{x} \in \mathbb{R}^d$ not equal to the zero vector. We are now ready to define the multivariate normal distribution; cf. the book [11] for many interesting details.

Definition 3.2 A d-dimensional random vector $\mathbf{X}$ is called *normally distributed with mean* $\boldsymbol{\mu} \in \mathbb{R}^d$ *and covariance matrix* Σ (a positive definite $d \times d$ matrix), if it has probability density

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right), \quad \mathbf{x} \in \mathbb{R}^d. \tag{3.2}$$

If $\boldsymbol{\mu} = 0$ and $\Sigma = I_d$ (I_d being the $d \times d$-identity matrix), we speak of a *d-dimensional standard normal vector*.

Note that one can also define normal distributions with only a positive semidefinite covariance matrix Σ (i.e., a symmetric matrix satisfying $\mathbf{x}^\top \Sigma \mathbf{x} \geq 0$ for all $\mathbf{x} \in \mathbb{R}^d$). One way to do this is by demanding that $\mathbf{X} = \boldsymbol{\mu} + A\mathbf{Y}$ where $\mathbf{Y}$ is standard normally distributed (with lower dimension) and A is chosen such that $AA^\top = \Sigma$.

The parameter $\boldsymbol{\mu}$ is the mean vector of $\mathbf{X}$ and changing it shifts the distribution (i.e., it changes the location of the distribution in a non-random way). Hence, it has nothing to do with the dependence structure between the vector components $X_1, \dots, X_d$, which therefore must be totally described by Σ.

Each diagonal element Σ_{ii} of the matrix Σ gives the variance of the corresponding ith coordinate X_i, whereas the off-diagonal element Σ_{ij} with $i \neq j$ gives the covariance of X_i and X_j, a dependence measure we shall investigate in detail below.

In Fig. 1 we depict the densities of several bivariate normal distributions. For the standard normal density the surface is very homogeneous (it is left invariant

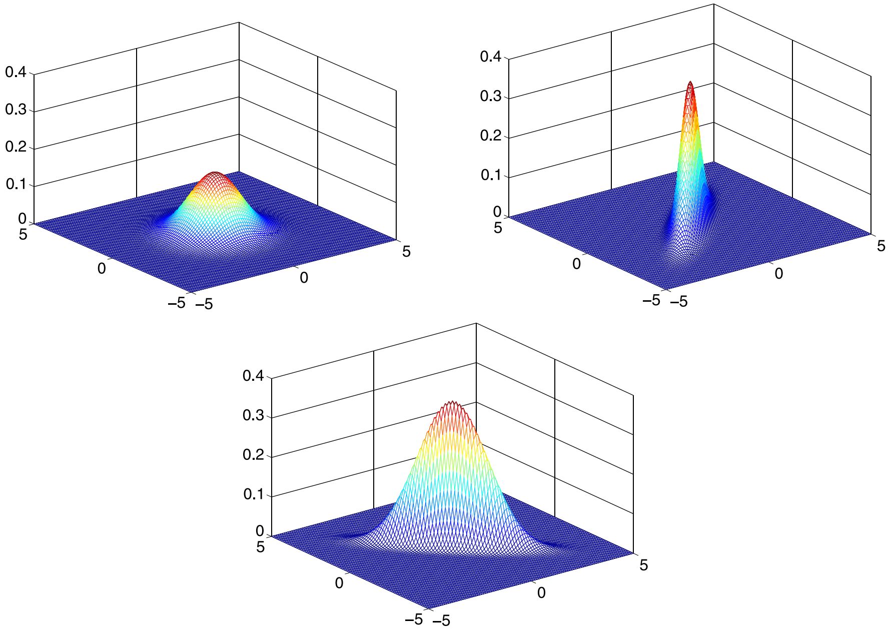


Fig. 1 Bivariate normal densities: standard normal (independent components; *upper left*), normal with variance 1 and covariance $\rho = 0.9$ (highly positively correlated; *upper right*), normal with variance 1 and covariance $\rho = -0.9$ (highly negatively correlated; *lower*)

by rotations), whereas in the two other cases the mass of the distribution (i.e., the area with a high value for the density) is concentrated around the diagonal (i.e., the line where $x_1 = x_2$), or the negative diagonal (i.e., the line where $x_1 = -x_2$), respectively. Intuitively it seems that in the standard normal distribution the two components X_1 and X_2 are rather independent, whereas in the other two cases they appear to be rather dependent. This intuition is indeed true.

However, there is more to be learned from these plots. A natural question is, what do the lines look like where the density has a fixed specified value; i.e., what are the sets of possible values (x_1, x_2) satisfying $f_{\mathbf{X}}(x_1, x_2) = c$ for some $c > 0$? From the plot of the density, we guess that for the standard normal density, these *contour lines* should be circles around the origin. Note that the standard normal density has its maximum at 0 with value $f_{\mathbf{X}}(0, 0) = 1/(2\pi)$. We calculate the following for $c \in (0, 1/(2\pi)]$ from (3.1) (by ln we denote the natural logarithm; i.e., the analytical inverse of the exponential function):

$$\begin{aligned} & f_{\mathbf{X}}(x_1, x_2) = c \\ \Leftrightarrow \quad & -\frac{1}{2}\left(x_1^2 + x_2^2\right) = \ln(2\pi c) \\ \Leftrightarrow \quad & x_1^2 + x_2^2 = -2\ln(2\pi c). \end{aligned}$$

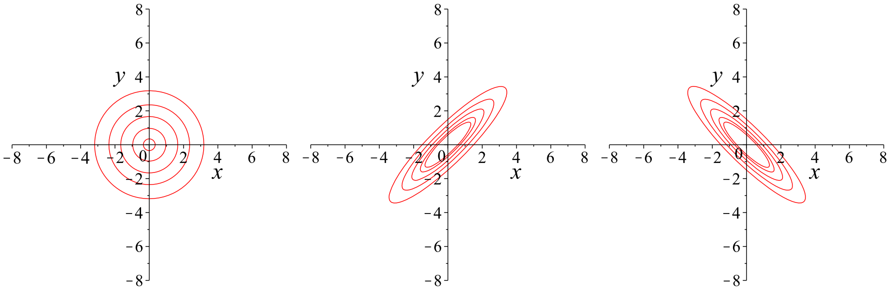


Fig. 2 Contour plots of bivariate normal densities: standard normal (independent components; *left*), normal with variance 1 and covariance $\rho = 0.9$ (highly positively correlated; *middle*), normal with variance 1 and covariance $\rho = -0.9$ (highly negatively correlated; *right*). The levels of the contours are $0.15, 0.1, 0.04, 0.01, 0.001$

From elementary geometry we recall that this last equation describes the circle around zero with radius $\sqrt{-2\ln(2\pi c)}$ (note that $2\pi c < 1$, and hence $\ln(2\pi c) < 0$).

In the general case (with arbitrary mean and covariance matrix) we may still assume that $\boldsymbol{\mu} = 0$, since the mean changes only the location, not the dependence structure. For arbitrary Σ the sets with equal values for the normal density can also be calculated and we obtain (again only for possible values of c) from Definition 3.2 and the formula for the explicit inversion of a 2×2 matrix:

$$\begin{aligned} & f_{\mathbf{X}}(x_1, x_2) = c \\ \Leftrightarrow \quad & \Sigma_{22}x_1^2 - 2\Sigma_{12}x_1x_2 + \Sigma_{11}x_2^2 = -2\det(\Sigma)\ln\big(2\pi\sqrt{\det(\Sigma)}c\big). \end{aligned}$$

Since this is again a quadratic equation, elementary geometry tells us that these sets are ellipses centred at the origin. As we shall also discuss in detail later on, the distributions where the contour lines of the density (the lines characterised by the density assuming the same value) are circles or, more generally, ellipses play a special role regarding the description of dependence.

4 Correlation as a Linear Dependence Measure

We now discuss the use of covariance or correlation as a measure of dependence. We start with a pair X, Y of random variables representing two different risks. Throughout this section we assume that all random variables have a finite variance; i.e., $E(X^2) < \infty$ (equivalently, $\int_{\mathbb{R}} x^2 f_X(x)dx < \infty$ if X has a density f_X).

Recall that the variance of a random variable X is given by $\operatorname{var}(X) = E((X - E(X))^2)$ and can be seen as a measure of the variability of the random variable or, in other words, how much the realisations of X tend to fluctuate around the mean value $E(X)$. Note that when X has a density f_X then its mean or expectation is $E(X) = \int_{\mathbb{R}} x f_X(x)dx$. The covariance of X and Y is given by $\operatorname{cov}(X, Y) = E((X -$

$E(X))(Y - E(Y))) = E(XY) - E(X)E(Y)$. From the first expression it is obvious that the covariance is a positive number if X and Y are "usually" both below or above their mean and negative if "usually" one is above its mean and one below.

The covariance carries information on the dependence, but is also affected by the variability (the typical spread around the mean) of the involved random variables. To get rid of the latter effect and to get a number measuring only dependence aspects one normalises the covariance by dividing the covariance by the product of the involved standard deviations (square roots of the variances).

Definition 4.1 (Correlation Coefficient) For two random variables with finite second moment the dependence measure

$$\rho(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{var}(X)\operatorname{var}(Y)}} \tag{4.1}$$

is called (*Pearson's*) *correlation coefficient*.

The correlation coefficient is usually estimated by its empirical version: given independent bivariate data $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ of joint observations from two random variables X and Y, respectively, the empirical correlation or correlation estimator is given by

$$\widehat{\rho}(X, Y) = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^n (X_i - \overline{X})^2 \sum_{i=1}^n (Y_i - \overline{Y})^2}}, \tag{4.2}$$

where

$$\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i \quad \text{and} \quad \overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i .$$

$\overline{X}$ is the empirical mean of the X_i and $\overline{Y}$ the empirical mean of the Y_i.

Classical results (the Cauchy-Schwarz inequality) ensure that the correlation of any two random variables has to be between -1 and 1 (as has also its empirical estimator), and for independent random variables $\operatorname{cov}(X, Y) = 0$ and, thus, the correlation $\rho(X, Y) = 0$ as well.

The correlation is a measure of linear dependence. In particular, perfect linear dependence is equivalent to $\rho(X, Y) = \pm 1$.

Theorem 4.2 *Two random variables X, Y are perfectly linearly dependent*; *i.e., $Y = aX + b$ with some $a \neq 0$ and $b \in \mathbb{R}$, if and only if $\rho(X, Y) = \pm 1$.*

Proof Assume first $Y = aX + b$. Then

$$\begin{aligned}\operatorname{cov}(X, Y) &= E\big(\big(X - E(X)\big)\big(aX + b - \big(aE(X) + b\big)\big)\big) = aE\big(\big(X - E(X)\big)^2\big) \\ &= a\operatorname{var}(X),\end{aligned}$$

$$\operatorname{var}(Y) = \operatorname{var}(aX + b) = a^2 \operatorname{var}(X)$$

and, hence, $\rho(X, Y) = a/\sqrt{a^2} = \pm 1$, depending on the sign of a.

To ease notation for the converse implication we set $\widetilde{X} = X - E(X)$ and $\widetilde{Y} = Y - E(Y)$ in the following. Assume now that

$$\rho(X, Y) = \frac{E(\widetilde{X}\widetilde{Y})}{\sqrt{E(\widetilde{X}^2)E(\widetilde{Y}^2)}} = \pm 1.$$

Then $E(\widetilde{X}^2)$, $E(\widetilde{Y}^2) > 0$ and we have that

$$E\big(\widetilde{Y}^2\big)\big(E\big(\widetilde{X}^2\big)E\big(\widetilde{Y}^2\big) - \big(E(\widetilde{X}\widetilde{Y})\big)^2\big) = 0.$$

However, calculations show that

$$E\big(\widetilde{Y}^2\big)\big(E\big(\widetilde{X}^2\big)E\big(\widetilde{Y}^2\big) - \big(E(\widetilde{X}\widetilde{Y})\big)^2\big) = E\big(\big(E\big(\widetilde{Y}^2\big)\widetilde{X} - E(\widetilde{X}\widetilde{Y})\widetilde{Y}\big)^2\big). \tag{4.3}$$

Since the expectation of a non-negative random variable is zero if and only if the random variable is zero (strictly speaking this has to hold only almost surely, but we ignore such technicalities), (4.3) implies that

$$Y - E(Y) = \frac{E(\widetilde{Y}^2)}{E(\widetilde{X}\widetilde{Y})}\big(X - E(X)\big)$$

and thus Y is of the form $aX + b$ as claimed. □

Proposition 4.3 (First Properties of Correlation) *Let X and Y be two random variables.*

(a) *Symmetry*:

$$\rho(X, Y) = \rho(Y, X).$$

(b) *Effect of linear transformations*:
For all $\alpha, \gamma \neq 0$ and $\beta, \delta \in \mathbb{R}$,

$$\rho(\alpha X + \beta, \gamma Y + \delta) = \operatorname{sign}(\alpha\gamma)\rho(X, Y),$$

where $\operatorname{sign}(x)$ *is equal to* $+1$ *for* $x > 0$ *and* -1 *for* $x < 0$. *Hence, the correlation is invariant under strictly increasing* linear *transformations (the case when* $\alpha, \gamma > 0$).

The concepts of covariance and correlation extend to multivariate random vectors as follows.

Definition 4.4 Let $\mathbf{X} = (X_1, \ldots, X_d)^\top$ be a d-dimensional and $\mathbf{Y} = (Y_1, \ldots, Y_m)^\top$ an m-dimensional random vector. Then we can take covariances and correlations between every pair of components of $\mathbf{X}$ and $\mathbf{Y}$ and summarize them in $d \times m$-matrices, called the *covariance matrix* and the *correlation matrix*:

$$\operatorname{cov}(\mathbf{X}, \mathbf{Y}) = \big(\operatorname{cov}(X_i, Y_j)\big)_{1\le i\le d, 1\le j\le m},$$
$$\operatorname{corr}(\mathbf{X}, \mathbf{Y}) = \big(\rho(X_i, Y_j)\big)_{1\le i\le d, 1\le j\le m}.$$

The covariance matrix of a random vector $\operatorname{cov}(\mathbf{X}, \mathbf{X})$ with itself is called the *covariance matrix of* $\mathbf{X}$ and we write $\operatorname{var}(\mathbf{X}) := \operatorname{cov}(\mathbf{X}, \mathbf{X})$.

Proposition 4.5 (Further Properties of Correlations and Covariances) *Let* $\mathbf{X} = (X_1, \dots, X_d)^\top$ *be a* d*-dimensional and* $\mathbf{Y} = (Y_1, \dots, Y_m)^\top$ *an* m*-dimensional random vector.*

(a) *Symmetry*:
$\operatorname{var}(\mathbf{X})$ *and* $\operatorname{corr}(\mathbf{X}, \mathbf{X})$ *are symmetric positive semi-definite matrices* (*cf. before Definition* 3.2).

(b) *Linear transformations*:

$$\operatorname{cov}(A\mathbf{X} + a, B\mathbf{Y} + b) = A \operatorname{cov}(\mathbf{X}, \mathbf{Y})B^\top$$

for every $n \times d$ *matrix* A, $k \times m$ *matrix* B *and every* $a \in \mathbb{R}^n$ *and* $b \in \mathbb{R}^k$.

(c) *Linear combinations*:
For every $a \in \mathbb{R}^d$ *the variance of the linear combination* $a^\top\mathbf{X}$ *is given by*

$$\operatorname{var}\big(a^\top\mathbf{X}\big) = a^\top \operatorname{cov}(\mathbf{X})a.$$

(d) *Additivity*:

$$\operatorname{cov}(\mathbf{X}, \mathbf{Y} + \mathbf{Z}) = \operatorname{cov}(\mathbf{X}, \mathbf{Y}) + \operatorname{cov}(\mathbf{X}, \mathbf{Z})$$

for every m*-dimensional random vector* $\mathbf{Z} = (Z_1, \dots, Z_d)^\top$.

Illustration 4.6 Suppose we model the water flow R (in litres per second) of a river at a certain point and assume that the river is formed by two independent rivers just a bit upstream. Let the water flow in the first river be R_1 and that in the second river R_2. Then $\operatorname{cov}(R_1, R_2) = \rho(R_1, R_2) = 0$ by the assumed independence. Clearly, it should hold that $R = R_1 + R_2$ (assuming some kind of equilibrium state). Thus $\operatorname{cov}(R, R_1) = \operatorname{cov}(R_1, R_1) + \operatorname{cov}(R_1, R_2) = \operatorname{var}(R_1)$ and hence

$$\rho(R, R_1) = \frac{\operatorname{var}(R_1)}{\sqrt{\operatorname{var}(R_1)\operatorname{var}(R)}} = \frac{\operatorname{var}(R_1)}{\sqrt{\operatorname{var}(R_1)(\operatorname{var}(R_1) + \operatorname{var}(R_2))}}$$
$$= \sqrt{\frac{\operatorname{var}(R_1)}{(\operatorname{var}(R_1) + \operatorname{var}(R_2))}}$$

and, likewise, if we replace R_1 by R_2. For example, if both original rivers; i.e., R_1 and R_2, have the same variance we get $\rho(R, R_1) = 1/\sqrt{2}$.

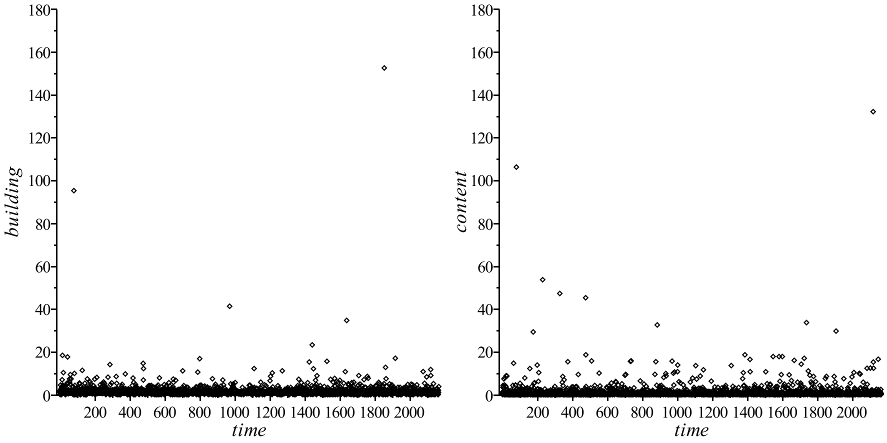


Fig. 3 Time series plot of the losses in building (*left*) and the losses in content of the Danish fire insurance data from 1980 to 1990. The time is in days starting January 3rd, 1980, leaving out weekends and holidays

Illustration 4.7 (Danish Fire) Throughout this paper we will illustrate the various dependence measures using a data set of Danish fire insurance claims from 1980 to 1990 available from http://www.ma.hw.ac.uk/~mcneil/data.html.

The original data set includes data on the losses of the fire insurance arising from the damage to the building, from the burnt content of the building, and from losses to profits (of companies in the burnt buildings). Since the last variable is zero in most cases, we consider only the losses of building and content. To avoid strange artefacts due to the fact that the data set considers only events where the total loss (sum of the loss in the three categories) exceeded one million Danish Kroner, we consider only events where both the losses in building and of content individually exceed this threshold.

In Fig. 3 we provide a time series plot of the data.

To assess the dependence we provide scatter plots of the loss data as well as the logarithms of the losses in Fig. 4. At the original scale it is hard to see what is going on in the majority of the observations, since they form a cloud at the origin and only the extreme events can be seen, for which it is hard to see any clear dependence structure. On the logarithmic scale one sees that there is no clear trend/dependence in the data, but that the two loss variables tend to behave similar and thus should be positively dependent. This can also be seen from the correlations which are 0.51 for the original data and 0.38 after taking logarithms.

Correlation is a very popular dependence measure. The reasons are that it can be easily estimated from data by its empirical version, and that it is the natural dependence measure for the multivariate normal distribution. In this model it describes the dependence of the random components completely, and also in the more general class of elliptical distributions.

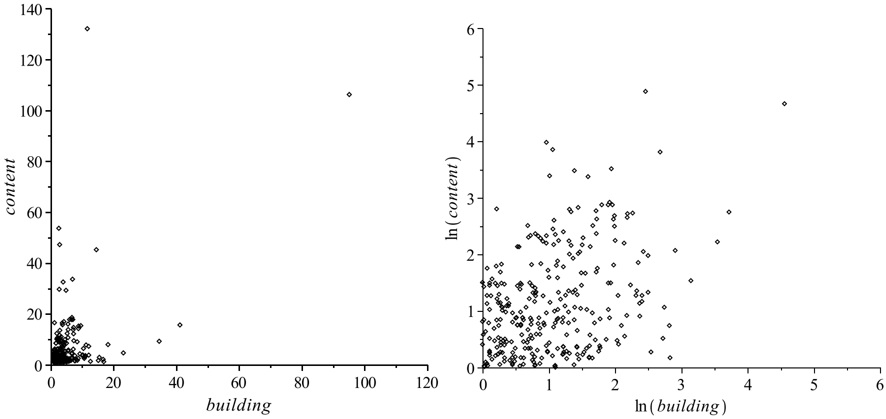


Fig. 4 Scatter plot of the Danish fire insurance data: losses of buildings and losses of content, original scale (*left*) and logarithmic scale (*right*)

4.1 Disadvantages of Correlation

Correlation has certain disadvantages that one should be aware of when using it.

(a) It is defined only when the variances of the random variables exist. In particular, for extreme risks this is not always guaranteed. A relevant example in the context of risk is the t-distribution with ν degrees of freedom with density $f(x) = c(1+x^2/\nu)^{-(d+\nu)/2}$, $x \in \mathbb{R}$. For two t-distributed random variables with $\nu \leq 2$ the correlation is not defined. Also for two Pareto-distributed random variables with densities $f_1(x) = \alpha_1/x^{\alpha_1+1}$, $x > 1$, and $f_2(x) = \alpha_2/x^{\alpha_2+1}$, $x > 1$, and shape parameters $\alpha_1 \leq 2$ or $\alpha_2 \leq 2$, the correlation is not defined.

(b) Two independent random variables with finite variances are uncorrelated. However, the converse is not true. There exists an abundance of cases where random variables are uncorrelated, but not independent.

On a simple level, if X is a standard normal random variable, and $Y = X^2$, then X and Y are obviously not independent, since X^2 is a function of X. However, $\text{cov}(X,Y) = \text{cov}(X,X^2) = E(X^3) - E(X)E(X^2) = 0$, since all odd moments of a normal random variable are equal to 0.

Examples on a more advanced level include variance mixtures of normal random variables (cf. Example 5.3) and, in a dynamic context, stochastic volatility models in finance and stochastic intermittency models in turbulent and other environmental data.

Only in special parametric models (the multivariate normal distribution is *the* typical example), does uncorrelatedness imply independence.

(c) Covariances and correlations depend on the distribution in a highly non-trivial way. For instance, if one knows only the correlation of X, Y, then nothing can be said about the correlation of $T(X), T(Y)$ for a non-linear increasing transformation T.

(d) The correlation depends on the whole distribution. However, in the context of risk one does not really care about the dependence for the "usual outcomes" but about the dependence of the extreme outcomes. The correlation thus typically provides at most very limited information about the dependence of risks.

4.2 Misconceptions of Correlation

Unfortunately, there are several popular misunderstandings regarding correlation which we shall explain now.

Misconception 1: "Marginals and Correlation Matrix Determine the Distribution" It is often wrongly thought that, if one knows the distributions of the random variables X_1 and X_2 and their correlation $\rho(X_1, X_2)$, then one knows already the bivariate distribution of the random vector $\mathbf{X} = (X_1, X_2)^\top$. This is false not just in general, but even in a normally distributed world. In particular, as we shall see in a moment, if X_1 and X_2 are known to be each standard normally distributed, and have correlation ρ, one cannot conclude that $(X_1, X_2)^\top$ is bivariate normally distributed with mean zero and covariance matrix $\Sigma = \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)$.

Illustration 4.8 Let X_1 be a standard normally distributed random variable and define X_2 by

$$X_2 = \begin{cases} X_1 & \text{if } |X_1| \le 1, \\ -X_1 & \text{if } |X_1| > 1. \end{cases}$$

Then X_2 is also standard normally distributed, because X_1 is and the standard normal distribution is symmetric around zero. Since both X_1 and X_2 have a finite variance, $\rho := \rho(X_1, X_2)$ exists and is some number in $(-1, 1)$, which is hard to compute explicitly. Note that it is clear that the correlation is different from ±1 because of Theorem 4.2. We now prove by contradiction that the random vector $(X_1, X_2)^T$ is not bivariate normally distributed. Thus, assume $(X_1, X_2)^T$ is bivariate normally distributed, then $X_1 + X_2$ is also normally distributed with mean 0 and variance $2 + 2\rho > 0$. However, from the construction of X_2 we see that

$$X_1 + X_2 = \begin{cases} 2X_1 & \text{if } |X_1| \le 1, \\ 0 & \text{if } |X_1| > 1. \end{cases}$$

Thus the probability that $X_1 + X_2$ is strictly bigger than two in absolute value is zero. Since this probability is strictly positive for every normally distributed random variable, we have the desired contradiction. Hence, our assumption that $(X_1, X_2)^T$ was bivariate normally distributed must be wrong.

Misconception 2: "In All Multivariate Models It Is Possible to Have All Values Between −1 and 1 as Correlation" Likewise, the belief is widespread that in every multivariate model one may have all values between -1 and 1 for the correlation. Unfortunately, not all combinations of valid pairwise correlations lead to a valid (i.e., positive semi-definite) overall correlation matrix.

However, this is not the only pitfall. Very often the model structure implies additional constraints on the correlation, such as having to be non-negative. The following is an example.

Illustration 4.9 Assume that an insurance company has sold insurance policies against damages by storm (S) and heavy rain (R). There are three types of insurance claims, those which regard damages by storm only, those which regard damages by heavy rain only, and those with both types of damages (caused e.g. by a thunderstorm with heavy rain and storm). We now want to model the number of claims for storm $S(t)$ which arrived up to time t (since the initial time 0), and the number of claims for rain $R(t)$ which arrived up to time t.

The classical insurance claim number model is a Poisson process (see e.g. Resnick [29]) for the arrivals of insurance claims. A Poisson process with rate (or frequency) $\lambda > 0$ is a counting process, where the number of claims at any time $t > 0$ is Poisson distributed with a mean linear in t with some rate $\lambda > 0$. We have $E(X(t)) = \lambda t$ and $\operatorname{var}(X(t)) = \lambda t$ for all times $t > 0$ for a Poisson process X. An alternative stochastic description of a Poisson process is as follows: it starts at zero at the initial time zero. After an exponentially distributed (with mean $1/\lambda$) waiting time, during which it remains 0, it jumps to one. Afterwards it remains again constant for an exponentially distributed (with mean $1/\lambda$) waiting time and then it jumps to two and so on. The rate λ gives the mean number of jumps (all of height one) in a unit time interval.

We use three independent Poisson processes, $\{N^R(t)\}_{t\geq 0}$ giving the arrival of claims regarding only heavy rain, $\{N^S(t)\}_{t\geq 0}$ giving the arrival of claims regarding only storm and $\{N^B(t)\}_{t\geq 0}$ giving the arrival of claims regarding both. The corresponding rates will be denoted λ^R, λ^S and λ^B. Clearly, we have $R(t) = N^R(t) + N^B(t)$ and $S(t) = N^S(t) + N^B(t)$ for $t \geq 0$ and we want to understand the dependence of $R(t)$ and $S(t)$. The process $R(t)$ is (as a sum of Poisson processes) again a Poisson process with rate (or frequency) $\lambda^S + \lambda^B$ and $S(t)$ is one with rate $\lambda^R + \lambda^B$. Hence, for all $t \geq 0$, we have

$$\rho\big(R(t), S(t)\big) = \frac{\operatorname{cov}(R(t), S(t))}{\sqrt{\operatorname{var}(R(t))\operatorname{var}(S(t))}} = \frac{\operatorname{var}(N^B(t))}{\sqrt{\operatorname{var}(R(t))\operatorname{var}(S(t))}}$$

$$= \frac{\lambda^B}{\sqrt{(\lambda^B + \lambda^R)(\lambda^B + \lambda^S)}}.$$

In this model the correlation can only be between 0 and 1.

Assume further that we have already done univariate modelling of both R, S and obtained Poisson processes with rates μ^R and μ^S and then consider the joint model. We must then have that $\lambda^B + \lambda^R = \mu^R$ and $\lambda^B + \lambda^S = \mu^S$ to be consistent with the univariate models. Hence, $\lambda^B \leq \min\{\mu^R, \mu^S\}$ is immediate, interpreting the rates as the frequencies of the arrival of claims. Going back to our correlation we get for all

$t \geq 0$ that

$$\rho\big(R(t), S(t)\big) = \frac{\lambda^B}{\sqrt{\mu^R \mu^S}} \leq \min\left\{\sqrt{\frac{\mu^R}{\mu^S}}, \sqrt{\frac{\mu^S}{\mu^R}}\right\}.$$

If $\mu^R \neq \mu^S$, the possible correlations are thus below an upper bound strictly smaller than one. This result has been obtained in the framework of Operational Risk in Böcker and Klüppelberg [16, Eq. (11)].

For more details on the problematic issues of correlation we refer to [1, 2].

5 Spherical and Elliptical Distributions

We have already seen that the contours of equal density are circles in the standard normal bivariate distribution and ellipses in the non-standard normal case. Likewise, one can show that in general dimensions the contours of equal density of the normal distribution are ellipsoids, and are spheres in the standard normal case (actually whenever all components are independent; i.e., all off-diagonal entries of the covariance matrix are zero, and have the same variance).

The spherical distributions extend the standard normal distribution $N_d(\mathbf{0}, I_d)$ (i.e., the distribution of d independent standard normal components). The density of a spherical distribution satisfies

$$f(\mathbf{x}) = \psi\big(\mathbf{x}^\top \mathbf{x}\big), \quad \mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$$

where $\psi : \mathbb{R} \to \mathbb{R}^+$ is an appropriate function.

Examples are the multivariate t-distribution with ν degrees of freedom with density $f(\mathbf{x}) = c(1 + \mathbf{x}^\top \mathbf{x}/\nu)^{-(d+\nu)/2}$ and the logistic distribution with density $f(\mathbf{x}) = c \exp(-\mathbf{x}^\top \mathbf{x})/(1 + \exp(-\mathbf{x}^\top \mathbf{x}))^2$. Here c are the norming constants, which guarantee the densities to integrate to 1. It should be noted that random variables with a non-normal joint distribution that is spherical are uncorrelated random variables, which however are not independent (see e.g. [24]).

There are various ways to think about a spherical distribution.

(i) From the densities above we see that the contours of equal density are circles in the bivariate models; i.e., "spheres" in arbitrary dimensions.

(ii) Equivalently, we can think of a spherical random vector $\mathbf{X}$ as having the same distribution under every orthogonal transformation; i.e., if we multiply it by a $d \times d$ matrix M with the property that $M^\top M = M M^\top = I_d$, then $M\mathbf{X}$ has the same distribution as $\mathbf{X}$.

(iii) Finally, a spherical random vector $\mathbf{X}$ has the same distribution as $R\mathbf{U}$, where $\mathbf{U}$ is uniformly distributed on the unit sphere $\mathcal{S}_{d-1} = \{\mathbf{s} \in \mathbb{R}^d : \mathbf{s}^\top \mathbf{s} = 1\}$, and R is a positive random variable, independent of $\mathbf{U}$.

Elliptical distributions generalize multivariate normal distributions $N_d(\boldsymbol{\mu}, \Sigma)$ with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ, and also have contours of equal density which are ellipsoids. Moreover, just as ellipsoids are linear transformations of spheres, elliptical distributions are obtained as linear transformations of spherical distributions.

For a general treatment of elliptical distributions we refer to Fang, Kotz, and Ng [4].

Definition 5.1 A random vector $\mathbf{X} \in \mathbb{R}^d$ has an *elliptical distribution* if there exist $\boldsymbol{\mu} \in \mathbb{R}^d$, a positive semi-definite $d \times d$ matrix $\Sigma = (\sigma_{ij})_{1 \le i,j \le d}$, a positive random variable G and a random vector $\mathbf{U}^{(d)} \sim \text{unif}\{\mathbf{s} \in \mathbb{R}^d : \mathbf{s}^\top \mathbf{s} = 1\}$ (i.e., $\mathbf{U}^{(q)}$ is uniformly distributed on the unit sphere in $\mathbb{R}^d$) independent of G such that $\mathbf{X}$ satisfies ($\overset{d}{=}$ means that the distributions of the random variables on both sides are equal)

$$\mathbf{X} \overset{d}{=} \boldsymbol{\mu} + GA\mathbf{U}^{(d)} \quad \text{with } A \in \mathbb{R}^{d \times d} \text{ and } AA^\top = \Sigma. \tag{5.1}$$

We write $\mathbf{X} \sim \mathcal{E}_d(\boldsymbol{\mu}, \Sigma, G)$.

The random variable G is called the *generating variable*. Furthermore, if the first moment exists, then $E(\mathbf{X}) = \boldsymbol{\mu}$, and if the second moment exists, then G can be chosen such that $\text{var}(\mathbf{X}) = \Sigma$.

Note that we write $\mathbf{X} \sim \mathcal{E}_d(\boldsymbol{\mu}, \Sigma)$ if we consider only quantities which do not depend on the concrete generating random variable G, and we denote $E(\mathbf{X}) = \boldsymbol{\mu}$, $\text{var}(\mathbf{X}) = \Sigma$, provided they exist.

Furthermore, note that in the following we always call $\Sigma = AA^\top$ the covariance matrix (its elements the covariances) of an elliptical distribution even if the second moments do not exist.

In elliptical models covariances and correlations are natural dependence measures. This is a consequence of the following properties:

Proposition 5.2 (Properties of Elliptical Distributions) *Let $\mathbf{X} \sim \mathcal{E}_d(\boldsymbol{\mu}, \Sigma)$ be elliptically distributed.*

(a) *Consider the map $T(\mathbf{X}) = B\mathbf{X} + \mathbf{b}$ for a $q \times d$-matrix B and a vector $\mathbf{b} \in \mathbb{R}^q$. Then $B\mathbf{X} + \mathbf{b} \sim \mathcal{E}_q(B\boldsymbol{\mu} + \mathbf{b}, B\Sigma B^\top)$.*

(b) *From this follows immediately that all marginal distributions of $\mathbf{X}$ are elliptical; in particular, the components of $\mathbf{X}$ are one-dimensional elliptical, which means they are symmetric around their means (or the median, if the mean does not exist).*

Moreover, for an arbitrary component X_i there are $a > 0, b \in \mathbb{R}$ such that $X_i \overset{d}{=} aX_1 + b$, where instead of X_1 we could have chosen any other component. Hence, in distribution any component can be realised as a linear transformation of one fixed component.

Let $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)^\top \sim \mathcal{E}_d(\boldsymbol{\mu}, \Sigma)$ *with* $\mathbf{X}_1 \in \mathbb{R}^p$, $\mathbf{X}_2 \in \mathbb{R}^q$ *with* $p + q = d$. *Let* $\boldsymbol{\mu} = (\mu_1, \mu_2)^\top$ *with* $\mu_1 \in \mathbb{R}^p$, $\mu_2 \in \mathbb{R}^q$, *and* $\Sigma = \left(\begin{smallmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{smallmatrix}\right)$. *Then*

$$\mathbf{X}_1 \sim \mathcal{E}_p(\mu_1, \Sigma_{11}) \quad \textit{and} \quad \mathbf{X}_2 \sim \mathcal{E}_q(\mu_2, \Sigma_{22}).$$

Hence, subvectors of elliptically distributed random vectors are again elliptically distributed, and the parameters are known explicitly.

(c) *Assume that* Σ *is positive definite. The conditional distribution of* $\mathbf{X}_1$ *given* $\mathbf{X}_2$ *is also elliptical:*

$$\mathbf{X}_1 \mid \mathbf{X}_2 \sim \mathcal{E}_p(\mu_{1|2}, \Sigma_{11|2}),$$

where $\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{X}_2 - \mu_2)$ *and* $\Sigma_{11|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

(d) *Every elliptical distribution is uniquely determined by the mean, the covariance matrix* Σ, *and the distribution of the generating random variable* G.

A very important class of elliptical distributions is given by the normal variance mixture models.

Example 5.3 (Normal Variance Mixture Model) (a) Let $\mathbf{X} \overset{d}{=} \boldsymbol{\mu} + \sqrt{W} A\mathbf{Z}$ with $\boldsymbol{\mu} \in \mathbb{R}^d$, $A \in \mathbb{R}^{d\times m}$ a matrix of rank $d < m$, $\mathbf{Z} \in \mathbb{R}^m$ a standard normal vector and $W > 0$ a random variable, independent of $\mathbf{Z}$. Then $\mathbf{X}$ is said to follow a normal variance mixture model, and one can show that the contours of equal density are ellipsoids, hence it is an elliptical distribution.

(b) In the situation of part (a), if W has an inverse gamma distribution with parameters $(\frac{\nu}{2}, \frac{\nu}{2})$, then for ν an integer, $\nu/W \sim \chi^2_\nu$, i.e. ν/W is χ^2 (chi-square) distributed with ν degrees of freedom. This implies that $\frac{1}{d}(\mathbf{X}-\boldsymbol{\mu})^\top\Sigma^{-1}(\mathbf{X}-\boldsymbol{\mu}) \sim \frac{\nu\chi^2_d}{d\chi^2_\nu}$, which is $F(d, \nu)$-distributed (recall that $\Sigma = AA^\top$).

Moreover, we have $\mathbf{X} - \boldsymbol{\mu} = \frac{A\mathbf{Z}}{\sqrt{W}} \sim \boldsymbol{t}_\nu(0, \Sigma)$; i.e., $\mathbf{X} - \boldsymbol{\mu}$ is a d-dimensional t-distributed vector with ν degrees of freedom. Further, if $\nu > 2$, then $\mathbf{X} - \boldsymbol{\mu}$ has covariance matrix $\frac{\nu}{\nu-2}\Sigma$. If $\nu \leq 2$ the covariance matrix does not exist.

Hence, the t-distribution—occurring frequently in statistics—is an example of a normal variance mixture. It is often used in risk management as an alternative to the normal distribution, because it puts more mass on large events (cf. Fig. 5) and, in its multivariate version, it allows for modelling joint large events (cf. Example 8.5(c)).

Some contour plots for the densities of t-distributions can be found in Fig. 5. As can be seen they are quite similar to the corresponding plots for the normal distribution in Fig. 2, but especially for small ν the density decays much more slowly than a normal density.

6 Rank Correlations

Correlations depend on the underlying distribution, and may even not exist (when there is no finite second moment). Non-parametric and robust alternatives have been

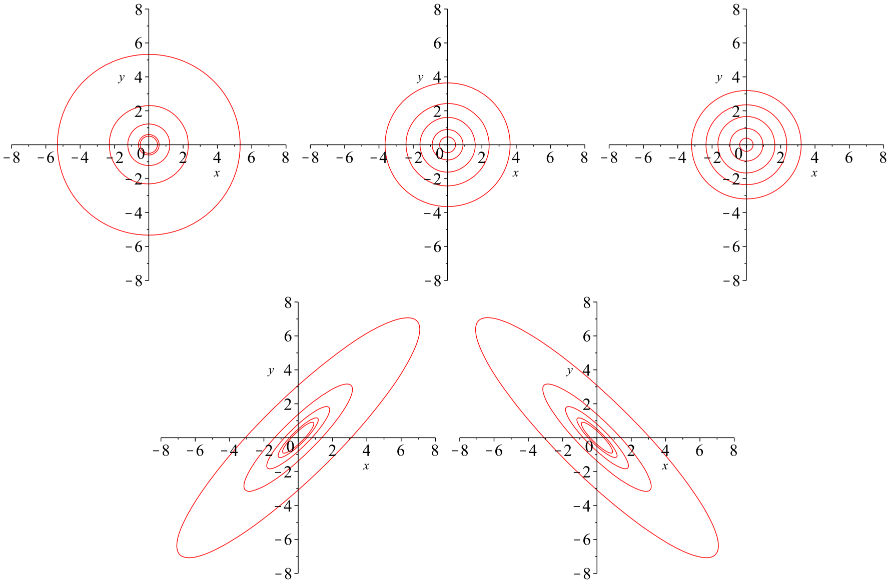


Fig. 5 Contour plots of bivariate t_ν-densities: *upper row*: uncorrelated components; i.e., Σ is the identity matrix; different degrees of freedom: $\nu = 1$ (*left*), $\nu = 10$ (*middle*), $\nu = 500$ (*right*). *Lower row* ($\nu = 1$): strongly correlated components, with $\rho = 0.9$ (*left*), and $\rho = -0.9$ (*right*). The levels for the individual contour lines are the same as in Fig. 2

proposed, which are based only on the ranks of the observations. Here ranking refers to a data transformation where numerical or ordinal values are replaced by their ranks. For instance, if numerical data 1.7, 9.3, 7.2 and 5.3 are observed, then the ranks of these data would be 1, 4, 3, 2. The actual sizes of the data are completely ignored. Obviously, ranking is not unique, when data of equal value are observed. There is a simple way how to deal with these so-called *ties*, and we explain this by an example. Assume that we observe 1.7, 7.2, 9.3, 7.2 and 5.3; then we would take the mean rank for the two equal observations, and obtain ranks 1, 3.5, 5, 3.5, 2. One deals similarly with 3 or more equal values.

Often this situation is excluded from the beginning by requiring that the underlying distribution has a density. Then (with probability 1) equal values do not happen in a sample.

Definition 6.1 (Spearman's Rank Correlation Coefficient) Let X, Y be random variables with continuous distribution functions F_1, F_2 and joint distribution function F. Let ρ be Pearson's correlation coefficient from Definition 4.1. Then *Spearman's rank correlation* is given by

$$\rho_S(X, Y) = \rho\big(F_1(X), F_2(Y)\big).$$

We have to explain why this is a rank correlation coefficient. Recall that for a distribution function F we denote by F^{-1} its generalized inverse function as defined in (1.1) and recall that F^{-1} is the analytic inverse of F, if F is strictly increasing.

First of all note that $F_1(X)$ is a random variable with values in $[0, 1]$. Moreover, since F_1 is continuous, $P(F_1(X) \le x) = P(X \le F_1^{-1}(x)) = F_1(F_1^{-1}(x)) = x$ for $x \in [0, 1]$. This implies that $F_1(X)$ is a standard uniform random variable (i.e., it is uniformly distributed on the interval $[0, 1]$). Consequently, ρ_S measures the correlation between two uniform random variables, and the original sizes of X and Y have become irrelevant.

One can say that rank correlations measure the degree of monotone dependence.

Let $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ be independent bivariate observations from two random variables X and Y, such that all the values of (X_i) and (Y_i) are different (there are no ties).

We estimate Spearman's rank correlation coefficient by its empirical version, which is based on replacing $F_1(X)$ and $F_2(Y)$ by their empirical versions. To this end the data $(X_1, Y_1), \dots, (X_n, Y_n)$ are converted into ranks, which we denote by $(\operatorname{rank}(X_i), \operatorname{rank}(Y_i))$ and the empirical correlation coefficient as given in (4.2) is calculated for these ranks.

The formula simplifies by virtue of the fact that $\frac{1}{n}\sum_{i=1}^n \operatorname{rank}(X_i) = \frac{1}{n}\sum_{i=1}^n i = \frac{n+1}{2}$, and

$$\sum_{i=1}^n \left(\operatorname{rank}(X_i) - \frac{n+1}{2}\right)^2 = \sum_{i=1}^n \left(\operatorname{rank}(Y_i) - \frac{n+1}{2}\right)^2 = \sum_{i=1}^n \left(i - \frac{n+1}{2}\right)^2$$
$$= \frac{1}{12} n\left(n^2 - 1\right).$$

Then the empirical Spearman's rank correlation coefficient is given by

$$\widehat{\rho}_S(X, Y) = \frac{1}{2} n\left(n^2 - 1\right) \sum_{i=1}^n \left(\operatorname{rank}(X_i) - \frac{n+1}{2}\right)\left(\operatorname{rank}(Y_i) - \frac{n+1}{2}\right).$$

Definition 6.2 (Kendall's Rank Correlation) Let (X_1, Y_1) and (X_2, Y_2) be independent random vectors with bivariate distribution function F. Then *Kendall's tau* is given by

$$\tau(X, Y) = P\big((X_1 - X_2)(Y_1 - Y_2) > 0\big) - P\big((X_1 - X_2)(Y_1 - Y_2) < 0\big).$$

The dependence Kendall's tau captures is better understood in its empirical version. Let $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ be a sample of bivariate observations from two random variables X and Y, such that all the values of (X_i), and respectively (Y_i), are different. Any pair of observations (X_i, Y_i) and (X_j, Y_j) are said to be *concordant*, if the ranks for both elements agree: that is, if both $X_i > X_j$ and $Y_i > Y_j$ or if both $X_i < X_j$ and $Y_i < Y_j$. They are said to be *discordant*, if $X_i > X_j$ and $Y_i < Y_j$ or if $X_i < X_j$ and $Y_i > Y_j$.

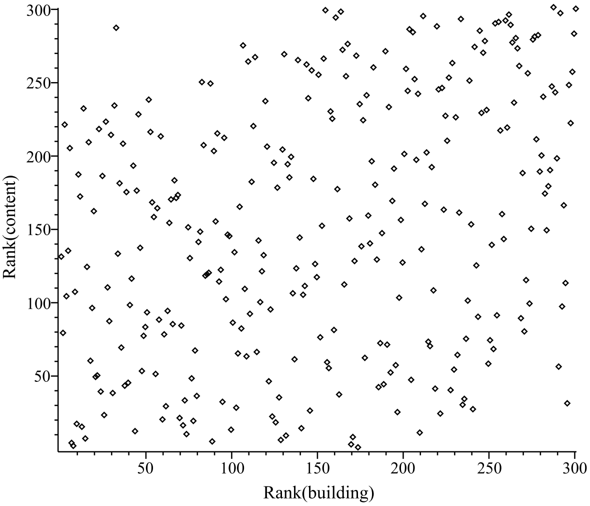


Fig. 6 Scatter plot of the Danish fire insurance data losses in building and losses in content after conversion to ranks

Definition 6.3 (Empirical Kendall's Rank Correlation Coefficient) The empirical version of Kendall's rank correlation is defined as:

$$\widehat{\tau} = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{\frac{1}{2}n(n-1)}$$

$$= \frac{2}{n(n-1)} \sum_{1 \leq i \leq j \leq n} \operatorname{sign}\big((X_i - X_j)(Y_i - Y_j)\big).$$

Note that the sign is equal to 1 whenever the two pairs are concordant, and it is -1, whenever the two pairs are discordant.

Rank correlation coefficients share some of the properties of Pearson's correlation coefficient: they are symmetric, lie between -1 and 1, and if X and Y are independent, they are equal to 0. Moreover, since they are based on ranks, rank correlations are invariant with respect to increasing transformations; i.e., if $T(x) \leq T(y)$ for all $x < y$, then $\rho_S(T(X), T(Y)) = \rho_S(X, Y)$, and the same holds for Kendall's tau.

Both Kendall's τ and Spearman's ρ can be calculated from the copula of a bivariate random vector with continuous marginal distributions (for a proof see Sect. 5.2.3 of McNeil, Frey, and Embrechts [8]); see next section for definitions and discussions of copulae. This means that both rank correlation coefficients are defined by the dependence structure only and not the marginal distributions.

Intuitively, both dependence measures check whether the ranks are similar, but there are important differences in what they actually measure, which are rather technical and thus beyond the scope of this introductory chapter (see [21, 27]).

Illustration 6.4 (Danish Fire Continued) In Fig. 6 the ranks of the losses in building are plotted against the ranks of the losses of content. The fact that there are very few points at the lower right and upper left corner hints again at positive dependence. Indeed, we obtain for the empirical versions of Spearman's ρ the estimate $\widehat{\rho}_S = 0.32$ and of Kendall's τ the estimate $\widehat{\tau} = 0.21$.

7 Copulae

The idea of modelling dependence in terms of ranks culminates in the concept of a copula. A copula describes the dependence structure completely and thus is in general a very complex object.

We start by recalling that for a random variable X with continuous distribution function F (recall that then we have, with probability 1, no ties in the observations) the transformed random variable $U := F(X)$ has a standard uniform distribution (i.e., is uniformly distributed on the interval [0, 1]).

This concept is now extended to a multivariate distribution as follows. Let $\mathbf{X} = (X_1, \dots, X_d)^\top$ be a random vector with distribution function F, and let F_j denote the marginal distribution function of X_j for $j = 1, \dots, d$. If all F_j are continuous functions, then we can do the same transformation as above, componentwise, which yields a random vector $(F_1(X_1), \dots, F_d(X_d))^\top$ taking values only in the unit cube $[0, 1]^d$. Note that all components of this vector are standard uniform random variables. This motivates the following definition.

Definition 7.1 (Copula) A copula is the joint distribution function of marginally uniformly distributed random variables. More precisely, if $U_1, \dots, U_d$ are $U(0, 1)$, then the function $C : [0, 1]^d \to [0, 1]$ defined by

$$C(u_1, \dots, u_d) = P(U_1 \le u_1, \dots, U_d \le u_d)$$

is a copula.

Applying this concept to the componentwise transformed random variables above, the vector $(F_1(X_1), \dots, F_d(X_d))^\top$ has distribution function given by

$$C_F(u_1, \dots, u_d) = P\big(F_1(X_1) \le u_1, \dots, F_d(X_d) \le u_d\big)$$

for $(u_1, \dots, u_d)^\top \in [0, 1]^d$. C_F is the copula of the vector $(X_1, \dots, X_d)^\top$.

In the way we have defined/constructed a copula above, it covers only the continuous case. The case of non-continuous random variables can be covered as well, but this becomes much more technical. A thorough introduction to copulae can be found in the book by Nelsen [9], for instance, or in [3, 8], which are of special interest in connection with risk modelling.

Before we discuss the use of copulae in risk analysis further, we present some examples. We formulate them for $d = 2$, and for most of the models it should be obvious, how they generalize to arbitrary dimension d.

Example 7.2 (Bivariate Copula Families) Let $u_1, u_2 \in [0,1]^2$.

(a) Independence copula:

$$C^{ind}(u_1, u_2) = u_1 u_2.$$

As the name already suggests, this is the copula of two independent random variables. Recall that two random variables are independent if and only if their joint distribution function is the product of the marginals. This is inherited by the copula.

(b) Copula of perfect dependence:

$$C^{dep}(u_1, u_2) = \min(u_1, u_2).$$

This copula models the situation, when the observations are perfectly dependent. For the two uniform random variables corresponding to the copula this means that they are identical. In general two random variables X, Y have the copula of perfect dependence if and only if there exists a random variable Z and two increasing functions f and g such that $X = f(Z)$ and $Y = g(Z)$. Note that intuitively this means that as soon as you know the value of one variable you also know the value of the other random variable for sure.

(c) Normal copula: for $\theta \in (-1, 1)$,

$$\begin{aligned} &C^{No}(u_1, u_2; \theta) \\ &\quad = \Phi_2\big(\Phi^{-1}(u_1), \Phi^{-1}(u_2)\big) \\ &\quad = \frac{1}{2\pi\sqrt{1-\theta^2}} \int_{-\infty}^{\Phi^{-1}(u_1)} \int_{-\infty}^{\Phi^{-1}(u_2)} \exp\left(\frac{-(x_1^2 - 2\theta x_1 x_2 + x_2^2)}{2(1-\theta^2)}\right) dx_1 dx_2, \end{aligned}$$

where Φ_2 and Φ denote the distribution functions of the bivariate and the univariate standard normal distribution, respectively, and Φ^{-1} is the inverse function of the cumulative standard normal distribution function Φ.

Again the name already tells us the idea behind this copula. It is the copula of two standard normally distributed random variables with correlation θ which are also jointly normally distributed. A sample from this copula can easily be obtained by drawing from a bivariate standard normal distribution with correlation θ and then applying the function Φ^{-1} to every coordinate.

Note that $\theta = 0$ gives the independence copula, whereas $\theta = 1$ gives the copula of perfect dependence. For $\theta = -1$ one obtains perfect negative dependence (i.e., the copula $\max(u_1 + u_2 - 1, 0)$ which, in contrast to the other examples, is a copula only for dimension $d = 2$).

As mentioned before and explained in more detail and by examples in Chap. 6, [20], extreme value models are important for risk management. When considering copula models in the context of bivariate extreme value models, so-called *extreme value copulae* occur. These copulae have to be of a very special form; i.e., their dependence structure can be represented in terms of a so-called Pickands dependence function A, a convex function satisfying $\max(s, 1-s) \leq A(s) \leq 1$ for all $s \in [0, 1]$;

see e.g., Beirlant, Goegebeur, Segers, and Teugels [14, Chap. 8.2.5]. In terms of such a Pickands dependence function an extreme value copula C has the form

$$C(u_1,u_2)=\exp\left\{\ln(u_1u_2)A\left(\frac{\ln(u_2)}{\ln(u_1u_2)}\right)\right\}. \tag{7.1}$$

Note that the right hand side is equal to u_1u_2 for the Pickands dependence function $A\equiv 1$; this is the independent case. A quantity often considered and estimated is the value in (7.1) for $u_1=u_2$; i.e. $A(\frac{1}{2})$. For symmetric copulae it is the minimum of A, hence gives a measure of maximal dependence in the model. We come back to this in Sect. 8.

Example 7.3 (Extreme Value Copulae and Their Pickands Dependence Function) Throughout $u_1,u_2\in[0,1]^2$ and $s\in[0,1]$.

(a) Gumbel copula:

Using the Pickands dependence function the Gumbel copula with parameter $\theta\in[1,\infty)$ is given by

$$A^{Gu}(s)=\left(s^\theta+(1-s)^\theta\right)^{1/\theta}.$$

Elementary calculations show that the Gumbel copula is thus

$$C^{Gu}(u_1,u_2)=\exp\left\{-\left(\left(-\ln(u_1)\right)^\theta+\left(-\ln(u_2)\right)^\theta\right)^{1/\theta}\right\}. \tag{7.2}$$

For $\theta=1$ the Gumbel copula is actually the independence copula, whereas for $\theta\to\infty$ the Gumbel copula converges to the copula of perfect dependence. Thus the Gumbel copula allows modelling a continuum of possible dependencies from independence to perfect positive dependence, giving a nice parametric model for different dependence scenarios.

(b) t-EV copula:

Using the Pickands dependence function the t-EV copula with parameter $\boldsymbol{\theta}=(\theta_1,\theta_2)\in(0,\infty)\times(-1,1)$ is given by

$$\begin{aligned}A^{t-EV}(s;\boldsymbol{\theta})=\,&st_{\theta_1+1}\left(\frac{(\frac{s}{1-s})^{1/\theta_1}-\theta_2}{\sqrt{1-\theta_2^2}}\sqrt{\theta_1+1}\right)\\&+(1-s)t_{\theta_1+1}\left(\frac{(\frac{1-s}{s})^{1/\theta_1}-\theta_2}{\sqrt{1-\theta_2^2}}\sqrt{\theta_1+1}\right),\end{aligned}$$

with t_ν for $\nu\in(0,\infty)$ representing the distribution function of the t_ν-distribution (i.e., the t-distribution with ν degrees of freedom). The t-EV copula (with "EV" standing for "extreme value") arises as the limiting dependence structure of componentwise maxima of independent and identically distributed bivariate t_{θ_1}-distributed random variables with the correlation of the underlying bivariate normal distribution being θ_2. For more details see e.g. [19].

Statistically, parametric copulae are rather easy to fit, since it is not necessary to specify marginal models. One can simply take the empirical distribution functions, plug them into a parametric copula model and estimate the copula parameters, for instance by likelihood methods. Various copula models are presented in Haug, Klüppelberg, and Peng [5], where also R codes for fitting such copula models are provided. The problem is obviously the choice of the parametric model.

Abstractly speaking a copula encodes the dependence structure of a d-dimensional random vector by transforming it to a d-dimensional random vector with standard uniform margins. In principle, one could just as well transform it to any other d-dimensional random vector with prescribed marginals to encode the dependence structure. So the question arises whether the use of a copula is the best way to transform data. Alternative transformations are indeed used in relation to some special applications. For instance, in reliability theory marginals have been transformed to normal random variables, which is admittedly not as easy as the transformation to uniform, since the normal distribution function is given as an integral, which cannot be calculated explicitly. See [10, 22, 26] for details.

Experts from extreme value theory often normalize marginals to standard extreme value distributions, when interested in the maximum of a sample. Typically the standard Fréchet distribution is used (see e.g. Proposition 5.10 of [29] for more details). Here the transformation is given by $-1/\ln(F(X))$, which has distribution function $P(-1/\ln(F(X)) \le z) = \exp\{-1/z\}1_{[0,\infty)}(z)$. When interested in the minimum of a sample, often the transformation is to the standard exponential distribution given by $F(x) = (1-e^{-x})1_{[0,\infty)}(x)$ (i.e., the transformation is $-\ln(1-F(X))$); cf. e.g. [23] for multivariate exponential distributions.

A Taylor expansion to the standard Fréchet distribution function gives $P(-1/\ln(F(X)) > z) \sim 1/z$ (equivalently, $zP(-1/\ln(F(X)) > z) \to 1$) as $z \to \infty$, so that large values of z happen with substantial probability (in particular compared to the normal distribution where $P(N(0,1) > z) \sim z\phi(z) = (\sqrt{2\pi}z)^{-1}\exp\{-(z^2/2)\}$ as $z \to \infty$ (ϕ denotes the standard normal density). Taking $z = 10$, one obtains for the Fréchet distribution the probability 0.09516258 and for the standard normal distribution 7.619853×10^{-24}. For the uniform distribution, no value larger than 1 can happen (with probability 1). As you can see in Fig. 7, it may be advantageous to transform data to Fréchet marginals when interested in the dependence structure of extreme events, as then the extremes really stick out.

Illustration 7.4 Because of the simple transformation in the marginals the use of copulae to model dependence has had a striking success in particular in the financial industry. The copula mostly applied has been the normal copula which means that in the end all dependence is as in a multivariate Gaussian situation and is completely described by the correlation matrix of the underlying multivariate Gaussian random variable.

For example, this model was used as a model for the probability of joint defaults—the probability that any two members (say A and B) of a pool of credits will both default within the next year or some other pre-specified period (i.e., the credit taker fails to pay the interest or the credit notional amount back). Denoting by T_A

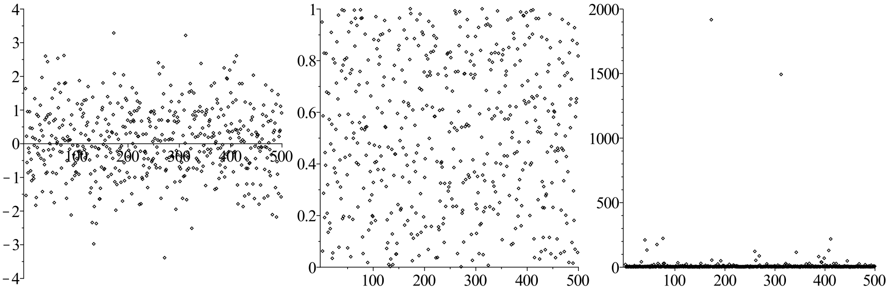


Fig. 7 Simulation of 500 independent and identically distributed standard normally distributed random variables (*left*) and their transformations to standard uniform (*middle*) and standard Fréchet random variables

the time when A defaults and likewise by T_B that B defaults, this model describes the probability that both credits will default as

$$P(T_A < 1, T_B < 1) = \Phi_2\big(\Phi^{-1}\big(F_A(1)\big), \Phi^{-1}\big(F_B(1)\big); \rho\big),$$

where F_A and F_B are the marginal distribution functions of the default times and ρ the correlation of the used normal copula.

This model, suggested in [25], was heavily blamed (and obviously before the subprime crisis heavily used) in a now famous article from 2009 (still to be found on the Internet at http://www.wired.com/techbiz/it/magazine/17-03/wp_quant) entitled "The Formula That Killed Wall Street". The reason is that in a bivariate (and likewise in a higher dimensional) normal model with correlation different from 1 the probability that both variables X and Y are very big at the same time is extremely small: asymptotically for $z \to \infty$ the events that $X > z$ and $Y > z$ become independent. Now it turned out during the subprime crisis that the dependence between different credits is much higher. In the US subprime credit market it became obvious that many more of those involved in the markets than the credit models predicted to be likely could not fulfil their obligations (to pay the interest, repay the principal etc.). The problem was that these credits had been pooled by the issuing banks and—sliced up into packets—sold to investors all over the world; the prices agreed upon in these sales were based usually on the above model (as were the triple-A ratings of some of these products by rating agencies). Additionally, many derivatives based upon them—credit default swaps or credit default options were originally designed as insurance against defaults—were traded and very often they were bought or sold not to insure oneself, but for purely speculative reasons. So when many credits started to default, financial institutions all over the world had to accept that their assets were worth much less than they had thought, which implied tremendous losses in particular for the financial industry. An interesting paper on how to model these risks more realistically is [18].

Consequently, the financial crisis of the last years is a clear warning that one should not use models without basic knowledge of what they can model and what they cannot. Model risk is abundant and needs a critical mind concerning the application of various models and the interpretation of their resulting outcome, when applied with care. In the above normal copula model dependence is modelled by the correlation of the underlying normal distribution. It has long been known that a normal copula is by no means a model that captures dependent risks: in a normal copula model very high risks are always independent (see Example 8.5).

As we have seen in Sect. 5 the elliptical distributions are natural extensions of multivariate normal distributions and are also characterised mainly by their mean and covariance structure, only that additionally a positive generating random variable comes into play. Likewise, we can extend the normal copula to an elliptical copula by using the copula corresponding to a general elliptical distribution.

Definition 7.5 (Elliptical Copula) We define an elliptical copula as the copula of $\mathbf{X} \sim \mathcal{E}_d(\boldsymbol{\mu}, \Sigma, G)$ and write $\mathcal{EC}_d(R, G)$ for short, where R is the correlation matrix of the elliptical distribution and G the generating random variable.

The notation $\mathcal{EC}_d(R, G)$ for an elliptical copula makes sense, since it is characterized by the generating variable G (unique up to a multiplicative constant) and the copula correlation matrix R. This follows as a simple consequence of the definition and the fact that copulae are invariant under strictly increasing transformations.

Example 7.6 (a) Let $\mathbf{Z}$ be a d-dimensional mean $\mathbf{0}$ normal vector with arbitrary covariance matrix Σ, and denote by Φ the one-dimensional standard normal distribution function, then the distribution of $(\Phi(Z_1), \ldots, \Phi(Z_d))$ is a Gaussian copula.

(b) Let $\mathbf{X} \sim \sqrt{\nu}\frac{\mathbf{Z}}{\sqrt{W}}$ with W being a χ^2-distributed random variable with ν degrees of freedom and $\mathbf{Z}$ a d-dimensional mean $\mathbf{0}$ normal vector with arbitrary covariance matrix Σ. So $\mathbf{X}$ follows a d-dimensional t-distribution with ν degrees of freedom and we write $\mathbf{X} \overset{d}{=} \boldsymbol{t}_\nu(\mathbf{0}, \Sigma)$; i.e., $\mathbf{X}$ is distributed as in Example 5.3(b). Denoting by t_ν the one-dimensional t-distribution with ν degrees of freedom, then the distribution of $(t_\nu(X_1), \ldots, t_\nu(X_d))$ is the corresponding copula, which we call a $\boldsymbol{t}_\nu$-copula.

In Fig. 8 we show the differences between the normal distribution, the t_4-distribution, and in Fig. 9 their copulae. Comparing the figures in the left column we see that, for the same normal margins, the dependence structure given by the t_4 copula yields more data in the left lower and right upper corners. The right column shows first that t-margins are heavier tailed than normal margins. Furthermore, for the t_4-distribution we see more data in the left lower and right upper corners than for the normal copula. Moreover, for the t_4-copula the data spread out more in direction of the right lower and left upper corners than for the normal copula.

Illustration 7.7 (Danish Fire Continued) In Fig. 6 the ranks of the losses in building are plotted against the ranks of the losses of content. Up to a normalization this is a plot of the copula (the data transformed to uniform margins as in Fig. 9). As we

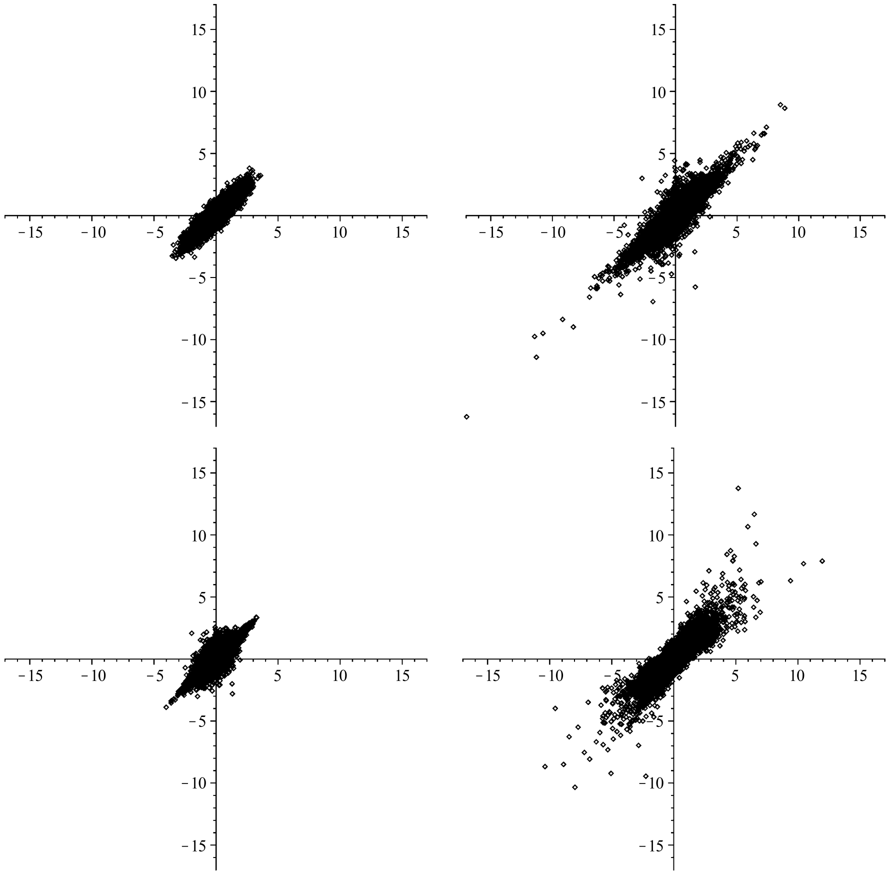


Fig. 8 *Upper row*: simulation of 10,000 bivariate normally distributed random variables (*left*) and bivariate t_4-distributed random variables (*right*). *Lower row*: simulation of 10,000 bivariate random variables with normal marginal distributions and a t_4-copula (*left*), and with t_4 marginal distributions and a normal copula (*right*). In all cases the correlation parameter was $\rho = 0.9$

already said, the fact that there are very few points at the lower right and upper left corner hints again at positive dependence.

Illustration 7.8 (Engineering Risk Analysis) Engineers often deal with complex systems with a large number of components. Suppose such a system consists of d components. As the consequence of a risky event Y (e.g. an accident, an earthquake, a tsunami, a hurricane or a cyber attack) each component can be damaged. Typically the degree of damage will be different for every component.

A realisation y of Y would give the strength of such events above. The damage done to component n is measured by a random variable X_n for $n = 1, \dots, d$ which gives the costs of repairing or when necessary replacing the component. Assume that all damage variables X_n have continuous distribution functions F_n with densities

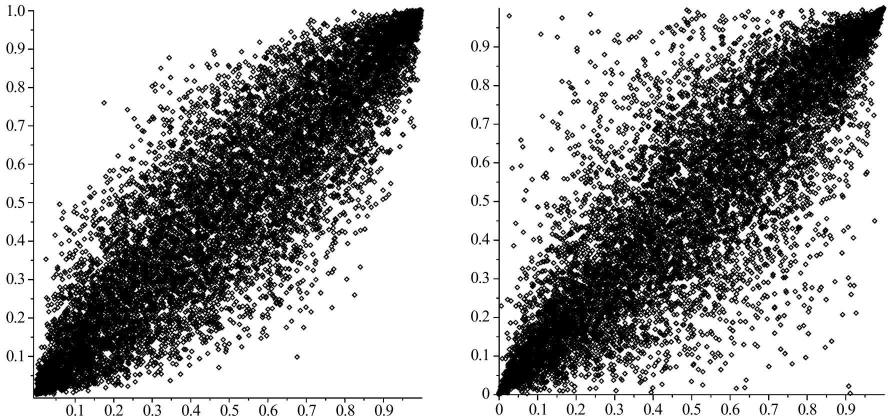


Fig. 9 The copulae corresponding to Fig. 8; i.e., normal copula (*left*) and t_4-copula (*right*)

f_{X_n} for $n = 1, \dots, d$, and that together with Y they have a joint density $f_{X_1,\dots,X_d,Y}$. Depending on the realised damage attributable to the risk event Y, summarized in the vector $(x_1, \dots, x_d)$, the monetary amount $K(x_1, \dots, x_d)$ is needed to repair the system; some components would have to be repaired, some to be replaced. Note that K could simply be the sum of the x_n, but we allow for more general functions, since cost reductions or increases occur when you have to repair/replace several components.

In engineering, risk is often calculated as expected costs due to possible damages. We calculate the expected costs for repairing the system as

$$
\begin{aligned}
E(K) &= \int_0^\infty \cdots \int_0^\infty K(x_1, \dots, x_d) f_{X_1,\dots,X_d}(x_1, \dots, x_d) dx_1 \cdots dx_d \\
&= \int_0^\infty \left(\int_0^\infty \cdots \int_0^\infty K(x_1, \dots, x_d) f_{X_1,\dots,X_d|Y}(x_1, \dots, x_d \mid y) dx_1 \cdots dx_d \right) f_Y(y) dy,
\end{aligned}
$$

where f_Y is the density of the risky event variable Y and $f_{X_1,\dots,X_d|Y}$ the joint density of the damages to the individual components given the risky event Y. From this calculation we see immediately that we need a model for the random vector taking the dependence structure between the damages to the different components $(X_1, \dots, X_d) \mid Y$ into account.

The dependence structure of $(X_1, \dots, X_d) \mid Y$ can be described via a copula. An unrealistic but simple scenario is the independence copula (i.e. we assume that the damages to the individual components are independent given Y). If additionally K is simply the sum of the individual damages, we obtain:

$$
E(K) = \int_0^\infty \left(\sum_{n=1}^{d} \int_0^\infty x_n f_{X_n|Y}(x_n \mid y) dx_n \right) f_Y(y) dy,
$$

where $f_{X_n|Y}$ is the conditional density of the damage in component n given the risky event Y.

Clearly these assumptions will be too simple in most real-life applications, because the damages to the individual components are most likely dependent given Y or the costs of repairing the system are not the sum of the costs of repairing/replacing the individual components.

Remark 7.9 (a) Whereas bivariate copula models are well-known in great detail, higher dimensional models are usually hard to analyse and to fit to real data, not least due to numerical problems when optimizing the likelihood function. Exceptions are the normal and t-copula models. A fairly new approach opens up the way to copulae of arbitrary dimension; cf. [12] and the book [7].

(c) In general the usage of copulae seems rather demanding at first and most statistical software does not include functions to handle copulae in their basic distributions. However, for many statistical programmes there are very well implemented and documented extensions available which make the use of copulae rather easy in applications. For example, for the programme R there are the packages `copula` and `fCopulae` available at http://cran.r-project.org/web/packages/. They include e.g. functions to handle Archimedean, elliptical and extreme value copulae.

(b) For a parsimonious model with respect to parameters, dimension reduction is an important first step. There exist many well-known methods (e.g. principal component analysis) in classical multivariate statistics. Therefore, the use of copulae often needs to be combined with such methods. Dimension-reduction methods based on elliptical copula models have been suggested in Klüppelberg and Kuhn [6].

8 Extremal Dependence Measures

As explained in Chap. 6, [20] extremal risks can be modelled and estimated in a stochastic framework. In contrast to Chap. 6, [20], in the present chapter we are concerned about joint extreme risks, which can be particularly dangerous. Hence it is of the utmost importance to model and assess the joint occurrences of extreme events correctly. In other words it is not important to get the dependence of the "typical" observations right, but one must get the dependence of the extreme events right. One of the first questions for a statistical model is, then, if it is likely to model joint extreme events.

In this section we briefly present models and methods to allow for a realistic assessment of the dependence of extremal events. An interesting collection of theoretical results and case studies for further reading is Reiss and Thomas [28]. Another very accessible book on extreme value statistics is Coles [17]; more advanced is Beirlant et al. [14].

One way to consider the question whether extremal events are dependent or not is by asking, what is the probability that a random variable Y assumes a large value given that we already know that another random variable X takes a large value. Consequently, one natural way to model extremal dependence is to consider the asymptotic behaviour of the probability that $Y > z$ given that $X > z$, as $z \to \infty$. If X and Y

are independent, we have that $P(Y > z \mid X > z) = P(Y > z) \to 0$ as $z \to \infty$. Thus we call any pair X, Y of random variables with $P(Y > z \mid X > z) \to 0$ as $z \to \infty$ *tail independent*. Intuitively this means that extreme events typically occur only in one variable, provided they occur. In contrast to this, we speak of *tail dependence* whenever the limit is non-zero, which implies that with a positive probability extreme events occur in both random variables at the same time. It turns out that this intuitive approach makes sense only when X, Y have the same distribution (or at least distributions with comparable tails). To account for this, one normalises the tails first using the same trick as we know already from the copulae. To be precise one defines tail dependence coefficients (for the upper tail) as follows. Again we invoke the quantile function from (1.1).

Definition 8.1 (Tail Dependence Coefficients) Let X, Y be two random variables with continuous distribution functions F_X and F_Y. The *upper tail dependence coefficient* of (X, Y) is defined by

$$\lambda_U = \lim_{\alpha \uparrow 1} P\big(F_Y(Y) > \alpha \mid F_X(X) > \alpha\big) = \lim_{\alpha \uparrow 1} P\big(Y > F_Y^{-1}(\alpha) \mid X > F_X^{-1}(\alpha)\big),$$

provided the limit exists ($\alpha \uparrow 1$ stands for taking the limit for α going to 1 from below). If $\lambda_U \in (0, 1]$, then X and Y are called *upper tail dependent*. If $\lambda_U = 0$, they are called *upper tail independent*.

Remark 8.2 (i) The assumption of continuous distributions is not really necessary, if one restricts the definition to $\lambda_U := \lim_{\alpha \uparrow 1} P(Y > F_Y^{-1}(\alpha) \mid X > F_X^{-1}(\alpha))$.

(ii) Noting that $P(F_Y(Y) > 1 - t \mid F_X(X) > 1 - t) = \frac{P(F_Y(Y)>1-t, F_X(X)>1-t)}{P(F_X(X)>1-t)}$ and $P(F_X(X) > 1 - t) = t$, we obtain the equivalent definition

$$\lambda_U = \lim_{t \to 0} t^{-1} P\big(F_X(X) > 1 - t, F_Y(Y) > 1 - t\big).$$

(iii) The link to the Value-at-Risk as defined in Definition 1.1(b) is obvious:

$$\lambda_U = \lim_{\alpha \uparrow 1} P\big(Y > \mathrm{VaR}_\alpha(Y) \mid X > \mathrm{VaR}_\alpha(X)\big).$$

One can show that the tail dependence is a copula property; i.e., the marginal distributions have no effect on the value of λ_U.

Theorem 8.3 *If X, Y have copula C, then*

$$\lambda_U = \lim_{\alpha \uparrow 1} \frac{1 - 2\alpha + C(\alpha, \alpha)}{1 - \alpha}. \tag{8.1}$$

Remark 8.4 Theorem 8.3 provides a useful link of λ_U to the Pickands dependence function:

$$\frac{1-2\alpha+C(\alpha,\alpha)}{1-\alpha}=\frac{1-2\alpha+\exp(2\ln(\alpha)A(\frac{1}{2}))}{1-\alpha}$$
$$=2\left(1-A\left(\frac{1}{2}\right)\frac{-\ln\alpha+o(\ln\alpha)}{1-\alpha}\right)$$

by a Taylor expansion of the exponential function around 0. Using l'Hospital's rule we calculate

$$\lim_{\alpha\uparrow 1}\frac{-\ln\alpha+o(\ln\alpha)}{1-\alpha}=\lim_{\alpha\uparrow 1}\frac{\frac{1}{\alpha}(1+o(1))}{\alpha}=1$$

giving

$$\lambda_U=2\left(1-A\left(\frac{1}{2}\right)\right). \tag{8.2}$$

For each copula model we can determine if it allows for tail dependence or not.

Example 8.5 (a) Since by Proposition 5.2(c) the conditional distribution of a bivariate Gaussian random vector is normal ($\mathcal{E}_p$ is in this case, of course, the normal distribution), the Gaussian copula (or Gaussian distribution) with correlation $\rho<1$ has

$$\lambda_U=2\lim_{x\to\infty}\left(1-\Phi\left(\frac{\sqrt{1-\rho}}{\sqrt{1+\rho}}x\right)\right)=0.$$

Hence, when using a Gaussian copula one always has tail independence unless one considers the degenerate situation where $\rho=1$. Therefore, one must never use the Gaussian copula when one wants to model phenomena where extreme events occur jointly in different variables. The financial industry has learnt this the hard way (see Illustration 7.4).

(b) For a Gumbel copula (8.2) gives $\lambda_U=2-2^{1/\theta}$. Hence, whenever $\theta>1$ we have a positive tail dependence and the tail dependence coefficient can assume any value in (0, 1).

(c) For the bivariate t_ν-copula with ν degrees of freedom and correlation $\rho\in[-1,1]$ one calculates using again (8.2)

$$\lambda_U=2\left(1-t_{\nu+1}\left(\frac{\sqrt{\nu+1}\sqrt{1-\rho}}{\sqrt{1+\rho}}\right)\right)$$

with $t_{\nu+1}$ being the distribution function of a t-distributed random variable with $\nu+1$ degrees of freedom. This implies that for every $\rho>-1$ the upper tail dependence coefficient $\lambda_U>0$; i.e., that even for negative correlation it is far more likely than in the Gaussian copula to have both variables large at the same time.

Illustration 8.6 (Danish Fire Continued) Estimation of the tail dependence coefficient is rather tricky, since it is an asymptotic property (an asymptotic conditional probability). Therefore, it may depend rather strongly on the choice of the threshold approximating this asymptotic. We refer to Haug et al. [5] for a detailed analysis of these issues. For the Danish fire data, [5] reports a value of $\widehat{\lambda_U} = 0.416$ for the tail dependence coefficient between the losses in building and content. Therefore, there is a non-negligible tail dependence and thus an insurance company needs to be prepared to meet large losses in its fire insurance for buildings and its insurance for the contents at the same time. Of course, intuitively this is not surprising.

To sum up our simple data example using the fire insurance data, we see that all dependence measures in this context report a positive dependence. But they focus on different aspects and thus the most adequate one should be used in any particular application. In particular, you should be aware that when using correlations, a simple order-preserving transformation such as taking logarithms may have a big impact, whereas it will have no effect, if the dependence measure depends only on the ranks (or the copula).

9 Food for Thought

We list some questions which should be seriously considered for every real risk problem at hand.

- Is my risk problem multivariate? What are the risk factors involved?
- Which techniques do I use to model dependence? Does risk occur from the data around the mean or rather from extreme events? Is it important to get the bulk of the data right or the extremes? Should I use all data or only extreme values for a statistical analysis?
- What model should I use? What does the model I use assume about the dependence structure?
- How will I deal with the model risk?
- How sensitive are the outcomes of my research to assumptions about dependence? Should I apply several models and check robustness of the outcomes by a sensitivity analysis?

Important Final Call: We could give only a brief introduction into dependence modelling and some related problems. Likewise, we could give an overview only over some techniques and a very limited number of examples without going into details. Much more can be found in the literature and in the end every application calls for a tailor-made model. Therefore, it may well be necessary to extend and adapt the existing techniques in line with what is needed for a concrete application.

10 Summary

In this paper we showed that the dependence structure matters critically when facing different risks. The overall risk may change completely when the dependence changes. We discussed various approaches to model the dependence structure. The most popular measure of dependence is correlation which, however, covers only linear effects and has other drawbacks and limitations. As alternative dependence measures we considered rank correlations and copulae. The latter are theoretically able to encode the complete dependence structure, but for a statistical risk analysis one chooses certain parametric families which may introduce severe limitations and also model risk; cf. Chap. 10, [13]. Furthermore, we explained that elliptical distributions are natural generalisations of the multivariate normal distribution where mainly the correlation structure matters. Finally, we introduced tail dependence and explained that it is of utmost importance in connection with risk modelling, because it captures the dependence of the extremes, which is what typically matters in risk assessment, risk evaluation and consequential risk handling, and which may be rather different from the dependence of the bulk of the observations.

Acknowledgements The authors thank Wilson Lauw for preparing large parts of the figures during his research internship at Ulm University. Financial support by the Institute for Advanced Study of the Technische Universität München (TUM-IAS) is gratefully acknowledged.

References

Selected Bibliography

1. P. Embrechts, A.J. McNeil, D. Straumann, Correlation: pitfalls and alternatives. Risk Mag. **May**, 69–71 (1999)
2. P. Embrechts, A.J. McNeil, D. Straumann, Correlation and dependence in risk management: properties and pitfalls, in *Risk Management: Value at Risk and Beyond*, ed. by M.H.A. Dempster (Cambridge Univ. Press, Cambridge, 2002), pp. 176–223
3. P. Embrechts, F. Lindskog, A.J. McNeil, Modelling dependence with copulas and applications to risk management, in *Handbook of Heavy Tailed Distributions in Finance*, ed. by S.T. Rachev (Elsevier, Amsterdam, 2003), pp. 329–384, Chap. 8
4. K.-T. Fang, S. Kotz, K.-W. Ng, *Symmetric Multivariate and Related Distributions* (Chapman & Hall, London, 1990)
5. S. Haug, C. Klüppelberg, L. Peng, Statistical models and methods for dependent insurance data. J. Korean Stat. Soc. **40**, 125–139 (2011)
6. C. Klüppelberg, G. Kuhn, Copula structure analysis. J. R. Stat. Soc., Ser. B, Stat. Methodol. **71**, 737–753 (2009)
7. D. Kurowicka, H. Joe (eds.), *Dependence Modeling: Vine Copula Handbook* (World Scientific, Singapore, 2010)
8. A. McNeil, R. Frey, P. Embrechts, *Quantitative Risk Management*. Princeton Series in Finance (Princeton University Press, Princeton, 2005)
9. R. Nelsen, *An Introduction to Copulas*, 2nd edn. Lecture Notes in Statistics, vol. 139 (Springer, New York, 2006)
10. D. Straub, Engineering risk assessment, in *Risk – A Multidisciplinary Introduction*, ed. by C. Klüppelberg, D. Straub, I. Welpe (2014)
11. Y.L. Tong, *The Multivariate Normal Distribution* (Springer, New York, 1990)

Additional Literature

12. K. Aas, C. Czado, A. Frigessi, H. Bakken, Pair-copula constructions of multiple dependence. Insur. Math. Econ. **44**(2), 182–198 (2009)
13. K.F. Bannör, M. Scherer, Model risk and uncertainty—illustrated with examples from mathematical finance, in *Risk – A Multidisciplinary Introduction*, ed. by C. Klüppelberg, D. Straub, I. Welpe (2014)
14. J. Beirlant, Y. Goegebeur, J. Segers, J. Teugels, *Statistics of Extremes* (Wiley, Chichester, 2004)
15. F. Biagini, T. Meyer-Brandis, G. Svindland, The mathematical concept of measuring risk, in *Risk – A Multidisciplinary Introduction* ed. by C. Klüppelberg, D. Straub, I. Welpe (2014)
16. K. Böcker, C. Klüppelberg, Modelling and measuring multivariate operational risk with Lévy copulas. J. Oper. Risk **3**(2), 3–27 (2008)
17. S. Coles, *An Introduction to Statistical Modeling of Extreme Values* (Springer, London, 2001)
18. R. Cont, Y.H. Kan, Statistical modeling of Credit Default Swaps Portfolios. Working paper (2011). Available at SSRN: http://ssrn.com/abstract=1771862
19. S. Demarta, A.J. McNeil, The t copula and related copulas. Int. Stat. Rev. **73**(1), 111–129 (2005)
20. V. Fasen, C. Klüppelberg, A. Menzel, Quantifying extreme risks, in *Risk – A Multidisciplinary Introduction*, ed. by C. Klüppelberg, D. Straub, I. Welpe (2014)
21. G.A. Fredricks, R.B. Nelsen, On the relationship between Spearman's rho and Kendall's tau for pairs of continuous random variables. J. Stat. Plan. Inference **137**(7), 2143–2150 (2007)
22. M. Hohenbichler, R. Rackwitz, Non-normal dependent vectors in structural safety. J. Eng. Mech. Div., ASCE **107**(6), 1227–1249 (1981)
23. H. Joe, Families of min-stable multivariate exponential and multivariate extreme value distributions. Stat. Probab. Lett. **9**(1), 75–81 (1990)
24. D. Kelker, Distribution theory of spherical distributions and a location-scale parameter generalization. Sankhyā, Indian J. Stat., Ser. A **32**, 419–438 (1970)
25. D.X. Li, On default correlation: a copula function approach. J. Fixed Income **9**, 43–54 (2000)
26. P.-L. Liu, A. Der Kiureghian, Multivariate distribution models with prescribed marginals and covariances. Probab. Eng. Mech. **1**(2), 105–112 (1986)
27. R.B. Nelsen, On measures of association as measures of positive dependence. Stat. Probab. Lett. **14**(4), 269–274 (1992)
28. R.-D. Reiss, M. Thomas, *Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields*, 2nd edn. (Birkhäuser, Basel, 2001)
29. S.I. Resnick, *Extreme Values, Regular Variation, and Point Processes.*, 2nd edn. (Springer, New York, 1987)

Chapter 10
Model Risk and Uncertainty—Illustrated with Examples from Mathematical Finance

Karl F. Bannör and Matthias Scherer

Stochastic modeling techniques have become increasingly popular during the last decades, particularly in mathematical finance since the groundbreaking work of Bachelier (Théorie de la spéculation, Gauthier-Villars, Paris, 1900), Samuelson (Ind. Manag. Rev. 6(2):13–39, 1965), and Black and Scholes (J. Polit. Econ. 81(3):637–654, 1973). Essentially, all models are wrong in the sense that they simplify reality. However, there are numerous models available to model particular phenomena of financial markets and calculated option prices, hedging strategies, portfolio allocations, etc. depend on the chosen model. This gives rise to the question which model to choose from the rich pool of available models and, second, how to determine the correct parameters after having selected some specific model class. Thus, one is exposed to both model and parameter risk (or uncertainty). In this survey, we first provide an inside view into the principles of stochastic modeling, illustrated with examples from mathematical finance. Afterwards, we define model risk and uncertainty according to Knight (Risk, uncertainty, and profit, Hart, Schaffner & Marx, Chicago, 1921) and present some methods how to deal with model risk and uncertainty.

Keywords Financial market models · Parameter risk · Model risk · Risk-capturing functionals · Convex risk measure

Mathematics Subject Classification (2010) 62C12 · 91G20

The Facts

- In natural sciences as physics, chemistry, and biology, laws of nature often support model building. In social sciences like economics, there may be no natural laws offering models.

K.F. Bannör (✉)
Deloitte & Touche GmbH, Rosenheimer Platz 4, 81669 Munich, Germany
e-mail: kbannoer@deloitte.de

M. Scherer
Mathematical Finance, Center for Mathematical Sciences, Technische Universität München,
Boltzmannstrasse 3, 85748 Garching bei München, Germany

C. Klüppelberg et al. (eds.), *Risk – A Multidisciplinary Introduction*,
DOI 10.1007/978-3-319-04486-6_10,

- Stochastic modeling tries to capture the stylized facts of the distribution of outcomes in concern.
- Often, there is considerable ambiguity which model (or, equivalently, which probability distribution) to choose.
- One distinguishes between model risk and model uncertainty, following the terminology of Knight [9].
- Model risk is a situation where one can quantify the likelihood of the validity of the different models to choose from, i.e. a probability distribution on the set of models is known.
- Model uncertainty is a situation where one does not have any additional information about the different models, i.e. a probability distribution on the set of models is unknown.

1 Stochastic Modeling of Real-World Phenomena

> *Die Theorie liefert viel, aber dem Geheimnis des Alten bringt sie uns doch nicht näher. Jedenfalls bin ich überzeugt davon, dass der nicht würfelt.*[1]*—Albert Einstein, Nobel Laureate in Physics*

Models from classical mechanics, as illustrated in Chap. 4 of Mainzer [38], often describe effects that have fully been studied. Hence, a deterministic functional relationship can be taken as a mathematical model for description.[2] In contrast, there exist many real-world phenomena that exhibit deterministic behavior, but the description of the deterministic behavior is much too complex, or the behavior is difficult to observe. In such cases, it has turned out to be a tractable way to move from deterministic modeling to stochastic modeling, enriching a deterministic functional relationship by accounting for different random states which may occur. These different random states are gathered in a stochastic basis, which is mathematically described by a probability space $(\Omega, \mathcal{F}, P)$.

Simplification Due to Stochasticity Stochasticity is often used to model deterministic phenomena in a tractable way such that the model still describes the *outcomes* of real-world phenomena (that might actually be deterministic in nature). Instead of modeling the deterministic and possibly complicated procedure which leads to the outcome, one focuses only on data concerning the outcome, analyzes the "distribution" of the outcomes, and finally one sets up a stochastic model which captures the distribution of the outcomes as realistic as possible.

A very easy but vivid example of a situation where specifying the deterministic behavior may be awkward is modeling the result of throwing a (fair) dice: obviously, throwing a dice is an action which can be described completely by classical

[1] Translation: the theory yields a lot, but it does not bring us closer to the secret of the old one [god]. Anyway, I am convinced that he [god] does not throw the dice.

[2] One prominent exception is the statistical approach to quantum physics.

mechanics. Shaking the dice in a dice cup is a mechanic procedure, where the dice turns when touching the walls of the dice cup, falls and rolls on the table, and eventually displays some number. But the whole procedure of shaking the dice and rolling is extremely complicated to model in the world of classical mechanics, since many different influences have to be taken into account (like, e.g., the shape and size of the dice cup and the dice, the different directions and magnitudes of the shaking, etc.). Such a deterministic model would be hard to determine, to set up, and even more difficult to evaluate.

If one, however, is only interested in the result, i.e. the thrown number, one might imagine a model which is much more simple and circumvents the difficulties of modeling such a situation with classical mechanics. The mixing procedure cannot be reproduced easily and as a result, every side of the dice occurs similarly often. Mathematically spoken, the relative share $r(j)$ of obtaining a fixed number $j \in \{1, 2, 3, 4, 5, 6\}$ is independent of the number j and since the relative shares have to add up to one, it follows that $r(j) \approx 1/6$ for all $j \in \{1, 2, 3, 4, 5, 6\}$. Hence, a probabilistic model describing the result of throwing a dice, which both models reality feasibly and yields a tractable situation, is to provide a stochastic basis in the following way: let $\Omega := \{1, 2, 3, 4, 5, 6\}$ be the state space of possible dice throw outcomes, $\mathcal{F} := \mathfrak{P}(\Omega)$ all possible combinations of outcomes, and $P : \mathcal{F} \to [0, 1]$ a probability measure defined via $P(\{j\}) = 1/6$ for all $j \in \{1, 2, 3, 4, 5, 6\}$. Then the probability space $(\Omega, \mathcal{F}, P)$ sufficiently describes the possible outcomes of a dice throw in an abstract, easy, and tractable manner.

Contrary to modeling the dice throw by classical mechanics, the stochastic model has simplified and abstracted tremendously from the original situation. The whole procedure of throwing the dice *physically* is completely disregarded. Instead, the stochastic model only focuses on the *result* of the dice throw and models it directly, which turns out to be much more tractable and also feasible from an empirical point of view.

A Detailed Excursion: Stochastic Modeling in Finance In physics and engineering, mathematical modeling of real-world phenomena goes back to Isaac Newton, Gottfried Wilhelm Leibniz, and even to the ancient Greeks. In contrast, in finance, mathematical and particularly stochastic modeling is a rather recent trend, starting with the seminal dissertation of Bachelier [16].

When regarding the financial world instead of modeling phenomena from classical mechanics, one immediately recognizes that the whole system is much more complex in the sense that many different forces drive the market, and their influence is of non-negligible order. When describing the fall of a stone to the ground in a laboratory, there are undoubtly also many different forces apart from earth gravitation that actually have some influence (e.g. the aerodynamic resistance, the gravitation of different objects in the laboratory). But their magnitude is so small compared to the magnitude of earth gravitation that not considering them eventually does not matter for a realistic model.

In contrast, when modeling financial markets (e.g. stock markets for the purpose of, e.g., option pricing), there are many different market participants that influence asset prices by their trade decisions. Hence, a model trying to capture the whole market microstructure with all interactions of market participants would be a monstruous, extremely complicated attempt with myriads of parameters. Thus, such an approach is only tractable under severe simplifications (similar to the dice example). But, additionally, there are several other reasons not to model the microstructure of financial markets.

- First, different to the dice example, financial markets cannot be put under laboratory conditions and therefore models cannot be tested reliably, i.e. experiments cannot be repeated.
- Second, due to the complexity of the operations, it is impossible to observe all market participant's behavior and interaction simultaneously.
- Third, many market participants exhibit irrational and erratic behavior which may be difficult to model even when modeling only a single market participant. There have been approaches as the celebrated "Prospect Theory" of Kahneman and Tversky [34][3] trying to provide a scope for such a kind of behavior, which still is ongoing research.
- Finally, and maybe most crucial, the whole system is dynamic, with new market participants entering and leaving the system. Even if one could observe the market participants' behavior and collect huge amounts of data, in every second, new market participants enter the financial markets and behave differently, such that predictions relying on historical data might not explain future market situations successfully.[4]

Hence, the typical approach to model stock markets is to disregard the market microstructure (which is, e.g., forgetting about the market participants action and interaction,[5] analog to forgetting about the mechanics when rolling the dice) and to model asset prices statistically.

To set up a sensible stochastic model for the price of, e.g., a stock or an index, one typically scrutinizes stylized facts of time series of the price process and tries to mimic these properties with stochastic models fulfilling as many of these stylized facts as possible. Compared to an ansatz focusing more on data (an extreme ansatz may be a non-parametric one only exploiting data), such a modeling paradigm allows to capture general movements. Furthermore, a stochastic model for a stock price should be tractable enough in the sense that it costs moderate effort to simulate the stock price and prices of related financial instruments (e.g. futures and options, see Hull [8] for an introduction into financial instruments) may be calculated in a (semi-)analytic way. With these requirements for a model, one starts to collect some stylized facts of time series of stock prices and obtains as first observations:

[3] Daniel Kahneman was awarded the Nobel Memorial Prize in Economic Sciences 2002 for his work on irrational behavior in economics.

[4] In financial markets, one can even argue that relying too much on collected data may result in overconfidence, since the data may not be representative any more to model future events.

[5] One should note that there are some approaches trying to capture the microstructure.

- The stock price process, abbreviated by $S = (S_t)_{t \geq 0}$, is always positive.
- Returns (yields) of stock prices are symmetrically scattered around 0 (or around somewhere close to 0) and behave roughly similar and uncorrelated of each other.

Taking the second stylized fact as a starting point, a possible tool for modeling stock returns seems to be the normal distribution, which is widely understood, mathematically tractable, and plays a prominent role in asymptotic statistics (cf. the central limit theorem). Furthermore, for small periods Δt, the discrete return

$$\frac{S_{t+\Delta t} - S_t}{S_t}$$

may comfortably be approximated by the difference of the logarithm $\log S_{t+\Delta t} - \log S_t$. Hence, a first idea might be to model logarithmic differences by i.i.d. normally distributed random variables. With this motivation and the notion of Brownian motion (we omit the formal definition due to technicalities, see Øksendal [10] for details), one arrives at modeling stock prices with a geometric Brownian motion (which goes back to Samuelson [44]), also often called the Black–Scholes model.[6]

Example 1.1 (Black–Scholes Model) A stock price $(S_t)_{t \geq 0}$ is modeled by a Black–Scholes model if it follows a geometric Brownian motion, i.e. its dynamics follow the stochastic differential equation[7]

$$\mathrm{d}S_t = \mu S_t \,\mathrm{d}t + \sigma S_t \,\mathrm{d}W_t, \qquad S_0 > 0,$$

where $(W_t)_{t \geq 0}$ is a standard Brownian motion. The parameter $\mu \in \mathbb{R}$ is called the *drift* of the stock price and the parameter $\sigma > 0$ is called the *volatility* of the stock price.

The Black–Scholes model allows for an easy and comprehensive interpretation: the whole model is parameterized by the drift and the volatility of the process. Since the model implies normally distributed stock returns, everyone who is familiar with the normal distribution can apply and handle the model. The drift parameter μ controls the average stock return, which grows linearly in μ. In terms of stock prices, μ is the (exponential) growth rate of the stock price. The higher the drift μ, the faster the stock price grows *on average*. On the other hand, the volatility parameter σ describes how the returns scatter around the average returns. When regarding the stock price instead of the returns, the volatility controls how much the stock price

[6] Actually, the model was not developed by Black and Scholes, but by Samuelson and was inspired by the seminal PhD thesis Bachelier [16]. Fischer Black and Myron Scholes derived tractable formulae for European options in this model and introduced the idea of replication in their seminal paper Black and Scholes [3]. This work, together with the inspired work of Robert Merton, resulted in awarding the Nobel Memorial Prize in Economic Sciences to Robert Merton and Myron Scholes in 1997. Fischer Black died already in 1995, thus he did not receive the prize.

[7] Stochastic processes are often described via stochastic differential equations (SDEs). For readers that are unfamiliar with SDEs, we recommend the introductory book of Öksendal [10].

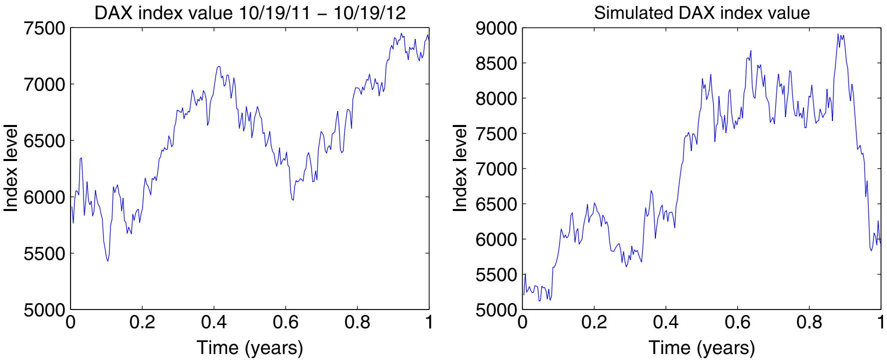


Fig. 1 Comparison: a time series of the DAX index level compared with a simulated path of the DAX in the Black–Scholes model

moves non-directionally. The higher the volatility σ, the more fluctuations the stock price exhibits.

For the pricing of options on the stock, one applies the *risk-neutral version* of the Black–Scholes model, where the drift equals the interest rate of a risk-free investment.

Obviously, the dynamics imposed by the Black–Scholes model are rough simplifications of real stock price dynamics. While the only source of randomness in the Black–Scholes model is the Brownian motion and all other ingrediences (i.e. drift and volatility) are deterministic, real stock prices are driven by an extremely complex market microstructure. Instead of modeling the whole market microstructure with the dynamics of action and interaction, one simply assumes that it suffices to reduce the complexity to the determination of two parameters—the drift and the volatility. In case of risk-neutral dynamics (which is the standard assumption when pricing options), the complexity is further reduced to the determination of one single parameter—the Black–Scholes volatility. On the other hand, trajectories which are simulated in the Black–Scholes model look somewhat like plots of time series of real stock prices (cf. Fig. 1). Furthermore, the simple structure ensures the tractability of the model, in particular, there exist closed-form pricing formulas for various kind of options, like the classical Black–Scholes formula for European calls and puts.

Taking a closer look on stock price time series as well as on stock price related data (e.g. option prices), one clearly sees that the Black–Scholes model is oversimplifying reality and some stylized facts may not be explained by the Black–Scholes model like the following (which are not exhaustive):

- Extremely high and low returns are more likely to occur in reality than the normal distribution implies (“heavy tails of returns”).
- Volatility is not constant, different market periods (high and low volatility) can be observed (“volatility clustering”).

- Downward price movements are typically accompanied by large undirectional movements ("leverage effect").
- Option prices do not follow the Black–Scholes model, implied volatilities[8] are non-constant ("smile effect").

Hence, different alternatives to (and extensions of) the Black–Scholes model have been developed to tackle the shortcomings of using simple geometric Brownian motion, introducing models based on different processes with heavier tails or stochastic volatility and/or jumps. One model that has become popular in practice is the Heston model, see Heston [7], it uses a Cox–Ingersoll–Ross square-root process[9] as stochastic volatility. We briefly sketch the ingredients of the Heston model.

Example 1.2 (Heston Model) A stock price $(S_t)_{t\geq 0}$ is modeled by a Heston model if its dynamics follow the coupled stochastic differential equations

$$\begin{aligned} \mathrm{d}S_t &= \mu S_t \,\mathrm{d}t + \sigma_t S_t \,\mathrm{d}W_t^{(1)}, \quad S_0 > 0, \\ \mathrm{d}\sigma_t^2 &= \kappa\left(\sigma_{\text{long}}^2 - \sigma_t^2\right)\mathrm{d}t + \xi\sigma_t \,\mathrm{d}W_t^{(2)}, \quad \sigma_0^2 > 0, \\ \mathrm{d}W_t^{(1)}\mathrm{d}W_t^{(2)} &= \rho \,\mathrm{d}t, \end{aligned}$$

where $(W_t^{(j)})_{t\geq 0}$, $j = 1, 2$ are correlated Brownian motions with correlation $\rho \in [-1, 1]$.

For further explanation, one can see that the general stock price dynamics resemble closely the dynamics of the Black–Scholes model, except for one fact: the volatility σ is not assumed to be constant any more, but is now a stochastic process itself (due to technical reasons, one models the "variance process" $(\sigma_t^2)_{t\geq 0}$ instead of the volatility process $(\sigma_t)_{t\geq 0}$). In particular, the noise in the stock price process is now time-dependent and has its own dynamics.

Assuming the dynamics of a Cox–Ingersoll–Ross square-root process for the variance process, one may see the following behavior of the variance:

- The variance process $(\sigma_t^2)_{t\geq 0}$ exhibits non-constant noise, which is governed by the parameter $\xi > 0$. This parameter is usually called the *vol-of-vol*.
- In the long run, the variance fluctuates around a fixed number, the *long-term variance*, which is controlled by the parameter $\sigma_{\text{long}}^2 > 0$.
- The variance process is mean-reverting to the long-term variance, i.e. if the variance is dragged away from its long-term level, it drifts back to the long-term variance. The *speed of mean reversion* is controlled by the parameter $\kappa > 0$.

[8] In the Black–Scholes model, for a given European option, there is a one-to-one relationship between volatilities and option prices. Hence, for options with known market prices, one can recalculate the *implied volatility* from the market prices. Usually, one can exhibit that for different options, the recalculated implied volatilities differ, which is a hint that the Black–Scholes model cannot explain the observed option prices.

[9] The name stems from the Cox–Ingersoll–Ross interest rate model.

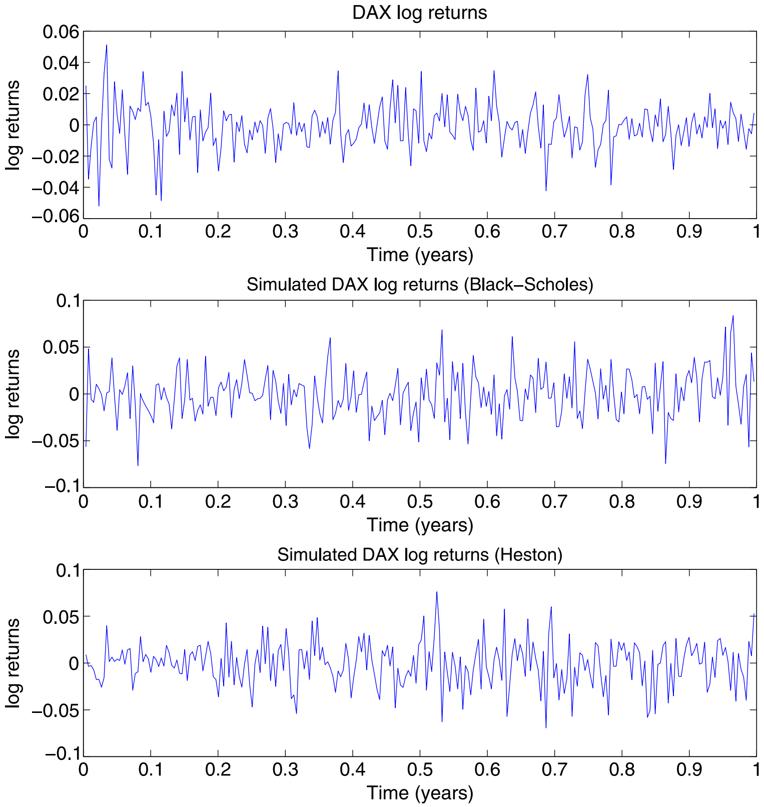


Fig. 2 Comparison: logarithmic returns of the DAX compared with simulated logarithmic returns from the Black–Scholes and the Heston model. One can see that the Black–Scholes model produces returns with regular noise, while the Heston model incorporates volatility clustering, i.e. there exist time periods of high and low fluctuations in the returns

- The correlation $\rho \in [-1, 1]$ describes the co-movement of the stock price and its variance. As described above, this can be used to account for the so-called "leverage effect", establishing that volatility movements and stock price movements have negative correlation.

The Heston model is a relatively simple extension of the Black–Scholes framework (replacing constant volatility by a variance process following a Cox–Ingersoll–Ross model) to model stock prices. But, unarguably, the Heston model overcomes some of the shortcomings of the Black–Scholes model that have been described above (cf. Fig. 2). By making volatility stochastic and time-dependent, it captures the non-constant behavior of volatility. Furthermore, incorporating correlation between the drivers of the stock price and variance processes allows to account for the leverage effect, i.e. for negative correlations ρ. One has to remark that these additional stylized facts come at the price of losing mathematical tractability: prices for some important options (e.g. European put and call options) cannot be calculated with simple formulae any more as in the Black–Scholes model, instead one has to rely on numerical algorithms as, e.g., techniques from Fourier analysis to obtain semi-analytic formulae as described in Carr and Madan [23].

2 Model Risk and Uncertainty

> *[T]here are known knowns; there are things we know that we know. There are known unknowns; that is to say there are things that, we now know we don't know. But there are also unknown unknowns—there are things we do not know, we don't know.—Donald Rumsfeld, United States Secretary of Defence 1975–1977, 2001–2006*

In the previous section, we have roughly outlined the main principles of mathematical modeling, in particular stochastic modeling where we will focus on below. Hence, if we refer to *modeling* in the remaining part of this survey, we always mean *stochastic modeling*.

When setting up a stochastic model, one often observes a complicated situation where the outcome in concern behaves in a more or less erratic manner. In some cases (like the dice example), a simple and accurate description may be provided easily. But, typically, the object to model is much more complicated (like the price process of a stock). Hence, it is not clear from the beginning that the choice of one stochastic model P is a good choice or a different model $\tilde{P}$ might be more suitable, like choosing either a Black–Scholes or a Heston model for stock prices. Typically, the quantity of interest is modeled by a random variable X or some stochastic process $(S_t)_{t\geq 0}$. Hence, a situation where modeling may be complex can be mathematically described as a situation where a whole set of probability measures $\mathcal{P}$ (which may typically be infinite) is available for modeling. Sometimes, the set of possible probability measures (i.e. different stochastic models) $\mathcal{P}$ may be parameterized in a canonical way by a parameter space Θ, i.e. $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$.

To provide a concise wording to different situations that may occur if different models $\mathcal{P}$ are available, we first make a short excursion into the literature. The seminal dissertation of Knight [9] analyzes the situation where different states $x_1, \dots, x_N$ are possible outcomes for X. Knight [9] distinguishes between two possible situations that may occur:

1. One knows the probability of each possible outcome $x_1, \dots, x_N$.
2. One does not know the probability of each possible outcome $x_1, \dots, x_N$.

The ladder situation, where hardly any information is available, is called *uncertainty* by Knight [9]. The former one, which at least allows for a probabilistic description, is called *risk*. Obviously, facing risk is a special case of uncertainty (since one could always forget about the probabilities) and a more comfortable situation compared to facing real uncertainty. One can try to deal with a risky situation by *risk management*, i.e. exploiting the information about the probabilities of the different outcomes $x_1, \dots, x_N$ and acting such that a certain risk functional is minimized.

Research from economics, but also from behavioral sciences like psychology and cognitive science, has shown that most people exhibit aversion towards both risk and uncertainty (often subsumed under the term *risk aversion*). A mathematical concept covering risk aversion (prefering situations of certainty over situations of risk) is described by the foundations of utility theory by von Neumann and Morgenstern [48] and furthermore by the introduction of the axioms of subjective expected utility by Savage [45]. Arrow [14] and Pratt [41] analyze risk aversion from an economic perspective. Concerning uncertainty, it has been shown that the concept of uncertainty aversion is available, describing that a situation of risk is generally prefered to a situation where true uncertainty is exhibited. This idea was promoted by Ellsberg [29], challenging the axioms of Savage, which was later reconciled in the works of Gilboa and Schmeidler [31].

Transfering the concepts of risk and uncertainty to stochastic modeling, the situation of having a whole set of models $\mathcal{P}$ to choose from for modeling is generally referred to as *model uncertainty*. If each model $P \in \mathcal{P}$ can be identified by a parameter θ from some parameter space Θ, one speaks about *parameter uncertainty*.[10] If we additionally have given a probability measure R on the set of possible models $\mathcal{P}$ (resp. on the parameter space Θ) which quantifies the probability of each model (resp. parameter) to be the right choice, then we are in a setting of *model risk* (resp. *parameter risk*), which can be considered as a special case of model (resp. parameter) uncertainty.

This is illustrated in Fig. 3.

Examples Model and parameter uncertainty arise in numerous situations. If one faces a complex situation where a stochastic model is applied, one is often ambiguous between different models to choose from. Even after having decided for a specific parametric model, the correct determination of the model's parameters is not straightforward and may result in different obstacles.

When stochastically modeling financial objects, there are myriads of possibilities to simplify, thus many different models are competing with each other. In option pricing, model risk (resp. uncertainty) should not be underestimated, as pointed out by Figlewski [5]. During the financial crisis of 2008, where massive misvaluation of portfolio credit instruments played an important role, this has been discussed in quite some detail among experts, but also in popular media as, e.g., Salmon [11].

Example 2.1 (Parameter Uncertainty in Financial Market Models) All models treated in Sect. 1 are exposed to parameter uncertainty. We will discuss later whether we experience *true parameter uncertainty* in the sense that no information about the parameters is known or we have *parameter risk*, i.e. we are able to quantify whether certain parameters are more likely than others.[11]

[10] From a purely mathematical point of view, distinguishing between model and parameter uncertainty is just up to a mapping $\Theta \to \mathcal{P}$ which may always be obtained for some set Θ. Often, the set Θ can be chosen such that treating different parameters $\theta \in \Theta$ allows for more convenient interpretation in the real world than treating the corresponding model P_θ.

[11] Besides parameter uncertainty, the chosen parametric models can also be incorrect, i.e. model uncertainty can occur.

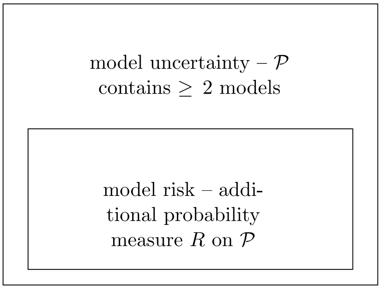

model uncertainty – $\mathcal{P}$ contains $\geq$ 2 models
model risk – additional probability measure R on $\mathcal{P}$

Fig. 3 Relationship between model uncertainty and risk. One can regard model risk as a "special case" of model uncertainty, since one can always ignore the probability measure R quantifying the likelihood of the different models

1. Examining the risk-neutral version of the Black–Scholes model, the dynamics of a stock price follow the stochastic differential equation

   $$\mathrm{d}S_t = r\,S_t\,\mathrm{d}t + \sigma\,S_t\,\mathrm{d}W_t, \qquad S_0 > 0,$$

   with $(W_t)_{t\geq 0}$ being Brownian motion, r the risk-free interest rate, and σ the stock's volatility. While the initial stock price S_0 and the risk-free rate r are usually available from market information, one does not have direct information about the volatility σ. Hence, a priori every positive number $\sigma > 0$ can be taken. Usually, one uses market data (e.g. estimation based on time series of stock prices, or fits the model to the prices of traded instruments) to specify the volatility σ.
2. In the (risk-neutral) Heston model, the stock price dynamics follow the coupled stochastic differential equations

   $$\begin{aligned} \mathrm{d}S_t &= r\,S_t\,\mathrm{d}t + \sigma_t\,S_t\,\mathrm{d}W_t^{(1)}, \qquad S_0 > 0, \\ \mathrm{d}\sigma_t^2 &= \kappa\left(\sigma_t^2 - \sigma_{\mathrm{long}}^2\right)\mathrm{d}t + \xi\sigma_t\,\mathrm{d}W_t^{(2)}, \qquad \sigma_0^2 > 0, \end{aligned}$$

   with $(W_t^{(j)})_{t\geq 0}$, $j = 1, 2$, being Brownian motions with correlation $\rho \in [-1, 1]$. Contrary to the Black–Scholes model, the number of unknown parameters is higher. Again, the initial stock price S_0 and the risk-free rate r are known by market quotation. On the other hand, the initial volatility σ_0, the mean reversion speed $\kappa > 0$, the long-term volatility $\sigma_{\mathrm{long}}^2 > 0$, the vol-of-vol $\xi > 0$, and the correlation $\rho \in [-1, 1]$ are typically not given and—different from the Black–Scholes case—their interpretation is more complicated. Hence, we face parameter risk concerning the parameters $\sigma_0, \kappa, \sigma_{\mathrm{long}}^2, \xi, \rho$.

Even across different models and when establishing perfect fits to market prices[12] of standard instruments (e.g. European call options), one obtains that there is still ambiguity and different models may cause different prices for non-standard options (as pointed out in Schoutens, Simons, and Tistaert [12]).

[12] One possibility to estimate the model parameters is to fit the parameters to known market prices of options.

3 Dealing with Model Risk

If history repeats itself, and the unexpected always happens, how incapable must Man be of learning from experience?—George Bernard Shaw, dramatist

Scrutinizing the available mathematical objects in presence of model (resp. parameter) risk, there exists more than only the set of different possible models $\mathcal{P}$. Additionally, one assumes that $\mathcal{P}$ is the state space of a probability space $(\mathcal{P}, \mathcal{F}^{\mathcal{P}}, R)$ where the probability measure R quantifies the probabilities that the different models $P \in \mathcal{P}$ are the correct models to choose. This delivers a lot of information which has to be analyzed carefully: first, for each stochastic model $P \in \mathcal{P}$, there are given probabilities for the different outcomes one has to deal with. Second, among all these models there is a second probability measure R assigning "weights" to the different models collected in the set $\mathcal{P}$.[13] In this case, one has numerous mathematical obstacles to tackle and to find the right way to incorporate model risk into quantities which may be of interest to be calculated like, e.g., prices of options.

From a statistical perspective, model risk can be regarded as an ansatz in the tradition of *Bayesian statistics*, where one main assumption is that the chosen model (or parameter) itself is random and the probability distribution on the possible models reflects subjective beliefs about the likelihood of the model. Opposed to this view, so-called *frequentist statistics* (going back to the seminal work of Fisher [30]) assumes that a true, but unknown, model (resp. parameter) exists and one cannot assign probabilities to different "candidate models". In history, there has been major dissent between these two philosophical approaches to statistics. A detailed critique and discussion of Bayesian and frequentist methods in statistics is beyond the scope of this article and we refer the interested reader to the books of Samaniego [43] and Bertsch McGrayne [18], but we give a short insight into the foundations of Bayesian statistics later in Sect. 3.2.

One situation where parameter risk traditionally occurs is parameter estimation from given data (e.g. time series of stock prices). In a standard procedure, disregarding parameter risk, one computes the derived estimators from the given data, i.e. calculates point estimates for the parameters. But from estimation theory, one knows that an estimator is a random object itself. Furthermore, an estimator may be biased. Hence, procedures that solely rely on using the point estimate disregard the parameter risk which arises through the estimator's distribution, e.g. its bias and variance.

Parameter estimation is a key step in every application where real data is analyzed. Hence, we present an example employing the Black–Scholes model where the estimator's distribution quantifies the parameter risk.

Example 3.1 (Parameter Risk from Estimation of the Black–Scholes Volatility) We consider a Black–Scholes setting as given in Example 1.1, where the volatility σ is

[13] Due to technical reasons, it may occur that the probability for all single models is zero, i.e. $R(\{P\}) = 0$ for all $P \in \mathcal{P}$.

the key parameter for option pricing. This parameter is not directly given by the market (different from the current stock price S_0 and the risk-free rate r). Hence, the determination of the volatility is a situation where one is exposed to parameter uncertainty. If the stock price actually follows a Black–Scholes model, it may be a sensible idea to estimate the volatility from time series data. Taking the logarithmic returns $x_1, \ldots, x_N$, $x_j = \log S_{t_j+\Delta t} - \log S_{t_j}$, $j = 1, \ldots, N$, one may choose the classical estimator for the variance (it may be more convenient to estimate the returns' variance), corrected for the frequency of the data Δt, which results in the estimator

$$\hat{\sigma}_N^2 = \frac{1}{\Delta t(N-1)} \sum_{j=1}^{N} (x_j - \bar{x})^2, \qquad \bar{x} = \frac{1}{N} \sum_{j=1}^{N} x_j$$

for the variance corresponding to the Black–Scholes volatility, which is consistent and asymptotically normal under very weak assumptions. Applying general theory from statistics, one obtains that, under the assumption of independent normally distributed returns and a true variance $\sigma_0^2 > 0$ (as the Black–Scholes model does), the distribution of the estimator is a χ^2-distribution up to some scaling. Hence, the distribution determining the parameter risk arising from the estimation risk of volatility (resp. variance) is essentially determined by the χ^2-distribution, provided that the true model is a Black–Scholes model with variance σ_0^2. The parameter space is given by $\Theta = \mathbb{R}_{>0}$ and the estimator's distribution R has density r given by

$$r(x) = \frac{(\Delta t(N-1))^{\frac{N-1}{2}}}{\Gamma(\frac{N-1}{2})(2\sigma_0^2)^{\frac{N-1}{2}}} x^{\frac{N-3}{2}} \exp\left(-\frac{x \Delta t(N-1)}{2\sigma_0^2} \right) \mathbf{1}_{\{x>0\}}.$$

3.1 Measuring and Quantifying Model Risk

As defined by Knight [9], the exposure to model risk is a situation where probabilities of different possible models are available. Hence, one should have mathematical instruments at hand to measure and/or to quantify model risk. Fortunately, for the general situation of the measurement and quantification of risk, a rich and mathematically rigorous theory of *risk measures*[14] has been developed, yielding numerous interesting results. For the specific purpose of treating model risk, the theory of risk measures can be transferred, specifically tailored, and applied to the model risk setting under concern. The theory of (convex) risk measures was originally designed for treating financial and actuarial risk, headed by the seminal paper Artzner, Delbaen, Eber, and Heath [1], we follow the red line of this survey and the model risk framework in a financial context.

[14] The terminology "risk measure" may be misleading from a mathematical point of view, since the functions that are proposed to be risk measures are not measures from a measure-theoretical point of view, but functionals.

To ensure a concise understanding, we recapitulate the proper definition of risk measures in a slightly more general setup. A special case of the definition can be found in the textbook Föllmer and Schied [6].

Definition 3.2 (Risk Measure, cf. Biagini, Meyer-Brandis, and Svindland [19], Chap. 5) Let $\mathcal{X}$ be a collection of random variables on a probability space $(\Omega, \mathcal{F}, P)$, i.e. risk-exposed quantities, let $\pi : \mathcal{H} \to \mathbb{R}$ be a linear mapping on a subcollection of random variables $\mathcal{H} \subset \mathcal{X}$ and let $\rho : \mathcal{X} \to \mathbb{R}$ be a function.

ρ is called a *risk measure* w.r.t. π, if ρ fulfills the following axioms:

- ρ is monotone, i.e. for $X, Y \in \mathcal{X}$ and $X \geq Y$, $\rho(X) \geq \rho(Y)$ holds;
- ρ is π-translation invariant, i.e. for $X \in \mathcal{X}$ and $Y \in \mathcal{H}$ the equality $\rho(X + Y) = \rho(X) + \pi(Y)$ holds.[15]

Furthermore, ρ may have additional properties which are often postulated:

- ρ is called *convex*, if for $X, Y \in \mathcal{X}$ and $\lambda \in [0, 1]$, $\rho(\lambda X + (1 - \lambda)Y) \leq \lambda\rho(X) + (1 - \lambda)\rho(Y)$ holds;
- ρ is called *coherent*, if it is convex and positively homogeneous, i.e. for $X \in \mathcal{X}$ and $c > 0$, $\rho(cX) = c\rho(X)$ holds;
- ρ is called *P-law-invariant*,[16] if the value of $\rho(X)$ only depends on the P-distribution of X, i.e. $\rho(X) = \rho(Y)$ holds if X and Y have the same distribution under P.

In practice, several "risk measures" are used. A traditional risk measure is, e.g, the variance, which was suggested for quantifying the risk of investments in portfolio theory in the seminal work of Markowitz [39].[17] However, the variance is not a risk measure in the sense of Definition 3.2, since it fails to be monotone.

The idea behind a risk measure is to compress all risk modeled by a random variable X into a single number $\rho(X)$. Obviously, this means that some information (i.e. the whole distribution of X) is lost and complexity is reduced, but it is a helpful and popular method to provide insight into risk for professional risk managers and to communicate to external audience. The convexity property translates into risk diversification: combining different risky quantities should not be penalized, i.e. the combined position cannot be riskier than the combination of the single positions. Furthermore, at first glance, the notion of π-translation invariance is rather unintuitive and difficult to understand: the interpretation is that the elements from $\mathcal{H}$ do not exhibit the kind of risk which is supposed to be measured ("risk-less positions"). Its risk quantification is solely determined by the linear mapping π, which is not risky

[15] In many cases, as discussed below, it is sufficient to think of $\mathcal{H}$ as the constants and of π as the identity function.

[16] If there is no ambiguity between different probability measures, the reference to the probability measure P is omitted.

[17] Harry M. Markowitz received the Nobel Memorial Prize in Economic Sciences 1990 for his groundbreaking research on portfolio theory.

by definition (since it does not exhibit risk diversification). In the original definition of convex risk measures, the subspace $\mathcal{H}$ only consists of the constant functions ("no risk") and the linear mapping π is simply the identity, i.e. $\pi(c) = c$.

The notion of risk measures was developed due to the shortcoming of classical risk measures as, e.g., quantiles (Value-at-Risk, often abbreviated by VaR), which in many cases did not exhibit desirable properties (e.g. VaR does not always support diversification). (Convex) risk measures provide a mathematically precise and rich framework for the measurement of risk, thus, it may also be adapted to measure model (resp. parameter) risk. The most popular non-trivial example of a convex risk measure is the Average-Value-at-Risk, which averages over the tails of a distribution and overcomes the shortfall of the Value-at-Risk being not convex.

The concrete implementation of the adaptation of the general framework of risk measures always depends on the setting what has to be measured, but, as a first idea, when a certain number $f(P)$ has to be calculated which depends on the probability measure $P \in \mathcal{P}$, it may be a sensible idea to apply the risk measure framework to the function f to provide a number accounting for the model (resp. parameter) risk.

Example: Option Pricing Incorporating Parameter Risk A canonical example where model/parameter risk arises is option pricing. For this task, one uses financial market models as described in Sect. 1 which heavily rely on parameters that are not directly observable on the markets. Hence, those parameters have to be estimated, either via time series analysis of financial data or via fitting to market prices of available instruments (e.g. call and put options). As pointed out in Example 3.1, the procedure of obtaining the parameters exposes one to parameter risk. If one wants to state a price for some option using a certain model, e.g. the Heston model, one should account for parameter risk in the chosen model.[18] For some option X, each parameter vector θ in a financial market model yields the risk-neutral price of the option X w.r.t. the parameter vector θ as an expectation $\mathbb{E}_\theta[X]$. But, different from the usual model output, option traders typically state two prices—a bid price (to which she or he is willing to buy the option) and an ask price (to which she or he sells the option). Hence, the key idea is that parameter risk is a crucial determinant for the width and location of the bid-ask spread.

Thus, for option pricing purposes, the notion of a (model) risk-capturing functional and risk-captured (ask and bid) prices are developed in Bannör and Scherer [17] using the theory of convex risk measures.

Definition 3.3 (Model Risk-Capturing Functional, Risk-Captured Prices) Let $\mathcal{Q}$ be a family of option pricing models[19] and let R be a probability measure on $\mathcal{Q}$. Let $\mathcal{D}$ denote all options X we seek to price, which additionally satisfy some technical conditions. Let furthermore ρ be a normalized, law invariant convex risk measure

[18] Actually, one should also account for model risk, but this may not be tractable any more.

[19] To remain consistent with the usual terminology from mathematical finance, we denote risk-neutral measures by Q and a set of different risk-neutral measures by $\mathcal{Q}$.

on some functions on $\mathcal{Q}$. Then the mapping $\Gamma : \mathcal{D} \to \mathbb{R}$, defined by

$$\Gamma(X) := \rho\big(Q \mapsto \mathbb{E}_Q[X]\big), \tag{3.1}$$

is called a *model risk-capturing functional* w.r.t. the distribution R. $\Gamma(X)$ is called the *risk-captured* (*ask*) *price of* X given the functional Γ. Furthermore, $\bar{\Gamma}(X) := -\Gamma(-X)$ is called the *risk-captured bid price of* X.

The definition of risk-captured prices is somewhat technical and involves many different requirements (mainly to ensure the existence of the objects we deal with), but, in principle, the concept of treating the number of interest—the option price—as a function of the random model and applying a risk measure to it remains the same. In this case, since the methodology is supposed to be used for option pricing purposes, some additional quantities are required (e.g. normalization) to ensure that the number $\Gamma(X)$ makes sense. Furthermore, convexity is crucial since model risk should be a risk that profits from risk diversification. Option traders always regard their positions from a portfolio point of view, quoting bid-ask prices according to their portfolio position (e.g. they give better prices for options fitting to their present position).

The definition of model (resp. parameter) risk-captured prices is related to the idea behind some other non-linear pricing ideas that were mainly used for pricing in incomplete markets (like, e.g., Carr, Geman, and Madan [24], Cherny and Madan [25]).

3.2 *Bayesian Treatment of Model Risk*

A popular mathematical tool, when confronted with model risk, is Bayesian statistics. The basic idea behind Bayesian statistics is that the relationship between distributions of different models and samples thereof is not static, but is a dynamic process where the knowledge of the model distribution is constantly enhanced/updated. In this case, the model (resp. the parameter) is regarded to be random as well. Hence, one of the key results of Bayesian statistics we will present here is how the model (resp. parameter) distribution is updated and learns from the collected samples. Summarizing, Bayesian methodology is about how to obtain a proper distribution on the models incorporating information about the data into the construction process. A standard reference on Bayesian theory is Bernardo and Smith [2], one can find more about Bayesian methods in Chap. 8 of Czado and Brechmann [27].

Bayes's theorem, going back to the English minister of the Presbyterian church Thomas Bayes, is—in its most basic form—a relationship of conditional probabilities. Interchanging the conditioning set with the set which is evaluated, the conditional probability can be easily derived. Formulated in a mathematically precise manner, Bayes's theorem states the following result:

Theorem 3.4 (Bayes's Theorem, General Version) *Let* $(\Omega, \mathcal{F}, P)$ *be a probability space and* $A, B \in \mathcal{F}$ *some events with* $P(A), P(B) > 0$. *Then the following relationship between the conditional probabilities of the considered events holds*:

$$P(B|A) = \frac{P(B)P(A|B)}{P(A)}.$$

At first glance, Bayes's theorem does not seem to have any interconnection with model risk and the application of Bayes's theorem towards model risk is not obvious. But when a distribution on the set of possible probability measures $\mathcal{P}$ is at hand, Bayes's theorem delivers an interesting interpretation of the relationship between the probability of outcomes and the probability of having the right model.

Therefore, let R be a probability distribution on the set of probability measures $\mathcal{P}$ quantifying the model risk, a joint probability measure Π living on the Cartesian product of the state space and the possible probability measures $\Omega \times \mathcal{P}$ may be defined on the "rectangle sets" via

$$\Pi(A \times B) := \int_B P(A) R(\mathrm{d}P) \tag{3.2}$$

for $A \times B \in \Omega \times \mathcal{P}$ (this measure may be extended to the whole product σ-algebra). The product measure Π can be interpreted as a probability measure which both incorporates possibilities of the outcomes and the different models. If we then apply Bayes's theorem to this situation, we obtain the following "model risk version" of Bayes's theorem.

Theorem 3.5 (Bayes's Theorem, Model Risk Version) *Let* Π *be defined as in* (3.2) *and* $\Pi(A \times \mathcal{P}) > 0$. *Then*

$$\Pi(\Omega \times B | A \times \mathcal{P}) = \frac{\Pi(\Omega \times B)\Pi(A \times \mathcal{P}|\Omega \times B)}{\Pi(A \times \mathcal{P})} = \frac{R(B)\Pi(A \times \mathcal{P}|\Omega \times B)}{\int_{\mathcal{P}} P(A) R(\mathrm{d}P)}$$

holds.

Defining suggestively $\Pi(A|B) := \Pi(A \times \mathcal{P}|\Omega \times B)$ as well as $\Pi(B|A) := \Pi(\Omega \times B|A \times \mathcal{P})$, one may summarize Theorem 3.5 via the handy expression

$$\Pi(B|A) = \frac{R(B)\Pi(A|B)}{\int P(A) R(\mathrm{d}P)}. \tag{3.3}$$

If we have a closer look on this formula, (3.3) reveals an interesting relationship between *model-intrinsic risk* (which is inherent in the different possible stochastic models $P \in \mathcal{P}$) and *model risk* (which is quantified by the probability measure R on the possible models $\mathcal{P}$). The probability that a set of stochastic models $B \subset \mathcal{P}$ is correct, given that a certain outcome $A \subset \Omega$ arrives, can be calculated by a fraction of the raw probability $R(B)$, corrected by a fraction which consists of the probability of the outcome A given the models B and the probability of A averaged over all possible models $\mathcal{P}$. Hence, starting with a probability measure R on $\mathcal{P}$ quantifying model risk, one may obtain some further information and correct for the outcome A. In particular, if $B = \{P_0\}$ consists only of the probability measure P_0 (with positive probability $R(\{P_0\}) > 0$), (3.3) reduces to the even simpler form

$$\Pi(P_0|A) = \frac{R(P_0)P_0(A)}{\int P(A)R(\mathrm{d}P)}. \tag{3.4}$$

In a model risk framework based on continuous risk, one often has that the probability for a single model P_0 is zero (i.e. risk that comes from Lebesgue-a.c. probability measures), so the convenient representation (3.4) is usually not available. But there is a way out to find a nice form for Bayes's theorem treating model risk: if we assume that a parameterization of the set of possible models $\mathcal{P} = (P_\theta)_{\theta\in\Theta}$ with $\Theta \subset \mathbb{R}^n$ is at hand, the model risk probability measure R has a density $r(\theta)$, and the random variable of interest $X : \Omega \to \mathbb{R}^d$ has density $p(x|\theta)$ under P_θ for all $\theta \in \Theta$, we obtain the classical model risk version of Bayes's theorem involving densities.

Theorem 3.6 (Bayes's Theorem, Parameter Risk Version with Densities) *Let r, $(p_\theta)_{\theta\in\Theta}$ be as above. Then the conditional density $r(\cdot|x)$ can be calculated via*

$$r(\theta|x) = \frac{r(\theta)p(x|\theta)}{\int_\Theta p(x|\theta)r(\theta)\,\mathrm{d}\theta}. \tag{3.5}$$

Theorem 3.6 suggests particularly that the distribution on the parameters (represented by the density r) can be updated and adjusted, given the information from the samples $x = (x_1, \ldots, x_d)$. This can be regarded as follows: one starts with a parameter distribution r[20] (which is usually called a priori distribution or *prior distribution*, since it is the distribution imposed without any further information) and observes samples $x_1, \ldots, x_d$ on the market. Now, the distribution r is adjusted to the observation of the sample $x = (x_1, \ldots, x_d)$. Roughly speaking, the weights on the parameters are adjusted according to the likelihood of the sample outcome $x = (x_1, \ldots, x_d)$. As a result, one obtains a new distribution represented by the density $r(\cdot|x)$ incorporating both the information which was given by the a priori distribution and the additional information contained in the samples $x_1, \ldots, x_d$. Consequently, the obtained distribution $r(\cdot|x)$ is called the *a posteriori distribution* or *posterior distribution* on the parameters given $x = (x_1, \ldots, x_d)$. The whole procedure is referred to as *Bayesian updating* or *Bayesian inference*, since the new information contained in the samples $x_1, \ldots, x_d$ causes the old beliefs of the parameter distribution (summarized in the a priori density r) to be updated, resulting in the a posteriori density $r(\cdot|x)$. Bayesian updating can be done constantly when new data is available. Often, the old posterior distribution then comes into play as the new prior distribution, which is again updated with information from new samples $\tilde{x} = (x_{d+1}, \ldots, x_{d+\tilde{d}})$. Figure 4 illustrates this updating procedure.

[20] In the following, we use the word *distribution* for abbreviation and mean the distribution induced by the respective density.

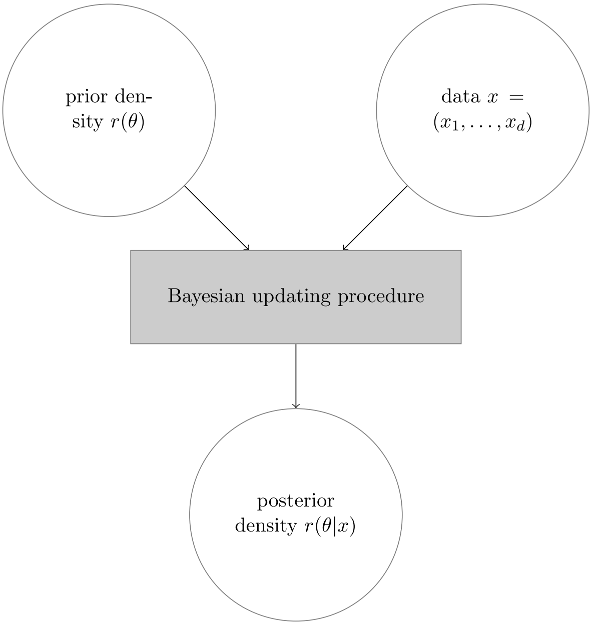

prior density $r(\theta)$
data $x = (x_1, \ldots, x_d)$
Bayesian updating procedure
posterior density $r(\theta|x)$

Fig. 4 This diagram illustrates the process which is done in Bayesian updating (here as a mathematical "black box"). The information from the prior density (*top left*) is merged with data samples (*top right*), resulting in a unified distribution (*bottom*)

Merging Expert Knowledge and Data Evidence into a Unified Framework A common application of the Bayesian updating process is when the input source is twofold: first, one has real-world data available for estimating parameters. A classical statistic paradigm would now solely rely on the given data, estimating the parameters and—if required—calculating the (asymptotic) distribution by using theory from mathematical statistics or resampling methods. But, in some cases, one wants to incorporate some expert judgement as well, particularly in case that the data may be difficult to judge (e.g. the data only reflects the recent past and some events not reflected in the past may happen in the future). Another case where one would like to incorporate expert judgements is when only very few data is available (like, e.g., operational risk events or corporate defaults) or a large fraction of data is outdated. For example, an option trader with long experience might impose a distribution on the parameters of a financial market model (e.g. Heston model) being subject to parameter risk (compare Example 2.1). Using a Bayesian updating procedure, one would use this distribution being the result of expert judgement as the *a priori distribution*. As a second step, one may use the Bayesian updating procedure and samples from financial market data (e.g. option prices) to adjust the expert view to real-world data.

One could also interpret Bayesian updating the other way round and start with a prior distribution that may be a sensible idea without having a closer look (prior distribution as "default distribution"). One can then use data or "expert estimates" to tilt the distribution towards results that are more in line with the data/the expert estimates.

As a result of applying Bayes's theorem in the version stated in Theorem 3.6, one obtains the *a posteriori distribution* integrating both the expert judgement as well as the data. Hence, loosely speaking, the a posteriori distribution may be regarded as a "merger" between the expert opinion and information extracted from data.

Examples The methodology of Bayesian updating has widely been exploited in practice. Due to its handyness in terms of mathematical formulae and its mathematical rigorousity, it is one of the first choices to obtain distributions on models and particularly parameters.

Example 3.7 (Black–Litterman Portfolio Selection) A popular application of Bayesian updating is the Black–Litterman approach to portfolio optimization, as described in Black and Litterman [20]. In classical Markowitz portfolio optimization, risk and return characteristics of different investments are purely estimated from data (e.g. time series, option prices). A clear drawback of this procedure is that the used data is backward looking and does not carry information about future developments. Hence, one would like to introduce some procedure where data is one input, but on the other hand some subjective market opinion may influence the result. One way to incorporate some "market opinion" additionally is to use a Bayesian approach. In this case, both subjective views of investment performance and risk (a priori distribution) as well as financial market data (typically time series of financial instrument prices) can be integrated by means of Bayesian updating. As a result, one obtains a new distribution for risks and returns which is used for portfolio optimization purposes, called Black–Litterman portfolio selection.

Also in option pricing, Bayesian methodology provides a framework to obtain a distribution on the parameters such that today's option prices can be merged with an external view, e.g. coming from expert judgement or exploiting "more probable" market information.

Example 3.8 (Bayesian Option Pricing) There have been several attempts to incorporate Bayesian ideas into option pricing, we only sketch few of them (a complete overview would be out of scope). As described above, option pricing is a situation where one is exposed to parameter risk (and, presumably, model risk). Hence, Bunnin, Guo, and Ren [22] and Gupta and Reisinger [32] both suggest to compute the posterior distribution via Bayesian updating incorporating new data like realizations from time series and (more forward-looking) prices of European options. Gupta and Reisinger [32] assume that put and call option prices follow a true model that is noised by independent error terms. A mathematical framework is suggested how this assumption can be interpreted in terms of a parameter prior distribution. In particular, a local volatility framework is used and it is assumed that in the short run, the market-implied Black–Scholes volatilities of the most popular options[21] are concise approximations for the local volatility.

[21] Usually, the options which are most traded are the options with a strike close to today's stock price, the so-called at-the-money options.

An interesting question remains from choosing the prior distribution. Once having done the Bayesian updating procedure several times, one may use the old obtained posterior density as the new prior to start with as described above. Later, we refer to the Bernstein–von Mises theorem treating the asymptotic impact of the prior distribution.

In insurance applications, one is often more involved with using time-series data due to more stationary conditions (e.g. fire claims or other insurance losses observe more stationary behavior as financial markets). Many textbooks as, e.g., Böcker [21], Klugman [35], Wüthrich and Merz [49] address Bayesian methods for risk management in insurance and finance.

4 Dealing with Model Uncertainty

In nichts zeigt sich der Mangel an mathematischer Bildung mehr, als in einer übertrieben genauen Rechnung.[22]*—Carl Friedrich Gauss, mathematician*

In some cases, it is a hard task to quantify the probability of certain models to be the true model. It may be even impossible to impose a probability measure R on the set of different models $\mathcal{P}$ from which one may choose. In these situations, one experiences *true model uncertainty*. In such a situation, one has much fewer alternatives than in case of model risk, where quantification may be done via different risk measures, as we have described earlier. Conversely, in case of model uncertainty, one is typically restricted to consider worst-case scenarios: if there is no additional information and we have complete ambiguity between different stochastic models represented by the set of probability measures $\mathcal{P}$, one has little choice to boil the “degree of model uncertainty” down to one number as we have done it in case of model risk.

Worst-Case Approaches Mostly, one seeks to calculate a number $f(P)$ (e.g. the price of some option) which depends on the chosen model $P \in \mathcal{P}$. Not having any further information at hand, the easiest way (and maybe the only feasible one—since everything in the scope of the model set $\mathcal{P}$ is possible) to quantify model uncertainty (as described for option pricing by Cont [4]) is to take the worst cases (resp. best cases) between the different models, namely

$$u = \sup_{P \in \mathcal{P}} f(P), \qquad l = \inf_{P \in \mathcal{P}} f(P).$$

Hence, the whole model uncertainty may be quantified by the difference of the two numbers

$$u - l = \sup_{P \in \mathcal{P}} f(P) - \inf_{P \in \mathcal{P}} f(P).$$

[22] Translation: nothing shows the lack of mathematical education more than an exaggeratedly exact calculation.

The difference between the extremes is an appropriate number to measure the (maximal) impact of model uncertainty on the quantity f. In case of model risk, i.e. the knowledge about the likelihood of each model, worst-case approaches can also be done. But, due to the additional knowledge, many other alternatives (as, e.g., convex risk measures) are possible.

Often, the number of interest $f(P)$ is the expectation of some random variable X w.r.t. the probability measure P (as in the case of option pricing). If this holds, the theory of convex risk measures (a standard reference is Föllmer and Schied [6]) immediately yields that the quantity

$$u(X) = \sup_{P \in \mathcal{P}} \mathbb{E}_P[X]$$

fulfills all the axioms of a coherent risk measure (without law invariance). One can go even further and define the *upper envelope* of a set of probability measures by defining

$$\mu_{\mathcal{P}}(A) := \sup_{P \in \mathcal{P}} P(A), \quad A \in \mathcal{F}.$$

In general, the upper envelope $\mu_{\mathcal{P}}$ is not a probability measure any more, but a submodular set function. Here, we can still define some integral, the Choquet integral w.r.t. $\mu_{\mathcal{P}}$, and the quantity $u(X)$ can be represented as a Choquet integral

$$u(X) = \int X \, \mathrm{d}\mu_{\mathcal{P}}.$$

The Choquet integral is a generalization of the regular integral and relaxes some properties, e.g. it is not linear any more in general, but preserves features as, e.g., monotonicity. The rich theory of Choquet integration, delivering many tools to work with, can be found in the compendium of Denneberg [28].

Examples

Worst-Case Option Pricing Cont [4] describes the situation when a set of risk-neutral probability measures $\mathcal{Q}$ is available, but one does not have any information which one to pick for the valuation of some option X. As described above, it is suggested to use a worst case ansatz and to deliver two prices

$$u(X) = \sup_{Q \in \mathcal{Q}} \mathbb{E}_Q[X] \quad \text{and} \quad l(X) = \inf_{Q \in \mathcal{Q}} \mathbb{E}_Q[X],$$

which can again be interpreted as bid-ask prices. As described above, the functional u fulfills the axioms of a coherent risk measure. Conversely, if there is a coherent risk measure ρ which is defined on a suitable collection of random variables, general theory immediately yields that it can be represented as the supremum of the expectation w.r.t. some "stress-test measures" $\mathcal{Q}$, i.e.

$$\rho(X) = \sup_{P \in \mathcal{Q}} \mathbb{E}_Q[X]$$

holds for a set of "stress-test measures" $\mathcal{Q}$ which are absolutely continuous w.r.t. the original measure P. Hence, in this sense, convex risk measures as treated in Sect. 3.1 can also provide a framework to measure model uncertainty.

In some cases (as, e.g., the calibration to market prices), one might have additional information about the trustworthyness of a model, contained in some "penalty function" $\alpha : \mathcal{Q} \to [0, \infty]$. In this case, Cont [4] suggests "penalized worst-case pricing" by setting the two option prices via

$$u(X) = \sup_{Q \in \mathcal{Q}} \mathbb{E}_Q[X] - \alpha(Q) \quad \text{and} \quad l(X) = \inf_{Q \in \mathcal{Q}} \mathbb{E}_Q[X] - \alpha(Q). \tag{4.1}$$

On the other hand, it can be shown that, in principle, every convex risk measure can be represented in the style of (4.1) (cf., e.g., Föllmer and Schied [6]). The very general framework developed by Cont [4] may be understood best by stating an example. One prominent example incorporating a rich class of pricing models is the uncertain volatility model by Avellaneda, Levy, and Paras [15].

Example 4.1 (Pricing with Uncertain Volatility) As described earlier, in a Black–Scholes model (cf. Example 1.1), the assumption of volatility being constant has caused numerous critique. Hence, stochastic volatility models (e.g. the Heston model presented in Example 1.2) have been developed. Again, these models assume certain characteristics of the volatility process. Another approach, leaving many degrees of freedom, was suggested by Avellaneda et al. [15]: volatility is introduced to be a stochastic process, living on a compact interval; i.e. the volatility process $(\sigma_t)_{t \geq 0}$ has its range in an interval $[\sigma_l, \sigma_u]$ with $\sigma_u > \sigma_l > 0$. The bounds σ_u, σ_l may be obtained from expert judgements or data like, e.g., available implied volatilities of liquid options. With these implicitly imposed models, Avellaneda et al. [15] develop an approach based on control theory methods to calculate model-free upper and lower bounds for the price of options.

Dependence Modeling Another situation where model uncertainty may arise is dependence modeling: often, different stochastic quantities that are related to each other (e.g. weight and height of persons) should be modeled jointly. Typically, this is modeled by assuming the realizations to come from a random vector $X = (X_1, \ldots, X_d)$. Assuming that the univariate distributions of the random variables $X_1, \ldots, X_d$ are known, one still has to determine the interconnection between the random variables, i.e. the dependence structure. Fortunately, Sklar's theorem provides that the dependence structure of any multivariate distribution may be separated from the univariate marginal distributions and any dependence structure corresponds to some copula, which is a multivariate distribution function with uniform marginals (see, e.g., Nelsen [40]). However, the set of copulas provides a broad and rich source with numerous dependence structures like, e.g., elliptical copulas or Archimedean copulas. Hence there are still infinitely many copulas to choose from, and sometimes there is little evidence about how the dependence structure may look like. More on this class of functions can be found in Chap. 9 of Klüppelberg and Stelzer [36].

In many cases, the choice of dependence structure is crucial for modeling events correctly, a vividly discussed example being portfolio default risk. On the eve of the financial crisis of 2008, there existed massive misvaluation of financial products called CDO (collateralized debt obligations) that were structured from, e.g., housing mortgages. The key principle of these products was to bundle several credits and redistribute the credit repayments and interest payments into different slices (so-called "tranches") in the following manner: in case of default, all defaults first reduce the notional of the most junior tranche. After elimination of the most junior tranche through defaults, the notional of the second-most junior tranche is reduced by occuring defaults and so on. As pointed out by Heitfield [33], the valuation of CDO tranches heavily relies on the imposed model of the dependence structure between the credit defaults. Predominantly, Gaussian copulas were used to account for the dependence, but Gaussian copulas were not able to capture important stylized facts like, e.g., contagion effects and tail dependence.

Typically, in case of *dependence uncertainty*, one wants to calculate a quantity $f_P(X_1, \dots, X_d)$ and is uncertain about the dependence structure (represented by a copula model P) of the random vector $(X_1, \dots, X_d)$. This means that the set of possible models $\mathcal{P}$ is constructed such that the univariate distributions of the random variables $X_1, \dots, X_d$ do not vary, but the dependence structure, which can be summarized by

$$\begin{aligned}\mathcal{P} := \big\{ & P \text{ probability measure on } (\Omega, \mathcal{F}) \text{ with} \\ & \text{fixed marginal distributions } P^{X_j} \sim F_j \big\}.\end{aligned}$$

The optimization problem to solve is to find upper and lower bounds (as in the proposal of Cont [4])

$$u(X_1, \dots, X_d) = \sup_{P \in \mathcal{P}} f_P(X_1, \dots, X_d),$$

$$l(X_1, \dots, X_d) = \inf_{P \in \mathcal{P}} f_P(X_1, \dots, X_d)$$

for functions f, which may include numerous applications, e.g. the calculation of risk measures of portfolios of financial instruments $X_1, \dots, X_d$. An important result from copula theory is that the set of copulas has upper and lower natural bounds, called the Fréchet–Hoeffding bounds. These can be interpreted (at least in dimension $d = 2$) as "complete positive dependence" (comonotonicity) and "complete negative dependence" (countermonotonicity). But the Fréchet–Hoeffding copula bounds are not necessarily the copulas[23] which produce the upper and lower bounds $u(X_1, \dots, X_d)$ resp. $l(X_1, \dots, X_d)$ for the quantity $f(X_1, \dots, X_d)$. Hence, the problem of determining the right dependence structure to approximate the upper and lower bounds has to be tackled mathematically.

[23] For $d > 2$, the lower Fréchet–Hoeffding bound is not even a copula.

Puccetti and Rüschendorf [42] present numerical and computational techniques to calculate upper and lower bounds for special functions f, including important examples like the Value-at-Risk (VaR) of portfolios. Using the fact that the empirical equivalent of copulas can be regarded as rearrangements, an algorithm is developed to calculate the bounds $u(X_1, \dots, X_d)$ resp. $l(X_1, \dots, X_d)$. In particular, it turns out that the comonotonicity copula (the upper Fréchet–Hoeffding bound) usually produces not the largest Value-at-Risk, but a copula that manages concentrating mass to the tail in a uniform manner.

In the bivariate case, there is another approach by Tankov [47], which refines the upper and lower bounds for a functional $f(X_1, X_2)$ when some information about the dependence (i.e. Kendall's Tau, a standardized association measure which is often more suitable than the correlation) is given. This is used to compute model-free bounds for bivariate options (e.g. best-of-two options), given a certain level of association measured by Kendall's Tau.

5 Food for Thoughts

This chapter intends to give a brief survey about model risk and uncertainty with a tilt towards financial topics, but, obviously, there are several questions that naturally arise.

- In this chapter, model risk and uncertainty is discussed in the context of mathematical finance. Obviously, also in natural sciences, model risk and uncertainty plays an important role. As a detailed example for a discussion of model risk and uncertainty in a natural sciences context, we refer to the book of Cooke [26].
- Convex risk measures are a tractable and well-studied class of risk functionals, but convexity (resp. subadditivity) may be an assumption that is too strong for real-life applications. Thus, there have been numerous generalizations and enhancements of convex risk measures incorporating weaker properties, like quasi-convexity or comonotone convexity (resp. subadditivity), studied in Song and Yan [46].
- When incorporating model (resp. parameter) risk by using convex risk measures, one might think about continuity properties of the computed numbers when imposing different kind of distributions on the parameters. In particular, one might want that if there is a sequence of distributions $(R_N)_{N \in \mathbb{N}}$ on the parameter set Θ converging to some limit distribution R_∞, the sequence of numbers capturing the model risk w.r.t. the distributions $(R_N)_{N \in \mathbb{N}}$ should eventually converge to the number capturing the model risk w.r.t. the distribution R_∞. An application would, e.g., be the distribution induced by some consistent estimator $\hat{\theta}_N$ converging to the "true" parameter. It turns out that, dependent on the risk measure, different types of convergence yield convergence for different classes of risk measures. Some ideas which risk measures behave as desired with weak convergence can be found in Bannör and Scherer [17], a detailed technical analysis about different topologies on probability measures that induce convergence of the risk measures is given in Krätschmer, Schied, and Zähle [37].

- In case of Bayesian methodology, one key problem is the choice of prior distribution. In some cases, the Bernstein–von Mises theorem states that in the asymptotics, the choice of prior distribution does not matter any more (e.g. van der Vaart [13]). Hence, the more iterations one does in the Bayesian updating procedure, one obtains more stable results (in case of drawing the sample from a stationary situation). Conversely, there are also situations where the Bernstein–von Mises theorem does not hold, which lead to criticism of the Bayesian methodology.

6 Summary

We presented an introduction to stochastic modeling and highlighted some problems concerned with model specification and the decision process which model to select. We defined and distinguished model uncertainty and risk, both are situations one typically faces when modeling complex objects as, e.g., financial markets, in a stochastic manner. We mentioned various examples, primarily from mathematical finance, where model and parameter risk and uncertainty play a prominent role. We have outlined methods based on convex risk measures dealing with both model risk and uncertainty, furthermore, we gave insight into Bayesian updating, which can be a helpful tool to refine parameter distributions in case of parameter risk.

Acknowledgement We thank Claudia Klüppelberg and Roger M. Cooke for valuable remarks on earlier versions of the manuscript.

References

Selected Bibliography

1. P. Artzner, F. Delbaen, J.-M. Eber, D. Heath, Coherent measures of risk. Math. Finance **9**(3), 203–228 (1999)
2. J.M. Bernardo, A.F.M. Smith, *Bayesian Theory*, 2nd edn. Wiley Series in Probability and Statistics (2007)
3. F. Black, M. Scholes, The pricing of options and corporate liabilities. J. Polit. Econ. **81**(3), 637–654 (1973)
4. R. Cont, Model uncertainty and its impact on the pricing of derivative instruments. Math. Finance **16**(3), 519–547 (2006)
5. S. Figlewski, Derivatives risks, old and new, in *Wharton-Brookings Papers on Financial Services* (1998)
6. H. Föllmer, A. Schied, *Stochastic Finance*, 2nd edn. (De Gruyter, Berlin, 2004)
7. S.L. Heston, A closed-form solution for options with stochastic volatility with applications to bond and currency options. Rev. Financ. Stud. **6**(2), 327–343 (1993)
8. J. Hull, *Options, Futures, & Other Derivatives* (Prentice Hall, New York, 2000)
9. F.H. Knight, *Risk, Uncertainty, and Profit* (Hart, Schaffner & Marx, Chicago, 1921)
10. B. Øksendal, *Stochastic Differential Equations: An Introduction with Applications* (Springer, New York, 2003)
11. F. Salmon, Recipe for disaster: the formula that killed Wall Street. Wired Mag. (2009)

12. W. Schoutens, E. Simons, J. Tistaert, A perfect calibration! Now what? Wilmott Mag. **3** (2004)
13. A.W. van der Vaart, *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics (2000)

Additional Literature

14. K.J. Arrow, *Aspects of the Theory of Risk-Bearing* (Yrjö Jahnssonin Säätiö, Helsinki, 1965)
15. M. Avellaneda, A. Levy, A. Paras, Pricing and hedging derivative securities in markets with uncertain volatilities. Appl. Math. Finance **2**, 73–88 (1995)
16. L. Bachelier, *Théorie de la spéculation* (Gauthier-Villars, Paris, 1900)
17. K.F. Bannör, M. Scherer, Capturing parameter risk with convex risk measures. Eur. Actuar. J. 1–36 (2013)
18. S. Bertsch McGrayne, *The Theory that Would Not die* (Yale University Press, New Haven, 2011)
19. F. Biagini, T. Meyer-Brandis, G. Svindland, The mathematical concept of measuring risk, in *Risk – A Multidisciplinary Introduction*, ed. by C. Klüppelberg, D. Straub, L. Welpe (2014)
20. F. Black, R. Litterman, Global portfolio optimization. Financ. Anal. J. **48**(5), 28–43 (1992)
21. K. Böcker, *Rethinking Risk Measurement and Reporting Volume I—Uncertainty, Bayesian Analysis and Expert*. Risk Books (2010)
22. F.O. Bunnin, Y. Guo, Y. Ren, Option pricing under model and parameter uncertainty using predictive densities. Stat. Comput. **12** (2000)
23. P. Carr, D. Madan, Option valuation using the fast Fourier transform. J. Comput. Finance **2**, 61–73 (1999)
24. P. Carr, H. Geman, D. Madan, Pricing and hedging in incomplete markets. J. Financ. Econ. **62**, 131–167 (2001)
25. A. Cherny, D. Madan, Markets as a counterparty: an introduction to conic finance. Int. J. Theor. Appl. Finance **13**(8), 1149–1177 (2010)
26. R.M. Cooke, *Uncertainty Modeling in Dose Response: Bench Testing Environmental Toxicity*. Statistics in Practice (Wiley, New York, 2009)
27. C. Czado, E. Brechmann, Bayesian risk analysis, in *Risk – A Multidisciplinary Introduction*, ed. by C. Klüppelberg, D. Straub, L. Welpe (2014)
28. D. Denneberg, *Non-additive Measure and Integral* (Kluwer Academic, Norwell, 1994)
29. D. Ellsberg, Risk, ambiguity, and the Savage axioms. Q. J. Econ. **75**(4), 643–669 (1961)
30. R.A. Fisher, On the mathematical foundations of theoretical statistics. Philos. Trans. R. Soc. A **222**, 309–368 (1922)
31. I. Gilboa, D. Schmeidler, Maxmin expected utility with non-unique prior. J. Math. Econ. **18**(2), 141–153 (1989)
32. A. Gupta, C. Reisinger, Robust calibration of financial models using Bayesian estimators. J. Comput. Finance (2012, to appear)
33. E. Heitfield, Parameter uncertainty and the credit risk of collateralized debt obligations. Available at SSRN 1190362 (2009)
34. D. Kahneman, A. Tversky, Prospect theory: an analysis of decision under risk. Econometrica **47**, 263–291 (1979)
35. S.A. Klugman, *Bayesian Statistics in Actuarial Science: With Emphasis on Credibility* (Springer, Berlin, 2011)
36. C. Klüppelberg, R. Stelzer, Dealing with dependent risks, in *Risk – A Multidisciplinary Introduction*, ed. by C. Klüppelberg, D. Straub, L. Welpe (2014)
37. V. Krätschmer, A. Schied, H. Zähle, Comparative and qualitative robustness for law-invariant risk measures (2012)
38. K. Mainzer, The new role of mathematical risk modelling and its importance for society, in *Risk – A Multidisciplinary Introduction*, ed. by C. Klüppelberg, D. Straub, L. Welpe (2014)

39. H.M. Markowitz, The utility of wealth. J. Polit. Econ. **60**, 151 (1952)
40. R.B. Nelsen, *An Introduction to Copulas*, 2nd edn. (Springer, Berlin, 2006)
41. J.W. Pratt, Risk aversion in the small and in the large. Econometrica **32**, 122–136 (1964)
42. G. Puccetti, L. Rüschendorf, Computation of sharp bounds on the distribution of a function of dependent risks. J. Comput. Appl. Math. **236**(7), 1833–1840 (2012)
43. F.J. Samaniego, *A Comparison of the Bayesian and Frequentist Approaches to Estimation* (Springer, Berlin, 2010)
44. P.A. Samuelson, Rational theory of warrant pricing. Ind. Manage. Rev. **6**(2), 13–39 (1965)
45. L.J. Savage, *The Foundations of Statistics* (Wiley, New York, 1954)
46. Y. Song, J.-A. Yan, Risk measures with comonotonic subadditivity or convexity and respecting stochastic orders. Insur. Math. Econ. **45**(3), 459–465 (2009)
47. P. Tankov, Improved Fréchet bounds and model-free pricing of multi-asset options. J. Appl. Probab. **48**(2), 389–403 (2011)
48. J. von Neumann, O. Morgenstern, *Theory of Games and Economic Behavior* (Princeton University Press, Princeton, 1944)
49. M.V. Wüthrich, M. Merz, *Stochastic Claims Reserving Methods in Insurance* (Wiley, New York, 2008)

Part III
Risk Treatment in Various Applications

Chapter 11
Cost-Benefit Analysis

Jutta Roosen

The basic rationale of cost-benefit analysis lies in the idea that things are worth doing if the benefits resulting from doing them outweigh their costs.
Amartya Sen (Cost-benefit analysis, The University of Chicago Press, Chicago, 2001, p. 98)

Every society is facing a number of risks and their regulation requires many considerations. From an economic standpoint, it can be said that risks impose a cost on society. Avoiding and regulating risks equally engenders costs. In order to help public decision makers come to terms with these trade-offs, economists have developed the method of cost-benefit analysis. It is based on the simple idea that things are worth doing when the benefits from doing them are greater than their costs. As simple as this basic idea is, as tricky and controversial are the implications of putting it into practice. Issues of controversy relate to valuing environmental benefits, determining the value of human health and life, balancing the interest of current and future generations by discounting, and dealing with the biases of subjective risk perception when defining a rational risk policy. This chapter will introduce the basic assumptions underlying cost-benefit analysis and the procedures involved in conducting one.

Keywords Cost-benefit analysis · Discounting · Non-market valuation · Value of a statistical life

Mathematics Subject Classification (2010) 91B06 · 91B15

J. Roosen (✉)
Chair of Marketing and Consumer Research, TUM School of Management, Technische Universität München, Alte Akademie 16, 85354 Freising-Weihenstephan, Germany
e-mail: jroosen@tum.de

C. Klüppelberg et al. (eds.), *Risk – A Multidisciplinary Introduction*,
DOI 10.1007/978-3-319-04486-6_11,

The Facts

- Cost-benefit analysis is rooted in the ethics of utilitarianism: things are of value because they are valued by humans in their pursuit of happiness and well-being.
- Cost-benefit analysis allows for the systematic consideration of all effects of a public project or policy. All costs and benefits are evaluated in monetary terms and hence are comparable.
- Cost-benefit analysis also considers the effects of projects on the environment, nature, and human health.
- The choice of the discount rate is of crucial importance for the outcome of a cost-benefit analysis. Small changes in the discount rate can lead to large changes in the outcome of the analysis. This is because the rate enters via an exponential discounting exercise.
- While the basic idea of cost-benefit analysis is widely accepted, many issues of implementation are hotly debated. Hence it is of crucial importance to make explicit all assumptions of the analysis and to make outcomes of different project valuations comparable.
- Cost-benefit analysis does not ignore the implications of behavioral sciences. It is a response to the bounded rationality of decision makers, trying to make public policy accountable to principles of rationality.

1 Introduction

Methylmercury, an organic form of mercury, is a toxic compound that alters fetal brain development when there is significant prenatal exposure (EFSA [18]). Exposure results from fish consumption and in particular concerns children of women who consume large amounts of fish before and during pregnancy. These children have a significant vulnerability to the adverse neurological effects of methylmercury (Budtz-Jorgensen et al. [13]). Levels of mercury in the environment have increased considerably over the last century. The most important anthropogenic sources of mercury are coal-fired power plants. When atmospheric mercury created during the coal burning process is deposited on surface water, bacteria convert it to the organic form, methylmercury. It then enters the food chain of aquatic life and accumulates in fish tissues. Moreover, methylmercury bio-accumulates in the food-chain leading to high mercury concentrations in predatory fish such as tuna, mackerel, and shark (Shimshack et al. [38]).

Because of its valuable nutrition properties (omega-3 fatty acids, proteins, vitamins and minerals) fish has taken centre stage in regulatory debates on food safety and nutrition (Caswell [14]). Policies dealing with methylmercury include power-plant regulation by capping mercury emissions. Because of the persistence of mercury in the environment, limiting emissions will not suffice for managing this risk and consumption advisories are an important means to limit exposure to contaminated fish for groups at risk (pregnant women, women of childbearing age and young children). Risk advisories, however, may also have spill-over effects to consumer groups not at risk (men and children at older ages), hence causing them to forego benefits of fish consumption such as omega-3s that are considered of importance to cardiovascular health. Balancing all these benefits, costs and risks of regulatory choices demand a careful analysis of all effects involved. In such cases, economists turn to cost-benefit analysis, comparing and aggregating all impacts valued in monetary terms.

Before entering into the description and discussion of cost-benefit analysis, a second example may be of interest. Pimentel et al. [31] have estimated the cost of soil erosion and benefits of conservation technologies. Soil erosion is a major environmental and agricultural problem around the world. Unsustainable agricultural practices lead to wind or water erosion that threatens the fertility of agricultural soils. Beyond this productivity impact, wind erosion can lead to pollution with fine-particulate matter, with an effect on human health because fine particles—solid or liquid—can get deep into the lungs and cause serious health problems such as aggravated asthma and increased respiratory symptoms (EPA [19]). Fine particles also lead to impaired views and damage material and countryside. Water erosion on the other hand can lead to soil run-off into streams and lakes and cause biological and recreational damage there. Again, public decision makers must consider what to do about soil erosion. Conservation practices are available, but may lead to reduced yields and profits in agriculture. Weighing of these costs and benefits of soil conservation can be done by cost-benefit analysis.

Cost-benefit analysis is a method of applied economics employed to evaluate public projects that involve investments and costs and returns over time. It helps the public decision maker to decide whether a project should be carried out or not or which project should be selected when several are under consideration. Basically the idea is to judge if public resources are used efficiently (not just effectively) and hence, if the costs of a project can be justified by its benefits. Cost-benefit analysis can be performed on all sorts of different projects. It can be applied to infrastructure projects such as the construction of a new road or railway. Even when judging educational projects, cost-benefit analysis can be useful. For example the OECD regularly publishes an evaluation of investment in higher education in its member countries (OECD [29]). Cost-benefit analysis also allows for assessing environmental regulation, such as a ban on specific pesticides or the conservation of an endangered species. One must however consider an important aspect of cost-benefit analysis: it is appropriate only for projects that are marginal for the public decision maker. Marginal means the project is relatively small in the overall portfolio of public projects. This is the case with the soil erosion or methylmercury example considered above. However, there may also be non-marginal projects such as those where the investment is of the size of a large share of the gross domestic product (GDP) of a country or projects for limiting and mitigating climate change, the impact of which may not be considered marginal. In these cases general assumptions of cost-benefit analysis on the risk aversion of the public decision maker and the income effect are no longer applicable. Hence particular care must be taken when conducting welfare assessments under such circumstances.

Cost-benefit analysis is also useful for health-related projects. For example, it can be used to address the question whether the benefits of cancer screening over large segments of the population outweigh the costs (cf. Chap. 17, [42]). Here ethical and methodological considerations on the evaluation of non-market goods become particularly thorny. Is there a monetary value to a human life, and how can we go about the evaluation of human life? In order to avoid such questions, medical professionals often prefer to replace cost-benefit analysis by cost-utility analysis or cost-effectiveness analysis. However, cost-benefit analysis not only provides information on the attractiveness of one medical treatment compared to other medical treatments. It helps the public decision maker to know if the investment in a health program is overall efficient, and furthermore if it is more efficient or less efficient compared to other projects such as investments in infrastructure or education.

The basic idea of cost-benefit analysis is that one can use monetary values to evaluate a project. That means that all costs and benefits are evaluated in terms of money. These monetary values are not only used for market aspects of the project, but also for non-market goods. Take the example of installing a wind farm. The flows of money involved are those of the investment costs at the beginning of the project. There are also running costs of maintaining the wind farm over time and there is certainly the benefit of the electricity generated. In this example, costs and benefits can easily be evaluated by using the market costs of the necessary investment, the maintenance costs and the generated electricity evaluated at market price. Certainly, a project is attractive if over the lifetime of the project the generated benefits are greater than the accruing costs, and there are a number of decision rules that can be used in order to verify if the project is attractive or not.

However, there are a number of difficulties that arise when doing cost-benefit analysis. First, costs as well as benefits occur over time. To make this flow of costs and benefits comparable over time, they have to be discounted to net present value (*NPV*). While discounting seems a simple mathematical exercise, the ethical implications of discounting are quite large. Economists have hence fiercely debated the choice of the appropriate discount rate. Second, benefits and also costs occurring in the future are uncertain. One hence has to resort to expected utility analysis (cf. Chap. 3, [40]). Finally, benefits and costs can also involve non-market goods. Consider again the example of the wind park. There may be the benefit of reduced CO_2-emissions versus damage to wildlife such as birds. For those non-market goods, market prices are not available and the value of the benefits and damages (costs) would need to be estimated. Section 4 of this chapter will briefly introduce the methods that are available for the valuation of non-market goods.

The idea of cost-benefit analysis has a long tradition and dates back to large engineering projects as illustrated by a publication by Jules Dupuit in the mid-19th century. It was formally introduced in the regulatory process in the United States for works by the Army Corps of Engineers by the Flood Control Act of 1936 (Persky [30]). In the great infrastructure projects before and during the New Deal policy of the Roosevelt administration the Army Corps of Engineers developed rules to assess projects, also accounting for non-economic impacts. Since then, the idea of cost-benefit analysis has evolved and while today a standard procedure in the United

States, it is gradually being practiced more in EU regulation. Since the treaty of Maastricht (1992), the Community has to take into account costs and benefits of action and lack thereof when preparing its policy on the environment. However, only slowly is a body of good practice being established in the European regulatory process (Renda [34]).

Before entering into the details of cost-benefit analysis, a few words are in place regarding its role in a book on risk and security. As illustrated by the examples above, many of the regulatory questions involve risks. However, these risks occur at the level of single individuals. For society as a whole many of these risks are expected damages that can be calculated as the probability multiplied by the size of damage. The general assumption is that the regulator is not risk averse (cf. Chap. 3, [40]) and that it suffices to consider expected costs and benefits. Stated otherwise and in reference to Chap. 1, [47], the US Army Corps of Engineers applied a deterministic concept of risk when introducing cost-benefit analysis in relation to their projects. Only very recently are probabilistic and equity considerations taking more room in the debates. In this sense, cost-benefit analysis can be seen as a tool to support better risk management, based on the results of thorough risk analysis.

2 Doing a Cost-Benefit Analysis

In principle, the process of cost-benefit analysis is similar to that of any private investment appraisal. However, the public decision maker, e.g. the government of a country, is a large decision maker and may take into account particular considerations for discounting the future in a wide portfolio of projects. Furthermore, because the public decision maker must also account for market failures such as those caused by public goods and externalities, specific considerations apply for the evaluation of costs and benefits.

Public goods are for example clean air, a bridge over a river, or a scenic view. A defining characteristic for public goods is that they are not excludable, i.e. everybody can consume them, and non-rival, i.e., an additional person consuming the good does not reduce the consumption of others. Take the example of clean air in a city. Everybody living in the city will benefit from good air quality (non-excludable). Clean air is also not depleted if other people consume it too, for instance tourists visiting the city, so clean air is a non-rival good. For a bridge crossing a river it is the same. Everybody can use it (non-excludable), and the utility of using it does not decline when others use it as well (non-rival).[1] Certainly in this example one could propose a road toll, so that only those who have paid can cross the bridge. As such the good would become excludable. However, given the non-rivalness in consumption a toll would lower social welfare. This is the rational for public goods being

[1] Non-rivalness may be limited by congestion. This can concern the example of the bridge when it comes to traffic jams or the example of clean, fresh air, when a small room with many people is considered.

provided by the public and paid for by taxes rather than by charging prices as is the case for private goods.

Externalities are closely related to public goods. They result from market activities; however, they are not valued in the market. Similar to public goods, they are not excludable. Externalities can be negative, such as the air pollution caused by a coal power plant. The plant operator will consider material capital and labor costs when taking production decisions. He will also consider the price of electricity that determines revenue. He will, however, not consider the air pollution caused by the plant, even if it leads to impaired views, respiratory diseases or accumulation of mercury deposits. These are costs that accrue to society, but these will not have to be covered by the plant operator. Hence these costs are external to the firm's production decision. An externality can also be positive. Take the example of a vaccine. A vaccine is used to protect an individual from a communicable disease; however, by taking the vaccine the disease pressure in a society can be considerably reduced so that other people also benefit. This balancing of private and public benefits is one explanation for the observation of decreasing vaccination rates after a disease is considered overcome, leading to new outbreaks such as the polio-outbreak reported in Central-Asia (WHO [46]). The individual benefit of vaccination has declined (lower probability to contract a disease), but by lower vaccination rates society may put this accomplishment at risk (increased probability for the disease to come back). Section 4 will explain how a value for public goods and externalities can be estimated. First of all, we will look at the economic foundations of cost-benefit analysis.

Cost-benefit analysis seeks to identify projects that make society better off. It is rooted in welfarism and utilitarianism, so what matters to society is the well-being of people. For example, nature as such can matter to social welfare, but only in so far as it matters to people and their utility.

Assume we have a social welfare function that is a functional of all the individual utility functions of a society consisting of N people, $W_0 = \Phi(U_1(c_1), U_2(c_2), \ldots, U_N(c_N))$. Here, c_i denotes an index of consumption for individual $i = 1, 2, \ldots, N$ in the society and $U_i(c_i)$ denotes the resulting level of utility for individual i (with $\frac{\partial U_i}{\partial c_i} > 0$ and $\frac{\partial^2 U_i}{\partial c^2} \leq 0$). One can think of c_i as a money-value index, an aggregate of current and future consumption including private and public goods.[2] The welfare function is assumed to be differentiable, increasing ($\frac{\partial W_0}{\partial U_i} > 0$) and concave ($\frac{\partial^2 W_0}{\partial U_i^2} \leq 0$) in individual utility levels.[3]

[2] Consumption can be taken as a conglomerate of all different consumption goods, including public goods and externalities. Alternatively, it could be a placeholder for a vector of all goods/bads consumed (including public goods, environmental amenities, health etc.) that affect human well-being. When considering only private goods lifetime consumption will be constraint by lifetime income, closely related to wealth of an individual, i.e., $U(W)$ in Chap. 3, [40]. However, in welfare economics, personal well-being is most often considered to depend on consumption (not income), because not all types of consumption require expenditures.

[3] The assumption of concavity implies a societal preference for equality. The closer the welfare function is to a linear function, the easier it is to balance utility of somebody very poor against utility of somebody very rich.

Now suppose a project A is implemented, changing the consumption levels of people by Δc_i. Hence social welfare if implementing project A becomes

$$W_A = \Phi\big(U_1(c_1 + \Delta c_1), U_2(c_2 + \Delta c_2), \dots, U_N(c_N + \Delta c_N)\big). \tag{1}$$

At the social welfare level, we can conclude that a project increases social welfare if $W_A - W_0 > 0$. There exists one major difficulty in this assessment: for many projects some members of society will gain and others will lose. This means there are people who gain with $U_i(\Delta c_i) > 0$ and people who lose with $U_i(\Delta c_i) < 0$. Cost-benefit analysis deals with the question of how to trade off utility increases by those who gain against utility decreases of those who lose.

Welfare analysis has been conceived to draw welfare conclusions in such situations where a public project/policy is under consideration. The social welfare function is a powerful analytical tool in this regard; however it is not easily determined. As a matter of fact, Arrow [9] has proven that a social welfare function may not even exist (for an accessible treatment the reader may consult Mueller [27, pp. 384–399]).[4] Hence in order to judge welfare impacts, the economist and mathematician Vilfredo Pareto (1848–1923) suggested a criterion to make such efficiency judgements. The *Pareto* criterion states that a project/policy is considered *welfare enhancing* if nobody loses and at least some members of society are made better off (the strong Pareto criterion). A weak version of the Pareto condition is that a policy change is desirable if everybody in society is made better off (Johansson [22]).[5]

The Pareto criterion makes a lot of sense. Everybody can probably agree to the weak version of the criterion, and if there is no envy, then there will also be no opposition against the strong version of the Pareto criterion. However, the Pareto criterion has one important weakness: most projects do not only have winners: some members of society will lose. Consider the example of an infrastructure project such as the construction of a new airport runway. While the region may benefit as a whole, those living close to the airport will suffer from augmented noise and pollution. Because of this need to weigh off gainers and losers, Hicks (1939) and Kaldor (1939) proposed a *compensation criterion* in two independent publications. Take the case where a project under consideration moves the economy from a consumption level $(c_1, c_2, \dots, c_N)$ to $(c_1 + \Delta c_1, c_2 + \Delta c_2, \dots, c_N + \Delta c_N)$. According to Kaldor the project is desirable, if it is *hypothetically* possible to redistribute income (and hence consumption) such that everybody becomes better off with the project than without the project. The Hicks criterion states that a project is desirable if it is not possible that the losers bribe the gainers to forego the project (Johansson [22]). In this sense both criteria take the stance that compensation must theoretically be possible.

[4] Kenneth J. Arrow received the Nobel prize jointly with John R. Hicks in 1972 "for their pioneering contributions to general economic equilibrium theory and welfare theory". In its award ceremony speech the prize committee stated with regard to Arrow's contribution "This conclusion, which is a rather discouraging one, as regards the dream of a perfect democracy, conflicted with the previously established welfare theory" [28].

[5] The weak Pareto criterion is weaker in the sense that every project that passes the weak test also passes the strong test, but not vice-versa. Obviously, fewer projects will pass the weak test.

The two versions of the compensation criterion take a different baseline perspective. Kaldor starts his analysis from the situation before the project, whereas Hicks considers the wealth distribution after the project.

The compensation criterion is quite useful because it now allows ranking projects that could not be ranked by the Pareto criterion. Note that the compensation criterion speaks only of the hypothetical possibility of compensation and not about implementation. This is because both authors (Kaldor and Hicks) as well as many other economists consider the question of efficiency separately from the question of distribution. They were foremost concerned with how resources should be used to achieve a maximum level of welfare.

The compensation criterion brings us close to the idea of cost-benefit analysis. However, one more crucial assumption is needed: cost-benefit analysis makes the assumption that the marginal utility of money is constant. What does this mean? It means that a loss of one Euro has the same impact on utility for all individuals. Hence we make the assumption that $\frac{\partial U_i}{\partial c_i} = \lambda$ for all i. In practical terms, it implies that taking away one Euro from one person and giving it to another person leads to changes in utility for these two people that net out.

The compensation criterion demands that those who gain can compensate those who lose to accept the project (or those who lose cannot bribe those who gain to forego the project) and hence requires that the question of distribution can potentially be solved in a way such that everybody is better off. As a result it is possible that everybody achieves a higher level of utility. If we furthermore assume a constant marginal utility of money, then we can state that the benefits of the project (in monetary terms) must be able to cover the costs of that project. Hence a project passes the cost-benefit test, if the benefits are greater than the costs.

Having discussed the economic foundation of cost-benefit analysis, the chapter will now consider procedural issues. That is, how to define a project and its effects? How to summarize benefits and costs over time and what is the appropriate discount rate? Finally, how to deal with uncertainty?

This section follows Hanley et al. [3] in dividing a typical cost-benefit analysis into six steps:

1. Definition of the project
2. Identification of the impacts of the project
3. Evaluation of the impacts
4. Calculation of the net present value
5. Application of the net present value test (or similar tests)
6. Conduct of a sensitivity analysis

All six steps are discussed one by one.

1. Definition of the Project There can be all types of projects considered via cost-benefit analysis. First it is important to define the limits of the project and its standing. "Standing" considers the issue of whose benefits and costs should count in cost-benefit analysis (Pearce et al. [5]). As a basic rule, all nationals should be included, whereas benefits and costs to non-nationals must be included according to specific considerations. Here it needs to be considered (a) if the project relates to international policy issues such as acid rain or climate change and (b) if there are ethical considerations for counting benefits and costs for non-nationals.

2. Identification of the Impacts of the Project A project has many implications on the use of resources and the creation of impacts. For example, if a coal power plant is constructed, electricity is generated (benefit) but air pollution may increase (cost). These costs of air pollution may be hard to estimate: they may include changes in human health and mortality (cf. the example of mercury exposure in the introduction). Labor and capital are used in the construction and contribute to the cost of the project. Alternatively, consider a new agricultural regulation that may limit the extent of soil erosion (benefit). Reduced soil erosion will have private benefits to land owners because the yield potential will be preserved for the future. Avoiding erosion also has a public benefit, because river-water quality will be improved. The costs of such a policy are born by farmers, who will have to invest in soil conservation technology.

3. Evaluation of the Impacts All identified impacts have to be valued in monetary terms. Suppose a project A leads to a flow of benefits and costs over time. The project starts in year $t = 0$ and runs for T years. We denote benefits evaluated as monetary benefits as B_t and costs as C_t for $t = 0, 1, \ldots, T$. In general, it is recommended to evaluate all benefits and costs in *real* monetary terms. That is, all money flows have to be deflated or evaluated at *current* prices ($t = 0$). The valuation of costs and benefits is easy if private goods are concerned. The value of these goods can be measured by their market price. The issue of valuing nonmarket goods (public goods, externalities etc.) is more complicated. Economists have proposed different valuation methods, notably the contingent valuation method, the hedonic pricing method and the travel-cost method. Those will be discussed later in this chapter; for now we assume that there are ways to calculate also the benefits and costs when market prices are not available.

For an example the reader is referred to Table 1. The project has a life-time of 20 years ($t = 0, \ldots, 19$) as shown in column 1. Columns 2–3 show the benefits and the costs of the project. The project is characterized by large investment costs in the first two years (100 each) and a major maintenance cost in year 10 (50). At the end the project ($T = 19$) is decommissioned with a cost of 50. Benefits accrue from year 2 onwards, with an exception of year 10, because the project has to be shut down for maintenance. After that the project is showing age with a decreasing flow of benefits over the second half of its lifetime.

4. Calculation of the Net Present Value At each point in time, the net value (NV) of the project can be calculated as $NV_t = B_t - C_t$.

As in any investment project, we have to account for the opportunity costs of time. This is done by discounting the flow of benefits and costs with the discount

Table 1 Example of cost-benefit analysis calculations, $d = 0.05$

t	B_t	C_t	NV_t	Discounted B_t	Discounted C_t	NPV	NPV at $i = r$
0	0	100	−100	0.00	100.00	−100.00	−100.00
1	0	100	−100	0.00	95.24	−95.24	−89.85
2	40	10	30	36.28	9.07	27.21	24.22
3	50	10	40	43.19	8.64	34.55	29.01
4	50	10	40	41.14	8.23	32.91	26.07
5	50	10	40	39.18	7.84	31.34	23.42
6	50	10	40	37.31	7.46	29.85	21.05
7	50	10	40	35.53	7.11	28.43	18.91
8	50	10	40	33.84	6.77	27.07	16.99
9	40	10	30	25.78	6.45	19.34	11.45
10	0	50	−50	0.00	30.70	−30.70	−17.15
11	40	10	30	23.39	5.85	17.54	9.24
12	40	10	30	22.27	5.57	16.71	8.30
13	40	10	30	21.21	5.30	15.91	7.46
14	40	10	30	20.20	5.05	15.15	6.70
15	30	10	20	14.43	4.81	9.62	4.02
16	30	10	20	13.74	4.58	9.16	3.61
17	20	10	10	8.73	4.36	4.36	1.62
18	20	10	10	8.31	4.16	4.16	1.46
19	0	50	−50	0.00	19.79	−19.79	−6.54
Sum				424.54	346.95	77.59	0

$NPV = 77.59$, $BCR = 1.22$, $IRR = 0.11$

Example taken from Conrad [15]

rate d. As can be seen in Table 1, discounted benefits are calculated as $(1 + d)^{-t} B_t$ and discounted costs as $(1 + d)^{-t} C_t$. In the example, the discount rate is set at $d = 0.05$. Because money that is invested in one project cannot be used otherwise, we have to account for this opportunity forgone. Because costs and benefits accrue over time, the net present value (NPV) for the project is calculated as follows:

$$NPV = \sum_{t=0}^{T} (1 + d)^{-t} (B_t - C_t). \tag{2}$$

Discounting occurs here with compound interest, i.e., using an exponential function. This leads to specific properties of discounted values and has triggered extensive discussions on the appropriate choice of the discount rate d. Section 3 of this chapter will be devoted to this issue. Taking the example in Table 1 with $d = 0.05$, the NPV results as 77.59.

Table 2 Decision criteria in cost-benefit analysis when deciding on a single project

Decision criteria	Formula	Decision rule
Net present value	$NPV = \sum_{t=0}^{T}(1+d)^{-t}(B_t - C_t)$	$NPV > 0$
Benefit-cost ratio	$BCR = \frac{\sum_{t=0}^{T}(1+d)^{-t}B_t}{\sum_{t=0}^{T}(1+d)^{-t}C_t}$	$BCR > 1$
Internal rate of return	$\sum_{t=0}^{T}(1+r)^{-t}(B_t - C_t) = 0$	$r > d$

5. Application of the Net Present Value Test (or Another Test) The net present value test considers if the net present value of a project is positive or not. Benefits are greater than costs and hence the project is socially desirable if

$$NPV = \sum_{t=0}^{T}(1+d)^{-t}(B_t - C_t) > 0. \tag{3}$$

In the example in Table 1, the $NPV = 77.59$ obviously passes the net present value test and the project should be implemented.

An alternative test would be to calculate the benefit-cost ratio (*BCR*) and to check if the ratio is greater than 1:

$$BCR = \frac{\sum_{t=0}^{T}(1+d)^{-t}B_t}{\sum_{t=0}^{T}(1+d)^{-t}C_t} > 1. \tag{4}$$

Referring again to Table 1, the *BCR* results as 1.22. Here again, the decision rule suggests implementing the project.

A third way to assess if a project is desirable is to calculate the internal rate of return (*IRR*) on the project. The *IRR* is defined as the discount rate, r, at which the *NPV* of the project would exactly be zero, that is:

$$NPV = \sum_{t=0}^{T}(1+r)^{-t}(B_t - C_t) = 0. \tag{5}$$

A project is then socially efficient if $r > d$, which means that the rate of return on the project is larger than the rate of time-preference of society. For the example in Table 1, the *IRR* is $r = 0.11$, which is greater than $d = 0.05$. Again the rule suggests implementing the project.

Table 2 summarizes the decision criteria when deciding on a single project.

One may wonder why there are many alternative decision rules. In principle they give the same result, but there are particular situations when one decision rule outperforms the others. In general, economists recommend using the net present value test.

When selecting projects with a limited budget, the *BCR* is useful. Let the available budget for investment be M. A public decision maker can choose between L mutually non-exclusive projects, each incurring an investment costs $I_l, l = 1, \ldots, L$ at the beginning of the project. E.g. for the project in Table 1 the investment costs would be the cost of 200 in years 1 and 2. Then sort all L projects by their *BCR*,

Table 3 Ranking projects

Project	Cost (C)	Benefits (B)	*NPV* (rank)	*BCR* (rank)
X	100	200	100 (1)	2.0 (3)
Y	50	110	60 (3)	2.2 (2)
Z	50	120	70 (2)	2.4 (1)

Source: Pearce et al. [5]

so that $BCR_1 \geq BCR_2 \geq \cdots \geq BCR_l \geq \cdots \geq BCR_L$. The projects selected should be $BCR_1, BCR_2, \ldots, BCR_{L0}$ such that $\sum_{i=1}^{L_0} I_l \leq M \leq \sum_{i=1}^{L_0+1} I_l$. That is the public decision maker should choose the projects with the largest *BCR* so that the available budget is sufficient to cover the investment costs of these projects.

Table 3 illustrates by example the advantage of the *BCR* rule when considering budget constraints.[6] There are three projects under consideration and they are not mutually exclusive. However, the public decision maker has a limited budget and hence can only realize projects that do not exceed a cost of 100. If projects were ranked according to the *NPV* criterion, project X would be ranked 1 and the available capital would be exhausted. According the *BCR* ranking, projects Y and Z would be realized (total cost equal 100). The total *NPV* of these projects is $130(= 60 + 70)$. This is higher than the *NPV* of project X alone (100).

The internal rate of return is often used to calculate the return on an investment and compare it across sectors. For example, in its report "Education at a Glance" the OECD regularly publishes internal rates of return for individuals obtaining higher education as part of initial education (e.g. OECD [29]). The private *IRR* for tertiary education in Germany for instance is 11.5 % for men and 8.4 % for women. This is slightly below the OECD average at 12.4 % and 11.5 %. Using the internal rate of return exempts the analyst from making assumptions regarding the discount rate. The problem though is that solving for r requires the solution to a higher degree polynomial that can have multiple solutions.

6. Conduct of a Sensitivity Analysis Typically, cost-benefit analysis requires making predictions for the future. How will benefits evolve and how will the costs? Cost-benefit analysis requires a lot of data, many based on estimates. Uncertainty can be found around the individual prices and also the physical and social impacts. Electricity prices may increase or decrease over time. Machinery wear may increase the maintenance costs of equipment. Weather and climate uncertainties may influence the agronomic yield impact of soil conservation policy.

Given all these uncertainties it is necessary to conduct a sensitivity analysis on all parameters that enter the cost-benefit analysis. For instance in the example of Table 1, do you come to the same conclusions, when annual benefits increase or decrease by 10 %? Would the project still be desirable, were the maintenance cost in year 10 to double? Sensitivity analysis helps to check the robustness of the results. It means repeating the same analysis with different value estimates. This can be done

[6] A mathematical proof would maximize net benefits subject to the constraints.

Table 4 The *NPV* in dependence of discount rate d and time t

Discount rate d	*NPV* of 100 Euros after ... years				
	1	5	10	50	100
0.03	97.09	86.26	74.41	22.81	5.20
0.05	95.24	78.35	61.39	8.72	0.76
0.10	90.91	62.09	38.55	0.85	0.01

by considering a limited number of scenarios. Or it can be done in a systematic way by Monte Carlo simulation.

3 The Discount Rate

One number of crucial importance in cost-benefit analysis is the discount rate. It is used to calculate the *NPV* and hence to make costs and benefits that accrue over time comparable to each other. This single number is one of the most debated issues in cost-benefit analysis overall.

Discounting occurs because cost-benefit analysis originates from welfare economics and individual preferences. Those are summarized in the utility function and it has been observed that individuals prefer now to later. The discounting of future benefits hence should use the rate that expresses this time preference. Discounting occurs also because when investing today, we forego the opportunity to invest tomorrow. This opportunity cost of time should be considered in the discount rate. For a general treatment of intertemporal decision making the interested reader is referred to microeconomics textbooks such as Varian ([43], Chap. 19).

A typical way to determine this discount rate is to use the interest rate on long-term government bonds. Government bonds are issued by a country in order to borrow money. The interest rate that the government has to pay for borrowing money also shows the opportunity cost of time for conducting public projects. Compared to individual borrowers, the government bears lower risk premia and interest on its bonds because it can raise taxes in order to redeem these bonds.

Because discounting with discount factor $(1+d)^{-t}$ results in compound of interest and hence is exponential, it is crucial to consider the effect of choosing the discount rate d. Table 4 provides an example discounting 100 Euros at three different rates (0.03, 0.05 and 0.10) and over different time periods (1 year up to 100 years).

It can be observed that a higher discount rate yields a lower *NPV* and the impact of this discounting is larger the longer the time span. When discounted at a rate of $d = 0.03$, 100 Euro in year 1 will have a present value of 97.09, but 100 Euro in year 100 only have a present value of 5.20. The effect is even stronger for a larger discount rate, so that the present value of 100 Euro in year 1 discounted at a rate of $d = 0.10$ will have a net value of 90.91.

Mathematically, this observation is quite obvious. On ethical grounds though, it leads to fierce debates. As shown in the example in Table 1, the typical cost structure of public projects implies that costs have to be born at the beginning of the project and benefits accrue only after a significant investment has been made. Hence, benefits are discounted for long time periods whereas costs are not. Certainly the public decision maker should account for the time preference of society and hence choose a positive discount rate. But what if projects run over very long time spans that cover several generations? This question has become the centre point of the discussion since economists have started to consider policies for mitigating and reducing climate change (for this debate see Arrow [10]; Stern [39]; Weitzman [45]; Gollier [20]; Gollier and Weitzman [21]). Climate change is an issue that may have implications for many generations to come and because of the implications of Table 4 the appropriate choice of the discount rate is of considerable importance and some argue for a discount rate of zero for reasons of intergenerational justice. Basically the question is whether we can justify that the costs of avoiding climate change born by the current generation are not as heavily discounted as are benefits enjoyed by future generations.

4 Estimating the Costs and Benefits of Nonmarket Goods

We have seen that cost-benefit analysis is quite similar to the appraisal of investment projects. The flow of costs and benefits is evaluated over the lifetime of the project and similar rules such as the *NPV* rule or the calculation of the *IRR* are applied. However, for public projects there is one important distinction. Often public goods and externalities are involved or even the primary motivation to start a public project.

Public goods and externalities were defined in Sect. 2 and by their very definition do not have a market value. Hence it is not possible to use market prices for evaluating costs and benefits related to them. But the fact that there are no market prices does not mean that there is no value. Economists have developed sophisticated methods to measure such values by estimating people's willingness to pay for non-market goods.

Consider the case of water quality in a lake. Water quality protects aquatic life, it enhances the scenic view for residents and tourists and it improves the quality of leisure activities in and on the lake like swimming or fishing. We return to the individual utility function to consider the value of such water quality. Utility in this case depends on the consumption of private goods that are accounted for by an individual's level of current wealth, W_i. It also depends on the environmental quality, e_0 (note that we drop the subscript i because water quality—a public good—is the same for everybody). Hence, utility can be described by $U_i(W_i, e_0)$. It is increasing in both arguments. Now suppose that water quality can be improved, e.g., through the regulation of run-off from agricultural fields into the lake by requiring a green corridor of 5 m along the fields bordering the lake and creeks running into it. Environmental quality would be enhanced to e_1. While the change in e refers to some kind of water quality indicator, we would be interested in the value of this water quality change to an individual. Since utilities cannot be compared across individuals, we would use a monetary evaluation of that change.

One measure of this utility change would ask for people's *willingness to pay* (*WTP*). This *WTP* is the maximum amount of money that people would be willing to forego to obtain the change in environmental quality. It is implicitly defined by the following equation:

$$U_i(W_i, e_0) = U_i(W_i - WTP_i, e_1). \tag{6}$$

On the left-hand side of the equation, a lower environmental quality e_0 leads to a lower utility compared to the right-hand side of the equation. The higher utility caused by the enhanced environmental quality e_1, however, is compensated through a decrease in wealth by the amount WTP_i. This WTP_i is the amount of money that individual i is willing to pay for the environmental improvement. Because it compensates the environmental improvement, it is also named *compensating variation*. This compensating variation may be relevant to determine the tax charge that people are willing to pay for agri-environmental programs that limit the amount of agricultural run-off. Note that water quality is a property of the water—hence the same for everybody. The value of that water quality measured in WTP_i, however, may differ between individuals.

Another way to measure the utility impact of such a policy is to use *willingness to accept* (*WTA*). It asks the question of what amount of money people are willing to accept to forgo the environmental improvement. In mathematical terms, this means

$$U_i(W_i + WTA_i, e_0) = U_i(W_i, e_1). \tag{7}$$

This increase in wealth by WTA_i on the left-hand side is equivalent to an increase in the environmental quality on the right hand side. It is hence also named the *equivalent variation*. On theoretical grounds, compensating and equivalent variation should be of similar size. However, empirical studies have found that *WTA* estimates are considerably larger than *WTP* estimates, in particular when environmental goods are concerned. Section 6 'Food for Thought' will return to this issue.

Now that we have a theoretical construct of the value of public goods and/or externalities, the question is how to quantify that value empirically. There are different methods available: the *hedonic valuation* method uses surrogate markets as does the *travel cost* method, which is often used for assessing the value of an environmental amenity such as lake water quality. Finally, people's values can be assessed using survey methods such as the *contingent valuation* approach. The hedonic valuation and travel cost method observe people's decisions in relation to the non-market good that is being evaluated. Hence these methods are called *revealed preference* methods, because they are based on preferences as they are revealed in people's decisions. The contingent evaluation method is based on *stated preferences*, that is, people state how they would decide in hypothetical scenarios that are described in the survey. The following paragraphs will describe the three methods in more detail.

The hedonic valuation method considers that goods consist of bundles of attributes (Lancaster [24]). The description of each of these attributes will define the value of the good. This product characteristics approach has been treated in a market setting by Rosen [35], hence establishing a basis for the hedonic valuation method. For example, if we seek an estimate of the *WTP* for lake water quality, we can refer to the housing market as a surrogate market. The price of a house may be determined by characteristics such as the size in square meters, the number of rooms, the age of the house and whether it has a garden. Also neighborhood characteristics may count, such as the distance to employment centers, the quality of the local school and public transport. Finally, environmental characteristics such as the view to a lake may also be an important determinant of the house price and it may change with the quality of the lake.

In order to find the value of such environmental quality characteristics, a regression analysis would link the prices of houses, P, to all these characteristics:

$$P = f(\textit{house-}, \textit{neighborhood- and environmental characteristics}). \tag{8}$$

The next section will provide an example for the hedonic valuation method when valuing risk to life and health.

The travel cost method uses people's travel choices to estimate the value of a public good such as lake water quality. It is one of the oldest environmental valuation techniques and it has been developed in the US in the context of valuing recreation in national parks (Hanley et al. [3]). The travel cost method makes use of the idea that environmental amenities are valuable for recreation activities and that recreation requires expenditures in terms of time and money. Monetary expenditures are needed for travel (car, gasoline, bus ticket) and also time is a scarce resource and hence has an opportunity cost. To implement such a travel cost method for the example above, visitors to the lake would be asked about the distance they had to travel and the time spent on travel and on the lake. Improvements in water quality could lead to an increase of visits by recreational fishermen and using the travel cost method the corresponding value could be estimated.

The contingent valuation method is a stated preference method. Here actual choices are not observed, but people are asked in surveys for their valuations. While economists generally prefer revealed preference methods to stated preference methods, the latter have some advantages. A salient feature is that preferences for nonexistent attributes can be elicited. For instance, if one is interested in a *WTP* for food safety as one characteristic of the food supply, the hedonic pricing method is hard to implement. In general, all food available on the market is considered safe and there are no explicit risk differences that experts and consumers would agree upon. In such cases it can be useful to estimate *WTP* using the contingent valuation method. When doing so, it is important to consider realistic scenarios in the valuation survey for ensuring that respondents do not misinterpret or ignore attributes. It is a characteristic of the contingent valuation method that it allows for the evaluation of hypothetical scenarios. This advantage is at the same time a major disadvantage, because elicited values may suffer from biases that are related to the hypothetical nature of the survey. This hypothetical bias is only one problem of the resulting estimates; other biases are related to the strategic behavior of the respondent and conceptual mistakes in conducting the survey. Economists have hence developed extensive toolkits to avoid the pitfalls of the contingent valuation method and the interested reader is referred to books such as the one by Carson and Mitchell [1].

The hypothetical valuation based on the characteristics approach has eventually led to the development of (hypothetical) choice experiments. Choice experiments (CE) have been developed in the context of transport studies and now have been brought into many different applications in environmental valuation (Adamowicz et al. [7]), marketing (Lusk et al. [25]) or medical treatments (Kjaer and Gyrd-Hansen [23]). In CEs respondents are asked to make repeated choices between different consumption bundles, which are described by different attributes. Typically, one of these attributes is price. This procedure enables the researcher to estimate *WTP* for each attribute considered in the CE.

This section introducing *WTP* as a concept for finding a monetary value for impacts that do not have a market value makes very apparent the anthropocentric welfare foundations of economics. Things are of value because people value them. The value may be related to the use of the resource (use value), but it may exist also other reasons (non-use values). Consider again the example of lake water quality given above. There may be some fish species, not used in commercial or recreational fishing, threatened by the deterioration of water quality. There is obviously no use value to the fish species, nevertheless there will be a loss if the species is lost. To take this loss into account in a cost-benefit analysis, economists consider aspects such as existence values (Hanley et al. [3]). The option value considers the value of preserving a resource (here a species), because it may become valuable in the future. It is hence part of the total value of preserving existence of the species. The existence value counts for the fact that the mere existence of a species is important to people. It may be motivated by selfish reasons or altruistic motives. For example, moral or religious reasons lead people to value the existence of a species or people may want to preserve the species for their children and grandchildren. While considerably widening the scope of economic values, these values still maintain the anthropocentric view that things are valuable because they are valued by humankind.

5 The Value of Risks to Life and Health

Sometimes projects involve also impacts on human health and hence human morbidity and mortality. As an example, let us look at the regulation of arsenic in drinking water in the United States (cf. Sunstein [41]; Raucher et al. [33] and references therein). Under US law, the Environmental Protection Agency (EPA) set a Maximum Contaminant Level (MCL) of 10 μg/L in drinking water. Respecting this MCL can require considerable water treatment costs. Because of economies of scale, the costs of the regulation per household are much larger in small communities compared to those in large communities. The benefit of the regulation is an estimated reduction in bladder and lung cancer cases. Based on EPA estimates, Raucher et al. [33] calculate that a reduction from 15 μg/L to 10 μg/L would avoid 4,450 cases of cancer per 1 million people exposed to the elevated level of arsenic, about half of which (53 %) would be fatal over a 70-year time span.[7]

In this example, the assessment of costs is relatively straightforward. What about the benefits? The benefits are the avoidance of a risk to human life. How can such benefits be valued in monetary terms?[8]

Many people would argue that a human life has infinite value or is even invaluable. Hence, no monetary value can be assigned to a human life saved. However, people take decisions every day that decide about their risk to health and life: a worker when she decides to eat healthily or not, a car driver when he decides to speed on the motorway or not, or a student when she decides to run a red light to turn up on time to an exam. All these decision have a small, albeit real impact on the probability of surviving the day. This observation has been used by economists to value life in terms of a small change of the likelihood of death due to the cause under consideration (cf. the deterministic approach to risk as explained in Chap. 1, [47]). Methods as described in Chap. 16, [8] can be used to estimate the risk at the population level and changes therein.

Public decision makers take such decisions many times. They decide to ban or not a pesticide that has been shown to have an impact on agricultural workers' and consumers' health. They modify or not a dangerous intersection in a city, so as to reduce the number of deaths due to traffic accidents. They decide to impose a speed limit or not. Economists have used the observed public decisions to calculate the implicit value of reducing risk to life and health from such data. That is, they looked at all sorts of regulations, the number of lives saved by these regulations and their costs. For instance Cropper et al. [17] analyzed the determinants of pesticide regulation decisions in the time span of 1975 to 1989 by the EPA. They show that the EPA indeed balances cost and benefits. However, the costs per cancer case avoided amount to $35 million for an applicator (farm worker) and to $60,000 for consumers of pesticide residues in food. That means saving the life of an agricultural worker costs as much as saving the life of more than 500 consumers. Such large differences in valuation lead to inefficiencies, and more lives would be saved if projects were selected on more rational grounds.

We link this value of reducing risk to life and health to the expected utility model that was introduced in Chap. 3, [40]. The model is based on Cook and Graham [16]. Assume that the state-dependent preferences of an individual are represented by a von-Neumann-Morgenstern utility function $U(W, H)$, where W denotes wealth and H denotes the individual's health state.[9] In this case, we consider two health states: $H = 0$ if the individual is dead and $H = 1$ if the person is alive. To simplify notation, let $U_0(W) = U(W, 0)$ and $U_1(W) = U(W, 1)$ and assume that $U_1(W) > U_0(W)$ for all W. This means at any level of wealth the utility is always higher when alive

[7] Sunstein [41] underlines the uncertainties related to the health damage estimation and states that the number of lives saved by the regulation may vary between 0 and 112.

[8] In their analysis, Raucher et al. [33] assume a Value of a Statistical Life of US-$7 million.

[9] Here we drop the subscript i to keep things simple.

rather than dead.[10] Let's also assume that utility increases with wealth, i.e., the first derivative $U_j' > 0$, but at a decreasing rate, i.e., the second derivative $U_j'' \leq 0$, for $j = 0, 1$.

Given the baseline mortality risk π, expected utility results as

$$E[U] = \pi U_0(W) + (1 - \pi)U_1(W). \tag{9}$$

The individual would be willing to forgo a part of his wealth W if offered the opportunity to reduce the health risk π by an amount p. We call the maximum amount of money that a person is willing to spend on the reduction of mortality risk *WTP*. As introduced in Sect. 4, *WTP* is defined such that the increase in expected utility due to the decrease in mortality risk is exactly offset by the decrease in utility because of the decrease in wealth. Mathematically stated:

$$\begin{aligned} &\pi U_0(W) + (1 - \pi)U_1(W) \\ &\quad = (\pi - p)U_0(W - \mathit{WTP}) + (1 - \pi + p)U_1(W - \mathit{WTP}). \end{aligned} \tag{10}$$

On the left hand side of the equation we see the expected utility before the change in mortality risk and on the right hand side we see the expected utility after the change. Under specific assumptions regarding risk preferences, it can be shown that this *WTP* is increasing as a function of the reduction of risk. That would mean people would be willing to pay more for projects that would reduce the mortality rate to a greater extent.

In valuing risk to life and health, researchers have mostly resorted to the hedonic valuation method. One market that has been used as surrogate market for the risk to life and health is the automobile market. It is based on the idea that road safety is valuable because it avoids deadly accidents. But how to value this benefit? What is it worth to people to be safe on the road? There is no market for road safety where you could find a price determined by the interplay of demand and supply. The hedonic valuation method would look for a market good that can serve as a surrogate market, that is, one that also values safety on the road. An obvious choice is the market for cars. Cars come in all sorts of brands and types and one characteristic is occupant safety in accidents. They are regularly tested by crash tests and reports can be found in relevant automobile magazines. The hedonic valuation method makes the assumption that the safety of people in a car during a car accident enters into the price of a car. Atkinson and Halverson [12] have done this in a publication in 1990. They estimate the value of reducing risk to life at $5 million per person (according to Viscusi and Aldy [44]). Another surrogate market for safety is the labor market. Jobs differ in their safety. A fire-fighter faces different risks compared to a white-collar worker. Differences in pay can be used to estimate workers' *WTA* risk using wage rates of different occupations correcting for educational and other job-related aspects. Viscusi and Aldy [44] give an overview of studies estimating the value of

[10] One could also argue that the utility of wealth when dead is zero. This would mean $U_0(W) = 0$. This assumption is often made. Relaxing the assumption means that we accommodate a bequest motive, that is people value bequeathing wealth to their children etc. at the end of their life.

reducing risk to life and health. Reviewing thirty studies based on US labor market data, they find estimates between \$0.5 and more than \$20[11] per life saved.

Returning to the example at the beginning of this section, you may still say, don't we all know that arsenic is a poison? Is there a reason not to get it out of the drinking water? One important aspect when managing risks and when conducting cost-benefit analysis is the issue of risk-risk trade-offs (Graham and Wiener [2]). Countervailing risks have to be considered. In a quest for better protection of the population, maximum contaminant levels of arsenic are fixed. However, the cost imposed on community water systems may be so high that other, more valuable opportunities for saving lives are foregone. Raucher et al. [33] discuss this point, in particular considering the lack of economies of scale of water treatment in small communities leading to higher cost per life saved.

6 Food for Thought

- The discount rate is crucial for projects that have intergenerational implications. Discuss the arguments in favour and against using a positive discount rate or a discount rate of zero when conducting cost-benefit analysis.
- In some countries the use of cost-benefit analysis is required for most policies but it is precluded when cancerogenic agents are the focus of the policy (e.g. pesticide bans etc.). Discuss the implications that such ruling may have.
- Consider and discuss reasons that may explain differences in *WTP* and *WTA* estimates.
- *WTP* evaluations are based on the subjective perceptions of the goods being evaluated. To exemplify the issues related to subjective risk perceptions, Pollack [32] has told the story of a town named Happyville. The citizens of Happyville have come to fear a contaminant in their drinking water well. The construction of water purification plant is hence proposed. The major of the city commissions a chemical analysis of the water in order to learn about the extent of water pollution. It turns out that there is no risk related to the water quality. Hence based on this objective risk evaluation no purification plant is needed as it would cause costs without creating a benefit. Nevertheless, the citizens of Happyville do not trust the scientific study and still insist on the purification plant.
- Using this or other examples, discuss the role of objective and subjective risk evaluation in willingness to pay estimates used in cost-benefit analysis. Are there reasons for considering subjective risk evaluations? How may this conclusion differ considering the possibility of people's reaction to the risk, e.g., drinking water risk compared to nuclear power risk? You may also refer to Salanie and Treich [36] to find arguments and look at Marette et al. [26] for an application.

[11] To make values comparable across different studies all results have been corrected for inflation to US-\$ values for the year 2000.

7 Summary

Managing risk in modern societies requires the regulation of impacts on human health and the environment. However, the extent of regulatory activities in many countries has made it necessary to consider what is ‘good’ regulation and what is not. Cost-benefit analysis can help to answer this question. This chapter has introduced the reader to the theory and practice of cost-benefit analysis. It made the underlying assumptions explicit, introduced the procedures step by step and discussed critical issues for empirical applications. Certainly, regulation is not only a question of efficiency. Other issues are at stake such as equity considerations, risk trade-offs, uncertainty about future impacts and irreversibilities, moral concerns regarding the limitations of utilitarianism to just name a few.

The role that cost-benefit analysis can play for good public decision making cannot be overrated. Arrow et al. [11] published a short note on the role that cost-benefit analysis can play in environmental health and safety regulation. They argue that cost-benefit analysis is useful for comparing the favorable and unfavorable effects of policies in a coherent manner. Considering the economic effects of different policies is very important for society and hence government agencies should not be precluded from taking such considerations into account. All assumptions made in a cost-benefit analysis should be made explicit and underlying uncertainties should be described. Cost-benefit analysis can hence help to identify efficient policies.

Despite the argument in favour of cost-benefit analysis, government agencies should also have the possibility to override the conclusion of the cost-benefit analysis, if there are good reasons to do so. Cost-benefit analysis can exemplify the cost to society of not following the result of a cost-benefit analysis and society will be able to judge, if the benefit sought is worth this cost. For instance, equity considerations may preclude the implementation of certain regulations, even if this comes at a cost on efficiency grounds. Also environmental projects may be rejected if they put species at risk even if this has been accounted for in the cost-benefit evaluation.

Sunstein [6] argues in favor of cost-benefit analysis because public choices are inherently complex. Humans are subject to limits of rationality in decision making (see Chap. 3, [40]). Why should policy makers and regulators be exempt from such irrationalities? In fact, most likely they are not. Doing cost-benefit analysis can save the public from irrational policy making and help to save resources for uses that are in the best interest of society.

Acknowledgement The author thanks Pierre Dehez, Charles Goldie, Claudia Klüppelberg and three anonymous reviewers for helpful comments on earlier drafts of this paper.

References

Selected Bibliography

1. R.T. Carson, R.C. Mitchell, *Using Surveys to Value Public Goods: The Contingent Valuation Method (Resources for the Future)* (Johns Hopkins University Press, Baltimore, 2000)

2. J. Graham, B. Wiener, *Risk Versus Risk: Tradeoffs of Protecting Health and the Environment* (Harvard University Press, Cambridge, 1995)
3. N. Hanley, J.F. Shogren, B. White, *Introduction to Environmental Economics* (Oxford University Press, Oxford, 2001)
4. E.J. Mishan, E. Quah, *Cost Benefit Analysis*, 5th edn. (Routledge, Abingdon, 2007)
5. D. Pearce, G. Atkinson, S. Mourato, *Cost-Benefit Analysis and the Environment: Recent Developments* (OECD, Paris, 2006)
6. C.R. Sunstein, *Risk and Reason: Safety, Law and the Environment* (Cambridge University Press, Cambridge, 2002)

Additional Literature

7. W. Adamowicz, P. Boxall, M. Williams, J. Louviere, Stated preferences approaches for measuring passive use values: choice experiments and contingent valuation. Am. J. Agric. Econ. **80**, 64–75 (1998)
8. D.P. Ankerst, V. Seifert-Klauss, M. Kiechle, Translational risk models, in *Risk – A Multidisciplinary Introduction*, ed. by C. Klüppelberg, D. Straub, I. Welpe (2014)
9. K.J. Arrow, *Social Choice and Individual Values* (Wiley, New York, 1951)
10. K.J. Arrow, Discounting, morality and gaming, in *Discounting and Intergenerational Equity*, ed. by P. Portney, J. Weyant (RFF Press, Washington, 1999)
11. K.J. Arrow, M.L. Cropper, G.C. Eads, R.W. Hahn, L.B. Lave, R.G. Noll, P.R. Portney, M. Russell, R. Schmalensee, V.K. Smith, R.N. Stavins, Is there a role for benefit-cost analysis in environmental, health and safety regulation? Science **April**, 222–223 (1996)
12. S.E. Atkinson, R. Halvorsen, The valuation of risk to life: evidence from the market for automobiles. Rev. Econ. Stat. **72**, 133–135 (1990)
13. E. Budtz-Jorgensen, N. Keiding, P. Grandjean, P. Weihe, Estimation of health effects of prenatal methylmercury exposure using structural equations model. Environ. Health **1**, 145–168 (2002)
14. J. Caswell, Quality assurance, information tracking and consumer labeling. Mar. Pollut. Bull. **53**, 650–656 (2006)
15. J.M. Conrad, *Resource Economics* (Cambridge University Press, Cambridge, 2000)
16. P.J. Cook, D.A. Graham, The demand for insurance and protection: the case of irreplaceable commodities. Q. J. Econ. **91**(1), 143–156 (1977)
17. M.L. Cropper, W.N. Evans, S.J. Beraradi, M.M. Ducla-Soares, P.R. Portney, The determinants of pesticide regulation. A statistical analysis of EPA decision making. J. Polit. Econ. **100**(1), 175–197 (1992)
18. EFSA (European Food Safety Authority), Opinion of the scientific panel on contaminants in the food chain on a request from the commission related to mercury and methyl mercury in food. EFSA J. **34**, 1–14 (2004). Available at http://www.efsa.eu.int. Accessed February 2006
19. EPA (Environmental Protection Agency). Six common pollutants. Particulate matter. Health (2012). Available at http://www.epa.gov/pm/health.html. Accessed September 15, 2012
20. C. Gollier, Discounting with fat-tailed economic growth. J. Risk Uncertain. **37**, 171–186 (2008)
21. C. Gollier, M.L. Weitzman, How should the distant future be discounted when discount rates are uncertain? Econ. Lett. **145**, 812–829 (2010)
22. P.-O. Johansson, *An Introduction to Modern Welfare Economics* (Cambridge University Press, Melbourne, 1991)
23. T. Kjaer, D. Gyrd-Hansen, Preference heterogeneity and choice of cardiac rehabilitation program: results from a discrete choice experiment. Health Policy **85**, 124–132 (2008)
24. K. Lancaster, *Consumer Demand: A New Approach* (Columbia University Press, New York, 1971)

25. J.L. Lusk, J. Roosen, J.A. Fox, Demand for beef from cattle administered growth hormones or fed genetically modified corn: a comparison of consumers in France, Germany, the United Kingdom, and the United States. Am. J. Agric. Econ. **85**, 16–29 (2003)
26. S. Marette, J. Roosen, S. Blanchemanche, The combination of lab and field experiments for benefit-cost analysis. J. Benefit-Cost Anal. **2**(3), 2 (2011)
27. D.C. Mueller, *Public Choice II* (Cambridge University Press, Cambridge, 1989)
28. Nobel Prize Committee, Committee of the Sveriges Riksbank prize in economic sciences in memory of Alfred Nobel (1972). Award ceremony speech delivered by professor Ragnar Bentzel. http://www.nobelprize.org/nobel_prizes/economics/laureates/1972/presentation-speech.html. Accessed September 15, 2012
29. OECD, *Education at a Glance 2011: OECD Indicators* (OECD, Paris, 2011). http://dx.doi.org/10.1787/eag-2011-en. Accessed October 22, 2011
30. J. Persky, Cost-benefit analysis and the classical creed. J. Econ. Perspect. **15**(4), 199–208 (2001)
31. D. Pimentel, C. Harvey, P. Resosudarmo, K. Sinclair, D. Kurz, M. McNair, S. Crist, L. Shpritz, L. Fitton, R. Saffouri, R. Blair, Environmental and economic costs of soil erosion and conservation benefits. Science **24**, 1117–1123 (1995)
32. R.A. Pollack, Imagined risks and cost-benefit analysis. Am. Econ. Rev. **88**, 376–380 (1998)
33. R.S. Raucher, S.J. Rubin, D. Crawford-Brown, M.M. Lawson, Benefit-cost analysis for drinking water standards: efficiency, equity, and affordability considerations in small communities. J. Benefit-Cost Anal. **2**(1) (2011)
34. A. Renda, *Impact Assessment in the EU: The State of the Art and the Art of the State* (Center for European Policy Studies, Brussels, 2006)
35. S. Rosen, Hedonic prices and implicit markets: product differentiation in pure competition. J. Polit. Econ. **81**, 34–55 (1974)
36. F. Salanie, N. Treich, Regulation in Happyville. Econ. J. **119**, 665–679 (2009)
37. A. Sen, The discipline of cost-benefit analysis, in *Cost-Benefit Analysis*, ed. by M.D. Adler, E.A. Posner (The University of Chicago Press, Chicago, 2001)
38. J.P. Shimshack, M.B. Ward, T.K.M. Beatty, Mercury advisories: information, education, and fish consumption. J. Environ. Econ. Manag. **53**, 158–179 (2007)
39. N. Stern, *Stern Review: The Economics of Climate Change* (Cambridge University Press, Cambridge, 2007)
40. D. Straub, I. Welpe, Decision-making under risk: a normative and behavioral perspective, in *Risk – A Multidisciplinary Introduction*, ed. by C. Klüppelberg, D. Straub, I. Welpe (2014)
41. C.R. Sunstein, The arithmetic of arsenic. Working Paper 01-10, AEI-Brookings Joint Center for Regulatory Studies, August (2001)
42. L. Thümer, U. Protzer, V. Seifert-Klauss, Risk reduction of cervical cancer through HPV screening and vaccination—assumptions and reality, in *Risk – A Multidisciplinary Introduction*, ed. by C. Klüppelberg, D. Straub, I. Welpe (2014)
43. H.R. Varian, *Intermediate Microeconomics. A Modern Approach* (Norton, New York, 2003)
44. W.K. Viscusi, J.E. Aldy, The value of a statistical life: a critical review of market estimates throughout the world. J. Risk Uncertain. **27**(1), 5–76 (2003)
45. M.L. Weitzman, Subjective expectations and asset-return puzzle. Am. Econ. Rev. **97**, 1102–1130 (2007)
46. WHO (2011) 11th meeting of the European technical advisory group of EXperts on immunization (ETAGE). Copenhagen, DM. Accessible at http://www.euro.who.int/__data/assets/pdf_file/0014/145400/e95120.pdf Accessed October 23, 2011
47. K. Zachmann, Risk in historical perspective: concepts, contexts, and conjunctions, in *Risk – A Multidisciplinary Introduction*, ed. by C. Klüppelberg, D. Straub, I. Welpe (2014)

Chapter 12
Engineering Risk Assessment

Daniel Straub

Engineers must make decisions or advise decisions makers in problems involving uncertainty and risk. Engineering risk assessments support engineers and scientists in this task, by providing a structured approach to understanding and modeling the risks. Such risk assessments are based on a quantitative engineering modeling approach, which differs from the actuarial approach to risk modeling. Because of limited data, engineers must utilize all available information from multiple sources, including physical and logical models, observed data and expert knowledge. This information is uncertain and often contradicting. The methods presented in this chapter help engineers to consistently combine this information to come up with best estimates of risk and optimal decision support. They also help the engineer in understanding the limitations and sensitivity of risk estimates and facilitate the communication and comparison of risks. Finally, they enable the definition of clear criteria for assessing the acceptability and optimality of engineering solutions to reducing risk.

Keywords Reliability · Decision making · Risk acceptance · Structural reliability · Bayesian networks

Mathematics Subject Classification (2010) 60K10 · 62C10 · 62P30 · 90B25

The Facts

- Risk assessment is a formalized way of identifying feasible and optimal actions in situations involving uncertainty and risk.
- Risk assessment includes the identification of risks, the analysis of risks and the assessment of optimality and acceptability of risks.

D. Straub (✉)
Engineering Risk Analysis Group, Faculty of Civil, Geo and Environmental Engineering, Technische Universität München, Theresienstr. 90, 80333 Munich, Germany
e-mail: straub@tum.de

C. Klüppelberg et al. (eds.), *Risk – A Multidisciplinary Introduction*,
DOI 10.1007/978-3-319-04486-6_12,

- In engineering, risk is often associated with systems for which no or limited data is available. An actuarial approach to risk analysis (based purely on failure/damage statistics) is thus not feasible and alternative methods are needed.
- Engineering risk analysis combines physical, chemical and other models with probabilistic models of uncertainty, which are derived from both data and expert knowledge.
- Because data is limited, inclusion of in-service observations is an important part of engineering risk analysis (e.g. using Bayesian methods).

1 Introduction

The single most important responsibility of the engineer is to make decisions or to provide advice on decision making related to technology and the environment. Examples of decisions that engineers are concerned with include:

- the selection of the height of a concrete slab in a building;
- the choice of a traffic regime at a road intersection;
- the choice of soil remediation measures at the site of an industrial facility;
- the selection of the structural system and material for a skyscraper;
- the choice of an inspection and monitoring regime for an aircraft;
- the decision on the location of a new railway line;
- the choice of a site and a concept for a nuclear waste deposit.

The above examples range from seemingly minor decisions to decisions that have a major impact on a society. In all these decision problems, the engineer aims at identifying the decision alternative that is the optimal one in accordance with a set of objectives, such as cost minimization and minimization of environmental impact. Ideally, the engineer can define an objective function and all the variables entering the objective function are known with certainty. In this case, the identification of the optimal decision becomes a trivial matter. An example of such a decision problem is the design of a column that should lead to the minimum cost under the condition that it complies with the relevant codes.

In most real situations, however, the engineer must consider different, often contradictory, objectives, and she must make the decisions under conditions of uncertainty. For example, in the case of the column design, the minimization of the cost might not be the only objective, but additionally a minimization of the environmental impact might be desirable. Furthermore, the column might be subject to blast loads that are not specified by the code and which are highly uncertain. In general, the larger the impact of the decision, the more it will be required to address conflicting objectives and uncertainty. In order to make rational choices under such circumstances and to be able to justify and communicate these choices, the engineer needs to be able to formalize the problem, in a similar way as she formalizes a structural design using the rules of mechanics. This is the aim of engineering risk analysis and assessment.

Risk assessment can be seen as a special case of general decision analysis that involves uncertain, adverse consequences. Even though it is possible to merely compute the risks without considering any decisions, it is important to realize that an effective risk assessment can only be carried out in the context of the decisions to be taken (by the engineer, her client and society). The formulation of the scope of a risk analysis strongly depends on the potential decisions to be made by the decision maker. As an example, an earthquake risk analysis for a building will be different if the client is an insurance company merely interested in setting a premium, in which case it might be sufficient to determine the expected value of the annual loss in economic terms with limited accuracy, or if the client is an owner interested in a safe home, in which case it might be desired that the analysis determines the expected loss of life and property damage for different alternative seismic retrofitting options.

2 Definition of Risk

Risk arises whenever there is uncertainty on potentially adverse system outcomes, such as the failure of a structural system, the contamination of ecological systems, traffic accidents, monetary losses. The risk associated with an event increases with increasing probability of the event and/or increasing consequences. This is intuitively understood.

Here, the following mathematical definition of risk is used:

$$\text{Risk} = \text{Expected adverse consequences.}$$

The term "expected" refers to the mathematical concept of the expected value. For the case of a single adverse event E, e.g. the event of a car crash, the risk $R(E)$ is computed as the product of the probability of the event $\Pr(E)$ with the consequences of the event $c(E)$:

$$R(E) = \Pr(E) \cdot c(E). \tag{1}$$

In most risk assessments, more than one possible adverse event (scenario) needs to be considered. The total expected risk is then computed by integration or summation over all possible scenarios and risk contributions. As an example, consider the risk due to flooding in an area A. The flood hazard is commonly described by the annual maximum discharge Q in the relevant river. Let $f_Q(q)$ be the probability density function (PDF) of Q, and let $c(q,x)$ be the economic consequences of a flood with discharge q at location x. The total economic risk in the area is then calculated by integrating over all possible values of Q (the scenarios) and by integrating over the total area A:

$$R = \int_{x \in A} \int_0^{\infty} c(q,x) f_Q(q) \mathrm{d}q \mathrm{d}x. \tag{2}$$

As obvious from these definitions, risk is expressed in the same dimension as the consequences, e.g., monetary values, number of fatalities, amount of toxic material. In many instances, it will be necessary or preferable to convert these consequences into an abstract utility value, to allow for a more consistent expression of the decision maker’s preferences under uncertainty (see Chap. 3, [47] for an introduction to utility theory).

In engineering applications, probability (as in Eqs. (1) and (2)) is generally a subjective value, following the Bayesian interpretation of probability. In some instances, the terms likelihood or belief are used instead of probability, but to avoid confusion we will always use the term probability here. The reason for the subjective interpretation is that in real engineering applications the conditions for the frequentist (sometimes falsely termed “objective”) interpretation of probability are not met. However, since decisions *must* be made, the engineer has no alternative to using her best estimate of the probabilities of events, which of course should be based on all available data and information. For this reason, Bayesian methods, which enable the combination of information from different sources, have a central role in engineering risk analysis.

3 Risk Assessment Procedure

A risk assessment is a formalized approach to determining and assessing the risk. When combined with the planning of actions, it is denoted risk-based decision making (or risk management). A procedure for risk analysis and management is illustrated in Fig. 1, adapted from Stewart and Melchers [4], and briefly outlined in the following.

Any risk assessment should commence with a definition of the context in which the analysis takes places. The risk analyst should state who the decision makers and the involved stakeholders are (client, society, governmental organizations, individuals) and it should be identified what their objectives and preferences are. Constraints and potentially influencing factors, including legal, financial, political, cultural and organizational aspects, should be determined. On this basis, the goals and the constraints of the risk analysis should be clearly stated. In particular, the criteria against which the risk is to be assessed must be defined at this stage (see Sect. 6 for examples) and agreed upon with the client.

In a next step, the investigated system must be clearly defined, as it is commonly done in a proper engineering analysis. The system is defined in terms of its physical extension (e.g. the area included in an environmental risk assessment), in terms of the potential hazards (which types of hazard are not included?) and in terms of its societal dimensions (e.g. the types of consequences that are to be considered). This definition should be established in collaboration with, and must be approved by, the client.

In a third step, a hazard scenario analysis is performed, aimed at identifying all relevant scenarios contributing to the risk. This includes an initial assessment of the risks associated with the scenarios. This crucial part of the analysis, which provides the basis for all the later analysis, is presented in more details in Sect. 4.

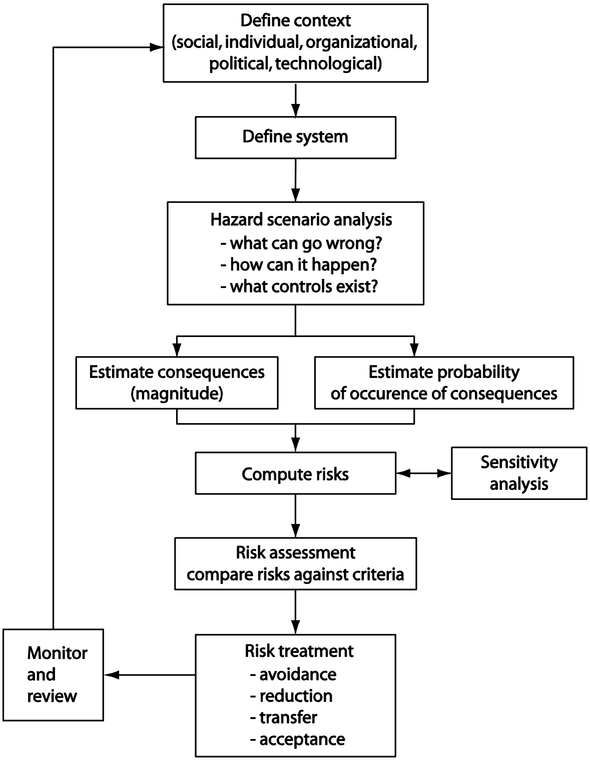


Fig. 1 Risk-based decision making/risk management procedure (adapted from Stewart and Melchers [4])

The hazard scenario analysis is followed by the quantitative risk analysis, which consists of estimating the probability of the identified adverse events as well as their consequences, by means of a variety of probabilistic modeling and analysis tools, which will be outlined in Sect. 5. These computations must generally be based on a number of assumptions. For this reason, it is essential that the computed risks are subject to a sensitivity analysis, in order to understand the influence of the assumptions on the final results. This may be followed by further analysis of crucial assumptions and a re-evaluation of the risks.

Finally, the risks are assessed, i.e. they are compared against the previously defined risk acceptance criteria (outlined in Sect. 6). At this stage, the results of the analysis are presented to the decision makers and, in some instances, to the stakeholders. On the basis of the risk assessment, strategies for treating the non-acceptable risks must be identified. Four different strategies are distinguished in Fig. 1:

- *Avoidance*: The system, or parts of the system, is no longer operated, thus reducing the associated risk to zero. In many instances, this is not an option.
- *Reduction*: The risks are reduced by introducing appropriate mitigation measures, which reduce the probability of events or their consequences. Examples

include modifications of the system itself, controlling the system through monitoring/inspection and early warning/evacuation procures.
- *Transfer*: Financial risks can be transferred through insurance or related financial instruments.
- *Acceptance*: In some instances, risks that do not comply with risk acceptance criteria must be accepted. Such acceptance should always be a temporary solution until other measures are adopted.

Following the implementation of the measures, it is required to monitor the efficiency of the measures and to review the risks after their implementation. If necessary, adjustments to the risk treatment strategy must be made.

Most elements of the risk assessment are changing with time, and ideally the risk analysis is set up in a dynamic manner, i.e. it is revised at regular intervals. Thereby, it is of importance that all the assumptions and computations made in the assessment are well documented, and that all information, including data, is well organized. This will highly facilitate an update of the risk assessment at future times, since a major portion of the budget for risk assessments is typically allocated to the collection and organization of data and information.

A set of application examples of engineering risk analyses can be found in Stewart and Melchers [4].

4 Hazard Scenario Analysis

A central part of any risk analysis is the hazard scenario analysis. In this phase of the analysis, all potential hazards and scenarios leading to damages must be identified, and suitable strategies to reach this goal must be implemented. These can vary strongly depending on the type of system and risks considered, on whether or not similar risk assessments were previously performed and on whether or not standardized procedures for the risk assessment of the considered system exist. An example of such a standardized procedure is the Probabilistic Safety Assessment methodology (PSA) developed for nuclear power plants (e.g. Beckjord et al. [11]; Apostolakis [7]).

4.1 Risk Screening

A key element of the hazard scenario analysis is a procedure for collecting the knowledge of relevant experts, which is typically achieved by organizing a meeting that includes engineers with relevant system-specific knowhow, personnel with field experience and risk analysts. Such meetings, which are sometimes termed risk screening meetings, can be understood as an organized brain-storming. In a first round, the participants are asked to envision everything that could possibly go wrong, however unlikely the scenario. It is important that the organizers of the meeting (the moderators) ensure that no scenarios are discarded at this point. In particular experienced practitioners tend to make arguments such as “this has never happened before”, and the moderators must make sure that no participant is discouraged by such comments. At this point in the process, even the highly unlikely scenarios can be of relevance. Clearly, such a meeting must be well structured and the moderators must be well prepared with background knowledge and all potentially relevant information (e.g. plans, maps, photographs, etc.).

4.2 Qualitative and Semi-quantitative Assessment of Risks

In conjunction with the risk screening, a first semi-quantitative estimation of the probability and the consequences of scenarios is made. To this end, it is common to define so-called risk matrixes, as illustrated in Fig. 2. The colors indicate the risk category. Since risk is the product of probability and consequence (Eq. (1)), the diagonals correspond to equi-risk lines if consequences and probability are plotted in log-log-scale, as is commonly done.

Here, the probability (or frequency) of events is grouped into classes (e.g., >0.1, 0.1–10^{-2}, 10^{-2}–10^{-3}, 10^{-3}–10^{-4}, $<10^{-4}$), as is the consequences of events. Often, separate risk matrixes are defined for different consequence categories (fatalities, financial consequences, ecological consequences). It is noted that many industrial companies and government agencies have such risk matrixes, but these are confidential in most cases, due to legal concerns.

To each scenario, as identified in the risk screening, is assigned a probability and a consequence class (or several consequence classes, one for each category). A useful strategy to facilitate this assignment is to illustrate each consequence class by some example scenarios. This is particularly relevant when the assignments are made by experts with limited experience in estimating probabilities.

At the end of the hazard scenario analysis, it must be determined, which of the scenarios are to be further studied in the detailed analysis. This is achieved by considering all identified scenarios and excluding those that are considered to be of acceptable risk (e.g. those that fall into the green area in the matrix of Fig. 2). In this process it is important that all the assumptions made are well documented. Furthermore, when deciding which risks to accept, the limited accuracy of the initial hazard scenario analysis must be accounted for; i.e., only those risks that cannot become relevant even with a more detailed analysis can be excluded.

In this context, often the so-called ALARP principle is invoked, which stands for “As low as reasonably possible”. It is common practice to divide the risk matrix into three regions: a region of acceptable risk, a region of inacceptable risk and in-between is the ALARP region, as shown in Fig. 2. All risks that are in the ALARP region should be reduced to a level “as low as reasonably possible”. This signifies that for all risk scenarios falling into this region, the risks should be optimized, typically through a cost-benefit analysis. This is further discussed in Sect. 6.

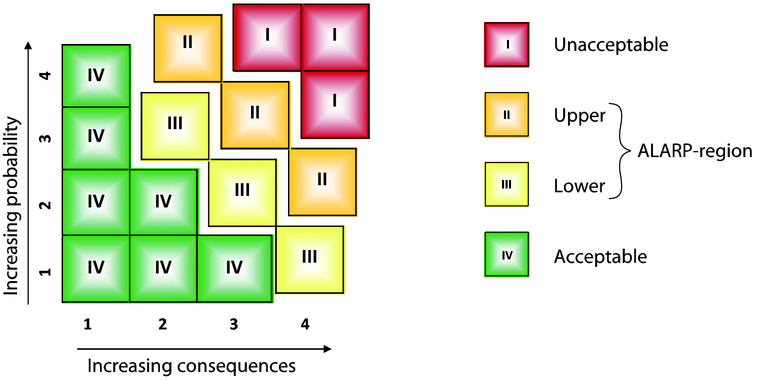


Fig. 2 Risk matrix

4.3 Logic Tree Analysis

As part of the initial assessment, as well as in detailed quantitative risk assessments, logic trees are often used for system representation. These are typically binary system representations; the most well-known are *fault trees* and *event trees*. Fault trees establish the relation between component failures and system failure events (the latter are called top-events). Event trees establish the consequences of system failures (top events) by laying out all possible event sequences following the system failure. These tools, their applications and their limitations are described in Chap. 13, [50].

5 Quantitative Risk Assessment

Quantitative risk assessment should be based on probabilistic methods. However, in risk analysis of anthropogenic systems, typically not enough data is available to determine a useful failure statistics. The reasons are

(a) that the number of such systems is often limited and failure rates are low;
(b) that the systems are subject to unique design, loading and operation conditions;
(c) that the systems are subject to common factors, introducing strong dependence among observations.

As an example, the rate of fatal accidents of European and US commercial airlines is in the order of 10^{-7}–10^{-8} per hour of flight (NTSB [33]), and accidents can reasonably well be modeled by a Poisson process. However, the failure rate varies depending on aircraft age, operator, and various other factors. Assume that our aim is to determine the probability of failure for a specific aircraft of Lufthansa during the next flight hour and compare it with the acceptable value of 10^{-8}. Within the Lufthansa fleet, in the past 10 years, no fatal accident of an aircraft in service occurred, and in the past 20 years, one fatal accident occurred. Even if all the other specific factors of this aircraft were neglected, any statistical estimate of the failure rate is highly uncertain. More importantly, the estimate of the failure rate will not provide us with useful information on how to reduce probability of failure, since the influence of the various factors that can be modified (inspection/maintenance procedures, flight operation procedures, aircraft design) is not and cannot be quantified using statistical methods alone. It is therefore necessary to combine statistical data with engineering models of the process. This is a central part of quantitative methods in engineering risk analysis.

The following sections outline a number of techniques available in engineering risk analysis for combining engineering models with stochastic models and data, all of which aim at providing the most accurate prediction of the probability of adverse events with the given information.

5.1 Statistics

Despite the fact that there is typically not sufficient data available, statistics remains an essential tool in engineering risk analysis. In particular, probabilistic models of input parameters must be determined; as an example, the statistics of rainfall precipitation are a required input to a flood risk analysis. In this and many other examples, an estimate of the extreme behavior is essential, i.e. extreme value statistics are of importance (see Chap. 6, [23]). Special focus must be put on an accurate assessment of the uncertainties involved, since the data basis is often insufficient; an excellent example of such uncertainty is given by Coles et al. [15].

When data is limited and statistical uncertainty is relevant, Bayesian statistics enables to consistently account for this uncertainty and to include it in the assessment. An introduction to Bayesian statistics is given in Chap. 8, [17]. In addition, it is often useful to combine the data with expert opinion, which is facilitated by Bayesian statistics, whereby the prior distributions are selected following the experts. However, care is needed in order not to use the information contained in the data twice, which can happen when the experts' opinions are based on the same data that are used to determine the posterior statistics.

5.2 Probabilistic Analysis of Engineering Models

In engineering, physical, chemical or logical models of the relevant processes are typically available. These are used to make predictions of the performance of given systems. Any model can be considered as a function g that establishes a relationship between inputs $\mathbf{X}$ and outputs $\mathbf{Y}$:

$$\mathbf{Y} = g(\mathbf{X}). \tag{3}$$

In many instances (and those are the situations of interest to us), all or some of the input variables $\mathbf{X} = [X_1; X_2; \ldots]$ are random. As a result, the outcome variables

$\mathbf{Y} = [Y_1; Y_2; \ldots]$ become random as well, even if the model (the function) is known with certainty. We are thus dealing with functions of random variables.

In addition, the function g itself can also be random, i.e. for a given $\mathbf{X}$, the vector of outcome variables $\mathbf{Y} = [Y_1; Y_2; \ldots]$ is random. This situation commonly occurs in the analysis of problems involving stochastic processes, e.g. in the analysis of dynamic systems with random excitation (e.g. cars, aircraft, structures under wind or earthquake excitation). Introductions to the analysis of such systems can be found in Lutes and Sarkani [31]. Here, we will restrict ourselves to problems in which g is a deterministic function. This does not imply that we assume the model to be perfect: model uncertainty can be included through additional random variables in $\mathbf{X}$.

Ideally, we compute the full probability distribution of $\mathbf{Y}$ exactly. However, this is only possible in few cases, as discussed below. In some instances, it is sufficient to compute moments of the distribution of $\mathbf{Y}$ instead of the full distribution, which significantly simplifies the problem. For most problems, however, it will be necessary to use approximation methods. These include Monte Carlo Simulation (MCS) and the class of Structural Reliability Methods (SRM), which also include advanced sampling techniques such as adaptive importance sampling and subset simulation. These SRM are presented in Sect. 5.3.

It should be noted that applied physical models are often numerical, e.g. Finite Element (FE) models. This implies that no analytical solution for $\mathbf{Y} = g(\mathbf{X})$ exists, and that obtaining values of $\mathbf{Y}$ can be costly (in terms of computation time). This has implications on the applicable methods for evaluating the characteristics of $\mathbf{Y}$.

Illustration 5.1 (Fatigue Model) For illustrational purposes, we consider the Palmgren-Miner model for material fatigue, which occurs in dynamically loaded structures such as aircraft, trains, cars, bridges and buildings. One of the tragic failures caused by material fatigue was the accident of the ICE train at Eschede, Germany in 1998, causing 101 fatalities. Fatigue damage can be measured in terms of a normalized damage D, which in the simplest form of the model is computed as

$$D = n\frac{1}{C}S^m. \tag{4}$$

Here, C and m are material parameters, S are the stress ranges due to constant cyclic loading and n are the number of stress cycles. Failure occurs when the damage exceeds 1, i.e. when $D \geq 1$.

We consider the case where C and S are random variables, i.e. we have $\mathbf{X} = [C; S]$ and $\mathbf{Y} = [D]$. We will use this model to illustrate the different concepts and solution strategies below.

The simplest class of models is the one of linear models, which can be generically written as

$$\mathbf{Y} = g(\mathbf{X}) = \mathbf{a}_0 + \mathbf{a}\mathbf{X}, \tag{5}$$

where $\mathbf{X}$ is a vector of length n_X, $\mathbf{Y}$ and $\mathbf{a}_0$ are vectors of length n_Y, and $\mathbf{a}$ is a $n_Y \times n_X$ matrix of coefficients. As is well known, for linear models, the mean and covariance of $\mathbf{Y}$ can be computed exactly (Papoulis and Pillai [34]). In the special case that the random variables $\mathbf{X}$ are multinormal (Gaussian) distributed, the random variables $\mathbf{Y}$ also have a multinormal distribution. This explains the popularity of linear Gaussian models: for these models, the full distribution of $\mathbf{Y}$ is readily obtained, since it is fully described by its mean and covariance.

It is noted that many non-linear models can be transformed into linear models, as illustrated in the following.

Illustration 5.2 (Fatigue Model) The non-linear model for material fatigue of Eq. (4) can be transformed into a linear model of C and S by taking the logarithm:

$$\ln D = \ln n - \ln C + m \ln S. \tag{6}$$

It follows that the mean of the logarithm of the fatigue damage is

$$\mathrm{E}[\ln D] = \ln n - \mathrm{E}[\ln C] + m\mathrm{E}[\ln S] \tag{7}$$

and its variance is

$$\mathrm{Var}[\ln D] = \mathrm{Var}[\ln C] + m^2\,\mathrm{Var}[\ln S]. \tag{8}$$

If C and S are lognormal distributed, then $\ln D$ is normal distributed and the probability of failure, $\Pr(D \geq 1) = \Pr(\ln D \geq 0)$ can be computed analytically.

In the case of non-linear engineering models, a common strategy in probabilistic analysis is to approximate the models by a linear or quadratic model, so-called first- and second-order approximations (e.g. Papoulis and Pillai [34], Straub [5]). Rarely, higher order approximations are also chosen. However, in risk analysis, it is commonly the extreme events that are of interest. In this case, the approximation of the function $g(\mathbf{X})$ around the expected value $\mathbf{M}_\mathbf{X}$ is generally not suitable. An alternative is to approximate $g(\mathbf{X})$ in the tail of the distribution, corresponding to the region of interest. Such an approach is pursued by structural reliability methods introduced in Sect. 5.3 below.

In theory, it is also possible to compute the exact distribution of $\mathbf{Y} = g(\mathbf{X})$. As is well known, when Y is a scalar one-to-one function of a single random variable X, then the distribution of $Y = g(X)$ is readily obtained as

$$f_Y(y) = f_X\left[g^{-1}(y)\right]\left|\frac{\mathrm{d}g^{-1}(y)}{\mathrm{d}y}\right|, \tag{9}$$

where g^{-1} is the inverse function of g. Solutions for general functions of one or more random variables are described in Papoulis and Pillai [34]. However, for most realistic models of engineering systems with several random variables, these solutions are not practical and approximate methods, such as the Monte Carlo simulation, are necessary.

5.2.1 Monte Carlo Approximation

With the availability of computers, a simple, intuitive and often effective approach to analyzing functions of random variables is Monte Carlo Simulation (MCS). It

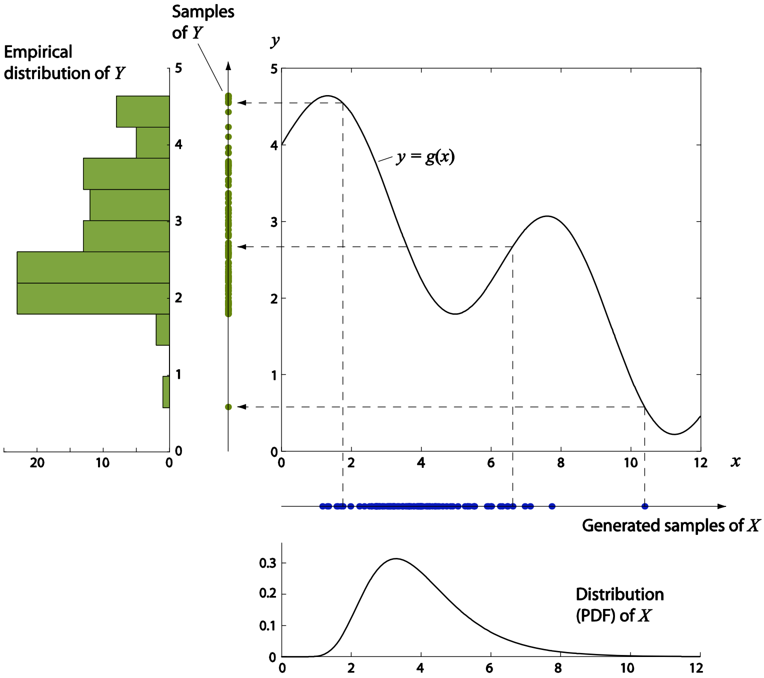


Fig. 3 Illustration of the Monte Carlo simulation approach to evaluating functions of random variables (from Straub [5])

proceeds by artificially generating samples $\mathbf{x}_i$, $i = 1, \ldots, n_s$ from the distribution of the input variables $\mathbf{X}$ and then evaluating the functions $\mathbf{y}_i = g(\mathbf{x}_i)$ for each sample value $\mathbf{x}_i$ separately. In this way, a set of samples $\mathbf{y}_i$, $i = 1, \ldots, n_s$ of the function values $\mathbf{Y}$ are generated, which provide an empirical estimate of the distribution of $\mathbf{Y}$. The principle of the MCS method is illustrated in Fig. 3 for the case of a scalar input variable X and a scalar output variable Y.

The MCS method is particularly useful when the function $\mathbf{Y} = g(\mathbf{X})$ must be evaluated numerically and when it is difficult or impossible to obtain the inverse function $g^{-1}(\mathbf{Y})$. In MCS, evaluation of the inverse function is not required.

A main advantage of MCS is its simplicity. For a given function $g(\mathbf{X})$, it consists of only three steps, which are readily performed with a few lines of computer code (in addition to the code required for evaluating $g(\mathbf{X})$). These are:

1. Generation of (pseudo-)random samples $\mathbf{x}_i$, $i = 1, \ldots, n_s$, of the input variables $\mathbf{X}$.
2. n_s evaluations of the function to $\mathbf{y}_i = g(\mathbf{x}_i)$.
3. Analysis of the generated samples $\mathbf{y}_i$ of $\mathbf{Y}$.

A more detailed introduction can be found e.g. in Rubinstein and Kroese [40] or Straub [6]. Here we only note that MCS is inefficient when computing the probability of rare events. When applying MCS with n_s samples to calculate the probability

p_F of an event F, the coefficient of variation of the MCS estimation error is approximately $(\sqrt{n_s p_F})^{-1}$. As an example, to compute a probability $p_F = 10^{-6}$, we need $n_s = 25 \times 10^6$ samples to achieve an accuracy of 20 %.

A variant of MCS, which is often more efficient, is importance sampling (IS), as described e.g. in Engelund and Rackwitz [21]. Instead of sampling randomly from the distribution of **X**, IS allows to concentrate samples of **X** in the region of interest. In risk analysis, this region typically corresponds to the values of **X** for which failure of the system occurs. The identification of this region is a non-trivial matter, which, however, is facilitated by structural reliability methods outlined in the next section.

5.3 Structural Reliability Methods

In risk analysis, we are mostly concerned with failure events that have small probabilities and for which the MCS approach is not efficient. For this reason, a class of methods called Structural Reliability Methods (SRM) have been developed since the 1970s (e.g. Rackwitz and Fiessler [39]; Der Kiureghian and Liu [18]). The following provides a brief outline of SRM, detailed introductions can be found e.g. in Ditlevsen and Madsen [20], Melchers [32] or Straub [6].

In SRM, the event of interest is described in terms of a so-called limit state function $g(\mathbf{X})$, where $\mathbf{X} = [X_1; X_2; \ldots; X_n]$ is the vector of random variables of the problem (the uncertain model input). By definition, the (failure) event F corresponds to

$$F = \big\{g(\mathbf{X}) \leq 0\big\}. \tag{10}$$

In this formulation, $\{g(\mathbf{X}) \leq 0\} = \Omega_F$ corresponds to a domain in the outcome space of **X**, whose surface is described by $\{g(\mathbf{X}) = 0\}$. The probability of the event F is thus identical to the probability of **X** taking a value within this domain. It can be computed by integrating the joint probability density function of **X**, denoted by $f(\mathbf{x})$, over Ω_F:

$$\Pr(F) = \int_{g(\mathbf{x}) \leq 0} f(\mathbf{x}) \mathrm{d}x_1 \mathrm{d}x_2 \cdots \mathrm{d}x_n. \tag{11}$$

The problem is illustrated in Fig. 4. For the case of two random variables, as in Fig. 4, numerical integration is straightforward, e.g. using quadrature rules. However, most methods for numerical integration have computation times that increase exponentially with the number of dimensions (one exception being MCS). Therefore, they are not suitable to solve the integral in Eq. (11) when the number of random variables is larger than 3 to 5.

All structural reliability methods aim at solving Eq. (11). All of these methods are approximations, and each method has its own advantages and disadvantages. Here, only the first-order reliability method (FORM) is briefly introduced, followed by a short outline of other methods.

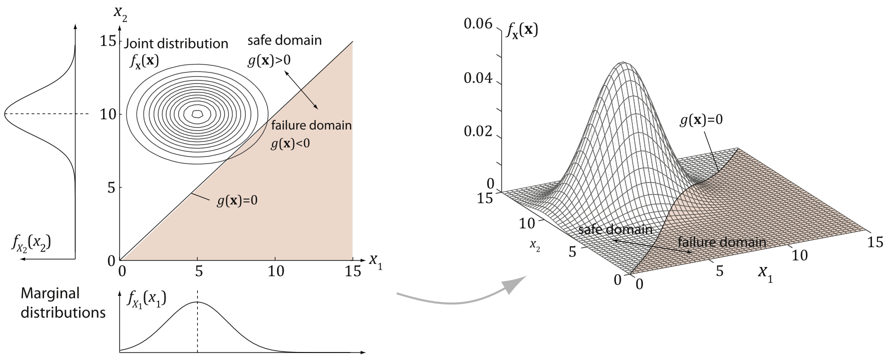


Fig. 4 Illustration of the general reliability problem, for the case of two random variables; *left*: contour plot of the joint PDF, *right*: 3D plot of the same joint PDF

Illustration 5.3 (Fatigue Failure) With the fatigue model introduced in Eq. (4), fatigue failure is modeled as the event of the damage D reaching or exceeding 1, i.e. $F = \{1 - D \leq 0\}$. It follows that the fatigue failure can be described by the following limit state function:

$$g(C, \mathit{\Delta} S) = 1 - n\frac{1}{C}S^m. \tag{12}$$

5.3.1 First-Order Reliability Method (FORM)

The method starts by transforming the problem from the original space of the random variables $\mathbf{X}$ to the space of standard normal random variables $\mathbf{U}$. If the joint distribution of $\mathbf{X}$ is of the Gaussian copula class, the Nataf transformation can be applied (Der Kiureghian and Liu [18]), if the joint distribution of $\mathbf{X}$ is of any arbitrary form, the Rosenblatt transformation can be used (Rackwitz and Fiessler [39]). The reader is referred to Ditlevsen and Madsen [20], Melchers [32] or Straub [6] for details. In the following, we let T denote this transformation, i.e.:

$$\mathbf{U} = \mathrm{T}(\mathbf{X}), \tag{13}$$

$$\mathbf{X} = \mathrm{T}^{-1}(\mathbf{U}). \tag{14}$$

The first basic idea of FORM is to transform the limit state function g to the space of standard normal random variables. Let G denote the new limit state function in standard normal space:

$$G(\mathbf{U}) = g\big(\mathrm{T}^{-1}(\mathbf{U})\big). \tag{15}$$

The transformation T is probability conserving, therefore we have that $\Pr(F) = \Pr(g(\mathbf{X}) \leq 0) = \Pr(G(\boldsymbol{U}) \leq 0)$. In analogy to Eq. (11), the probability of the failure event F is now computed by

$$\Pr(F) = \int_{G(\mathbf{u})\leq 0} \phi(\mathbf{u})\mathrm{d}u_1\mathrm{d}u_2\cdots\mathrm{d}u_n, \tag{16}$$

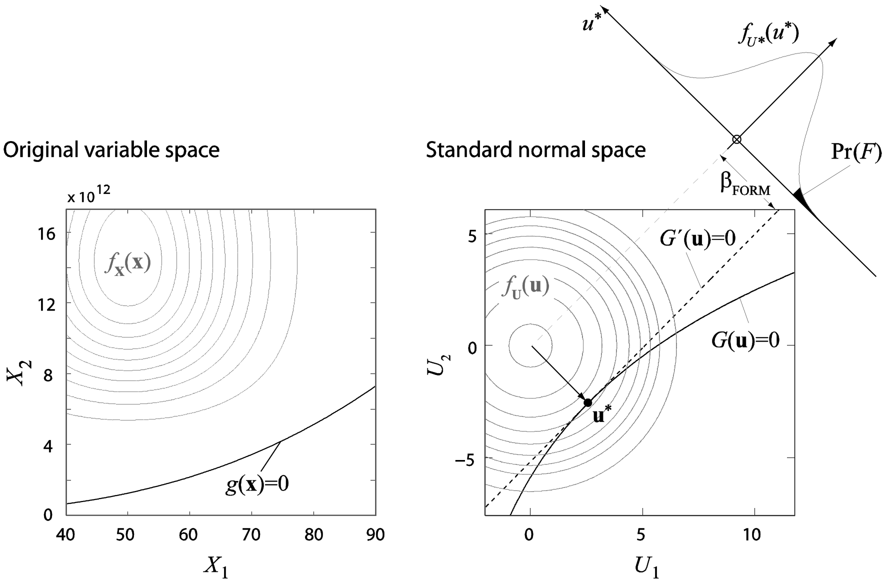


Fig. 5 Design point and linear approximation of the limit state surface. *Left side*: original random variable space; *right side*: standard normal space

where ϕ is the standard multivariate normal PDF.

The second basic idea of FORM is to approximate the limit state function $G(\mathbf{U})$ by a first-order Taylor expansion at the expansion point $\mathbf{u}^*$, denoted by $G'(\mathbf{U})$. To limit the approximation error, the expansion point is selected as the point in the failure domain with the highest probability content, the so-called Most Likely Failure Point (MLFP). Because the standard multivariate normal PDF ϕ is rotation-symmetric around the origin, the MLFP is equal to the point on the failure surface $G(\mathbf{U}) = 0$ that is the closest to the origin (provided that $\Pr(F) < 0.5$). The identification of the expansion point therefore corresponds to a constrained minimization problem:

$$\mathbf{u^*} = \arg\min \|\mathbf{u}\| \quad \text{subject to} \quad G(\mathbf{u}) = 0, \tag{17}$$

where $\|\mathbf{u}\| = \sqrt{\mathbf{u}^{\mathrm{T}}\mathbf{u}}$ is the Euclidian norm of the vector $\mathbf{u}$, which corresponds to the distance of u from the origin. The notation arg min stands for "the argument that gives the minimum value of".

Figure 5 illustrates the transformation of the limit state surface and the approximation by a hyperplane at the MLFP for the case of two random variables (in which case the hyperplane reduces to a line).

With this approximation, the limit state surface is approximated by its tangent at the design point, see Fig. 5. In FORM, the integration over the domain $\{G(\mathbf{u}) \leq 0\}$ is thus replaced by the integration over a half space defined by the tangent $\{G'(\mathbf{u}) = 0\}$.

Every marginal distribution of the standard multivariate normal distribution is a standard normal distribution. Therefore, the marginal probability distribution of $\mathbf{U}$

in the direction perpendicular to the linearized limit state surface is also a standard normal distribution, as illustrated in Fig. 5. It should be clear from the illustration that the probability of failure is fully defined by the distance $\beta_{\mathrm{FORM}} = \|\mathbf{u}^*\|$ between the origin and the MLFP as

$$\Pr(F) \approx \Pr\big(G'(\mathbf{U}) \leq 0\big) = \Phi(-\beta_{\mathrm{FORM}}). \tag{18}$$

Here, Φ is the standard normal cumulative distribution function (CDF). β_{FORM} is known as the FORM reliability index.

The FORM solution is independent of the problem dimension, i.e. the n-dimensional integration always reduces to an evaluation of the standard normal CDF. The difficulty in FORM is the identification of the MLFP, $\mathbf{u}^*$ i.e. the solution of the optimization problem of Eq. (17). Optimized algorithms exist for this purpose. Furthermore, specialized response surface methods have been developed to limit the number of calls of the function $g(\mathbf{X})$, e.g. Bucher and Bourgund [14] or Sudret [49].

FORM is surprisingly accurate for a wide range of problems, but the accuracy is obviously dependent on how strongly non-linear the limit state function is. For this reason, it is recommended to check improve the accuracy of FORM by performing an additional importance sampling, in which the sampling density is centered around the MLFP, e.g. Rackwitz [37]. Many other strategies exist, e.g. a second-order approximation (Breitung [13]) or a novel efficient simulation technique based on Markov Chain Monte Carlo (Au and Beck [8]), and the interested reader is referred to the literature provided in the bibliography.

Illustration 5.4 (Fatigue Failure) The fatigue failure is described by the limit state function in Eq. (12), $g(C, S) = 1 - nC^{-1}S^m$. We assume the following model for the parameters (all random variables are independent).

Because C and S are statistically independent, they can be transformed separately from $\mathbf{X}$ to $\mathbf{U}$-space, by requiring that $F_{Xi}(x_i) = \Phi[\mathrm{T}(x_i)]$. It follows that the inverse transformation T^{-1} from standard normal space is:

$$C = \exp(U_C \sigma_{\ln C} + \mu_{\ln C}),$$
$$S = U_S \sigma_S + \mu_S.$$

Consequently, the limit state function in standard normal space is obtained by inserting the above expressions in Eq. (12):

$$G(\mathbf{U}) = 1 - \frac{n}{\exp(U_C \sigma_{\ln C} + \mu_{\ln C})} (U_S \sigma_S + \mu_S)^m. \tag{19}$$

The original and the transformed limit state functions are those shown earlier in Fig. 5, where $X_1 = S$ and $X_2 = C$.

With the parameters of Table 1, the MLFP is found according to Eq. (17) as

$$\mathbf{u}^* = [2.59; -2.55].$$

(This can be verified graphically in Fig. 5.) The corresponding FORM reliability index is $\beta_{\mathrm{FORM}} = \|\mathbf{u}^*\| = 3.63$ and the FORM estimate of the probability of failure

Table 1 Parameters of the fatigue model

Variable	Distribution	CDF[a]	Parameters[b]
C	lognormal	$\Phi\left[(\ln c - \mu_{\ln C})/\sigma_{\ln C}\right]$	$\mu_{\ln C} = 30.5$, $\sigma_{\ln C} = 0.45$
S	normal	$\Phi\left[(s - \mu_S)/\sigma_S\right]$	$\mu_S = 50$, $\sigma_S = 12.5$
m	deterministic	–	$m = 3$
n	deterministic	–	$n = 10^7$

[a] Φ is the standard normal CDF

[b] All dimensions are corresponding to mm and N

is found as:

$$\Pr(F) \approx \Phi(-\beta_{\text{FORM}}) = 1.4 \times 10^{-4}.$$

For comparison, the exact solution found by direct numerical integration is $\Pr(F) = 1.3 \times 10^{-4}$. (By observing the shape of the linear approximation in Fig. 5, it should be clear that FORM slightly overestimates the reliability.)

5.4 System Reliability

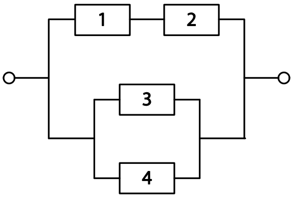
In the above sections it was assumed that the event of interest is described by a parametric function of a number of random variables $\mathbf{X}$. In many instances, however, the event of interest corresponds to a system failure event that can be described by a logical function of component failure events. As a simple example, consider the failure of an aircraft with four engines. The aircraft is still operational with one engine, and the system failure F_S can thus be expressed as the intersection of the component failures $F_i : F_S = \bigcap_{i=1}^{4} F_i$. (Such a system is known as a parallel system.)

The probability of component failure can often be determined from data, either from experimental tests or—preferably—from in-service failure data. The probability of the system failure is then determined based on the component failure probability and the logical model of the system.

If the components as well as the system are expressed by binary states (failure/survival), then the relation between component states and system state can be modeled by reliability block diagrams, an example of which is given in Fig. 6. The system fails whenever there is no path between the beginning and the end of the block diagram. It is noted that other logic trees, in particular fault trees (Chap. 13, [50]), can be converted into such reliability block diagrams.

The analysis of binary systems is commonly done by identifying the so-called minimal cut sets. A cut set is a set of components that leads to system failure (it "cuts" the diagram in two) and a minimum cut set is one in which no subset is a cut set (i.e. all components must fail to cause failure of the system). For the example of Fig. 6, there are two minimal cut sets: $\{1, 3, 4\}$ and $\{2, 3, 4\}$.

Fig. 6 Example of a reliability block diagram for a system with four components



The dual to cut sets are link sets: a link set is a set of components that ensure the system to work, and a minimum link set is one where no subset is a link set (i.e. all components are necessary for the system to function). The minimum link sets of the system in Fig. 6 are: $\{1, 2\}$, $\{3\}$ and $\{4\}$. It is pointed out that the identification of minimal link sets or cut sets is non-trivial, and can become computationally infeasible for large and complex systems.

There are two basic types of systems: the parallel system and the series system. In the parallel system, all components are set in parallel, i.e. the system fails only if all components fail: $F_S = \bigcap_{i=1}^{n} F_i$. In the series system, all components are set in series, i.e. the system fails as soon as one components fails: $F_S = \bigcup_{i=1}^{n} F_i$. For a general system, failure can be described by considering each minimal cut set as a parallel system (all components must fail for failure to occur), and the system as a series system of its minimal cut sets (the system fails as soon as one cut set fails). It follows that system failure is:

$$F_S = \bigcup_{k=1}^{n_k} \bigcap_{i \in Ck} F_i, \tag{20}$$

wherein n_k is the number of minimal cut sets and C_k is the index set describing the kth minimal cut set. For the example of Fig. 6, it is: $F_S = (F_1 \cap F_3 \cap F_4) \cup (F_2 \cap F_3 \cap F_4)$. By applying the distributive law, this can be reformulated to $F_S = (F_1 \cup F_2) \cap F_3 \cap F_4$. (Alternatively, this formulation can be obtained directly from the minimal link set formulation.)

For known cut sets, the system failure probability $\Pr(F_S)$ can be computed as a function of the individual component failure probabilities $\Pr(F_i)$, $i = 1, \ldots, n$ when component failure events are statistically independent. (As an example, if all components of the system shown in Fig. 6 are independent and have identical failure probability $\Pr(F_i) = 0.1$, then the probability of system failure is $\Pr(F_S) = 0.0019$.) This assumption of independence does not hold for most applications, and it is then necessary to know the probabilities of intersections, such as $\Pr(F_i \cap F_j)$. In this case, exact computation is only possible for small systems or when the dependence structure can be expressed in a simple form (e.g. when dependences are caused by common influencing factors). However, approximate solutions based on simulation (e.g. MCS) exist, or bounds can be computed (e.g. Song and Der Kiureghian [42]).

The computation of system reliability is a broad discipline, in particular when including also non-binary (i.e. multi-state) systems. The interested reader is referred to the monographs on system reliability by Barlow and Proschan [10] and Høyland and Rausand [26].

5.5 Bayesian Updating

Bayesian analysis is an important tool in engineering risk analysis, since it facilitates the consistent combination of information from various sources, which is crucial when the amount of data is limited. As an example, there is large uncertainty associated with tunnel construction because of random geology, but prior to and during the construction information is gathered from the site, e.g. by observing deformations or measuring groundwater flow. These allow the experienced engineer to adjust the project to minimize risks. Bayesian updating can formalize this process of assessing the risk conditional on such observations (e.g. Straub [44], Papaioannou and Straub [3]).

Bayesian updating of the probability of an event F with an observation event Z is based on the rule of Bayes:

$$\Pr(F \mid Z) = \frac{1}{\Pr(Z)} \Pr(Z \mid F) \Pr(F). \tag{21}$$

Here, $\Pr(F)$ is the a-priori probability of F (i.e. before the observation Z); $\Pr(F \mid Z)$ is the conditional a-posteriori probability of F (i.e. conditional on the observation Z); the conditional probability $\Pr(Z \mid F)$ is the so-called likelihood, which describes the information content of Z with respect to F; $\Pr(Z)$ is the a-priori probability of making the observation Z, which is obtained by normalization. Bayesian updating can be performed repetitively. Consider the case where we make two observations Z_1 and Z_2 sequentially. Firstly, the probability of F is updated with the observation Z_1 following Eq. (21). Secondly, the updated probability $\Pr(F \mid Z_1)$ becomes the new prior probability, and the conditional $\Pr(F \mid Z_1 \cap Z_2)$ is calculated from Eq. (21) where $\Pr(F)$ is replaced with $\Pr(F \mid Z_1)$.

Bayes' rule is at the heart of Bayesian statistics, as introduced in Chap. 8, [17]. The reader is referred to that chapter for details on the practical implementation of Eq. (21) in that context. There are two practical differences between the application in Bayesian statistics and in engineering risk assessment: (a) Unlike in Bayesian statistics, where the prior probability distribution is often weakly informative, in risk assessment the prior probability $\Pr(F)$ is generally informative, as it is based on the available models of the process. (b) In engineering risk assessment, the event F is often described by complex probabilistic models (often based on engineering models, as outlined earlier). Therefore, different computational approaches are required than in Bayesian statistics (e.g. the use of MCMC is often inefficient). The methods are often based on structural reliability methods, but other methods like Bayesian networks are also becoming popular. The reader is referred to Straub [43, 44] for examples of such methods.

Illustration 5.5 (Updating of Fatigue Reliability and Risk) A common strategy to reduce the risk due to fatigue failures is to perform regular inspections of the fatigue-sensitive structural details. Trains, aircrafts, turbines, bridges and many other structures undergo regular inspection, which are costly due to the inspection cost and the

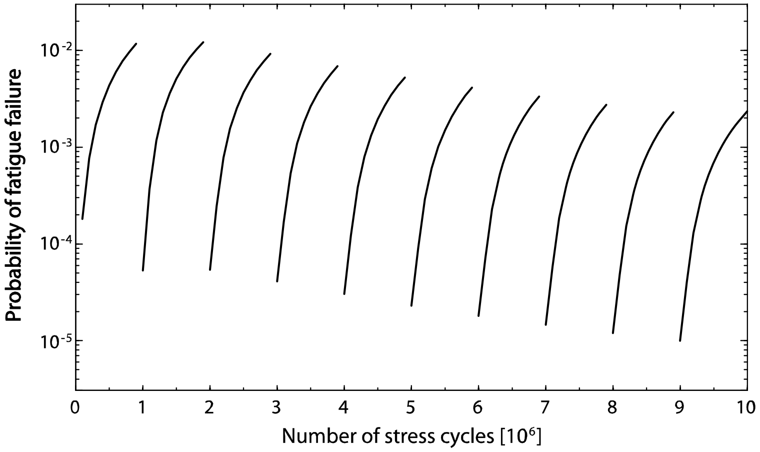


Fig. 7 Bayesian updating of the probability of fatigue failure with inspection results: inspections are performed in intervals of 10^6 stress cycles, all inspections result in no-identification of defects; taken from Straub [43]

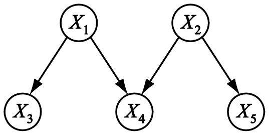
downtime of the system. (As an example, basic checks of commercial aircrafts are performed approximately every 500–800 flight hours.) For these reasons, there is a strong interest in optimizing these inspections, which requires quantifying the effect of inspections on the probability of failure (e.g. Straub and Faber [46]).

Fatigue inspections check whether or not cracks are present in the material. When defects are found, they are repaired. When no defects are found, the probability of failure is decreased, purely due to the reduction of the uncertainty. The quality of the inspection is described by so-called Probability of Detection (PoD) functions, which describe the probability of detecting a defect as a function of the defect size. To update the probability of failure, the likelihood function is constructed by combining this PoD function with physical models describing crack growth. The latter are a function of multiple random variables. In this way, Bayes' rule can be used to update the probability of failure after every inspection that results in not finding a defect. An exemplarily result is shown in Fig. 7.

5.6 Bayesian Networks

Bayesian networks (BNs), also known as Bayesian belief networks, are probabilistic models that facilitate efficient representation of the dependence structure among random variables by graphical means. BNs have been developed since the 1980s, mostly in the field of artificial intelligence, for representing probabilistic information and reasoning (Russell and Norwig [41]). They have found applications in many fields such as statistical modeling, language processing, image recognition and machine learning, and have increasingly been applied in engineering risk analysis. Recent applications in this field are reported, e.g., in Fris-Hansen [24], Faber et al. [22],

Fig. 8 A simple Bayesian network



Grêt-Regamey and Straub [25], Straub [43], Bensi et al. [12]. A general introduction to BN can be found in the textbook by Jensen and Nielsen [28].

In a nutshell, BNs model a joint probability distribution of a set of random variables $\mathbf{X} = [X_1, \ldots, X_n]$. Each random variable is represented by a node in the BN, and the links between them represent the dependence structure among the variables. If all $\mathbf{X}$ are discrete, they are fully described by their joint probability mass function (PMF), $p(\mathbf{x})$. The size of the joint outcome space of $\mathbf{X}$ for which $p(\mathbf{x})$ must be defined increases exponentially with the number of variables, but the BN enables an efficient modeling by factoring the joint probability distribution into conditional (local) distributions for each variable given its *parents*. Parents of a variable X_i are all random variables that have links pointing to X_i. A simple BN with five variables is illustrated in Fig. 8, where X_1 is a parent of X_3 and X_4, and X_2 is a parent of X_4 and X_5.

The joint PMF for this network is given as

$$p(\mathbf{x}) = p(x_1, x_2, \ldots, x_5) = p(x_1)p(x_2)p(x_3 \mid x_1)p(x_4 \mid x_1, x_2)p(x_5 \mid x_2) \qquad (22)$$

which can be written in the compact and general form

$$p(\mathbf{x}) = \prod_{i=1}^{n} p\big[x_i \mid pa(X_i)\big] \qquad (23)$$

where $pa(X_i)$ denotes the set of parents of X_i.

The decomposition of the joint PMF into the conditional PMFs of each variable given its parents, $p[x_i \mid pa(X_i)]$, is motivated by the *d-separation* rules (Pearl [36]), which describe the independence assumptions encoded in the graphical structure of the BN. However, the BN definition of the joint PMF according to Eq. (23) is quite intuitive even to the lay engineer with little understanding of the theory. To understand the efficiency of the BN representation, consider the case where each variable in the BN of Fig. 8 has 10 outcome states. To directly represent the joint PMF $p(\mathbf{x})$, it is necessary to specify 10^5 probability values (the size of the outcome space of $\mathbf{X}$). However, with the decomposition according to Eq. (22), it is sufficient to specify $10 + 10 + 10^2 + 10^3 + 10^2 = 1220$ probability values (e.g. for specifying $p(x_5 \mid x_2)$ for all combinations of X_2 and X_5, 10^2 values are required). Therefore, even for this simple example, the required information for specifying the problem is reduced by two orders of magnitude.

To efficiently compute marginal and conditional probabilities of variables in the network (the *inference* process), the conditional independence properties can also be exploited. Global computations involving $p(\mathbf{x})$ can be replaced by local computations. For the case that all random variables are discrete and/or linear combinations of Gaussian random variables, exact inference algorithms exist, but finding

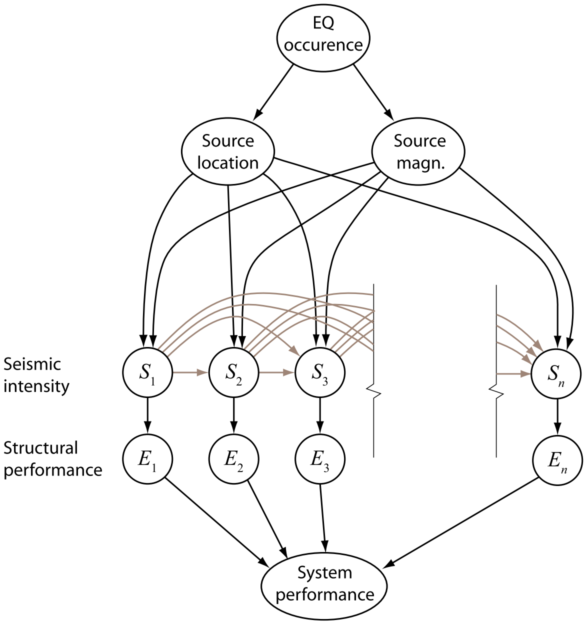


Fig. 9 BN model for seismic risk analysis of an infrastructure system (Straub et al. [48])

optimal computation strategies in a specific BN is a NP-hard task. Alternatively, sampling methods can be used to evaluate BNs. The latter can also be applied to BNs with continuous random variables. An accessible introduction to all these inference algorithms is provided by Jensen and Nielsen [28]. It is noted that a variety of software exists for constructing and evaluating BNs, many of which are available for free (e.g. the Genie software, developed at the University of Pittsburgh: http://genie.sis.pitt.edu/).

The BN has several features that make it highly useful in engineering risk analysis:

(a) Its graphical form provides a concise representation of statistical dependence that can be understood also by non-experts.
(b) The decomposition of the problem into local conditional distributions corresponds to the way complex risk analyses are performed. Combining different probabilistic models within one single BN model is often straightforward.
(c) As its name suggests, the BN is efficient for Bayesian updating when new information becomes available.

BNs are a powerful modeling framework when it is possible to exploit conditional independence among random variables. This is the case for most applications of engineering risk analysis, where the relation among random variables is often characterized by causal relations (A causes B). One example of such a dependence structure is given in Fig. 9. The dependence between the seismic intensities S_i at multiple sites i due to common earthquake source characteristics can be modeled efficiently with the BN. It also captures the assumption that the performance E_i of

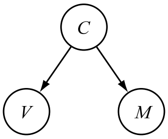


Fig. 10 The causal network for the corrosion inspection problem

infrastructure elements (bridges, pipelines, etc.) depend only on the seismic intensity at their site. However, in the example given in Fig. 9 it is also observed that, when including spatial correlation between the seismic intensity at different locations, a large number of links are needed (indicated in grey). This is one example of a dependence that is not efficiently represented by a BN. In most instances, suitable modeling strategies can avoid such types of dependences (e.g. Straub and Der Kiureghian [45]).

BN can directly be extended to decision graphs, to assess the effect of mitigation actions on the risk, and to optimize decisions following the classical decision theory (Chap. 3, [47]).

Illustration 5.6 (Corrosion Inspection) To determine the risk due to corrosion of the reinforcement in a reinforced concrete structure, a so-called "half-cell potential measurement" is performed to identify corrosion activity, together with a visual inspection of the concrete surface. Let us denote the condition of the element by C, with $\{C = 0\}$ being the event of no corrosion and $\{C = 1\}$ the event of corrosion. V is the visual inspection, with $\{V = 0\}$ the event of no visible corrosion and $\{V = 1\}$ the event of visible corrosion. M is the outcome of a half-cell potential measurement with $\{M = 0\}$ being the event of no-indication and $\{M = 1\}$ the event of indication.

It is reasonable to assume that for given condition of the element, the outcome of the measurement is independent of the visual inspection. Therefore, the causal network for this problem is shown in Fig. 10.

The conditional probability mass functions required for the specification of the network can be summarized in so-called conditional probability tables:

Event	Probability
$\{C = 0\}$	0.8
$\{C = 1\}$	0.2

Event	Probability conditional on	
	$\{C = 0\}$	$\{C = 1\}$
$\{V = 0\}$	1	0.5
$\{V = 1\}$	0	0.5

Event	Probability conditional on	
	$\{C = 0\}$	$\{C = 1\}$
$\{M = 0\}$	0.8	0.15
$\{M = 1\}$	0.2	0.85

These probability models can be obtained from deterioration models and past experience with the inspections. With these specifications, it is possible to compute the probability of corrosion conditional on different measurement/observation outcomes. It is:

Event	Probability conditional on			
	$\{V=0\}, \{M=0\}$	$\{V=0\}, \{M=1\}$	$\{V=1\}, \{M=0\}$	$\{V=1\}, \{M=1\}$
$\{C=0\}$	0.977	0.653	0	0
$\{C=1\}$	0.023	0.347	1	1

For this simple example, the computations are trivial and can easily be performed by hand. As an example, it is:

$$\begin{aligned}\Pr(C=1 \mid V=0 \cap M=1) &= \frac{\Pr(C=1 \cap V=0 \cap M=1)}{\Pr(V=0 \cap M=1)}\\ &= \frac{\Pr(C=1)\Pr(V=0 \mid C=1)\Pr(M=1 \mid C=1)}{\sum_{i=0}^{1}\Pr(C=i)\Pr(V=0 \mid C=i)\Pr(M=1 \mid C=i)}\\ &= \frac{0.2\times 0.5\times 0.85}{0.8\times 1.0\times 0.2+0.2\times 0.5\times 0.85}=0.347.\end{aligned}$$

Note that this corresponds to the application of Bayes' rule.

5.7 Sensitivity Analysis

One of the most important parts of any risk analysis is the investigation of the sensitivity of the computed risks to changes in the model parameters and assumptions. In engineering risk analysis, it is often necessary to make relatively crude assumptions on certain model parameters, due to the lack of detailed information or models. It is therefore essential that the sensitivity of the computed risks to these assumptions is quantified.

A sensitivity analysis essentially consists in re-running the risk computations for different input parameters. If the number of parameters is large and/or the risk model is computationally demanding, these re-runs must be limited to a few cases, which have to be selected using engineering judgment. Also, sensitivity measures from probabilistic calculations (e.g. using FORM or MCS) can be used (e.g. Cooke and van Noortwijk [16]), but it must be considered that these measures are local, i.e. for non-linear models they reflect only the effect of small changes in the assumptions.

It is also noted that many risk analyses are notional, which means that they do not compute the real risks, but compute the risk conditional on certain idealized assumptions. In particular, the effect of human error is often excluded from quantitative risk computations, due to the difficulty in modeling such errors. In this case, the computed value cannot be compared against absolute risk criteria, but the model it is still useful to assess the sensitivity of the risk to influencing factors and model assumptions. By means of sensitivity analyses, it is possible to pre-evaluate different mitigation strategies.

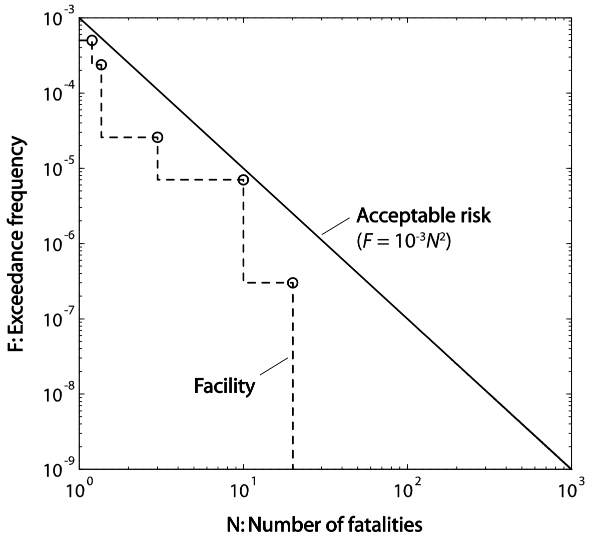


Fig. 11 Acceptable risks for chemical plants in the Netherlands, together with an exemplary F–N curve for a facility (with acceptable risk)

6 Risk Acceptance and Optimization

Once risks are computed, they must be compared against acceptance criteria. Often, multiple risk acceptance criteria (RAC) must be considered. On the one hand, criteria may be defined separately for different consequence classes (fatalities and health effects, economical, environmental). Also, it is often distinguished between individual risk (i.e. the risk accrued by one specific individual), and societal risk (the average risk in a society). The former applies e.g. to the workers in a facility or to inhabitants nearby, the latter to a member of the general public who is exposed only infrequently. On the other hand, RAC may be defined separately by the different stakeholders involved. The operator of the facility and the regulatory bodies may each have their own RAC, whereby the latter are mostly concerned with life and health risks, and increasingly with environmental risks.

RAC can be expressed in different formats, depending on the type of risk considered. It is common to express the acceptable *individual* safety risk in terms of the probability of an individual dying due to an accident during a reference time period. The acceptable *societal* safety risk is often expressed in terms of so-called F–N diagrams, where F stands for exceedance frequency and N stands for the number of fatalities. Figure 11 shows the acceptable societal risk for chemical and process plants in the Netherlands (Jongejan [29]), together with a fictitious curve for a facility. To understand this diagram, consider the point ($N = 10^1$, $F = 6 \times 10^{-6}$): this point signifies that events with $N = 10$ or more are estimated to occur with an annual frequency of 6×10^{-6}. The risk of an activity is acceptable when the entire curve is to the left of the acceptability criterion.

Risk acceptance criteria can be derived by means of different fundamental principles. It is often distinguished between:

(a) Expressed preferences: with this approach, RAC are obtained directly by asking the relevant stakeholders. The difficulty with this approach is that risk levels are often abstract values that are difficult to understand by most individuals and organizations.
(b) Revealed preferences: RAC are derived from the risk that is implicitly accepted by current activities. As an example, when assessing a new system, it can be stated that any risk that is lower or equal to the risk of the present system is acceptable. This is the most commonly applied approach in engineering.
(c) Optimization: RAC can be derived by identifying optimal risk levels, as discussed in Sect. 6.1 below. This allows regulators to require that risks are reduced to a level that can be achieved with reasonable efforts (the ALARP principle outlined in Sect. 4.2).

Existing RAC are often obtained by a combination of the above principles. For example, it is common to derive acceptance criteria from current practice, but then adjust the criteria using optimization principles, e.g. using more stringent criteria for risks where mitigation costs are low. (This approach was followed in deriving the target reliability values provided in Annex B of Eurocode 0 (DIN [19]).) Furthermore, RAC from the public (such as the one shown in Fig. 11) often represent a public consensus, and are derived based on processes involving scientists and engineers, but also representatives of governmental bodies and politicians.

For further examples and details on risk acceptance criteria, the reader is referred to Paté-Cornell [35], Aven and Vinnem [9] and Jongejan [29].

6.1 Optimization

When making decisions involving risk, one should aim at making *optimal* decisions. On the one hand, it is desirable to reduce risks as much as possible; on the other hand, one should use as little resources (money, material, time) as possible for risk reduction. This leads to a classical optimization problem, aiming at finding the optimal trade-off between risk and resources spent for risk reduction, which is illustrated in Fig. 12. The optimal decision is the one minimizing the expected cost, the optimal risk is the one associated with the decision leading to the minimal total expected cost.

Optimization principles can be used to derive absolute RAC (e.g. Rackwitz [38]), or they can be invoked by requiring that risks are reduced to an optimal level, following the ALARP principle. Such an approach is pursued by the UK Health and Safety Executive, which is the regulatory body in the UK (HSE [27]).

The optimization approach requires that all consequences and costs are expressed in the same unit, which is typically a monetary unit. If safety risks are involved, this requires quantifying the value of a statistical life (e.g. Lentz [30]). While this is not without controversy, such an approach is necessary if it is to be ensured that resources are distributed optimally among different activities within a society (for further discussion see Sect. 2.4 in Chap. 3, [47]).

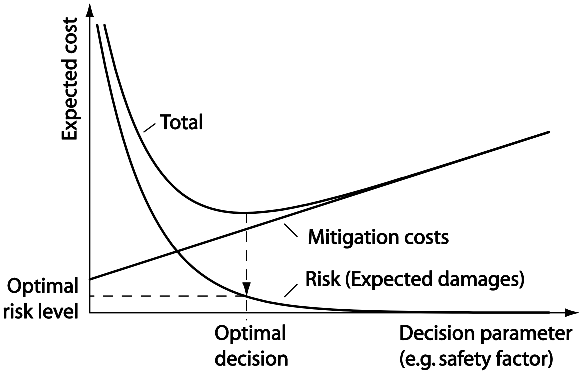


Fig. 12 Trade-off between risk and mitigation cost

7 Food for Thought

- How can we combine an engineering model, which is based on physical principles, with observed data?
- An open question in many risk analyses is how to quantify the effect of human and organizational factors.
- Discuss the context and the system definition of a risk assessment for a nuclear waste depository.
- Why do we differentiate between individual risks and societal risks?
- Engineers must often make decisions involving potentially large consequences and fatalities on the basis of limited information. How can the engineer sleep well at night?
- What is the principle of FORM?
- Why is a linear or quadratic approximation of the performance function around the mean value not suitable to compute the risk of fatigue failure of an aircraft?
- If you need to advise on which of two alternative designs for a train axle should be selected, how would you proceed?
- Often, the most difficult part of an engineering risk assessment is to explain the methods and the results to lay people and even other engineers, due to their difficulties in understanding probability. How can one approach this?

8 Summary

This chapter outlines a framework for engineering risk assessment, with a particular emphasis on quantitative methods. A general procedure is introduced, including system definition, hazard identification, risk analysis, sensitivity analysis, risk assessment and mitigation. Thereafter, it is focused on the quantitative modeling of risk in engineering, which differs from the actuarial approach by combining probabilistic engineering models (typically physical and/or chemical models) with empirical data and sometimes expert knowledge. This is illustrated by brief examples. A brief outline of risk acceptance and optimality in the context of engineering applications concludes the chapter.

References

Selected Bibliography

1. T. Aven, *Foundations of Risk Analysis. A Knowledge and Decision-Oriented Perspective* (Wiley, Chichester, 2005)
2. J.R. Benjamin, C.A. Cornell, *Probability, Statistics, and Decision for Civil Engineers* (McGraw-Hill, New York, 1970)
3. I. Papaioannou, D. Straub, Reliability updating in geotechnical engineering including spatial variability of soil. Comput. Geotech. **42**, 44–51 (2012)
4. M.G. Stewart, R.E. Melchers, *Probabilistic Risk Assessment of Engineering Systems* (Chapman & Hall, London, 1997)
5. Straub, Lecture notes in engineering risk analysis. TU München (2011)
6. Straub, Lecture notes in structural reliability methods. TU München (2011)

Additional Literature

7. G.E. Apostolakis, How useful is quantitative risk assessment? Risk Anal. **24**(3), 515–520 (2004)
8. S.-K. Au, J.L. Beck, Estimation of small failure probabilities in high dimensions by subset simulation. Probab. Eng. Mech. **16**, 263–277 (2001)
9. T. Aven, J.E. Vinnem, On the use of risk acceptance criteria in the offshore oil and gas industry. Reliab. Eng. Syst. Saf. **90**(1), 15–24 (2005)
10. R.E. Barlow, F. Proschan, *Mathematical Theory of Reliability*. Classics in Applied Mathematics, vol. 17 (SIAM, Philadelphia, 1996)
11. E.S. Beckjord, M.A. Cunningham, J.A. Murphy, Probabilistic safety assessment development in the United States 1972–1990. Reliab. Eng. Syst. Saf. **39**(2), 159–170 (1993)
12. M.T. Bensi, A. Der Kiureghian, D. Straub, A Bayesian network methodology for infrastructure seismic risk assessment and decision support. PEER Report 2011/02, Pacific Earthquake Engineering Research Center, University of California, Berkeley (2011)
13. K. Breitung, Asymptotic approximations for multinormal integrals. J. Eng. Mech., Trans. ASCE **110**(3), 357–366 (1984)
14. C.G. Bucher, U. Bourgund, A fast and efficient response surface approach for structural reliability problems. Struct. Saf. **7**(1), 57–66 (1990)
15. S. Coles, L.R. Pericchi, S. Sisson, A fully probabilistic approach to extreme rainfall modeling. J. Hydrol. **273**(1–4), 35–50 (2003)
16. R.M. Cooke, J.M. van Noortwijk, Local probabilistic sensitivity measures for comparing FORM and Monte Carlo calculations illustrated with dike ring reliability calculations. Comput. Phys. Commun. **117**(1–2), 86–98 (1999)
17. C. Czado, E.C. Brechmann, Bayesian risk analysis, in *Risk – A Multidisciplinary Introduction*, ed. by C. Klüppelberg, D. Straub, I. Welpe (2014)
18. A. Der Kiureghian, P.-L. Liu, Structural reliability under incomplete probability information. J. Eng. Mech., Trans. ASCE **112**(1), 85–104 (1986)
19. DIN, Eurocode 0—basis of structural design (EN 1990:2002). Deutsches Institut für Normung e.V. (2001)

20. O. Ditlevsen, H.O. Madsen, *Structural Reliability Methods* (Wiley, New York, 1996)
21. S. Engelund, R. Rackwitz, A benchmark study on importance sampling techniques in structural reliability. Struct. Saf. **12**(4), 255–276 (1993)
22. M.H. Faber, I.B. Kroon, E. Kragh, D. Bayly, P. Decosemaeker, Risk assessment of decommissioning options using Bayesian networks. J. Offshore Mech. Arct. Eng. **124**(4), 231–238 (2002)
23. V. Fasen, C. Klüppelberg, A. Menzel, Quantifying extreme risks, in *Risk – A Multidisciplinary Introduction*, ed. by C. Klüppelberg, D. Straub, I. Welpe (2014)
24. A. Friis-Hansen, Bayesian networks as a decision support tool in marine applications. PhD thesis, DTU, Lyngby, Denmark (2000)
25. A. Grêt-Regamey, D. Straub, Spatially explicit avalanche risk assessment linking Bayesian networks to a GIS. Nat. Hazards Earth Syst. Sci. **6**(6), 911–926 (2006)
26. A. Høyland, M. Rausand, *System Reliability Theory. Models and Statistical Methods*. A Wiley-Interscience Publication (Wiley, New York, 1994)
27. HSE, *Reducing Risks, Protecting People. HSE's Decision-Making Process*. JHSE Books (Health and Safety Executive, Liverpool, 2001)
28. F.V. Jensen, T.D. Nielsen, *Bayesian Networks and Decision Graphs*. Information Science and Statistics (Springer, New York, 2007)
29. R.B. Jongejan, How safe is safe enough? The government's response to industrial and flood risks. PhD thesis, TU, Delft, NL (2008)
30. A. Lentz, Acceptability of civil engineering decisions involving human consequences. PhD thesis, TU München (2007)
31. L.D. Lutes, S. Sarkani, *Random Vibrations. Analysis of Structural and Mechanical Systems* (Elsevier/Butterworth/Heinemann, Amsterdam, 2004)
32. R.E. Melchers, *Structural Reliability Analysis and Prediction* (Wiley, New York, 1999)
33. NTSB, Aviation accident statistics. National Transportation Safety Board, US (2010). Retrieved June 19, 2011
34. A. Papoulis, S.U. Pillai, *Probability, Random Variables, and Stochastic Processes* (McGraw-Hill, Boston, 2009)
35. M.E. Paté-Cornell, Quantitative safety goals for risk management of industrial facilities. Struct. Saf. **13**(3), 145–157 (1994)
36. J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. The Morgan Kaufmann Series in Representation and Reasoning (Morgan Kaufmann, San Mateo, 1988)
37. R. Rackwitz, Reliability analysis—a review and some perspectives. Struct. Saf. **23**(4), 365–395 (2001)
38. R. Rackwitz, Optimal and acceptable technical facilities involving risks. Risk Anal. **24**(3), 675–695 (2004)
39. R. Rackwitz, B. Fiessler, Structural reliability under combined load sequences. Comput. Struct. **9**, 489–494 (1978)
40. R.Y. Rubinstein, D.P. Kroese, *Simulation and the Monte Carlo Method* (Wiley-Interscience, New York, 2007)
41. S.J. Russell, P. Norvig, *Artificial Intelligence: A Modern Approach* (Prentice-Hall, Englewood Cliffs, 2003)
42. J. Song, A. Der Kiureghian, Bounds on system reliability by linear programming. J. Eng. Mech., Trans. ASCE **129**(6), 627–636 (2003)
43. D. Straub, Stochastic modeling of deterioration processes through dynamic Bayesian networks. J. Eng. Mech., Trans. ASCE **135**(10), 1089–1099 (2009)
44. D. Straub, Reliability updating with equality information. Probab. Eng. Mech. **26**(2), 254–258 (2011)
45. D. Straub, A. Der Kiureghian, Bayesian network enhanced with structural reliability methods. Part A: theory. J. Eng. Mech., Trans. ASCE **136**(10), 1248–1258 (2010)
46. D. Straub, M.H. Faber, Risk based inspection planning for structural systems. Struct. Saf. **27**(4), 335–355 (2005)

47. D. Straub, I. Welpe, Decision-making under risk: a normative and behavioral perspective, in *Risk – A Multidisciplinary Introduction*, ed. by C. Klüppelberg, D. Straub, I. Welpe (2014)
48. D. Straub, M.T. Bensi, A. Der Kiureghian, Spatial modeling of earthquake hazard and infrastructure performance through Bayesian networks, in *Proc. ASCE Engineering Mechanics '08 Conference*, University of Minnesota, Minneapolis (2008)
49. B. Sudret, Meta-models for structural reliability and uncertainty quantification, in *Proc. Asian-Pacific Symposium on Structural Reliability and Its Applications*, Singapore (2012)
50. B. Vogel-Heuser, S. Rösch, Integrated modeling of complex production automation systems to increase dependability, in *Risk – A Multidisciplinary Introduction*, ed. by C. Klüppelberg, D. Straub, I. Welpe (2014)

Chapter 13
Integrated Modeling of Complex Production Automation Systems to Increase Dependability

Birgit Vogel-Heuser and Susanne Rösch

In current practice, analysis and development of the same mechatronic component are performed separately for both functional and nonfunctional (for definition see Sect. 3) aspects, often by different engineers and/or engineering teams, and specified in different modeling languages. This gap between the development processes of the different aspects of components leads, on the one hand, to inefficient system development processes and additional iterations between functional and nonfunctional design. On the other hand, it makes for a neglected opportunity to increase system dependability during runtime (Avizienis et al. in IEEE Trans. Dependable Sec. Comput. 1(1):11–33, 2004). By building on basic engineering information, for instance by integrating models containing selected information about a system into its control code, dynamic reconfiguration during runtime helps to increase dependability and reduce risk. Risk in this chapter is defined according to Bertsche as the "product of severity of damage and probability of occurrence" (Bertsche et al. in Zuverlässigkeit mechatronischer Systeme. Grundlagen und Bewertung in frühen Entwicklungsphasen, Springer, Berlin, 2009, p. 55) and the term dependability is used according to Avizienis et al. (IEEE Trans. Dependable Sec. Comput. 1(1):11–33, 2004): "dependability is an integrating concept that encompasses the following attributes:

- *availability* (availability in this context is considered as "the degree to which a system or component is operational and accessible when required for use, often expressed as a probability" (IEEE Std. 610.12-1990, IEEE standard glossary of software engineering terminology, The Institute of Electrical and Electronics Engineers, USA, 1990)): readiness for correct service;
- *reliability*: continuity of correct service;
- *safety*: absence of catastrophic consequences on the user(s) and the environment;

B. Vogel-Heuser (✉)
Chair of Automation and Information Systems, Department of Mechanical Engineering, Technische Universität München, Boltzmannstr. 15, 85748 Garching bei München, Germany
e-mail: vogel-heuser@ais.mw.tum.de

S. Rösch
Automation and Information Systems, Department of Mechanical Engineering, Technische Universität München, Boltzmannstr. 15, 85748 Garching bei München, Germany

C. Klüppelberg et al. (eds.), *Risk – A Multidisciplinary Introduction*,
DOI 10.1007/978-3-319-04486-6_13,

- *integrity*: absence of improper system alterations;
- *maintainability*: ability to undergo modifications and repairs" (Avizienis et al. in IEEE Trans. Dependable Sec. Comput. 1(1):11–33, 2004, p. 13).

Another important term used in this chapter is Quality of Service (QoS). This term has been used recently for different domains. In this chapter QoS is used for the quality that can be assumed when using a substitute strategy to replace another service.

This chapter contributes to the design of system availability, reliability, and safety, focusing on complex production automation systems and highlighting the results by introducing application examples from the control of a continuous thermo-hydraulic particle board press.

Keywords Automation systems · Model-based system and software engineering · Integrated modeling · Safety and functional analysis · Dynamic reconfiguration

The Facts

- Reduced time to market, and lowering of costs for product automation systems, require concurrent engineering.
- Traditional modeling methods do not support integrated development of functional and safety aspects for production automation systems.
- Additional integration of basic engineering models into the control code can be used to increase dependability of production automation systems during runtime by incorporating those models into the control code as a knowledge base for intelligent adaptive behavior.
- Integration of the different functional views of a production automation system, i.e. mechanical, electrical/electronic and software, with their constraints and restrictions will support model based dynamic reconfiguration.

1 Introduction

Today, suppliers of mechatronic products face stronger competition worldwide, resulting in a need for reduced time to market. This leads to decreasing duration times for a project and decreasing start up times, which directly influence plant manufacturers and their automation suppliers. Due to the need for reduction of project duration and time to market, concurrent and simultaneous engineering have become more important (see Fig. 1). Therefore, automation suppliers require support during the whole engineering life cycle in a more efficient way. This applies not only to the design phase as such, but to the entire life cycle. Starting about five years ago [14], forced by competition through globalization, the phases in life cycles of production automation systems, i.e. concept phase, design phase up to construction, needed to

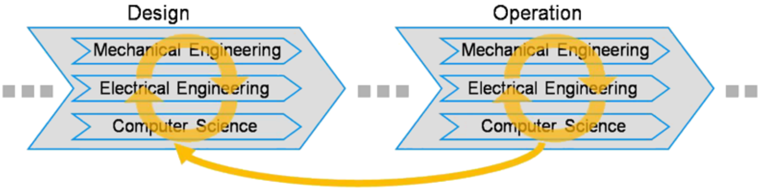


Fig. 1 Concurrent Simultaneous Engineering for design, optimization and operation (CSE) [14]

be shortened and better integrated. Up to now there has been a lack of method and tool integration. The result is a rupture in the engineering work flow between the different phases and the different disciplines.

Another challenge focused on during the last 5 to 10 years is the integration of the different views of functional aspects of a production automation system, such as mechanical, electrical/electronic, and software. Regarding these different aspects a specific challenge can be identified: the necessity to integrate the different disciplines to achieve a more appropriate solution, taking into account all aspects of a system.

During the last few years methods and technologies have been generated to allow the first promising developments of such an integration:

- From computer science, meta modeling has been introduced and is now widely spread in engineering domains, as are model coupling techniques [18, 19] and tools. *Eclipse* for example is one opportunity for the coupling of models of different engineering phases and of different engineering disciplines.
- From the discipline of automation, *AutomationML*, containing a high-level description of the topology of a system within *CAEX*, and several lower-level descriptions for specific aspects of a system such as *PLCOpen XML*, has been introduced as an XML-based description approach for engineering information [5].
- Embedded systems in production automation systems become more powerful computationally, which is a prerequisite for the use of model information and the implementation of intelligent algorithms for adaptive control systems during runtime.
- Last but not least, model based engineering is becoming more popular in the different disciplines of product and production automation, that being the prerequisite for acceptance of re-use and model coupling.

This chapter is organized as follows: first the different views from production automation are introduced and explained using a continuous thermo-hydraulic particle board press as application example. Section three highlights the modeling of functional and nonfunctional requirements, which need to be integrated into the engineering approach and the whole life cycle. Section four presents a first attempt to integrate safety and functional design by mapping the traditional safety models, for example Failure Mode and Effects Analysis (FMEA) and Fault Tree Analysis (FTA) to the functional models using an object oriented approach. The example mentioned above, a real industrial application, is given as an evaluation example. The benefit of integrating basic model engineering information into the control code for dynamic reconfiguration during runtime is demonstrated in section five, based on the assumption that an integrated approach is to be used. This approach allows the control system to cope with malfunctions and, under given safety and operational constraints, to adapt its behavior autonomously. Section six presents some new ideas, and seven summarizes the results and gives our outlook on future work.

2 Different Views on Production Automation Systems

Thramboulidis [12] introduces a three-view model, modeled in the Systems Modeling Language (SysML).[1] The three views of Thramboulidis comprise software engineering, mechanical engineering, and electrical engineering. The central "+1" model is the mechatronic system (MTS) model which is specified using SysML. Thramboulidis et al. highlight that safety is one aspect of the MTS +1 view [13]. They claim that as a result of the synergistic modeling and the integration it is possible to make extensive dependability predictions during the development phases. Unfortunately, this is only a first concept and has not yet been implemented or proven. As a limiting prerequisite, Thramboulidis et al.'s modeling approach requires that all disciplines agree on the same component interfaces, which is rarely the case in industry. Li et al. [4] focus on the integration of mechanical models and models of information technology using SysML.

Wannagat and Vogel-Heuser [16] and Schütz and Wannagat [11] introduce a different three-view model for modeling production automation systems. It is suggested to view systems in the perspective of the technical process, the technical system, and the automation control system. The model supports the interdisciplinary work of different disciplines. Dividing the plant into different domain-specific views enables the description of the different disciplines with one modeling language such as SysML. The model allows each domain-specific engineer to have his/her own view on their components as well as a view of interfaces depicted as relations to components of other disciplines and their requirements. Schütz and Vogel-Heuser [10] use the three views and SysML as an approach to model energy aspects integrated into the functional models.

According to Wannagat and Vogel-Heuser [16], the three views can be described as follows: The technical system relates to the mechanical parts of a plant; therefore it contains information about the layout and the connections of the mechatronic components, as well as energy and material flows between them. In the automation control system controllers, networks, sensors, and actuators are included. Thramboulidis separates this view into software and electrical engineering, which is modeled as a sub-layer in our concept. The technical process itself describes the manufacturing of the product, taking account of the chronological order and all physical

[1] SysML is an extension of the Unified Modeling Language (UML), defined by the OMG, to satisfy the requirements of system engineers. In particular it offers "a semantic foundation for modeling system requirements, behavior, structure, and parametrics, ..." [26].

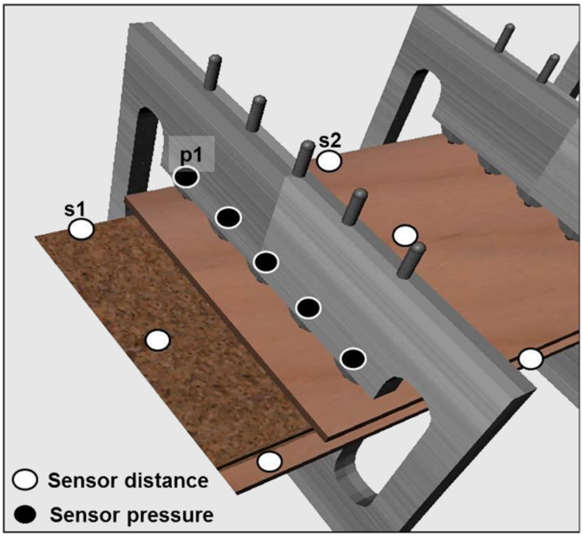


Fig. 2 Process and Instrumentation Diagram of the continuous thermo-hydraulic particle board press as application example for basic engineering information and model base for dynamic reconfiguration [16]

changes made during the process, e.g. chemical or pharmaceutical processes. In our opinion this view is essential because it represents the actual purpose of the system and is not addressed in Thamboulidis et al.'s approach. All three views allow for specialized observation of the whole system and its components. One component contains all aspects that have been assigned to it in the different views. This way the component establishes a connection concerning the content of the different views, enabling the analysis and comparison of all aspects.

Illustration 2.1 Sample application of a continuous thermo-hydraulic particle board press according to [16].

A model of a continuous production process, as basic engineering information, will be introduced in this section. With this model dynamic reconfiguration in order to increase availability will be demonstrated in Sect. 5. The continuous thermo-hydraulic particle board press is a real industrial application (Fig. 2). It is composed of up to 80 separately controlled frames (in Fig. 2 two frames are depicted). Each frame consists of 5 separately controlled cylinders with sensors for pressure (p_1) and distance (s_1, s_2).

The technical process (Fig. 3) is modeled as an internal block diagram (SysML). It shows the different sections of a continuous thermo-hydraulic particle board press from a technologist's point of view. The raw material for the particle board (wooden fibers with glue, i.e. mat) is fed into the press on the left side and will be heated and pressed. The different sections are modeled with regard to different technological functionality. From the initial description an activity diagram (Fig. 4) of this technical process is designed showing the three sections of the press. They are modeled using so-called swim lanes, which are used for structuring activity diagrams.

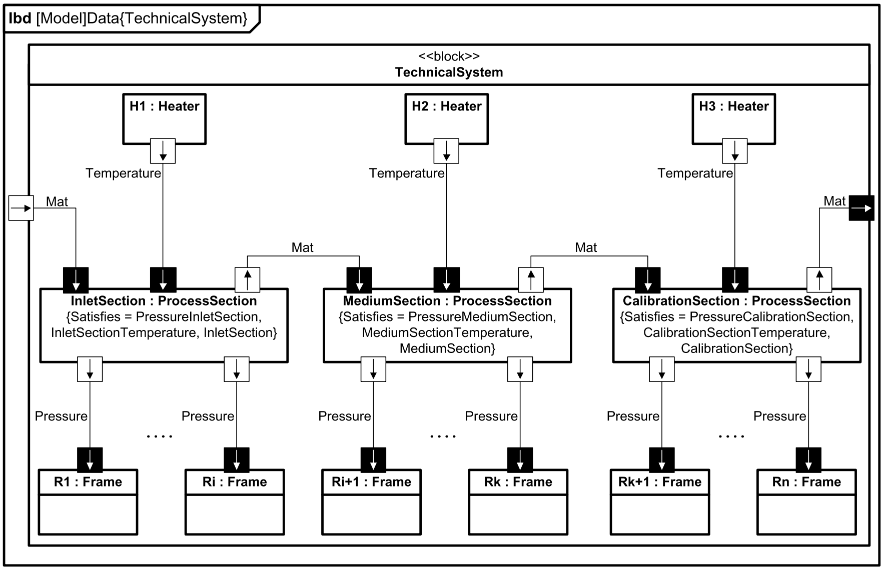

ibd [Model]Data{TechnicalSystem}
<<block>>
TechnicalSystem
H1 : Heater
H2 : Heater
H3 : Heater
Temperature
Temperature
Temperature
Mat
Mat
Mat
Mat
InletSection : ProcessSection
{Satisfies = PressureInletSection, InletSectionTemperature, InletSection}
MediumSection : ProcessSection
{Satisfies = PressureMediumSection, MediumSectionTemperature, MediumSection}
CalibrationSection : ProcessSection
{Satisfies = PressureCalibrationSection, CalibrationSectionTemperature, CalibrationSection}
Pressure
Pressure
Pressure
Pressure
Pressure
Pressure
....
....
....
R1 : Frame
Ri : Frame
Ri+1 : Frame
Rk : Frame
Rk+1 : Frame
Rn : Frame

Fig. 3 Internal block diagram of the technical process view (*top level*) of the application example: continuous thermo-hydraulic particle board press [16]

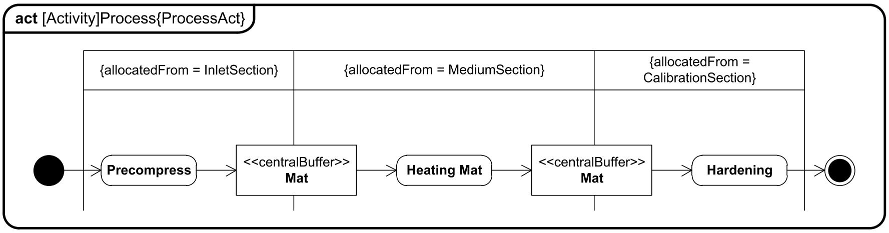

act [Activity]Process{ProcessAct}
{allocatedFrom = InletSection}
{allocatedFrom = MediumSection}
{allocatedFrom = CalibrationSection}
Precompress
<<centralBuffer>>
Mat
Heating Mat
<<centralBuffer>>
Mat
Hardening

Fig. 4 Activity diagram of the technical process (*top level*) [16]

The technical system consists of a generator for electric power supply, the hydraulic system and its interface to the mat (technical process view), the hydraulic main valve and the five valves as well as pressure cylinders of each frame as can be derived from Fig. 2. The mat (bottom right Fig. 5) is pressed by these cylinders, indicated by connectors from each cylinder to the mat representing the force (F). The cylinders increase the pressure as soon as the valves are opened (connector Pressure). These structural aspects of the technical system are depicted in the internal block diagram (Fig. 5).

The automation control system (Fig. 6) represents the chosen automation concept with a classical automation device—a Programmable Logic Controller (PLC (S7))—connected via a bus coupler to the input and output connectors of the single frame of the press. Some of the components are only partially indicated behind a similar component so as to show better the most important connec-

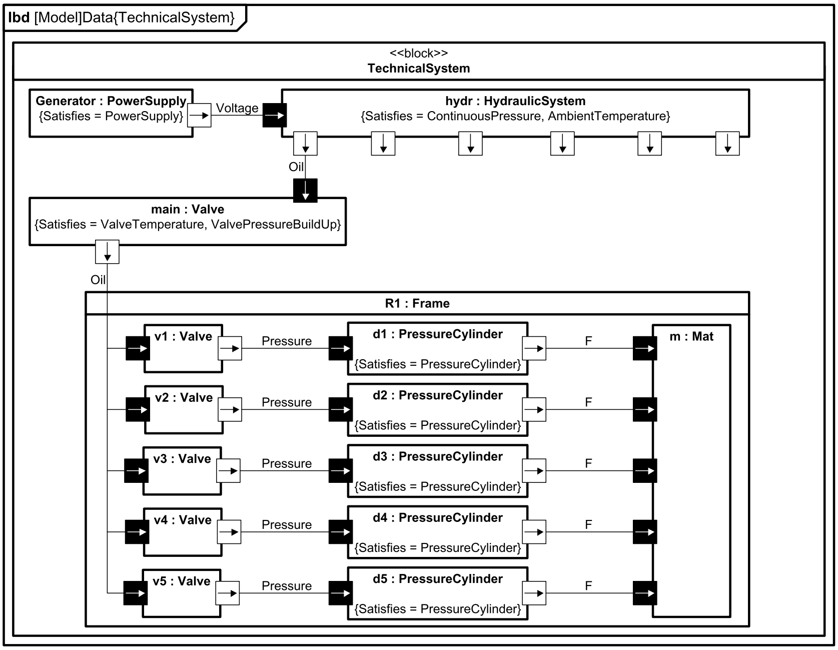

ibd [Model]Data{TechnicalSystem}
<<block>>
TechnicalSystem
Generator : PowerSupply
{Satisfies = PowerSupply}
Voltage
hydr : HydraulicSystem
{Satisfies = ContinuousPressure, AmbientTemperature}
Oil
main : Valve
{Satisfies = ValveTemperature, ValvePressureBuildUp}
Oil
R1 : Frame
v1 : Valve
Pressure
d1 : PressureCylinder
{Satisfies = PressureCylinder}
F
m : Mat
v2 : Valve
Pressure
d2 : PressureCylinder
{Satisfies = PressureCylinder}
F
v3 : Valve
Pressure
d3 : PressureCylinder
{Satisfies = PressureCylinder}
F
v4 : Valve
Pressure
d4 : PressureCylinder
{Satisfies = PressureCylinder}
F
v5 : Valve
Pressure
d5 : PressureCylinder
{Satisfies = PressureCylinder}
F

Fig. 5 Internal block diagram technical system, excerpt with five cylinders in one press frame [16]

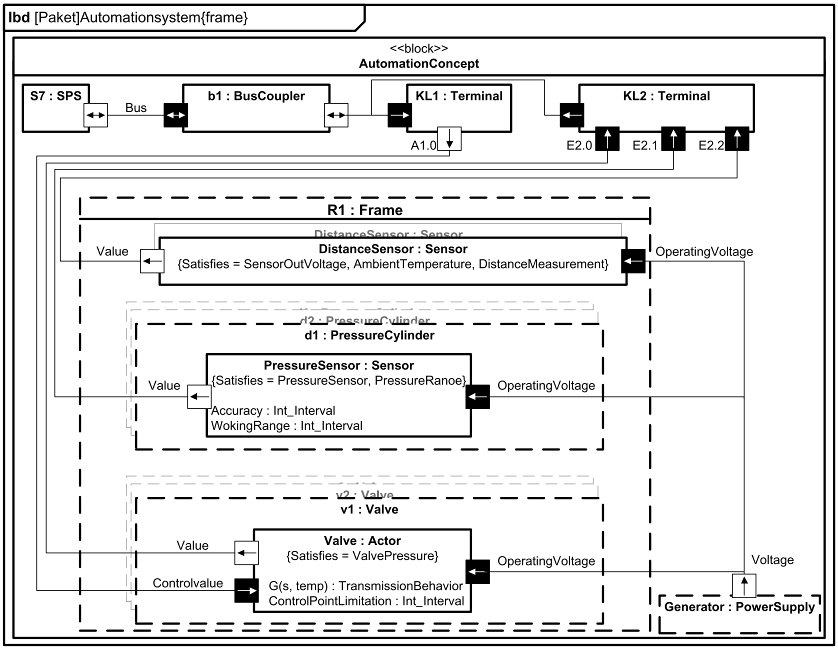


Fig. 6 Internal block diagram of the automation concept

tions within the automation concept. The generator (bottom right), the pressure cylinders, the valve and the frame show the connection to the technical system via a dashed line because they are not part of the automation control system; they belong to another view. Their sensors and actuators, however, belong to the automation system. By representing the most important links between different views using dashed lines the domain engineer has, on the one hand, an overview of his aspects, which helps to reduce complexity, and, on the other hand, the understanding of related components between the disciplines supported by the links.

Recent concepts allow us to generate code for automation applications out of more detailed SysML models [10]. After introducing the three views of the disciplines as one prerequisite, the modeling of requirements will be discussed, regarding also the three different views for one small application example.

3 Modeling Functional and Nonfunctional Requirements

Nonfunctional requirements in general are classified in [1] and from a software engineering point of view in [24]. Dependability and security (Fig. 7) as well as interoperability, maintainability, and time constraints are such nonfunctional requirements. Quality of Service (QoS) has been used recently for different domains (see Sect. 5.1.2). In the case of the continuous thermo-hydraulic particle board press, the application example introduced in this book chapter, QoS comes in where a sensor may be replaced by a calculated one in case of a malfunction (dynamic reconfiguration, Sect. 5). The replacement strategy increases the dependability of the plant under the prerequisite that a minimum required product quality can still be produced with the replacements (virtual sensors).

In our approach functional and nonfunctional requirements are included as constraints to the different views of the production automation system on different levels of detail and granularity starting from a sensor or an actuator up to the entire plant. In Sect. 5.1.2 the tolerance model is introduced as a means to trace whether the required reliability will be maintained, and whether the required probability of a given quality will be reached. If requirements apply to the behavior (dynamic) they need to be modeled in one of the behavior diagrams of the SysML such as the activity diagram. In those diagrams the requirements can be connected to the matching activities fulfilling these requirements.

Illustration 3.1 Temperature and lifespan requirements on the continuous thermo-hydraulic particle board press.

In the requirements model, which consists of different requirement diagrams (documents) according to the three views on the system (upper part of Fig. 8), the requirements of the different domains are described. The domain “technical process” requires the compliance with a defined temperature profile (up to 280 °C)

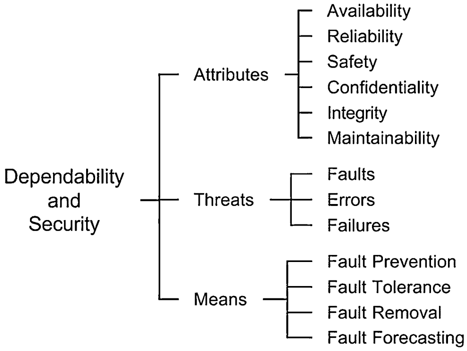


Fig. 7 The dependability and security tree [1]

inside the inlet zone of the press, whereas the domains "automation system" and "technical system" restrict the temperature to ranges in which their components cannot be damaged (automation system: up to 60 °C; technical system: up to 300 °C).

These functional requirements are traced to corresponding constraint blocks using the "satisfy" relation (middle part of Fig. 8). The constraints are instanced inside the blocks which describe the modules of the plant that has to fulfill these requirements by means of a *composite aggregation* (lower part of Fig. 8). For example the block "InletZone" instances the constraint that is linked to the requirement defining a temperature profile. To show how nonfunctional requirements are modeled, a requirement for the valve to reach its defined lifespan is modeled in Fig. 8. The nonfunctional requirement "Lifespan valve" is modeled the same way the functional requirements are modeled. It is linked to a constraint limiting the frequency of opening and closing the valve and is also categorized as part of the technical system.

A heating valve—being part of the automation system of the press—instances the constraint that restricts the temperature due to the limitations of its electronic components ("tempPAS"). Additionally, it references the constraints expressing the limitations resulting from requirements of the other two domains by means of *shared aggregations* ("TempInletZone" and "TempHeatExch"). With this information a parametric diagram can be modeled that describes the limitations regarding the temperature, which were originally stated in the requirements diagrams (documents). The SysML parametric diagram of a heating valve (Fig. 9) within the continuous thermo-hydraulic particle board press contains the different domains' limitations regarding temperature affecting this module. The constraints that are referenced by the valve only by shared aggregations are displayed with dashed lines. This indicates that the valve does not constrain the temperature to these limitations; however, the limitations of other modules of the continuous thermo-hydraulic particle board press ("InletZone" and "HeatExchanger") affect the valve. The temperature outside the valve is composed of the three modules' temperatures. To calculate the temperature inside the valve, thus the tem-

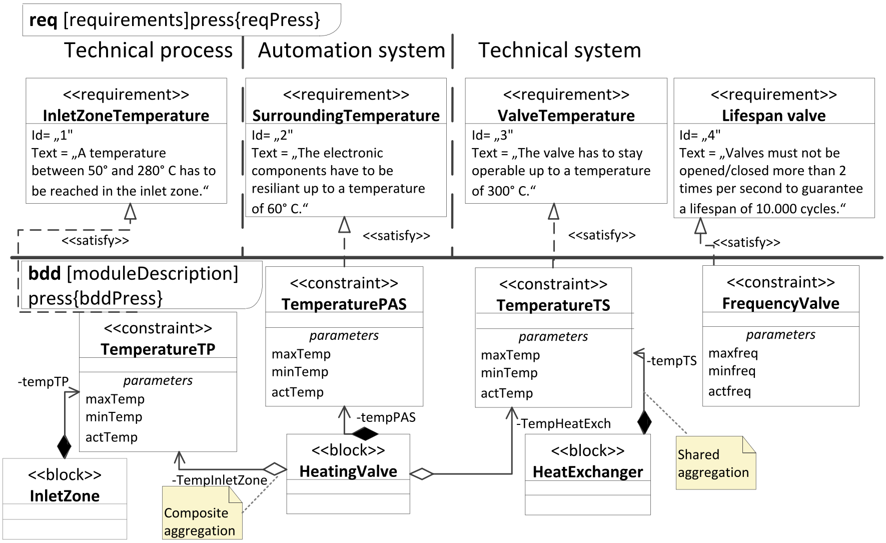


Fig. 8 Requirements and block definition diagram for the requirements concerning temperature

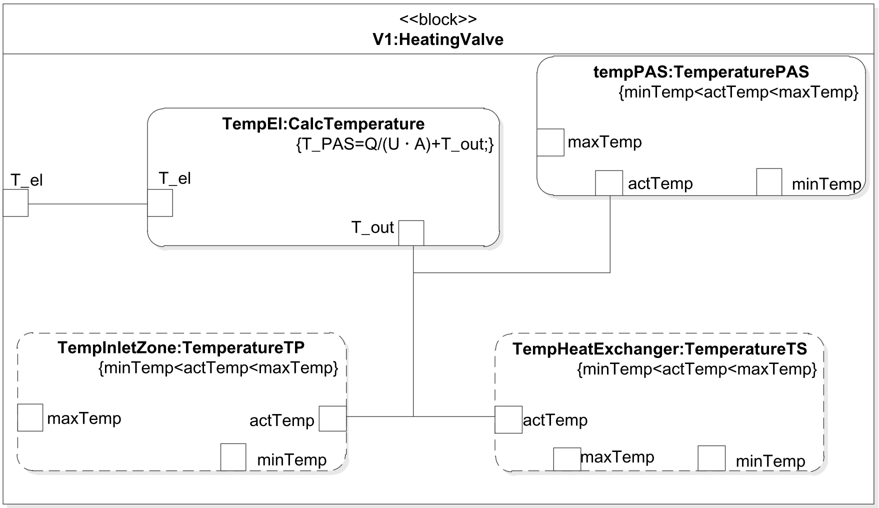


Fig. 9 Parametric diagram depicting the temperature of the heating valve

perature actually affecting the electrical components, the constraint “CalcTemperaturePAS”, which expresses a formula for the heat transmission through the surface, is used.

The SysML models are the basis for coping with faults in production automation systems (see Sect. 5).

4 Integration of Potential Malfunctions and Faults as an Integrated Part of Production Automation Systems' Behavior

The task of risk analysis is to identify and evaluate risk for the entire system (see Chap. 12, [25]) at an early stage, based on the weaknesses of individual components of a system. In the first part of this section the traditional safety analysis approaches, which analyze and evaluate faults, errors, and failures, are introduced, i.e., Failure Mode and Effects Analysis (FMEA) and Fault Tree Analysis (FTA). We further discuss their application as part of production automation systems design.

Laprie et al. define faults, errors, and failures in relation to provided services or functions. An "adjudged or hypothesized cause of an error is called a fault" [1, p. 13]. The "definition of an error is the part of the total state of the system that may lead to its subsequent service failure" [1, p. 13], which means an error does not inevitably lead to a failure. A "failure is an event that occurs when the delivered service deviates from correct service" [1, p. 13].

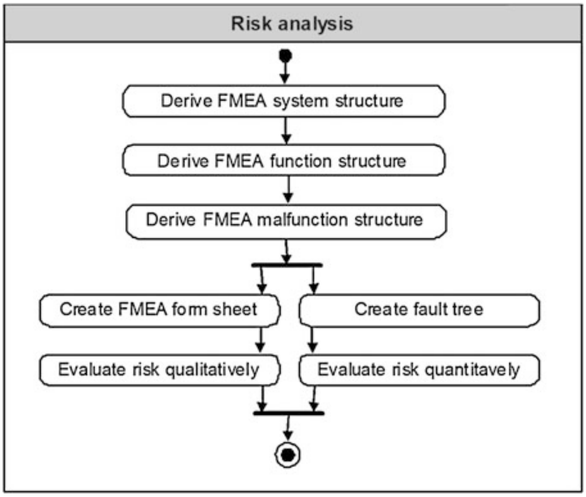
In the automotive domain the current standard follows VDA [27] to check the safety of systems using FMEA [22] and FTA [21].

Rink[2] proposes a method that combines requirement analysis, system modeling, design of control functions, and creation of a safety concept and risk analysis of the system using an integrated object-oriented system model similar to the approach proposed in Sect. 2. The model takes both the desirable and the undesirable behavior of the system into account. Hence a universal system model is created that can be used for the design of functions as well as for the safety concept and risk analysis. Objects that are modeled as components within the system model are determined, based on the physical structure. The components' parameters and state variables appear as attributes in the object specification. The desirable and the undesirable behaviors are described as operations and graphically represented in state charts and activity diagrams. Interactions between objects are depicted by collaboration diagrams. The requirements for the nominal behavior are the basis for the design of control functions. The safety concept must recognize possible system errors and, if necessary, activate appropriate replacement functions so the system maintains or reaches a safe state again. Rink presents a use case-oriented approach for risk analysis generating the structures of the risk analysis based on mapping rules from an object-oriented system. The risk analysis starts with the construction of the FMEA system structure based on object and class diagrams of the object-oriented system model (Fig. 10).

Then the *FMEA function structure* is derived from the state charts, the activity, and the collaboration diagrams that model the desired behavior. On the basis of the undesired behavior of the system model the *FMEA malfunction structure* is created. To facilitate the generation of the FMEA-function and malfunction structure from the object-oriented system model, it is necessary to limit the variety of the means of description available in (UML/SysML) using modeling guidelines. These guidelines

[2] In the following the results of a research cooperation with an automotive company will be introduced which refers to the PhD of Anton Rink [6–9].

Fig. 10 Steps and procedure during risk analysis [7]


Risk analysis
Derive FMEA system structure
Derive FMEA function structure
Derive FMEA malfunction structure
Create FMEA form sheet
Create fault tree
Evaluate risk qualitatively
Evaluate risk quantitavely

are the basis for the development and use of efficient transformation algorithms to generate the structures of the risk analysis. Both the qualitative and the quantitative risk analysis are performed by means of FMEA form sheets (see Sect. 4.1, Fig. 13) and fault trees (Sect. 4.2, Fig. 14). Depending on the results of the analysis, measures to optimize the control functions and the safety concept are taken in order to meet the requirements regarding availability and safety.

4.1 Failure Mode and Effects Analysis (FMEA)

FMEA is used to analyze whether all malfunctions of the control function have been detected in the safety concept and whether measures to avoid critical states are available. For this purpose possible errors of a system and its control function are registered and in the filled-in FMEA form sheet (see Fig. 13). With support of a fault simulation, consequences of errors and their impact on a system are determined. Due to the object-oriented structure of the system model, specific information from the FMEA forms can be considered as attributes of the objects in the object-oriented system model. Hence a system model which can be used for the functional design, the creation of the safety plan, and the risk analysis is developed.

In accordance with Bertsche et al. [2] the FMEA follows five steps:

1. Creation of a hierarchical system structure out of system elements (system structure tree)
2. Description of the functions and the function structure (function structure tree)
3. Implementation of the malfunction analysis, e.g. detection of possible errors, causes of failures and failure sequences (malfunction structure)
4. Risk evaluation in the FMEA form sheet
5. System optimization with the goal of avoiding malfunctions or reducing risks

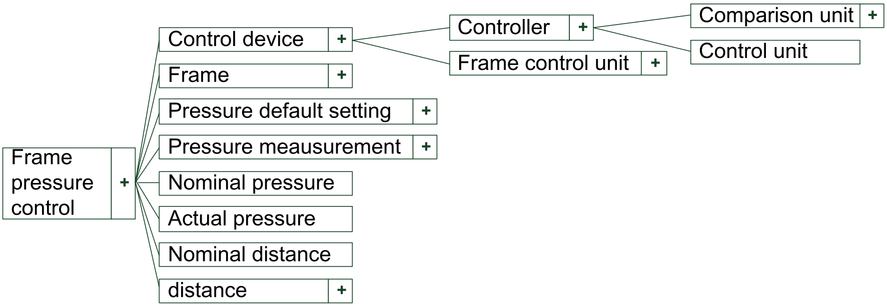


Fig. 11 FMEA—system structure tree showing the system structure of the frame pressure control; objects are modeled as system elements of the system "frame pressure control"

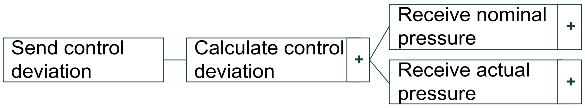


Fig. 12 FMEA—function structure tree (of step 2, Fig. 10) showing the function structure section of the control deviation calculation [8]

Illustration 4.1 Sample application: FMEA "pressure control" for the continuous thermo-hydraulic particle board press.

The FMEA system structure (Fig. 11) is derived from class and object process diagrams by depicting objects as system elements to bridge the gap between an object oriented approach and the safety analysis with FMEA (explained later in this illustration). Object and activity diagrams define the function structure.

Starting from the system function structure tree (Fig. 12) the malfunction structure can be determined. Figure 13 shows the completed FMEA form sheet for the system element "comparison unit", which can be found in Fig. 11 as a subcomponent of "controller". The function analyzed in this particular form sheet is called "Receive nominal pressure", which is only one of the functions that are fulfilled by the comparison unit. All possible malfunctions of the system are identified and entered into the FMEA form sheet (Fig. 13). By means of a fault simulation the failure sequences and their effects on the comparison unit are recognized. In this example a potential failure sequence begins with the comparison unit storing the actual pressure at too high a level. Subsequently, the pressure is assumed to be too high and a negative control difference is calculated. A potential effect of this failure is that the control receives a nominal pressure that is too low. If an error causes a critical state, a high value (maximum 10) is filled into the failure severity column (S). The next step is to check if one of the surveillance functions is available for the recognition of the failure and to evaluate the quality the detection of the failure has. In case of an absolute certainty that the failure is detected, the minimum value of 1 is filled into the column "detection probability" (D). If there is a substitute strategy, which allows avoiding the negative effects of the failure, a positive rating (low value) in

System element: Comparison unit								
Function: Receive nominal pressure								
Potential failure mode	S	Potential effect(s) of failure	Potential cause(s) of failure	Measures of prevention	O	Measures of detection	D	RPN
Store actual pressure too high	10	Received nominal pressure too high	Faulty controller specification of programming	Review and test procedures, estimate actual pressure	3	Checking and establishment of plausibility of the incoming and outgoing variables	3	90
> Store pressure too high	10							
>> Negative control difference	10		Controller hardware defect	Hardware tests	2	Checking of hardware during runtime	2	40
Store actual pressure too low	5	Received nominal pressure too low	Faulty controller specification of programming	Review and test procedures, estimate actual pressure	3	Checking and establishment of plausibility of the incoming and outgoing variables	3	45
> Store pressure too low	5							
>> Positive control difference	5		Controller hardware defect	Hardware tests	2	Checking of hardware during runtime	2	20
Don't store actual pressure	5	Nominal pressure not received	Controller hardware defect	Hardware tests, estimate actual pressure	2	Checking of hardware during runtime	2	20

Fig. 13 FMEA—form sheet for the function "Receive nominal pressure" of the system element "comparison unit"

the column of occurrence probability of the failure (O) is recorded. In this example it is possible to model the particle board press and to estimate the nominal pressure in a simulation. Therefore the plausibility of the value can be checked, the failure can be detected, and the substitute strategy of estimating the actual value can be applied. The weaknesses are ranked by risk priority numbers (RPN), which are derived for each failure by calculating the product of risk importance, detection probability and occurrence probability. Possible risk priority numbers are values between 1 (no weakness) and 1000 (extremely critical weakness).

4.2 *Fault Tree Analysis (FTA)*

While FMEA is used for qualitative failure analyses, the FTA enables quantitative analyses [3]. Thus it complements the FMEA and is especially effective when used in combination with it. The FTA analysis is classified as a Top-Down Analysis [2]. With the FTA [21] all possible failure combinations leading to an undesirable state are detected. If the failure occurrence of the individual failures is known, the failure occurrence for the considered undesirable event can be calculated. In consequence if all of the individual failure rates are known, the failure rate of the different system failures may be calculated. The aim of the FTA is to define the failure combinations that lead to undesirable events and their occurrence frequency. The fault tree for an undesirable event can be derived from the malfunction structure by extending the cause-effect information with logical information. If a fault tree contains exclusively OR, AND and NOT links, the frequency of occurrence of an undesirable event can be determined with the application of Boolean algebra and probability theory. In order to determine the probabilities of failures, the parameters of the failure behavior must be established. The FTA is especially useful when identifying critical failure paths. It allows in particular the consideration of the effects of multiple failures. If an FMEA is already available, this information can be used as a basis for the FTA as a possible form of failure [2]. In what follows, the FTA will be discussed for the continuous thermo-hydraulic particle board press.

Illustration 4.2 FTA: "pressure control" for the continuous thermo-hydraulic particle board press.

One segment of the fault tree, that for the case of the pressure control not operating regularly (chosen wording "Do_not_calculate_correct_control_deviation"), is given in Fig. 14. In a fault tree the object name is given at the first position and the name for the operation responsible for the undesirable event at the second position (e.g. Fig. 14: ComparisonUnit.Do_not_calculate_correct_control_deviation). As already mentioned, individual object failures are linked by Boolean operators in such a way that the failure occurrence can be calculated using Boolean algebra and probability theory [20]. The occurrence of failures of the pressure control (p_{logic}) is calculated using the failure probabilities (p_1, p_2, p_3, p_4, p_5) of the different component failures for bus inlets, control logic, and the nominal pressure determination (see Eq. (1)).

$$p_{\text{logic}} = \bigl(1-(1-p_1)\cdot(1-p_2)\cdot(1-p_3)\bigr)\cdot\bigl(1-(1-p_4)\cdot(1-p_5)\bigr). \tag{1}$$

The fault tree (Fig. 14) may also be derived from the technical system. If the nominal value of the pressure cannot be calculated, for example, the cause may be a bus problem (p_3, Message was not received), the encoder not working (p_2), or a faulty configuration or programming of the unit (p_1, reception of actual pressure too high/too low). The nominal pressure cannot be estimated if the frame model is not modeled (p_5) or the actual distance (p_4) is not received. The whole control deviation cannot be calculated if both the nominal pressure is not calculated and the estimated pressure is not available.

On the basis of qualitative (FMEA) and quantitative (FTA) risk assessment, measures to optimize system safety are initiated in a way that risks of individual failures and probabilities of system failure do not exceed specified limits. In the approach proposed by Rink, faults cannot be compensated according to the safety concept. If the unavailability of a function is detected the system is turned into a safe state until the error is corrected and the user is informed in addition. In the next section we

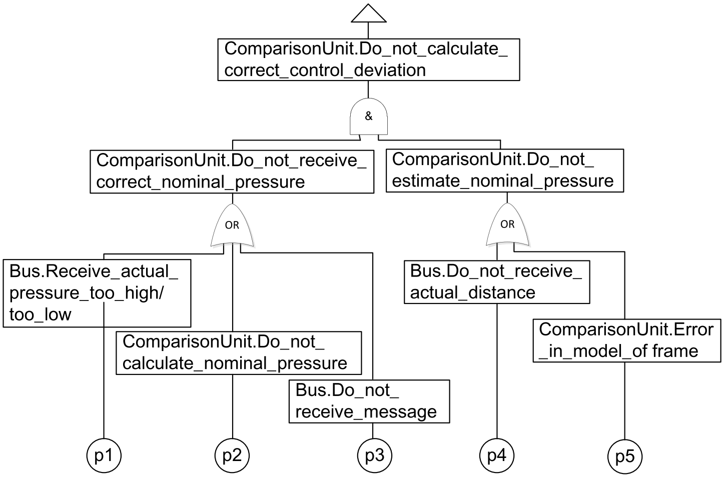

ComparisonUnit.Do_not_calculate_correct_control_deviation
&
ComparisonUnit.Do_not_receive_correct_nominal_pressure
ComparisonUnit.Do_not_estimate_nominal_pressure
OR
OR
Bus.Receive_actual_pressure_too_high/too_low
ComparisonUnit.Do_not_calculate_nominal_pressure
Bus.Do_not_receive_message
Bus.Do_not_receive_actual_distance
ComparisonUnit.Error_in_model_of frame
p1
p2
p3
p4
p5

Fig. 14 Fault tree for the malfunction “Do not calculate correct control deviation”

argue why this strategy is appropriate and accepted for the automotive domain, but is not applicable to machine and production automation.

5 Coping with Faults in Production Automation Systems

After stops reached during the production process in production automation, restarts are time consuming and not always done easily. For this reason they are executed only when absolutely necessary, such as in situations of fire, hazard, or a long shut down as a result of cleaning and maintenance procedures. This is the case for the particle board press introduced in Sect. 2. Therefore, mechanisms are required to cope with the failures of one or more subsystems, devices, or sensors, and to operate the plant with a possible lower product quality until a regular shut down can be scheduled.

5.1 Adaptive Control by an Agent Based Approach

An[3] approach well suited to coping with failures of one or more subsystems is dynamic reconfiguration during runtime based on adaptive control systems using agent-oriented software. By implementing additional engineering models into the control code a basis for adaptive behavior can be reached. In the following the agent

[3] In the following the results of research of Wannagat and Vogel-Heuser will be presented which refers to [15–17].

definition of the VDI/VDE guideline 2653 is applied: "An agent is an encapsulated (hardware/software) entity with specified objectives. An agent endeavors to reach these objectives through its autonomous behavior, in interacting with its environment and with other agents. Agents represent a modeling concept for solving technical problems independently of a specific form of realization" [28].

A main part of the agents' duties in the context of increasing availability of a control system is to detect, analyze, and handle faults. In the present state of the art, agents merely focus on instrument fault detection by using analytical redundancy between different measurement points. For every real sensor, which has functional dependencies to other sensors, we calculate additional virtual sensors using values of neighboring real sensors. The virtual sensors are used to validate the corresponding measurements and detect faults (parity space approach). If there is more than one virtual or real sensor available at one measurement-point, it is possible to detect a single fault (isolation). In principle, the virtual sensor values will never be as precise as real sensor values, but the information is beneficial for fault diagnosis purposes as well as for substitution. In the case of fault diagnosis the lack in precision may cause false alarms or a lack in sensitivity. In the case of substituting a real sensor by a virtual one, this loss of precision is relevant for closed loop controls and for the whole control strategy of the production process. Therefore it is insufficient just to calculate virtual sensors and to substitute for faulty real sensors. Additionally, the consequences of such substitutions and the constraints of the control system have to be taken into consideration. An agent knowledge base contains two main components—constraints and knowledge. Constraints define the margin of the action space that is used by the agents to take decisions. Use of basic engineering models representing knowledge about systems allows choice of alternative means to cope with failures at a certain point within the action space.

5.1.1 Requirements and Boundaries to Define an Action Space

The requirements defined for the agents, such as time and dependability as nonfunctional requirements, are the basis for specifying the action space (Fig. 15) and for defining the goals together with the parameters to achieve them.

In the first step, the requirements relating to contained modules, interfaces, and connections are collected separately for each of the system views presented in Sect. 2. Unlike the components, which can be related to more than one view, these requirements are strictly related to their own view. The same component can be viewed under different aspects. For example a valve is an actuator for the automation control system, it has a mechanical representation in the technical system, and it controls the flow rate from the technical process view. This leads to two advantages: firstly, it reduces the complexity because the requirements survey is distributed to three views for each element and secondly, it is the first integration of the requirements of different views in the same element (see Sect. 3).

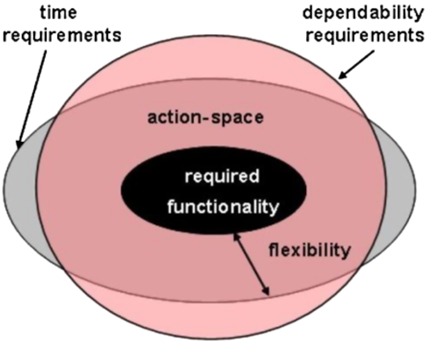


Fig. 15 Boundaries of the agent's action space [16]

In the second step, relations between functional and nonfunctional requirements defined in the first step need to be analyzed and linked to each other. The predefined functions of the modules are linked to the appropriate requirements and boundaries regardless of the three views. The result is a network of requirements and their relations, which makes it easy for the developer to get an overview of functionalities, requirements, and boundaries.

In the third step the developer evaluates if the desired flexibility may be reached with the predefined boundaries, and specifies how much flexibility the agents should have. The action space allows the agents to navigate and to achieve their goal within the appointed boundaries. Every parameter has to be checked regarding its influence on the time delay according to real time requirements and the failure probability of its related functionality in matters of dependability. Using these relations the agents are able to achieve their goal by changing the relevant parameters.

5.1.2 Diagnosis and Fault Management

To fulfill the requirements concerning the reliability of a plant, the agents have to know the effect of their actions as well as estimate the significance of a changing environment regarding their requirements within the whole system.

Similar to an FTA using the given modular structure, the developer is able to specify the relationship between the functionality of a module and its sub modules (Fig. 16). The goal is to specify the probability for correct execution of each function based on functions of the related subsystem, until the basic elements of the control system at the bottom level are reached. In this way it is possible to calculate the probability of a malfunction at the time a quality of a sensor measurement changes. The tolerance model is similar to a fault tree, but focuses not only on the top event, the function or the malfunction of the system and its probability, but also on the quality similar to a QoS with which the operation will continue under the prerequisite of a given replacement strategy (discussed in Sect. 5.1.3). This is realized by software agents, who implement the knowledge about the relation between the

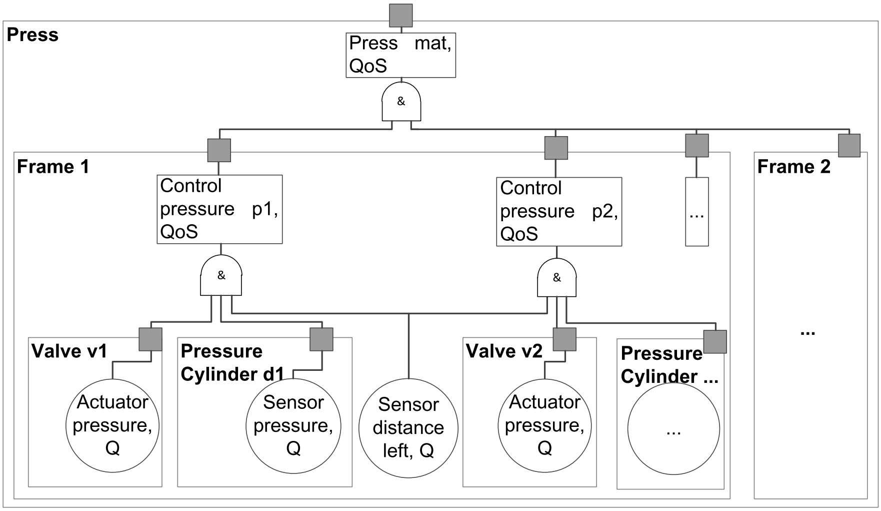


Fig. 16 Tolerance model combined with modular structure (*boxes* show module, *small grey squared boxes* depict interfaces between modules)

quality of a control system element, for example the quality of the sensors' measurements, the precision of the actuators' actions, and the functionality of a component. This relation is used to calculate both the risk of a failure regarding the observed changes and the effect of possible counteractions such as replacement of a sensor by a virtual (calculated) sensor value. Each agent is able to observe all elements of the control system and to relate the real values to the calculated values of its internal system model.

5.1.3 The Knowledge Base

Next, a knowledge base, which allows detecting sensor failures, calculating a surrogate value, and estimating the resulting precision at runtime, will be introduced. One important point for the design of such a knowledge base is that it is easy to design and implement in a Programmable Logic Controller (PLC) environment so as to be calculated during runtime. A very simple and powerful notation for this purpose, which is well known in the domain of automation, is the directed graph [20]. In this graph, each node represents a measurement point. It is equipped with a value source that can be either a real or a virtual sensor. A quality value at each node describes the accuracy of the measured or calculated value. The quality value ranges continuously from 0 to 1. The edges of the graph describe functional correlations (f, Fig. 17) between the measurement points and represent the analytical dependencies which are used to calculate virtual sensor values at runtime. The directions of the arrows indicate sensor values that are appropriate for a substitution (Fig. 17). In Fig. 17 on the right side, the sensors used within the quality model are shown.

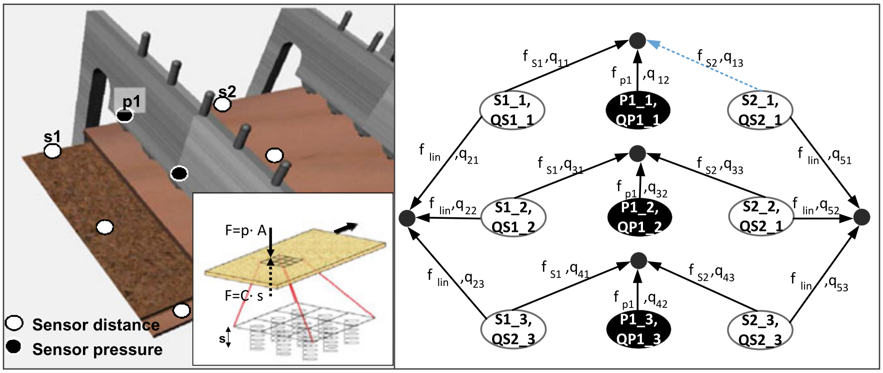


Fig. 17 Analytical dependencies of sensors in the hydraulic press in the quality model [15]

Sensor "$S2\_1$" can be calculated, using the function "f_{S2}" (edge) and the sensor values "$P1\_1$" and "$S1\_1$", for example. The function "f_{S2}" expresses the dependency between the thickness of the incoming material into the frame concerned (one of the sensors s_1), the pressure of the hydraulic cylinder (one of the sensors p_1), and the thickness of the outgoing material, using a spring model. The spring constant (C, Fig. 17, bottom left) represents the elasticity of the material and depends on the values of the actual temperature, the density, and the humidity of the wood. The black dots are used if more than one sensor is required to calculate a virtual sensor (Fig. 17, right).

It is possible to use virtual sensors as source for other virtual sensors and the probability for that rises with the number of failures and corresponding substitutions. The precision of virtual sensor values is possibly reduced by inaccurate models and time aspects, e.g. dead time or delays because of the underlying measurement, the field bus, or the calculation of virtual sensor values in the PLC. Reduced precision lowers the quality of the virtual sensors compared with original measurements. This loss of precision is given by a quality factor q similar to QoS for a measuring value (q, Fig. 17, right) which is bound to every arrow of the graph and described by values between 0 and 1. It represents the decreased quality (precision) by using this calculated virtual sensor instead of the real sensor. In addition a quality value Q (Q, Fig. 17, see QoS Sect. 3) at every node represents the precision of the real sensor itself. Therefore, the quality (precision) "q_{13}" of a virtual sensor replacing "$S2\_1$" may be calculated by the precision of the real sensor quality value "$QS1\_1$" of "$S1\_1$" and the quality value "$QP1\_1$" of the real pressure "$P1\_1$" multiplied by the loss of the quality because of the replacement represented by "q_{21}" as quality factor.

Furthermore, the agents use the quality value of a virtual sensor to determine the effect on the availability of the plant operation and to compare it with the given requirements and constraints. The reliability of sensor values is evident for processing automated production systems. Although the substitution of real sensors with calculated virtual sensors increases readiness in case of partial faults, it risks the accuracy of the process flow. While the correctness of possible alternative strategies for static systems is determined during development time, an agent based dynamic system decides this during runtime. Both have to decide whether the production process can be continued with replaced, calculated values or if it has to be suspended. The threshold, which defines the agent's decision, is oriented by a user defined safety requirement for the specific part of the process or the technical system. The loss of a sensor triggers the reconfiguration of the control behavior automatically. The automated result may be a single parameter adjustment or an immediate shut down of an entire plant and is characterized as dynamic reconfiguration at runtime.

The introduced concept was evaluated successfully by applying it to the particle board press. The reconfiguration took a maximum of two PLC cycles to adjust. Further information about the application is given in [17].

6 Food for Thoughts

There is still a great need for engineering support for adaptive control systems in manufacturing, to increase dependability as well as their interaction with the human as the operator, and by that adaptable behavior for individuals. Our contributions in the future will focus on these tasks. Besides this, a lack in visualization of adaptive control systems is clearly apparent and needs to be covered in future research. As the acceptance of any new technology is based on transparency and trust, we need to find appropriate visualization patterns to open the black box of intelligent behavior for operators, to gain acceptance and trust. On the other hand we need to provide automatic models derived from engineering data to reduce modeling effort for industrial application. This is another challenging application-oriented research topic, because it is strongly related to different domain-specific models in different tools and their interfaces.

7 Summary

The introduced methods and measures of an integrated safety and functional analysis, including the different views in a production automation system, help to reduce time to market and cost, as mentioned in the introduction. There is still a lack of meta models for the different disciplines and domains as well as in a support of integrating different models under the constraint of changes in variants and versions. In tool support some deficiencies must be accepted as well, but the basic concepts are available, which provide the foundation for necessary improvements. We have introduced three first steps on the road to an integrated modeling of safe and dependable complex systems: the cross-disciplinary modeling with SysML, a first approach of integrating safety and functional modeling, and finally the adaptive behavior based on basic engineering models to support dynamic reconfiguration of manufacturing systems during runtime.

References

Selected Bibliography

1. A. Avizienis, J.-C. Laprie, B. Randell, C.E. Landwehr, Basic concepts and taxonomy of dependable and secure computing. IEEE Trans. Dependable Sec. Comput. **1**(1), 11–33 (2004)
2. B. Bertsche, P. Göhner, U. Jensen, W. Schinköthe, H.-J. Wunderlich, *Zuverlässigkeit mechatronischer Systeme. Grundlagen und Bewertung in frühen Entwicklungsphasen (Foundations for a Reliability Evaluation of Mechatronic Systems. Reliability of Mechatronic Systems).* VDI-Book (Springer, Berlin, 2009), pp. 7–45
3. R. Lauber, P. Göhner, *Prozessautomatisierung (Process Automation), 1* (Springer, Berlin, 1999)
4. F. Li, G. Bayrak, K. Kernschmidt, B. Vogel-Heuser, Specification of the requirements to support information technology-cycles in the machine and plant manufacturing industry, in *14th IFAC Symposium on Information Control Problems in Manufacturing (INCOM)* (2012)
5. A. Lüder, L. Hundt, A. Keibel, Description of manufacturing processes using AutomationML, in *2010 IEEE Conference on Emerging Technologies and Factory Automation (ETFA)* (2010), pp. 1–8
6. A. Rink, Entwicklung einer Methode für die systemtechnische Auslegung verteilter und sicherheitskritischer Führungsfunktionen für Fahrzeugantriebe (Development of a method for the systemic interpretation of distributed, safety critical functions for vehicle engines). PhD Thesis, Bergische Universität Gesamthochschule Wuppertal, Wuppertal, Germany (2002)
7. A. Rink, Systemtechnische Auslegung sicherheitskritischer Führungsfunktionen für Fahrzeugantriebe (System-oriented design of safety critical control functions for vehicle drives). Autom.tech. Prax. **11**, 66–74 (2002)
8. A. Rink, B. Vogel, Objektorientierte Methode für die Auslegung eines verteilten Automatisierungssystems unter Sicherheitsaspekten (Object-oriented method with safety aspects for the design of a distributed automation system), in *Tagungsband zur Fachtagung Verteilte Automatisierung: Modelle und Methoden für Entwurf, Verifikation, Engineering und Instrumentierung*, 22.–23.3.2000, Magdeburg (2000), pp. 266–272
9. A. Rink, B. Vogel-Heuser, Auslegung von Führungsfunktionen für einen Fahrzeugantrieb anhand eines integrierten Prozessmodells unter Berücksichtigung der Sicherheitsaspekte (Design of a key function for a vehicle drive using an integrated process model taking safety aspects under consideration), in *VDI Kompetenzfeld Informationstechnik, Software-Engineering in der Praxis, Tagung*, 28.02.–01.03.2002, Düsseldorf. VDI-Berichte, vol. 1666 (VDI Verlag, Düsseldorf, 2002)
10. D. Schütz, B. Vogel-Heuser, Modellintegration von Verhaltens- und energetischen Aspekten für mechatronische Module (Integrated modeling of module behavior and energy aspects in mechatronics). Automatisierungstechnik **59**(1), 33–41 (2011)
11. D. Schütz, A. Wannagat, Domänenspezifische Modellierung für automatisierungstechnische Anlagen mit Hilfe der SysML (Domain specific modelling of technical facilities using SysML). Autom.tech. Prax. **3**, 54–62 (2009)
12. K. Thramboulidis, The $3 + 1$ SysML view-model in model integrated mechatronics. J. Softw. Eng. Appl. **3**, 109–118 (2010)
13. K. Thramboulidis, D. Soliman, G. Frey, Towards an automated verification process for industrial safety applications, in *IEEE Conference on Automation Science and Engineering* (2011), pp. 482–487
14. B. Vogel-Heuser, G. Kegel, K. Bender, K. Wucherer, Global information architecture for industrial automation. Autom.tech. Prax. **1**, 108–115 (2009)
15. A. Wannagat, Entwicklung und Evaluation agentenorientierter Automatisierungssysteme zur Erhöhung der Flexibilität und Zuverlässigkeit von Produktionsanlagen (Development and evaluation of agent oriented automation systems to increase flexibility and reliability of production plants). PhD Thesis, Technische Universität München, Munich, Germany (2010)

16. A. Wannagat, B. Vogel-Heuser, Agent oriented software-development for networked embedded systems with real time and dependability requirements the domain of automation, in *Proc. of 17th IFAC World Congress* (2008), pp. 4144–4149
17. A. Wannagat, B. Vogel-Heuser, Increasing flexibility and availability of manufacturing systems—dynamic reconfiguration of automation software at runtime on sensor faults, in *Proc. of the 9th IFAC Workshop on Intelligent Manufacturing Systems* (2008)

Additional Literature

18. J. Bézivin, Model driven engineering: an emerging technical space. Lect. Notes Comput. Sci. **4143**, 36–64 (2006)
19. J. Bézivin, On the unification power of models. Softw. Syst. Model. **4**, 171–188 (2005)
20. G. Chartrand, Directed graphs as mathematical models, in *Introductory Graph Theory* (Daver, New York, 1985), pp. 16–19
21. Deutsches Institut für Normung, *DIN 25424: Fehlerbaumanalyse (Fault Tree Analysis)* (Beuth, Berlin, 1981), Part 1; (1990), Part 2
22. Deutsches Institut für Normung, *DIN EN 60812: Analysetechniken für die Funktionsfähigkeit von Systemen – Verfahren für die Fehlzustandsart- und -auswirkungsanalyse (FMEA) (Analysis for the Functionality of Systems—Method for the Failure Mode and Effects Analysis)* (Beuth, Berlin, 2006)
23. IEEE Std. 610.12-1990, IEEE standard glossary of software engineering terminology. The Institute of Electrical and Electronics Engineers, USA (1990)
24. ISO/IEC FCD 25010, Systems and software engineering—software product quality requirements and evaluation (SQuaRE)—quality models for software product quality and system quality in use
25. D. Straub, Engineering risk assessment, in *Risk – A Multidisciplinary Introduction*, ed. by C. Klüppelberg, D. Straub, I. Welpe (2014)
26. SysML specification, 06/2012: available at http://www.omg.org/spec/SysML/1.3/PDF/
27. VDA, *Qualitätsmanagement in der Automobilindustrie, Sicherung der Qualität vor Serieneinsatz, System – FMEA (VDA: Quality Management in the Automotive Industry, Securing of Quality Prior to Production*. VDA Brochure, vol. 4, 1st edn. (VDA, Frankfurt am Main, 1996), Part 2
28. VDI/VDE, *Guideline 2653: Agents in Industrial Automation, Part I* (Beuth, Berlin, 2010)

Chapter 14
Information Technology Risks: An Interdisciplinary Challenge

Michael Schermann, Manuel Wiesche, Stefan Hoermann, and Helmut Krcmar

This chapter introduces students to general concepts and theoretical foundations of managing risks induced by developing and using information technology (IT risks). This chapter first provides an overview of the broad nature of IT risks. We introduce categories of IT risks to illustrate its diverse and heterogeneous causes and consequences as well as possible strategies required to balance the risks and benefits of information systems. Second, we illustrate the interdisciplinary challenges that come with managing IT risks on the most researched form of IT risk, namely IT project risks. We discuss the subjectivity of IT risks, various IT risk assessment techniques, outline the process of managing IT project risks, and introduce the dynamics of IT project risks. Third, we present five perspectives on IT risks as a fruitful lens to structure the variety of topics in IT risk research. Using these five perspectives as a framework, we present the most frequently cited IT risk research papers and theories. We conclude with an IT risk research agenda that posits worthwhile avenues for advancing the understanding and control of IT risks.

Keywords Information systems · IT risk · IT risk management · IT projects · Information technology

The Facts

- As information technology (IT) becomes ubiquitous, IT risks become an issue of all stakeholders of an organization. The perspective of the stakeholder determines the impact and magnitude of IT risks. Hence, there is no objective measure for IT risks.
- IT risks come into effect when IT impairs the goals of an organization. For instance, a faulty hard disk is not an IT risk per se until a travel agent is no longer able to book air flights.

M. Schermann (✉) · M. Wiesche · S. Hoermann · H. Krcmar
Chair for Information Systems, Department of Informatics, Technische Universität München, Boltzmannstr. 3, 85748 Garching bei München, Germany
e-mail: Michael.Schermann@in.tum.de

C. Klüppelberg et al. (eds.), *Risk – A Multidisciplinary Introduction*,
DOI 10.1007/978-3-319-04486-6_14,

- The term “IT risk” covers a wide range of issues such as: hacking attacks due to insecure software, loss of revenues due to faulty hardware, legacy systems that make organizations dependent on outdated hardware and software, customers that do not trust electronic commerce websites.
- IT risk research and practice have developed a variety of risk analysis techniques to cover the range of potential IT risks. Checklists help to identify recurring IT risks. Delphi studies support the prioritization of IT risk mitigation measures. Benchmarks illuminate shortcomings in running data processing centers.
- IT risks are bound to a specific situation. A malfunctioning online shop may not have a huge impact at 2 a.m. but consequences may be severe in the weeks before Christmas.
- IT and thus IT risks change at a fast-paced rate. Data leaks due to lost mobile phones or laptops were not an issue several years ago. IT risks are characterized by an arms race between IT risks and mitigation solutions. Again and again, online banking solutions need new security measures.
- The most researched IT risks are IT project risks; that is, risks that occur during the development of new software, hardware, and IT services. Thus, IT project risks serve as an exemplary illustration for the interdisciplinary challenges of handling IT risks.
- IT project risks include technical, social, and organizational aspects. IT projects develop new technology that have unintended side effects. The projects’ progress is impaired by weak customer engagement. Project stakeholders may have conflicting views on the project requirements, which often result in extensive completion delays.
- IT risks propagate through organizations. The strategic goal of reducing expenditures often forces organizations to outsource their IT to IT service providers. The service provider upgrades the outsourced information systems resulting in incompatible interfaces. This lack of control affects the IT enablement of critical business processes and raises new requirements that delay strategic IT projects. Finally, the daily IT operations are impaired by communication barriers and unexpected additional efforts.
- Reflecting the diverse nature of IT risks, IT researchers apply theories from many disciplines. IT investment decisions are grounded in decision-making theory while security risks are resolved by transferring methods from engineering disciplines.

1 Introduction

Information systems are entanglements of information technology (hardware and software), people, and organizations. Our fast-changing and technologically progressing economies, societies, and organizations result in complex risks induced by information technology (IT risks) that we are just beginning to understand [7]. The following examples illustrate the complexity of the nature of IT risks:

- Using the wireless Internet connection, the CEO of an international corporation is finalizing an important email on an upcoming merger in his hotel room. However, being jetlagged, he forgets to establish a secure connection and sends the email over the publicly accessible Internet connection of the hotel. A journalist following the CEO because of rumors about the merger eavesdrops on the Internet traffic of the hotel and intercepts the CEO's email. The content of the email circulates to journalists, analysts, and competitors causing the multi-billion dollar merger to fail.
- The social network Facebook collects information on the private and professional lives of its users to market advertising space on the Facebook platform. However, several privacy issues and concerns have heightened the awareness of potential risks from using Facebook on a private or organizational level. Organizations that prohibit the use of Facebook at the workplace must deal with the efforts their employees use to circumvent technological measures of blocking Facebook.
- The project of constructing a nation-wide billing system in Germany for toll roads that involved satellite-based vehicle tracking was delayed by almost three years and far exceeded the planned budget. Mal-specified and faulty communication and privacy concerns raised by non-governmental agencies caused excessive delays, budget overruns and legal actions. Today, the system is operating very effectively and other countries are interested in adopting it.

Managing risks induced by developing and using information systems has been an on-going challenge for practitioners and researchers alike [8]. The increasing importance of information systems in every aspect of our lives makes IT risk research a highly relevant and fruitful ground for interdisciplinary research. This chapter presents an overview of two important streams of IT risk research. In the first stream, researchers categorize important sources of IT risks, such as IT projects or IT operations [9]. In the second stream, researchers study the important steps of managing IT risks [10]. We illustrate both streams of IT risk research using the example of IT project risk management. Next, we sketch the theoretical foundations of IT risk research. The chapter concludes with a presentation of our thoughts on an agenda for interdisciplinary IT risk research.

2 Sources of IT Risks: Where Do IT Risks Come from?

Figure 1 shows important categories of IT risks as they occur in the various stages of interaction between information systems and business processes. In general, information systems provide the most value if they are aligned with the strategic objectives of the organization.

Strategic IT alignment risks originate from situations and events in which information systems do not align with the strategic objectives of the organization. A prominent example for strategic alignment risk stems from the banking industry.

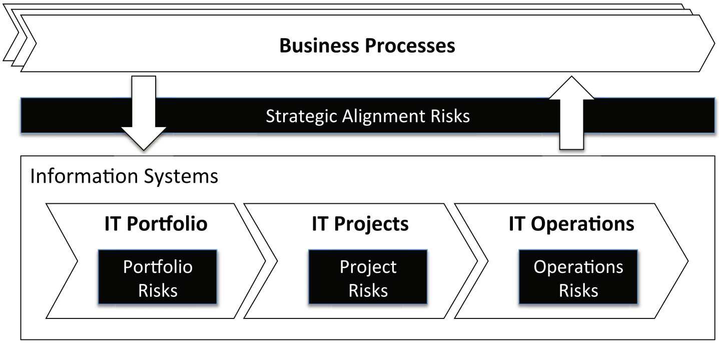


Fig. 1 Sources of IT risks [12]

The advent of mobile banking has drastically changed customers' banking behavior. Most banks struggled with the subsequent business processes because the underlying information systems were inflexible and could not adequately serve these business processes. Typical banking information systems were designed with the highest security standards. This resulted in information systems that were sealed off from the outside world. Customer interaction with these systems was unthinkable. Hence, banks were forced to invest significant sums in the renewal of their information systems. Other strategic alignment risks stem from using (at the time) new technology to support business processes such as IT for e-commerce, financial risk management, support in decision-making, and knowledge management. New information technologies are associated with high uncertainty about the actual capabilities, unintended implications, and their potential business value. For instance, during the rise of e-commerce technologies, risks stemmed from a lack of understanding online consumer behavior [11]. Similar risks are induced by electronic data exchange between organizations and strategic information processing [12].

In contrast to poor strategic decision making, *IT portfolio risks* refer to situations in which the IT department makes bad decisions about what kind of IT should be used and which information systems are necessary to enable business processes. For instance, portfolio risks often arise from outsourcing IT functions [13]. During outsourcing endeavors, organizations usually switch to the information systems of the IT service providers. If future requirements cannot be mapped to these information systems, organizations need to invest in expensive workarounds with poorer performance. For inter-organizational systems, portfolio risks become even more complex and demand cooperation on several levels. This means, organizations need to agree on a shared set of information technologies to establish value chains. More fundamental portfolio risks include IT investment decisions and a missing fit between IT and the corporate culture [14]. For example, while some organizations easily include social networks in their corporate culture, others struggle with deriving value from it.

IT operations risks describe undesired events from a lack of availability, integrity, or confidentiality. Operations risks stem from the failure or misuse of IT [15]. Large-scale invasions by viruses prevent employees from conducting even the most basic duties such as answering emails or receiving purchase orders. Operations risks can be further divided into two categories: new and unknown risks and known but unsolved risks. In known risks, the degree of uncertainty is relatively low and the number of risks occurring is relatively high. This makes it easier to quantify probability and impact of the considered risks. New and unknown risks usually occur with the emergence of new technologies.

IT project risks describe undesired events during designing, developing, and implementing new information systems [16, 17]. For instance, often stakeholders are not able to define a stable set of requirements. Even after beginning the programming of the information systems, stakeholders change requirements. This results in additional programming efforts that delay project completion. We discuss project risks in detail later in the chapter.

3 Steps of IT Risk Management: How Can One Handle IT Risks?

The goals of IT risk research are to understand what causes IT risks, what are the consequences of IT risks, and how does one deal with IT risks in the most effective manner. Figure 2 illustrates five important steps in handling IT risks based on the 'risk thermostat' by Adams [18] and provides a map for the various research areas on IT risk management. The idea of presenting risk management as a 'thermostat' highlights that the activities of risk management (risk identification, risk assessment, risk mitigation, etc.) are highly intertwined and should not be perceived as an ordered process (though it is being presented that way in most of the literature). Furthermore, the 'thermostat' illustrates that risk mitigation activities do not only affect the original perception and assessment of risks but also the originally stated objectives [18]. Hence, risk management should be seen as a tool to balance objectives and risk with appropriate risk mitigation interventions.

For the first step in the process of risk management, researchers study *IT failures* to understand their specific causes. To do so, they develop explanations of why such failures occur and identify indicators that allow practitioners to identify the associated IT risks as early as possible [17]. For instance, delayed and cost-exceeding software development projects occur from employing immature information technologies. Here, an early indicator would be difficulty in procuring project staffing, i.e., the project manager is not able to find software developers that have experience with the particular technology. Unsecure software often originates in development errors or misuse of information systems [15].

A large body of IT risk research focuses on advancing our understanding of and capabilities for *IT risk assessment*. This literature adopts a general definition of risk from other disciplines [19]. IT risks are events with a probability of occurrence and with either an established or estimated negative impact on the objectives of

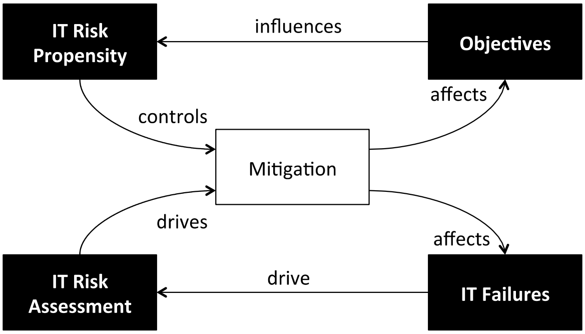


Fig. 2 Steps of IT risk management [18]

stakeholders [20–22]. The challenge with IT risk assessment lies in the subjectivity of IT risks: the stakeholders' perspective determine the impact and the magnitude of IT risks. Hence, it is difficult to establish an objective measure for IT risks.

For the third step in the process of risk management, IT risk research investigates how organizations manage their risk appetite through various levels of *IT risk propensities* that control behavior and decision-making. This stream of research focuses on the integration of risk management in the organization's strategy or strategic decisions on IT through analytical systems or long-term planning and decision support systems. In general, researchers study the decision makers' risk taking behavior [13, 14].

For the fourth step in the process of risk management, IT risk research studies the relationship of risk behavior and the *objectives* of IT endeavors. IT risks come into effect when IT impairs the objectives of an organization. For instance, a faulty hard disk is not an IT risk per se until it hinders a travel agent from booking flights. This literature views IT risks as variations in (often uncertain) outcomes of IT endeavors [11, 23, 24].

The fifth step in the process of risk management, *IT risk mitigation*, is about the design, implementation, and operation measures that help reduce the probability or the impact of IT risks. Here, the major challenges stem from integrating these measures in the business processes. Usually, risk mitigation measures such as entering passwords or using encryptions are perceived as burdensome. Hence, raising the security awareness and ensuring compliance with risk mitigation measures is pivotal in this step of handling IT risks [15].

In sum, the five steps of IT risk management present important fields for studying risks in IT and highlight the intertwined and complex nature of IT risk. The structure of Fig. 2 highlights the dynamics of IT risks and risk management. Effective risk mitigation activities are highly dependent on contextual factors. The variety and interplay between the four perspectives illustrates the challenges of understanding and establishing effective risk mitigation mechanisms in organizations [25]. In the next chapter, we will illustrate these steps using the example of IT project risk management.

4 The Example of IT Project Risk Management

IT project risk management is the most prominent stream of research in IT risk research. Hence, herewith follows an in-depth presentation on the state of knowledge on this topic.

Identifying Causes and Explanations of Failure: The Subjectivity of IT Project Risk The research to date on project failures is inconclusive. The well-known and widely cited Standish Group [26] report that around 68 % of the sampled IT projects are considered as failures (24 %) or challenged (44 %) in regards to either budget, completion schedule, or scope. Other researchers report different results. Sauer et al. [27], for example, find that about 67 % of the analyzed projects met budget, schedule and scope expectations. Based on the common understanding that risk denotes the probability and the loss associated with an unsatisfactory outcome (e.g., [21]), the question arises, 'What exactly renders an outcome unsatisfactory?'. The answer to this question largely depends on the respective stakeholder's expectation or objectives concerning the project. Stakeholders typically comprise the project manager, the project team members, the customer, the user, and the project sponsors. Depending on which perspective one takes, objectives, unsatisfactory outcomes, and thus risks, can vary. For instance, a software development project manager might strive for schedule, budget and scope objectives whereas the customer considers a high user acceptance rate more important. Similarly, for the project manager, an unsatisfactory outcome might be schedule and budget overrun or scope constraints (e.g. unstable requirements) while for the customer unsatisfactory outcomes refer to anything that impedes user acceptance (e.g. an unintuitive graphical user interface). In sum, the multidimensional nature of project success drives our understanding of risk [28]. The perspective of stakeholders determines the impact and magnitude of IT risks.

Assessing the Technical, Social, and Organizational Domains of IT Project Risks The literature describes project risks by grouping them according to common characteristics [8]. This grouping enables researchers to establish checklists of common risks. Although discussed controversially in literature, such checklists provide an easy and low cost approach to identifying risks in a project and are thus popular in research and practice. Table 1 shows a sample of existing studies on IT project risks.

Risks in IT projects can be grouped into three risk domains: the social subsystem, the technical subsystem, and the organizational subsystem. While the latter domain refers to the project management capabilities of the project team and the planning/control techniques applied by the project manager, the social subsystem domain comprises an unstable or highly political social context and users unable or not willing to contribute to project success. The technical subsystem domain captures risks related to unstable requirements, high project complexity and new or unfamiliar technology.

Figure 3 shows empirical evidence on how IT project risks affect the success of the project in terms of process performance (How well does the project

Table 1 Common risks in IT projects ranked by importance [30]

Rank	Schmidt et al. [9]		Kappelman et al. [29]		Hoermann et al. [30]	
1	Lack of effective project management skills	P	Lack of top management support	S	Inadequate technical infrastructure	T
2	Lack of top management commitment	S	Lack of documented requirements	P	Customer expectations	S
3	Lack of required skills in project personnel	P	Weak project manager	P	Core development dependencies	T
4	Not managing change properly	P	No change control process (change management)	P	Complex system architecture	T
5	No planning or inadequate planning	P	No stakeholder involvement and/or participation	S	Post go live approach not defined	P
6	Misunderstanding the requirements	P	Ineffective schedule planning and/or management	P	Customer financial obligations	S
7	Artificial deadlines	P	Weak commitment of project team	P	Expected performance issues	T
8	Failure to gain user commitment	S	Communication breakdown among stakeholders	S	Customer inability to undertake project	S
9	Lack of frozen requirements	P	Team members lack requisite knowledge and/or skills	P	Non-T&M payment terms	S
10	Lack of people skills in project leadership	P	Subject matter experts are overscheduled	P	Functionality gaps	T

T: Technical subsystem, S: Social subsystem, P: Project management subsystem

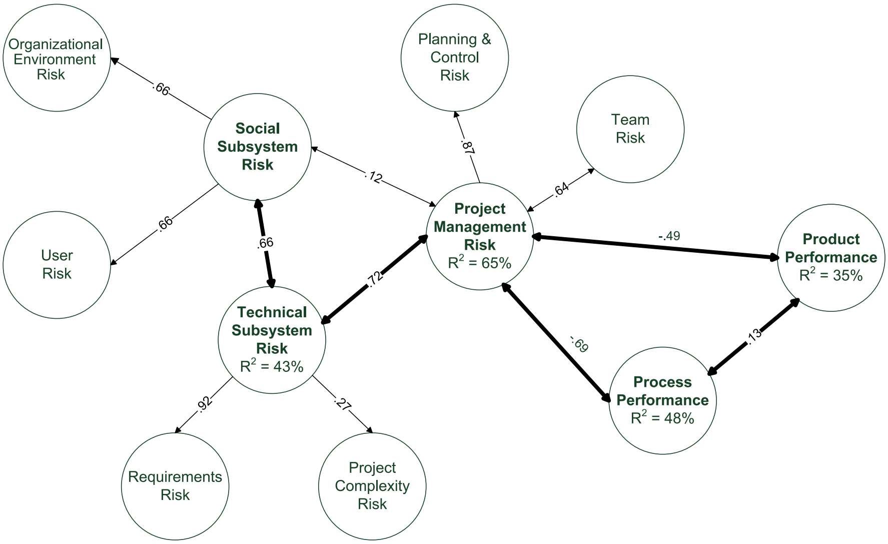

Organizational Environment Risk
.66
Social Subsystem Risk
.66
User Risk
.66
Planning & Control Risk
.87
Team Risk
.64
.12
Project Management Risk R² = 65%
-.49
Product Performance R² = 35%
.72
Technical Subsystem Risk R² = 43%
-.69
.13
Process Performance R² = 48%
.92
.27
Requirements Risk
Project Complexity Risk

Fig. 3 Effects of IT risk domains on project performance [31]

proceed?) and product performance (How well does the result match the objectives?).

Mitigating IT Projects Risks: Towards Standards of IT Project Risk Management Most authors acknowledge risk management as an integral part of IT projects, especially when it comes to managing large and complex projects. Boehm [21] introduces the concept of risk exposure (defined as probability and impact of an unsatisfactory outcome) to software development projects and characterizes risk management as a process comprising the six steps: risk identification, risk analysis, risk prioritization, risk management planning, risk resolution, and risk monitoring (see Fig. 4).

Instead of describing the risk management process in detail, Lyytinen, Mathiassen, and Ropponen [16] provide a framework to evaluate project risk management approaches as a distinct form of organizational behavior. The framework comprises three distinct environments (the management environment, the project environment, and the system environment), which are linked by the (risk) management process and the development process and help to organize risk management activities in a systematic and comprehensive way.

Understanding the Risk Propensity: The Dynamics of IT Project Risks In addition to the question which risks appear in IT projects and how can these risks be organized, the question of when they appear and how they evolve is also of substantial interest to IT project managers and researchers. Alter et al. [20] discuss several potential limitations of extant research on IT project risk, one of them being the 'frequent omission of the temporal nature of risk'. As the authors state, risks are likely to have different temporal patterns; not only their importance but also the points of time at which they occur can vary over the project life cycle.

In an earlier study, Alter et al. [8] studied the temporal aspect of IT project risks and suggested that linking them to project phases and consequently adapting project risk management increases the likelihood of successful IT projects. The authors identify eight risks and allocate them to seven project phases depending on when their effects become apparent. The identified risks include: non-existent or unwilling users; multiple users and designers; disappearing users, designers or maintainers; inability to specify the purpose or usage pattern in advance; lack or loss of support; lack of prior experience with similar systems; inability to predict and cushion the impact on all parties; and technical problems or cost-effectiveness issues. Alter et al. [8] map these risks to particular project phases and propose several risk-reducing strategies.

In a more recent study, Gemino et al. [32] introduce a temporal model of IT project performance that classifies IT project risks into a priori risks and emergent risks. While the a priori risks are associated with structural elements of the project and with knowledge resources available to the project team, emergent risks denote deficiencies in organizational support or result from the volatility of projects. A project manager might estimate a priori risks before the start of the project; emergent risks only become apparent during particular project phases. Using structural

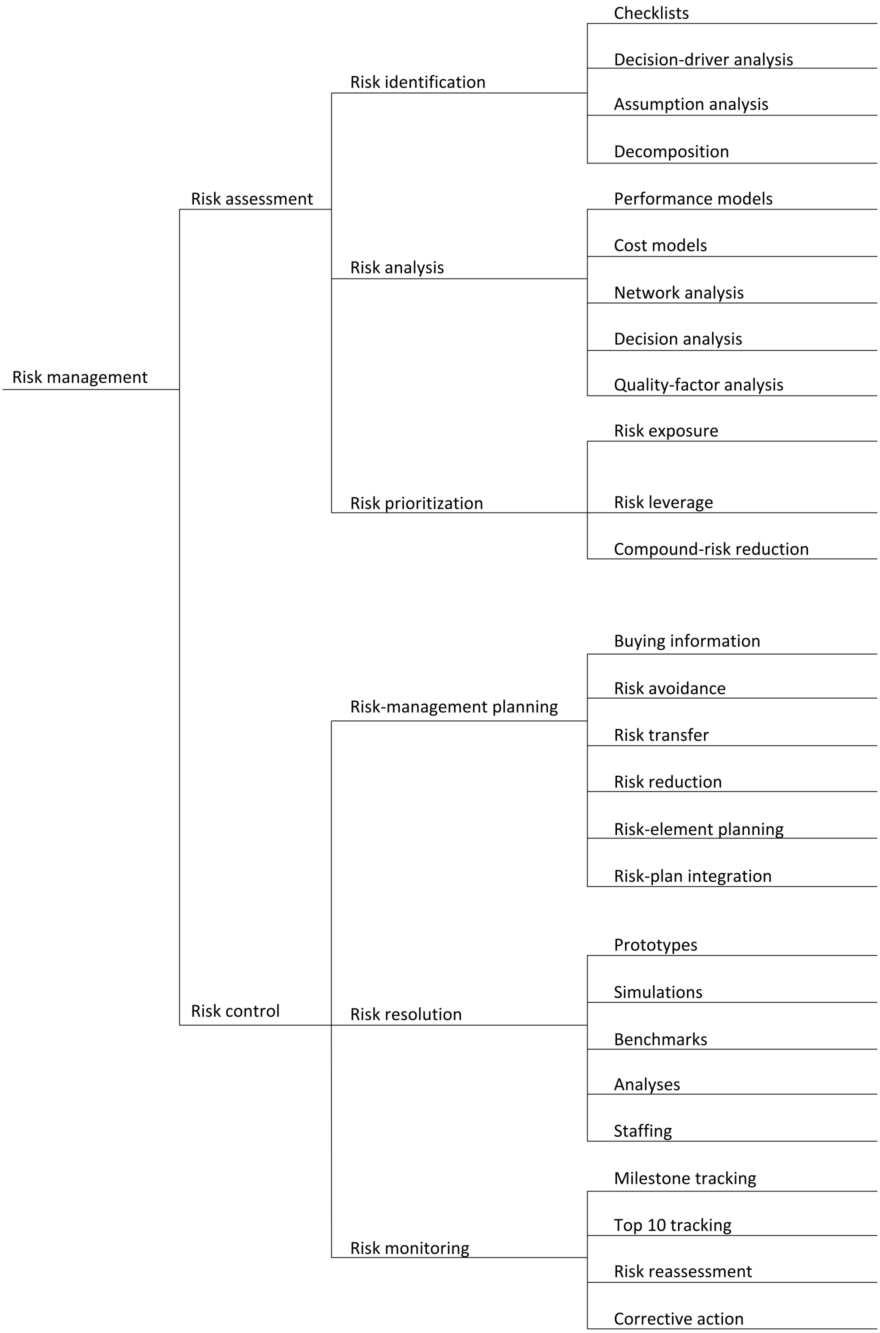


Fig. 4 Project risk management [21]

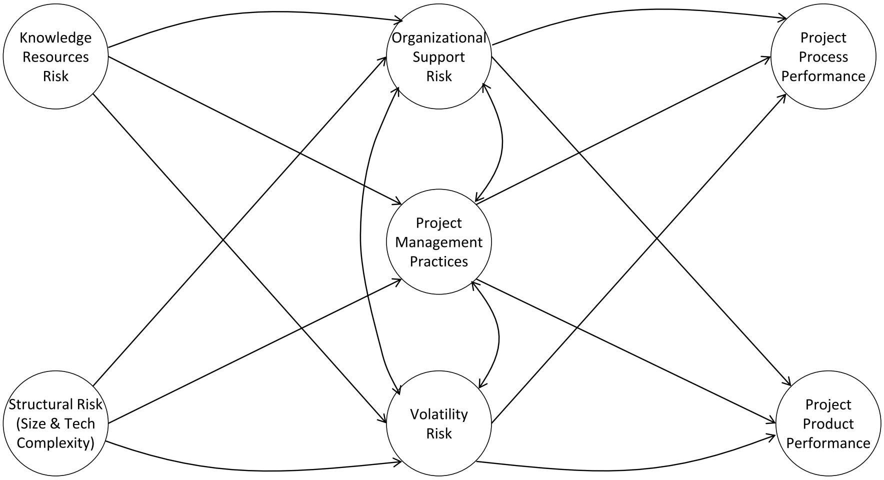

Knowledge Resources Risk
Organizational Support Risk
Project Process Performance
Project Management Practices
Structural Risk (Size & Tech Complexity)
Volatility Risk
Project Product Performance

Fig. 5 The temporal model of IT project performance [32]

equation modeling, the authors show that their model offers an improved explanatory power over traditional models of performance, partly resulting from the temporal perspective on IT project risks (see Fig. 5).

Safeguarding the Organizations' Objectives: The Benefit of IT Project Risk Management The benefits of project risk management are difficult to express in financial or other quantitative terms [33]. This means that practitioners of project risk management usually need to justify any effort associated with risk management. Other stakeholders often perceive risk management as an effort that comes on top of an already heavy operational workload. Subsequently, they try to resist or even avoid risk management. A considerable amount of research attempts to provide an empirical verification of the benefits of project risk management (e.g. [17, 24, 31]). Barki et al. [34], derive a contingency model, and hypothesize that project success is affected by the fit between the project's risk profile and its risk management profile. These authors conducted a survey of IT project managers to assess 75 Canadian IT projects in terms of system quality and cost gap (constructs for project success), as well as internal integration, user participation, and formal planning (constructs for the risk management practices). A project's risk exposure can be assessed using Barki et al.'s [23] instrument which comprises 23 risk variables. Analysis of a correlation between the degree of fit between a project's risk profile and its risk management profile and the performance measures indicates that projects that better adapt to their degree of risk exposure usually perform better. In a methodologically quite different action research approach, Baskerville and Stage [35] apply risk analysis to improve the managerial control over prototyping projects. By defining risks, specifying their consequences, assigning priorities, and selecting resolution strategies, the authors suggest that risk management can help improve the communication among users and developers, point out difficulties in maintaining the original project

Table 2 The 10 most-cited publications on IT risk

Citation	Discipline	Domain	Focus
Boehm, 1991 [21]	CS	IT projects	Risk management
Jarvenpaa et al., 2000 [38]	IT	E-commerce	Consumer trust
McFarlan, 1981 [36]	IT	IT projects	Risk management
Alter and Ginzberg, 1978 [8]	IT	IT projects	Risk models
Pavlou, 2003 [39]	IT	E-commerce	Consumer trust
Charette, 1989 [22]	CS	IT projects	Risk management
Barki et al., 1993 [23]	IT	IT projects	Risk factors
Nidumolu, 1995 [24]	IT	IT projects	Risk models
Keil et al., 1998 [37]	IT	IT projects	Risk factors
McKnight et al., 2003 [40]	IT	E-commerce	Consumer trust

CS = Computer science; IT = Information systems research

plan, and get a clearer picture on the status of the project. Using this approach, the authors reported few if any disruptions from the identified risks during the course of the project.

5 Theoretical Foundations of IT Risk

In this section, we discuss the most-cited publications as of April 2011 as a starting point for students who wish to explore IT risk research. On the one hand, reflecting the diverse nature of IT risks, IT researchers apply theories from many disciplines. On the other hand, the field of IT risk research is still very young and thus lacks original theories. This makes IT risk research a very promising ground for interdisciplinary research. The following sections serve as starting points to IT risk research.

Starting Point: The 10 Most-Cited Publications on IT Risk One of the most prominent publications on IT risks originates from the discipline of Computer Science (CS) (see Table 2). Boehm's [21] publication on risks in software development practices provides risk examples, recommendations for best practice, and principles for effective risk management to prevent software project disasters. Other pieces of research address the separate and aggregated assessment of project risks to ensure proper decision-making [36] and strategies for coping with uncertainty in management information systems development projects [8, 22]. Extensive research exists on the effects of coordination mechanisms and risk drivers on project performance [24], the detailed elements, which influence failure in developing systems (tasks, structure, technology, and actors), and lists of software risk factors and mitigation strategies for specific risks [23]. In the field of risk factors, IT project research concentrates on the effects of risk management and environmental factors on risk components, determination and prioritization of risk lists in IT projects [37].

Table 3 Most-cited theories in IT risk research

Citation	Domain	Topic
Williamson, 1979 [41]	Transaction cost economics	Transaction costs vary for firms, markets, and contractors depending on situational setting
Mayer et al., 1995 [44]	Trust	Propose a model of antecedents and outcomes of trust, incorporating trustor, trustee, and the role of uncertainty
Boehm, 1988 [45]	Software development	Proposes a spiral model for software development which consists of four phases of activities and incorporates elements of specification- and prototype-driven processes
Davis, 1982 [46]	Technology acceptance	Develops strategies for determining requirements for IT development on both an organizational and individual level
Grover et al., 1996 [43]	Outsourcing	Determine the importance of service quality and partnership within outsourcing relationships
DeLone and McLean, 1992 [42]	IT Success	Propose an integrated model of information systems success, including the impact of system quality and information quality on organizations
Ang and Straub, 1998 [47]	Outsourcing	Identify the economic determinants of IT outsourcing to incorporate outsourcing decisions in the strategy of an organization
Ganesan, 1994 [48]	Buyer-seller relationships	Finds mutual dependence and trust as determining factors for marketing endeavors under a given timely horizon
Zucker, 1986 [49]	Trust	Discusses processes, contingencies, and institutions as central elements of trust production
Zmud, 1986 [50]	Software development	Develops an approach for staffing, planning, and controlling software development
Akerlof, 1970 [51]	Buyer-seller relationships	Discusses the role of information uncertainty regarding quality heterogeneity in buyer-seller relationships
March and Simon, 1958 [52]	Organization theory	Discuss the motivational and affective aspects of human behavior, and cognition processes in organizations

The second area of research concentrates on the role of IT for on-line transactions. Existing research provides various perspectives on the role of consumer trust in e-commerce transactions. Research provides four high-level constructs: disposition to trust, institution-based trust, trusting beliefs, and trusting intentions for developing and empirically validating measures for a multidisciplinary and multidimensional model of trust in e-commerce [40]. Research also exists on the role of organizational size and popularity on trustworthiness and risk perception [38], and intention to transact and on-line transaction behavior as key drivers for engaging consumers in on-line transactions [39].

Starting Point: The Core Theories of IT Risk IT risk research is grounded in an interdisciplinary set of theories from organizational behavior, management, and

IT. Examples of such theories are thoughts on transaction cost economics which propose: transaction costs vary for firms, markets, and contractors depending on the situational setting [41], IT success is a multi-dimensional construct including influencing factors such as system quality and information quality [42], and the importance of service quality and partnership within outsourcing relationships [43]. Such theories represent the historical development of IT as a socio-economical discipline. Table 3 provides an overview of the most-cited theories.

6 Towards Interdisciplinary IT Risk Research

We return to the examples given in the introduction to illustrate the multidimensional character of IT risks. The impact of IT and the associated risk are continuously produced and reinterpreted on all levels of society. In the first example, the CEO chose convenience over security by sending confidential emails over an unsecured network. Through deliberately ignoring security advice, the CEO renders as useless the risk mitigation strategies of his company. In his work setting at a hotel, the risk of eavesdropping conflicted with achieving his objective, which was to communicate a message intended to be received by a specific, targeted group.

Facebook accumulates mass data through continuously adding new functionality, such as geo-coding of messages, and people begin using it in unanticipated ways, such as organizing political uprisings. Hence, governments and institutions have begun to criticize Facebook because of either privacy concerns or a sense of loss of control. Despite any real or perceived issues of privacy, people and organizations increasingly use Facebook to communicate. Similarly, appropriate mitigation strategies are the temporary result of agreement among many stakeholders within and across an organization. In the case of the billing systems for road tolls, the project should be considered a total failure according to typical project success measures. However, the steady governmental income and the ease of use of the system on an organizational level have led to reinterpretations of the project. In light of the system's success, even the privacy concerns on a societal level took a back seat in the public discussion.

To cover these aspects of IT risks, a multi-disciplinary body of theory is necessary. Therefore, we identified and reviewed publications on risk outside of the IT discipline. We analyzed these publications using qualitative data analysis and present these central publications on risk and discuss their potential for advancing IT risk research.

What Are Elements of IT Risks? Other disciplines discuss the fundamental elements of risk in great detail. For instance, Kahneman and Tversky ([53], cited 572 times per year) theorize about biases and the role of heuristics in individual risk perception. On a societal level, Beck ([54], cited 592 times per year) analyzes the structures and social systems of communicating risks as well as reaching societal consensus on risks. Still, research on the elements of risks in information systems provides promising ground for advancing a commonly shared understanding of IT risk. This issue is highlighted by the fact that no established and commonly shared definitions of "IT risk" exist [20].

What Are Measures of IT Risks? Measures of risks are an enduring topic in other disciplines. Artzner et al. ([55], cited 226 times per year) develop risk measures for financial markets. Similarly, Sharpe et al. ([56], cited 169 times per year) measure the effect of adding assets to a financial portfolio. By contrast, Slovic ([57], cited 142 times per year) explores contortions in measuring risk perception in groups. Kahneman and Tversky ([53], cited 572 times per year) show that utility is an inappropriate measure for risks. Although many authors question the applicability of financial risk measures to IT risks [58], some IT authors show their applicability in the domains of contract portfolios of IT services [59]. The research of Slovic [57] and Kahneman and Tversky [53] provides valuable insights into measuring qualitative risk as it is often suggested in the IT project management literature.

What Are Acceptable IT Risks? Given limited resources for risk mitigation, an important challenge in risk management is determining acceptable levels of risk. Here again, other disciplines provide promising trains of thought. Jorion ([60], cited 206 times per year) introduces the value at risk measure to determine acceptable levels of risk in the financial domain. Criticism against transferability to other domains has been expressed [58]. However, IT researchers have begun to explore the use of value of risk to determine acceptable levels of project risks [61]. Research on the social acceptability of risks offers valuable insight on risk. Douglas ([62], cited 148 times per year) explores the collaborative interpretation of acceptable risks by diverging stakeholders. IT researchers increasingly argue that strategic IT decisions under risk, successful IT projects, and collaboration risks in IT need to evolve from a single dimensional (shareholder) perspective to a multi-dimensional (stakeholder) perspective. Using the body of knowledge on risk research in sociology and thoughts on acceptable risk could provide a fresh perspective and help to develop theories with potential to bring about significant progress in risk research in IT.

What Are the Benefits of Risky Behavior with IT? Knight ([19], first edition from 1921), has been cited 732 times per year across disciplines, which makes the publication one of the fundamental and most-influential publications on risk. Knight's [19] main argument is that coping with unknown risks determines the success of economic organizations. Thus, organizations that mitigate risks effectively are able to allocate more resources to dealing with uncertain issues. Zuckerman ([63], cited 483 times per year) explores the psychological mechanisms for taking risks. His view provides a fresh perspective on risk for the IT discipline where risk is commonly associated with negative effects, failures, and loss (e.g. [23]). Beck ([64], cited 170 times per year) analyzes the potentials of transparent and open societal processes that construct shared understanding of risk and uncertainty. Research in IT could fundamentally benefit by incorporating the notion of uncertainty in risk research. This would shift the focus from risk exposure as a basis of decision making to situations where the probability distribution of a random outcome is unknown. Measures could be developed to cope with new and unknown risks effectively, such as through early warning systems. Unfortunately, many risk incidents incorporate a high degree of uncertainty and often lack the necessary number of empirical incidents to soundly predict the underlying distribution.

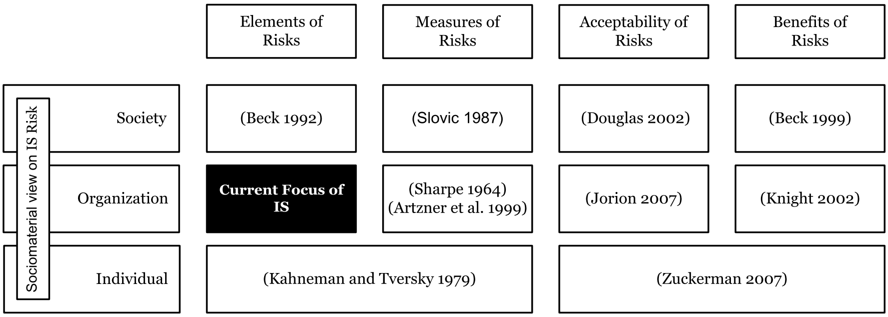
In sum, this chapter provides an overview of IT risk research, outlines the existing body of knowledge on IT risk research, and identifies promising areas for future research. With information systems becoming ubiquitous, IT risks permeate every aspect of life and effective risk mitigation increasingly requires an interdisciplinary approach.

7 Food for Thought

- Collect IT risks from newspapers and press releases. Identify what caused the IT risk, what mitigation activities where taken, and what was the damage or loss of the project.
- Consider the case of a faulty airline check-in system that was not online for a day and a half. The faulty system caused quite a stir among customers and the press but an analysis two months after the incident showed that the actual damage was way below €250,000. Discuss and develop an explanation.
- Discuss the statement of a CIO of a major corporation: “IT risks are a daily issue but without IT risks I would be afraid we would be behind our competition”.
- Discuss the case of the billing systems for road tolls. First, stakeholders, press, and public opinion considered the project to be a total failure. Two years after the project was completed, the steady incomes on the governmental level as well as the system’s ease of use have led to reinterpretations of the project. Today the system is being exported to other countries.
- Develop an IT risk assessment for the risk of hackers entering the billing system of a large online shopping system and stealing 100,000 sets of credit card information. Develop the risk assessment from the perspective of a person affected by this incident and from the perspective of the provider of the online shopping systems.

8 Summary

To operationalize the advancement of IT risk research, we first conceptualize three levels of research inquiry as one dimension of a research agenda. On the individual level, risk research focuses on the mechanisms of risk perception and the subjective assessment of risks. On the organizational level, risk research focuses on



	Elements of Risks	Measures of Risks	Acceptability of Risks	Benefits of Risks
Society	(Beck 1992)	(Slovic 1987)	(Douglas 2002)	(Beck 1999)
Organization	**Current Focus of IS**	(Sharpe 1964) (Artzner et al. 1999)	(Jorion 2007)	(Knight 2002)
Individual	(Kahneman and Tversky 1979)		(Zuckerman 2007)	

Fig. 6 Starting points for interdisciplinary IT risk research

managing risks as a function to achieve organizational goals. On a societal level, risk research focuses on the social construction processes that lead to either consensual or conflicting norms and practices for coping with risks. The other dimension of the research agenda consists of the four bodies of theoretical foundations of risk research, which we discussed above. Figure 6 shows the research agenda along with seminal publications as starting points toward interdisciplinary IT risk research.

References

Selected Bibliography

1. S. Alter, S. Sherer, A general, but readily adaptable model of information system risk. Commun. AIS **2004**(14), 1–28 (2004)
2. H. Barki, S. Rivard, J. Talbot, Toward an assessment of software development risk. J. Manag. Inf. Syst. **10**(2), 203–225 (1993)
3. R. Charette, *Software Engineering Risk Analysis and Management* (Multiscience Press, New York, 1989)
4. R.K. Rainer Jr., C.A. Snyder, H.H. Carr, Risk analysis for information technology. J. Manag. Inf. Syst. **8**(1), 129–147 (1991)
5. D.W. Straub, R.J. Welke, Coping with systems risk: security planning models for management decision making. MIS Q. **22**(4), 441–469 (1998)
6. L. Wallace, M. Keil, A. Rai, How software project risk affects project performance: an investigation of the dimensions of risk and an exploratory model. Decis. Sci. **35**(2), 289–321 (2004)

Additional Literature

7. M. Wiesche et al., Classifying information systems risks: what have we learned so far? in *46th Hawaii International Conference on Systems Science (HICSS 2013)*, Maui, HI, USA (2013)

8. S. Alter, M. Ginzberg, Managing uncertainty in MIS implementation. Sloan Manag. Rev. **20**(1), 23–31 (1978)
9. R. Schmidt et al., Identifying software project risks: an international Delphi study. J. Manag. Inf. Syst. **17**, 5–36 (2001)
10. J.R.K. Rainer, C.A. Snyder, H.H. Carr, Risk analysis for information technology. J. Manag. Inf. Syst. **8**(1), 129–147 (1991)
11. P.A. Pavlou, D. Gefen, Psychological contract violation in online marketplaces: antecedents, consequences, and moderating role. Inf. Syst. Res. **16**(4), 372–399 (2005)
12. M. Junginger, *Wertorientierte Steuerung von Risiken im Informationsmanagement* (Universität Hohenheim, Stuttgart, 2004)
13. C.L. Iacovou, R. Nakatsu, A risk profile of offshore-outsourced development projects. Commun. ACM **51**(6), 89–94 (2008)
14. M. Benaroch, Y. Lichtenstein, K. Robinson, Real options in information technology risk management: an empirical validation of risk-option relationships. MIS Q. **30**(4), 827–864 (2006)
15. D.W. Straub, R.J. Welke, Coping with systems risk: security planning models for management decision making. MIS Q. **22**(4), 441–469 (1998)
16. K. Lyytinen, L. Mathiassen, J. Ropponen, A framework for software risk management. J. Inf. Technol. **11**(4), 275–285 (1996)
17. J. Ropponen, K. Lyytinen, Can software risk management improve system development: an exploratory study. Eur. J. Inf. Syst. **6**(1), 41 (1997)
18. J. Adams, *Risk* (Routledge, Oxford, 1995)
19. F.H. Knight, *Risk, Uncertainty and Profit* (BeardBooks, Washington, 2002)
20. S. Alter, S. Sherer, A general, but readily adaptable model of information system risk. Commun. AIS **2004**(14), 1–28 (2004)
21. B. Boehm, Software risk management: principles and practices. IEEE Softw. **8**(1), 32–41 (1991)
22. R. Charette, *Software Engineering Risk Analysis and Management* (Multiscience Press, New York, 1989)
23. H. Barki, S. Rivard, J. Talbot, Toward an assessment of software development risk. J. Manag. Inf. Syst. **10**(2), 203–225 (1993)
24. S. Nidumolu, The effect of coordination and uncertainty on software project performance: residual performance risk as an intervening variable. Inf. Syst. Res. **6**(3), 191 (1995)
25. M. Schermann, *Risk Service Engineering: Informationsmodelle für das Risikomanagement* (Gabler, Wiesbaden, 2011)
26. The Standish Group, *CHAOS Summary for 2010* (The Standish Group, Boston, 2010)
27. C. Sauer, A. Gemino, B. Reich, The impact of size and volatility on IT project performance. Commun. ACM **50**(11), 79–84 (2007)
28. A. Shenhar et al., Project success: a multidimensional strategic concept. Long Range Plan. **34**(6), 699–725 (2001)
29. L. Kappelman, R. McKeeman, L. Zhang, Early warning signs of IT project failure: the dominant dozen. Int. J. Proj. Manag. **23**, 31–37 (2006)
30. S. Hoermann, M. Schermann, H. Krcmar, Towards understanding the relative importance of risk factors in IS projects. A quantitative perspective, in *18th European Conference on Information Systems*, Pretoria, South Africa (2010)
31. L. Wallace, M. Keil, A. Rai, How software project risk affects project performance: an investigation of the dimensions of risk and an exploratory model. Decis. Sci. **35**(2), 289–321 (2004)
32. A. Gemino, B. Reich, C. Sauer, A temporal model of information technology project performance. J. Manag. Inf. Syst. **24**(3), 9–44 (2007)
33. K. de Bakker, A. Boonstra, H. Wortmann, Does risk management contribute to IT project success? A meta-analysis of empirical evidence. Int. J. Proj. Manag. **28**(5), 493–503 (2010)
34. H. Barki, S. Rivard, J. Talbot, An integrative contingency model of software project risk management. J. Manag. Inf. Syst. **17**(4), 37–69 (2001)

35. R. Baskerville, J. Stage, Controlling prototype development through risk analysis. MIS Q. **20**(4), 481–504 (1996)
36. F.W. McFarlan, Portfolio approach to information systems. Harv. Bus. Rev. **59**(5), 142–151 (1981)
37. M. Keil et al., A framework for identifying software project risks. Commun. ACM **41**(11), 76–83 (1998)
38. S.L. Jarvenpaa, N. Tractinsky, M. Vitale, Consumer trust in an Internet store. Inf. Technol. Manag. **1**(1–2), 45–71 (2000)
39. P.A. Pavlou, Consumer acceptance of electronic commerce: integrating trust and risk with the technology acceptance model. Int. J. Electron. Commer. **7**(3), 101–134 (2003)
40. D.H. McKnight, V. Choudhury, C. Kacmar, Developing and validating trust measures for e-commerce: an integrative typology. Inf. Syst. Res. **13**(3), 334–359 (2003)
41. O.E. Williamson, Transaction-cost economics: the governance of contractual relations. J. Law Econ. **22**(2), 1–30 (1979)
42. W.H. DeLone, E.R. McLean, Information systems success: the quest for the dependent variable. Inf. Syst. Res. **3**(1), 60–95 (1992)
43. V. Grover, M.J. Cheon, J.T.C. Teng, The effect of service quality and partnership on the outsourcing of information systems functions. J. Manag. Inf. Syst. **12**(4), 89–116 (1996)
44. R.C. Mayer, J.H. Davis, F.D. Schoorman, An integrative model of organizational trust. Acad. Manag. Rev. **20**(3), 709–734 (1995)
45. B.W. Boehm, A spiral model of software development and enhancement. IEEE Comput. **21**(5), 61–72 (1988)
46. G.B. Davis, Strategies for information requirements determination. IBM Syst. J. **21**(1), 4–30 (1982)
47. S. Ang, D. Straub, Production and transaction economies and IS outsourcing: a study of the US banking industry. MIS Q. **22**(4), 535–552 (1998)
48. S. Ganesan, Determinants of long-term orientation in buyer-seller relationships. J. Mark. **58**(2), 1–19 (1994)
49. L.G. Zucker, Production of trust: institutional sources of economic structure, 1840–1920. Res. Organ. Behav. **8**, 53–111 (1986)
50. R. Zmud, Management of large software development efforts. MIS Q. **4**(2), 45–55 (1980)
51. G.A. Akerlof, The market for "lemons": quality uncertainty and the market mechanism. Q. J. Econ. **84**(3), 488–500 (1970)
52. J. March, H. Simon, *Organizations* (Wiley, New York, 1958)
53. D. Kahneman, A. Tversky, Prospect theory: an analysis of decision under risk. Econom., J. Econom. Soc. **47**(2), 263–291 (1979)
54. U. Beck, *Risk Society: Towards a New Modernity* (Sage, Frankfurt am Main, 1992)
55. P. Artzner et al., Coherent measures of risk. Math. Finance **9**(3), 203–228 (1999)
56. W.F. Sharpe, Capital asset prices: a theory of market equilibrium under conditions of risk. J. Finance **19**(3), 425–442 (1964)
57. P. Slovic, Perception of risk. Science **236**(4799), 280 (1987)
58. D.B. Parker, Risks of risk-based security. Commun. ACM **50**(3), 120 (2007)
59. R.J. Kauffman, R. Sougstad, Risk management of contract portfolios in IT services: the profit-at-risk approach. J. Manag. Inf. Syst. **25**(1), 17–48 (2008)
60. P. Jorion, *Value at Risk: The New Benchmark for Managing Financial Risk*, vol. 2 (McGraw-Hill, New York, 2007)
61. M. Sutter et al., Calculating the conditional value at risk in IS projects: towards a single measure of project risk, in *19th European Conference on Information Systems (ECIS)*, Helsinki, Finland (2011)
62. M. Douglas, *Risk and Blame: Essays in Cultural Theory* (Routledge, New York, 2002)
63. M. Zuckerman, *Sensation Seeking and Risk* (American Psychological Association, Washington, 2007)
64. U. Beck, *World Risk Society* (Polity Press, Cambridge, 1999)

Chapter 15
Risk Issues in Developing Novel User Interfaces for Human-Computer Interaction

Gudrun Klinker, Manuel Huber, and Marcus Tönnis

When new user interfaces or information visualization schemes are developed for complex information processing systems, it is not readily clear how much they do, in fact, support and improve users' understanding and use of such systems. Is a new interface better than an older one? In what respect, and in which situations? To provide answers to such questions, user testing schemes are employed. This chapter reports on a range of risks pertaining to the design and implementation of user interfaces in general, and to newly emerging interfaces (3-dimensionally, immersive, mobile) in particular.

Keywords Human-computer interaction · Design-risk · Miscommunication risk · Augmented reality · Usability testing

The Facts

- In our modern society, much of what we do is at least partially supported, guided or influenced by information stored, simulated or analyzed in computers.
- It is important that everybody is able to use and understand such virtual information without mental or physical barriers.
- Research in human-computer interaction strives towards finding suitable interaction paradigms that support people in using computer information in their daily activities—for both personal and professional use.
- Research on suitable interaction metaphors analyzes risks of miscommunication between humans and computers.
- Potential miscommunication can be a challenge (e.g., in games), a nuisance (in uncritical situations) or a physical danger (in life-critical situations).

G. Klinker (✉) · M. Huber · M. Tönnis
Fachgebiet Augmented Reality, Department of Informatics, Technische Universität München,
Boltzmannstr. 3, 85748 Garching bei München, Germany
e-mail: klinker@in.tum.de

C. Klüppelberg et al. (eds.), *Risk – A Multidisciplinary Introduction*,
DOI 10.1007/978-3-319-04486-6_15,

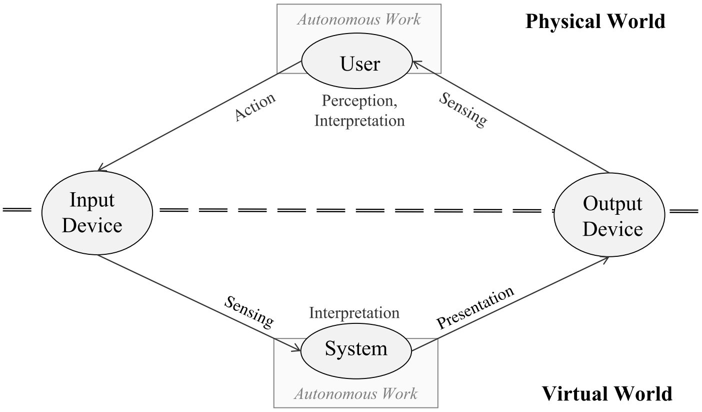


Fig. 1 Cycle of interaction based on human and computer-based perception, interpretation and action/presentation (adapted and extended from [2])

1 Introduction

When users interact with computer systems, they, as well as their real, physical environment, get in contact with the virtual world in the computer, as shown in the interaction cyle adapted from Bowman et al. [2] in Fig. 1. At the interface between the physical and virtual worlds are input and output devices that sense human actions via dedicated sensors as input signals, interpret and act upon them before rendering suitable output signals on displays. These, in turn, are perceived by the users and interpreted before a new interaction cycle starts.

Risks of misperception, misinterpretation and mispresentation exist at all stages in this interaction cycle. Users may not know well enough what actions they have to perform and how carefully they need to act them out such that the system can decipher them unambiguously. Sensor systems suffer from noise and various physical limitations. Furthermore, interpretation algorithms may be lacking some of the physical-world context when they analyze their input data, resulting in false positive and negative decisions in their command recognitions. The subsequently generated visualizations may fall short of representing the wealth of available information with appropriate clarity and detail on the available display hardware. Users may overlook important issues in the visualizations, or they may draw wrong conclusions because they are not familiar with the metaphors that were used.

For these reasons, human-computer interfaces need to be tested thoroughly and repeatedly to minimize the risk of miscommunication. In user-centered approaches, various different testing methods are applied throughout the entire product design and development life cycle. Yet, such testing has its own set of risky fallacies. The subsequent sections address each of these issues in detail. We begin with a brief description of a few current developments.

2 Examples

In recent years, user interfaces have progressed rapidly. They move away from the well-established WIMP[1] style of the Desktop metaphor that provides direct manipulation on a raster display, as described in the seminal text book by Shneiderman and Plaisant [13] towards highly immersive, multi-modal and multi-media, ubiquitous or mobile multi-touch-based interfaces (see, for example, Myers, Hudson, and Pausch [43] for further reading). Technological advances, regarding speed, resolution and accuracy of sensing devices, have recently triggered a number of novel user interface schemes to find their way into commodity devices, such as smartphones and game consoles. This section presents a few examples of such novel, post-WIMP user interfaces, as described by van Dam [16], and briefly glimpses at associated current user interaction issues.

2.1 Multi-touch

Very prominently, novel devices, such as smartphones, tablets[2,3] and larger surfaces[4] provide (*multi-*)*touch* input facilities: one or more users can jointly manipulate several virtual objects on small or large screens by touching them with one or more fingers.

The left picture in Fig. 2 shows three users collaboratively solving a Sudoku game on a large tabletop surface, presented by Echtler [30]. In the right picture, an ambulant incident officer uses a multi-touch tablet PC to monitor and organize the actions of a medical relief unit during the triage process of a catastrophic event, discussed in Nestler [44] (see also Iserson and Moskop [37]).

Issues: Some interaction schemes, such as a pinching gesture to resize an object, are becoming commonly understood. Yet, beyond such basic schemes, there is not yet a generally accepted way of moving, grouping and manipulating objects via multi-touch. We investigate suitable multi-touch use on a heavy, rugged device while a user is holding it in two hands (Coskun et al. [5]).

2.2 Mobility, Augmented Reality

By tracking users, *mobile* location-based services or ubiquitous computing (Weiser [17]) and *augmented reality* (*AR*) (Azuma et al. [1]) provide users with computer

[1] Windows, Icons, Menus, Pointers.

[2] http://www.apple.com/iphone/ (accessed 2012-02-26).

[3] http://www.android.com (accessed 2012-02-26).

[4] http://www.microsoft.com/surface (accesses 2012-02-26).

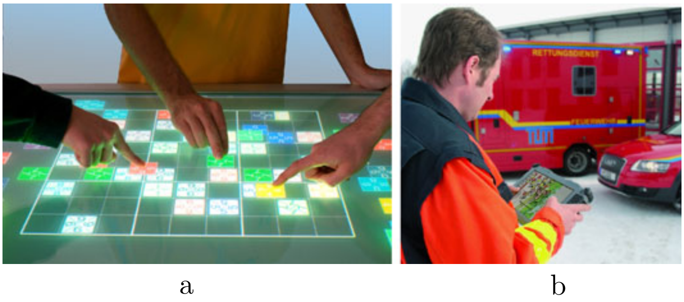


Fig. 2 Multi-touch interaction. (**a**) A collaborative sudoku game. (**b**) Coordinating support during a catastrophic event on a tablet PC

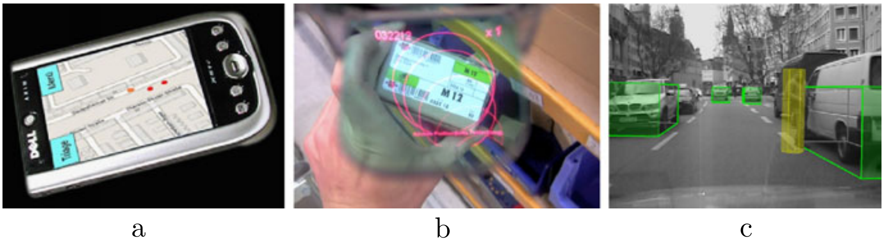


Fig. 3 Navigation assistance on mobile devices. (**a**) A bird's eye view on a PDA. (**b**) Logistics: ego-centric tunnel in a head-mounted display, leading to an object. (**c**) Ego-centric driver assistance in a car

information directly based on where they currently are and what they do. With AR, users see such information three-dimensionally embedded into their physical environment.

The left picture in Fig. 3 shows a PDA-based (2D) navigation assistant for each member of a rescue team in catastrophic events (Nestler [44]). The red dots indicate injured patients who need help urgently. Assignments of patients to rescuers are coordinated by the ambulant incident officer (right picture of Fig. 2), as well as collaboratively in the rescue center on a multi-touch table, such as in the left picture Fig. 2. The central picture of Fig. 3 shows an AR-based (3D) navigation assistant for commissioning tasks in large warehouses (Schwerdtfeger [52]). The logistics workers wear a head-mounted display which shows a tunnel (pink rings) that reaches from the display to the shelf. The right picture comes from a car driver assistance application. It indicates the locations of potential obstacles in the car's drive path, as detected by the on-board sensors (Tönnis et al. [58]). Current research investigates how such information can be presented to the driver: in a central information display, by warning sounds, vibrations, or potentially directly in the driver's view in a head-up display.

Mobile user interfaces raise several critical *issues*. Since users are seeing such information while they also participate in activities of their physical environment, they must not be distracted from looming dangers. Human-computer interaction

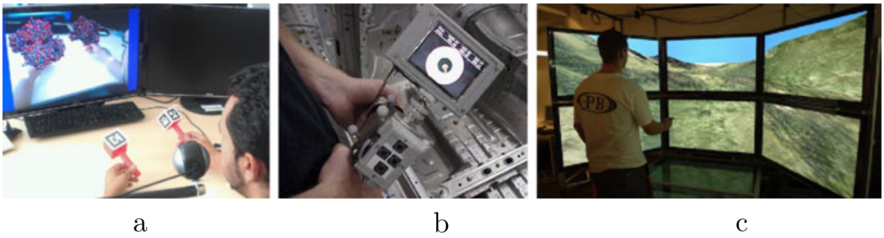


Fig. 4 Tangible interaction. (**a**) Augmented chemical reactions. (**b**) Intelligent welding gun. (**c**) Phone-based terrain exploration in a flexibly reconfigurable virtual environment

in time-critical or dangerous settings must ensure that secondary tasks, such as responding to computer information systems, do not overwhelm users such that they ignore primary tasks, such as attending to a patient (Nestler [44]) or evading physical obstacles (Tönnis [15]). Such issues will become even more urgent when AR is used in mobile settings and users have to operate in complicated physical settings, such as adapting their motion to uneven or slippery surfaces.

2.3 Tangible Interaction, Three-Dimensional User Interaction

By tracking not only users but also physical objects in a three-dimensional physical environment, *three-dimensional user interaction* (*3DUI*) beyond mouse, keyboard or multi-touch surface has become possible (Bowman et al. [2]). *Tangible user interfaces* (*TUIs*) (Ishii and Ullmer [9]) allow users to affect virtual worlds in AR and *virtual reality* (*VR*) applications by manipulating physical objects (Burdea and Coiffet [4]).

Examples of such tangible interaction are shown in Fig. 4. In the left picture, a user investigates and controls bonding activity between atoms of two molecules in a chemical simulation by rotating and moving the molecules via two sticks, one in each hand. The molecules change shape depending on their proximity to one another and the exerted energy fields. Special user gestures, such as holding both hands still for some time, establish and finalize proposed bonds between the molecules (Maier et al. [42]). In the central picture, a user holds a welding gun to attach a number of studs to a car frame. Using a notch and bead metaphor, a display on top of the gun indicates by an arrow inside several concentric rings where the next welding position is. When the user moves the gun to this location, a virtual ball becomes visible; when it fills the center ring completely, the gun is in perfect welding position (Echtler et al. [31]). In this case again, the (tangible) gun is tracked. Extra commands such as the welding itself and the selection of the next stud from a list are activated by special triggers and buttons on the gun. In the right picture, a user flies through a large virtual terrain by moving a smartphone in his hands like a toy airplane. Steering

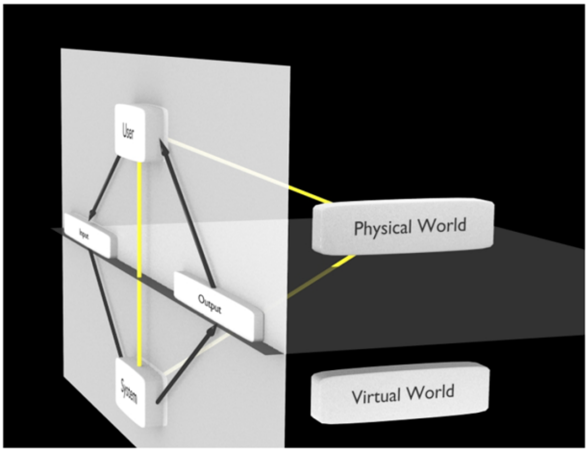


Fig. 5 Interaction triangle between human, computer and physical environment (extension of Fig. 1: the *white plane* shows Bowman's interaction cycle, the perpendicular *dark plane* separates the physical world from the virtual world, with the user, other people, objects and input/output devices residing on the physical side)

commands are derived from the accelerometers in the phone and thumb gestures on the built-in small multi-touch display. This is not a one-to-one mapping from the phone motion to the virtual flight path since the user can move only slightly in front of the screen. Current research investigates what motion gestures are most useful for users to navigate as quickly or precisely as possible along an intended path in a large virtual environment (Benzina et al. [24], Tönnis, Benzina, and Klinker [57]).

Research on three-dimensional human-computer interaction needs to determine the most suitable combination of interaction facilities, such as object or user motion (gestures), buttons, voice commands and more (Bowman et al. [2], Sandor [50]). Furthermore, research needs to determine where in the vicinity of a user these facilities exist (fixed within the environment or attached to the user or to an object) (Feiner et al. [32]).

An important *issue* regards the question how users can immerse deeply into exploring a high-dimensional space of simulated or measured data without being distracted by the human-computer interfaces. Visualization, simulation, virtual animation and interaction need to be so intuitive that the computer becomes virtually invisible (Norman [46]). The computer and human become partners in exploring and analyzing the information, with the computer amplifying human intelligence (Brooks [3]).

This human-computer partnership, embedded within a physical setting, is the overarching issue across all areas of multi-touch, mobile, AR-related or virtual *human-computer interaction* (*HCI*). *Human users*, the *physical world* including sensors and displays, and the *computer system* (including a virtual world full of simulations and animations), form an intricate triangular relationship (Fig. 5). Each corner of this triangle has its own set of errors or risks, all of which need to be dealt with in order to determine good human-computer interfaces.

3 Computer-Related Risks

The development and enhancement of human-computer interfaces, such as those presented in Sect. 2, has various sources of uncertainty, as represented by edges and nodes in Fig. 1. Uncertainties in measurements, interpretations, and presentations or actions result in the risk of miscommunication between humans and machines, which, in turn bear the risk of harmful consequences, if the user interfaces control computer programs with significant impact on our lives. This and the following section describe issues related to these risks. Tables 1–6 relate them to the application examples of Sect. 2.

Figure 1 shows the full interaction cycle between humans and computers (Bowman et al. [2]). This section presents issues pertaining to the uncertainties on the computer side, i.e., issues a computer system designer has to take into account when conceiving, building and testing the system hardware, the computer algorithms, and the underlying concepts. These are represented in the lower part of Fig. 1.

3.1 Sensing

Sensing is the first technical step in human-computer interaction. It is represented by the lower left arrow in Fig. 1. It receives the original input from human users and provides it to the computer system. Table 1 describes sensing requirements and impacts for the exemplary post-WIMP interfaces that were presented in Sect. 2.

To determine user input, suitable sensors need to be installed and registered, and their sensing properties need to be calibrated. Even though this is an issue with any kind of user interface, the following section focuses on issues pertaining to trackers for the multi-touch and tangible interfaces that have been presented in Sect. 2. Position $\mathbf{p} = (x, y, z)^T$ and orientation $\mathbf{r} = (\theta, \phi, \psi)^T$ of a user or an object are generally described as a pose $\mathbf{X} = (\mathbf{p}, \mathbf{q})^T$ with six degrees of freedom in a three-dimensional environment.[5] Orientation corresponds to rotations around three axes that can be provided as Euler angles, in matrix notation or as quaternions. Different fields refer to the rotation angles in different terms, such as yaw, pitch and roll for aircrafts or azimuth, elevation and tilt in astronomy.

Several physical principles can be used to determine and track object poses: optical, inertial, electro-magnetic, acoustic, radio-based, or mechanical tracking (Welch and Foxlin [18]). Each such principle suffers from errors that are generally classified and handled in a number of ways. To some extent, sensor error may be characterized as white noise $\mathcal{N}(0, \Sigma)$, following a Gaussian distribution with mean value 0 and covariance matrix Σ that depends on sensor-internal imprecision with respect to all six degrees of freedom. This is the accumulation of many unknown physical sources and is summarized according to the central limit theorem. Yet, not all

[5] Two-dimensional multi-touch surfaces require three degrees of freedom with $\mathbf{p} = (x, y)^T$ and maximally one rotation angle θ.

Table 1 Computer sensing issues in exemplary applications

Applications	Interfaces	Issues
Sudoku game	Multi-touch, TUIs (Figs. 2a, 4c)	Sensing and system reaction must be immediate and very robust and reliable such that users are able to build up an intuition for proper system control [30].
Catastrophic events	Multi-touch, Mobile (Figs. 2b, 3a)	System reliability is very important and devices thus need special protection (ruggedization), reducing e.g. the sensitivity of the touchscreen. Input signals are thus rather noisy [5, 44].
Logistics	Mobile, AR (Fig. 3b)	The position of the logistics worker (esp. the head pose) must be determined both precisely and robustly across a wide area of a warehouse, requiring complex tracking setups [48]. Further input devices (such as push buttons on a belt) are needed for system control. They must be usable intuitively and blindly [52].
Driver assistance	Mobile, AR (Fig. 3c)	Apart from user input, a large amount to environmental sensor data is required [58]. Relevant measurements include for example the friction coefficient of the street at the relevant section in order to correctly compute and display the breaking distance. For driver assistance, the state of the driver also has to be taken into account. To this end, eye and head tracking are becoming integrated into driver cabins.
Augmented chemical reactions	AR, TUIs (Fig. 4a)	The application uses a “desktop-AR” setup, consisting of tangible objects with markers, in front of a desktop monitor and a camera close to the user’s head [42]. The optical marker tracking algorithm must be fast, precise and robust against partial occlusions of the markers. The camera must be close to the user’s eyes in order to minimize discrepancies between the fields of view of the camera and the user.
Intelligent welding gun	Mobile, TUIs (Fig. 4b)	Stud welding requires sub-millimeter precision and thus very precise tracking. Further input methods for system control must be accessible during the (mobile) welding process [31].
Terrain exploration	TUIs, VR (Fig. 4c)	Sensing of flying gestures requires suitably accurate tracking data, either from a stationary tracker in a fixed VR setup, or from mobile inertial or touch sensors, e.g. built into mobile phones or tangible objects [24]. Thus this represents a tradeoff between high tracking fidelity with high setup costs and mediocre tracking quality without setup costs.

influences average out that way. Some, rather specific errors contribute systematic deviations from the true mean pose of an object, resulting in an inaccurate pose estimate with a systematic offset in position and/or orientation. This may stem from misaligned, or imprecisely placed sensors in an environment, such as a camera after someone has bumped into it. It can also stem from inaccurate depth measurements, if, for instance, a camera possesses an automatic zooming function. A third cause of inaccurate pose estimations is temporal measurement lag. Careful calibration and registration procedures both in the spatial and in the temporal domain are required in order to obtain precise and accurate input data (Huber [8], Keitler [39]).

Further problems arise from the physical limitations of sensors. Camera-based tracking fails when the direct line of sight to a tracked object is lost, e.g. because the object is temporarily occluded by another object due to current object or user motion or when an object leaves the field of view of a camera. Inertial sensors suffer from drift. Field-sensing devices, such as electro-magnetic trackers, compasses or radio-based trackers, loose precision when unforeseen further sources, such as magnetic objects, are added to the environment.

Due to individual sensor limitations, hybrid combinations of sensors are investigated. Many concepts of sensor fusion exist in probabilistic robotics (Thrun, Burgard, and Fox [56]), including Kalman filtering and particle filters. An important aspect involves the construction of robust, redundant sensor networks that combine mobile and stationary sensors (Pustka et al. [48]). In such networks, pose estimations from different sensors are transformed back and forth between different sensor coordinate systems, using forward and backward propagation, with respect to both geometry and sensor errors (Bauer [21]).

Furthermore, diligent calibrations and registrations of sensors and physical reference targets are required, to be performed by a *tracking engineer*. The degree of quality of this work, as well as of the sensors involved, has a serious impact on the performance of the entire human-computer interaction system, to the extent that poor quality may render the system dysfunctional. There is the risk that quality can vary over time, with users not being aware of the current quality level. In applications, such as medical surgery (Bauernschmitt et al. [22]) or high precision metrology (Keitler [39], Luhmann [40]), the current quality has to be checked frequently.

3.2 Interpretation

Interpretation is the central computational step in human-computer interaction. It is represented by the bottom circle in Fig. 1. Table 2 describes computer interpretation issues for the exemplary post-WIMP interfaces that were presented in Sect. 2.

Interpretation receives low-level information, such as a continuous flow of pose data, from the sensors and analyzes it in order to derive higher-level interpretations of users' intended commands to the system, as well as the current state of a changing physical environment. In the background, the computer system then does, what it is best at. According to the received commands, it accesses and analyzes data, computes new results and/or simulates situations within the constraints of a given model and according to further sensor input that monitors aspects of the physical world. Finally, the computer system then provides its analysis to the computer output component to generate appropriate output (visualizations and/or actions).

The complexity of interpreting user input depends on the number of different commands or steering controls that need to be distinguished. In principle, commands can be discrete events or they can relate to continuous control. For continuous control, the stream of tracked user or object poses is transformed into manipulation

Table 2 Computer interpretation issues and solutions in exemplary applications

Applications	Interfaces	Issues
Sudoku game	Multi-touch, TUIs	Due to similarities between multi-touch gestures, these are occasionally misinterpreted. E.g., there can be confusions between two-finger zooming and rotation.
Catastrophic events	Multi-touch, Mobile	Gestures are sometimes not sensed well on a ruggedized device and thus not recognized. Continuous signals can thus be interrupted. Extrapolations from history into a limited future are able to bridge small sensor gaps [5].
Logistics	Mobile, AR	To reduce the risk of wrong interpretations, control input is collected via a rotary dial with a push button [52].
Driver assistance	Mobile, AR	Car traffic scenes—especially at large intersections or in varying weather conditions are extremely complex and variable. The analysis of such sensor data is a long-standing research issue in robotics. For driver assistance, head tracking [58] and the determination of the driver's current state of mind (e.g., from glancing behavior) is added.
Augmented chemical reactions	AR, TUIs	Several different gestures are necessary to select and confirm one of many potential bonds between molecule models. These may differ only minimally from each other. Accordingly, the recognition may misinterpret these gestures [42].
Intelligent welding gun	Mobile, TUIs	The main input device is the welding gun itself. The risk of wrong gesture interpretation is avoided by using menu-based interaction.
Terrain exploration	TUIs, VR	There is no unique mapping from 6 DoF pose tracking to the control of the flight path of an airplane—with only 4 parameters. Proper transfer functions need to be defined [24].

of virtual objects according to predefined transfer functions. Steering results are directly related to the tracking quality of the sensors. Yet, the transfer functions may provide some filtering such as damping to reduce jitter that is due to sensor noise.

For recognizing discrete commands, the system designer needs to know how many different commands exist and how they can be distinguished. This means that a vocabulary of gestures needs to be considered, with each gesture being associated with distinctive features. Using machine learning and pattern recognition techniques, sequences of user poses (i.e.: user gestures) need to be compared and clearly separated from each other. For each gesture, the features form a cluster in some measurement space. Clusters from different gestures should not overlap—better: there should be a wide gap between clusters such that they can be distinguished even under the presence of noise. To this end, recognition algorithms need to be designed that derive appropriately distinguishable properties (the measurement space) from pose sequences. In addition to distinguishing between gestures, also false alarms (false positives) need to be considered and discarded.

The risk of misinterpreting gestures arises from non-unique situations, i.e. situations for which noisy pose sequences cannot be associated clearly with exactly one command. Another risk comes from the fact that the underlying world model for the design of appropriate gestures and the reaction to measurements of physical events may not have been complete: in real-world physical settings, situations may arise that lead to unintended gestures that are interpreted inappropriately in the limited scope of a computer application because the overall context was not fully modeled and understood (Neumann [11]).

3.3 Presentation

Presentation is the last computational step in human-computer interaction. It is represented by the lower right arrow in Fig. 1. It receives interpretations of the current user input and of the state of the physical environment from the interpretation component as well as the results of the background work, such as the current state of a simulation. It generates appropriate visualizations on output devices, to be perceived by the users. Table 3 describes computer interpretation issues for the exemplary post-WIMP interfaces that were presented in Sect. 2.

The amount of information that is acquired, generated and processed inside computers can be tremendous. The information presentation component is concerned with issues (1) what to show, (2) how to show it and (3) where to show it. Information presentation may involve hundreds of different attributes in a high-dimensional property space. Objects can have very intricate relations to one another, forming clusters, correlations, anti-correlations etc.

The first question is, *what to show*. Data reduction schemes such as projections from high-dimensional spaces to lower dimensions, as well as selections and combinations or rearrangements of dimensions using, e.g. principal component analysis are employed. There is a risk that important information is omitted or hidden in accumulations or projections along an attribute dimension. Interactive data exploration schemes and automatic data mining are part of an answer to such risks, allowing users to poke at the data and massage it until they are convinced that they have observed and explored all relevant aspects. Yet, short of performing an exhaustive search, little guarantee can be given that the entire body of information has been presented in all possible combinations of and along all attribute dimensions. An emerging concept of information selection for mobile applications concerns context dependency. It cannot be formulated as succinctly in mathematical terms but rather depends strongly on the interpretation of sensor data and the assumed world (context) model—bearing the risk that such model may not be complete and interpretations thus deficient (see Sect. 3.2).

The next question is, *how to show* the selected information. Information visualization and scientific visualization are concerned with developing schemes to present a wealth of information to users, bringing out the essential details without loosing the overview of the general context (Spence [14], Bederson and Shneiderman [23], Nielson, Hagen, and Müller [45], Tufte [59]). Quite a number of concepts exist on how to represent information in perceivable (visual, aural or tactile) form, by mapping attribute dimensions to the dimensions of a representation scheme. Information can be represented both for individual objects (object visualization) and

Table 3 Computer presentation issues

Applications	Interfaces	Issues
Sudoku game	Multi-touch, TUIs	Since the display is partially occluded by the hand during a multi-touch interaction, users may not be aware of all provided feedback. In programs that make heavy use of motion-sensors as input, showing detailed information could lead the user to either a bad performance or ignoring the information because the user needs to stop the motion of the display to be able to read the information properly.
Catastrophic events	Multi-touch, Mobile	The system shows a map plus icons of victims, rescue personnell, ambulances etc. Issues involve how to provide both an overview and detail and how to arrange the icons in dense areas where there is not enough space to show them side-by-side. Another issue is the suitable presentation of aggregate information. The entire presentation is shown on a large multi-touch table. Specialized views are also shown on a handheld tablet PC and on mobile phones to support and coordinate activities on-site [20].
Logistics	Mobile, AR	The logistics application shows a tunnel that directs the user to the picking target. The curvature of the tunnel reflects the distance and the turning angle to the target. Additionally, the application identifies the relevant target at the destination [52].
Driver assistance	Mobile, AR	Information related to neighboring or approaching traffic participants can be shown in a variety of modalities and at several places in a car, e.g. in a central display, a head-up display, as well as via sound or a vibrating seat, steering wheel or gas pedal. Examplary information includes driving directions, augmented onto the street, required breaking distances, the drive path, as well as the direction of looming dangers that the driver should attend to [15].
Augmented chemical reactions	AR, TUIs	Aside from showing the current state of two molecules, it is important to show some or all potential further bonds, as well as the status of current gesture analysis and recognition (e.g. the upcoming timeout for a “holding still” gesture) [42].
Intelligent welding gun	Mobile, TUIs	To guide the welder with the required accuracy, not only navigational aids are displayed, but also accuracy indicators that show the deviation of the current position of the tip of the welding gun from the targeted welding spot. Furthermore, the target position is presented textually at the top of the display—as a global orientation aid [31].
Terrain exploration	TUIs, VR	To immerse viewers into the terrain data, it is spread across an arrangement of 3 walls of a fully recoverable cave, FRAVE [57]. Each wall consists of two large stereo displays. Two further displays are placed on the floor, equipped with multitouch sensing facilities.

for statistical aggreations with respect to attribute dimensions (attribute visualization), such as histograms, scatter plots, and parallel coordinates plots. To this end, two or three spatial dimensions and the temporal dimension (animations or interactive, iterative steering) can be used to layout data geometrically as framing dimension that span the spatial layout of a representation scheme. These spatial dimensions can also be recursively sub-arranged (nested) to show blocks of data from further dimensions—contained dimensions representing information in each cell of the framing dimensions. Visualization schemes for contained dimensions are, for example, color, semi-transparent presentations, texts and glyphs with special shapes, orientations etc. The use of special structures such as trees or graphs leads to further well-established options. In many mobile and AR-based applications, the huge amount of data is not as much an issue as the question how to find suitable three-dimensional metaphors to relate the virtual data to the physical world of the user without occluding too much of the environment. Should the information be represented in a first-person perspective (ego-centric view) or in a bird's-eye perspective (exocentric view) (Bowman et al. [2])? What are suitable metaphors to indicate information behind a physical object—i.e., information that is currently occluded (Dey, Cunningham, and Sandor [29])? The x-ray metaphor is not immediately intuitive since it is inconsistent with physical reality. Other questions discuss how to represent operational information, social information about groups of objects or people or abstract background information that does not have a unique spatial connotation.

The final question is, *on what physical displays and where in the environment to show* the information. Representational layouts of information depend on the device, as well as on the available compute power and network bandwidth. Current presentation schemes vary from large, detailed presentations on combinations of multiple wide screens, such as a CAVE in VR over large single screens to desktop systems, tablet solutions and tiny displays on smartphones, with and without audio support (Artinger et al. [20], MacWilliams et al. [41], Sandor and Klinker [51]). Display characteristics such as the resolution, the dynamic range and color gamut of a display, the field of view and field of regard that it subtains in front of a user, and its current pose play an important part in devising an information presentation concept for the human-computer-interaction aspects of a computer application (Bowman et al. [2]). Providing interactivity, e.g. via WIMP-based devices, multi-touch, or tracking also influences the information presentation schemes since the UI also needs to have visual representations on the display in form of GUIs, virtual hands, icons or avatars.

Design decisions with respect to these issues bear many risks—yet those are generally not directly related to the technical issues that are presented in this section but rather to human issues and thus will be discussed in the following sections.

4 Human Issues

Following the discussion on uncertainties on the computer side, this section presents issues pertaining to the uncertainties on the human side: human sensing, perception and interpretation, and action. These are represented in the upper part of Fig. 1.

Human issues are not risks in themselves. Yet, they need to be considered as human factors when designing the computer interaction schemes of Sect. 3. To this end, they become the focus of user-centered design and testing schemes that are presented in Sect. 5.

4.1 Human Sensing

Human sensing is the first step on the human side of human-computer interaction. It is represented by the upper right arrow in Fig. 1. It describes the instant when humans sense computer output with their innate sensory organs (Eysenck and Keane [6]). Table 4 describes human sensing issues for the exemplary post-WIMP interfaces that were presented in Sect. 2.

Human sensing has general properties and limitations, as well as special limitations of individuals, depending on age, health and other factors. In the following, the discussion is restricted to visual sensing. The human retina has two kinds of sensory cells: cones and rods. They respond to light stimuli in the spectrum of "visible light". Cones have three different pigmentations that make them sensitive to different wavelengths and allow humans to see colors. They need significant amounts of light in order to respond. That's why color vision works well in broad daylight, but not at night. Color blindness is caused by deficient pigmentation, e.g., when green pigments are missing. Rods, on the other hand, work at low light levels. Yet, they respond to the entire visible spectrum rather than to subranges of wavelengths. Thus, they allow humans to see at night time—in grayscale rather than in color (Eysenck and Keane [6], Gregory [7]).

An important property of human eyes is foveal acuity. The retinal focus of the eye, the fovea, contains cones with very high density. Humans can see very acutely with this part of the eye, whereas vision in the remaining areas of the retina (peripheral vision) is decreasingly acute with increasing distance from the fovea. Yet, the peripheral area is known to help humans in detecting object motion. It also provides a wider, coarse overview of the physical environment for a field of view of about 175 degrees.

A further important issue of human vision is depth perception. A large number of monoscopic and stereoscopic depth cues exist. One of the most important ones is stereopsis. With two eyes, humans see objects in front of them twice, with a horizontal offset (disparity) on each retina. The disparity depends on the distance between the eyes and the distance of an object from the eyes. The smaller the distance the larger the disparity. By triangulating, the human brain is able to estimate depth from disparity, up to a distance of about 6 meters (Gregory [7]).

Further important cues are summarized under the term adaption, describing oculomotor (muscle) activity in the eye. Humans converge their eyes inward such that the most important object is seen in the foveal areas of both eyes with the highest acuity. This converging eye rotation is used by the brain as an additional source of

Table 4 Human sensing issues

Applications	Interfaces	Issues
Sudoku game	Multi-touch, TUIs	Multi-touch tables can be rather large; inspecting objects at the other side of the table is less precise than at close distance. Objects/text can also be upside-down relative to the current viewpoint.
Catastrophic events	Multi-touch, Mobile	This application takes place in a very stressful physical environment. Users are possibly bombarded by many physical sensations, such as bright or very dim illumination, loud sounds, and physical obstacles. Virtual presentations need to be shown with the right amount of contrast such that they are neither too strong nor too weak.
Logistics	Mobile, AR	Virtual navigation information is shown in an optically see-through head mounted display (HMD). Such displays present information at a specific virtual distance from the user's eyes—typically approximately at arms' length. If the object is not close to the virtual presentation distance, users cannot focus simultaneously on a physical object and on its annotation. Furthermore, if the current background if similar to the presentation in the HMD, the virtual information cannot be perceived well.
Driver assistance	Mobile, AR	When information is displayed inside the car, users need to adapt their eyes, both for brightness and for focal distance when they look at the information and also when they go back to the road. This takes valuable time that can be critical when reacting to physical dangers. Head-up displays show information outside the car, thereby reducing the adaptation time.
Augmented chemical reactions	AR, TUIs	The molecules are shown on a common desktop monitor, providing only monoscopic depth cues and motion parallax. In this case, people have shown problems seeing the 3D structure of the molecule well. There can also be a hand-eye coordination problem.
Intelligent welding gun	Mobile, TUIs	The welder needs to look at color images on a small screen. The navigational arrows are shown in 3D. Depth perception might be an issue. Yet, motion parallax is a dominant cue since the welder and the gun are mobile. Furthermore, at welding time, the physical car frame provides strong haptic cues.
Terrain exploration	TUIs, VR	To trigger the human ability to see stereoscopically in three dimensions, each screen of the FRAVE shows two versions of the scene, one for each eye of a tracked user. He/she has to wear shutter glasses that are synchronized to all displays simultaneously such that each eye only sees the version dedicated to its viewpoint.

depth information. Accomodation is a further depth cue. Muscles contract or dilate the lens in the human eye to allow it to focus on objects at different distances. Accomodation also contributes to the brain's estimation of object distances.

In normal physical settings, stereopsis and oculomotor cues all contribute to a consistant depth perception of objects in front of a person's eyes.

In human-computer interaction, *computer displays present representations of virtual information that are subject to human sensing capabilities and limitations*. It is important to account for potential color blindness, as well as for the fact that presentations can only be seen sharply in the very small foveal area of each eye. Thus, most parts of a computer presentation are seen without high resolution. It takes humans time to actively move their eyes across important areas of a display in order to see each area with the foveal area of their eyes. Eyes jump in saccades between different display areas. This may be in conflict with computer animations that expect users to focus on a certain display area at a particular instant in time. Thus, there is a risk that parts of a presentation are not seen with sufficient acuity by a user, since these areas were not within the foveal area at that moment.

Peripheral vision must also be considered with care (Jones et al. [38]). In many cases, it is not integrated into information presentation schemes and devices. Since peripheral vision provides humans with overview and advance notice of potential hazards in their environment, lack of peripheral input can result in risky or tiring situations: if a head-mounted display is closed around the eyes it shuts off users' peripheral view of the physical world. If open, but not covered by the display, there is a visible seam between the virtual and physical world. Furthermore, users have to rotate their heads continuously from side to side to see the information that is geometrically related to areas that are not covered by a small field of view in a wide geometric range (Rolland and Fuchs [49]).

Another issue is a sensory mismatch of depth cues for three-dimensional presentations of virtual objects in a stereoscopic display (Bowman et al. [2]). Here, convergence and stereopsis—induced by unnatural viewing conditions involving shutters, polarized or red-green glasses in front of a user's eyes—provide a depth impression that is inconsistent with accomodation: the eyes focus on the display surface rather than on the simulated depth of the virtual object. Such sensory mismatch is a problem for VR and also for AR, using head-mounted displays. Depending on the situation and the physical constitution of the user, one cue may dominate over another, thereby inducing the respective depth impression. Yet, there is the risk of users getting head aches or suffering from simulator sickness (especially, if motion cues are involved). Furthermore, there is a risk of potential after effects, i.e. the brain may adjust to this sensory mismatch and remain in this stay even when the user deals with physical objects—with thus reduced sensory ability.

If, on the other hand, a video-based presentation scheme is used that replaces the *optically see-through* direct integration of virtual information into video streams of the real world (e.g. in a mobile phone on a stationary display, or on opaque head-mounted glasses), humans suffer from reduced hand-eye coordination since the viewpoint of the camera does not coincide perfectly with their eyes.

4.2 Perception

Human perception is the central step on the human side of human-computer interaction. It is represented by the top circle in Fig. 1. It describes the process when humans attend to a sensed computer output, become aware of it and thus perceive

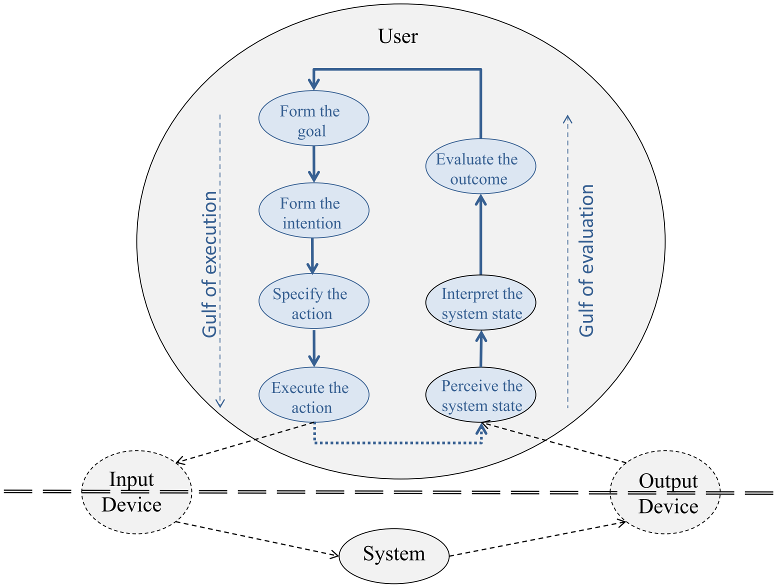


Fig. 6 Stages of action as a dynamic process of execution and evaluation during human-computer interaction (adapted and extended from [12])

it (Eysenck and Keane [6]). According to the perceived computer output, they reason about the world. They analyze the situation and draw conclusions, forming a goal and intention towards performing the next action. In the background, humans may meanwhile perform a number of independent autonomous tasks that are related to the project but do not need computer support. Eventually, they return to the human-computer interaction cycle, with the intention to provide input to the computer. Table 5 describes human perception issues for the exemplary post-WIMP interfaces that were presented in Sect. 2.

A seminal schematic of human cognitive work during human-computer-interaction has been presented by Norman [12]. Figure 6 shows Norman's action cycle, integrated into the cycle of interaction of Fig. 1. Norman's cycle spells out human cognitive activity in more detail, focusing on perception and interpretation as the central part. Norman connotates the transitions both from the sensing phase (Sect. 4.1) and to the action phase (Sect. 4.3) with metaphorical gulfs that users have to cross and that might overwhelm them with great or even unsurmountable challenges: the gulf of evaluation, i.e., perceiving and fully understanding the current state of a computer program from its current and past output, and the gulf of execution, i.e.: deciding what next step to take and how to convey this to the computer via its input system.

Perception requires humans to attend to stimuli. By doing so, humans may pay less attention to other stimuli. This situation is known as *perceptual tunneling* (Wickens and Hollands [19]) or *inattentional blindness*. Furthermore, experiments have demonstrated that humans can be completely oblivious to changes in parts of

Table 5 Human perception and interpretation issues

Applications	Interfaces	Issues
Sudoku game	Multi-touch, TUIs	On a large multitouch table, players may not be able to perceive all simultaneous changes on the board, if multiple players are interacting with the game (local focus, change blindness, cognitive capture).
Catastrophic events	Multi-touch, Mobile	Important information events can be missed by the rescue workers e.g. while treating the patient (change blindness). To avoid wrong decisions, each change should be clearly visible continuously, even if the rescuer missed the change when it occurred.
Logistics	Mobile, AR	The visual augmentations in an HMD may overwhelm commissioners such that they do not pay enough attention to the physical environment and run into physical obstacles (inattentional blindness).
Driver assistance	Mobile, AR	Information presentation in cars needs to be kept to a minimum to ensure that drivers are not overly distracted from the physical world. In tests, drivers have shown symptoms of perceptual tunnelling from as little as a single polygonal presentation of the expected drive path "contact analog" in a head-up display: in an experiment in a driving simulator, they drove significantly faster when seeing the drive path than when seeing only a line indicating the current breaking distance [15]. Furthermore, many users also suffer from simulator sickness, due to a very strong visual stereo impression inconsistent with their haptic sense (balance, proprioception).
Augmented chemical reactions	AR, TUIs	When interacting with large molecules, users may be overwhelmed by the amount of information and simulated activity. While a user is holding still and watching the simulation, a bonding activity may get activated accidentally (change blindness, cognitive capture). The system must ensure that the user is properly informed suitably.
Intelligent welding gun	Mobile, TUIs	Welders may be so fascinated by the AR-based guidance on the display that they ignore their physical environment and bump into obstacles (perceptual tunneling).
Terrain exploration	TUIs, VR	Due to a very strong stereo impression that does not match haptic sensing, users often suffer from simulator sickness.

their surroundings that are not within their current focus—a problem called *change blindness* (Steinicke et al. [54]). In addition to ignoring stimuli, humans here also exhibit problems memorizing recent images with the level of detail that is required to compare them agains newly incoming stimuli. Another well-known problem is *cognitive capture*. In this case, humans are so absorbed by a cognitive task that they ignore new, unexpected stimuli, as shown by the stunning video experiment of a gorilla walking through a basket ball game without being perceived by a large number of test persons.[6] Even further, humans may suffer from conceptual or *cognitive*

[6] D. Simonis and C. Chabris. Selective Attention Test. http://www.youtube.com/watch?v=vJG698U2Mvo, 1999. (Accessed 2012-03-02.)

overload, i.e. they receive so much information that they are unable to deal with it. The result can be *cognitive tunneling* (Wickens and Hollands [19]): humans become unable to make decisions due to information overload. They may then restrict themselves to pursue only a very limited subset of available options.

These are serious *concerns for information visualization and human-computer interaction schemes* since the mere fact that information has been presented cannot be taken as a guarantee that users have actually perceived and understood it. This covers some of the aspects of Norman's gulf of evaluation. Beyond being overwhelmed by too much information, users may also experience a gulf of evaluation due to poor, misleading representation schemes. Reasons could lie in sensing difficulties (color blindness, poor resolution), or in the choice of confusing presentation metaphors.

AR and VR have different strategies and goals towards dealing with human perceptual limitations. VR strives towards generating a virtual immersive experience by exploiting human limitations, such as perceptual tunnelling, change blindness, and cognitive capture—just as magicians when they confuse spectators with their tricks. Users are expected to overlook of suppress the perception of cues that tell them that the physical reality is different from the virtual experience. VR faces the danger of simulator sickness or a non-perfect sense of presence when sensing mismatches are not sufficiently strongly overwhelmed by the sense that is intended to be dominant. AR, on the other hand, needs to ensure that users co-exist safely with their physical environment. To this end, virtual information must not overwhelm the user's senses to an extent that physical reality is ignored. Virtual distractions may lead to a lack of situation awareness due to perceptual tunnelling, information overload or cognitive capture—with potential physical harm to the user. In evaluations of AR-related applications, users need to be assessed regarding their level of distraction, e.g. by requiring them to simultaneously perform activities related to the physical environment while also interacting with virtual information. The amount of distraction is determined via eye-tracking, analysis of the response time, and the amount of errors.

After interpretating sensor input, users face the gulf of execution when planning the next action. To this end, users need to be aware of the options that the input devices of the user interface offer. These options either need to be learned and memorized from manuals of from trial and error experiences, or the interface must be flexible enough to allow natural, spontaneous human input (such as natural speech or natural gestures). It is crucial that the user interface allows users to interact with as little contemplation of available options as possible. Users may get lost and confused in poor, unclear and inconsistent input schemes.

4.3 Action

The third, final step on the human side of the human-computer interaction cycle involves executing/performing the planned action. It is represented in Fig. 1 by the

Table 6 Human action issues

Applications	Interfaces	Issues
Sudoku game	Multi-touch, TUIs	At a large multi-touch table or in a large virtual 3D space, players may not always be able to directly touch and manipulate an object on the far side of where they are. They either have to move, or the system needs to provide metaphors for non-linear motion or manipulation via indirect pointing or selection, using a prop [2].
Catastrophic events	Multi-touch, Mobile	Collaborative control via a multi-touch table raises issues similar to the games scenario above. For the mobile part, the ruggedized tablet PC is very heavy and needs to be carried with two hands. Thus, only the thumbs are able to touch the multi-touch area of the tablet—along the vertical rims of the device. The application takes these limitations into account to design special "thumbs-only" metaphors to select and manipulate icons on the map while also scrolling and zooming the map [5].
Logistics	Mobile, AR	The user interface of the logistics application was deliberately kept simple and uses only a rotary encoder and a push-button. Still, picking incorrect items cannot be prevented by the system. Also a certain amount of training is necessary to adapt to the system.
Driver assistance	Mobile, AR	Car drivers need to be able to reach all computer control elements with their limbs while sitting in their seat. Suitable arrangements and semantic grouping of knobs and dials around the primary control area involving the stearing wheel and the gas, break and clutch pedal are major issues of modern car design.
Augmented chemical reactions	AR, TUIs	A common problem is that different users perceive different gestures as natural or intuitive. For example shaking a tangible object may be interpreted as removing the current molecule from the tangible by one user, or as moving to the next selected atom in the molecule by another user.
Intelligent welding gun	Mobile, TUIs	Apart from welding the studs, the only user actions concern setup and selection of welding scenarios as well as the inspection of previously recorded data. For deployment in an industrial environment, this interface was kept simple and robust. It has been in industrial use for several years; the welder(s) did not report problems with the required human action.
Terrain exploration	TUIs, VR	Flying a toy airplane is, in principle, well understood by most people. Yet, the relationship between the physical actions (recorded in a 6 DoF space) to the 4 DoF flight control of an airplane is not easy to grasp. Experiments show that it requires significant physical talent and gaming experience to isolate the important parameters from the redundant ones.

upper left arrow. It takes a user's planned action and tranforms it into a physically measurable action, suitable to the input devices of the system. Table 6 describes human action issues for the exemplary post-WIMP interfaces that were presented in Sect. 2.

Even when users have decided what action to take, it requires skill, experience and dexterity to actually perform the necessary physical action, depending on the input devices. The action may need to be executed with specific speed or precision, e.g. when typing on a keyboard, pointing (double clicking) with a mouse, pen or finger, speaking a command, looking at an object (glance control), or performing a free-form 3D gesture. Proper execution may require high reactive skills, good gross or fine motor control or even a well-developed sense of balance or rhythm (e.g. when interacting via a balance board in a sports game).

Users may not be fully aware of the assumptions and requirements of the input system, regarding the speed and precision for the intended actions to be recognizable by the system. Furthermore, even if the requirements are clear, it is not always easy for users to act according to the input specifications—or it may be uncomfortable or straining for them to perform the actions.

Depending on the application, this may be a thrilling and interesting challenge (games of skill), a frustrating hinderance (e.g. in office applications), or a potential source of danger (e.g. in safety-critical situations). Such issues are topics in ergonomics and human factors research (Shneiderman and Plaisant [13]).

Keyboard layouts and pointing devices have been analyzed, regarding users' ability to produce fast and/or precise input. As a prominent example, Fitts' law describes a relationship between the size of a target and its distance from a user's current pointing position (Fitts [33]): the larger a target, the faster can users move across a long distance to hit it easily. Vice versa, the shorter the distance to the target, the smaller can the target be—an essential aspect for designing layouts of icons on desktop-style graphical user interfaces. The GOMS model (Card, Moran, and Newell [27]) describes user interaction as an interplay of goals, operators, methods, selection rules. It was designed to help dividing interactive tasks into series of small actions in order to predict the time required to perform complex tasks. For example, this has been done for typing, using the keystroke-level model (KLM) (Card, Moran, and Newell [26]).

The so-called QWERTY keyboard[7] is a negative example: more than a century ago, the arrangement of keys was not designed to improve humans' typing speed but rather to keep the physical hammers from jamming.

There are also many evaluations of pointing devices. Critical distinctions exist between the concepts of direct pointing/touching versus indirect pointing. Direct pointing and touching, e.g. with a pen, with one's fingers on a multi-touch surface, or in augmented reality and tangible interaction, provides users with a direct association between their action and the visual object/icon that they are manipulating. In its purest form, the performed action has a one-to-one mapping to the intended manipulation, such as moving, rotating or enlarging a virtual photo that is shown on an interactive table, or manipulating a physical object. Yet, fingers or pens may not provide sufficient precision and accuracy when selecting very small objects—probably within a densely populated neighborhood of further objects—and/or when intending to perform minuscule manipulations. Indirect pointing e.g. with a computer mouse, on the other hand, allows much more precise selection and

[7] Called so due to its arrangement of keys in the upper row: Q-W-E-R-T-Y (English version).

manipulation—especially when they can be performed in conjunction with sufficiently large widgets such as scroll bars and dials. Yet, the direct association between user action and resulting object manipulation is missing. Users have to familiarize themselves with a mismatch of the position and direction of their physical manipulation with respect to consequences in the virtual computer world. For example, novice users of a computer mouse (such as very young children) have been observed lifting the mouse vertically upwards (rather than pushing it horizontally on a table) when trying to control upward cursor motion on the vertical screen of a desktop monitor.

Another critical issue are the dimensions of the physical interaction space versus the virtual world. Direct manipulation such as multi-touch interaction cannot work in locations (e.g. on very high screens on a wall) that users cannot reach. Furthermore, even if a location can be reached, users may not always want to move across extended distances. To this avail, rate control and non-uniform mappings between user action and virtual interpretation have been established (Bowman et al. [2], Shneiderman and Plaisant [13]).

5 Testing Issues

The previous Sects. 3 and 4 have presented and discussed a large number of issues and uncertainties pertaining to the design and the implementation of suitable interfaces for human-computer interaction. There is a huge parameter space of design options with many alternatives. At the outset, it is not clear which design choice is better than another one—or even optimal with respect to some criterion. Moreover, criteria and options may change over time, due to improving computational, sensing and presentation facilities of computers, as well as due to evolving cultural backgrounds on the human side regarding the ease of understanding upcoming interaction metaphors.

For each interaction concept, there exists the risk of misunderstanding and misinterpretation. Depending on the application, such miscommunication may be a challenge, a nuisance, or a source of danger, bearing potential harm to the user and/or the environment. Independently of the severity of the consequences, it is mandatory for the design of human-computer interaction systems to be accompanied by dedicated evaluation procedures, from project conception to product delivery in a *user-centered design process*. The evaluations are typically conducted empirically, using hypothesis-based testing procedures with a specified level of significance and associated alpha and beta errors (Sirkin [53], Swan, Ellis, and Adelstein [55]).

5.1 Evaluation Design

During the entire process of conceiving, building and finalizing a human-computer interaction concept, the current state of the design and implementation needs to undergo continuous evaluation. This is not a one-step task. Rather, design, prototypical implementation, evaluation and re-design build upon one another in ever-continuing circles (Bowman et al. [2], Shneiderman and Plaisant [13], Chandler and Chandler [28]).

Yet, the test designs may change over time, depending on the maturity of the system, as well as on the urgency of obtaining a preliminary appreciation of vague ideas vs. an in-depth comparison of well-thought-through metaphors or devices.

5.1.1 Strategies and Methods for Different Process Phases

A number of different evaluation approaches exist that are appropriate in different phases of a project and/or serve different evaluation strategies (Bowman et al. [2], Shneiderman and Plaisant [13]). Evaluation designers use a palette of different approaches during the course of the project.

In the very early, conceptual phase of a project and while the very first prototypes are being built, first evaluations and feedback are often acquired via *expert reviews*: the ideas and concepts are presented to a small number of experts at the example of use cases, e.g. by walking them mentally through the intended interaction processes or by demonstrating a rudimentary prototype. Feedback is gathered via questionnaires or interviews. If possible, experts will also be asked for heuristic evaluations, relating the current ideas to known guidelines and cases of best practice in the field. Such early feedback is valuable in cutting back on ideas that experts can quickly identify as unsuitable, based on their background expertise.

When early prototypes become available, *usability testing* becomes an option. Initial tests are typically conducted as *formative evaluations*, involving only few test persons and investigating only a small, well-selected subset of issues to form the base for a consistent interface design. As with expert reviews, designers can retrieve quick and very valuable feedback from such small evaluations: typically, very few test runs suffice to indicate the initial, major issues that need to be improved (Schwerdtfeger [52]). At later project phases—especially shortly before release, more substantial *summative evaluations* are conducted to sum up thorough comparisons of all options. The next Sect. 5.2 presents usability testing in detail. Usability tests are typically surrounded by demographic and subjective questionnaires, as well as closing interviews. Those are the topic of Sects. 5.3 and 5.4. In principle, the entire design space needs to be evaluated at this point. Yet, some simplifications are typically made for the sake of reduced complexity (Chandler and Chandler [28]).

At a later, more mature phase, larger target groups of users are also increasingly involved via *user surveys* and *acceptance tests*. A good example is the early release of beta-versions of computer systems, e.g. before the roll-out of a new game (Chandler and Chandler [28]).

Finally, after product release, feedback is gathered online, e.g. in newsgroups, as well as via telephone call centers, further acceptance tests, and user surveys.

5.1.2 Evaluation Criteria

A number of different metrics can be used to compare and evaluate the quality of different designs (Bowman et al. [2]).

For computer systems, system performance, such as the average frame rate or latency, or the network delay is of utmost importance. For visual systems, optical distortions of cameras and displays, the provided field of view and the resolution are further important evaluation criteria. These can be measured and compared without much user involvement. Yet, they also need to be considered in the context of user-centered evaluations since different performance metrics may have a large impact on the user-based test results.

The next set of evaluation criteria focuses on task performance: how fast can a user reach a specific location? What accuracy is being achieved? How many errors do users make when selecting or manipulating an object? Further criteria are the speed of learning a concept, the spatial awareness a user has gained when interacting with objects in a three-dimensional space and the degree of distraction induced by the system (e.g. by analyzing users' eye movements: when did they look where?). These are metrics that can be measured objectively, using automatic procedures. Data is collected during a test run and stored for subsequent statistical analysis. These criteria are describing the pragmatic quality (PQ) of a user interface, i.e. its effectiveness and efficiency.

The final set of evaluation criteria deals with subjective metrics involving user satisfaction—the so-called hedonic quality (HQ) of a system (Hassenzahl, Kekez, and Burmester [36]). In questionnaires, users are asked to describe their perceived ease of use, ease of learning and their satisfaction during the interaction process on a given scale. Further parameters, related to novel three-dimensional user interfaces are related to users' sense of presence in a virtual environment, and their degree of comfort (simulator sickness), pertaining to the elaborations in Sect. 4.1 on user accomodation, adaption, and potential after effects. How long does it take users when they subjected to optical illusions (sensory mismatches) during an experiment to re-adapt to the true physical interpretation of their senses after an experiment?

5.2 Usability Testing

Usability testing has gained much attention and importance. It represents the attempt to parameterize all important issues of a human-computer interaction approach systematically, describing them as a set of factors (dimensions). If these factors are independent, and if there are no additional, confounding factors, different approaches can be compared by letting a sufficiently large group of representative users interact with the computer in all different variants.

In order for such testing to be successful (i.e. to produce significant results), great care has to be taken to design a good test plan. In the following, several aspects of the physical environment, the underlying concept, and the established process and experimental structure are presented.

5.2.1 Test Setup in a Usability Lab

A usability lab typically consists of two areas that should be separated from each other as much as possible.

The first area is set up for the test person to interact with the system, as designed in the test plan. The environment may be as simple as a desktop monitor with WIMP-style interaction devices, or as complicated as an immersive, multimedia, three-dimensional driving or flight simulator, possibly even integrated into a motion platform. It may also be mobile, e.g. integrated into a real car driving in real traffic, or a mobile phone in pedestrian applications, such as in an augmented reality context. In all cases, the setup should be as realistic as possible, and the test person should be disturbed as little as possible while the experiment is running. The test area should be instrumented with extra recording equipment, such as microphones, cameras, eye trackers etc—in order to store as much information as possible about the course of the experiments and especially about the users' actions, reactions, gestures, mimics and side remarks. Such data can be invaluable during the post-analysis step when questions arise because a particular experiment has unusual results (i.e., outliers).

A second area is arranged for the person in charge of running the experiment. The experimenter should not influence the test person. Thus, the areas should be separated—ideally by a wall with a semi-transparent window.

Contrary to these well-established standards, the evaluation of novel user interfaces (e.g. for augmented reality) may require arrangements that conflict with prior guidelines of best practise. Schwerdtfeger argued in his dissertation that, for AR-based user interfaces in a logistics application, it was more reasonable to interrupt test persons when they consistently went astray than to let them fail during the entire experiment—since the reason for such errors was often related to poor calibrations of the optical-see-through display on their heads or to a basic misunderstanding of some aspect of the very novel hardware and interaction metaphors they were exposed to Schwerdtfeger [52].

5.2.2 Process of Collecting the Data

When preparing and conducting user tests, utmost care has to be taken to apply proper procedure—such that the results are not unneccessarily tainted and thereby rendered unusable. It is extremely difficult and costly to rerun an experiment: test persons will not react the same when they are exposed to the same interface a second time. Acquiring new test persons is time consuming and difficult.

Thus, much care must be taken during the planning phase of the experiment. All potential aspects that might have an influence have to be identified and either explicitly discarded from the test design or accounted for as one of the parameters under evaluation (see Sect. 5.2.3).

During the experimental part, proper procedure has to be set up and executed for each test person. A well-established procedure consists of greeting and introducing each newly arriving test person to the test setup in a predefined way (possibly using rehearsed sentences) such as not to bias persons at this stage. Typically test persons fill out a demographic questionnaire requesting information about general human factors (age, sex, …) as well as special factors (color-blindness, familiarity with novel user interfaces or games, …). The test in itself may consist of one or several parts, exposing test persons to different kinds of user interfaces or to different scenarios. Inbetween, further questionnaires may ask people about subjective impressions (see Sect. 5.3). The experimental session closes after the last set of experiments—possibly with a further subjective questionnaire and/or with a standardized or open interview (see Sect. 5.4).

After all test persons have participated in the experiment, the collected data is analyzed, filtered, and subjected to statistical analysis tools for hypothesis testing (Sirkin [53]). The results are compiled into a report.

5.2.3 Multi-factorial Design

As stated in Sect. 5.2.2, the design of a usability test has to account for all parameters that have a potential impact on the results. Multiple parameters are modeled by *multi-factorial design* approaches (Mukerjee and Wu [10]).

One or more measurement functions $t_d = f_d(x, y, z, \ldots)$ are established that describe criteria listed in Sect. 5.1.2, t_d, as functions f_d of parameters $x, y, z, \ldots$. The metrics t_d are *dependent variables* since they are the result of running testing with respect to varying $x, y, z, \ldots$. The parameters $x, y, z, \ldots$ are independent variables—so-called *factors*. Each can assume values—also called *levels*—within a predefined range.

When testing a user interface with respect to metric t_d, all independent variables need to be checked with respect to all their levels. Thus, the design space of the user interface, with respect to the given evaluation criterion, is the cross product of all factors. Its cardinality is the product of the cardinality of all level ranges: $\|T_d\| = \|X\| \times \|Y\| \times \|Z\| \times \cdots$. In a practical example, this means: test designers want to compare a novel multi-touch interface to a traditional mouse-based interface. This results in an independent variable UI-TYPE with levels MOUSE and MULTI-TOUCH. At the same time, the designers want to explore the benefit of sound and thus introduce an independent variable SOUND with levels SOUND-ON and SOUND-OFF. This creates $2 \times 2 = 4$ variants of user interfaces that need to be compared to one another in a statistical test procedure. If the designers were to include one more UI-TYPE level, PEN, the space of UI variants would extend to $2 \times 3 = 6$ variants. Evaluations have to test each of these variants in their experiments.

In addition to these *planned factors*, experiments may also be subject to unwanted—yet unavoidable—further factors, so-called *confounding factors*. Examples are learning effects, user fatigue or simulator sickness. This means that, if test persons are requested to participate in experiments for more than one variant, the sequencing of the variants may have an impact on the results since test persons may learn something about the scenario in the first test run (e.g. about the traffic situation in a driving simulation) that they can exploit during the second test run with a second variant of the user interface. They may thus perform better in the second test—not due to a superior user interface but due to learning effects. Conversely, fatigue or simulator sickness may have a negative impact on the results that also needs to be discounted: test persons may perform better in the first run than in subsequent runs.

To discount the effects of such confounding factors, the sequencing of the evaluations for different variants needs to be permuted between different test persons—requiring $n!$ different sequencing plans for n variants. For the given example of 6 variants, this means that $6! = 720$ test persons are needed for the evaluations, one for each permutation of the 6 variants. Without level PEN for factor UI-TYPE, only $4! = 24$ permutations need to be compared. This small example exemplifies the critical impact of introducing more levels to a factor: the resulting design space rapidly explodes to unmanageable numbers of user interface variants that all need to be compared systematically in the test design, requiring rapidly increasing numbers of test persons.

5.2.4 Experimental Structure

Critical to successful evaluations is the proper selection of test persons. These should correspond to the population of the targeted final users of an application. If the new user interface is expected to be helpful across many applications, test users must be drawn from a wide, diverse background. In most cases, it is not reasonable to recruit test persons only from the immediate, close circle of friends and co-workers since such group might be rather homogeneous regarding age, education, sex and experience with computers. On the other hand, some initial problems with a novel user interface might be so universal that they are criticized by nearly everybody—except for the developer of the interface. In such cases, first formative usability tests may be conducted with colleagues and friends who are more easily accessible than a non-biased, well-balanced broad group of representative test persons. When reporting on an evaluation, it is critical to describe the demographic constitution of the selected test group and to present the rationale why these people were selected.

As discussed in the previous Sect. 5.2.3, several variants of a user interface need to be compared. To this end, all variants have to be tested with the same depth, i.e.: test persons have to be organized in groups for each such variant. Two approaches exist for organizing such groups of test persons.

In a so-called *between-subject test design*, test persons are assigned to different groups, with each group testing exactly one user interface. This approach has the advantage that no learning effects of fatigue can occur since users participate in only one evaluation. A disadvantage of this approach is the need to balance all test groups such they all have a demographically similar distribution with respect to age, sex, etc. To ensure well-balanced (i.e., *unbiased*) test groups, a large number of test persons are required.

Alternatively, in a so-called *within-subject test design*, each test person is asked to work with all variants. In this setup, demographic bias is not as much of an issue—especially for initial formative evaluations. Yet, confounding factors such as learning or fatigue are a considerable problem (see Sect. 5.2.3). In order to discount biases due to confounding factors, the variants need to be presented to different test persons in permuted order. Still, the problem remains that test persons may be overly strained and exasperated from very long series of experiments (possibly with interleaved subjective questionnaires). Thus, the design of the test procedure per variant should be kept as short as possible.

5.3 Subjective Evaluations

When working with a particular user interface, *objective* measurements of user performance are generally accompanied by questionnaires. Increasingly, such *subjective* user feedback, is becoming more and more essential to the success of a novel product or computer system. Concepts such as user satisfaction and user experience are becoming central issues in user interface design.

Yet, it is not easy to measure subjective feedback from test persons. Research in psychology and human factors has established a number of standardized questionnaires that have been the result of thorough investigations how to pose questions such that individually differing degrees of emotions can be discounted.

The *NASA Task Load Index* (*TLX*) (Hart and Staveland [34]) measures mental workload. To this end, users are asked to grade the mental, physical and temporal demand of the system, as well as their appreciation of their own performance, and the amount of effort and frustration they experienced during the test. For each of these six criterions, test persons are asked to indicate their rating on a 20 point scale ranging from VERY LOW to VERY HIGH.

The *System Usability Scale* (*SUS*) (Brooke [25]) determines the effectiveness, efficiency and satisfaction of test persons working with a particular user interface. Users are requested to comment on ten standardized statements on a 5 point scale ranging from STRONGLY DISAGREE to STRONGLY AGREE, resulting in a score in the range from 0 to 100.

Finally, a method using a *Semantic Differential* (Osgood, Suci, and Tannenbaum [47]) and the *AttrakDiff* (Hassenzahl, Burmester, and Koller [35]) test use a list of opposing attribute pairs with a 5 point scale to elicit indications from the test persons, regarding ideas and affective attitudes which are associated with an interface.

5.4 Interviews and Anecdotal Use

Despite all attempts towards gathering objective or subjective measurements from test persons that can be quantitatively evaluated, verbal feedback and anecdotal usage are invaluables forms of in-depth information. Especially in the early phases of developing a novel human-computer interaction concept, a large number of issues are undefined. It is generally impossible to properly design a test setup that covers all of these issues. Careful inspection and recording of the test persons' every move and interrogating them about any observed moment of confusion or irritation sheds light on a sizeable number of essential problems that need to be mastered before large-scale systematic tests lead to conclusive results.

To this end, interviews are conducted in all stages of interface design and implementation. They can be associated with presentations, demonstrations and exhibits during expert reviews, small-scale usability studies, or random quests for user feedback. Interviews can be completely unstructured. Yet, already known issues are also cast into a systematic sequence of questions to be answered as part of the interviews by every interviewee.

6 Food for Thought

This chapter has reported on a large number of things that can go wrong in human-computer communication. Just as in human to human communication, there is much potential for misunderstanding. Humans can misinterpret computer presentations and animations, due to misleading metaphors or simply due to misled attention and overload. Conversely, machines can misinterpret human input commands, due to noisy sensor data or due to unprecisely performed human actions.

In conversations between humans, we are aware—to some extent—of potential misunderstandings, and we thus also communicate on a meta-level about the course of conversation with one another. We question, ascertain and reassure that the important issues have been clearly conveyed. We also express level of completeness when we simplify complicated matters for didactic matters such that the communication partner can comprehend an issue gradually over time. How can computer interfaces communicate on such a meta level, in parallel to conducting the principle exchange of information? This a topic of increasing importance, pertaining to issues of uncertainty analysis and uncertainty visualization.

Another important issue covers consciencious resource management—both on the human side and for the computer. Neither one has unlimited resources to perceive, interpret and act/present. There may be shortages of sensing/perception power, as well as memory and processing capacity. Human-computer interaction systems need to be aware of such resource limitations. Communication processes need to take explicitly into account that a communication partner may be currently overwhelmed by information and that it is, thus, better, to slow down and maybe even to keep quiet for a while. How can computers monitor resource shortages and adapt their communication strategy accordingly? This in another topic that is requires increasing attention.

7 Summary

This chapter has presented risks and issues of potential miscommunication on the basis of the interaction cycle by Bowman et al. [2]. For each step along this cycle, the chapter has discussed a number of critical issues and related them in associated tables to experiences that were made in the FAR-lab for when building and evaluating novel user interfaces for a number of applications. The chapter closes with a presentation of the most critical issues for planning proper test designs to evaluate novel human-computer interaction concepts in a human-centered approach.

Acknowledgements The discussion on user interfaces in this chapter is the result of nearly two decades of work on novel user interfaces related in particular to Augmented Reality. Many people have participated in the work and have influenced the authors in many ways. The authors (particularly the first author) would like to thank the many current and past members of the Fachgebiet Augmented Reality (FAR), as well as the labs at the Fraunhofer Institute for Computer Graphics (IGD) and the European Computer-industry Research Centre (ECRC) prior to FAR for sharing and jointly developing many of the ideas. The authors are grateful to many collaborators across a large number of publicly or privately funded research projects, such as TUMMIC, Presenccia, KAUST, Crumbs, FORLOG, Speedup, IGSSE, Trackframe, the Chair for Computer-Aided Medical Procedures (CAMP) and the Chair for Applied Software Engineering.

References

Selected Bibliography

1. R. Azuma, Y. Baillot, R. Behringer, S. Feiner, S. Julier, B. MacIntyre, Recent advances in augmented reality. IEEE Comput. Graph. Appl. **21**, 34–47 (2001). doi:10.1109/38.963459. http://portal.acm.org/citation.cfm?id=616073.618862
2. D.A. Bowman, E. Kruijff, J.J. LaViola, I. Poupyrev, *3D User Interfaces: Theory and Practice* (Addison-Wesley/Longman, Redwood City, 2004). ISBN 0201758679
3. F.P. Brooks Jr., The computer scientist as toolsmith II. Commun. ACM **39**, 61–68 (1996). http://doi.acm.org/10.1145/227234.227243
4. G.C. Burdea, P. Coiffet, *Virtual Reality Technology*, 2nd edn. (Wiley, Hoboken, 2003). ISBN 0471360899
5. T. Coskun, A. Benzina, E. Artinger, C. Binder, G. Klinker, User-centered development of UI elements for selecting items on a digital map designed for heavy rugged tablet PCs in mass casualty incidents, in *Proceedings of ACM SIGHIT International Health Informatics Symposium (IHI 2012)* (ACM, New York, 2012)
6. M.W. Eysenck, M.T. Keane, *Cognitive Psychology*, 6th edn. (Psychology Press, New York, 2010). ISBN 978-1-84169-539-6
7. R.L. Gregory, *Eye and Brain, the Psychology of Seeing*, 5th edn. (Princeton University Press, Princeton, 1997)
8. M. Huber, Parasitic tracking for augmented reality. Dissertation, Technische Universität München, München (Nov. 2011)
9. H. Ishii, B. Ullmer, Tangible bits: towards seamless interfaces between people, bits and atoms, in *CHI 97* (1997), pp. 234–241. citeseer.ist.psu.edu/ishii97tangible.html
10. R. Mukerjee, C.F.J. Wu, *A Modern Theory of Factorial Design*. Springer Series in Statistics (Springer, Heidelberg, 2006). ISBN 978-0387319919

11. P.G. Neumann, *Computer Related Risks* (ACM Press/Addison-Wesley, New York, 1995). ISBN 0-201-55805-X
12. D.A. Norman, *The Design of Everyday Things*, 2nd edn. First Basic (Perseus Books Group, Jackson, 2002). ISBN 0465067107
13. B. Shneiderman, C. Plaisant, *Designing the User Interface: Strategies for Effective Human-Computer Interaction*, 4th edn. (Pearson, Boston, 2005). ISBN 0321197860. http://www.gbv.de/dms/ilmenau/toc/492668051.PDF
14. R. Spence, *Information Visualization: Design for Interaction*, 2nd edn. (Prentice-Hall, Upper Saddle River, 2007). ISBN 0132065509
15. M. Tönnis, Towards automotive augmented reality. Dissertation, Technische Universität München, München (Nov. 2008)
16. A. van Dam, Post-WIMP user interfaces. Commun. ACM **40**, 63–67 (1997). http://doi.acm.org/10.1145/253671.253708
17. M. Weiser, Ubiquitous computing. Computer **26**, 71–72 (1993). http://doi.ieeecomputersociety.org/10.1109/2.237456
18. G. Welch, E. Foxlin, Motion tracking: no silver bullet, but a respectable arsenal. IEEE Comput. Graph. Appl. **22**(6), 24–38 (2002)
19. C.D. Wickens, J.G. Hollands, *Engineering Psychology and Human Performance*, 3rd edn. (Prentice-Hall, Upper Saddle River, 2000)

Additional Literature

20. E. Artinger, T. Coskun, S. Nestler, M. Mähler, Y. Yildirim-Krannig, F. Wucholt, F. Echtler, G. Klinker, Creating a common operation picture in realtime with user-centered interfaces for mass casualty incidents, in *Proceedings of the 4th International Workshop for Situation Recognition and Medical Data Analysis in Pervasive Health Environments (PervaSense), PervaSense'12* (2012). ICST.org
21. M. Bauer, Tracking errors in augmented reality. Dissertation, Technische Universität München, München (Sept. 2007)
22. R. Bauernschmitt, M. Feuerstein, J. Traub, E.U. Schirmbeck, G. Klinker, R. Lange, Optimal port placement and enhanced guidance in robotically assisted cardiac surgery. Surg. Endosc. **21**(4), 684–687 (2007). doi:10.1007/s00464-006-9057-z
23. B.B. Bederson, B. Shneiderman, *The Craft of Information Visualization: Readings and Reflections* (Morgan Kaufmann, San Francisco, 2003). ISBN 1558609156
24. A. Benzina, M. Tönnis, G. Klinker, M. Ashry, Phone-based motion control in VR: analysis of degrees of freedom, in *CHI Annual Conference on Human Factors in Computing Systems. Extended Abstracts (CHI EA)'11* (ACM, New York, 2011), pp. 1519–1524. ISBN 978-1-4503-0268-5. http://doi.acm.org/10.1145/1979742.1979801
25. J. Brooke, System usability scale (SUS): a quick-and-dirty method of system evaluation user information. Technical Report, Digital Equipment Corporation Ltd., Reading, UK (1986)
26. S.K. Card, T.P. Moran, A. Newell, The keystroke-level model for user performance time with interactive systems. Commun. ACM **23**(7), 396–410 (1980)
27. S.K. Card, T.P. Moran, A. Newell (eds.), *The Psychology of Human-Computer Interaction* (CDC Press, San Francisco, 1983)
28. H.M. Chandler, R. Chandler, *Fundamentals of Game Development* (Jones and Bartlett, Boston, 2009)
29. A. Dey, A. Cunningham, C. Sandor, Evaluating depth perception of photorealistic mixed reality visualizations for occluded objects in outdoor environments, in *3DUI* (2010), pp. 127–128
30. F. Echtler, Tangible information displays. Dissertation, Technische Universität München, München (Nov. 2009)

31. F. Echtler, F. Sturm, K. Kindermann, G. Klinker, J. Stilla, J. Trilk, H. Najafi, The intelligent welding gun: augmented reality for experimental vehicle construction, in *Virtual and Augmented Reality Applications in Manufacturing*, ed. by S. Ong, A. Nee (Springer, New York, 2003), Chap. 17
32. S. Feiner, B. MacIntyre, M. Haupt, E. Solomon, Windows on the world: 2D windows for 3D augmented reality, in *Proceedings of the 6th Annual ACM Symposium on User Interface Software and Technology, UIST '93*, New York, NY, USA (ACM, New York, 1993), pp. 145–155. ISBN 0-89791-628-X. http://doi.acm.org/10.1145/168642.168657
33. P.M. Fitts, The information capacity of the human motor system in controlling the amplitude of movement. J. Exp. Psychol. **47**(6), 381–391 (1954)
34. S.G. Hart, L.E. Staveland, Development of NASA-TLX task load index: results of empirical and theoretical research, in *Human Mental Workload*, ed. by P.A. Hancock, N. Meshkati (North Holland, Amsterdam, 1988)
35. M. Hassenzahl, M. Burmester, F. Koller, AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität, in *Mensch und Computer 2003: Interaktion in Bewegung*, ed. by G. Szwillus, J. Ziegler (Teubner, Leipzig, 2003)
36. M. Hassenzahl, R. Kekez, M. Burmester, The importance of a software's pragmatic quality depends on usage modes, in *Proceedings of the 6th International Conference on Work with Display Units (WWDU'02)*, Berlin, Germany (ERGONOMIC Institut fuer Arbeits- und Sozialforschung, Berlin, 2003), pp. 275–276
37. K.V. Iserson, J.C. Moskop, Triage in medicine, part I: concept, history, and types. Ann. Emerg. Med. **49**(3), 275–281 (2007)
38. J.A. Jones, J.E.I. Swan, G. Singh, S.R. Ellis, Peripheral visual information and its effect on the perception of egocentric depth in virtual and augmented environments, in *VR* (2011), pp. 215–216
39. P. Keitler, Management of tracking and tracking accuracy in industrial augmented reality environments. Dissertation, Technische Universität München, München (Apr. 2011)
40. T. Luhmann, Accuracy limits in photogrammetry, in *Proceedings of the Workshop Traceability in Large Scale Metrology* (2006)
41. A. MacWilliams, C. Sandor, M. Wagner, M. Bauer, G. Klinker, B. Brügge, Herding sheep: live system development for distributed augmented reality, in *ISMAR* (2003), pp. 123–132
42. P. Maier, M. Tönnis, G. Klinker, A. Raith, M. Drees, F. Kühn, What do you do when two hands are not enough? Interactive selection of bonds between pairs of tangible molecules, in *Proceedings of the 5th IEEE Symposium on 3D User Interfaces (3D UI)* (2010), pp. 83–90
43. B. Myers, S.E. Hudson, R. Pausch, Past, present, and future of user interface software tools. ACM Trans. Comput.-Hum. Interact. **7**, 3–28 (2000). http://doi.acm.org/10.1145/344949.344959
44. S. Nestler, Konzeption, Implementierung und Evaluierung von Benutzerschnittstellen für lebensbedrohliche, zeitkritische und instabile Situationen. Dissertation, Technische Universität München, München (July 2010)
45. G.M. Nielson, H. Hagen, H. Müller, *Scientific Visualization: Overviews, Methodologies, and Techniques* (IEEE Comput. Soc., Los Alamitos, 1997)
46. D.A. Norman, *The Invisible Computer* (MIT Press, Cambridge, 1998). ISBN 0262140659
47. C.E. Osgood, G.J. Suci, P.H. Tannenbaum, *The Measurement of Meaning* (University of Illinois Press, Champaign, 1957)
48. D. Pustka, M. Huber, C. Waechter, F. Echtler, P. Keitler, J. Newman, D. Schmalstieg, G. Klinker, Automatic configuration of pervasive sensor networks for augmented reality. IEEE Pervasive Comput. **10**(3), 68–79 (2011). http://doi.ieeecomputersociety.org/10.1109/MPRV.2010.50
49. J.P. Rolland, H. Fuchs, Optical versus video see-through head-mounted displays in medical visualization. Presence **9**(3), 287–309 (2000)
50. C. Sandor, A software toolkit and authoring tools for user interfaces in ubiquitous augmented reality. Dissertation, Technische Universität München, München (Oct. 2005)

51. C. Sandor, G. Klinker, A rapid prototyping software infrastructure for user interfaces in ubiquitous augmented reality. Pers. Ubiquitous Comput. **9**(3), 169–185 (2005)
52. B. Schwerdtfeger, Pick-by-vision: bringing HMD-based augmented reality into the warehouse. Dissertation, Technische Universität München, München (July 2010)
53. R.M. Sirkin, *Statistics for the Social Sciences*, 3rd edn. (SAGE, Thousand Oaks, 2006)
54. F. Steinicke, G. Bruder, K. Hinrichs, P. Willemsen, Change blindness phenomena for stereoscopic projection systems, in *Proc. IEEE Virtual Reality Conference (VR'10)* (2010), pp. 187–194
55. J.E.I. Swan, S.R. Ellis, A.B. Adelstein, Conducting human-subject experiments with virtual and augmented reality, in *Tutorial at the IEEE Virtual Reality Conference (VR'07)* (2007). http://www.cse.msstate.edu/~swan/teaching/tutorials/Swan-VR2007-Tutorial.pdf
56. S. Thrun, W. Burgard, D. Fox, *Probabilistic Robotics* (MIT Press, Cambridge, 2006)
57. M. Tönnis, A. Benzina, G. Klinker, Utilizing consumer 3D TV hardware for a flexibly reconfigurable visualization system. Technical report TUM-I-11-13, Technische Universität München (2011)
58. M. Tönnis, R. Lindl, L. Walchshäusl, G. Klinker, Visualization of spatial sensor data in the context of automotive environment perception systems, in *Proceedings of the 6th International Symposium on Mixed and Augmented Reality (ISMAR)* (2007)
59. E.R. Tufte, *The Visual Display of Quantitative Information*, 2nd edn. (Graphics Press, Cheshire, 2001). ISBN 0961392142. http://www.amazon.com/Visual-Display-Quantitative-Information-2nd/dp/0961392142%3FSubscriptionId%3D192BW6DQ43CK9FN0ZGG2%26tag%3Dws%26linkCode%3Dxm2%26camp%3D2025%26creative%3D165953%26creativeASIN%3D0961392142

Chapter 16
Translational Risk Models

Donna Pauler Ankerst, Vanadin Seifert-Klauss, and Marion Kiechle

With rapid progression of computing and other technological advances, the practice of modern medicine has moved from primarily anecdotal to largely quantitative. With due credit to the Internet and the new cyber-society, individuals have taken a more active role in the decision-making process concerning their health, from deciding whether or not to get screened for a disease to which treatment is best for their specific clinical profile. Treating physicians are more connected with latest medical breakthroughs through vast dissemination via the Internet. Statistical prediction models assembled on large well-designed cohorts, multiply validated and easily accessible through online calculators play a role in translating basic science results to implementation in the community for public health benefit. This chapter describes the risk model building process that forms the basis of modern medical decision-making, from statistical estimation to validation and implementation on the Internet. The early diagnosis of cancer is used as the context to illustrate principles, though the concepts immediately transcend to other disciplines as concluding examples in forestry and finance will show.

Keywords Logistic regression · Calibration · Discrimination · Prediction · Validation

D.P. Ankerst (✉)
Biostatistics, Center for Mathematical Sciences, Technische Universität München,
Boltzmannstr. 3, 85748 Garching bei München, Germany
e-mail: ankerst@tum.de

D.P. Ankerst
Health Science Center at San Antonio, University of Texas, 7703 Floyd Curl Drive, San Antonio,
TX 78229, USA

V. Seifert-Klauss
Gynaecology, Department of Medicine, Klinikum Rechts der Isar, Technische Universität
München, Ismaninger Str. 22, 81675 Munich, Germany

M. Kiechle
Chair of Gynaecology, Department of Medicine, Klinikum Rechts der Isar, Technische
Universität München, Ismaninger Str. 22, 81675 Munich, Germany

C. Klüppelberg et al. (eds.), *Risk – A Multidisciplinary Introduction*,
DOI 10.1007/978-3-319-04486-6_16,

The Facts

- Logistic regression for predicting disease from multiple risk factors, or more generally a dichotomous outcome from covariates, will be covered.
- Latest developments in external validation, including its distinct components of discrimination, calibration and net benefit, will be reviewed.
- How risk prediction tools get converted to online risk calculators will be discussed.

1 Introduction

Risks provide the currency by which doctors, patients and individual members of the general population communicate and make informed decisions regarding health. Examples of daily media encounters with risk include claims that diets rich in fruits and vegetables reduce the risk of heart disease, smoking increases the risk of lung cancer, or one glass of red wine per day reduces blood pressure, just to name a few. Increasingly it has been recognized that risks agglomerate from a multitude of factors rather than being the product of any single factor acting in isolation, for example, that both diet and exercise work more effectively in combination to reduce the risk of cardiovascular disease. Experience has also revealed that there is uncertainty underpinning estimates of risk, in other words, that one study may undo a previous study report on risk. Most people obtain information concerning health risk either passively through the media, or more actively, through the Internet. These sources in turn obtain their information from peer-reviewed published scientific studies and hence often serve as the translators of basic science to public use. The scientific studies have typically involved observation of a cohort or group of voluntary participants under the relevant controlled or uncontrolled environments of a clinical trial or observational study, respectively, followed by subsequent observation of outcome and an observed statistically significant association between risk factors, interventions and outcomes. Statisticians, epidemiologists and other quantitative scientists scrutinize the findings from such cohorts, determining which biases may have been at play that could ultimately limit validity of the findings or generalizability to populations beyond that on which the studies have been performed. For example, a risk model constructed primarily from people of one ethnicity may not apply to people of other ethnicities. They build risk models when applicable, and validate them in other cohorts, a process sometimes taking years beyond the already many years invested in conducting the original study collecting the data and unfortunately, sometimes resulting in failure to validate, thus limiting the scientific impact of the original study reporting a positive finding.

This chapter describes the process beginning with the end of a published study reporting a significant association between risk factors and outcomes and ending with implementation of a risk prediction model for public use via the Internet. The general concept of building a risk model applies to a vast variety of applications,

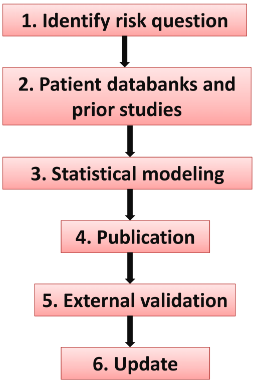


Fig. 1 The risk building paradigm

data and statistical models, the details of which cannot be delivered in a single chapter or even a single textbook. Hence, to concretely illustrate the general concepts, a specific application of the online Prostate Cancer Prevention Trial Risk Calculator (PCPTRC) for detection of prostate cancer will be used throughout. In later sections of the chapter, an overview of the most prominently used risk models in prostate, breast, lung, colorectal, and ovarian cancer will be reviewed, followed by generalizations to other diseases and other disciplines, such as ecology, business and finance.

2 Building Risk Models

The fundamental paradigm for building risk models is shown in Fig. 1 and proceeds as follows. First a population- or clinically-relevant risk question is identified. Second, the appropriate data sources are located that could be used to address the risk question. For example, existing data repositories could be mined retrospectively to see what risk factors led to development of a disease or outcome, or a new study could be designed to prospectively follow individuals. Once the data are collected, the third step requires statistical analysis, entailing data cleaning, fitting of an appropriate model, selection of covariates to include in the model, adjustments for potential biases in the data collection and then internal testing of model performance. Once this is complete, the model is published in order to disseminate the results, for the media, for further validation, which is the fifth step, and hopefully ultimately for use by the public. Technologies and risk factors change over time and new biomarkers of disease (biological entities that can be measured in the blood or urine) or new risk factors are continually discovered. Therefore the last step of the process is the continual task of keeping a risk model contemporary. To give an example of the process a brief overview of the PCPTRC from its conception to ongoing efforts to update as new biomarkers for prostate cancer are discovered is illustrated. Details for the specific steps of Fig. 1 will be more thoroughly described in subsequent sections of the chapter.

Example 2.1 We illustrate the risk building paradigm through the PCPTRC.

(1) (*Identify risk question*) Prostate cancer has the highest incidence of all cancers affecting U.S. men and is the second leading cause of cancer-related death behind lung cancer [10]. Prostate-specific antigen (PSA) is the leading blood test used for the early detection of prostate cancer and it is now common for men over 55 years of age to undergo routine screening for prostate cancer. Of concern to older men is given their PSA values and other clinical test results, what is their risk of prostate cancer? If sufficiently high, then they might be advised to undergo prostate biopsy, a more invasive diagnostic procedure.

(2) (*Patient databanks and prior studies*) The PCPTRC was developed based on analysis of data from 5519 placebo arm participants who had undergone annual PSA and digital rectal examination (DRE) screening as part of the 7-year Prostate Cancer Prevention Trial (PCPT) [24]. All PCPT participants were requested to undergo prostate biopsy, both during the trial when prompted by a PSA value exceeding 4 ng/mL or abnormal digital rectal exam (DRE) result and at the end of the trial regardless of PSA and DRE findings. The latter aspect made the PCPT cohort unique in the world in having prostate cancer status ascertained by biopsy even among men who did not meet the clinical criteria for recommendation to biopsy.

(3) (*Statistical modeling*) For predicting prostate cancer outcome all potential risk factors measured on participants during the trial were identified, including age, family history of prostate cancer in a first degree relative, whether or not a prior prostate biopsy had been performed that was negative for prostate cancer, race, ethnicity, and PSA and DRE outcomes within one year prior to the biopsy result used in the analysis. Participants could have multiple biopsies up until either a positive cancer diagnosis or the end-of-study required biopsy; only the last biopsy of each participant was used. Logistic regression was used to statistically model the association between the multiple risk factors to the outcome, prostate cancer or not, on biopsy. A separate logistic regression was performed for the association of risk factors to high grade prostate cancer, defined as prostate cancer with Gleason grade ≥ 7. High grade cancer is a particularly aggressive form of cancer that is more often associated with mortality.

(4) (*Publication*) The PCPTRC appeared online as soon as the algorithm for the PCPTRC appeared by [24].

(5) (*External validation*) Accuracy of the PCPTRC has been validated in a range of external populations, from healthy populations undergoing annual screening with men referred to prostate biopsy for elevated PSA or abnormal DRE, similar in art to the PCPT [16], to clinical populations where men underwent biopsy based on clinical symptoms [5, 7, 8, 15].

(6) (*Update*) The PCPT was enhanced in 2008 to include a new urine marker for prostate cancer, PCA3 [2] and due to the online posting of the updated calculator, soon thereafter externally validated [18]. It has recently been updated to include the biomarkers percent free PSA and [−2]proPSA (two recently discovered relatives of PSA that are also measurable in the blood) and externally

validated [3]. Minor updates to the PCPTRC were made to tailor for men currently taking finasteride [25] and to incorporate body mass index [12].

3 Statistical Models

Risk calculators can predict a range of outcome types, such as the probability of having prostate cancer on biopsy, of developing breast cancer in the next 5 years, or surviving past 10 years post-treatment for a disease. Accordingly they are built on the appropriate statistical model for the outcome of interest. For example, Cox's proportional hazards model is a popular model for predicting times to an event with possible censoring (end of follow-up time preceding occurrence of the event) because it incorporates risk factors and makes minimal assumptions on the baseline event hazard rate. Logistic regression is commonly used to predict binary events, and will be used to illustrate the principles throughout this chapter. The principles of model selection and testing applied to logistic regression easily extend to other statistical models for other outcome types.

Logistic regression is a method for relating multiple risk factors $X_1, \ldots, X_p$ assembled in a vector $X = (X_1, \ldots, X_p)$ to a dichotomous outcome Y which will be assumed to take the value of 1 for a bad outcome, such as disease and 0 for the opposing good outcome, such as no disease. Specifically it relates the log odds of the bad outcome ($Y = 1$) to the risk factors X through the relationship:

$$\log \frac{P(Y=1|X)}{1-P(Y=1|X)} = \alpha + \beta' X, \tag{1}$$

where in the formula, log denotes the natural logarithm (base e), α is an intercept, and β a vector of log odds ratios, one for each risk factor assembled in X.

To understand why β defines log odds ratios, it is helpful to consider the simple scenario of just one dichotomous risk factor, X, taking the value 1 for an unfavorable risk factor value versus 0 for a favorable risk factor value. Based on the logistic model, the odds of the bad outcome based on the single risk factor X is:

$$\frac{P(Y=1|X)}{1-P(Y=1|X)} = \exp\{\alpha + \beta X\},$$

where exp denotes the exponential function. This equation implies that for the individual with risk factor X, the probability of the bad outcome is a multiple, $\exp\{\alpha + \beta X\}$, times the probability of the good outcome. The odds of the bad outcome for individuals with the unfavorable risk factor ($X = 1$) and favorable risk factor ($X = 0$) are given by:

$$\frac{P(Y=1|X=1)}{1-P(Y=1|X=1)} = \exp\{\alpha + \beta\}, \qquad \frac{P(Y=1|X=0)}{1-P(Y=1|X=0)} = \exp\{\alpha\},$$

respectively. From these expressions one might expect β to be greater than 0 since individuals with the unfavorable risk factor should have a higher probability, and hence odds, of the bad outcome than individuals with the favorable outcome. The ratio of odds for individuals with the unfavorable ($X = 1$) to favorable ($X = 0$) risk factor describes the magnitude by which the odds of the bad outcome accordingly changes:

$$\frac{\frac{P(Y=1|X=1)}{1-P(Y=1|X=1)}}{\frac{P(Y=1|X=0)}{1-P(Y=1|X=0)}} = \frac{\exp\{\alpha+\beta\}}{\exp\{\alpha\}} = \exp\{\beta\}.$$

The simplified expression, $\exp\{\beta\}$, is the odds ratio (OR) for individuals with the unfavorable compared to favorable risk factor.

For the case of a single predictor X that is continuous rather than dichotomous, the OR gives the ratio of odds of outcome Y for a unit-increase in X (to see this compute the OR for $X = x + 1$ compared to $X = x$). An OR > 1 implies that an increase in the risk factor increases the odds of the bad outcome, OR < 1 means it decreases the odds and OR $= 1$ means it has no impact. From the relationship above, $\beta = \log\{\exp\{\beta\}\}$ is the log odds ratio (log OR), and values of $\beta > 0$, <0, and $=0$ have the same interpretations as for OR > 1, <1 and $=1$, respectively.

If X were a categorical risk factor with more than two levels, such as race with levels African American, Caucasian, and Other, the logistic model can still be fit by choosing one level as a reference (say Caucasian) and then returning two odds ratios, one for the comparison of each of the remaining levels, African Americans and Other, to the reference. In many aspects such as this logistic regression operates similarly as for linear regression. In the general case of multiple risk factors (1), $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ is the vector of respective log OR's for each of the multiple risk factors comprising $X = (X_1, X_2, \ldots, X_p)$:

$$\log \frac{P(Y=1|X)}{1-P(Y=1|X)} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p.$$

The interpretation of each parameter β_i is the log OR corresponding to a unit increase in the respective risk variable X_i, with all other risk variables in the model held constant.

Statistical packages return estimates of log odds ratios (β's), their standard errors, and p-values for tests of the null hypotheses that they equal 0 (no effect) versus two-sided alternative hypotheses that they do not equal 0. From these approximate 95 percent confidence intervals for log ORs can be constructed as (estimated log OR) ± 1.96 × (standard error); to obtain estimates and confidence intervals for the OR's, take the exponent of the estimates and 95 percent confidence interval bounds, respectively.

Typically many risk factors or individual characteristics, including demographic, environmental or other variables are available for potential inclusion in the model and additionally more complicated relationships between the variables and outcomes can be modeled using transformations and interactions. Therefore, a variety of model selection techniques are available in statistical packages, many of which automatically sort through large numbers of models. Some of the most commonly used model selection techniques are based on finding the model with the lowest Akaike's information criterion, AIC $= -2 \times$ maximized log likelihood $+ 2 \times$ number of parameters, or the lowest Bayesian Information criterion, BIC = $-2 \times$ maximized log likelihood $+ \log(\text{sample size}) \times$ number of parameters [1, 19]. The first terms of both criteria seek to find the model maximizing goodness of fit to the developmental data set, while the second terms penalize for over-parameterization, with the BIC tending to penalize more, and hence selecting smaller models with fewer parameters than the AIC on average.

As a preliminary indication of how the model may validate, internal validation can be performed by splitting the dataset into a training and test set or into a group of equally sized subsets that alternatively serve as training and test sets. A logistic regression model is fit from scratch on the training set and then evaluated on the test set using any one of the metrics to be defined later for external validation. For multiple splits of the dataset, test set performances are simply averaged. Bootstrapping, repeated random sampling with replacement of test and training sets, can also be used.

Example 3.1 The BIC and average cross-validated area underneath the receiver operating characteristic curve (AUC) were used to find the optimal multivariable logistic regression model relating potential risk factors to prostate cancer outcome on biopsy on data from the 5519 PCPT placebo arm participants used to develop the PCPTRC [24]. The AUC is a rank-based measure of how well a risk model discriminates the bad outcomes it aims to predict from the good outcomes and will be further defined in later sections. For the PCPT, 4-fold internal cross-validation was implemented, whereby the developmental dataset of 5519 observations was randomly partitioned into four subsets, three of size 1380 and one of size 1379, with randomization stratified to keep the proportion of prostate cancer cases between 20 % and 23 % in each subset. Over 50 models, some including two-way interactions, were evaluated by a combination of forward, backward and stepwise selection and subjective measures, such as including only statistically significant effects at the 0.05 level. BIC and cross-validated AUC values were tabulated for each of the models. The model with the lowest BIC value contained only main effects and no interactions among risk factors and was also one of the models with lowest AUC values. Therefore this model was selected to form the PCPTRC. The final selected logistic regression model contained four risk factors: PSA (OR = 2.34 for logPSA), DRE (OR = 2.47), family history of prostate cancer (OR = 1.31) and history of a prior negative prostate biopsy (OR = 0.64). All were statistically significant with a p-value less than 0.001 except for family history with a p-value of 0.002 [24].

4 External Validation

Once a risk model has been constructed, it is critical to evaluate its performance on a population independent to that on which it was developed. Internal validation, evaluating the model on the same population as on which the model was developed, even though it has been split into separate training and test sets, is not enough, since unmeasurable cohort effects will still favorably bias the performance of the model compared to what might be achieved in a completely distinct cohort collected elsewhere. A variety of evaluation methods for risk models have been proposed in the literature, and these can be grouped into those that measure discrimination, calibration, or both. Recent reviews detangle the different objectives of the many metrics currently employed to evaluate risk prediction models [20, 21]. All of these metrics require an external validation cohort or data set, whereby all individuals in the cohort have all risk factors X required for evaluation of the risk prediction model and the true outcome Y.

For missing covariates X, [9] showed by simulation that imputation results in less biased estimates of validation metrics than other currently used practices of either excluding the entire patient from the analysis or throwing the covariate out of a model. The current state of the art in imputation for X is based on specification of full conditional distributions for missing covariates and termed Multivariate Imputation by Chained Equations (MICE) and implementable in the R statistical package [26]. For missing outcomes Y in logistic regression, verification bias algorithms, which repeatedly impute the missing Y values using the assumed logistic regression form can be used if the missing data mechanism is assumed to be missing-at-random (MAR), meaning that the reason for missing data does not depend on the missing outcome value [4, 23].

4.1 Discrimination

Discrimination metrics focus on how well risk prediction models perform if used as the basis for making binary decisions as to whether individuals will have bad or good outcomes, sometimes referred to as hard classifications. Moving from a risk prediction, varying from 0 % to 100 %, to a positive versus negative decision on the bad outcome requires selection of a threshold r such that a risk above r corresponds to a positive test and below r, a negative test. The misclassification rate, or number of wrong test results made, is calculated separately for the subpopulations with bad and good outcomes.

How successfully the risk prediction predicts bad outcomes is termed sensitivity and on the external validation set is estimated by the percent positive tests among the bad outcomes:

$$\text{Sensitivity}(r) = \frac{\text{Number of bad outcomes with risk} > r}{\text{Number bad outcomes}},$$

where sensitivity is indexed by r as a reminder that it depends on the user-selected threshold r. How successfully the risk prediction tests negative for the good outcomes is termed specificity and is accordingly estimated by:

$$\text{Specificity}(r) = \frac{\text{Number of good outcomes with risk} \leq r}{\text{Number good outcomes}}.$$

The higher the sensitivity and specificity at any threshold r the better the risk prediction tool is. The problem is that as r increases from 0 % to 100 % specificity increases from 0 % to 100 % while sensitivity decreases from 100 % to 0 %, so that finding the threshold r that simultaneously optimizes sensitivity and specificity is difficult to achieve in practice. Sensitivity is often referred to as the true positive rate (TPR), one minus the sensitivity as the false negative rate (FNR) and one minus the specificity as the false positive rate (FPR).

The receiver operating characteristic (ROC) curve provides a summary of sensitivity and specificity for all choices of r ranging from 0 % to 100 %; it typically displays sensitivity on the y-axis and false positive rates on the x-axis, with both axes ranging from 0 % to 100 % [22]. The higher the ROC curve, the better its capacity for distinguishing bad from good outcomes. An appealing feature of ROC curves is that they are invariant with respect to measurement scales, for example, risks and the logits of risks (1) will yield the same ROC curve. This makes ROC curves particularly useful when comparing tests on completely different measurement scales, for example for directly comparing risk predictions from a model to the leading risk factor or covariate in the risk prediction model. Finally, as rank-based measures, ROC curves are by definition independent of disease prevalence in external validation set and hence can be applied to the case-control study situation in addition to prospective studies. An interesting single summary of the ROC curve is the area under the ROC curve (AUC), which in addition, conveniently holds the intuitive definition as the probability that a randomly chosen individual with a bad outcome has a higher risk prediction than a randomly chosen individual with a good outcome. The AUC ranges from a minimum at 50 %, implying predictive power of the risk prediction tool no better than flipping a coin to a maximum of 100 % for a perfectly discriminating risk prediction tool.

As seen by their formulas sensitivities and specificities for each threshold r can be calculated by just computing the appropriate sample proportions in the external validation set. The AUC is equivalent to the non-parametric Mann-Whitney or Wilcoxon rank sum statistic for comparing two populations and is hence easily computable using standard statistical software. The Wilcoxon test can be used for testing the null hypothesis that the AUC equals 0.5 versus the alternative that it exceeds 0.5. External packages can be imported into the statistical package R for computing the AUC and for performing various statistical tests for comparing AUCs of multiple tests.

Example 4.1 In 2009 the generalizability of the PCPTRC, which had been developed on a cohort of primarily Caucasian, healthy and elderly men, for potential applicability to other populations was investigated. The Early Detection Research Network (EDRN) clinical cohort comprised 645 men, some younger than members of the PCPT cohort, who had been referred to multiple urology practices across 5 states in the northeastern U.S. and had received a prostate biopsy due to some clinical indication, including persistent elevated PSA or abnormal DRE [7]. PCPTRC risks were calculated for each member of the EDRN cohort and compared to the actual clinical outcome on biopsy using sensitivities, specificities and the AUC. The PCPTRC demonstrated statistically significant superior discrimination for detecting prostate cancer cases compared to PSA (AUC = 69.1 % compared to 65.5 %, respectively, p-value = 0.009), and the ROC curve for the PCPTRC consistently fell at or above that for PSA for all false positive rates, with the greatest difference for false positive rates less than 25 %. For example, the thresholds of the PCPT Risk Calculator and PSA which obtained a false positive rate of 20 % were 48.4 % and 6.9 ng/mL, respectively (Table 2 of [7]). One can view these as two alternative tests for referral to further intensive diagnostic testing by prostate biopsy, each with equal specificities: the PCPTRC refers a patient to prostate biopsy if his PCPT risk exceeds approximately 50 % and the PSA test if his PSA exceeds 6.9 ng/mL. If these two diagnostic tests had been implemented in the EDRN population to "rule in" patients who should undergo prostate biopsy and "rule out" patients who should not, the PCPTRC would have correctly referred 47.1 % of the prostate cancer cases (sensitivity) and the PSA test 35.4 % of the prostate cancer cases. Although better than PSA, the PCPTRC would still have missed 50 % of the prostate cancer cases! Insisting that 80 % of prostate cancer cases get caught for both tests would have meant that the thresholds for referral would have had to be lowered, to 38.0 % and 4.0 ng/mL, for the PCPTRC and PSA test, respectively (Table 3 of [7]). But this would have approximately halved the specificity of both tests, to 40.3 % for the PCPT Risk Calculator and 44.1 % for PSA. In other words, approximately 60 % of the men who did not have prostate cancer would have been referred to a prostate biopsy unnecessarily (false positive rate), an error rate unacceptable from a public screening perspective.

4.2 Calibration

Calibration concerns itself with how close predicted risks from a model are to actual risks observed in an external validation population. Observed risks should match predicted risks among homogenous groups defined by the same risk profile. However, obtaining an external validation set large enough to have enough individuals with the same risk factors in order to make comparisons quickly becomes infeasible as the number of risk factors increases; hence approximations are made by further grouping.

One of the most commonly used calibration tests is based on an approximation to Pearson's chi-square goodness-of-fit test recommended by Lemeshow and Hosmer [11]:

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - n_i\pi_i)^2}{n_i\pi_i(1-\pi_i)},$$

where the validation set has been partitioned into k equally-sized groups (typically $k = 10$ with groups defined by deciles of the predicted risks in the validation set), O_i are the observed numbers of bad outcomes in the groups, n_i the observed numbers of participants in the groups, and π_i the mean risks of the groups. Under the null hypothesis, observed outcomes O_i are close to expected outcomes $n_i\pi_i$, hence X^2

should be small. Asymptotically under the null hypothesis, the X^2 statistic follows a chi-square distribution with k degrees of freedom.

An alternative measure of calibration, which measures reliability, was proposed by [6] and elaborated upon in [13]. The approach requires logistic regression of the outcomes ($Y_i = 0$ good outcome, $Y_i = 1$ bad outcome) on the logit of the predicted risks (π_i) as covariates for the $i = 1, \ldots, N$ individuals in the validation set:

$$\log \frac{P(Y_i = 1)}{1 - P(Y_i = 1)} = \alpha + \beta \log \frac{\pi_i}{1 - \pi_i}.$$

A perfect match of predicted to actual risks would occur when $\alpha = 0$ and $\beta = 1$. Therefore, a test of the composite null hypothesis $H_0 : \alpha = 0, \beta = 1$ provides an overall reliability test for the predictions. More specifically, the intercept α controls the calibration of the model, which is most clearly seen when $\beta = 1$. When $\beta = 1$, $\alpha < 0$ implies the predicted risks are too high and $\alpha > 0$, too low. When $\beta \neq 1$, noting that for $\pi_i = 0.5$:

$$\log \frac{P(Y_i = 1)}{1 - P(Y_i = 1)} = \alpha,$$

one can interpret the intercept α as a calibration measure at $\pi_i = 0.5$. The slope parameter β is referred to as the refinement parameter: $\beta > 1$ implies the predicted risks do not vary enough, $0 < \beta < 1$ they vary too much, and $\beta < 0$ they show the wrong direction. Therefore, additional tests of calibration given appropriate refinement, $H_0 : \alpha = 0 | \beta = 1$, and of refinement given appropriate calibration, $H_0 : \beta = 1 | \alpha = 0$, can be performed.

Example 4.2 In [7] it was reported that the average PCPTRC risk over all 645 men of the EDRN cohort was 45.1 %, which is fairly high in keeping with the nature of the cohort as elicited from multiple Urology practices. As a first indication of calibration the average PCPTRC risk among the cohort should correspond to the actual percent of the cohort that did have prostate cancer on biopsy. The percentage of the 645 men in the EDRN cohort diagnosed with prostate cancer was 43.4 %, fairly close to the average PCPTRC risk, providing a crude indication of calibration.

As an exploratory and primarily descriptive analysis of calibration among specific risk groups, Table 4 of [7] assessed the degree to which the PCPTRC calibrated to actual risks for specific subgroups, such as for Caucasians, African Americans, men with a positive family history and men with PSA less than 4.0 ng/mL. Across all subgroups the average PCPTRC risk never varied by more than approximately 5 or 6 percentage points from the observed risk but there were some subgroups where PCPTRC risks were better calibrated to actual risks than others. For example, among the 47 African American participants in the cohort, 51.1 % had prostate cancer but the average PCPTRC risk among these men was only 45.4 %. Application of the Lemeshow and Hosmer test of calibration yielded a p-value of 0.10, not rejecting the null hypothesis of a good fit at the 0.05 level of statistical significance. Cox's logistic regression of observed prostate cancer status on logits of predicted PCPTRC risks was also performed. The composite hypothesis test of reliability was not performed; however, the intercept from the logistic regression was estimated as -0.014

with standard error 0.091 and the slope by 1.291 with standard error 0.159. Separate 95 % confidence intervals for these estimates overlapped with 0 and 1, respectively, indicating that predicted PCPTRC risks were reliable estimates of observed risks in the EDRN population.

4.3 Net Benefit

Discrimination and calibration metrics objectively summarize accuracy but do not provide information as to which thresholds of a prediction model might be useful for basing clinical decisions. Towards this end, Vickers and Elkin [27] proposed a measure of net benefit justified through a layman's decision analysis framework that does not rely on user-specified costs associated with various outcomes as full-blown decision analyses typically do. As with the other accuracy measures, net benefit is evaluated on an external cohort to the one on which the risk model was developed as the expectation over the true and false positive counts:

$$\begin{aligned}\text{NetBenefit}(\text{Cohort}, r) = {} & \frac{\text{True Positive Count}(\text{Cohort}, r)}{\text{Sample Size}(\text{Cohort})} \\ & - \frac{\text{False Positive Count}(\text{Cohort}, r)}{\text{Sample Size}(\text{Cohort})}\left(\frac{r}{1-r}\right),\end{aligned}$$

where for emphasis dependencies on the chosen cohort and user-selected cutoff r are included in the definitions. The expression for the net benefit can be rewritten to show that it is also a function of the discrimination measures sensitivity and 1-specificity, TPR(Cohort, r) and FPR(Cohort, r), respectively, evaluated on the cohort and weighted by the proportions of bad outcomes (% Bad Outcomes(Cohort)) and good outcomes (% Good Outcomes(Cohort)) in the cohort:

$$\begin{aligned}\text{NetBenefit}(\text{Cohort}, r) = {} & \text{TPR}(\text{Cohort}, r) \times \%\ \text{Bad Outcomes}(\text{Cohort}) \\ & - \text{FPR}(\text{Cohort}, r) \times \%\ \text{Good Outcomes}(\text{Cohort}) \\ & \times \left(\frac{r}{1-r}\right).\end{aligned}$$

This expression illustrates further the dependence of the net benefit on the evaluation cohort. The discrimination metrics TPR and FPR already tend to depend on the cohort, net benefit further relies on how prevalent the disease is in the evaluation cohort. In other words, for two cohorts with the same operating characteristics of a prediction model, the cohort with a higher disease prevalence will demonstrate higher net benefit for using the prediction tool for clinical decisions.

Vickers and Elkin suggested evaluating the net benefit over all possible thresholds r of the prediction model ranging from 0 to 1. The specific numbers obtained for the net benefit can be difficult to interpret in isolation so they also recommended overlaid decision curves for the strategies of referring no patients to action or all patient's to action regardless of the threshold r selected. For these curves the last expression $[r/(1-r)]$ remains the same but the TPR and FPR are calculated based on the test rule that assigns no patients positive (in other words, $r > 1$) and all patients positive (in other words, $r < 0$). For taking no action, the TPR and FPR are identically 0 so the net benefit curve for taking no action is the horizontal line at 0 across all thresholds r. For taking action on all patients the TPR and FPR are 1 and the net benefit curve is % Disease(Cohort) − % NotDisease(Cohort)$[r/(1-r)]$, which seemingly ironically, still depends on the threshold r, but that is an artifact from the derivation of relative values of false positive results used in the derivation of the net benefit.

4.4 Overall Performance Measures

There are overall measures that combine discrimination and calibration that have been proposed for evaluating risk models but these have not gained widespread use largely for two reasons. Firstly, they have awkward properties because of the dichotomous nature of the outcome predicted by a continuous measure and secondly, they do not have an intuitive clinical interpretation.

The ubiquitous R^2 measure of proportion of variability explained by a linear regression of a continuous outcome Y on a series of variables has been extended to the case of generalized linear models, including logistic regression, where Y is dichotomous in the form of Nagelkerke's R^2 [14]:

$$R^2 = 1 - \exp\left[-\frac{2}{n}\left\{l(\hat{\beta}) - l(0)\right\}\right] = 1 - \left[\frac{L(0)}{L(\hat{\beta})}\right]^{2/n},$$

where $L(\cdot)$ and $l(\cdot)$ are the likelihood and loglikelihood functions, respectively, defined at the maximized values of β, the logistic regression log odds ratios, and for a null model with no covariates ($\beta = 0$). The problem with this measure is that for dichotomous outcomes it has a maximum less than 1, so is not as easy to judge as for continuous outcomes, where the maximum of R^2 is 1. Modifications by the max obtainable R^2 have been proposed but these are awkward to implement in practice. Therefore the criterion has not become widely used outside the case of linear regression with Normal outcomes.

A similar metric extended for dichotomous outcomes that has not found widespread use is the Brier score, which is simply the squared difference between the 0–1 Y outcomes and predictions from the model:

$$\text{Brier score} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{p}_i)^2.$$

Predictions are good if the Brier score is small but squared Euclidean distance between a dichotomous outcome Y and a continuous predictor p is not intuitive and will give coarse results for small sample sizes n. The Brier score also obtains a maximum less than one and similarly, attempts to correct it are awkward [21].

4.5 Integrated Discrimination Index

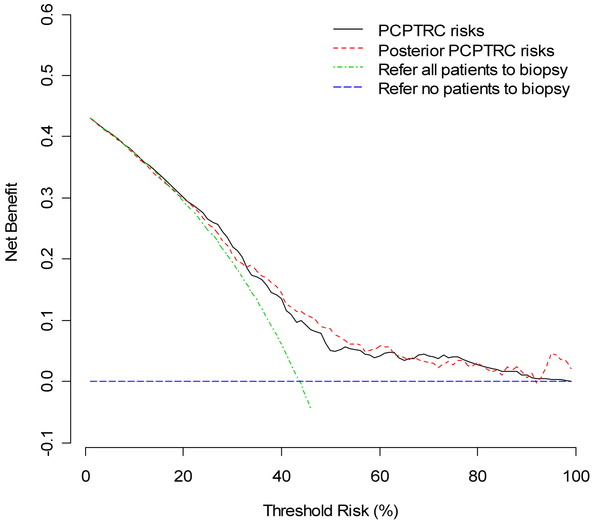
Noting shortcomings in the AUC for comparing risk prediction tools, Pencina and colleagues [17] proposed the integrated discrimination index (IDI) for comparing risk predictions from a new model to risk predictions from an old model that is simply the difference in discrimination slopes between the new and old predictions as proposed by Yates [28]:

$$\begin{aligned} IDI = & \left(\frac{1}{n_{events}} \sum_{i=1}^{n_{events}} p_{new,i} - \frac{1}{n_{nonevents}} \sum_{i=1}^{n_{nonevents}} p_{new,i} \right) \\ & - \left(\frac{1}{n_{events}} \sum_{i=1}^{n_{events}} p_{old,i} - \frac{1}{n_{nonevents}} \sum_{i=1}^{n_{nonevents}} p_{old,i} \right), \end{aligned}$$

where n_{events} are the number of events, or bad outcomes, and $n_{nonevents}$ are the number of non-events, or good outcomes, and the summations sum over the predicted probabilities from the new and old models as subscripted on the n's. The logic of the IDI is clear, a good prediction model should provide higher estimated risks among the bad outcomes in the validation set compared to the good outcomes, how good is determined by the discrimination slopes of the models. A positive IDI would indicate a new model has better discrimination slope than the old.

Example 4.3 Ankerst and colleagues [30] have developed an extension of the PCPTRC to incorporate the novel prostate cancer markers % freePSA and [−2]proPSA, which are both obtainable by blood tests. The methodology for updating the PCPTRC for new markers is based on Bayes algorithm for updating the prior odds of prostate cancer, which in this case are based on PCPTRC risks, via likelihood ratios of the distributions of the new marker among prostate cancer cases and controls to obtain posterior odds and hence updated posterior risks for prostate cancer; for more details see [29]. The updated PCPTRC is now available online at the same location as the PCPTRC. A developmental dataset of 474 participants in the San Antonio Biomarkers of Risk (SABOR) study were used to build the updated PCPTRC and the model was validated on an external EDRN dataset comprising 575 men. The IDI for comparing the new updated PCPTRC incorporating the two new markers to the standard PCPTRC evaluated on the EDRN validation set was 6.3 % (95 % CI 3.0 to 9.6 %), indicating a statistically significant positive improvement to using the updated model. Figure 2 compares the net benefit curves of the updated PCPTRC model (called posterior PCPTRC risks), the original PCPTRC, and the strategies of simply referring all men or no men to prostate biopsy irrespective of any risk prediction model. The benefit curves indicated benefit of using both the PCPTRC and updated PCPTRC for situations where risk thresholds exceeding 20 % for both of these rules would be used for referral to biopsy over the blanket rule of referring all men in the EDRN cohort to prostate biopsy, but no clear benefit of the more complicated updated PCPTRC to the standard PCPTRC in this region. Both the standard and updated PCPTRC provided benefit over the rule referring no patients to biopsy.

Fig. 2 Net benefit curves



5 Food for Thought

To illustrate fundamental principles this chapter has focused on risk models for cancer diagnosis, but similar principles apply for all aspects of cancer treatment and follow-up care, and easily extend to prediction problems in other disciplines outside of medicine.

For projecting risks over expanded time periods, such as the 10-year risk of heart attacks or other cardiovascular events in elderly people, models incorporating risks of death from other causes, referred to in the medical literature as competing risks, need to be implemented. Prognostic models refer to models used to predict treatment outcomes, such as how long a patient can expect to live after a given a treatment is administered. They may rely on other methodologies than logistic regression, such as the Cox proportional hazards models which are appropriate for handling the commonly occurring censored survival times. A censored survival time refers to the case where the exact date of death of a patient is not observed; it is only known that the patient has lived a certain number of years, such as up until the end of the clinical trial. Projections from such models can be used as a basis for making treatment decisions, by favoring treatments that have the longest survival period for specific patient clinical characteristics. The picture is not unidimensional as benefits in survival might be offset by loss in quality of life due to side effects. More complicated decision functions incorporating multiple outcomes are required for weighing the cost-benefits of competing treatment options. Increasingly, models addressing multiple long-term effects of preventative or curative treatments for cancer, such as higher incidences of ovarian and endometrial cancer in women taking tamoxifen, are being implemented in order to provide a unified picture of the pros and cons of various actions, providing many avenues for research in risk prediction for the future of medicine.

The principles of model building and online prediction outlined in this chapter also directly apply to other disciplines, including forestry, ecology, informatics and finance. For forest management, [32] have developed an online tool called SILVA that projects growth of trees over expanded time periods. Their model accounts for man-induced thinning, mortality, and other natural- and human-induced impacts. Using an expanded database of 40000 trees observed at 5-year intervals in Bavaria, Boeck and colleagues [31] implemented logistic regression to update the mortality simulator module of the SILVA program. Similar collaborations with ecologists are working towards prediction of microhabitats on trees representing biological diversity, a concept of interest to forest conservationalists. Informatic scientists are using the techniques to develop online predictions of project margins for large and complex software development portfolios. One can foresee similar applications for online predictions of financial success indicators based on the types of advanced time series models used in that field.

6 Summary

This chapter has detailed the step-by-step program by which risk prediction models are built, using as one illustration construction of the PCPTRC, one of the currently most widely used prostate cancer risk calculators. The importance of external validation across multiple cohorts pushing the envelope in terms of generalizability of the risk tool has been emphasized, as well as the separate components of validation which address discrimination, calibration, and net benefit. As risk prediction tools are typically founded on once-in-a-lifetime large well-designed studies, methodologies are needed for updating them based on new data and risk factor discoveries based on smaller more recent studies. This chapter has discussed the need for comparing existing to updated risk models, using the integrated discrimination index as one possible measure. To end a summary on risk prediction tools in current use was provided along with extensions to other outcomes in medicine and applications in other disciplines such as forestry and finance.

References

Selected Bibliography

1. H. Akaike, A new look at the statistical model identification. IEEE Trans. Autom. Control **19**, 716–723 (1974)
2. D.P. Ankerst, J. Groskopf, J.R. Day et al., Predicting prostate cancer risk through incorporation of prostate cancer gene 3. J. Urol. **180**, 1303–1308 (2008)
3. D.P. Ankerst, T. Koniarski, Y. Liang et al., Updating risk prediction tools: a case study in prostate cancer. Biom. J. **54**, 127–142 (2012)

4. C.B. Begg, R.A. Greenes, Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics **39**, 207–215 (1983)
5. V. Cavadas, L. Osório, F. Sabell, F. Teves, F. Branco, M. Silva-Ramos, Prostate cancer prevention trial and European randomized study of screening for prostate cancer risk calculators: a performance comparison in a contemporary screened cohort. Eur. Urol. **58**, 551–558 (2010)
6. D.R. Cox, Two further applications of a model for binary regression. Biometrika **45**, 562–565 (1958)
7. S.J. Eyre, D.P. Ankerst, J.T. Wei et al., Validation in a multiple urology practice setting of the prostate cancer prevention trial calculator for predicting prostate cancer detection. J. Urol. **182**, 2653–2658 (2009)
8. D.J. Hernandez, M. Han, E.B. Humphreys et al., Predicting the outcome of prostate biopsy: comparison of a novel logistic regression-based model, the prostate cancer risk calculator, and prostate-specific antigen level alone. BJU Int. **103**, 609–614 (2009)
9. K.J.M. Janssen, A.R.T. Donders, F.E. Harrell Jr. et al., Missing covariate data in medical research: to impute is better than to ignore. J. Clin. Epidemiol. **63**, 721–727 (2010)
10. A. Jemal, R. Siegel, J. Xu, E. Ward, Cancer statistics, 2010. CA Cancer J. Clin. **60**, 277–300 (2010)
11. S. Lemeshow, D.W. Hosmer Jr., A review of goodness of fit statistics for use in the development of logistic regression models. Am. J. Epidemiol. **115**, 92–106 (1982)
12. Y. Liang, D.P. Ankerst, M. Sanchez, R.J. Leach, I.M. Thompson, Body mass index adjusted prostate-specific antigen and its application for prostate cancer screening. Urology **76**, 1268.e1–1268.e6 (2010)
13. M.E. Mille, S.L. Hui, W.M. Tierney, Validation techniques for logistic regression models. Stat. Med. **10**, 1213–1226 (1991)
14. N.J. Nagelkerke, A note on a general definition of the coefficient of determination. Biometrika **78**, 691–692 (1991)
15. C.T. Nguyen, C. Yu, A. Moussa, M.W. Kattan, J.S. Jones, Performance of prostate cancer prevention trial risk calculator in a contemporary cohort screened for prostate cancer and diagnosed by extended prostate biopsy. J. Urol. **183**, 529–533 (2010)
16. D.J. Parekh, D.P. Ankerst, B.A. Higgins et al., External validation of the prostate cancer prevention trial risk calculator in a screened population. Urology **68**, 1153–1155 (2006)
17. M.J. Pencina, R.B. D'Agostino Sr., R.B. D'Agostino Jr., R.S. Vasan, Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat. Med. **27**, 157–172 (2008)
18. S. Perdonà, V. Cavadas, G.D. Lorenzo et al., Prostate cancer detection in the grey area of prostate-specific antigen below 10 ng/ml: head-to-head comparison of the updated PCPT calculator and Chun's nomogram, two risk estimators incorporating prostate cancer antigen 3. Eur. Urol. **59**, e1–e4 (2011)
19. G. Schwarz, Estimating the dimension of a model. Ann. Stat. **6**, 461–464 (1978)
20. E.W. Steyerberg, *Clinical Prediction Models* (Springer, New York, 2010)
21. E.W. Steyerberg, A.J. Vickers, N.R. Cook et al., Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology **21**, 128–138 (2010)
22. J.A. Swets, R.M. Pickett, *Evaluation of Diagnostic Systems: methods from Signal Detection Theory* (Academic Press, New York, 1982)
23. I.M. Thompson, D.P. Ankerst, C. Chi et al., The operating characteristics of prostate-specific antigen in a population with initial PSA of 3.0 ng/ml or lower. JAMA **294**, 66–70 (2005)
24. I.M. Thompson, D.P. Ankerst, C. Chi et al., Assessing prostate cancer risk: results from the prostate cancer prevention trial. J. Natl. Cancer Inst. **98**, 529–534 (2006)
25. I.M. Thompson, D.P. Ankerst, C. Chi et al., Prediction of prostate cancer for patients receiving finasteride: results from the prostate cancer prevention trial. J. Clin. Oncol. **25**, 3076–3081 (2007)
26. S. van Buuren, Multiple imputation of discrete and continuous data by fully conditional specification. Stat. Methods Med. Res. **16**, 219–242 (2007)

27. A.J. Vickers, E.B. Elkin, Decision curve analysis: a novel method for evaluating prediction models. Med. Decis. Mak. **26**, 565–574 (2006)
28. J.F. Yates, External correspondence: decomposition of the mean probability score. Organ. Behav. Hum. Perform. **30**, 132–156 (1982)

Additional Literature

29. D.P. Ankerst, J. Groskopf, J.R. Day et al., Predicting prostate cancer risk through incorporation of prostate cancer gene 3. J. Urol. **180**, 1303–1308 (2008)
30. D.P. Ankerst, T. Koniarski, Y. Liang et al., Updating risk prediction tools: a case study in prostate cancer. Biom. J. **54**, 127–142 (2012)
31. A. Boeck, J. Dieler, P. Biber, H. Pretzsch, D.P. Ankerst, Predicting tree mortality for European beech in southern Germany using spatially-explicit competition indices. For. Sci. (in press)
32. H. Pretzsch, P. Biber, J. Dursky, The single tree-based stand simulator SILVA: construction, application and evaluation. For. Ecol. Manag. **162**, 3–21 (2002)

Chapter 17
Risk Reduction of Cervical Cancer Through HPV Screening and Vaccination—Assumptions and Reality

Leonore Thümer, Ulrike Protzer, and Vanadin Seifert-Klauss

Cancer is a leading cause of death worldwide. It is estimated that about 20 % of all cancer deaths are originally caused by infectious diseases, making them an important risk factor. The classical approach to lower cancer risk has been the screening for precancerous lesions. In addition, screening for primary infections has become an option in order to evaluate individual cancer risks and to offer an intensive follow-up and intervention for patients who have tested positive. Furthermore, vaccines have been developed to prevent high risk infections and subsequent malignant diseases in the first place. Prospective epidemiological studies are necessary to evaluate the effect of prevention methods on cancer incidences but will give results only after a very long follow-up period. Therefore, mathematical models for risk prediction and risk reduction will be helpful tools to determine the effectiveness of screening and prevention programs. In this chapter we discuss cervical cancer as an example of a malignancy which may in many cases be preventable. During the last decades cervical cancer screening was based on cytological abnormalities. Since human papillomaviruses (HPV) have been identified to be the main risk factor for cervical cancer, the detection of HPV DNA in cells of the cervix has been investigated as a surrogate marker for high cancer risk. Here, we give an overview about the epidemiology and natural course of cervical cancer and HPV infections. We discuss benefits and limitations of current screening and prevention options which include cytology, histology, HPV detection, and HPV vaccination. Finally, we make special emphasis on the complex factors that need to be considered when developing mathematical models for prediction of risk reduction of cancer rates.

L. Thümer (✉)
Virology, Department of Medicine, Technische Universität München, Trogerstr. 30,
81675 Munich, Germany
e-mail: Leonore.Thuemer@virologie.med.tum.de

U. Protzer
Chair of Virology, Department of Medicine, Technische Universität München, Trogerstr. 30,
81675 Munich, Germany

V. Seifert-Klauss
Gynaecology, Department of Medicine, Klinikum Rechts der Isar, Technische Universität
München, Ismaninger Str. 22, 81675 Munich, Germany

C. Klüppelberg et al. (eds.), *Risk – A Multidisciplinary Introduction*,
DOI 10.1007/978-3-319-04486-6_17,

Keywords Human papillomaviruses · HPV · Cervical cancer · Cancer risk · Prevention

The Facts

- Cervical cancer screening is currently based on cytology, and HPV detection is used for modulating the use of other diagnostic procedures.
- Cytology based screening detects abnormalities in cervical cells and predicts the likelihood of precancerous lesions.
- HPV based screening detects DNA of high risk HPV in cervical cells.
- The main risk factor for cervical cancer is a persistent infection with high risk HPV, especially HPV 16 and 18, which are found in 70 % of cervical cancer cases.
- The main limitation of current screening strategies is the low positive predictive value because transient HPV infections and cytological abnormalities are frequent.
- HPV DNA tests have a higher sensitivity but lower specificity for precancerous lesions than cytology based tests.
- A vaccine against HPV 16 and 18 is available.
- Comprehensive mathematical models will help to evaluate the effects on risk reduction of different prevention strategies.
- High participation rates in prevention programs are essential to their success, as the major social risk factor for cervical cancer is non-participation in screening.

1 Cervical Cancer

Cervical cancer is the second most common cancer in women worldwide, with about 500,000 new cases each year. In Germany, each year over 6000 women are diagnosed with cervical cancer, the 5 year survival rate is around 60 %. Cervical cancer arises in the narrow portion of the uterus that joins to the vagina. Most cancers are squamous cell carcinomas deriving from flattened epithelial cells, the second most common type are adenocarcinomas deriving from glandular epithelial cells.

Since cervical cancer is a slow-developing disease progressing through several precancerous stages that can be detected and treated before they develop to invasive cancer, it is a potentially preventable disease. Screening programs based on detecting cytological and histological abnormalities have reduced incidence rates of invasive cervical cancer successfully by 50–80 % during the past decades. In Germany, the incidence of cervical cancer was reduced from 40/100,000 new cases per year in 1971 to 14/100,000 new cases in 2004 [3]. The incidence of precancerous lesions which put women at risk to develop cervical cancer, however, is 50 to 100 times higher than that of actual cancer development. Currently, German health authorities recommend an annual cervical cancer screening with the so called Papanicolaou test (PAP test) for women starting at age 20. The PAP test is based on a sample of cervical cells taken with a swab to detect cytological abnormalities and to classify cancer risk.

Since it was discovered that a persistent infection with high risk human papillomaviruses (HPV) is the central cause of cervical cancer, HPV DNA detection in cervical swabs has been investigated as a new screening marker. In comparison to PAP tests, HPV DNA based screening reaches a higher sensitivity to detect precancerous lesions. However, HPV tests tend to have a lower specificity because transient HPV infections are relatively frequent in women, and because both HPV infection and cytological abnormalities regress in most cases without progressing to cancer.

In comparison to other European countries, German cervical cancer incidences are still in the upper third range, which has raised the question of improving national screening and prevention strategies. This touches the subject of how many women at risk participate in a screening program (see Sect. 5.2), but has also raised discussions to which extent HPV DNA tests should be included into national screening strategies. HPV tests could be performed in combination with PAP tests or as an alternative for primary screening. Another option of preventing cervical cancer is vaccinating young women against HPV so that the risk of an HPV infection is reduced in the first place. The different strategies, HPV versus PAP screening and HPV vaccination, are investigated in large prospective epidemiological and clinical studies. Since these studies require large cohorts and take many years or even decades to generate data on risk reduction of cancer rates, mathematical models can help to assess the medical and cost-effectiveness of the different strategies.

2 Ideal Screening Test Characteristics

The purpose of a screening program is to stratify cancer risk for each individual patient and to offer intervention for those who are at high risk to develop cancer. The ultimate goal of screening is the reduction of the individual and population risk for cancer and to reduce mortality rates due to cancer.

An ideal screening test is on the one hand expected to identify every patient who will develop cancer and on the other hand to give a negative result to all patients who will stay healthy. However, it seems impossible to develop tests which such high sensitivity and specificity. In reality a sensitive test will always give false positive results leading to anxiety, distress and unnecessary treatment or follow-up at high costs. A highly specific test will give false negative results missing early cancer diagnosis resulting in loss of confidence and potential legal consequences. The following aspects are of central interest to determine whether a screening method is efficient in terms of risk stratification, benefits and costs for each patient and for the health system:

- Sensitivity: In how many sick patients will the test be false negative and the diagnosis will be missed?

- Specificity: How many healthy patients will be tested false positive and thus be over-diagnosed?
- Positive predictive value: How many patients with a positive screening result will develop cancer?
- Negative predictive value: How many patients with a negative screening result will stay healthy?
- Therapeutic consequences:
    - How effective is the resulting health care intervention?
    - How many patients with positive screening do you need to treat in order to save one patient from cancer? At what costs?
    - What are the risks for the patient in case of over-diagnosis and over-treatment? What costs does this raise?
- Cohort: Who takes part in the screening, at which age, at which intervals?
- Adherence: Is screening accepted and performed as being recommended?
- Health education: Is the population well informed about benefits and risks?

3 Screening with Cytology (PAP Test)

The so-called PAP test was developed by G. Papanicolaou and is used to detect cytological abnormalities in cervical cells and to stratify cancer risk. During a gynecological exam, single and clustered cells are wiped from the cervix surface with a cotton swab and from the cervical canal with a little brush onto two separate glass slides. The obtained cells are stained and inspected for abnormalities by experienced cytologists and are classified as PAP I–IV. This cytological classification expresses the likelihood of the presence of histological abnormalities, so-called epithelial dysplasia or cervical intraepithelial neoplasia (CIN grades 1–3) as lined out in Table 1. Misclassification of CIN likelihood by cytology happens in 8–10 % of cases, and can be due to (non-HPV) infections, incorrect timing or incomplete collection of the cells or incorrect handling, conservation or staining of the cytology specimen. The verification of cytological abnormalities depends upon a tissue sample (histology), in which the layering of cells can be visualized. Such tissue samples are either a biopsy or a conization specimen. The first is usually small which raises the problem of sufficiently representative sampling. Conization requires general anaesthesia and prolonged healing and leads to the shortening of the cervix by approximately 1 cm. On the other hand, conization is often not only a diagnostic but also a therapeutic procedure in which the lesion is removed. Table 1 shows the different systems of nomenclature which exist for cytological classification and which predict the likelihood of cervical intraepithelial neoplasia (CIN).

Table 1 Comparison between cytology (Munich and Bethesda nomenclature) and histology (WHO nomenclature) in cervical cancer screening

Cytology			Histology
Munich nomenclature (Central Europe) PAP test		Bethesda system (USA) Squamous intraepithelial lesion	WHO nomenclature Cervical intraepithelial neoplasia
PAP I	Normal		
PAP II	Mild inflammatory, degenerative or metaplastic changes		
PAP III	Undetermined result: severe inflammatory, degenerative changes, suspicious glandular cells; cannot exclude dysplasia, carcinoma in situ or malignancy	ASC-US: Atypical squamous cells of undetermined significance ASC-H: Atypical squamous cells—cannot exclude HSIL	
PAP IIID	Mild to moderate dysplasia	LSIL: Low grade squamous intraepithelial lesion	CIN 1: Low-grade intraepithelial neoplasia (mild dysplasia)
			CIN 2: Moderate intraepithelial neoplasia (moderate dysplasia)
PAP IVa	Severe dysplasia or carcinoma in situ	HSIL: High grade squamous intraepithelial lesion	CIN 3: High-grade intraepithelial neoplasia (severe dysplasia or carcinoma in situ)
PAP IVb	Severe dysplasia or carcinoma in situ, cannot exclude invasive carcinoma		
PAP V	Invasive carcinoma		

3.1 Epithelial Dysplasia—A Precancerous Lesions?

Dysplasia is by definition a *histological* diagnosis, requiring the analysis of a specimen of tissue with layered cells which is obtained by a biopsy or conization. Dysplasia is defined as cell changes within a multi-layered epithelium and is graded according to whether these cell changes occur only in the superficial layers of the epithelium (cervical intraepithelial neoplasia grade 1) or also in the intermediate (CIN 2) or basal layers (CIN 3). The nearer the cell changes are to the basal membrane, the higher the risk of progression to cancer. If the basal membrane is involved and cell changes are found beneath it, the neoplasia is no longer intraepithelial, but invasive. CIN 1 lesions regress spontaneously in up to 80 % of cases over the course of one year, CIN 2 lesions may still regress in 40 %, while CIN 3 lesions carry a likelihood of 90 % of progressing to become invasive cervical cancer, even if the lesions may be very small. This is why only CIN 3, which includes cervical cancer

Table 2 A proposed algorithm of dealing with screening results in Germany (graded after G. Papanicolaou) in the S2k-guidelines by the working group colposcopy and infection in the German society of gynecologists (DGGG). Colposcopy: inspection of the cervix with an enlarging optical device and two diagnostic staining methods to detect abnormalities

Cytology result	hrHPV result	Cytology checkup	Other diagnostic procedures
PAP I/II	Negative	12 months	–
	Positive	12 months	HPV control simultaneously If still hrHPV positive or cytology abnormal: colposcopy
PAP II W (unclear result)	Negative	12 months	HPV control
	Positive	6 months	HPV control simultaneously If still hrHPV positive or cytology abnormal: colposcopy
PAP III		4 weeks	
PAP III D (first time)	Negative	6 months	HPV control
	Positive	3–6 months	HPV control simultaneously If still hrHPV positive: colposcopy and biopsy
PAP III D (repeatedly)	Negative	6 months	HPV control Colposcopy after 12 months
	Positive	–	Colposcopy and biopsy
PAP IVa +	Positive or negative	–	Colposcopy and biopsy

in situ, is considered to be an obligatory precancerous lesion, requiring surgery soon after detection.

4 Screening with HPV Test

The principle of HPV testing is to find nucleic acids of high risk human papillomaviruses (hrHPV) in cervical cells, by means of molecular methods. HPV can be detected in cell material obtained from the cervix with brush swabs (similar to cytology) during a gynecological exam. In most of the tests used, probes can check whether any of a certain panel of HPV types is present. High risk HPV detection is currently used for modulating the use of other diagnostic procedures (see Table 2). It is also being discussed as an alternative for primary screening or for triaging cytology intervals in conjunction with cytology.

4.1 HPV Infection

In the 1970s an infection with human papillomaviruses (HPV) was identified as a necessary but not sufficient condition to develop invasive cervical cancer, and in subsequent epidemiological studies HPV was detected in over 99 % of cases of invasive carcinoma of the cervix [4, 21]. Of the more than 150 HPV types known to date, only 10–15 types have been associated with lesions that can progress to cancer and are therefore classified as high risk HPV (hrHPV). HPV 16 and 18 are the types most commonly found in cervical cancer and are deemed to be the most aggressive. The other hrHPV include types 31, 33, 35, 39, 45, 51, 52, 56, 58 and 59. Infections with multiple (low and high risk) HPV types are possible. A meta-analysis of 243 studies involving 30,848 cases of invasive cervical cancer worldwide from 1990 to 2010 found some type of hrHPV to be present in 90 % of all tumors. HPV16 was found in 57 % of all cases and HPV18 in 16 %, with both types contributing to over 70 % of all invasive cervical cancers. Multiple types were detected in 11 % of all affected women [16].

Human papillomaviruses are transmitted by direct contact with infection foci and can cause genital and skin warts as well as papillomas within the respiratory tract. However, mostly the infection is asymptomatic. HPV infection is one of the most common sexually transmitted diseases: worldwide the overall age-adjusted prevalence of current HPV infection is about 10 % in women, and up to 30 % in young women below 25 years of age [5]. HPV positivity largely depends on the sexual behavior of the woman and her partner, with a risk increase of approximately 4 % per new sexual partner. The cumulative lifetime prevalence, which reflects the number of people infected with HPV at least once in their life, can reach up to 80 % [14].

In most cases HPV are spontaneously cleared due to a gradual development of an effective immune response. In a prospective study following 608 college women at 6 months intervals for three years, it was shown that the cumulative 3-year incidence of HPV infection was 43 % and that the median duration of new infections was 8 months [10]. In general, it is assumed that 60–80 % of all HPV infections are cleared within 12 months, and 90 % within 24 months [18]. In cells infected with HPV, mild cytological abnormalities can be found which frequently regress spontaneously. Only in 10 % of cases a persistent HPV infection develops which can induce a transformation of the cervical epithelium. Dysplasia may develop in 1–10 % of women having a persistent HPV infection over the course of 0.5–5 years and which may progress through several stages to cancer with a latency of several years.

Risk factors to develop cervical cancer clearly relate to a high probability of acquiring a genital infection with HPV, such as a high number of sexual partners or a partner with many differing sexual partners, and early age at first sexual intercourse. Other risk factors which might determine the progression from HPV infection to precancerous lesions reflect an inability to build up a sufficient immune response against HPV, such as immunosuppression due to infection with the human immunodeficiency virus (HIV) or medication, a high number of pregnancies, or smoking. The use of condoms has been associated with a high rate of regression of dysplasia [11].

4.2 Molecular Tests in Clinical Trials

For risk stratification on cervical cancer based on HPV tests it has to be kept in mind, that infection with HPV is frequent especially in women under 30, mostly transient, and that only a persistent infection with high risk human papillomaviruses (hrHPV) represents a risk for cancer. In addition, only a small proportion of patients with a persistent hrHPV infection develop cancer. From 1 million women infected with hrHPV, only 8000 will develop carcinoma in situ, and only 1600 will develop invasive cervical cancer.

There are different molecular methods available for detecting HPV DNA in a sample of cervical cells, which is collected in a similar way as the PAP test. The Hybrid Capture 2 (hc2) assay detects DNA from 13 high risk HPV types, however an exact determination of the HPV type is not possible. This assay proved to be of great clinical value in many important clinical trials [17]. In women with abnormal cytology the hc2 HPV assay showed a higher sensitivity, with similar specificity for CIN 2 or worse (CIN 2+) compared to cytology alone [8]. Primary screening with the hc2 for hrHPV for identifying CIN 2+ lesions proved to be—compared to cytology—more sensitive (around 99 %) but significantly less specific (around 28 %). The low specificity rate is due to the transient nature of many HPV infections and to cross-reactivity of the probe cocktail with non-high risk HPV types, both of which do not cause cellular abnormalities. Next generation assays are focusing on the detection of hrHPV DNA including the genotyping for HPV16 and HPV18, since 70 % of all cervical cancer cases are associated with these two genotypes. Superiority of HPV16 and HPV18 testing in terms of positive predictive rates for CIN 3+ was proven by a long term study following 20,810 women aged 30 and older. 10-year cumulative incidence rates of CIN 3 or invasive cancer were 17.2 % among women positive for HPV16 and 13.6 % among women positive for HPV18. In comparison, only 3 % of women who tested positive for hrHPV other than HPV16/18 developed CIN 3 or invasive cancer after 10 years [13]. Figure 1 illustrates the results of the study performed on HPV genotyping. Based on these data, assays for clinical routine use were developed which give results on positivity for one of 14 hrHPV types, and specific positivity for HPV16 and HPV18. Alternative HPV assays detecting RNA as a marker of transcription of oncogenic viral genes are being investigated, which have a higher specificity (around 70 %) and confer a higher positive predictive value (around 50 %) [20]. Since these tests are expensive and show a lower sensitivity (around 74 %), they can only serve as second-line screening in case of a positive HPV DNA test.

5 Finding the Optimal Strategy for Cervical Cancer Screening

In the case of cervical cancer screening, a high sensitivity and high negative predictive value can be reached because practically most cancers are caused by an HPV infection and can be identified as dysplasia in early stages. However, it is far more

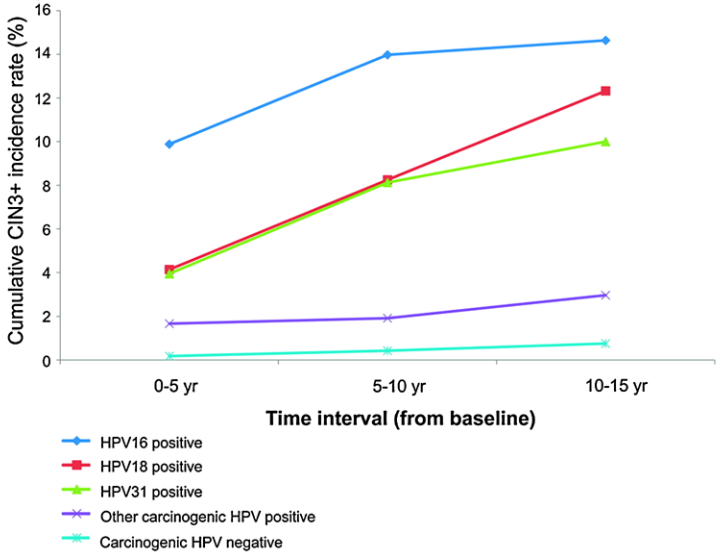


Fig. 1 Cumulative incidence rate of cervical intraepithelial neoplasia grade 3 or invasive cervical cancer (CIN 3+) over 15 years following a single HPV test. The risk estimates show primarily that in this older cohort (20,000 women, average age approximately 35 years), absence of hrHPV (hrHPV negative) predicts very low risk of subsequent CIN 3+. Baseline test positivity for HPV16, HPV18, or HPV31 was most strongly linked to subsequent CIN 3+. Note that the *y*-axis ends at 16 %. From: [19]

difficult to reach high specificity and a high positive predictive value. Currently available screening methods analyze markers that are relatively frequent in women: about 3–4 % of PAP smears are abnormal and 10 % of screened women (up to 30 % below the age of 30) are positive for HPV, with 2.6 % positive for HPV16 [5]. Abnormal cytology, HPV infection and mild precancerous lesions regress spontaneously in most cases and only a low percentage of women with persistent abnormalities and hrHPV infection develop cancer after a long period of time. Women with positive test results on HPV require time-intensive counselling, follow-up and shorter screening intervals, since they carry a risk factor but do not know whether they will develop a disorder at all, nor when this disorder might disappear. On the other hand, thanks to the high negative predictive value of HPV tests, screening intervals might be extended in case of HPV negativity, which can contribute to lowering screening costs.

5.1 PAP Test Versus HPV Test in Clinical Trials

Screening with PAP tests has successfully lowered incidence rates of cervical cancer during the last decades. However, it needs individual interpretation and thus is to some extent subjective. In comparison to cytology, HPV testing provides an objective test outcome that is highly reproducible and test procedures can be easily automated [7]. Most importantly, a recent meta-analysis showed that HPV DNA testing achieves higher sensitivity (relative sensitivity increase 33 %) than PAP tests to detect high-grade CIN and invasive cancer but lower specificity (relative reduction in specificity 6 %) when compared to cytology [8]. Therefore it has been proposed to use HPV testing as an addition or even alternative for cytology based screening.

Study results have led to the recommendation to use high risk HPV screening together with the PAP test for women aged 30 and older as a primary screen in order to identify women who need to be monitored for cytological and histological abnormalities more closely than others. However, the TOMBOLA study assessing the value of a single HPV test in women with mild and borderline cytological abnormalities, proposed to shift this age limit towards 40 [6]. The study, which included 4439 women from the UK aged 20–59, showed that an additional test on hrHPV, increased the positive predictive value for CIN 3+ from 9.7 % to 17.5 %. Specificity rates increased in older ages, leading to the conclusion that in women over 40, a negative HPV test could rule out further investigation. The study also found 22 % of women of all ages with CIN 3+ to be HPV negative. This of course raises concerns regarding the reliability of HPV negativity and the quality of HPV tests being used.

To evaluate whether HPV test or PAP test could better predict who is at low risk and who is at high risk to develop cancer, a large prospective study with a screening interval of 3 years was conducted in 300,000 women aged 30 and older and published in 2011 [12]. In the 16,757 women who tested positive for HPV, the presence of an abnormal cytology result greatly increased the cumulative incidence of CIN 3+ over 5 years from 5.9 % to 12.1 %. By contrast, abnormal cytology did not increase the 5-year risk of CIN 3+ for women negative by HPV testing to a substantial level (0.16 % vs 0.86 %). 73 % of the women positive by HPV testing had no cytological abnormality in PAP tests, but when biopsied still had a high rate of CIN 3+ and accounted for 30 % of the cancers and even 63 % of the adenocarcinomas detected in this study. Based on these results, the authors discussed whether testing for HPV without adjunctive cytology might be sufficiently sensitive for primary screening for cervical cancer.

In 2011, the German institution for evaluation of health system quality and economic concerns (IQWIG) evaluated six large population-based studies using different screening strategies for at least 3 years. The experts were unable to recommend a certain screening strategy, despite documented benefits of HPV test based screening alone or in combination with cytology compared to cytology alone. Benefits were documented for earlier diagnosis of CIN 3 and early cervical cancer, but did not show better survival. One of the reasons for not giving a recommendation by the institute was that a possible disadvantage through this new screening strategy could not be evaluated since no study data existed on primary screening with HPV tests alone (outside the present cytology-based screening system). However, the large study by Katki et al. [12] suggesting benefits of HPV screening has to be taken into account by further re-evaluations.

5.2 General Limitations of Screening

In 2007, 34 million women in Germany would have been eligible for cervical cancer screening, but only 21.8 million PAP smears were performed, 15.8 million in the course of screening. Low participation rates are a major reason for the limitations to any screening: in the United States, 50 % of women with a new diagnosis of cervical cancer have never had cervical cytology screening, another 10 % had not been screened within 5 years before diagnosis [1], and women who are immigrants to the US from countries where cytology screening is not the norm are an especially high-risk group [2]. Very recently, the German government has announced to change the organisation of cervical cancer screening from mere eligibility towards personal invitation mailings to patients.

The feeling of insecurity when facing repeat examinations in the absence of therapeutic measures may lead to mistrust in the medical system providing such screening. Women confronted with a positive result of either an HPV test or cytology need to be informed about potential consequences. Since HPV positivity is more common than abnormal cytology findings, the need for information increases. HPV is a risk factor, which *may* or *may not* lead to cytological abnormalities over the course of several years. The following relevant basic information should be provided to women on cytology and HPV testing in case of a positive result, according to German regulatory authorities: Nature and origin of an abnormal cytology, natural course of HPV infection and associated cell changes, HPV types (low risk/high risk), route(s) of infection, prevalence, latency, regression, effects on the sexual partner, management options and their consequences for fertility, and risk of cervical carcinoma. The need and capacities for counselling and management in case of any abnormal test finding must be considered along with resources supplying such counselling when discussing a change of mode of screening, as the large number of positive risk factor results will lead to a massive increase in demand for information.

6 Vaccine Against HPV

Vaccination is traditionally the most cost-effective approach to prevent infections and subsequent diseases. Two vaccines are currently available for primary prevention of HPV infection: one bivalent including types 16 and 18, and one quadrivalent vaccine covering additionally types 6 and 11, which are frequent in genital warts. It is recommended to apply three vaccine doses in girls before their first sexual intercourse. In Germany, health insurances cover the costs for girls between 12 and 17 years of age. Other countries have extended the recommended vaccination age for girls from 9 to 26 years, and even included vaccination of boys into their recommendations in order to achieve herd immunity.

Clinical studies proved the vaccine to be well tolerated and highly immunogenic. Most efficacy studies focused on the protection from HPV-related intraepithelial lesions and persistent HPV infection by the HPV types used in the vaccine. It was shown that the vaccine provided 100 % protection from persistent HPV16 infection over 17 months, and 94 % protection after 3.5 years [9, 15]. Studies with the clinically more relevant endpoints neoplasia and invasive cancer are still ongoing and results will not be available before 2020.

The success of an HPV vaccine clearly depends on the strength of the immune response induced, the HPV types included, cross-protection to untargeted HPV types and selection of the vaccinees (age and gender). Other important aspects are vaccine costs and the ability to deliver vaccines to countries with low health care resources where cancer screening rates are usually low. However, since current vaccine formulations neither protect against all high risk HPV types nor allow treating persistent infection, screening programs must continue, but intervals in vaccinated women might be extended.

7 Mathematical Models on Risk Reduction and Cost-Effectiveness

Since health economic resources are limited, decisions to introduce a new prevention strategy into a national prevention program need to be based on the evaluation of its clinical impact in terms of risk reduction and on cost-effectiveness. Prospective clinical studies are cost-intensive and require a long duration to show results on relevant clinical outcomes. So called health technology assessments (HTA) could represent important tools for public health decision makers to establish priorities on health care choices. HTA use mathematical modeling to determine risk reduction by certain strategies and their costs, taking into account comprehensive clinical, epidemiological and economic data.

In 2010, a health technology assessment was performed by German public health scientists in order to evaluate the long-term effectiveness and cost-effectiveness of HPV DNA testing as a primary screening method for cervical cancer in the context of the German health system [24]. The HTA evaluated 18 different strategies which varied in the combination and intervals of PAP and HPV testing (Table 3). Medical effectiveness was determined by the reduction of lifetime-risk for cervical cancer, reduction of mortality by cervical cancer and gained life-expectancy. Cost-effectiveness was calculated as life-time costs of a certain strategy including costs for screening and therapy, and the incremental cost-effectiveness-ratio (ICER), which represents the costs per gained life-year (LYG) in comparison to a less effective strategy. The goal of the HTA was to identify the optimal strategy for cervical cancer screening and to give recommendations for German health care decision makers.

The HTA used a Markov model with a hypothetical cohort of 15 year old women who moved through different states of HPV infection, cervical precancer and cancer over the course of a lifetime, simulating transition probabilities from one state to another. The report aimed at covering the complex factors involved in risk prediction and risk reduction of cervical cancer. Clinical, epidemiological and economic data from Germany was used as the basis of the model including the natural course and mortality of the disease, incidences of HPV infection, cervical cancer and its precancerous lesions, as well as common practice and costs of subsequent diagnostic and therapeutic options. In a base case analysis, test accuracy data was derived from international meta-analyses: sensitivity of PAP test and HPV test for CIN 3+ was set at 72 % and 98 %, specificity at 95 % and 92 %, reflecting the higher sensitivity but lower specificity of HPV DNA testing. In a scenario analysis, which took into account data on test accuracy from a German screening study, the sensitivity of PAP test was set at only 46 % for CIN 3+. The model was validated by comparing outcomes of model prediction with data from German cancer registries and literature.

Table 3 shows the medical effectiveness of the different strategies in the base case analysis compared to no screening: reduction of cervical cancer risk by screening varied between 53 % and 97 %, reduction of mortality between 61 % and 99 %, and gained life-expectancy between 56 and 91 undiscounted life days per woman. Annual PAP screening as currently being performed in Germany reduced the risk of cervical cancer by 93 %. HPV screening starting at age 30 combined with PAP screening for women aged 20 to 29, both at 2-year intervals, reduced the risk of cervical cancer by a comparable 91 %. In the scenario analysis with the lower PAP sensitivity, risk reduction with annual PAP tests was calculated to be only 78 %, and therefore at a significant lower level than HPV testing strategies at 1, 2, or 3 year intervals starting at age 30.

Cost-effectiveness was evaluated by determining the ICER of the dominating strategies, calculating the ratios of incremental costs and incremental life-expectancy, as represented by the slopes in Fig. 2. Biennial PAP screening between age 20 and 29 combined with biennial HPV screening starting at age 30, being equally effective as annual PAP screening, resulted in an ICER of 28,400 Eur/LYG in the base case analysis. Screening at 1 year intervals both for PAP and HPV testing would reduce risk of cervical cancer by only an additional 6 % at an ICER of 155,500 Eur/LYG. Acceptance of ICER depends on each society, no limits are applied in Germany. However, the WHO recommends ICERs not to exceed 3 times the BIP per person, for Germany this would correspond to 90,000 Eur/LYG. In the scenario analysis, increasing the interval of HPV testing to 3 years would reduce costs significantly with a reduction of cancer risk by 83 %, and therefore still at a higher level than annual PAP tests. Furthermore, an additional analysis revealed that starting screening at a later age allows health care resources to be saved without a relevant loss of effectiveness. The authors of the HTA therefore concluded that the optimal strategy for cervical screening in Germany could be performing biennial PAP tests between ages 25 and 29 followed by biennial HPV tests age 30 and above. Main limitations of the HTA were not considering life expectancy adjusted to quality of life, as screening results and precancer treatment might cause psychological distress and adverse events, which might also affect cost-effectiveness-ratios. Furthermore, adherence rates to screening were only estimated and not based on detailed data, neither were the effects of different HPV types considered in the model. This raises the need for further studies, as higher participation rates or reduced rates

Table 3 Results of a German HTA using a Markov model. Medical effectiveness of 17 strategies on cervical cancer screening compared to no screening. Before: age 20–29 [24]

Strategy	Risk reduction Cervical cancer vs. no screening (%)	Risk reduction Mortality vs. no screening (%)	Gained life expectancy vs. no screening (days per woman)
1-year-interval			
HPV, 1y, age 30; before PAP, 1y	97.4	98.7	91
PAP, 1y, age 20	92.7	96.1	88.7
2-year-interval			
HPV+PAP triage, 2y, age 30; before PAP, 2y	91.7	95.0	87.7
HPV+PAP, 2y, age 30; before PAP, 2y	91.6	94.9	87.6
HPV, 2y, age 30; before PAP, 1y	91.4	94.8	87.6
HPV, 2y age 30; before PAP, 2y	91.2	94.6	87.4
PAP, 2y, age 20	80.5	86.8	80.0
3-year-interval			
HPV+PAP triage, 3y, age 30; before PAP, 2y	84.8	89.6	82.7
HPV+PAP, 3y, age 30; before PAP, 2y	84.7	89.5	82.5
HPV, 3y, age 30; before PAP, 1y	84.7	89.4	82.7
HPV, 3y, age 30; before PAP, 2y	84.1	89.0	82.1
PAP; 3y, age 20	69.8	77.2	71.0
5-year-interval			
HPV+PAP triage, 5y, age 30; before PAP, 2y	72.2	78.2	72.7
HPV+PAP, 5y, age 30; before PAP, 2y	72.1	78.0	72.5
HPV, 5y, age 30; before PAP, 1y	72.0	77.9	72.8
HPV, 5y, age 30; before PAP, 2y	71.3	77.3	71.9
PAP, 5y, age 20	53.3	60.7	55.9

of HPV infection (as expected by HPV vaccination) might allow to extend screening intervals. Models taking into account quality-of-life data, effects of HPV vaccination and the heterogeneity of different HPV types might better evaluate long-term and cost effectiveness and might improve decision-analytic modeling.

Another HTA was performed in Italy in 2007, in order to assess the clinical and economic impact of the bivalent HPV vaccine in comparison to screening only [23]. Additionally to data on the epidemiology, screening and treatment of HPV infection and its related diseases, the authors took into account efficacy and costs of the HPV vaccine as well as women's knowledge and attitudes toward screening and vaccination. With the help of a systematic review and meta-analysis of efficacy studies on the bivalent HPV vaccine, the prevention rate of a persistent HPV infection was estimated to be 87 % for HPV16 and 78 % for HPV18. Mathematical modeling showed

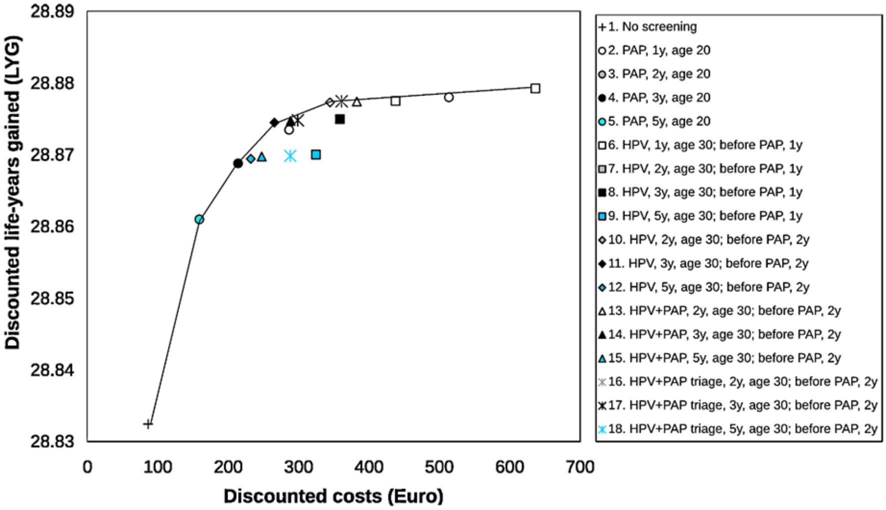


Fig. 2 Results from a German HTA using a Markov model. Cost-effectiveness is based on incremental life-expectancy versus incremental costs. Before: age 20–29 [24]

that vaccinating against HPV plus screening was the best strategy for risk reduction and could reduce the incidence and mortality of cervical cancer by 67 % in comparison to screening only, with an ICER of 22,000 Euros per quality adjusted life year. Analysis of questionnaires revealed a broad interest of women toward vaccination. On the other hand, adherence to screening was found to vary greatly between regions, with only 14 % to 70 % of women repeating the PAP test every 3 years as being recommended by Italian health authorities.

The results of these reports reflect the great value that HTAs can provide for evaluating clinical and economic impacts of new prevention strategies in national health programs. However, in order to allocate health care resources efficiently and successfully, it is essential to strengthen organisational and social involvements like educational campaigns on health care choices throughout the population.

8 Food for Thought

On the one hand, carcinoma of the cervix is the second most common cancer of women in the world, and a persistent infection with high risk HPV types is a major risk factor. On the other hand, this risk factor may or may not lead to cytological abnormalities over the course of several years. Most HPV infections are cleared within 12 months and 90 % within 24 months. Other infections, such as HIV-infection increase the likelihood of disease (a big problem in Africa).

Can assumptions based on African data be transferred to Germany (a country with much lower incidence of cervical cancer)?

Table 3 shows a hypothetical cohort of women under 20 years of age. The different stages of infection, precancerous and cancer disease are modeled based on assumptions from clinical epidemiologic and economic data in Germany. But test accuracy data were derived from international meta-analyses.

How can mathematical models be constructed for an entity so difficult to predict?

The most widely used tests check whether any of a certain panel of HPV types is present. They cannot check whether more than one HPV type is present. Since more refined HPV tests recently allow the identification of more than one HPV type, knowledge can be expected in the future on how combinations of various HPV types act together.

Possibly certain combinations are particularly dangerous?

Apart from specifics applying to the virus type, it is likely that factors affecting transmission of mucosal human papillomavirus play a role.

How can immune-biological factors of the host (patient) be incorporated in models?

9 Summary

Cervical cancer is a slow-developing disease progressing through several precancerous stages. Screening programs based on detecting cytological and histological abnormalities have successfully lowered cancer incidence rates in the last decades. Since it was discovered that a persistent infection with high risk human papillomaviruses is the major risk for cervical cancer, HPV detection in cervical swabs has been investigated as a new screening marker. High sensitivity and high negative predictive values can be reached with currently available screening methods, but specificity and positive predictive values of cancer risk tend to be low because both HPV infection and cytological abnormalities occur frequently and regress in most cases. Discussions are still ongoing whether HPV screening should be used additionally to or instead of cytology based screening. Since large prospective epidemiological and clinical studies are expensive and take many years or even decades to generate data on how prevention programs influence cancer rates, mathematical models will be helpful tools to determine cancer risks as well as (cost-)effectiveness of prevention programs. Risk prediction and estimations of risk reduction must be based on a comprehensive analysis, considering the various risk factors and probabilities of transition to precancerous lesions and invasive cancer, as well as the diagnostic and therapeutic consequences and their costs. Assumptions must also take into account the variable rates of participation in prevention programs, which largely depend on the public knowledge about benefits and limitations of screening and vaccination. A multidisciplinary team including medical doctors, health economists, mathematicians, public health experts and sociologists is needed to determine the risks and evaluate the effectiveness of prevention strategies comprehensively.

References

Selected Bibliography

1. ACOG Practice Bulletin No. 109: Cervical cytology screening. Obstet. Gynecol. 114, 1409–1420 (2009)
2. ACOG Committee Opinion No. 425: Health care for undocumented immigrants. Obstet. Gynecol. 113, 251–254 (2009)
3. AWMF, Diagnostik und Therapie des Zervixkarzinoms. Interdisziplinäre Leitlinie der Deutschen Krebsgesellschaft e.V. und der Deutschen Gesellschaft für Gynäkologie und Geburtshilfe. AWMF online (2008)
4. F.X. Bosch, M.M. Manos, N. Muñoz, M. Sherman, A.M. Jansen, J. Peto, M.H. Schiffman, V. Moreno, R. Kurman, K.V. Shah, Prevalence of human papillomavirus in cervical cancer: a worldwide perspective. international biological study on cervical cancer (IBSCC) study group. J. Natl. Cancer Inst. **87**, 796–802 (1995)
5. F.X. Bosch, A.N. Burchell, M. Schiffman, A.R. Giuliano, S. de Sanjose, L. Bruni, G. Tortolero-Luna, S.K. Kjaer, N. Muñoz, Epidemiology and natural history of human papillomavirus infections and type-specific implications in cervical neoplasia. Vaccine, Suppl. **26**(10), K1–K16 (2008)
6. S. Cotton, L. Sharp, J. Little, M. Cruickshank, R. Seth, L. Smart, I. Duncan, K. Harrild, K. Neal, N. Waugh, The role of human papillomavirus testing in the management of women with low-grade abnormalities: multicentre randomised controlled trial. BJOG **117**, 645–659 (2010)
7. T. Cox, J. Cuzick, HPV DNA testing in cervical cancer screening: from evidence to policies. Gynecol. Oncol. **103**, 8–11 (2006)
8. J. Cuzick, M. Arbyn, R. Sankaranarayanan, V. Tsu, G. Ronco, M.-H. Mayrand, J. Dillner, C.J.L.M. Meijer, Overview of human papillomavirus-based and other novel options for cervical cancer screening in developed and developing countries. Vaccine, Suppl. **26**(10), K29–K41 (2008)
9. D.M. Harper, E.L. Franco, C.M. Wheeler, A.-B. Moscicki, B. Romanowski, C.M. Roteli-Martins, D. Jenkins, A. Schuind, S.A. Costa Clemens, G. Dubin, Sustained efficacy up to 4.5 years of a bivalent L1 virus-like particle vaccine against human papillomavirus types 16 and 18: follow-up from a randomised control trial. Lancet **367**, 1247–1255 (2006)
10. G.Y. Ho, R. Bierman, L. Beardsley, C.J. Chang, R.D. Burk, Natural history of cervicovaginal papillomavirus infection in young women. N. Engl. J. Med. **338**, 423–428 (1998)
11. C.J.A. Hogewoning, M.C.G. Bleeker, A.J.C. van den Brule, F.J. Voorhorst, P.J.F. Snijders, J. Berkhof, P.J. Westenend, C.J.L.M. Meijer, Condom use promotes regression of cervical intraepithelial neoplasia and clearance of human papillomavirus: a randomized clinical trial. Int. J. Cancer **107**, 811–816 (2003)
12. H. Katki, W.K. Kinney, B. Fetterman, T. Lorey, N.E. Poitras, L. Cheung, F. Demuth, M. Schiffman, S. Wacholder, P.E. Castle, Cervical cancer risk for women undergoing concurrent testing for human papillomavirus and cervical cytology: a population-based study in routine clinical practice. Lancet Oncol. **12**, 663–672 (2011)
13. M.J. Khan, P.E. Castle, A.T. Lorincz, S. Wacholder, M. Sherman, D.R. Scott, B.B. Rush, A.G. Glass, M. Schiffman, The elevated 10-year risk of cervical precancer and cancer in women with human papillomavirus (HPV) type 16 or 18 and the possible utility of type-specific HPV testing in clinical practice. J. Natl. Cancer Inst. **97**, 1072–1079 (2005)
14. L. Koutsky, Epidemiology of genital human papillomavirus infection. Am. J. Med. **102**, 3–8 (1997)
15. L.A. Koutsky, K.A. Ault, C.M. Wheeler, D.R. Brown, E. Barr, F.B. Alvarez, L.M. Chiacchierini, K.U. Jansen, A controlled trial of a human papillomavirus type 16 vaccine. N. Engl. J. Med. **347**, 1645–1651 (2002)

16. N. Li, S. Franceschi, R. Howell-Jones, P.J.F. Snijders, G.M. Clifford, Human papillomavirus type distribution in 30,848 invasive cervical cancers worldwide: variation by geographical region, histological type and year of publication. Int. J. Cancer **128**, 927–935 (2011)
17. M. Poljak, B.J. Kocjan, Commercially available assays for multiplex detection of alpha human papillomaviruses. Expert Rev. Anticancer Ther. **8**, 1139–1162 (2010)
18. M. Schiffman, P.E. Castle, J. Jeronimo, A.C. Rodriguez, S. Wacholder, Human papillomavirus and cervical cancer. Lancet **370**, 890–907 (2007)
19. M. Schiffman, N. Wentzensen, S. Wacholder, W. Kinney, J.C. Gage, P.E. Castle, Human papillomavirus testing in the prevention of cervical cancer. J. Natl. Cancer Inst. **103**, 368–383 (2011)
20. A. Szarewski, L. Ambroisine, L. Cadman, J. Austin, L. Ho, G. Terry, S. Liddle, R. Dina, J. McCarthy, H. Buckley, C. Bergeron, P. Soutter, D. Lyons, J. Cuzick, Comparison of predictors for high-grade cervical intraepithelial neoplasia in women with abnormal smears. Cancer Epidemiol. Biomark. Prev. **17**, 3033–3042 (2008)
21. J.M. Walboomers, M.V. Jacobs, M.M. Manos, F.X. Bosch, J.A. Kummer, K.V. Shah, P.J. Snijders, J. Peto, C.J. Meijer, N. Muñoz, Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. J. Pathol. **189**, 12–19 (1999)

Additional Literature

22. Z. Hel, E. Stringer, J. Mestecky, Sex steroid hormones, hormonal contraception and the immunobiology of human immunodeficiency virus-1 infection. Endocr. Rev. **31**(1), 79–97 (2010)
23. G. La Torre, C. de Waure, G. Chiaradia, A. Mannocci, S. Capri, W. Ricciardi, The health technology assessment of bivalent HPV vaccine cervarix in Italy. Vaccine **28**, 3379–3384 (2010)
24. G. Scroczynski, P. Schnell-Inderst, N. Mühlberger, K. Lang, P. Aidelsburger, J. Wasem, T. Mittendorf, J. Engel, P. Hillemanns, K.-U. Petry, A. Krämer, U. Siebert, *Entscheidungsanalytische Modellierung zur Evaluation der Langzeit-Effektivität und Kosten-Effektivität des Einsatzes der HPV-DNA-Diagnostik Im Rahmen der Zervixkarzinomfrüherkennung in Deutschland*. Schriftenreihe Health Technology Assessment (HTA) in der Bundesrepublik Deutschland, vol. 98 (2010)
25. N.J. Veldhuijzen, P.J. Snijders, P. Reiss, C.J. Meijer, J.H. van de Wijgert, Factors affecting transmission of mucosal human papillomavirus. Lancet Infect. Dis. **10**(12), 862–874 (2010)