Texts in Applied Mathematics 70

Ping Yan
Gerardo Chowell

Quantitative Methods for Investigating Infectious Disease Outbreaks

EXTRAS ONLINE

Springer

Texts in Applied Mathematics

Volume 70

The mathematization of all sciences, the fading of traditional scientific boundaries, the impact of computer technology, the growing importance of computer modeling and the necessity of scientific planning all create the need both in education and research for books that are introductory to and abreast of these developments. The aim of this series is to provide such textbooks in applied mathematics for the student scientist. Books should be well illustrated and have clear exposition and sound pedagogy. Large number of examples and exercises at varying levels are recommended. TAM publishes textbooks suitable for advanced undergraduate and beginning graduate courses, and complements the Applied Mathematical Sciences (AMS) series, which focuses on advanced textbooks and research-level monographs.

More information about this series at http://www.springer.com/series/1214

Ping Yan • Gerardo Chowell

Quantitative Methods for Investigating Infectious Disease Outbreaks

Springer

Ping Yan
Infectious Diseases Prevention
and Control Branch
Public Health Agency of Canada
Ottawa, ON, Canada

Department of Statistics
and Actuarial Science
Faculty of Mathematics
University of Waterloo
Waterloo, ON, Canada

Gerardo Chowell
School of Public Health
Georgia State University
Atlanta, GA, USA

ISSN 0939-2475 ISSN 2196-9949 (electronic)
Texts in Applied Mathematics
ISBN 978-3-030-21922-2 ISBN 978-3-030-21923-9 (eBook)
https://doi.org/10.1007/978-3-030-21923-9

Mathematics Subject Classification: Primary: 92D30; 92C60; 60J28; 60K20; 60K37; 62P10. Secondary: 97M6; 37N2

This Springer imprint is published by the registered company Springer Nature Switzerland AG.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

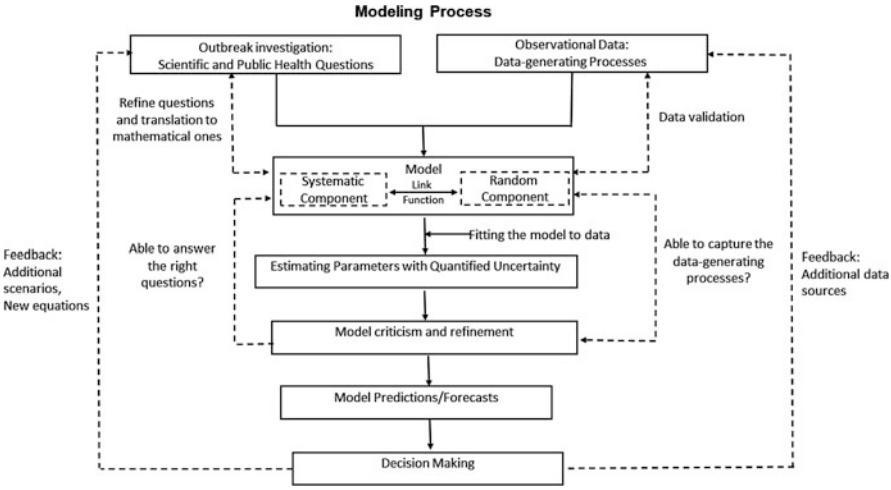
To Louise, Genevieve, and Veronique
To Pia and Catalina

Preface

Mathematical and statistical models and methods can play a central role in outbreak investigations and in public health decision-making. The purpose of this book is to provide readers with balanced perspectives between theory and practice. To provide insight between models driven by scientific hypotheses intended to characterize the agent-host-environment interface in complex disease transmission dynamics, and models driven by observational data intended to capture the data-generating process; and between the unobservable variables predicted by most disease transmission dynamic models and data collected based on observed outcomes. As for prerequisites, before embarking into Chaps. 2–4 of this book, the readers will need an essential understanding of random variables, distribution theory, and stochastic processes (see, for instance, the textbook by Ross (2019)).

The modeling process in this book is illustrated in the following flowchart. Unlike most other scientific investigations, in which questions are formulated and data arise from experiments to address those questions, data arising from outbreak investigations are mostly observational and collected by different agencies for a variety of purposes. In this book, we put equal emphasis on answering the right questions and understanding the data-generating processes.

We started our collaboration in 2003 when we met at a modeling workshop focused on social responses to bioterrorism involving infectious agents, organized by the Center for Discrete Mathematics and Theoretical Computer Science (DIMACS) at Rutgers University. In subsequent years, we actively participated in and co-organized some of the workshops and summer schools on disease modeling supported by the Mathematics of Information Technology and Complex Systems (Canada); the Pacific Institute for the Mathematical Sciences; the Fields Institute; the Banff International Research Station for Mathematical Innovation and Discovery; the Simon A. Levin Mathematical, Computational, and Modeling Science Center at Arizona State University; the Centro Internacional de Ciencias, Cuernavaca, Mexico; and the Centre for Disease Modelling at York University, Georgia State University, University of British Columbia, and University of Alberta, among others.



Much of our research on integrating mathematical and statistical models in infectious disease outbreak investigations has been motivated by discussions among applied mathematicians and statistical scientists during these workshops. This book contains materials from our own presentations and lecture notes, our ideas and views based on personal communications throughout these years, and, more importantly, inspirations from questions during the workshops and summer schools from world-renowned scientists as well as young researchers and graduate students.

We would like to thank the following mentors and long-term collaborators: Carlos Castillo-Chavez, Jerald F. Lawless, Fred Brauer, Mac Hyman, Nicholas Hengartner, Paul W. Fenimore, Hiroshi Nishiura, Cecile Viboud, Lone Simonsen, Mark Miller, Selcuk Candan, Charles Perrings, Alexandra Smirnova, Jianhong Wu; colleagues at the Public Health Agency of Canada, Donald Sutherland, Chris Archibald, Dena Schanzer, Fan Zhang, and Pascal Michel; colleagues at the Georgia State University School of Public Health, Michael Eriksen and Richard Rothenberg; and colleagues at the University of Waterloo, Mary Thompson and Charmaine Dean. We are thankful to Donna Chernyk, our Springer Editor, for providing detailed guidance throughout the publication process.

Special thanks go to graduate students Kimberlyn Roosa, Amna Tariq, and Yiseul Lee in the Department of Population Health Sciences, Georgia State University School of Public Health, for their help in proofreading and editing.

Finally, we would like to thank our families for their support and understanding.

Ottawa, ON, Canada — Ping Yan
Atlanta, GA, USA — Gerardo Chowell

Contents

Chapter 1
Introduction

Infectious diseases ranging from respiratory (influenza, common cold, tuberculosis, the respiratory syncytial virus), vector-borne (plague, malaria, dengue, chikungunya, and Zika) to sexually transmitted (the human immunodeficiency virus, syphilis) have historically affected the human population in profound ways. For example, the Great Plague, well known as the Black Death, was caused by the bacterium *Yersinia pestis* and killed up to 200 million people in Eurasia and about 30–60% of Europe's population during a 5-year span in the fourteenth century. At the time, the plague infection was thought to be due to some "bad air", but it was not discovered that bites of infected fleas were behind the pandemic until late 1890s. If the human civilization had known about the transmission mechanisms behind the plague infections, the epidemic's impact on morbidity and mortality could have been mitigated through basic public health interventions. This is to say that knowledge of the transmission processes and the natural history of infectious diseases in different environments represents invaluable actionable information for thwarting the spread of infectious diseases.

Fortunately, over the years human civilization has made great strides in increasing our understanding of the transmission dynamics of emerging and re-emerging infectious diseases. For instance, John Snow, known as the father of modern epidemiology, mapped the location of cholera cases during the 1854 epidemic in Soho, London, and made the link between the spatial distribution of cholera cases and a pump that he hypothesized as the source of the disease (Fig. 1.1). Following his observations, the pump was removed to avoid further exposures, and the number of cases subsided.

One remarkable and definite shift to the germ theory occurred during the "golden bacteriology" era during the second half of the nineteenth century. In fact, the 1889–1990 influenza pandemic is arguably the first influenza pandemic that occurred in a new and progressive state of knowledge about infectious disease transmission. This pandemic is better known as the "Russian Flu" because the rapid global spread of the pandemic virus can be traced back to Saint Petersburg, Russia in October

P. Yan, G. Chowell, *Quantitative Methods for Investigating Infectious Disease Outbreaks*, Texts in Applied Mathematics 70,
https://doi.org/10.1007/978-3-030-21923-9_1

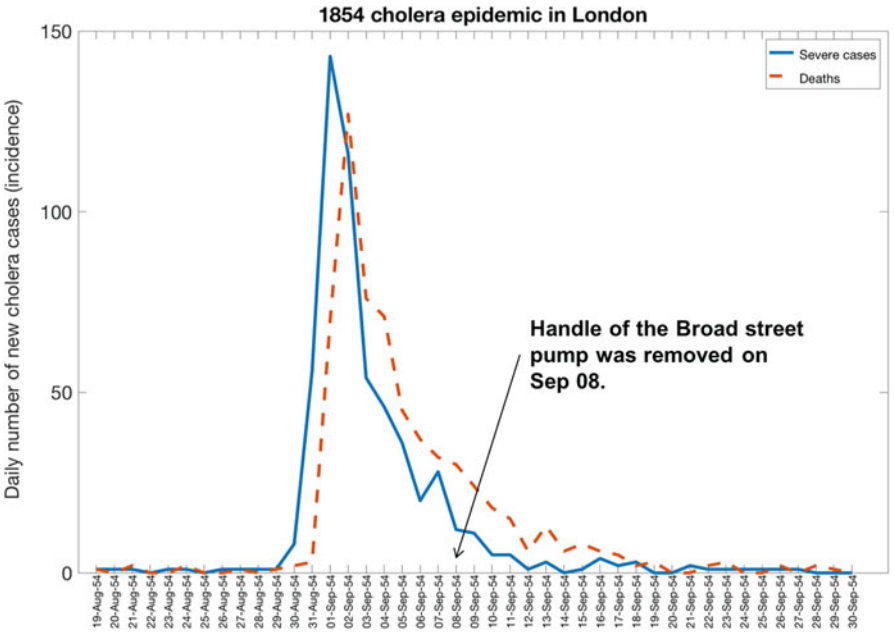


Fig. 1.1 The number of new cholera cases during the 1854 epidemic in Soho, London

1889 (Valleron et al. 2010). Moreover, it was the first pandemic to unfold in a world connected by rail and maritime transportation; it spread across Europe in approximately 6 weeks, with an estimated mean speed at 394 km/week (Valleron et al. 2010) and circulated around the world in just 4 months (Valleron et al. 2010).

Following the 1889–1990 influenza pandemic, in 1918 a novel influenza virus struck the world and killed 20–100 million people, a figure that easily exceeds the death toll associated with World War I (Johnson and Mueller 2002; Dahal et al. 2017; Mills et al. 2004). In the USA alone, about 675,000 people succumbed to the 1918 pandemic virus (Fig. 1.2). However, it was not discovered until years later that an influenza virus was responsible for this pandemic. One hundred years after the 1918 pandemic, we not only remember this devastating historic health disaster, it also serves as a stark reminder of the public health impact that the influenza virus continues to exert on the global population. The 1918 "Spanish Flu" pandemic represents one of the most important case studies for pandemic preparedness available today. However, locating death records to reconstruct the mortality impact of this pandemic requires the arduous task of searching for these documents in old cemeteries, public archives, parishes, and church records (Alonso et al. 2018; Simonsen et al. 2018).

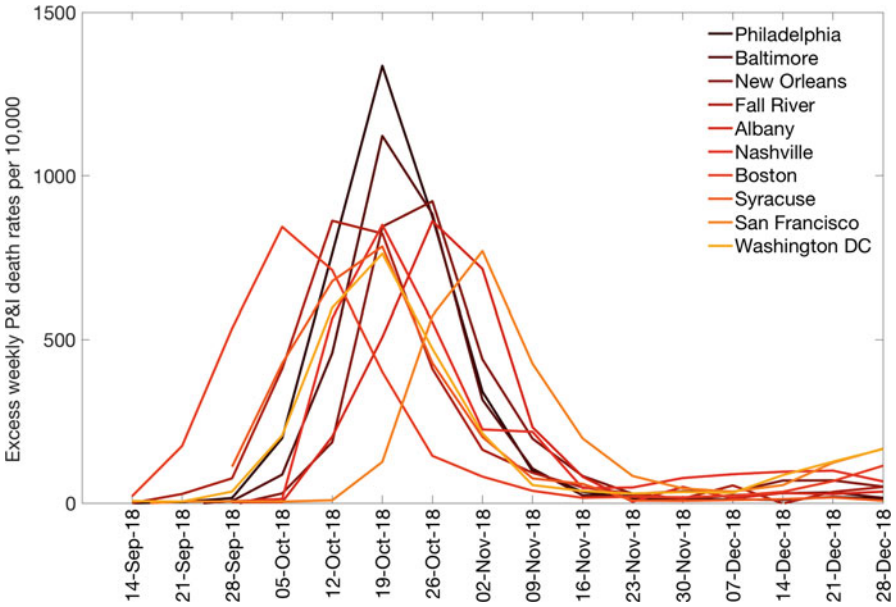


Fig. 1.2 Excess death rate associated with the 1918 influenza pandemic in US cities that exhibited the highest peak excess death rates

The application of mathematical and statistical tools to investigate and forecast evolving epidemics and pandemics has increased significantly during the last couple of decades from ∼50 to >800 publications per year (Fig. 1.3). The worldwide epidemic of acquired immunodeficiency syndrome (AIDS), caused by the human immunodeficiency virus (HIV), started in the early 1980s and accelerated the applications and developments of mathematical and statistical models. This contributed to the understanding of factors that promote transmission of HIV and of strategies for preventing transmission. While the number of studies that apply mathematical modeling to study infectious disease dynamics has rapidly increased over the last two decades (Fig. 1.3), the great majority of those studies are still associated with HIV/AIDS, although this trend has declined somewhat during the last decade, followed by tuberculosis. In addition, the number of studies associated with emerging infectious diseases such as Ebola, dengue, chikungunya, and Zika has been increasing during the last 5–10 years as a result of recent regional and global epidemics (Fig. 1.3).

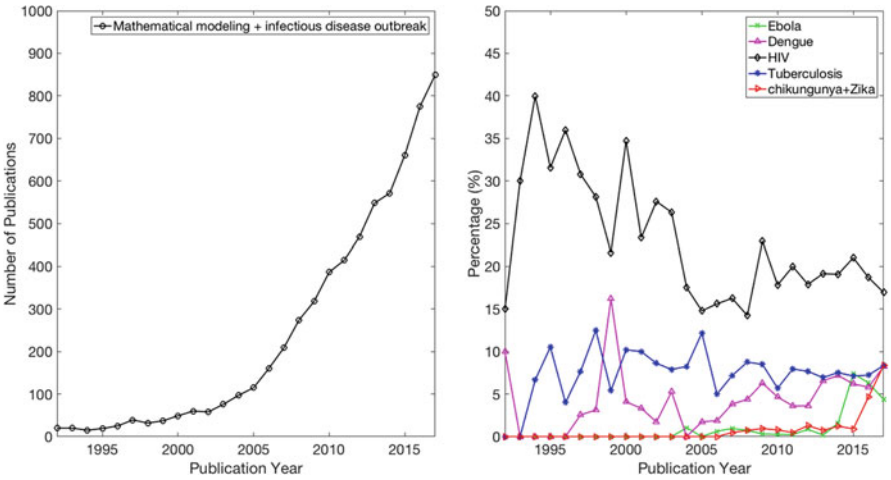


Fig. 1.3 Number of publications on mathematical modeling and infectious disease (left panel) and the fraction of those publications related to different infectious diseases (right panel) by publication year

1.1 The Motivation

Mathematical modeling plays an important role in ordering our thoughts and sharpening vague intuitive notions. Initial models are verbal descriptions that tend to become insufficient as soon as the scenarios become complicated. Mathematics provides a powerful language that forces us to be logically consistent and explicit about assumptions.

Over the years, we have encountered very interesting, inspiring, and challenging discussions at the end of workshops on infectious disease modeling with the following recurrent themes:

1. While most disease transmission models predict an expected exponential growth at the beginning of the epidemic, empirical data often exhibit sub-exponential growth patterns (Viboud et al. 2016). How do we best characterize these non-unique sub-exponential growth functions in the context of infectious disease modeling?
2. Are there many, even infinitely many, mechanisms that lead to the same or very similar sub-exponential growth functions?
3. Does a slower than expected initial growth at the beginning of the epidemic imply a smaller value of the basic reproduction number R_0, a key quantity in the field of infectious disease epidemiology (Anderson and May 1991; Diekmann and Heesterbeek 2000; Brauer 2006), as suggested by many transmission models?
4. What exactly does it mean when we say "deterministic models approximate their stochastic counterparts by the law-of-large numbers"? Are we referring to a population that is infinitely large or something else?

5. Which features of the population-based models, in which the exponential distribution is assumed at the individual levels, can be generalized with non-exponential distributions?
6. Regarding effectiveness of control measures against the spread of diseases, even if imperfect implementation in terms of coverage or compliance has been explicitly taken into account in the models, empirical observations often leave us with impressions that the control measure that "looks good" in theory "do not work at all" in practice. Are there more theories that could capture this phenomenon?
7. How do we reconcile the quantities as predicted by disease transmission models with observed data from outbreak investigations and public health surveillance?
8. The need for precise definitions of verbal descriptions in quantitative analyses. For instance,
    - What do we mean by "a case" when data from outbreak investigations and surveillance are presented as time-series of "number of cases"?
    - Are "generation intervals" consistently defined across literature in epidemiology and infectious disease models?
    - How do we characterize and compare "variability" among random variables, such as the infectious periods or the numbers of secondary infections transmitted by a primary infector?
9. What do we mean by "non-identifiability" when fitting models to data?

Of models formulated in mathematical languages, there are different types that are designed for different purposes.

Broadly speaking, there are mathematical models aimed at facilitating our understanding of the medical, biological, ecological, and social interactions that manifest the outbreaks and epidemics of infectious diseases in order to gain insight into specific questions or to generate theories about what must or might happen; and there are statistical models aimed at capturing the data generation process, for detecting general patterns, predicting epidemic trajectories, managing control strategies, or simply describing epidemic trends. Within both mathematical and statistical models, there are models designed at the population level in a phenomenological way versus models that are individual-based with which researchers aim to capture relevant mechanistic processes.

Individual-based models start from descriptions or assumptions about the evolution of the infectiousness and the natural history of the disease progression within an infected host. These include models for the latent periods, the infectious periods, the incubation periods, recovery, mortality, and so on. Some of the individual-based models also combine social contacts with the evolution of the infectiousness in terms of infectious contacts (Dietz 1995).

Phenomenological models can be deterministic or stochastic and include transmission dynamics models formulated using differential equations or stochastic processes as well as empirical growth functions, such as the generalized logistic growth models. Transmission dynamics models depend on tacit assumptions at the individual level.

The developments of many new statistical models and methods in the study of infectious diseases were driven by the HIV epidemic (Brookmeyer and Gail 1994). Data arising from infectious disease investigations pose unique challenges in classic statistical theory and practice because disease outbreak data do not arise from designed experiments. Each outbreak cannot be repeated naturally under identical conditions, whereas the large amount and multiple sources of clinical data, outbreak investigation data from non-conventional surveys, public health surveillance, and observational data from prevalence and incidence cohorts are collected addressing the same outbreak event. Before statistical methods are used to understand and control the epidemic, statistical models are needed to address the data generation processes, which not only include the epidemiologic and biologic processes that give rise to the disease outbreaks, but also the processes that dictate how data are observed and how a "case" is documented and reported.

When talking about "fitting the model to data," we tend to think of one type of model designed for a specific purpose. However, fitting a dynamic mathematical model to observed outbreak data (e.g., for the purpose of estimating important transmission parameters) involves all three levels of models: the population phenomenological model which depends on tacit assumptions of the individual-based model nested within it, and the statistical model that links the disease transmission process to the data generation process. Very often in practice, these different types of models are considered simultaneously even without the investigators' consciousness.

Driven by the HIV epidemic that started in the late 1970s, the outbreaks of the severe acute respiratory syndrome (SARS) in 2003, pandemic influenza preparedness, and preparedness for other emerging and re-emerging epidemics, the literature on infectious disease modeling has flourished during the past 40 years. However, most articles are confined within subdisciplines according to model characteristics and research focus. While the field of mathematical epidemiology has a long history (e.g., Ross 1911, 1928; Anderson and May 1991; Diekmann and Heesterbeek 2000; Keeling and Rohani 2008; Sattenspiel 2009; Allen 2010; Vynnycky and White 2010; Becker 2015; Andersson and Britton 2012; Manfredi and D'Onofrio 2013; Kermack and McKendrick 1927; Brauer 2006; Brauer and Castillo-Chávez 2001), formal efforts at connecting mathematical models with epidemiological data with the goal of calibrating models for predictive/forecasting purposes have only started to take hold during the last decade (Chretien et al. 2015; Biggerstaff et al. 2016; Chowell 2017; Viboud et al. 2018).

1.2 Structure of the Book with Brief Summary

Chapter 2 provides a review of basic concepts of probability and statistical models for the distributions of continuous lifetime data, closely related to individual-based models that describe the evolution of infectiousness and the natural history of the disease progression. We re-tell the story from a different angle with emphases on the shapes of hazard functions and tail properties of the lifetime distributions

instead of repeating the subject commonly found in a typical textbook on survival analysis. These characteristics have profound impacts on outcomes of the transmission dynamic models at the population level. We will discuss and compare two lifetime random variables, both in terms of magnitude and variability, together with the Laplace transform of lifetime distributions. These concepts will provide the foundations for most of the remaining chapters.

Chapter 3 addresses the distributions of random counts and counting processes, which are closely related to population-based phenomenological models. Section 3.2 provides a framework that links the continuous lifetime distributions at the individual level to the distributions of random counts at the population level. It also provides a historical account. Contemporary discussions on "super-spreading events" as seen in outbreak investigation data in SARS-like diseases are typically associated with transmissions along highly heterogeneous networks characterized by long tailed degree distributions (Lloyd-Smith et al. 2005). Similarly, in the context of incurring accidents, publications in actuarial science journals can be traced back to debates on proneness, contagion, or spells in the first half of the twentieth century that gave rise to important models such as the mixed-Poisson process and the Yule process. Section 3.3 lays the foundation for measuring the evolution of random counts over time, which are key measurements in all population-based models.

Chapter 4 focuses on behaviors of a disease outbreak during the initial phase, immediately after a single (or very few) infected individual are "seeded" into a very large susceptible population. The first part discusses extinction versus growth and relationships among three key parameters: the basic reproduction number R_0, the initial (exponential) growth rate r, and the probability of extinction δ are made and established. With the notion of the "prevalence cohort" (Fig. 4.8), we re-write the classic Lotka equation (4.36) as (4.40) under the assumptions about homogeneous mixing. It reveals that:

1. R_0 only depends on the average value of the infectious periods regardless of the variance or the exact distribution. In models without natural births and deaths in the population, the value of R_0 is not affected by the presence or absence of latent periods.
2. The probability of extinction δ depends on the specific distribution of the infectious periods but is not affected by the presence or absence of latent periods.
3. If the infectious disease does not become extinct during the first few generations, the initial (exponential) growth rate r depends on specific distributions for both the latent periods and the infectious periods.
4. Each of the mathematical relationships between R_0 and δ, and between R_0 and r, as found in the literature, is under a set of strict assumptions on the social contact process and the progression of infectiousness within infected individuals.

Therefore,

1. Given the fixed value $R_0 > 1$ and the infectious periods distribution, the model with latent periods has a smaller initial growth rate r than the one without.

2. Given the fixed value $R_0 > 1$ and the latent periods distribution, the more variable the infectious periods, the smaller the value of r.
3. Without specifying the distributions of the latent periods and the infectious periods, there is no order between the values of r and of R_0.
4. If $R_0 > 1$, without specifying the distribution of the number of secondary infections generated by the primary infectious individual (through the distribution of the infectious periods), there is no order between the values of δ and of R_0.
5. There is a direct relationship between r and δ, rarely mentioned in the literature, that $r = \beta(1 - \delta)$, provided that there is no latent period and that the number of infections produced by a typical infectious individual during a time interval of length x is Poisson distributed with mean value βx. This relationship does not depend on the distribution of the infectious period.

The second part of Chap. 4 emphasizes that the three parameters R_0, δ, and r are intrinsic in the sense that they represent the state of the system at (disease-free) equilibrium when the initially infected individuals are seeded. Section 4.5 presents growth patterns that are most likely to happen when the system moves away from the equilibrium condition. Many discussions are on empirically observed slower growth patterns that largely deviate from the exponential growth assumption (Chowell et al. 2015; Chowell 2017). We attempt to precisely define the sub-exponential growth functions in the context of infectious disease transmission and enlist several assumptions about the transmission dynamics that all lead to such early growth pattern, from the depletion of the susceptible population to scaling of epidemic growth shaped by various factors and their combination including the level of contact clustering and reactive behavior changes (Chowell et al. 2016) and to unobservable individual-level heterogeneity. A special sub-exponential growth function of the form, $(1 + rvt)^{1/v}$, $r, t > 0, 0 < v \leq 1$, is introduced in Chap. 4 which frequently appears in later chapters (6, 8 and 9) in examples and discussions.

Chapters 5 and 6 discuss compartment models when the outbreak moves beyond the initial phase. Much of Chap. 5 is the synthesis of previously published literature on both stochastic and deterministic transmission dynamic models, with our added perspectives. Our interest is to generalize some of the features of these models beyond the assumptions based on the exponential distribution on durations of various stages, and beyond the simple generalizations such as the Erlang distribution (which is a subset of the gamma distribution characterized by smaller variances compared to the exponential distribution with equal mean values). These discussions start in Sect. 5.5.2 and continue in Sect. 6.2.1. In these discussions, Laplace transforms of probability distributions are extensively used as tools to calculate transition probabilities among compartments and average durations within compartments. They are valid for arbitrary distributions without specific assumptions of these distributions. When these distributions are exponential, general results in Sects. 5.5.2 and 6.2.1 return to those published in the literature, such as the expression of the reproduction number as the non-negative eigenvalue of the next generation matrix (van den Driessche and Watmough 2008) as well as in examples in these sections.

We also point out a transcendental relationship among (4.43), (5.66), and (6.24). In these expressions, the Laplace transforms are tools to compare distributions ranked by variability which lead to Propositions 27 and 28 along with discussions in subsequent paragraphs.

Other distinct topics in Chap. 5 are empirical models to describe population-based phenomena without "mechanically" modeling the transmission dynamics at the level of individuals and interactions among individuals. These models are useful for curve fitting, as used in examples later in Chap. 8.

Models in Chap. 6 are more complex and involve intervention measures during the epidemic. Section 6.3 demonstrates a potential application of these models in the context of preparedness for an influenza-like acute respiratory infectious disease with numerical illustrations in hypothetical race-to-treat scenarios and with limited treatment supply. Section 6.5 discusses the impact of unobservable heterogeneity in treatment rates on effectiveness. This section addresses Question 6 in Sect. 1.1. We also draw the attention of the expression $(1 + \phi x v)^{-1/v}$ in (6.31) which echoes the sub-exponential growth function $(1 + rvt)^{1/v}$ in Chap. 4. This is because in both cases, a frailty model from survival analysis is used to model the unobservable heterogeneity among individuals.

Chapter 7 addresses Question 7, 8, and 9 in Sect. 1.1 and serves as a transition between the theoretical topics in previous chapters and Chaps. 8 and 9. The focus is on the data generating processes and statistical issues of fitting models to data. As repeatedly emphasized in Chaps. 4–6, population-based models involve tacit assumptions at the level of individuals, such as the exponential, gamma, or other distributions of the infectious periods. These are conceptual models to address general issues and general patterns, such as the prediction of "incidence" according to time at infection (which is usually unobservable). On the other hand, statistical models address the data generating processes, which include the epidemiology aspects but also the observational schemes, including "case definition," surveillance and reporting, and adjustments for observational biases. In each model, choices are made with respect to which aspects of "the real world" should be included in the description of the model and which should be ignored. These choices not only depend on the perceived importance of various factors, but also on the purpose of each of these models. Frequently, fitting a mathematical model, such as a transmission model, to data collected from surveillance and outbreak investigations involves three types of models (assumptions) that take place at the same time. This requires "nesting" one type of model within another. For example, the statistical model that describes data may involve assumptions of the mean and variance, and in some instances, the assumptions of specific distributions such as Poisson or negative binomial. In addition, the model also handles observational biases such as adjustment for reporting delays (Sect. 7.3). The mean of the statistical model may be a function of time with unknown parameters. This function may involve convolution structures, such as back-calculation (Sect. 7.4), to connect predictions from a conceptual model to expected values of observable outcomes. The conceptual model is thus embedded inside a statistical model. However, this will inevitably result in statistical issues such as non-identifiability (Sect. 7.2). This section mainly

discusses concepts, with a few examples as well as some simple methods where applicable. This is an important field that needs more research and development.

Chapters 8 and 9 focus more heavily on applications, although some models not covered in Chaps. 5 and 6 are presented such as metapopulation spatial models and individual-based network models (Chap. 9). Examples presented are based on a case study for the 2016 Zika epidemic in Antioquia, Colombia (Sect. 8.3), a case study of the 2016 epidemic of yellow fever in two areas of Angola: Luanda (the capital) and Huambo (Sect. 8.4), and a case study of the 2014 Ebola outbreak in Mali (Sect. 9.4).

Chapter 2
Shapes of Hazard Functions and Lifetime Distributions

The main focus of this book is to address phenomenological questions regarding the spread of infectious diseases at the population level. Examples of such questions include:

1. If one or a few infected individuals are "seeded" in a large and completely susceptible population, will it only lead to a handful of infected individuals and the (small) outbreak burns out; or will it lead to an "explosive" (large) outbreak that results in a significant proportion of the population infected?
   (a) If the outcome is the former, what is the expected total number of infected individuals and what is the expected time to extinction?
   (b) If the outcome is the latter, how fast will it grow?
2. In a large outbreak, can we predict the peak burden of the disease and the timing of the peak? How about the long-term outcomes? Will it simply go away after a single wave or a few repeated waves, or will it settle down at some equilibrium state and the epidemic becomes endemic?
3. What about the effects of control measures, such as public health interventions including quarantine, isolation, or pharmaceutical treatments and vaccination?

These phenomenological questions will be addressed by different phenomenological models at the population level (Chap. 4 and onwards). Almost all of these models involve tacit assumptions at the level of individuals. In most diseases transmission models that will be discusses in Chapters 5-6 of this book, the hidden assumptions are: (i) all individuals have equal chances to make random contacts with each other; (ii) a typical infected individual has an infectious period that is exponentially distributed. The simplest model is the SIR (Susceptible-Infected-Recovered) model, associated with an infectious stage and a constant recovery rate. This model has produced many theoretical results along the entire history of an

P. Yan, G. Chowell, *Quantitative Methods for Investigating Infectious Disease Outbreaks*, Texts in Applied Mathematics 70,
https://doi.org/10.1007/978-3-030-21923-9_2

epidemic, from the initial seeding of an infected individual in a very large population until the end of the epidemic when the last infected individual recovers (e.g., Brauer 2008; Allen 2010).

Questions naturally arise, such as what would a non-exponentially distributed infectious period, e.g., increasing or decreasing recovery rate (as a function of time from infection), do to the predicted outcomes? Will some of the predictions be altered and other predictions remain unchanged? If yes, which are they?

Although this chapter covers the same materials as in many classic survival analysis textbooks (e.g., Cox and Oaks 1987; Lawless 2003; etc.), it is organized from the perspective like that in Marshall and Olkin (2007) with more focus on the shapes of hazard functions, the tail properties, and the comparison of variabilities. The main purpose is to closely examine the assumptions at the level of individuals in disease transmission models. This gives the preparation needed in later discussions in Chaps. 4–6.

We call a continuous non-negative random variable denoted by $X \geq 0$ the "lifetime," which is a terminology commonly used in classic textbooks such as Lawless (2003) and Marshall and Olkin (2007).

With respect to infectious disease modelling, the lifetime X is a duration, arising from: (i) the natural history of infectiousness of an infected individual (e.g., latent and infectious periods); (ii) the natural history of clinical manifestation (e.g., incubation period and duration of illness); and (iii) the reaction time of the public health system (e.g., how long it takes to detect an infection or to isolate an infections individual). Assumptions with respect to these durations are used to construct probability models, which in turn give rise to the distributions of random counts (Chap. 3) in the phenomenological models that are related to prevalences and incidences of disease transmission.

2.1 Definitions and the Scale Parameter

2.1.1 The Hazard Function, the Distribution Functions, and Some Commonly Used Summary Measures

For the (absolutely) continuous random variable $X \geq 0$, the hazard function is defined by

$$h_X(x) = \lim_{\delta \to 0} \frac{\Pr\{x < X \leq x + \delta | X > x\}}{\delta}, \tag{2.1}$$

satisfying $h_X(x) \geq 0$ and $\int_0^x h_X(x)dx < \infty$, for some x.

As a duration, X is always associated with an initial event and a subsequent event. The hazard function plays a central role in survival analysis and in industrial reliability that measures the instantaneous probability of the occurrence of the

subsequent event, given that it has not occurred by time t. It has also been called the hazard rate function, or the failure rate function, by different authors in the literature. It is one of the most important quantities underlying every aspect of infectious disease models.

The probability density function (p.d.f.), the cumulative distribution function (c.d.f.), and the survival function are defined by

$$f_X(x) = \lim_{\delta \to 0} \frac{\Pr\{x < X \leq x + \delta\}}{\delta},$$

$$F_X(x) = \Pr\{X \leq x\} = \int_0^x f_X(t)dt,$$

$$\overline{F}_X(x) = \Pr\{X > x\} = \int_x^\infty f_X(t)dt = 1 - F_X(x)$$

respectively. The p.d.f. satisfies (i) $f_X(x) \geq 0$; and (ii) $\int_0^\infty f_X(x)dx = 1$. The c.d.f. $F_X(x)$ satisfies $F_X(0) = 0$, monotonically increasing, and $F_X(\infty) = 1$. The hazard function can be written as $h_X(x) = f_X(x)/\overline{F}_X(x)$.

The following quantities are commonly used summary measures:

1. For $X \geq 0$, the expected (mean) value of X is denoted by $E[X]$ or μ_X. It is also called the first moment of X. It has the following equivalent expressions

$$E[X] = \int_0^\infty x dF_X(x) = \int_0^\infty x f_X(x)dx = \int_0^\infty \overline{F}_X(x)dx. \tag{2.2}$$

2. The qth-quantile of the distribution is defined by $x_q > 0$, such that $F_X(x_q) = q$, $0 < q \leq 1$. When $q = 0.5$, we call $x_{0.5}$ the *median* of the distribution, satisfying $F_X(x_{0.5}) = \overline{F}_X(x_{0.5}) = 0.5$.
3. The variance of X is defined by

$$var[X] = E[(X - \mu_X)^2] = \int_0^\infty (x - \mu_X)^2 dF_X(x).$$

Let $\Phi(x)$ be a Borel function. Then $\Phi(X)$ is a random variable, but not necessarily non-negative. It can be also shown that the first moment of $\Phi(X)$ is

$$E[\Phi(X)] = \int_0^\infty \Phi(x) dF_X(x) = \int_0^\infty \Phi(x) f_X(x)dx. \tag{2.3}$$

The variance is a special case when $\Phi(x) = (x - \mu_X)^2$. Later in this chapter and in later chapters of this book, different classes of $\Phi(x)$ will be used, such as the class of monotone functions, the class of convex or concave functions, etc.

2.1.2 The Scale Parameter

Let $X_0 \geq 0$ be a standard lifetime with hazard function $h_0(x)$ and survival function $\overline{F}_0(x)$ and p.d.f. $f_0(x)$. For the moment, we assume that these functions do not involve any parameters. We define the lifetime $X = \frac{X_0}{\lambda}$, $\lambda > 0$ by transforming the time scale. We call λ the scale parameter. The survival function for X is

$$\overline{F}_X(x;\lambda) = \Pr\{X_0 > \lambda x\} = \overline{F}_0(\lambda x). \tag{2.4}$$

The p.d.f. and the hazard function become

$$f_X(x;\lambda) = \lambda f_0(\lambda x),\ h_X(x;\lambda) = \lambda h_0(\lambda x).$$

Conversely, for any lifetime distribution with survival function $\overline{F}_X(x;\lambda)$ satisfying (2.4), λ is a scale parameter, and one can always re-scale the lifetime $X_0 = \lambda X$ so that the distribution for X_0 has scale parameter 1.

The lifetime distribution may involve multiple parameters, with one scale parameter λ and a vector of additional parameters $\underline{\theta}$. In this case, the general expressions are

$$\overline{F}_X(x;\lambda,\underline{\theta}) = \overline{F}_0(\lambda x;\underline{\theta}),\ f_X(x;\lambda,\underline{\theta}) = \lambda f_0(\lambda x;\underline{\theta}),\ h_X(x;\lambda,\underline{\theta}) = \lambda h_0(\lambda x;\underline{\theta}).$$

Because $h_X(x;\lambda,\underline{\theta}) = \lambda h_0(\lambda x;\underline{\theta})$, the scale parameter λ does not alter the shape of the hazard function characterized by other parameters in $\underline{\theta}$. In other words, by re-scaling both the x-axis and the y-axis, $h_X(x;\lambda)$ and $h_0(x)$ are the same. Therefore, without losing generality, we let $\lambda = 1$ and use the shape of the hazard function $h_0(x;\underline{\theta})$ to develop some commonly used parametric lifetime distributions.

2.2 The Shapes of Hazard Functions

In phenomenological models describing transmission dynamics in populations, implicit assumptions are made on the shapes of hazard functions concerning durations arising from the natural history of infectiousness of an infected individual (e.g., latent and infectious periods) as well as clinical aspects. Very often these hazard functions are called *rates* and are assumed to be constants.

Therefore, this chapter is organized differently from many similar chapters in most survival analysis textbooks which define the lifetime distributions first and then discuss the properties of the hazard functions. We would like to emphasize how different shapes of the hazard functions define and characterize important properties of lifetime distributions. More importantly, when these distributions are applied to durations in stochastic and deterministic disease transmission models to be discussed in later chapters, such as the latent and the infectious periods, we

investigate how different shapes of the hazard functions determine the expected outcomes predicted by these models.

The hazard function uniquely determines the p.d.f. and the survival function via the following relationships

$$\overline{F}_X(x) = \exp\left(-\int_0^x h_X(u)du\right), \tag{2.5}$$

$$f_X(x) = h_X(x)\exp\left(-\int_0^x h_X(u)du\right). \tag{2.6}$$

To ensure that $\overline{F}_X(x) \to 0$ as $x \to \infty$, one adds another condition for the hazard function $\int_0^\infty h_X(x)dx = \infty$.

In other words, one can choose any non-negative continuous function satisfying $h_X(x) \geq 0$, $\int_0^x h_X(x)dx < \infty$ and $\int_0^\infty h_X(x)dx = \infty$ as a hazard function to define a continuous time distribution. The shapes of the hazard functions can be constant, monotone, non-monotone or with very complex forms.

For example, demographers are very familiar with the bath-tub shaped hazard functions to describe the human life span. It decreases sharply for the first few years to reflect the infant mortality, followed by decades of a low level approximately constant hazard rate and then rises as a convex function to reflect aging. Such bath-tub shaped hazard functions are estimated from demographic data and used to construct the survival function $\overline{F}_X(x)$ and predict the human life expectancy. Similar bath-tub shaped hazard functions are also seen in industrial reliability, where the first decreasing phase characterizes the "break-in" period of a new product and the later increasing phase characterizes the "worn-out" process. In Exercise 2.1, we will define a bath-tub shaped hazard function and ask readers to construct the distribution functions and calculate the expected value.

2.2.1 The Constant Hazard Function and the Exponential Distribution

This section starts with the simplest shape, the constant hazard functions. When the hazard function is constant, we commonly call it the hazard rate, or simply, the rate. Without losing generality, we assume $h_0(x) = 1$.

From (2.5), the survival function corresponding to the standard exponential distribution is:

$$\overline{F}_0(x) = \ f_0(x) = e^{-x}. \tag{2.7}$$

Re-scaling the time by a scale parameter $\lambda > 0$, $X = \frac{X_0}{\lambda}$ is distributed according to the exponential distribution with rate λ, given by

$$h_X(x;\lambda) = \lambda, \ \overline{F}_X(x;\lambda) = e^{-\lambda x}. \tag{2.8}$$

The mean and variance of the exponential distribution are

$$E[X] = \frac{1}{\lambda},\ var[X] = \frac{1}{\lambda^2}. \tag{2.9}$$

The median for the exponential distribution is $x_{0.5} = \frac{1}{\lambda}(\log 2) < E[X]$. Therefore,

1. the constant (hazard) rate defines the exponential distribution;
2. the rate λ is also the scale parameter of the exponential distribution;
3. the mean value of the exponential distribution is the reciprocal of the rate.

The exponential distribution has been widely used in infectious disease models, even in models that are deterministic. For example, dynamics infectious disease transmission models (Chap. 5) can be either modelled stochastically as a continuous time Markov chain or deterministically as a system of ordinary differential equations. The common feature in these models is the assumption of the constant recovery rate. It implies that all infected individuals have independently and identically distributed infectious periods following the exponential distribution, with the mean equal to the reciprocal of the recovery rate.

Note that the statement "*the mean equal the reciprocal of the rate*" is implicitly associated with the assumption of the exponential distribution.

2.2.2 Monotonic Hazard Functions Without Upper Limit

Monotonic hazard functions are natural extensions of the constant hazard rates. Take the incubation periods for example, the natural question is whether the hazard of developing clinical symptoms remains constant regardless of the time elapsed since time of infection, or is it an increasing function of time since infection.

The Weibull Distribution One of the generalizations of $h_0(x) = 1$ is a class of monotone hazard functions defined by the power function

$$h_0(x; \varsigma) = \varsigma x^{\varsigma - 1},\ \varsigma > 0. \tag{2.10}$$

The parameter ς determines the shape of $h_0(x)$. When $\varsigma > 1$, $h_0(x; \varsigma)$ monotonically increases to infinity and when $\varsigma < 1$, it decreases to zero, as $x \to \infty$. When $\varsigma = 1$, $h_0(x; \varsigma) = 1$. We call ς the shape parameter.

From (2.5), the survival function is

$$\overline{F}_0(x; \varsigma) = \exp\left(-x^{\varsigma}\right). \tag{2.11}$$

Re-scaling the time by a scale parameter $\lambda > 0$, $X = \frac{X_0}{\lambda}$ is distributed according to

$$h_X(x; \lambda, \varsigma) = \lambda\varsigma\,(\lambda x)^{\varsigma-1} \text{ and } \overline{F}_X(x; \lambda, \varsigma) = e^{-(\lambda x)^{\varsigma}}. \tag{2.12}$$

This is a Weibull distribution with scale parameter λ and shape parameter $\varsigma > 0$. An attractive feature of the Weibull distribution is that the complementary logarithm of the survival function is a linear function of the logarithm of time.

$$\log\left[-\log \overline{F}_X(x; \lambda, \varsigma)\right] = \varsigma \log(\lambda x)\,.$$

The Weibull distribution has mean and variance:

$$E[X] = \frac{1}{\lambda}\Gamma\left(1 + \frac{1}{\varsigma}\right), \tag{2.13}$$

$$var[X] = \frac{1}{\lambda^2}\left[\Gamma\left(1 + \frac{2}{\varsigma}\right) - \Gamma\left(1 + \frac{1}{\varsigma}\right)^2\right]. \tag{2.14}$$

The median is $x_{0.5} = \frac{1}{\lambda}(\log 2)^{\frac{1}{\varsigma}} < E[X]$. The exponential distribution is a special case of the Weibull distribution with $\varsigma = 1$.

Figure 2.1 illustrates the Weibull distribution according to the shape parameter with a standardized time scale. Example 1 will show a Weibull distribution with increasing hazard function used to model the incubation period from HIV to AIDS based on data collected during the 1980s.

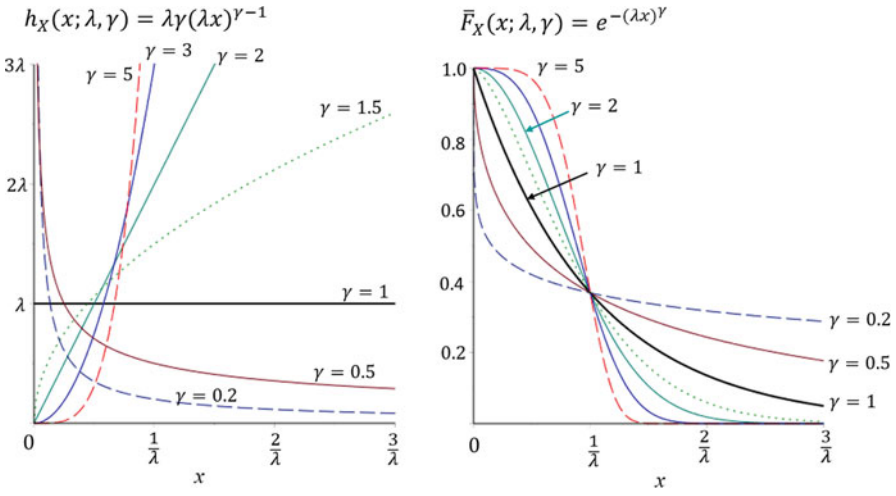


Fig. 2.1 Illustrations of the hazard and the survivor functions for the Weibull distribution by shape parameter ς with standardized time scale

The Gompertz Distribution As an alternative to (2.10), $h_0(x;\varsigma)$ may be modelled by $h_0(x;\varsigma) = \varsigma e^x$ where ς is also a shape parameter. However, it is not a generalization of $h_0(x) = 1$. The distribution for $X = \frac{X_0}{\lambda}$ is called the Gompertz distribution, with its survival function, the p.d.f., and the hazard function given by

$$\overline{F}_X(x;\lambda,\varsigma) = e^{-\varsigma(e^{\lambda x}-1)},\ f_X(x;\lambda,\varsigma) = \lambda\varsigma e^{\lambda x} e^{-\varsigma(e^{\lambda x}-1)},$$
$$h_X(x;\lambda,\kappa) = \lambda\varsigma e^{\lambda x}.$$

This distribution is less commonly used in applications. One of the difficulties is that the moments cannot be written in closed form. Another difficulty is that it is often difficult to fit such a distribution to data. However, worth pointing out the connection of the hazard function of this distribution with the ordinary differential equation $\frac{d}{dx}h(x) = \lambda h(x)$, $x > 0$, which gives a physical rationale for describing human mortality due to aging.

2.2.3 *Hazard Functions that Converge to a Positive Constant as $x \to \infty$*

Distributions defined by hazard functions that converge to a positive constant as $x \to \infty$ are said to have an exponential tail (see more discussion on tail properties in Sect. 2.3.)

The Gamma Distribution

The most commonly used monotone hazard function with such behavior is

$$h_0(x;\kappa) = \frac{x^{\kappa-1}e^{-x}}{\Gamma(\kappa) - \int_0^x u^{\kappa-1}e^{-u}du},\ \kappa > 0. \tag{2.15}$$

The shape parameter is κ. If $\kappa > 1$, $h_0(x;\kappa)$ monotonically increases and approaches the limit $\lim_{x\to\infty} h_0(x;\kappa) = 1$. If $\kappa < 1$, it monotonically decreases and approaches the limit $\lim_{x\to\infty} h_0(x;\kappa) = 1$. When $\kappa = 1$, it returns to the constant hazard $h_0(x;\kappa) = 1$. The survival function and the p.d.f. are

$$\overline{F}_0(x;\kappa) = 1 - \frac{1}{\Gamma(\kappa)}\int_0^x u^{\kappa-1}e^{-u}du,\ f_0(x;\kappa) = \frac{x^{\kappa-1}e^{-x}}{\Gamma(\kappa)},\ \kappa > 0. \tag{2.16}$$

After re-scaling, $X = \frac{X_0}{\lambda}$ is distributed according to

$$\begin{aligned} h_X(x;\lambda,\kappa) &= \frac{\lambda\,(\lambda x)^{\kappa-1}\,e^{-\lambda x}}{\Gamma(\kappa) - \int_0^{\lambda x} u^{\kappa-1}e^{-u}du}, \\ \overline{F}_X(x;\lambda,\kappa) &= 1 - \frac{1}{\Gamma(\kappa)}\int_0^{\lambda x} u^{\kappa-1}e^{-u}du, \\ f_X(x;\lambda,\kappa) &= \frac{\lambda\,(\lambda x)^{\kappa-1}}{\Gamma(\kappa)}e^{-\lambda x}. \end{aligned} \tag{2.17}$$

This is a two-parameter gamma distribution, which reduces to the exponential distribution when $\kappa = 1$. Although both the hazard and the survival functions involve the incomplete gamma function $\int_0^x u^{\kappa-1}e^{-u}du$, the gamma distribution has many desirable features that make it a very convenient choice in infectious disease transmission models. Some of these features are:

1. The gamma distribution with integer valued shape parameter $\kappa = 1, 2, \ldots$ is called the Erlang distribution. The Erlang distribution with mean value μ can be obtained as the sum of κ independently and exponentially distributed lifetimes, each with mean value μ/κ. This feature makes it a popular choice for ordinary differential equation models to handle non-exponential distributions using the linear chain reduction trick (Smith 2011). It is worth noticing that, for Erlang distributions with $\kappa \geq 2$, the variance is always smaller than the exponential distribution with the same mean value.
2. By re-parameterizing $\mu = \frac{\kappa}{\lambda}$ and $\kappa = \kappa$, the mean and the variance of the gamma distribution are expressed as

$$E[X] = \mu,\ var[X] = \frac{\mu^2}{\kappa}.$$

   Given a finite mean value μ, κ ranks the variance. The variance approaches zero as $\kappa \to \infty$ on the one hand, and approaches infinity as $\kappa \to 0$ on the other hand.
3. Unlike the Weibull distribution, the shape parameter κ in the gamma distribution ranks both the hazard function and the survival function according to the order as shown in Fig. 2.2. These are important stochastic orders comparing the magnitudes of lifetime distributions, which will be discussed in more detail in Sect. 2.5.1.
4. The p.d.f. of the gamma distribution has a very flexible form, from a highly skewed shape with a long tail (J-shape) when $\kappa < 1$, to the negative exponential function as $\kappa = 1$ and towards a bell-shape when $\kappa > 1$, as shown in Fig. 2.3.

The gamma distribution has a simple explicit form of the Laplace transform which will be discussed in detail in Sect. 2.4. It is a very useful tool in disease models involving convolutions. It can also be used to compare variability of lifetimes. For example, how variability of the latent period or the infectious period of the infected

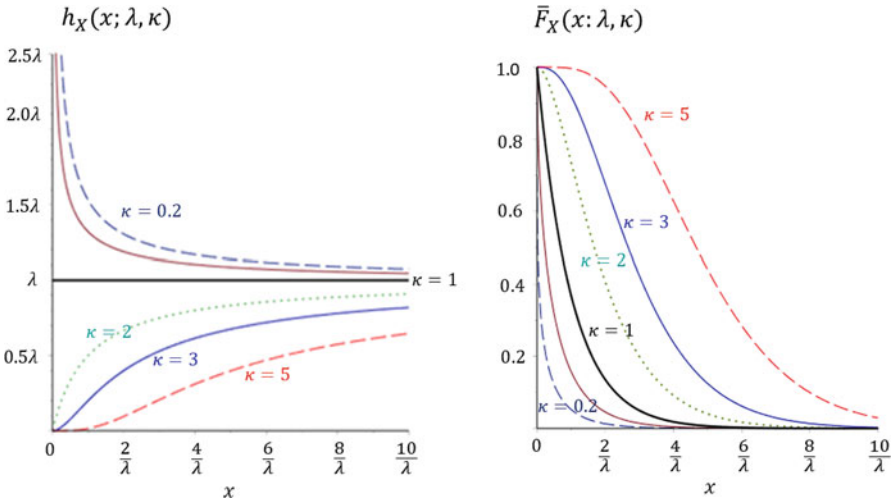


Fig. 2.2 Illustrations of the hazard and the survival functions of the gamma distribution

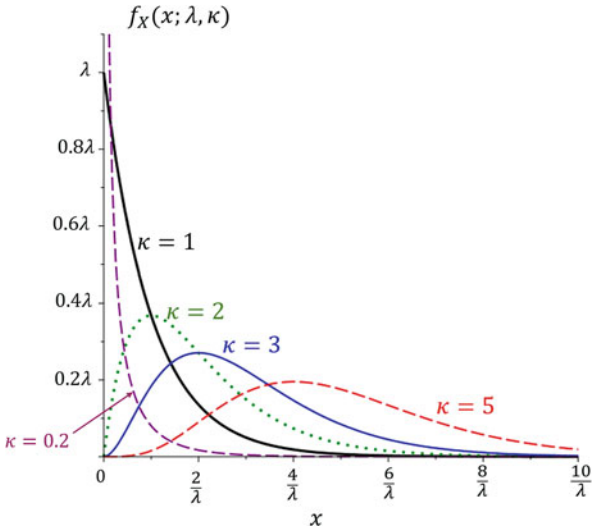


Fig. 2.3 Flexible shapes of the p.d.f. of the gamma distribution

individuals may affect the behavior of an epidemic at its initial growth phase; how variability of the latent period or the infectious period of the infected individuals may affect the effectiveness of certain public health measures aimed at controlling the spread of an epidemic. The Laplace transform is a powerful tool in frailty models, which is a random effect model for unobservable heterogeneity. Thus, applications of Laplace transforms, especially those with the gamma distribution, will be discussed in many chapters throughout this book.

The Inverse-Gaussian Distribution: Non-monotone and Converge to a Positive Constant as $x \to \infty$

As an alternative to the gamma distribution, the inverse-Gaussian distribution also has an exponential tail. Let λ be the scale parameter and κ the shape parameter, the p.d.f. and the survival function are

$$\begin{aligned} f_X(x;\lambda,\kappa) &= \frac{\lambda\kappa}{\sqrt{2\pi(\lambda x)^3}} \exp\left(-\frac{(\lambda x-\kappa)^2}{2\lambda x}\right) \\ \overline{F}_X(x;\lambda,\kappa) &= \Phi\left(\frac{\lambda x-\kappa}{\sqrt{\lambda x}}\right) - e^{2\kappa}\,\Phi\left(-\frac{\lambda x+\kappa}{\sqrt{\lambda x}}\right) \end{aligned} \quad \lambda,\kappa > 0 \tag{2.18}$$

where $\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-\frac{w^2}{2}} dw$ is the standard normal distribution function.

The hazard function $h_X(x;\lambda,\kappa) = \frac{f_X(x;\lambda,\kappa)}{\overline{F}_X(x;\lambda,\kappa)}$ increases from zero to a maximum and then decreases (for certain range κ) to an asymptotic value $\frac{\lambda}{2}$ as shown in Fig. 2.4.

The inverse-Gaussian distribution shares many of the features of the gamma distribution.

1. Like the gamma distribution, the shape parameter κ ranks both the hazard function and the survival function.
2. By re-parametrization $\mu = \frac{\kappa}{\lambda}$ and $\kappa = \kappa$, the mean and the variance are also expressed as

$$E[X] = \mu,\ var[X] = \frac{\mu^2}{\kappa}.$$

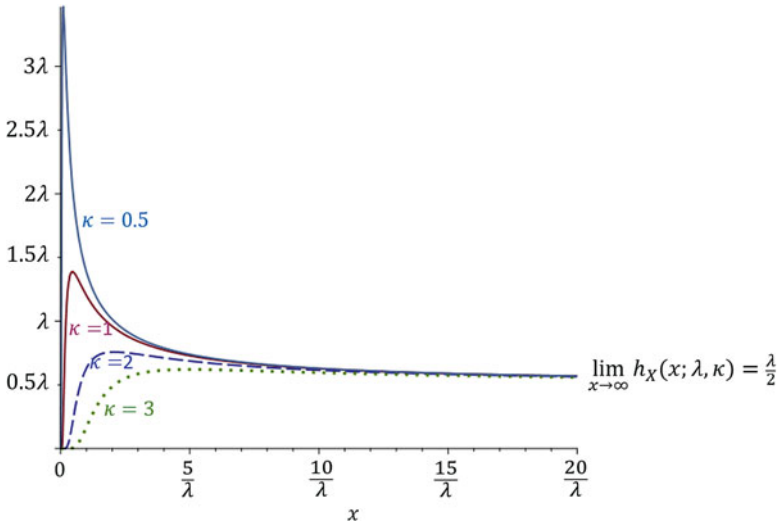


Fig. 2.4 Shapes of the hazard function of the inverse-Gaussian distribution

Given finite mean value μ, κ ranks the variance. The variance approaches zero as $\kappa \to \infty$ on the one hand, and approaches infinity as $\kappa \to 0$ on the other hand.

3. Like the gamma distribution, the inverse-Gaussian distribution also has an explicit form of the Laplace transform, making it useful to study many important aspects of infectious disease dynamics.

Unlike the gamma distribution, the inverse-Gaussian distribution does not include the exponential distribution as a special case. However, at $\kappa = 1$, the first two moments are $E[X] = \mu$ and $var[X] = \mu^2$ which are the same as those of the exponential distribution.

2.2.4 *Two Empirical Distributions for Disease Progression Characterized by Non-monotone Hazard Functions*

Empirical evidence suggests that if an individual is still in her/his latent period after a long time since exposure, this individual is more likely to remain non-infectious. Similarly, if after a long time since infection an individual is still symptom free, this individual is more likely to remain symptom free. This suggests that a hazard function may initially increase but eventually decreases to zero after reaching a maximum value.

The Log-Normal Distribution as a Model for the Incubation Period

The incubation period is the duration from the time of infection until the time of developing clinical symptoms within an infected individual. Sartwell (1966) studied various infectious diseases and found that the incubation periods of acute infectious diseases tend to follow the log-normal distribution. The validity of the log-normal assumption for the incubation periods has been only supported by empirical evidence. Many studies were carried out by epidemiologists in Japan (Nishiura 2007) through testing of goodness-of-fit to acute infectious disease data. Such empirical evidence, to the best, supports the idea that the incubation period tends to follow a distribution that is right skewed with a long tail, as characterized by the shape of the hazard function.

Under the convention used in this chapter, by setting the scale parameter $\lambda = 1$, the survival function and the p.d.f. of the log-normal distribution with a shape parameter $\varsigma > 0$ are

$$\overline{F}_0(x;\varsigma) = 1 - \Phi(\varsigma \log x),\ f_0(x;\varsigma) = \frac{1}{\sqrt{2\pi}} \frac{\varsigma}{x} e^{-\frac{(\varsigma \log x)^2}{2}} \tag{2.19}$$

where $\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-\frac{w^2}{2}} dw$ is the normal distribution function. The hazard function is therefore

$$h_0(x; \varsigma) = \frac{\frac{1}{\sqrt{2\pi}} \frac{\varsigma}{x} \exp\left\{-\frac{(\varsigma \log x)^2}{2}\right\}}{1 - \Phi(\varsigma \log x)}, \ \varsigma > 0. \tag{2.20}$$

It can be shown that for $\varsigma > 0$, (i) $h_0(0; \varsigma) = 0$; (ii) $h_0(x; \varsigma)$ increases to a maximum then approaches zero monotonically as $x \to \infty$. The turning point occurs at x^* such that $h_0(x^*; \varsigma) = \frac{1}{x^*}\left(\varsigma^2 \log(\varsigma x^*) + 1\right)$. After re-scaling, $X = \frac{X_0}{\lambda}$ is distributed according to

$$\overline{F}_X(x; \lambda, \varsigma) = 1 - \Phi(\varsigma \log(\lambda x)), \ f_X(x; \lambda, \varsigma) = \frac{\varsigma}{\sqrt{2\pi} x} e^{-\frac{(\varsigma \log(\lambda x))^2}{2}}. \tag{2.21}$$

The log-normal distribution has mean and variance:

$$E[X] = \frac{1}{\lambda} \exp\left(\frac{1}{2\varsigma^2}\right), \ var[X] = \frac{1}{\lambda^2} e^{\frac{1}{\varsigma^2}} \left(e^{\frac{1}{\varsigma^2}} - 1\right).$$

The log-normal distribution is linked to the normal distribution for $Y = \log X$. Letting $\mu_Y = \log \lambda$ and $\sigma = \varsigma^{-1}$, it can be shown that Y follows the normal distribution $N(\mu_Y, \sigma^2)$. This clearly demonstrates the advantage of fitting the log-normal distribution to data in statistical analysis.

The Log-Logistic Distribution

The name log-logistic distribution is due to its link with the logistic distribution for $Y = \log X$. Where data might support the log-normal distribution, it is equally suitable to consider the log-logistic distribution as an alternative. Setting the scale parameter $\lambda = 1$, the hazard function is

$$h_0(x; \varsigma) = \frac{\varsigma x^{\varsigma - 1}}{1 + x^{\varsigma}}, \ \varsigma > 0. \tag{2.22}$$

It can be shown that for the shape parameter $\varsigma > 1$, (i) $h_0(0; \varsigma) = 0$; (ii) $h_0(x; \varsigma)$ increases to a maximum at $x^* = (\varsigma - 1)^{\frac{1}{\varsigma}}$ then approaches zero monotonically as $x \to \infty$. From (2.5), the survival function is

$$\overline{F}_0(x; \varsigma) = \frac{1}{1 + x^{\varsigma}}. \tag{2.23}$$

After re-scaling, $X = \frac{X_0}{\lambda}$ is distributed according to

$$h_X(x;\lambda,\varsigma) = \frac{\lambda\varsigma\,(\lambda x)^{\varsigma-1}}{1+(\lambda x)^{\varsigma}},\ \overline{F}_X(x;\lambda,\varsigma) = \frac{1}{1+(\lambda x)^{\varsigma}}. \tag{2.24}$$

The mean $E[X]$ and the variance $var[X]$ are

$$E[X] = \frac{1}{\lambda}\Gamma\left(1+\frac{1}{\varsigma}\right)\Gamma\left(1-\frac{1}{\varsigma}\right),\ \text{if } \varsigma > 1$$

$$var[X] = \frac{1}{\lambda^2}\left[\Gamma\left(1+\frac{2}{\varsigma}\right)\Gamma\left(1-\frac{2}{\varsigma}\right) - \Gamma\left(1+\frac{1}{\varsigma}\right)^2\Gamma\left(1-\frac{1}{\varsigma}\right)^2\right],\ \text{if } \varsigma > 2.$$

When the shape parameter $\varsigma \le 2$, the hazard function decreases to zero very fast and the survival function decreases to zero at a very slow speed, so that the tails are "heavy" enough to preclude the existence of a finite mean or variance (see more discussion in Sect. 2.3.2). When $\varsigma \le 2$, the variance of the log-logistic distribution does not exist. When $\varsigma \le 1$, the mean value does not exist. If the infectious period follows such a distribution, one of the most important parameters in disease transmission models, the basic reproduction number, is not defined and many fundamental theories based on the existence of an epidemic threshold will no longer hold.

When $0 < \varsigma < 1$, the log-logistic distribution is also called the Pareto-III distribution (Marshall and Olkin 2007) when it is regarded as an extension of the Pareto distribution (see Sect. 2.3.2) with hazard function monotonically decreasing to zero.

The Log-Logistic Distribution vs. the Log-Normal Distribution For both the log-logistic and the log-normal distributions, the median is $x_{0.5} = \lambda^{-1}$. The log-logistic distribution closely approximates the log-normal distribution. Both models may fit equally well to observed data in many applications. Suppose that X_1 follows a log-normal distribution as defined in (2.21) with scale parameter λ_1 and shape parameter ς_1. One can find a log-logistic distributed X_2 with the same scale parameter $\lambda_2 = \lambda_1$ so that these two distributions have the same median, and then calibrate the shape parameter ς_2 in the log-logistic distribution so that the two distributions are in close agreement. For instance, one can choose a quantile near the tail end, such as the 0.95th quantile $x_{0.95}$, and calibrate ς_2 so that both distributions have the same 0.95th quantile. This can be done by solving the equation $\frac{1}{1+(x_{0.95}/\lambda_1)^{\varsigma_2}} = 0.05$. Figure 2.5 compares the two distributions such that both have the same median and the same quantile $x_{0.95}$.

The Log-Logistic Distribution vs. the Weibull Distribution While the similarity of the log-logistic distribution and the log-normal distribution is empirical, the connection between the log-logistic distribution and the Weibull distribution is profound. It is worth noticing that the numerator of the hazard function of the

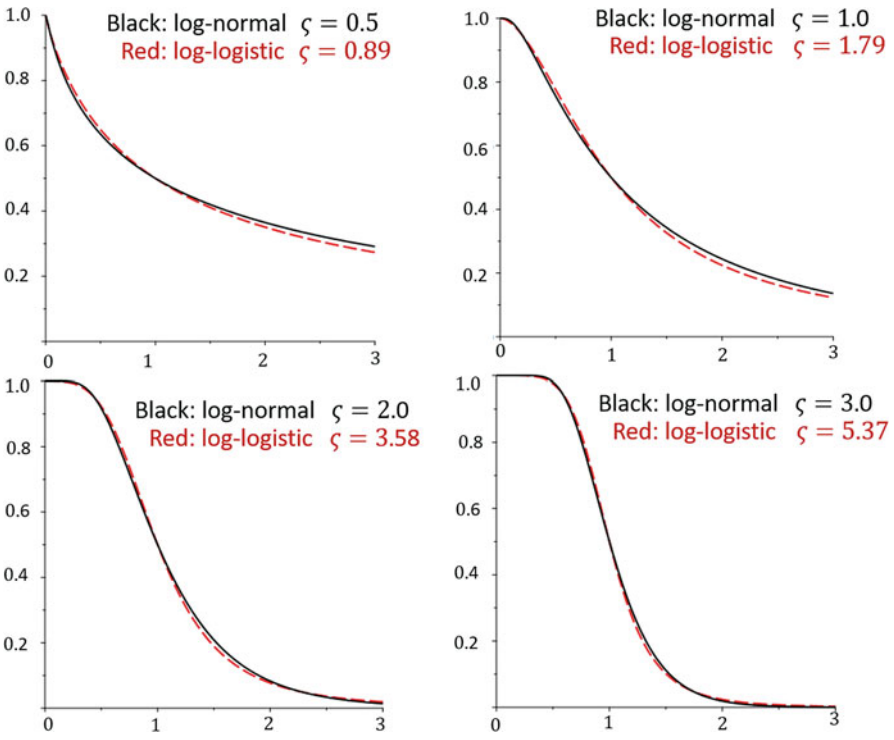


Fig. 2.5 Compare log-logistic and log-normal survival functions with median = 1 at various shape parameters

log-logistic distribution, $\lambda\varsigma\,(\lambda x)^{\varsigma-1}$, is the hazard function of the Weibull distribution in (2.12); and $(\lambda x)^{\varsigma}$ in the denominator of the hazard function of the log-logistic distribution is the cumulative hazard function of the Weibull distribution. This is not a coincidence. Later in Sect. 2.6, we shall see that the hazard function of the log-logistic distribution can be derived as a special case of a random mixture of heterogeneous individuals. Each of them has an individual hazard function of the Weibull distribution. Biologically speaking, it might be natural to assume a monotonically increasing hazard function at the individual level to model lifetimes such as the incubation period, as well as the latent period or the infectious period. However, due to unobservable individual heterogeneity, at a cohort level, these lifetimes tend to behave with a "dampened" hazard function by a random effect (Sect. 2.6).

Nishiura (2007) pointed out that, with respect to modelling incubation periods, the log-normal distribution, which empirically mimics the log-logistic distribution, often fits well with data from acute diseases but not as well for fitting data arising from chronic diseases. For diseases with long incubation periods, the Weibull distribution is often preferred in the literature. In fact, for the log-logistic distribution

with shape parameter $\varsigma > 2$, the early part of the hazard function and the corresponding survival function are in close agreement with that of the Weibull distribution.

Example 1 In the 1980s, results from cohort studies were used to guide the selection of models for the incubation distribution from HIV infection to AIDS symptoms. These studies are: (1) analysis on 458 hemophiliacs for age > 20 (Goedert et al. 1989), (2) data from the San Francisco City Clinic Cohort of homosexual men enrolled in hepatitis vaccine study (Hessol et al. 1989) (3) a cohort of 468 seroconverters who were injecting drug users (Italian Seroconversion Study 1992), (4) an Italian cohort of 952 seropositive males and females (Gauvreau et al. 1994); as well as estimation by deconvoluting the AIDS incidence data in San Francisco using epidemiological surveys to reconstruct the distribution of incubation times (Bacchetti and Moss 1989). The maximum follow-up time of these cohort was only 11 years. These studies provided a general picture that the probability of developing AIDS within the first 2 years of seroconversion is less than 3%. The cumulative probability of developing AIDS within 7 years after seroconversion is approximately 25% and the cumulative probability of developing AIDS within 10 years approaches 50%. These are shown by the Kaplan-Meier estimates in Fig. 2.6. A widely used model is a Weibull distribution with scale parameter $\lambda = 0.00211$ and shape parameter $\varsigma = 2.516$ (Brookmeyer and Goedert 1989). This distribution suggests a monotonically increasing hazard function. A log-logistic model with scale parameter $\lambda = 0.1$ and shape parameter $\varsigma = 3.08$ was used by Lui et al. (1988) to describe HIV incubation among gay men and is in close agreement with the aforementioned Weibull distribution for the first 10 years after sero-conversion. Due to the non-monotone hazard function of the log-logistic distribution, HIV progression after 10 years is much slower (with a dampened hazard function) than that suggested by the Weibull model.

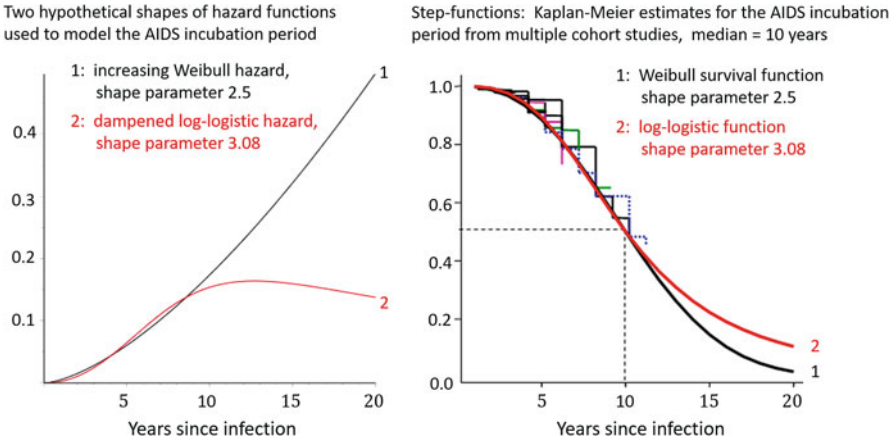


Fig. 2.6 Compare the Weibull distribution and the log-logistic distribution, both with median = 10 years, as models for the incubation period from HIV to AIDS

2.2.5 *Parametric Lifetime Distributions with More than Two Parameters*

One may create distributions with more flexible shapes of the hazard functions by introducing more parameters. For instance, (2.22) can be generalized by involving another shape parameter κ and hence $h_0(x;\varsigma,\kappa)=\frac{\varsigma\kappa x^{\varsigma-1}}{1+x^{\varsigma}}$. If X_0 has a distribution given by such a hazard function, then the hazard function for $X=\frac{X_0}{\lambda}$ has three parameters $f_X(x;\lambda,\varsigma,\kappa)=\frac{\lambda\varsigma\kappa(\lambda x)^{\varsigma-1}}{(1+(\lambda x)^{\varsigma})^{\kappa+1}}$, $\overline{F}_X(x;\lambda,\varsigma,\kappa)=\frac{1}{(1+(\lambda x)^{\varsigma})^{\kappa}}$, and $h_X(x;\lambda,\varsigma,\kappa)=\frac{\lambda\varsigma\kappa(\lambda x)^{\varsigma-1}}{1+(\lambda x)^{\varsigma}}$. This gives the Pareto IV distribution (also known as the Burr's distribution) which includes the log-logistic distribution ($\kappa=1$), the Pareto I and II distributions ($\varsigma=1$), the Weibull distribution ($\kappa\to\infty$), and the exponential distribution ($\kappa\to\infty,\varsigma=1$). Lawless (2003) provides more detailed accounts on this distribution.

Another three-parameter distribution is the generalized-gamma distribution which generalizes (2.17) with p.d.f. $f_X(x;\lambda,\kappa)=\frac{\lambda\varsigma(\lambda x)^{\varsigma\kappa-1}}{\Gamma(\kappa)}e^{-(\lambda x)^{\varsigma}}$. This distribution includes the exponential ($\varsigma=\kappa=1$), the gamma($\varsigma=1$), and the Weibull ($\kappa=1$) distributions. The log-normal distribution is also a limiting case as $\kappa\to\infty$ (Meeker and Escobar 1998, p. 100; Kalbfleisch and Prentice 2002, p. 37).

One may also create hazard functions with other shapes. Hazard functions with complicated shapes can be also constructed as continuous piece-wise functions (Exercise 2.1). The biology of the infection ultimately determines the most appropriate forms for the hazard functions for disease modelling. For most part of the book, we only use parametric lifetime distributions involving no more than two parameters.

2.3 The Residual Life Distribution and the Tail Property

2.3.1 *The Residual Life Distribution as Uniquely Determined by the Hazard Function*

The *residual life* distribution for X is defined through a conditional survival function denoted by the conditional probability

$$\overline{F}_X(x|\tau)=\frac{\overline{F}_X(\tau+x)}{\overline{F}_X(\tau)}=\Pr\{X>\tau+x|X>\tau\}$$

The residual life distribution is uniquely determined by the hazard function $h_X(x)$, as

$$\overline{F}_X(x|\tau) = \exp\left(-\int_{\tau}^{\tau+x} h_X(u)du\right). \tag{2.25}$$

The Shape of Hazard Functions and the Tail Property

The tail property refers to the behavior of the survival function of the residual time.

1. If $h_X(x)$ is strictly increasing with $h_X(x) \to \infty$, then $\overline{F}_X(x|\tau)$ is a decreasing function for τ, $\lim_{\tau\to\infty} \overline{F}_X(x|\tau) = 0$, for any $x > 0$.
2. If $h_X(x) \to \lambda$, the distribution has *exponential tail* $\lim_{\tau\to\infty} \overline{F}_X(x|\tau) = e^{-\lambda x}$, for any $x > 0$.
3. If there exists $x^* \geq 0$ such that for $x > x^*$, $h_X(x)$ is a decreasing function of x with $\lim_{x\to\infty} h_X(x) = 0$, then

$$\lim_{\tau\to\infty} \overline{F}_X(x|\tau) = 1, \text{ for any } x > 0. \tag{2.26}$$

An intuitive interpretation for (2.26) is that, if X ever exceeds a large value, it is likely to exceed any larger value. A distribution is said to be *heavy tailed* if satisfying (2.26).

2.3.2 *Some Highly Skewed, Heavy Tailed Distributions*

In modelling disease transmission, one often encounters a situation in which the majority of individuals are associated with very small values of X, the distribution is highly skewed, leaving few individuals with very large values.

Of the distributions that we have covered, the exponential distribution, the gamma distribution when $\kappa < 1$, the Weibull distribution when $\varsigma < 1$ and the log-logistic distribution when $\varsigma \leq 1$ all have a common feature that the hazard function is non-increasing and $\frac{d^2}{dx^2} \log \overline{F}_X(x) = -\frac{d}{dx} h_X(x) \geq 0$. In other words, $\overline{F}_X(x)$ is log-convex. Some of the survival functions decrease to zero at very slow speeds, so that their tails are "heavy" enough to preclude the existence of finite mean and variance. In fact, it needs not have finite moment in any positive order (Marshall and Olkin 2007, Proposition 4.C.12).

The Pareto Distributions

A family of highly skewed distributions is the Pareto distributions. It has special importance in infectious disease modelling. If the infectious period follows such

a distribution, many fundamental theories based on the existence of an epidemic threshold will no longer hold. The following survival functions are Pareto distributions, with terminology from Arnold (1983),

$$\overline{F}_X(x) = \begin{cases} (1+\lambda x)^{-1}, & \lambda > 0 & \text{Pareto I} \\ (1+\lambda x)^{-\kappa}, & \lambda, \kappa > 0 & \text{Pareto II} \\ (1+(\lambda x)^{\varsigma})^{-1}, & \lambda > 0,\ 0 < \varsigma < 1 & \text{Pareto III} \\ (1+(\lambda x)^{\varsigma})^{-\kappa}, & \lambda, \kappa > 0,\ 0 < \varsigma < 1 & \text{Pareto IV} \end{cases}. \tag{2.27}$$

For the rest of the book, we use Pareto-II for the Pareto distribution, which includes Pareto-I as a special case. Pareto-III is a subset of the log-logistic distribution. Pareto-IV generalizes Pareto-I, II, and III.

The Pareto distributions are sometimes called the power-law distributions. In general, *power-law* distributions can be defined as distributions such that for some $\varsigma > 0$, $A > 0$, $\lim_{x\to\infty} \frac{\overline{F}_X(x)}{x^{\varsigma+1}} = A$.

Sometimes one also compares the tail properties against the exponential versus the power-law shaped tails.

1. A distribution is *sub-exponential*, if $\lim_{x\to\infty} \frac{\overline{F}_X(x)}{\exp(-\lambda x)} = \infty$. It has a heavier tail (goes to zero more slowly) compared to an exponential tail. All heavy tail distributions are sub-exponential.
2. *Power-law* distributions have heavy tail distributions, but not all sub-exponential distributions are power-law.

We summarize some of the distributions that we have discussed in Table 2.1 where $I(\kappa, x) = \frac{1}{\Gamma(\kappa)} \int_0^x u^{\kappa-1} e^{-u} du$ is the incomplete gamma function.

2.4 The Laplace Transform for Life Distributions

We use the notation $L[\phi](s) = \int_{-\infty}^{\infty} e^{-sx} \phi(x) dx$ for the Laplace transform with respect to a function $\phi(x)$ provided that the integration exists. In mathematical dynamic models, the Laplace transform is well known in its usefulness in solving systems of linear ordinary differential equations with constant coefficients, as well as in differential-difference equations and partial differential equations. We refer to Bellman and Roth (1984) for detailed accounts. Another important application is in solving the renewal-type equations of the form

$$u(t) = v(t) + \int_0^t f(x) u(t-x) dx, \tag{2.28}$$

where $v(t)$ is a uniformly bounded function. In terms of Laplace transform, this equation can be re-written by $L[u](s) = \frac{L[v](s)}{1-L[f](s)}$. This usefulness is significant both in deterministic dynamic models and in stochastic processes. In deterministic

Table 2.1 Compare survival functions, hazard functions, residual life distributions, and tail properties for some distributions

	Survival func. $\overline{F}_X(x)$	Hazard func. $h_X(x)$	Tail property $\overline{F}_X(x\|\tau)$
Exponential	$\exp\{-(\lambda x)\}$	λ	Exponential $= \exp\{-(\lambda x)\}$
Gamma	$1 - I(\kappa, \lambda x)$	$\to \lambda$	Exponential $\to \exp\{-(\lambda x)\}$ as $\tau \to \infty$
Weibull $\varsigma < 1$	$\exp\{-(\lambda x)^{\varsigma}\}$	Decreases for all $x > 0$ $\lim_{x\to\infty} \lambda\varsigma\,(\lambda x)^{\varsigma-1} = 0$	Sub-exponential, not power-law $\to 1$ as $\tau \to \infty$
Log-logistic	$(1 + (\lambda x)^{\varsigma})^{-1}$	Decreases for $x > \frac{1}{\lambda}(\varsigma - 1)^{\frac{1}{\varsigma}}$ $\lim_{x\to\infty} \frac{\lambda\varsigma(\lambda x)^{\varsigma-1}}{1+(\lambda x)^{\varsigma}} = 0$	Sub-exponential, not power-law $\to 1$ as $\tau \to \infty$
Log-normal	$1 - \Phi(\varsigma \log(\lambda x))$	Decreases for $x > x^*$ (exists) $\to 0$ as $x \to \infty$	Sub-exponential, not power-law $\to 1$ as $\tau \to \infty$
Pareto II	$(1 + \lambda x)^{-\kappa}$	Decreases for all $x > 0$ $\lim_{x\to\infty} \frac{\lambda\kappa}{1+\lambda x} = 0$	Sub-exponential, power-law $\to 1$ as $\tau \to \infty$

models, if $v(t) = \frac{d}{dt}u(t)$, then (2.28) is well recognized as the Volterra integro-differential equation in mathematical biology. The renewal-type equation (2.28) plays an important role in stochastic models, especially those based on renewal processes.

With respect to lifetime distributions, where $f(x)$ is a p.d.f. of a lifetime X, the Laplace transform $L[f](s) = \int_0^\infty e^{-sx} f(x)dx$ always exists. We also simplify the language so that $L[f](s)$ is sometimes referred to as the Laplace transform of the lifetime X.

The Laplace transform will be a useful tool throughout this book. Here we summarize some of its useful features.

2.4.1 *Laplace Transform of the Sum of Two Independent Random Variables*

Suppose two non-negative random variables X_1 and X_2 are independent, with p.d.f.'s $f_1(t)$ and $f_2(t)$, respectively. The p.d.f. of the sum X_1+X_2 is the convolution $g(t) = \int_0^\infty f_1(u) f_2(t-u)du$, with Laplace form being the multiplication of two separate Laplace transforms:

$$L[g](s) = L[f_1](s)L[f_2](s) = \left(\int_0^\infty e^{-st} f_1(t)dt\right)\left(\int_0^\infty e^{-st} f_2(t)dt\right). \tag{2.29}$$

2.4.2 Moment Generating Property

If all moments for X exist and denote $\mu_n = E[X^n]$, then

$$L[f](s) = \sum_{n=0}^{\infty}(-1)^n \frac{s^n}{n!}\mu_n. \tag{2.30}$$

The nth moment is $\mu_n = (-1)^n \frac{d^n}{ds^n} L[f](s)|_{s=0}$. In particular, the mean $\mu_1 = -\frac{d}{ds} L[f](s)|_{s=0}$, and the second moment $\mu_2 = \frac{d^2}{ds^2} L[f](s)|_{s=0}$.

2.4.3 As a Probability Comparing X Against an Exponentially Distributed Lifetime Y

Let X be a lifetime with p.d.f. $f(x)$ and Y be an exponentially distributed lifetime with a scale parameter (also the hazard rate) s such that $\Pr(Y > y) = e^{-sy}$, then the probability $\Pr(Y > X)$ is the Laplace transform

$$\Pr(Y > X) = \int_0^\infty e^{-sx} f(x)dx = L[f](s). \tag{2.31}$$

2.4.4 Laplace Transform as a Survival Function

In (2.31), the survival function of the exponential distribution e^{-st} was viewed as a function of t with s being the scale parameter. It can be also viewed as a function of s with t being the scale parameter. Then $L[f](s)$ is a monotonically decreasing function of s satisfying $L[f](0) = 1$ and approaches zero as $s \to \infty$, a property of a survival function. It also has a valid interpretation as a survival function (see Sect. 2.6), with some important properties.

A function f with domain $(0, \infty)$ is said to be completely monotonic if it possesses derivatives $f^{(n)}(x) = \frac{d^n}{dx^n} f(x)$ for all $n = 0, 1, 2, 3, \ldots$ and if

$$(-1)^n f^{(n)}(x) \geq 0 \tag{2.32}$$

for all $x > 0$. The Laplace transform $L[f](s)$ is completely monotonic because $(-1)^n \frac{d^n}{ds^n} L[f](s) = \int_0^\infty x^n e^{-sx} f(x)dx \geq 0$ for all $s > 0$.

It has been shown a survival function is completely monotonic if and only if it has a completely monotonic density (Marshall and Olkin 2007). As a survival function $L[f](s)$, its density $-\frac{d}{ds} L[f](s) = \int_0^\infty x e^{-sx} f(x)dx$ is completely monotonic. A completely monotonic function is log-convex (Marshall and Olkin 2007). Hence, $\log L[f](s)$ is a convex function for all $s > 0$.

2.5 Comparing Two Lifetimes X_1 and X_2

When both X_1 and X_2 are random, comparisons between X_1 and X_2 need to be carefully defined and examined. Shaked and Shanthikumar (2007) and Marshall and Olkin (2007) are excellent monographs with extensive definitions, propositions, and discussions. This section will only cover two aspects: the comparison of magnitude and the comparison of variability between two lifetimes.

2.5.1 Comparing Magnitudes

When comparing two lifetimes X_1 and X_2, people often use "average" as a single summary measure. Such an ordering is defined by $E[X_1] \leq E[X_2]$ (should these mean values exist) and is denoted by $X_1 \leq_{ave} X_2$ to reflect that it is an ordering "on average." One of the most commonly used summary measures is the *Life Expectancy*, whether the life expectancy in population A is longer than the life expectancy in population B.

Using a single summary measure to compare two random variables with large variations can be problematic and subject to mis-interpretations. Figure 2.7 illustrates two crossings survival functions, in which $\mu_2 = \int_0^\infty \overline{F}_2(x)dx$ is twice as $\mu_1 = \int_0^\infty \overline{F}_1(x)dx$. Therefore by definition, $X_1 \leq_{ave} X_2$. On the other hand, if we examine the median values, it turns out that the median for X_1 is twice the value of the median for X_2. So which is larger?

Although Fig. 2.7 is artificially created, there is no shortage of real-life examples where survival functions cross. For instance, Li et al. (2015) provide examples from clinical trials of crossing survival curves, as reproduced in Fig. 2.8.

A natural and stronger ordering is the stochastic order.

Definition 2 X_1 is smaller than X_2 in *stochastic order, denoted as* $X_1 \leq_{st} X_2$, the following statements are equivalent:

1. if corresponding survival functions $\overline{F}_1(x) \leq \overline{F}_2(x)$ for all $x > 0$;
2. if $E[\Phi(X_1)] \leq E[\Phi(X_2)]$ for all increasing functions Φ such that the expectations exist;
3. if $\Pr\{\Phi(X_1) > x\} \leq \Pr\{\Phi(X_2) > x\}$ for all increasing functions Φ, for all x.

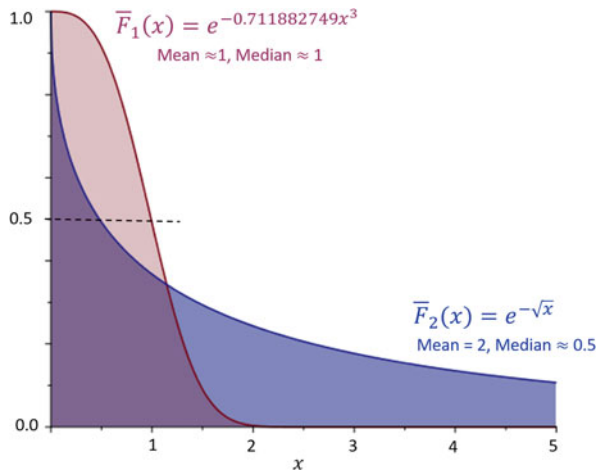


Fig. 2.7 Illustration of two crossing survival functions: the one with twice the mean value has half the median value of the other

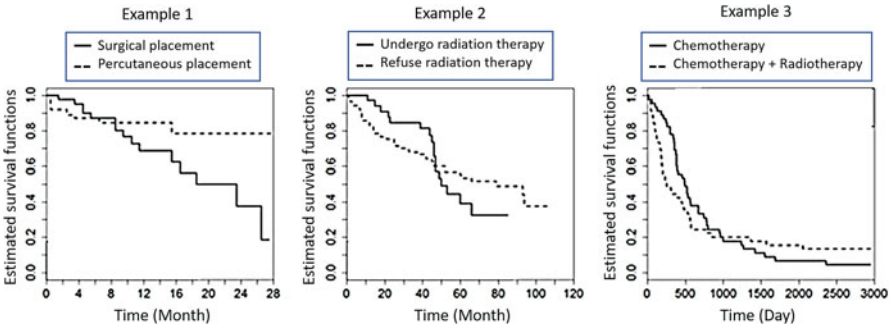


Fig. 2.8 Examples of crossing survival functions. Source: Li et al. Li et al. (2015)

For the proof of equivalence in the definition statements, we refer readers to Proposition 2.A.2. in Marshall and Olkin (2007). It follows that, by taking $\Phi(x) = x^r$, if $X_1 \leq_{st} X_2$, then $E[X_1^r] \leq E[X_2^r]$ for $r \geq 0$. Hence $X_1 \leq_{st} X_2 \Rightarrow X_1 \leq_{ave} X_2$.

Recall the definition of the scale parameter $\overline{F}_X(x;\lambda) = \Pr\{X_0 > \lambda x\} = \overline{F}_0(\lambda x)$, then within the same distribution family, keeping other parameters fixed, the scale parameter ranks the distribution according to stochastic order. Consequently, the log-linear model known as the accelerated life time model (Lawless 2003) of the form

$$\log \lambda = \alpha + \beta_1 z_1 + \cdots + \beta_m z_m$$

compares lifetimes according to stochastic order.

However, the stochastic ordering is still a partial ordering. The stochastic order is not preserved under residual life distribution, that is, $\overline{F}_1(x) \leq \overline{F}_2(x)$ for all x does not lead to $\overline{F}_1(x|\tau) \leq \overline{F}_2(x|\tau)$ for all x and τ. A counterexample can be given by comparing a log-logistic distribution with a similar log-logistic distribution by shifting the location. A stronger ordering that preserves the stochastic order under residual life distribution is the ordering that compares the hazard functions because of (2.25).

Stochastic and Hazard Rate Ordering of Lifetimes

Definition 3 X_1 is smaller than X_2 in hazard rate order, *denoted as* $X_1 \leq_{hr} X_2$, the following statements are equivalent:

1. if corresponding hazard functions $h_1(x) \geq h_2(x)$, for all $x > 0$;
2. if $\overline{F}_1(x|\tau) \leq \overline{F}_2(x|\tau)$ for all x and τ;
3. if the ratio $\overline{F}_1(x)/\overline{F}_2(x)$ is a decreasing function of x.

Example 4 Suppose that X_1 follows a log-logistic distribution with shape parameter $\varsigma = 4$ such that $\overline{F}_1(x) = (1 + x^4)^{-1}$ and $X_2 = X_1 + 2$ so that $\overline{F}_2(x) = 1$ if $x < 2$; $\left(1 + (x-2)^4\right)^{-1}$ if $x \geq 2$. Clearly, $\overline{F}_1(x) \leq \overline{F}_2(x)$ for all x and hence $X_1 \leq_{st} X_2$. However, they do not follow hazard rate order. Figure 2.9 shows the case when $\tau = 2$ that all the three equivalent statements in Definition 3 are not met.

2.5.2 Comparing Variabilities

A General Description of Variability Is Based on "Majorization"

If X_1 and X_2 have equal mean values (should they exist), a general description of variability is based on "majorization." Let f_1 and f_2 be the corresponding p.d.f.s for X_1 and X_2, the verbal description for X_2 *being more dispersed (spread out) than* X_1 is reflected in Fig. 2.10 about the change of signs between f_1 and f_2 and their corresponding survival functions $\overline{F}_1$ and $\overline{F}_2$.

The following two definitions are equivalent (Marshall and Olkin 2007). In verbal terms:

Definition 5 $X_1 \leq_{cv} X_2$ if and only if $E[X_1] = E[X_2]$ plus the following two statements:

1. $f_2(x) - f_1(x)$ has two sign changes and the sign sequence is: $+, -, +$ (see Fig. 2.10).
2. $\overline{F}_1(x) - \overline{F}_2(x)$ has one sign change and the sign sequence is: $+, -$ (see Fig. 2.10).

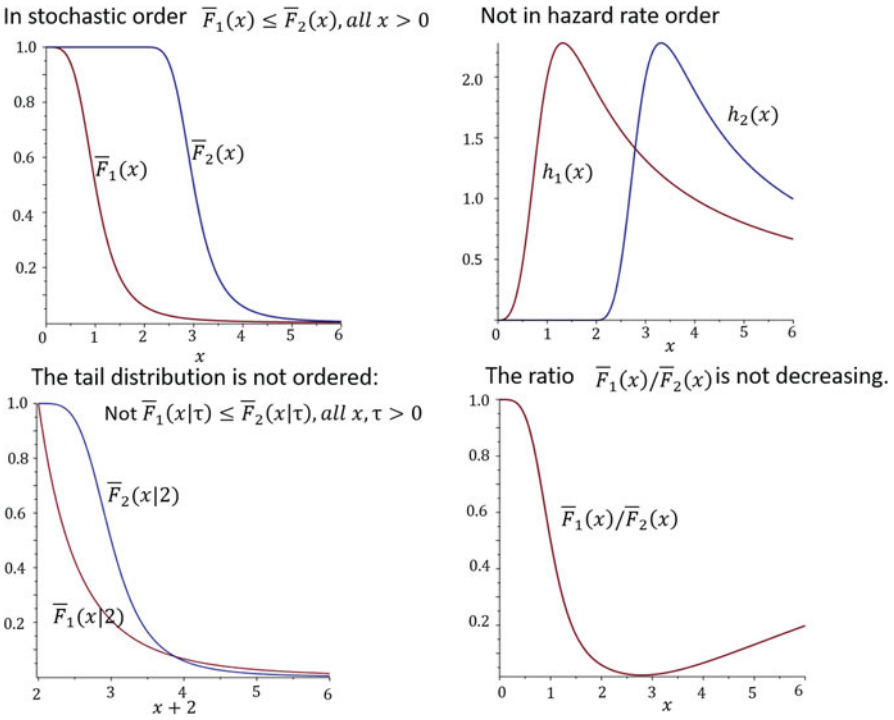


Fig. 2.9 Illustration of Example 4: two lifetimes satisfying stochastic ordering but not hazard rate ordering

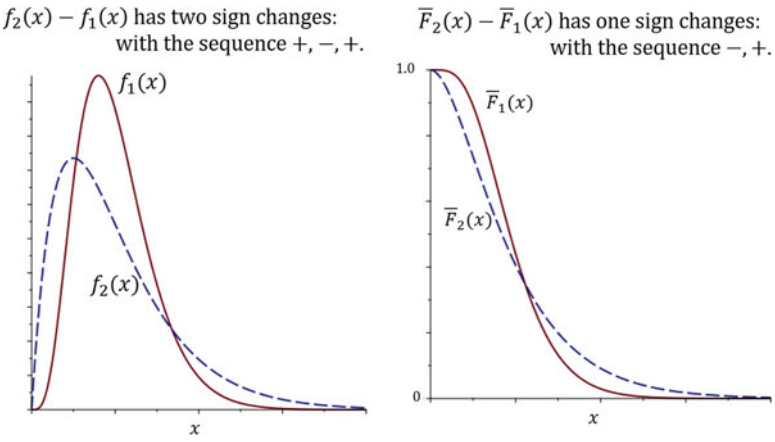


Fig. 2.10 Verbal and graphic presentation for the convex order showing that X_2 is more "spread out" than X_1

Equivalently, in mathematical terms:

Definition 6 $X_1 \leq_{cv} X_2$ if $E[\Psi(X_1)] \leq E[\Psi(X_2)]$ for all convex functions $\Psi(x)$ for which these expectations exist.

The order $X_1 \leq_{cv} X_2$ is called the convex order. The convex order implies the ordering according to variance $var(X) = E[(X-\mu)^2]$ because x^2 is a convex function.

Laplace Transform as an Order to Compare Variability of Lifetimes

The convex order also implies the ordering according to the Laplace transform $E[e^{-sX}]$ for all $s > 0$, because e^{-sx} is a convex function. Therefore, a weaker variability order than the convex order is the Laplace transform order. It has been proposed to compare the variability of two lifetimes (Stoyan 1983).

Let $L[\overline{F}](s) = \int_0^\infty e^{-sx}\overline{F}(x)dx$ be the Laplace transform for the survival function, it can be shown through integration by parts that

$$L[\overline{F}](s) = \frac{1}{s}\left[1 - L[f](s)\right].$$

Definition 7 X_1 is smaller than X_2 in Laplace transform order, denoted as $X_1 \leq_{Lt} X_2$, if the corresponding Laplace transforms $L[f_1](s) \geq L[f_2](s)$, or equivalently, $L[\overline{F}_1](s) \leq L[\overline{F}_2](s)$.

Laplace Transforms of the Gamma and Inverse-Gaussian Distributions

Gamma Distribution with Mean μ and Variance μ^2/κ For the gamma distribution (2.17), by re-parametrization $\mu = \kappa/\lambda$ and $\kappa = \kappa$, the Laplace transforms have explicit expressions

$$\begin{aligned} L[f](s) &= \left(1 + \tfrac{s\mu}{\kappa}\right)^{-\kappa}, \\ L[\overline{F}](s) &= \tfrac{1}{s}\left[1 - \left(1 + \tfrac{s\mu}{\kappa}\right)^{-\kappa}\right]. \end{aligned} \tag{2.33}$$

When $\kappa = 1$, the Gamma distribution reduces to the exponential distribution with $L[f](s) = (1+s\mu)^{-1}$ and $L[\overline{F}](s) = \frac{\mu}{1+s\mu}$.

Another special case is when $\kappa \to \infty$, the resulting distribution is degenerated to a single point μ with $var[X] \to 0$. In this case, $L[f](s) = e^{-s\mu}$ and $L[\overline{F}](s) = \frac{1}{s}\left(1 - e^{-s\mu}\right)$.

When using the Laplace transform order to compare variability, sometimes the Laplace transform of the gamma distribution is used as the benchmark to compare variability of other lifetimes.

Definition 8 A survival function $\overline{F}_X(x)$ with finite mean $\mu = \int_0^\infty \overline{F}_X(x)dx$ is said to belong to the L_κ-class of distributions, if

$$L[\overline{F}](s) \geq \frac{1}{s}\left[1 - \left(1 + \frac{\mu s}{\kappa}\right)^{-\kappa}\right], \text{ for all } s > 0. \tag{2.34}$$

Definition 9 The L_1-class of distributions are called the L-class of distributions, satisfying

$$L[\overline{F}](s) \geq \frac{\mu}{1 + \mu s}, \text{ for all } s > 0. \tag{2.35}$$

All L-class distributions are larger in Laplace transform order compared to the exponential distributions with the same mean value. All gamma distributions with shape parameter $\kappa \geq 1$ belong to the L-class distributions. All L_κ-class distributions are larger in Laplace transform order compared to the gamma distribution with shape parameter κ of the same mean value.

Even for distributions of which the Laplace transform cannot be written explicitly, Laplace transform order comparisons still can be made. These are the cases for the Weibull and the log-normal distributions.

Following Klar (2002), the Weibull distribution with shape parameter $\varsigma \geq 1$ belongs to the L_1-class, which is larger in Laplace transform order compared to the exponential distributions with the same mean value.

Let X_1 follow a log-normal distribution and X_2 follow a gamma distribution. Both have the same mean μ, but with variances $var[X_1] = \mu^2\left(e^{\frac{1}{\varsigma^2}} - 1\right)$ and $var[X_2] = \mu^2/\kappa$, respectively. If the shape parameter ς in the log-normal distribution satisfies $\varsigma^{-2} \leq \log\left(1 + \frac{1}{\kappa}\right)$, from Klar (2002), the log-normal distribution belongs to the L_κ-class. In fact,

$$\varsigma^{-2} \leq \log\left(1 + \frac{1}{\kappa}\right) \Longleftrightarrow var[X_1] \leq var[X_2].$$

The Inverse-Gaussian Distribution with Mean μ and Variance μ^2/κ For the inverse-Gaussian distribution (2.18), by re-parametrization $\mu = \kappa/\lambda$ and $\kappa = \kappa$, the Laplace transforms have explicit expressions

$$\begin{aligned} L[f](s) &= \exp\left\{\kappa - \kappa\sqrt{\tfrac{2\mu_I}{\kappa}s + 1}\right\} \\ L[\overline{F}](s) &= \tfrac{1}{s}\left(1 - \exp\left\{\kappa - \kappa\sqrt{\tfrac{2\mu_I}{\kappa}s + 1}\right\}\right) \end{aligned} \tag{2.36}$$

The inverse-Gaussian distribution belongs to the L_κ-class. That is, given the same mean and variance, the Laplace transform of the inverse-Gaussian distribution $L[f](s)$ is always smaller than that for the gamma distribution with the same μ and κ.

More generally (Klar 2002), if X_1 follows the inverse-Gaussian distribution and X_2 follows the gamma distribution, if κ is the shape parameter of the inverse-Gaussian distribution and κ_G is the shape parameter of the gamma distribution, when $\kappa \geq \kappa_G$,

$$\exp\left\{\kappa - \kappa\sqrt{\frac{2\mu_I}{\kappa}s + 1}\right\} \leq \left(1 + \frac{s\mu_I}{\kappa_G}\right)^{-\kappa_G}, \text{ for all } s > 0.$$

2.6 Mixture of Distributions and Frailty Models

Let us consider a situation that the distribution for random variable X_i associated with individual i is specified by c.d.f. $F_X(x|\theta_i)$, such that X_i's are not identically distributed. In some cases, this heterogeneity can be observed through a vector of covariates $\underline{z}$, say, such as gender, birth date, height, etc. A common practice in statistics is to model θ_i as a function of $\underline{z}$ via a generalized linear model $\eta(\theta_i) = \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_q z_q$, where $\eta(\cdot)$ is a link function such that $-\infty < \eta(\theta_i) < \infty$.

If the heterogeneity is not observable, one assumes that θ_i varies among individuals as independently and identically distributed (i.i.d.) random variables with c.d.f. $U(\theta)$, such that at the population level, one may model X arising from a distribution given by

$$\begin{aligned} F_X(x) &= \textstyle\int_{\theta\in\Theta} F_X(x|\theta)dU(\theta), \\ \overline{F}_X(x) &= \textstyle\int_{\theta\in\Theta} \overline{F}_X(x|\theta)dU(\theta), \\ f_X(x) &= \textstyle\int_{\theta\in\Theta} f_X(x|\theta)dU(\theta), \end{aligned} \tag{2.37}$$

should these integrations exist. Similar presentations are not true for the hazard function.

The distribution $U(\theta)$ is called the mixing distribution. Throughout this book, we use the notation

$$\mathcal{P}(x|\theta) \underset{\theta}{\wedge} U(\theta) = \int_{\theta\in\Theta} \mathcal{P}(x|\theta)dU(\theta) \tag{2.38}$$

for the mixture of distributions, where $\mathcal{P}(x|\theta)$ represents the distribution given the fixed value of θ, which can be expressed either as $F_X(x|\theta)$, $\overline{F}_X(x|\theta)$ or $f_X(x|\theta)$ in (2.37). This notation is widely adopted in the literature, for instance, Johnson et al. (1993), Devroye (1992), Karlis and Xekalaki (2005), among others.

Many of the life distributions we have introduced so far have a mixture representation.

1. The Weibull distribution can be represented as a mixture between a uniform distribution defined over the interval $[0, \theta^{\frac{1}{\beta}}]$ and θ is distributed according to a gamma distribution with shape parameter 2 (Walker and Stephens 1999). That is, $Uniform[0, \theta^{\frac{1}{\varsigma}}] \underset{\theta}{\wedge} U(\theta)$ where the p.d.f. $u(\theta) = U'(\theta) = \lambda^2 \theta e^{-\lambda\theta}$.
2. The gamma distribution with shape parameter $\kappa < 1$ can be represented as a mixture of an exponential distribution (Gleser 1989), $Exp(x|\theta) \underset{\theta}{\wedge} U(\theta)$, where the p.d.f. of the exponential distribution is $\theta e^{-\theta x}$ and the p.d.f. of $U(\theta)$ is

$$u(\theta|\lambda, \kappa) = \begin{cases} \frac{\lambda^\kappa}{\theta(\theta-\lambda)^\kappa \Gamma(1-\kappa)\Gamma(\kappa)}, & \theta \geq \lambda \\ 0, & \text{otherwise.} \end{cases}$$

3. The Laplace transform $L[f](s) = \int_0^\infty e^{-sx} f(x)dx$ with respect to p.d.f. $f(x)$ is a survival function arising as a mixture of the exponential distribution with $f(x)$ as its mixture distribution.

In fact, with the exception of the degenerate distributions (i.e., the c.d.f. takes only the values 0 and 1), all distributions have non-trivial mixture representations, and such representations are not unique. In applications, if there exist one or more natural representations, it is important to recognize them and use them to help understanding of the underlying stochastic mechanisms.

2.6.1 *Frailty and Dampened Hazard Functions*

The proportional hazard model $h(x|\theta) = \theta h_*(x)$ is used to model heterogeneity should individuals follow different hazard functions, where $h_*(x) > 0$ is a baseline hazard function and $\theta > 0$ is the frailty parameter associated with the value θ_i for individual i. It can be alternatively written in terms of survival functions $\overline{F}(x|\theta) = e^{-\theta H_*(t)}$ where $H_*(x) = \int_0^x h_*(u)du$. A commonly used proportional hazard model is the log-linear model so that

$$\log \theta = \beta_1 z_1 + \cdots + \beta_m z_m$$

provided that individual heterogeneity is observable through a vector of covariates $(z_1, \ldots, z_m)$.

The frailty model is a random effect model for unobservable heterogeneity. It assumes that θ is random with mean value $E(\theta) = 1$ and probability density function (p.d.f.) $u(\theta)$. If there is no heterogeneity, then $u(\theta)$ degenerates to a point $\theta \equiv 1$ with no variation and the survival function is $\overline{F}_*(x) = e^{-H_*(x)}$.

When the population is composed of a mixture of heterogeneous individuals, the survival function arises from a mixed distribution:

$$\overline{F}^{(mixed)}(x) = \int_0^\infty \overline{F}(x|\theta)u(\theta)d\theta = \overline{F}(x|\theta) \underset{\theta}{\wedge} U(\theta).$$

In this case, $\overline{F}(x|\theta) = e^{-\theta H_*(t)}$ and

$$\overline{F}^{(mixed)}(x) = \int_0^\infty e^{-\theta H_*(x)}u(\theta)d\theta = L[u](H_*(x)), \tag{2.39}$$

where $L[u](s) = \int_0^\infty e^{-s\theta}u(\theta)d\theta$ is the Laplace transform with respect to $u(\theta)$ and $L[u](H_0(x))$ is $L[u](s)$ evaluated at $s = H_*(x)$.

The hazard function is

$$h^{(mixed)}(x) = -\frac{d}{dx}\log \overline{F}^{(mixed)}(x) = -\frac{d}{dx}\log L[u](H_*(x)). \tag{2.40}$$

The importance of the Laplace transform in the context of frailty modelling was pointed out in Hougaard (1984).

The cumulative hazard function $H_*(x)$ is a monotonically increasing function from zero to infinity and $s = H_*(x)$ can be regarded as a transformed time. As a result, $\overline{F}^{(mixed)}(x)$ in (2.39) is log-convex with respect to transformed time $H_*(x)$ and $h^{(mixed)}(x)$ is a decreasing function with respect to transformed time $H_*(x)$. This leads to the phenomenon of the dampened hazard function, as previously shown in Fig. 2.6, due to the random effect of the frailty model.

Frailty Models with Gamma Distributed $u(\theta)$

We choose the Gamma distribution for $u(\theta)$ with $E[\theta] = 1, var[\theta] = \kappa^{-1} = v$. The Laplace transform in (2.33), by letting $\mu = 1$ and $\kappa^{-1} = v$, becomes

$$L[u](s) = (1 + vs)^{-1/v} = \frac{\kappa^\kappa}{(\kappa + s)^\kappa} \tag{2.41}$$

and $-\frac{d}{ds}\log L[u](s) = \frac{1}{1+vs} = \frac{\kappa}{\kappa+s}$.

From (2.39) and (2.40),

$$\begin{aligned} \overline{F}^{(mixed)}(x) &= (1 + vH_*(x))^{-1/v} = \tfrac{\kappa^\kappa}{(\kappa+H_*(x))^\kappa}, \\ h^{(mixed)}(x) &= \tfrac{h_*(x)}{1+vH_*(x)} = \tfrac{\kappa}{\kappa+H_*(x)}h_*(x). \end{aligned} \tag{2.42}$$

In the second expression, $h^{(mixed)}(x)$ is the baseline hazard function $h_*(x)$ dampened by a factor $(1 + vH_*(x))^{-1}$.

The Frailty Model for the Weibull Distribution with Gamma Mixture The baseline hazard function of the Weibull distribution is $h_*(x;\lambda,\varsigma)=\lambda\varsigma\,(\lambda x)^{\varsigma-1}$. The cumulative hazard function is $H_*(x)=(\lambda x)^{\varsigma}$. With the gamma mixture $u(\theta)$ with $E[\theta]=1$, $var[\theta]=\kappa^{-1}=v$, (2.42) becomes

$$\overline{F}^{(mixed)}(x)=\left(1+v\,(\lambda x)^{\varsigma}\right)^{-1/v}=\frac{\kappa^{\kappa}}{(\kappa+(\lambda x)^{\varsigma})^{\kappa}},$$

$$h^{(mixed)}(x)=\frac{\lambda\varsigma\,(\lambda x)^{\varsigma-1}}{1+v\,(\lambda x)^{\varsigma}}=\frac{\kappa\lambda\varsigma\,(\lambda x)^{\varsigma-1}}{\kappa+(\lambda x)^{\varsigma}}.$$

This distribution is the Pareto IV distribution (Dubey 1968, 1969).

When $var[\theta]=1$, the mixture distribution is exponential and the baseline distribution is Weibull,

$$\overline{F}^{(mixed)}(x)=\left(1+(\lambda x)^{\varsigma}\right)^{-1}\text{ and }h^{(mixed)}(x)=\frac{\lambda\varsigma\,(\lambda x)^{\varsigma-1}}{1+(\lambda x)^{\varsigma}}$$

which returns to the log-logistic distribution (2.24), also known as the Pareto-III distribution.

On the other hand, if the baseline distribution is exponential and the mixture distribution is gamma, then

$$\overline{F}^{(mixed)}(x)=(1+v\,(\lambda x))^{-1/v}=\frac{\kappa^{\kappa}}{(\kappa+(\lambda x))^{\kappa}},$$

$$h^{(mixed)}(x)=\frac{\lambda}{1+v\,(\lambda x)}=\frac{\kappa\lambda}{\kappa+(\lambda x)}. \tag{2.43}$$

These are Pareto-I distribution (when $\kappa=1$) and Pareto-II distribution.

Figure 2.11 illustrates $h^{(mixed)}(x)=\frac{\kappa\lambda\varsigma(\lambda x)^{\varsigma-1}}{\kappa+(\lambda x)^{\varsigma}}$ and $\overline{F}^{(mixed)}(x)=\frac{\kappa^{\kappa}}{(\kappa+(\lambda x)^{\varsigma})^{\kappa}}$ at $\varsigma=2$. In this case the baseline hazard $h_*(x;\lambda,\varsigma)=2\lambda^2x$ is a linear function of x. The time scale for these plots is standardized according to the scale parameter λ. The limiting case $v=\kappa^{-1}\to 0$ returns to the baseline distribution.

In the special case of a constant baseline hazard rate λ, $\overline{F}^{(mixed)}(x)$ is log-convex with respect to x while λ is a scale parameter. In this case, $h^{(mixed)}(x)$ is monotonically decreasing starting from $h^{(mixed)}(0)=\lambda$.

Proposition 10 (Marshall and Olkin (2007)) *If $\overline{F}_X(x|\theta)=e^{-\lambda\theta x}$, then $\overline{F}^{(mixed)}(x)=\int_{\theta\in\Theta}\overline{F}_X(x|\theta)dU(\theta)$ is log-convex and the corresponding hazard function $h_X(x)=-\frac{d}{dx}\log\overline{F}(x)$ is decreasing.*

In fact, the stronger stated is that, for any non-trivial mixture distribution with p.d.f. $u(\theta)$ defined on $(0,\infty)$ with $E[\theta]=1$, $\overline{F}^{(mixed)}(x)=\int_0^{\infty}e^{-\lambda\theta x}u(\theta)d\theta=$

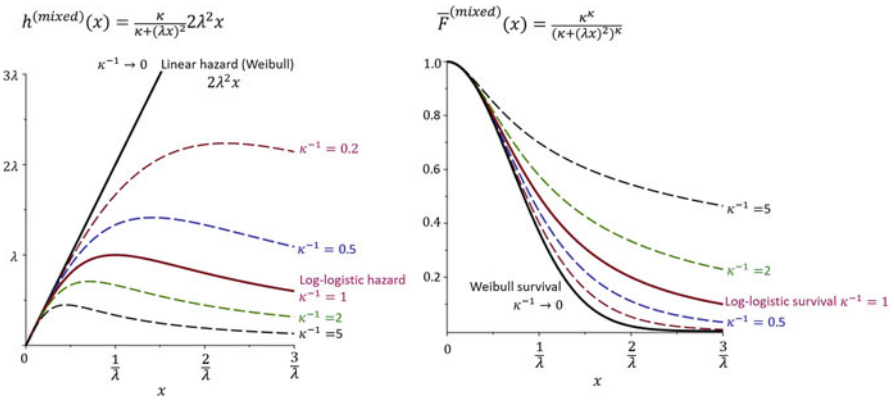


Fig. 2.11 Illustration of the mixture distribution of the Weibull distribution with gamma distributed mixture through the frailty model

$L[u](\lambda x)$ and the Laplace transform is completely monotonic. Since the corresponding hazard function can be expressed as

$$h^{(mixed)}(x) = -\frac{\frac{d}{dx} L[u](\lambda x)}{L[u](\lambda x)}.$$

One can easily verify that $h^{(mixed)}(x)$ is completely monotonic. Hesselager et al. (1998) have put forward the following theorem: *"A distribution with a completely monotonic hazard function is a mixed exponential distribution."*

2.7 Problems and Supplements

2.1 Define the following piece-wise hazard function for the lifetime X with a bathtub shape

$$h_X(x) = \begin{cases} \frac{1+a}{1+x}\mu, & 0 \le x \le a \\ \mu, & a < x \le b, \ \mu > 0, \ \theta > 0; \ 0 < a < b < \infty. \\ \mu e^{\theta(x-b)}, & b < x < \infty \end{cases}$$

(a) Use (2.5) to write down the survival function $\overline{F}(x)$.
(b) Let $a = 2$, $b = 50$, $\mu = 0.002$, and $\theta = 0.09$, plot $h_X(x)$ and $\overline{F}_X(x)$ for $0 \le x \le 100$.
(c) What is the meaning of $\int_0^\infty \overline{F}_X(x)dx$? Calculate $\int_0^\infty \overline{F}_X(x)dx$ given the above parameters values.

2.2 The *residual life* distribution for X is defined through a conditional survival function

$$\overline{F}_X(x|\tau) = \Pr\{X > \tau + x | X > \tau\} = \frac{\overline{F}_X(\tau + x)}{\overline{F}_X(\tau)},\ x \geq 0.$$

(a) Show that $\overline{F}_X(x|\tau)$ is uniquely determined by the hazard function $h_X(x)$, as $\overline{F}_X(x|\tau) = \exp\left(-\int_\tau^{\tau+x} h_X(u)du\right)$.

(b) Show that, given $X > \tau$, the residual life has p.d.f. and hazard function given by $f_X(x|\tau) = \frac{f_X(\tau+x)}{\overline{F}_X(\tau)}$ and $h_X(x|\tau) = h_X(\tau + x)$.

(c) The mean residual is defined as $m(\tau) = E[X - \tau | X > \tau]$. Show that if $m(\tau)$ exists

$$m(\tau) = \int_0^\infty \overline{F}_X(x|\tau)dx = \frac{\int_\tau^\infty \overline{F}_X(x)dx}{\overline{F}_X(\tau)}.$$

(d) Using the hazard function in 2.1 and the parameters values in 2.1(b), calculate $m(\tau)$ at $\tau = 25$, 45 and 65.

(e) For two lifetime variables X_1 and X_2, with hazard functions $h_1(x)$ and $h_2(x)$, survival functions $\overline{F}_1(x)$ and $\overline{F}_2(x)$, *residual life survival functions* $\overline{F}_1(x|\tau)$ and $\overline{F}_2(x|\tau)$, respectively, show that the following statements are equivalent:

(i) $h_1(x) \geq h_2(x)$, for all $x > 0$;
(ii) $\overline{F}_1(x|\tau) \leq \overline{F}_2(x|\tau)$ for all x and τ;
(iii) $\overline{F}_1(x)/\overline{F}_2(x)$ is a decreasing function of x.

2.3 If two non-negative random variables X_1 and X_2 are independent, with p.d.f.'s $f_1(x)$ and $f_2(x)$, respectively, show that

(a) the p.d.f. of the sum $X_1 + X_2$ is the convolution $g(x) = \int_0^\infty f_1(u) f_2(x - u)du$;

(b) the Laplace transform of $g(x)$ is the multiplication of two separate Laplace transforms:

$$L[g](s) = L[f_1](s)L[f_2](s) = \left(\int_0^\infty e^{-sx} f_1(x)dx\right)\left(\int_0^\infty e^{-sx} f_2(x)dx\right).$$

2.4 Consider a disease that has a latent period T_E with hazard function $h_E(x)$, where x is measured from the time of infection. An infected individual is not infectious during this period. By the end of the latent period, the infected individual starts to be infectious. Meanwhile, a public health control measure starts immediately to isolate infected individuals. Successfully isolated individuals do not pose a risk of transmitting the disease. The time to isolation is denoted by T_c with hazard function $h_c(x)$. We assume that T_c and T_I are independent.

(a) Show that the probability that an infected individual progresses to become infectious is

$$\Pr(T_c > T_E) = \int_0^\infty h_E(x) e^{-\int_0^\infty [h_c(x)+h_E(x)]dx} dx.$$

(b) If the isolation rate is constant, $h_c(x) = \phi > 0$ and the p.d.f. of T_E is $f_E(x)$, show that $\Pr(T_c > T_E) = L[f_E](\phi)$, where $L[f](s)$ denotes the Laplace transform for the p.d.f. $f(x)$. Write down the expression for $\Pr(T_c > T_E)$ when T_E is exponentially distributed with rate $\alpha > 0$.

(c) Given the constant isolation rate ϕ, we compare the latent periods with the same average value μ_E. Is it true that the more variable the latent period, the more effective the isolation as a public health intervention?

2.5 For individuals who are infectious, we assume there is a natural infectious period T_I. By the end of the infectious period, the infected individual recovers and is no longer infectious. Let $h_I(x)$, $f_I(x)$, and $\overline{F}_I(x)$ denote the hazard function, the p.d.f., and the survival function, respectively; and x is measured from the start of infectiousness.

(a) Show that the mean infectious period $\mu_I = \int_0^\infty x f_I(x) dx$ can be expressed as $\mu_I = \int_0^\infty \overline{F}_I(x) dx$.

(b) The isolation measure also effectively reduces disease transmission by shortening the infectious period. Let T_c be the time from the start of infectiousness to isolation, with hazard function $h_c(x)$. Let $T = \min(T_c, T_I)$ be the effective infectious period. Show that T is smaller than T_I in stochastic order.

(c) If the isolation rate is constant, $h_c(x) = \phi$, show that the effective mean infectious period is

$$\overline{\mu}_I = L[\overline{F}_I](\phi)$$

where $L[\overline{F}](s)$ is the Laplace transform with respect to the survival function $\overline{F}_I(x)$.

(d) Show that $\phi L[\overline{F}_I](\phi) = 1 - L[f_I](\phi)$. Both $L[f_I](\phi)$ and $\phi L[\overline{F}_I](\phi)$ are meaningful probabilities. What probabilities do they represent?

(e) Given the constant isolation rate ϕ, we compare the infectious periods with the same average value μ_I. Is it true that the more variable the infectious period, the more effective the isolation as a public health intervention?

2.6 Assuming a random sample of n individuals with lifetimes $X_1, \cdots, X_n$, independently and identically distributed lifetimes with hazard function $h(x)$, p.d.f. $f(x)$, and survival function $\overline{F}(x)$. The order statistics is given by $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$.

(a) Show that the p.d.f. of $X_{(1)}$ is $nf(x)\overline{F}(x)^{n-1}$ and the hazard function of $X_{(1)}$ is $nh(x)$.

(b) We assume that at calendar time t, there are $C(t) = n$ infected individuals in the (infinitely large) susceptible population. All individuals are identical and each individual has a constant rate r to infect a susceptible individual, independently from other infected individuals. What is the instantaneous growth rate of $C(t)$ at time t, expressed as $\lim_{h\to 0} \Pr\{C(t+h) = n+1 | C(t) = n\} = ?$

(c) Now we assume that infected individuals are different. Infected individual i carries an intrinsic rate $z_i r$ to produce a new infection, where r is a baseline rate and z_i is a frailty variable. Individual heterogeneity is unobservable. We assume z_i as i.i.d. random variables with mean value $E(z) = 1$ and p.d.f. $\xi(z)$. At the beginning of the epidemic, we initially seed i_0 infected individuals and assume that at time t, the susceptible population is still infinitely large. Show that the instantaneous growth rate of $C(t)$ at time t is a function of time $\lim_{h\to 0} \Pr\{C(t+h) = n+1 | C(t) = n\} = \rho(t) = -\frac{d}{dt} \log L[\xi](rt)$, where $L[\xi](s)$ is the Laplace transform of $\xi(z)$.

(d) Let $\xi(z)$ be the p.d.f. of the Gamma distribution with $E(z) = 1$ and variance $var[z] = v > 0$, show that

$$\rho(t) = r/(1 + rvt), v > 0.$$

Viewing $\rho(t)$ as a hazard function of a lifetime distribution, which distribution does it correspond to?

Chapter 3
Random Counts and Counting Processes

We now turn our attention to the population level dynamics and ask phenomenological questions. First, many important measures in the study of infectious diseases are count variables N, taking integer values $n = 0, 1, 2, \ldots$. For example:

1. the number of infectious contacts, defined as contacts at which a transmission of infection takes place (Dietz 1995), made by an infected individual throughout its entire infectious period;
2. the number of new infections in a population in a given time period;
3. the final size of an epidemic, defined as the cumulative number of infections in a population when no more infected or susceptible individuals are left;
4. the number of infectious individuals in a specific epidemiological stage of the disease at a given time (prevalence);
5. the count of generations (as discrete time), of which, the initial infectious case introduced into a completely susceptible population occurs at generation zero; those directly infected by this case make the first generation of cases, and so on;
6. the number of connections in a social network.

We consider an infectious disease epidemic in a population as a realization of a stochastic process and these count variables are outcomes of such a process. We call them random counts. We start this chapter by reviewing some important distributions for random counts. We relate these discrete distributions with continuous time stochastic processes, especially counting processes, to reveal different stochastic mechanisms that manifest these random counts.

P. Yan, G. Chowell, *Quantitative Methods for Investigating Infectious Disease Outbreaks*, Texts in Applied Mathematics 70,
https://doi.org/10.1007/978-3-030-21923-9_3

3.1 Some Important Distributions For Random Counts

3.1.1 The Probability Functions and Related Quantities

The distribution for the random count N is the probability mass function (p.m.f.)

$$f_n = \Pr\{N = n\},\ n = 0, 1, 2, \cdots,$$

satisfying (i) $f_n > 0$; (ii) $\sum_{N=0}^{\infty} f_n = 1$. The survivor function is defined as

$$\overline{F}_n = \Pr\{N \geq n\} = \sum_{j=n}^{\infty} f_j$$

so that $f_n = \overline{F}_n - \overline{F}_{n+1}$. The hazard function for N is defined as

$$h_n = \frac{\Pr(N = n)}{\Pr(N \geq n)},\ n = 0, 1, 2, \cdots.$$

The discrete hazard function is bounded by $0 \leq h_n \leq 1$, which is unlike the continuous distributions where the hazard function satisfies $0 < h(x) < \infty$. It can be shown that $1 - h_n = \frac{\overline{F}_{n+1}}{\overline{F}_n}$, $n = 0, 1, 2, \cdots$, so that

$$\overline{F}_n = \prod_{j=0}^{n-1}(1 - h_j) \text{ and } f_n = h_n \prod_{j=0}^{n-1}(1 - h_j),\ n = 0, 1, 2, \cdots. \tag{3.1}$$

Analogous to the *residual life* distribution in Chap. 2, we consider the residual count distribution by the condition probability, and it can be determined by the tail part of the hazard function h_n.

$$\overline{F}_{x|n} = \frac{\overline{F}_{n+x}}{\overline{F}_n} = \frac{\Pr(N \geq n + x)}{\Pr(N \geq n)} = \prod_{j=n}^{n+x-1}(1 - h_j).$$

The rth moment of N is defined as $E[N^r] = \sum_{n=0}^{\infty} n^r f_n$. It can be further shown that the mean can be written as $E[N] = \sum_{n=1}^{\infty} \overline{F}_n$ and the mean residual time can be written as

$$E(N - n|N \geq n) = \sum_{x=1}^{\infty} \overline{F}_{x|n} = \sum_{x=1}^{\infty} \prod_{j=n}^{n+x-1}(1 - h_j) = \sum_{k \geq n} \prod_{j=n}^{k}(1 - h_j). \tag{3.2}$$

The tail property is usually studied as the limits of the hazard function and the mean residual life as $n \to \infty$. It can be immediately shown that

1. if $\lim_{n\to\infty} h_n = 1$, then $\lim_{n\to\infty} E(N-n|N \geq n) = 0$.
2. if $\lim_{n\to\infty} h_n = c < 1$, then $\lim_{n\to\infty} E(N-n|N \geq n) = (1-c)/c$.
3. if $\lim_{n\to\infty} h_n = 0$, then $\lim_{n\to\infty} E(N-n|N \geq n) = \infty$.

It is convenient to study the distribution of random counts through the probability generating function (p.g.f.). Under the assumptions that N has finite mean and finite variance, the p.g.f. is defined by

$$G_N(s) = E\left(s^N\right) = \sum_{n=0}^{\infty} s^n f_n. \tag{3.3}$$

This power series converges absolutely for all $|s| \leq 1$. If the distribution of N is not degenerated to a point mass, that is, there is no $n \geq 0$ such that $\Pr(N = n) = 1$, then the p.g.f. is strictly increasing for $s \in [0, 1]$ and is strictly convex. It satisfies

$$G_N(0) = \Pr\{N = 0\},\ G_N(1) = 1,\ G'_N(s) > 0,\ G''_N(s) > 0. \tag{3.4}$$

If the p.m.f. f_n is given, $G_N(s)$ is uniquely defined through (3.3). The probability generating function is a tool to generate probabilities. Provided that $G_N(s)$ is a smooth function of s with higher order of derivatives, then

$$f_n = \frac{1}{n!} G_N^{(n)}(0),\ n = 0, 1, 2, \cdots, \tag{3.5}$$

where $G_N^{(n)}(0) = \frac{d^n}{ds^n} G_N(s)\Big|_{s=0}$, so that the p.m.f. f_n can be uniquely generated through $G_N(s)$. The probability generating function can also be used to generate moments. The mean and variance of N are

$$E[N] = G'_N(1),\quad var[N] = G''_N(1) + G'_N(1) - \left(G'_N(1)\right)^2. \tag{3.6}$$

In general, the mth factorial moment is $E[N(N-1)\cdots(N-m+1)] = \frac{d^m}{ds^m} G_N(s)\Big|_{s=1}$.

3.1.2 Two Classes of Distributions

The Power Series Distributions

A very broad class of count distributions is defined by the power series

$$f_n = \Pr\{N = n\} = \frac{a_n \theta^n}{A(\theta)},\quad n = 0, 1, 2, \ldots;\ \theta > 0, \tag{3.7}$$

where $A(\theta)$ is the normalization factor $A(\theta) = \sum_{j=0}^{\infty} a_j \theta^j < \infty$ and θ is the canonical parameter. It has the recursive formulae $\frac{f_{n+1}}{f_n} = \frac{a_{n+1}}{a_n}\theta$. The hazard function has the form

$$h_n = \frac{a_n \theta^n}{\sum_{j=n}^{\infty} a_j \theta^j}, \quad n = 0, 1, 2, \ldots; \ \theta > 0.$$

For general properties of this distribution, we refer to Chap. 2 of Johnson et al. (1993). This distribution has been applied in the study of the transmission of infectious diseases, such as Farrington and Grant (1999), Farrington et al. (2003), among others.

The p.g.f. corresponding to (3.7) is

$$G_N(s) = \frac{A(s\theta)}{A(\theta)}. \tag{3.8}$$

Consequently, the mean and variance, according to (3.6) can be expressed by

$$E[N] = \theta \frac{A'(\theta)}{A(\theta)} = \theta \frac{d}{d\theta} \log A(\theta),$$

$$var[N] = E[N] + \theta^2 \frac{d^2}{d\theta^2} \log A(\theta).$$

One of the very useful properties of the power series distribution is that, if $(N_1, N_2, \ldots, N_m)$ are independent and identically distributed from the same power series distribution, then $T = \sum_{i=1}^{m} N_i$ is the minimal sufficient statistic for θ.

We introduce some count distributions under this class: the Poisson distribution with $A(\theta) = e^{\theta}$, $\theta > 0$; the negative binomial distribution (including the geometric distribution as a special case) with $A(\theta) = (1-\theta)^{-\kappa}$, $\kappa > 0$ and $0 < \theta < 1$; and the logarithmic distribution with $A(\theta) = -\log(1-\theta)$, $0 < \theta < 1$.

1. The Poisson Distribution A power series distribution with $A(\theta) = e^{\theta}$, $\theta > 0$ leads to $a_n = \frac{1}{n!}$ and

$$f_n = \frac{1}{n!}\theta^n e^{-\theta}, \ n = 0, 1, 2, \ldots \tag{3.9}$$

with recursive formulae $\frac{f_{n+1}}{f_n} = \frac{\theta}{n+1}$, $n = 0, 1, 2, \ldots$. The Poisson distribution is determined by a single rate parameter θ to model the number of events occurring in a time interval or space area so that the probability of zero event occurring in a unit of time or space is $\Pr(N = 0) = e^{-\theta}$.

Remark In infectious disease models, it corresponds to "homogeneous mixing" in a very large population so that each individual has equal chance to make contacts with everyone else at rate λ and the probability of transmission per contact between a pair of susceptible–infectious individuals is also a constant $p \in (0, 1]$. In such

case, the number of infectious contacts during a time interval $(t, t+x]$ follows a Poisson distribution with $\theta = \lambda p x$.

The p.g.f. of the Poisson distribution is $G_N(s) = e^{-\theta(1-s)}$. The mean and variance are $E[N] = var[N] = \theta$. Kosambi (1949) showed that among power series distributions, $E[N] = var[N]$ characterizes the Poisson distribution. The hazard function is monotonically increasing and given by

$$h_n = \frac{1}{n!}\frac{\theta^n}{\sum_{j=n}^{\infty}\frac{1}{j!}\theta^j} = \left(1 + \frac{\theta}{n+1} + \frac{\theta^2}{(n+2)(n+1)} + \cdots\right)^{-1}.$$

Hence $\lim_{n\to\infty} h_n = 1$ and $\lim_{n\to\infty} E(N-n|N \geq n) = 0$.

2. The Geometric Distribution The geometric distribution is a power series distribution with $A(\theta) = (1-\theta)^{-1}$, $0 < \theta < 1$ with $a_n = 1$, $n = 0, 1, 2, \ldots$. Its p.m.f. and p.g.f. are

$$f_n = \theta^n(1-\theta) \text{ and } G_N(s) = \frac{1-\theta}{1-s\theta},\ n = 0, 1, 2, \ldots. \tag{3.10}$$

The mean and variance are

$$E[N] = \frac{\theta}{1-\theta}$$
$$var[N] = \frac{\theta}{(1-\theta)^2} = E[N] + E[N]^2.$$

It has a constant hazard function $h_n = 1-\theta$, $n = 0, 1, 2, \ldots$. and the mean residual life

$$E(N-n|N \geq n) = \sum_{k=1}^{\infty}\theta^k = \frac{\theta}{1-\theta} = E[N],$$

which is the memoryless property. It is the discrete analogue to the exponential distribution for the continuous lifetime and sometimes it is re-parameterized by $1-\theta = e^{-\tau}$ with p.m.f expressed by

$$f_n = \left(1-e^{-\tau}\right)^n e^{-\tau}, n = 0, 1, 2, \ldots. \tag{3.11}$$

Remark In many SIS, SIR, SEIR, SEIRS models, the number of contacts per unit of time follows a Poisson distribution, and the infectious period is exponentially distributed. In such cases, the total number of secondary infections produced by an initial infectious individual "seeded" into an infinitely large susceptible population during his/her entire infectious period, denoted by N, follows a geometric distribution.

3. The Negative Binomial Distribution The negative binomial distribution generalizes the geometric distribution as a power series distribution with $A(\theta) = (1-\theta)^{-\kappa}$, $\kappa > 0$ and $0 < \theta < 1$. It leads to $a_n = \binom{n+\kappa-1}{\kappa-1} = \frac{\Gamma(n+\kappa)}{\Gamma(\kappa)\Gamma(n+1)}$, $n = 0, 1, 2, \ldots$ and

$$f_n = \frac{\Gamma(n+\kappa)}{\Gamma(\kappa)\,\Gamma(n+1)}\theta^n(1-\theta)^\kappa,\ n = 0, 1, 2, \ldots \tag{3.12}$$

with recursive formulae $\frac{f_{n+1}}{f_n} = \frac{n+\kappa}{n+1}\theta$, $n = 0, 1, 2, \ldots$. The p.g.f. is

$$G_N(s) = \left(\frac{1-\theta}{1-s\theta}\right)^\kappa,\ n = 0, 1, 2, \ldots.$$

The hazard function is

$$\begin{aligned} h_n &= \left(\sum_{m=0}^{\infty}\frac{\Gamma(m+n+\kappa)\,\Gamma(n+1)}{\Gamma(n+\kappa)\,\Gamma(m+n+1)}\theta^m\right)^{-1} \\ &= \left(1+\frac{n+\kappa}{n+1}\theta+\frac{(n+\kappa+1)(n+\kappa)}{(n+2)(n+1)}\theta^2+\cdots\right)^{-1}. \end{aligned}$$

For finite $\kappa < \infty$ and $0 < \theta < 1$, $0 < \lim_{n\to\infty} h_n = 1-\theta < 1$. The limiting case of the mean residual life is

$$\lim_{n\to\infty} E(N-n|N \geq n) = \sum_{k=1}^{\infty}\theta^k = \frac{\theta}{1-\theta}.$$

It is regarded as the discrete analogue to the gamma distribution for the continuous lifetime.

The mean and variance of the negative binomial distribution are

$$\begin{aligned} E[N] &= \theta\frac{d}{d\theta}\log A(\theta) = \frac{\theta\kappa}{1-\theta} \\ var[N] &= \frac{\kappa\theta}{(1-\theta)^2} = E[N]+\frac{1}{\kappa}E[N]^2. \end{aligned}$$

Meanwhile, $\Pr(N=0) = (1-\theta)^\kappa$. When κ becomes small, it increases both the probability of observing $\{N=0\}$ and the variance, so that $\Pr(N=0) \to 1$ and $var[N] \to \infty$ as $\kappa \to 0$. By removing the zero observations, the conditional probability $\Pr(N=n|N>0)$ gives the zero-truncated negative binomial distribution

$$\Pr(N=n|N>0) = \frac{\Gamma(n+\kappa)}{\Gamma(\kappa)\,\Gamma(n+1)}\frac{\theta^n(1-\theta)^\kappa}{1-(1-\theta)^\kappa}. \tag{3.13}$$

4. The Logarithmic Distribution The power series distribution with respect to $A(\theta) = -\log(1-\theta)$, $0 < \theta < 1$ gives $a_n = n^{-1}$ for $n = 1, 2, \ldots$ and

$$f_n = \frac{n^{-1}\theta^n}{-\log(1-\theta)}, \quad n = 1, 2, \ldots.$$

It is the limiting case of (3.13) as $\lim_{\kappa \to 0} \frac{\Gamma(n+\kappa)}{n!} = n^{-1}$ and $\lim_{\kappa \to 0} \frac{1}{\Gamma(\kappa)} \frac{(1-\theta)^\kappa}{1-(1-\theta)^\kappa} = -\frac{1}{\log(1-\theta)}$. The recursive formulae is $\frac{f_{n+1}}{f_n} = \frac{n}{n+1}\theta$.

Since $\theta < 1$, f_n decreases as n increases. The p.g.f. is $G_N(s) = \frac{\log(1-s\theta)}{\log(1-\theta)}$. The mean and variance are

$$E[N] = \frac{-1}{\log(1-\theta)} \frac{\theta}{1-\theta}, \; var[N] = -\frac{\theta\,(\theta + \log(1-\theta))}{(1-\theta)^2 \log^2(1-\theta)}.$$

This distribution is only defined at integer values $n = 1, 2, \ldots$. The hazard function is

$$h_n = \frac{\theta^n/n}{\sum_{j=n}^{\infty} \theta^j/j} = \left(1 + \frac{n}{n+1}\theta + \frac{n}{n+2}\theta^2 + \frac{n}{n+3}\theta^3 + \cdots\right)^{-1}$$

satisfying $0 < \lim_{n \to \infty} h_n = 1 - \theta < 1$. Like the negative binomial distribution, the limiting case of the mean residual life is

$$\lim_{n \to \infty} E(N - n | N \geq n) = \sum_{k=1}^{\infty} \theta^k = \frac{\theta}{1-\theta}.$$

The Power-Law Distributions

A different class of count distributions is characterized by very long tails that are very distinct compared to the distributions discussed under the power series class, such as the Poisson, geometric, negative binomial and logarithmic distributions. A power-law distribution is any distribution, continuous or discrete, such that $\lim_{x \to \infty} \frac{\overline{F}(x)}{x^{\theta+1}} = A$ for some $\theta > 0, A > 0$. For certain parameter range of θ, the tail can be so long such that the moments may not exist. This type of distribution has gained much attention in infectious disease modeling.

Remark Let the random count N be the number of secondary infections produced by an initial infectious individual "seeded" into an infinitely large susceptible population during its entire infectious period. If the distribution of N has a very long tail following the power-law property, it is possible that its mean $E[N]$ is infinite. In this case, one of the key parameters in disease transmission models, $R_0 = E[N]$, is undefined.

1. Zipf Distribution The distribution with p.m.f. given by (3.14) was empirically developed in the study of linguistics by Zipf (1949) and has been historically referred to as the Zipf distribution.

$$f_n = \Pr\{N = n\} = \frac{n^{-(\theta+1)}}{\zeta(\theta+1)}, \; n = 1, 2, 3, \ldots, \tag{3.14}$$

where $\zeta(\theta) = \sum_{n=1}^{\infty} \frac{1}{n^{\theta}}$ is the Riemann zeta function, which is finite when $\theta > 0$. The hazard function is

$$h_n = \frac{\Pr\{N = n\}}{\Pr\{N \geq n\}} = \frac{n^{-(\theta+1)}}{\sum_{k \geq n} k^{-(\theta+1)}} = \left(\sum_{k=0}^{\infty} \left(\frac{n}{n+k} \right)^{\theta+1} \right)^{-1} \to 0, \text{ as } n \to \infty.$$

Therefore, $\lim_{n\to\infty} E(N - n | N \geq n) = \infty$. The rth moment can be calculated by

$$E[N^r] = \frac{\zeta(\theta - r + 1)}{\zeta(\theta + 1)} < \infty, \; r < \theta.$$

If $r \geq \theta$, the rth moment is infinite. Hence, if $\theta \leq 1$, the mean $E[N]$ is infinite; if $\theta \leq 2$, the variance is infinite. The Pareto-II distribution with p.d.f. $f(x) = \theta (x+1)^{-(\theta+1)}$ can be considered the continuous analogue to the distribution given by (3.14). Conversely, (3.14) has sometimes been referred to as the discrete time Pareto distribution.

Remark Liljeros et al. (2003) used this distribution to model the web of sexual contacts in the study of sexually transmitted infections. In the literature for the study of complex networks, the term "scale-free" was coined by physicist Barabási and Albert (1999) referring to networks with degree distributions given by (3.14) or tails that converge to that of (3.14).

2. Other Power-Law Distributions The Zipf distribution (3.14) has so far remained as an empirical model with an explicit power-law expression without a direct theoretical interpretation. The importance of the power-law is to model the tail of a highly skewed distribution rather than for the entire distribution; the Zipf distribution may yield a poor fit to real data, as shown by Stumpf et al. (2005). The following alternative distributions with the power-law tail property have direct theoretical interpretations and can be closely related and derived from stochastic mechanisms in disease transmission dynamics.

Simon (1955) describes a class of the distribution of the form $f_n = AB(n, \theta+1)$ having the power-law property, where A and θ are constants and $B(n, \theta+1)$ is the Beta function

$$B(n, \theta+1) = \frac{\Gamma(n)\Gamma(\theta+1)}{\Gamma(n+\theta+1)}, \; n = 1, 2, \cdots.$$

By the well-known property of the Gamma function, for any constant k, $\frac{\Gamma(n)}{\Gamma(n+k)} \sim n^{-k}$ as $n \to \infty$. Therefore, $f_n = AB(n, \theta+1) \sim \Gamma(\theta+1)\, n^{-(\theta+1)}$. Letting $A = \theta$, this distribution is the Yule distribution with p.m.f.

$$f_n = \theta \frac{\Gamma(n+1)\,\Gamma(\theta+1)}{\Gamma(n+\theta+2)},\ n = 0, 1, 2, \cdots . \tag{3.15}$$

The mean is $E[N] = \frac{1}{\theta-1}$ which exists only if $\theta > 1$. The variance is $var[N] = \frac{\theta^2}{(\theta-2)(\theta-1)^2}$ which only exists if $\theta > 2$. These are the same property as that for (3.14). Unlike (3.14), in which the hazard function cannot be written in a simple form, the hazard function for the Yule distribution is

$$h_n = \frac{\theta}{n+\theta+1}.$$

This is analogous to the hazard function of the continuous Pareto distribution. Some authors prefer to call the Yule distribution the discrete Pareto distribution (Xekalaki and Panaretos 2006).

The Waring distribution extends the Yule distribution by an additional parameter κ:

$$f_n = \theta \frac{\Gamma(n+\kappa)\Gamma(\theta+\kappa)}{\Gamma(\kappa)\Gamma(n+\theta+\kappa+1)},\ n = 0, 1, 2, \cdots . \tag{3.16}$$

When n is sufficiently large, the parameter κ plays little role. Using the Barnes expansion (Johnson et al. 1993, p. 6),

$$\frac{\Gamma(n+\kappa)}{\Gamma(n+\kappa+\theta+1)} \approx \frac{1}{n^{\theta+1}}\left(1 - \frac{(\theta+1)(\theta+2\kappa)}{2n} + \cdots\right).$$

The generalized Waring distribution includes one more parameter:

$$f_n = \frac{\Gamma(\theta+\rho)\,\Gamma(\kappa+\theta)}{\Gamma(\theta)\Gamma(\rho)\,\Gamma(\kappa)} \frac{\Gamma(n+\kappa)\Gamma(n+\rho)}{\Gamma(n+1)\Gamma(n+\kappa+\theta+\rho)} \approx \frac{\Gamma(\theta+\rho)\,\Gamma(\kappa+\theta)}{\Gamma(\theta)\Gamma(\rho)\,\Gamma(\kappa)} n^{-(1+\theta)}.$$

These are all power-law distributions with the same tail property. They are more closely related to different stochastic mechanisms in disease transmission dynamics.

3.2 Random Count Distributions as Generated by Stochastic Disease Transmission Models

The discrete distributions for random counts often arise as marginal distributions of complex stochastic processes with respect to disease transmission. They involve the continuous lifetime distributions in the preceding chapter representing the

natural history of disease progression as well as point processes characterizing social contacts. Different processes may manifest the same marginal distribution for certain random count data, but they behave very differently regarding other characteristics of the transmission dynamic.

Simon (1962) provides a historical overview of different stochastic mechanisms that manifest the negative binomial distribution, in terms of accident proneness and contagion. In contemporary literature, Xekalaki (2014) gives a comprehensive synopsis of probability models for random counts by looking at their origins, motivation, and applications.

3.2.1 Mixture of Poisson Distributions and Processes

Greenwood and Yule (1920) developed the concept of "accident proneness," assuming the occurrence of accidents at individual level may be modeled as "pure chance" according to the Poisson distribution $f_n = \frac{1}{n!}\theta^n e^{-\theta}$, but individuals have constant but unequal probabilities of having an accident. This leads to the mixed-Poisson distribution to model such individual heterogeneity.

A mixed-Poisson distribution is constructed by $f(n|\theta) = \frac{\theta^n}{n!}e^{-\theta}$ and a mixing distribution $U(\theta)$. Using the notation in Karlis and Xekalaki (2005), we denote the mixed-Poisson distribution as

$$Poisson(\theta) \underset{\theta}{\wedge} u(\theta).$$

It is a discrete distribution with p.m.f. and p.g.f.

$$f_n = \frac{1}{n!}\int_0^\infty \theta^n e^{-\theta} dU(\theta), \tag{3.17}$$

$$G_N(s) = \int_0^\infty e^{-\theta(1-s)} dU(\theta) = L[u](1-s), \tag{3.18}$$

where $L[u](s) = \int_0^\infty e^{-s\theta} dU(\theta)$ is the Laplace transform of the mixing distribution. Some of the important properties are:

1. The mean value of N is $\mu = E[\theta] = \int_0^\infty \theta dU(\theta)$.
2. The variance is $var[N] = \mu + var[\theta]$ with extra-Poisson variation $var[\theta]$.
3. The probability of observing $\{N = 0\}$ is always higher in a mixed-Poisson distribution than in a simple Poisson distribution with the same mean (Feller 1943). More generally, if $f_n^{(1)}$ is the p.m.f. of the mixed-Poisson distribution given by (3.17) and $f_n^{(2)}$ is the p.m.f. of the Poisson distribution given by (3.9) and the two distributions have the same mean, then $f_n^{(1)} - f_n^{(2)}$ has exactly two sign changes $+, -, +$. This implies a mixed-Poisson distribution gives a higher probability for $\{N = 0\}$ and has a longer right tail (Shaked 1980).

4. Similar to the convex order to compare variability of two lifetimes with equal means (Chap. 2 of this book), Shaked (1980) showed that for every convex function $c(x)$, it holds that

$$\sum c(n) f_n^{(1)} \geq \sum c(n) f_n^{(2)}.$$

5. When the p.d.f. of the mixing distribution $u(\theta)$ exists, the shape of the p.m.f. of a mixed-Poisson distribution exhibits a resemblance to that of $u(\theta)$ (Karlis and Xekalaki 2005). Lynch (1988) extended the findings in Shaked (1980) and proved that mixing carries the form of the mixing distribution over to the resulting mixed distribution in general.

The Poisson distribution is closely related to the Poisson process.

Definition 11 Let $K(t)$ denote the cumulative number of events occurring during the time interval $[0, t]$, and the event history is $\mathcal{H}_t = \{K(u) : 0 \leq u \leq t^-\}$. The process $\{K(t) : t \geq 0\}$ is a time-homogeneous Poisson process of infinitesimal rate β, if for small $h > 0$, all $t > 0$, and the following conditions hold:

1. $K(0) = 0$;
2. $\Pr\{K(t) - K(t-h) = 1|\mathcal{H}_{t-h}\} = \beta h + o(h)$;
3. $\Pr\{K(t) - K(t-h) > 1|\mathcal{H}_{t-h}\} = o(h)$.

The marginal distribution of $K(t)$ is

$$\Pr(K(t) = k) = \frac{(\beta t)^k}{k!} e^{-\beta t} \sim Poisson(\beta t). \tag{3.19}$$

The mixed-Poisson process is thus defined by the random effect on the infinitesimal risk β such that the marginal distribution of $K(t)$ becomes

$$Poisson(\beta t) \underset{\beta}{\wedge} u(\beta).$$

The Negative Binomial Distribution as a Mixed-Poisson Distribution

The negative binomial distribution (3.12) has the bounded canonical parameter $\theta \in [0, 1]$. When used to model counts data, it typically arises in re-parameterized forms reflecting different origins of the stochastic mechanisms. The logit transform $\log \frac{\theta}{1-\theta} = \log \mu - \log \kappa$, $\mu > 0$ leads to the expression

$$f_n = \frac{\Gamma(n+\kappa)}{\Gamma(\kappa)\Gamma(n+1)} \left(\frac{\mu}{\kappa+\mu}\right)^n \left(\frac{\kappa}{\kappa+\mu}\right)^\kappa, \; n = 0, 1, 2, \cdots \tag{3.20}$$

with mean and variance $E[N] = \mu$ and $var[N] = \mu + \mu^2/\kappa$. Greenwood and Yule (1920) show that the above expression is $Poisson(\theta) \underset{\theta}{\wedge} u(\theta)$, where $u(\theta)$

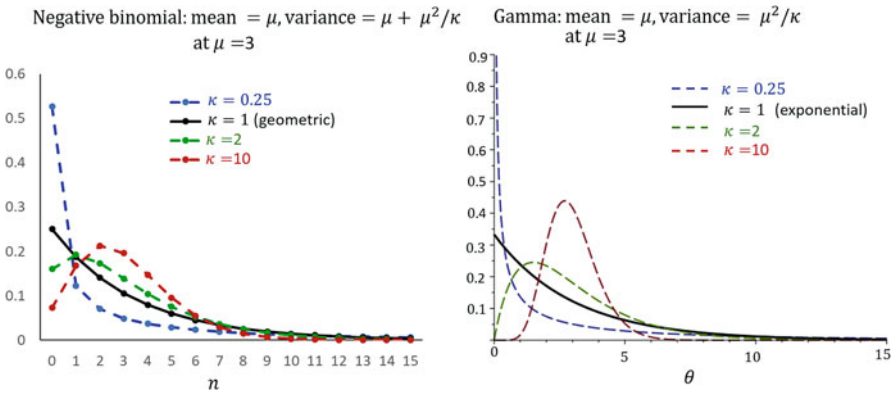


Fig. 3.1 Resemblance of the shape of the negative binomial distribution as a mixed-Poisson distribution and the shape of the gamma distribution as its mixing distribution under the same mean

is the gamma distribution with $E[\theta] = \mu$ and $var[\theta] = \mu^2/\kappa$. The hazard function corresponding to (3.20) approaches a constant $\lim_{n\to\infty} h_n = \frac{\kappa}{\kappa+\mu}$ and the mean residual life approaches $\lim_{n\to\infty} E(N - n|N \geq n) = \mu/\kappa$. These are the memoryless properties exhibited by the exponential tail. Figure 3.1 shows the resemblance of the p.m.f. of negative binomial distributions and the p.d.f. of the gamma distribution when compared at the same mean value $\mu = 3$. It also shows the "two crossings" in Shaked (1980). The shape parameter κ ranks the variability of both distributions according to convex order.

The Negative Binomial Distribution for the Infectious Contacts A very important random count is the number of infectious contacts produced by a typical infected individual during the entire infectious period, denoted by N. We examine the following negative binomial distribution for N :

$$f_n = \frac{\Gamma(n+\kappa)}{\Gamma(\kappa)\,\Gamma(n+1)} \left(\frac{\beta\mu_I}{\kappa+\beta\mu_I}\right)^n \left(\frac{\kappa}{\kappa+\beta\mu_I}\right)^\kappa, \; n = 0, 1, 2, \cdots, \tag{3.21}$$

where β is the infinitesimal rate of a Poisson process, and μ_I is the mean value of the infectious period. The mean and variance of N are $E[N] = \beta\mu_I$ and $var[N] = \beta\mu_I + \beta^2\mu_I^2/\kappa$. This distribution arises from at least two different mechanisms:

1. As a mixed-Poisson distribution $Poisson(\beta\mu_I) \underset{\beta}{\wedge} u(\beta)$. In this interpretation, we assume that the infectious period $T_I = \mu_I$ is not random. An infected individual i, who became infectious at time 0, produces infectious contacts according to a constant rate β_i so that the number of infectious contacts produced during the entire infectious period time interval $[0, \mu_I]$ is $Poisson(\beta_i\mu_I)$. Infected individuals have different infectious contact rates (i.e., proneness). If β_i is gamma distributed with mean value β and variance β^2/κ, then N has the negative

binomial distribution (3.21). The p.g.f. of N in $Poisson(\beta\mu_I) \underset{\beta}{\wedge} u(\beta)$ can be expressed by

$$G_N(s) = \int_0^\infty e^{-\beta\mu_I(1-s)} u(\beta) d\beta = L[u](\mu_I(1-s)), \tag{3.22}$$

where $L[u](s)$ is the Laplace transform of $u(\beta)$.

2. As a stopped Poisson process. In this interpretation, individuals are homogeneous. Each individual produces infectious contacts according to the same rate β. The number of infectious contacts produced during the time interval $[0, t]$ follows a Poisson distribution with mean value $\theta = \beta t$, as long as this individual is still infectious at time t. We assume the infectious period is a random variable T_I. It can be shown that, if T_I has a gamma distribution so that $E[T_I] = \mu_I$ and $var[T_I] = \mu_I^2/\kappa$, the number of infectious contacts produced by a typical infected individual during the entire infectious period follows the negative binomial distribution which is also expressed by (3.21). The p.g.f. of N in the stopped Poisson process is expressed by

$$G_N(s) = \int_0^\infty e^{-\beta x(1-s)} f_I(x) dx = L[f_I](\beta(1-s)), \tag{3.23}$$

where $f_I(x)$ is the p.d.f. of the infectious period and $L[f_I](s)$ is the Laplace transform of the infectious period.

The two interpretations have very different underlying assumptions. As far as the distribution for N is concerned, data arising from a negative binomial distribution (3.21) cannot identify whether the underlying mechanism corresponds to (3.22) or (3.23). It will be shown later in this book that, some key features of the epidemic are determined by the distribution N itself, regardless of the underlying stochastic mechanisms whereas some other features of the epidemic are dependent on whether the underlying mechanism corresponds to (3.23) or (3.22).

Other Mixed-Poisson Distributions

To obtain mixed-Poisson distributions, direct evaluation of the integrand (3.17) is difficult with the exception of a few mixing distributions. Similarly, to obtain the p.g.f. using (3.18), it is only feasible for the mixing distribution $U(\theta)$ of which the Laplace transform can be written explicitly, such as the gamma distribution and the inverse-Gaussian distribution.

Panjer and Willmot (1982) and Willmot (1993) show that, for several mixed Poisson distributions, a recursive formula can be obtained. A large number of Poisson mixtures have been developed. For an extensive review, see Karlis and Xekalaki (2005), in which Table 1 provides more than 30 different mixed-Poisson distributions along with references. Some of these distributions are more relevant in

disease modeling, whereas others have found their applications in actuarial sciences and social sciences.

We introduce a generalization of the mixed-Poisson distribution in the context of the number of infectious contacts with two levels of mixing.

Let the cumulative number of infectious contacts generated by a typical infected individual during the time interval $[0, t]$ follow the Poisson distribution $K(t) \sim Poisson(\xi t)$, which is the marginal distribution of a Poisson process $\{K(t) : t \geq 0\}$. We assume individual heterogeneity (e.g., proneness) and consider the mixed-Poisson process $Poisson(\xi t) \underset{\xi}{\wedge} u(\xi)$ and $u(\xi)$ follows the exponential distribution with $E[\xi] = \beta$. Then by time t, the marginal distribution of $K(t)$ arising from a mixed-Poisson process follows the geometric distribution with p.m.f. $\left(\frac{\beta t}{1+\beta t}\right)^n \left(\frac{1}{1+\beta t}\right)$, provided that at time t the individual is still infectious. We further assume that the infectious period T_I is random, and in particular, following a Pareto-II distribution with p.d.f. $\beta\theta\,(\beta t + 1)^{-(\theta+1)}$. Then the number of infectious contacts produced by this individual during the entire infectious period is $N = K(T_I)$ and has p.m.f.

$$\begin{aligned}\Pr(N = n) &= \beta\theta \int_0^\infty \left(\frac{\beta t}{1+\beta t}\right)^n \left(\frac{1}{1+\beta t}\right) (\beta t + 1)^{-(\theta+1)}\, dt \qquad (3.24)\\ &= \theta \int_0^\infty \left(\frac{x}{1+x}\right)^n \left(\frac{1}{1+x}\right) (x+1)^{-(\theta+1)}\, dx\\ &= \theta \frac{\Gamma(n+1)\,\Gamma(\theta+1)}{\Gamma(n+\theta+2)},\ n = 0, 1, 2, \cdots\end{aligned}$$

which is the Yule distribution (3.15). The last identity can be easily proven by rewriting $\varsigma = \frac{1}{1+x}$ such that $dx = -\varsigma^{-2} d\varsigma$ and $\theta \left(\frac{1}{x+1}\right)^{\theta+1} = \theta \varsigma^{\theta+1}$. In this case, $\int_0^1 (1-\varsigma)^n\, \varsigma\theta\varsigma^{\theta-1} d\varsigma = \theta \frac{\Gamma(n+1)\Gamma(\theta+1)}{\Gamma(n+\theta+2)}$.

An immediate generalization is to assume that $u(\xi)$ follows the gamma distribution with $E[\xi] = \beta$ and $var[\xi] = \beta^2/\kappa$ so that the marginal distribution of $K(t)$ arising from a mixed-Poisson process follows a negative binomial distribution. The infectious period T_I still follows the Pareto-II distribution with p.d.f. $\frac{\beta\theta}{\kappa}\left(1 + \frac{\beta t}{\kappa}\right)^{-(\theta+1)}$. The resulting distribution for $N = K(T_I)$ is

$$\begin{aligned}&\frac{\Gamma(n+\kappa)}{\Gamma(n+1)\Gamma(\kappa)} \int_0^\infty \left(\frac{\beta t}{\kappa+\beta t}\right)^n \left(\frac{\kappa}{\kappa+\beta t}\right)^\kappa \frac{\beta\theta}{\kappa} \left(1 + \frac{\beta t}{\kappa}\right)^{-(\theta+1)} dt \qquad (3.25)\\ &= \frac{\Gamma(n+\kappa)}{\Gamma(n+1)\Gamma(\kappa)} \int_0^1 (1-\varsigma)^n\, \varsigma^\kappa \theta \varsigma^{\theta-1} d\varsigma\\ &= \theta \frac{\Gamma(n+\kappa)\Gamma(\theta+\kappa)}{\Gamma(\kappa)\Gamma(n+\theta+\kappa+1)},,\ n = 0, 1, 2, \cdots\end{aligned}$$

which is (3.16).

Therefore the Yule and the Waring distributions can be interpreted as mixed-Poisson distributions with two levels of mixing. They belong to the family of generalized negative binomial convolution as defined in Bondesson (1979).

3.2.2 *Highly Skewed Data: Proneness, Contagion, or Spells?*

Count data arising from infectious diseases tend to be over-dispersed (variance is greater than the mean). One frequent manifestation of over-dispersed data is that the incidence of zero counts is greater than expected for the Poisson distribution. The negative binomial distribution is more flexible than the Poisson distribution, but it is better for over-dispersed count data that are not necessarily heavy tailed. Figure 3.2

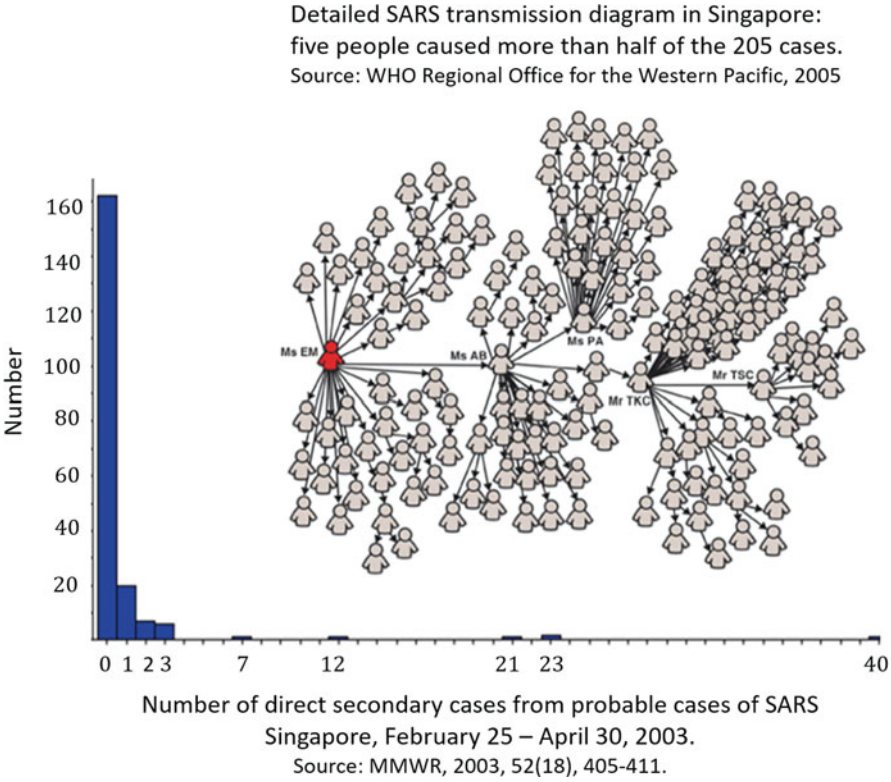


Fig. 3.2 Illustration of a count distribution with both a large frequency of zeros and a very heavy tail

shows that, out of 205 confirmed SARS cases in Singapore between February 25 and April 30 2003, 163 cases produced zero secondary transmissions, whereas 5 cases were likely responsible for more than half of the 205 cases. Lord and Geedipally (2011) showed that, for data with a large frequency of zeros and a very heavy tail, the Poisson distribution tends to underestimate the number of zeros given the mean of the data, while the NB distributions may overestimate zeros, but underestimate observations with a count.

Both the Waring distribution (3.16) and the Yule distribution as its special case $\kappa = 1$ are highly skewed distributions exhibiting such a tail property. It can be derived as a Beta mixture of the negative binomial distribution in its canonical form $\frac{\Gamma(n+\kappa)}{\Gamma(n+1)\Gamma(\kappa)}(1-\varsigma)^n \varsigma^\kappa$ assuming the canonical parameter $\varsigma \in (0,1)$ having the Beta distribution with p.d.f. $u(\varsigma) = \theta \varsigma^{\theta-1}$, $0 < \varsigma < 1$ so that

$$\frac{\Gamma(n+\kappa)}{\Gamma(n+1)\Gamma(\kappa)} \int_0^1 \varsigma^\kappa (1-\varsigma)^n \left(\theta \varsigma^{\theta-1}\right) d\varsigma = \theta \frac{\Gamma(n+\kappa)\Gamma(\theta+\kappa)}{\Gamma(\kappa)\Gamma(n+\theta+\kappa+1)}. \tag{3.26}$$

This formulation can be further written into two different ways:

1. The canonical parameter ς can be re-parameterized via the logit link function $\log\left(\frac{1-\varsigma}{\varsigma}\right) = \log\frac{\beta}{\kappa} + \log t$. If the infectious contact process $\{K(t)\}$ has the marginal distribution following negative binomial distribution expressed as:

$$\Pr\{K(t) = n\} = \frac{\Gamma(n+\kappa)}{\Gamma(n+1)\Gamma(\kappa)} \left(\frac{\beta t}{\kappa+\beta t}\right)^n \left(\frac{\kappa}{\kappa+\beta t}\right)^\kappa, \tag{3.27}$$

   assuming the infectious period T_I following the Pareto-II distribution with p.d.f. $\frac{\beta\theta}{\kappa}\left(1+\frac{\beta t}{\kappa}\right)^{-(\theta+1)}$, then (3.25) becomes (3.26).
2. Alternatively, the canonical parameter ς can be re-parameterized via the complementary logarithm link function $\log(-\log(1-\varsigma)) = \log\frac{\beta}{\kappa} + \log t$. If the infectious contact process $\{K(t)\}$ has the marginal distribution following negative binomial distribution expressed as:

$$\Pr\{K(t) = n\} = \frac{\Gamma(n+\kappa)}{\Gamma(n+1)\Gamma(\kappa)} \left(1-e^{-\beta t/\kappa}\right)^n e^{-\beta t}, \tag{3.28}$$

   assuming an exponentially distributed infectious period with p.d.f. $f_I(t) = \frac{1}{\mu_I} e^{-t/\mu_I}$. In this case, $N = K(T_I)$ has the distribution given by

$$\Pr(N = n) = \int_0^\infty \frac{\Gamma(n+\kappa)}{\Gamma(n+1)\Gamma(\kappa)} \left(1-e^{-\beta t/\kappa}\right)^n e^{-\beta t} \left(\frac{1}{\mu_I} e^{-t/\mu_I}\right) dt. \tag{3.29}$$

   Letting $\varsigma = e^{-\beta t/\kappa}$ and $\theta = \frac{\kappa}{\beta\mu_I}$, (3.29) also returns to (3.26).

The marginal distribution expressed by (3.27) corresponds to the gamma mixture of the Poisson process assuming individual heterogeneity (e.g., proneness) among infectious contacts. It assumes that the intensity rate of producing a new infection at individual level is independent from the past history, but individuals are heterogeneous due to unequal transmission rates or randomness in the infectious period. At the population level such that the count numbers in two non-overlapping intervals are correlated. This is the "proneness" argument.

The marginal distribution expressed by (3.28) corresponds to a completely different stochastic process for the infectious contacts. It arises from a linear pure birth process (Bartlett 1955; Bhattacharya and Waymire 1990; Allen 2010), corresponding to the "contagion" argument (Irwin 1941; McKendrick 1925). It assumes that initially all individuals have the same probability of incurring an accident, but later this probability changes by each accident sustained. In the infectious disease context, if an infectious individual has produced k infectious contacts by time x, the hazard for producing the $(k+1)$th contact is a linear increasing function of k; thus, the more infectious contacts it produces, the more likely it produces more infectious contacts. This phenomena is sometimes referred to as preferential contacts in the literature related to modeling social contact networks (Barabási and Albert 1999). At the population level, the count numbers in two non-overlapping intervals are independent.

Consider a linear pure birth process (the Yule process) with

$$\Pr\{K(t+h) = n+1 | K(t) = n\} = \beta(n+1)h + o(h),\ n = 0, 1, 2, \cdots. \tag{3.30}$$

According to (3.30), given that an infectious individual has produced n infectious contacts by time t, the hazard for producing the $(n+1)$th contact is an increasing function of n. Starting at $t = 0$, corresponding to the beginning of the infectious period for an individual, the waiting time to producing the first infectious contact is exponentially distributed with mean $E[X_1] = \frac{1}{\beta}$; conditioning on the first infectious contact, the waiting time to the second infectious contact is exponentially distributed with mean $E[X_2] = \frac{1}{2\beta}$; $\cdots$; and conditioning on an infectious individual who has produced n infectious contacts, the waiting time to producing the $(n+1)$th infectious contact is exponentially distributed with mean $E[X_{n+1}] = \frac{1}{(n+1)\beta}$. On the surface, it looks as if the more infectious contacts it produces, the more likely it produces more infectious contacts.

More generally, a pure birth process is defined by the transition probability

$$\Pr\{K(t+h) = n+1 | K(t) = n\} = (\beta_1 n + \beta_2)h + o(h). \tag{3.31}$$

Conditioning on $K(t) = n$, the instantaneous rate of producing the next infectious contact during $[t, t+h)$ can be considered as an independent competing risk hazard: either from a global environment with constant rate β_2, or from a clustered

environment with non-constant rate $\beta_1 n$ and $\beta_1 \neq \beta_2$. The hazard of producing an infectious contact at time t is

$$\lim_{h\to 0} \frac{\Pr\{K(t+h) = n+1 | K(t) = n\}}{h} = \beta_1 n + \beta_2 = \beta\left(n/\kappa + 1\right), \tag{3.32}$$

where $\kappa = \beta_2/\beta_1$. If $\{K(t)\}$ arises as a linear pure birth process given by (3.31), from page 274 of Bhattacharya and Waymire (1990), the marginal distribution for $K(t)$ follows the negative binomial distribution given by (3.28).

When $\kappa \to \infty$, the linear pure birth process reduces to a Poisson process with $\beta = \beta_2$

$$\Pr\{K(x+h) = n+1 | K(x) = n\} = \beta h + o(h),\ n = 0, 1, 2, \cdots. \tag{3.33}$$

When $\kappa = 1$, the linear pure birth process reduces to the Yule process given by (3.30) and (3.28) reduces to the geometric distribution

$$\Pr\{K(t) = n\} = \left(1 - e^{-\beta t}\right)^n e^{-\beta t}. \tag{3.34}$$

On the surface, (3.28) is simply a re-parameterization of (3.27). The fundamental difference is the underlying process that may have substantial implications on certain features of disease transmission.

Example 12 Let's consider the case that each infected individual has an infectious period $T_I = \mu_I$ which is not random. The distribution of the number of infectious contacts produced by this individual during the entire infectious period is $N = K(T_I)$ and follows a negative binomial distribution, which could be either expressed as

$$\Pr\{N = n\} = \frac{\Gamma(n+\kappa)}{\Gamma(\kappa)\,\Gamma(n+1)} \left(\frac{\beta\mu_I}{\kappa + \beta\mu_I}\right)^n \left(\frac{\kappa}{\kappa + \beta\mu_I}\right)^\kappa,\ n = 0, 1, 2, \cdots \tag{3.35}$$

derived from (3.27) or expressed as

$$\Pr\{N = n\} = \frac{\Gamma(n+\kappa)}{\Gamma(n+1)\Gamma(\kappa)} \left(1 - e^{-\beta\mu_I/\kappa}\right)^n e^{-\beta\mu_I},\ n = 0, 1, 2, \cdots \tag{3.36}$$

derived from (3.28). The mean value $E[N]$ defines the basic reproduction number in epidemiology, $R_0 = E[N]$. With respect to (3.35), the mean value $R_0 = \beta\mu_I$ is proportional to the infectious period. The parameter κ ranks the variance of the individuals heterogeneity regarding the transmission rate but it has no effect on the basic reproduction number. However, with respect to (3.36),

$$\begin{aligned} R_0 = E[N] &= \kappa \left(e^{\beta\mu_I/\kappa} - 1\right) \\ &= \beta\mu_I + \frac{1}{2\kappa}(\beta\mu_I)^2 + \frac{1}{6\kappa^2}(\beta\mu_I)^3 + O\left((\beta\mu_I)^4\right) \\ &\to \beta\mu_I, \quad \text{as } \kappa \to \infty. \end{aligned}$$

As the "contagion" factor $1/\kappa$ increases, so does the value of R_0.

In addition to heterogeneity (proneness) and contagion arguments, there are other stochastic mechanisms that manifest the same distribution. *Clustering* is another possibility. Cresswell and Froggatt (1963) introduced such a model in the accident context whereby the number of injuries by a person involved can be thought as a random sum

$$M = Y_1 + Y_2 + \cdots + Y_N,$$

where Y_i are i.i.d. random numbers with mean $E[Y] = \mu_Y$, variance $var[Y] = \sigma_Y^2$, and N follows a Poisson distribution with mean equal to its variance $E[N] = var[N]$. As a result,

$$\begin{aligned} E[M] &= \mu_Y E[N] \\ var[M] &= (\mu_Y^2 + \sigma_Y^2)E[N] \end{aligned}$$

and data are over-dispersed if $\mu_Y^2 + \sigma_Y^2 > 1$. In the original work of Cresswell and Froggatt (1963), Y_i is the number of injuries from the ith accident, and each person is liable to *spells* of weak performance during which the accidents occur. When Y_i follows a logarithmic distribution, the total number of injuries M follows a negative binomial distribution. Xekalaki (1983) discusses spells model in terms of a generalized Waring distribution. Xekalaki and Zografi (2008) developed the generalized Waring process and show that such a process could also be interpretated in the context of a spells model and used in modeling temporally evolving data.

Without going further into details of the spells model, we point out that it is also highly relevant to disease modeling. Taking the number of infectious contacts generated by a typical infected individual as example, spells may arise from spatial clustering in external environment and only during such spell the infected individual is exposed to a large number of susceptible individuals or arise internally in terms of fluctuating infectiousness.

The above discussion reveals the following important points:

1. It seems to be quite common that count data from infectious diseases tend to exhibit a highly skewed tail property and power-law, similar to the distribution in Fig. 3.2.

2. A highly skewed distribution that describes the number of infectious contacts produced by a typical individual during the entire infectious period could be
   (a) due to a highly skewed infectious period distribution such as the Pareto distribution in (3.24), whereas the infectious contact process has moderate heterogeneity in terms of the transmission rate β resulting in a geometric distributed marginal distribution for $K(t)$;
   (b) or due to the contagion factor in the infectious contact process modeled by a linear pure birth process, whereas the infectious period is exponentially distributed as expressed by (3.29);
   (c) or due to spells either arising externally from spatial clustering or internally as infectiousness fluctuates;
   (d) the count data themselves cannot distinguish these mechanisms.
3. The power-law distributions given by (3.14)–(3.16) involve an important parameter $\theta > 0$. If $\theta \leq 1$, the mean $E[N]$ is infinite. Therefore, if the number of infectious contacts produced by this individual during the entire infectious period is best fitted by such distributions, there is a possibility that R_0 cannot be defined.

3.3 General Formulation of a Counting Process

A counting process generates random counts. It is a stochastic process $\{K(t) : t \geq 0\}$ with $K(0) = 0$ and $K(t) < \infty$, whose paths are with probability one right-continuous, piecewise constant, and have only jump discontinuities, with jumps of size $+1$. For comprehensive reading, we recommend Fleming and Harrington (1991) and Andersen et al. (1993).

A counting process $\{K(t) : t \geq 0\}$ is adopted to the history $\mathcal{H}_t = \{K(u), 0 < u \leq t^-\}$ which contains the information generated by the process $K(t)$ on $[0, t^-)$. $\mathcal{H}_t$ may depend on the history of more than one correlated processes on $[0, t^-)$, especially counting processes arising from the transmission of infectious diseases. For example, in the stochastic susceptible-infectious-recovered (SIR) model (Chap. 5), at any given time t, the numbers of susceptible, infectious, and recovered individuals $\{S(t), I(t), R(t)\}$ are all random counts. The cumulative number of infected individuals $C(t) = I(t) + R(t)$ and the number of recovered individuals $R(t)$ by time t form two counting processes $\{C(t) : t \geq 0\}$ and $\{R(t) : t \geq 0\}$. Both are adopted to the history $\mathcal{H}_t = \left\{[S(u), I(u)] : 0 \leq u \leq t^-\right\}$. However, not all random counts are generated from counting processes. For instance, the number of infectious individuals at time t, $I(t)$ is a random count but the process $\{I(t) : t \geq 0\}$ is not a counting process.

A counting process $\{K(t) : t \geq 0\}$ is jointly specified by several features.

1. *The marginal distribution*: The number of events during a time interval $[t_1, t_2)$ is $K(t_2, t_1) = K(t_2) - K(t_1) \geq 0$ and the cumulative number of events by time t, $K(t) = K(t, 0)$, are random counts.

(a) A counting process is said to have *independent increments* if the number of events in disjoint intervals are independent.
(b) A counting process is said to have *stationary increments* if the distribution of the number of events that occur in any time interval depends only on the length of the time interval, that is, $K(x+t,x) = K(t)$.

2. *The intensity*: The instantaneous intensity of the increments is defined as

$$\begin{aligned}\lambda(t) &= \frac{1}{dt}\Pr\{K(t+dt) - K(t) = 1|\mathcal{H}_t\} \\ &= \frac{1}{dt}E[K(t+dt) - K(t)|\mathcal{H}_t] = \frac{d}{dt}E[K(t)|\mathcal{H}_t].\end{aligned} \tag{3.37}$$

In general, $\{\lambda(t), t \geq 0\}$ is a stochastic process because it is a function of stochastic events in the past given by $\{\mathcal{H}_t : t \geq 0\}$. The cumulative intensity $\{\Lambda(t) = \int_0^t \lambda(u)du : t \geq 0\}$ is also a stochastic process adopted to the same history $\{\mathcal{H}_t : t \geq 0\}$. $\{\Lambda(t) : t \geq 0\}$ is called the *compensator* of the counting process $\{K(t) : t \geq 0\}$. A counting process is often modeled by specifying the intensity process.

(a) The intensity process $\lambda(t)$ may be a deterministic function of time denoted by $\beta(t) \geq 0$ rather than a stochastic process. In this case, it is not dependent on the past. A special case is $\lambda(t) \equiv \beta$, which equals stationary increment.
(b) The intensity process $\lambda(t)$ may be modeled by a multiplicative model $\lambda(t) = \beta(t)\xi$, where $\beta(t)$ is a deterministic function of time t and ξ is a random variable but not a stochastic process. A further special case is $\beta(t) \equiv \beta$. In this case, $E[\lambda(t)]$ is constant, implying a stationary increment.
(c) The intensity process $\lambda(t)$ may be modeled by a multiplicative model $\lambda(t) = \beta(t)W(t)$, where $\beta(t)$ is a deterministic function of time t and $\{W(t) : t \geq 0\}$ is a stochastic process adopted to the history $\mathcal{H}_t$. In later chapters with respect to dynamic disease transmission models, we shall see models such as $\lambda(t) = \beta K(t)$, $\lambda(t) = \beta K(t)[n - K(t)]$, $\lambda(t) = \beta(t)K(t)[n - K(t)]$, $\beta > 0$, etc.

3. *Gaps between successive events*: For a counting process, let us denote X_1 the time to the first event, and for $k \geq 1$, X_k the time between the $(k-1)$th and the kth events. We further let $Y_1 = X_1$, $Y_2 = X_1 + X_2, \cdots, Y_k = X_1 + X_2 + \cdots + X_k$ to denote the times that events occur. Some of the key features of the counting process are studied through continuous lifetime distributions for gaps between successive events $X_1, X_2, \cdots, X_k, \cdots$ and the timing of events $Y_1, Y_2, \cdots, Y_k, \cdots$.

A typical sample path of a counting process is illustrated schematically in Fig. 3.3.

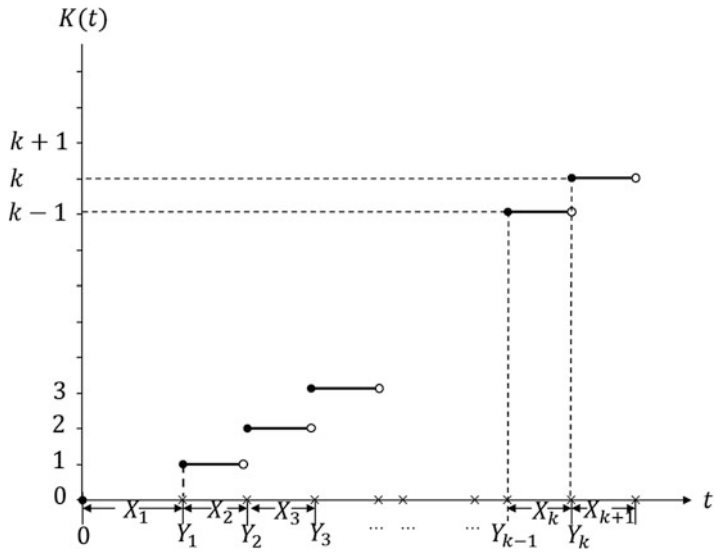


Fig. 3.3 Diagram for random variables representing time of events

3.3.1 Review of Some of the Counting Processes that Have Been Mentioned Earlier

The Time-Homogeneous Poisson Process Given by Definition 11

The time-homogeneous Poisson process can be considered as a canonical counting process. It has several equivalent definitions. It simultaneously satisfies:

1. The marginal distribution $K(t) = K(t, 0)$ has the Poisson distribution

$$\Pr(K(t) = n) = \frac{(\beta t)^n}{n!} e^{-\beta t}.$$

2. The process is Markovian because the future sample path only depends on $K(t)$ as given by Definition 11.
3. The process has stationary increment, i.e., the marginal distribution for $K(x + t, x)$ is identical to that of $K(t)$ for any $x \geq 0$.
4. The process has independent increments.
5. The gaps between successive events $X_1, X_2, \cdots, X_k, \cdots$ are i.i.d. exponentially distributed with mean $E[X] = 1/\beta$ and hazard rate $h(x) = \beta$.
6. The waiting time to the kth event Y_k has a gamma distribution with p.d.f. $f(y) = \frac{\beta}{\Gamma(k)} (\beta y)^{k-1} e^{-\beta y}$.

Counting Processes with the Negative Binomial Distribution as the Marginal Distribution for Count Numbers

Unlike the Poisson process, named after the marginal distribution of the count numbers, the "negative binomial processes" is a collective term of several counting processes whose marginal distributions can be expressed as a negative binomial distribution. In previous subsections, we have seen two negative binomial presentations for $K(t)$, given by (3.27) and (3.28), arising from two different underlying stochastic mechanisms. There are other counting processes that give rise to the negative binomially distributed marginal distributions.

Gamma Mixture of the Poisson Process The marginal distribution given by (3.27) typically arises from a mixed-Poisson process in which the marginal distribution is $Poisson(\beta t) \underset{\beta}{\wedge} u(\beta)$, where $u(\beta)$ follows a gamma distribution. One can re-write the original Poisson process as $Poisson(\lambda t)$, in which, $\lambda = \beta\xi$ with β being a constant and ξ the random effect, which is gamma distributed with mean $E[\xi] = 1$ and variance $var[\xi] = 1/\kappa$. The p.d.f. is $u(\xi) = \frac{\kappa}{\Gamma(\kappa)} (\xi\kappa)^{\kappa-1} e^{-\xi\kappa}$. The marginal distribution is

$$\begin{aligned} \Pr(K(t) = n) &= \int_0^\infty \frac{(\beta\xi t)^n}{n!} e^{-\beta\xi t} \frac{\kappa}{\Gamma(\kappa)} (\xi\kappa)^{\kappa-1} e^{-\xi\kappa} d\xi \\ &= \frac{\Gamma(n+\kappa)}{\Gamma(n+1)\Gamma(\kappa)} \left(\frac{\beta t}{\kappa+\beta t}\right)^n \left(\frac{\kappa}{\kappa+\beta t}\right)^\kappa. \end{aligned} \tag{3.38}$$

The intensity $\lambda = \beta\xi$ is a random variable with p.d.f. $u_\lambda(\lambda) = \frac{1}{\beta} \frac{\kappa}{\Gamma(\kappa)} \left(\frac{\lambda}{\beta}\kappa\right)^{\kappa-1} e^{-\frac{\lambda}{\beta}\kappa}$ and mean value $E[\lambda] = \beta$. However, it is dependent on $\{\mathcal{H}_t : t \geq 0\}$ through the value of $K(t)$. Given $K(t) = n$, the conditional probability for the intensity for λ is

$$\begin{aligned} f(\lambda|K(t) = n) &= \frac{\Pr(K(t) = n|\lambda(t) = \lambda) u_\lambda(\lambda)}{\Pr(K(t) = n)} \\ &= \frac{\frac{(\lambda t)^n}{\Gamma(n+1)} e^{-\lambda t} \frac{1}{\beta} \frac{\kappa}{\Gamma(\kappa)} \left(\frac{\lambda}{\beta}\kappa\right)^{\kappa-1} e^{-\frac{\lambda}{\beta}\kappa}}{\frac{\Gamma(n+\kappa)}{\Gamma(n+1)\Gamma(\kappa)} \left(\frac{\beta t}{\kappa+\beta t}\right)^n \left(\frac{\kappa}{\kappa+\beta t}\right)^\kappa} \\ &= \frac{1}{\lambda\Gamma(n+\kappa)} \left(\lambda \frac{\kappa+\beta t}{\beta}\right)^{n+\kappa} e^{-\lambda\frac{\kappa+\beta t}{\beta}}. \end{aligned} \tag{3.39}$$

The conditional expectation is

$$E[\lambda|K(t) = n] = \frac{1}{\Gamma(n+\kappa)} \int_0^\infty \left(\lambda \frac{\kappa+\beta t}{\beta}\right)^{n+\kappa} e^{-\lambda\frac{\kappa+\beta t}{\beta}} d\lambda = \frac{\beta(n+\kappa)}{\kappa+\beta t}. \tag{3.40}$$

Because it depends on $\{\mathcal{H}_t : t \geq 0\}$ only through $K(t) = n$, it is still Markovian. However, this process does not have independent increments. In fact, if $(s_1, t_1]$, $(s_2, t_2]$ are two non-overlapping intervals and let $x_1 = t_1 - s_1$, $x_2 = t_2 - s_2$, then $cov\,[K(s_1, t_1), K(s_2, t_2)] = \kappa^{-1}\beta^2 x_1 x_2$ (see Cook and Lawless 2007, p. 37). With respect to the gaps between successive events $X_1, X_2, \cdots, X_k, \cdots$, they are identically distributed and X_1 arises from a gamma mixture of the exponential distribution, which is a Pareto-II distribution with hazard rate $h(x) = \frac{\kappa\beta}{\kappa+\beta x}$. Note that when $\kappa \to \infty$, this process returns to the Poisson process, with $\lim_{\kappa\to\infty} \frac{\beta(n+\kappa)}{\kappa+\beta t} = \lim_{\kappa\to\infty} \frac{\kappa\beta}{\kappa+\beta x} = \beta$. However, the gaps are not independent.

Many of the above statements are also true for mixed-Poisson processes with arbitrary p.d.f. $u(\xi)$ with $E[\xi] = 1$ and $u_\lambda(\lambda) = \frac{1}{\beta}u(\lambda/\beta)$. The general expressions for (3.38)–(3.40) become

$$\Pr(K(t) = n) = \frac{(\beta t)^n}{\Gamma(n+1)} \int_0^\infty \xi^n e^{-\beta\xi t} u(\xi) d\xi,$$

$$f(\lambda | K(t) = n) = \frac{\lambda^n e^{-\lambda t} u_\lambda(\lambda)}{\int_0^\infty \lambda^n e^{-\lambda t} u_\lambda(\lambda) d\lambda},$$

$$E[\lambda | K(t) = n] = \frac{\int_0^\infty \lambda^{n+1} e^{-\lambda t} u_\lambda(\lambda) d\lambda}{\int_0^\infty \lambda^n e^{-\lambda t} u_\lambda(\lambda) d\lambda},$$

respectively.

1. These processes have stationary increments because at any given time t, $E[K(t)] = E[E[K(t)|\xi]] = \beta t$, hence $\frac{d}{dt}E[K(t)] = E[\lambda] = \beta$.
2. These processes do not have independent increments. One only needs to show that $\Pr\{K(s) = n, K(s+t) - K(s) = l\} \neq \Pr\{K(s) = n\} \Pr\{K(s+t) - K(s) = l\}$. In fact,

$$\begin{aligned}
&\Pr\{K(s) = n, K(s+t) - K(s) = l\} \\
&\quad = \int_0^\infty \Pr\{K(s) = n, K(s+t) - K(s) = l | \xi\} u(\xi) d\xi \\
&\quad = \int_0^\infty \frac{(\xi s)^n}{n!} e^{-\xi s} \frac{(\xi t)^l}{l!} e^{-\xi t} u(\xi) d\xi \\
&\quad \neq \int_0^\infty \frac{(\xi s)^n}{n!} e^{-\xi s} u(\xi) d\xi \int_0^\infty \frac{(\xi t)^l}{l!} e^{-\xi t} u(\xi) d\xi \\
&\quad = \Pr\{K(s) = n\} \Pr\{K(s+t) - K(s) = l\}.
\end{aligned}$$

3. The gaps between successive events $X_1, X_2, \cdots, X_k, \cdots$ are identically distributed but not independent. The survival function for X_1 is $\overline{F}(x) = \int_0^\infty e^{-\beta\xi x} u(\xi) d\xi = L[u](\beta x)$, where $L[u](s) = \int_0^\infty e^{-s\xi} u(\xi) d\xi$ is the Laplace transform with respect to $u(\xi)$, and the hazard function is $h(x) = -\frac{d}{dx} \log L[u](\beta x)$ (see Chap. 2).

The Linear Birth Process as a Negative Binomial Process The marginal distribution given by (3.28) is also negative binomial. The intensity $\{\lambda(t), t \geq 0\}$ is a stochastic process because it is a function of stochastic events in the past given by $\{\mathcal{H}_t : t \geq 0\}$. From (3.32), $\lambda(t)$ depends on the state $K(t) = n$ at time t, and hence

$$\lambda(t) = \beta_1 K(t) + \beta_2 = \beta\left(K(t)/\kappa + 1\right),$$

where $\kappa = \beta_2/\beta_1$.

1. This process does not have stationary increments because $E[K(t)] = \kappa\left(e^{\beta t/\kappa} - 1\right)$ and $\frac{d}{dt}E[K(t)] = E[\lambda(t)] = \beta e^{\frac{t}{\kappa}\beta}$.
2. This process does not have independent increment. This is directly from the definition, $\Pr\{K(t+h) - K(t) = 1 | K(t) = n\} = \beta(n/\kappa + 1)h + o(h)$ which depends on $K(t) = n$.
3. With respect to the gaps between successive events $X_1, X_2, \cdots, X_k, \cdots$, they are independent but not identically distributed. Starting at $K(0) = 0$, X_1 is exponentially distributed with mean $E[X_1] = 1/\beta$; X_2 is exponentially distributed with mean $E[X_2] = \frac{\kappa}{\beta(\kappa+1)}$; the gap between the nth and the $(n+1)$th events X_{n+1} is exponentially distributed with mean $E[X_{n+1}] = \frac{\kappa}{\beta(n+\kappa)}$; etc.

A Linear Birth Process for Initial Exponential Growth of an Epidemic in a Population The following counting process has been used in the literature as an approximate model for the exponential growth in an infinitely large population during the initial phase of an epidemic when the depletion of the susceptible individuals is negligible. Let $\{C(t) : t \geq 0\}$ be a counting process, where $C(t)$ is the cumulative number of infections at time t. The conditional probabilities are specified by

$$\begin{aligned}
&\Pr\{C(t+h) = n+1 | C(t) = n\} = rnh + o(h) \\
&\Pr\{C(t+h) = n | C(t) = n\} = 1 - rnh + o(h), \quad n = 1, 2, \ldots \\
&\Pr\{C(t+h) > n+1 | C(t) = n\} = o(h).
\end{aligned} \tag{3.41}$$

This process corresponds to $\beta_2 = 0$ in (3.28) but also modified so that $n \geq 1$. The rate $r > 0$ is called the Malthusian number.

From page 250 of Allen (2010), given the initial condition $C(0) = n_0$, the marginal distribution of $C(t)$ at time t follows the negative binomial distribution with the exponential growth $E[C(t)] = n_0 e^{rt}$. The variance also grows exponentially over time: $var[C(t)] = n_0 e^{2rt}(1 - e^{-rt})$.

With respect to the intensity process $\lambda(t) = \frac{d}{dt}E[C(t)|\mathcal{H}_t]$, it is Markovian, depending on $\{\mathcal{H}_t : t \geq 0\}$ only through the value of $C(t)$ at time t. It can be written as $\lambda(t) = rC(t)$. Hence $\frac{d}{dt}E[C(t)] = rE[C(t)]$. This argument corresponds to the deterministic model $C_d'(t) = rC_d(t)$, where $C_d(t)$ is a deterministic function for the cumulative infections that approximates the mean value $E[C(t)]$.

The gap between the nth and the $(n+1)$th infection in the population, X_{n+1}, is exponentially distributed with mean $E[X_{n+1}] = \frac{1}{nr}$. This model assumes that each of the infected individuals at time t contributes a constant rate r to producing a new infection during $(t, t+h]$ independently. Given $C(t) = n$, the time to producing the next new infection is the first of the n order statistics of independently distributed exponential distributions with rate r, which is also exponentially distributed with rate nr.

3.3.2 Martingales and Their Relations with Counting Processes

This advanced topic is beyond the scope of this book. We only provide a brief note because it is considered one of the statistical models used to fit complex disease transmission models to time-series data. Readers may skip this section for the time being and come back to it when referenced.

Martingale: A Very Brief Introduction

Let $\{M(t) : t \geq 0\}$ be a right-continuous stochastic process (Fig. 3.3) with left-hand limits adopted to the filtration history $\mathcal{H}_t$. We say $\{M(t) : t > 0\}$ is a martingale if (i) $E\,|M(t)| < \infty$ for all $t < \infty$; (ii) $E\,\{M(t+x)|\mathcal{H}_t\} = M(t)$, a.s. for all $x \geq 0, t \geq 0$. If we replace (ii) above by $E\,\{M(t+x)|\mathcal{H}_t\} \geq M(t)$, almost surely, for all $x \geq 0, t \geq 0$, we call $\{M(t) : t > 0\}$ a submartingale.

The *Doob–Meyer decomposition theorem* states that, for any submartingale $\{K(t) : t \geq 0\}$, there exists a unique predictable increasing process $\{\Lambda(t) : t \geq 0\}$, called a *compensator*, such that $\Lambda(0) = 0$, a.s., $E[\Lambda(t)] < \infty$ for any t, and $\{M(t) = K(t) - \Lambda(t) : t > 0\}$ is a martingale. The proof of this theorem is beyond the scope of this book.

The trend of a submartingale tends to increase over time. The trend of a martingale tends to be constant over time. A consequence is $E\,\{M(t)|\mathcal{H}_0\} = M(0)$, a.s. for all $t \geq 0$. When $M(0) \equiv 0$, we have $E\,\{M(t)\} = 0$, a.s. for every t. In this case, we call $\{M(t) : t \geq 0\}$ a zero-mean martingale.

Variance and Covariance Processes for Zero-Mean Martingales

For a zero-mean martingale $\{M(t) : t \geq 0\}$, the variation process is a stochastic process $\{\langle M\rangle\,(t) : t \geq 0\}$ which makes $\left\{M^2(t) - \langle M\rangle\,(t)\right\}$ a zero-mean martingale. For two zero-mean martingales $\{M_1(t)\}$ and $\{M_2(t)\}$, the covariate process is a stochastic process $\{\langle M_1, M_2\rangle\,(t) : t \geq 0\}$ which makes $\{M_1(t)M_2(t) - \langle M_1, M_2\rangle\,(t)(t)\}$ a zero-mean martingale.

The $\int HdM$ Martingale Transform

The $\int HdM$ martingale transform is a useful tool. We refer the readers to Fleming and Harrington (1991) and Andersen et al. (1993) for the theory. Without presenting the proof, for a martingale of the form $\{M(t) = K(t) - \Lambda(t) : t \geq 0\}$ satisfying $E[K(t)] < \infty$ for all $t < \infty$, let $\{H(t) : t \geq 0\}$ be a bounded predictable process, then the process given by $\int_0^t H(x)dM(x)$ is a martingale. Here the integral $\int_s^t fdM$ represents the Stieltjes integration of the sum of the values of f at jump times of $M(x)$ in the interval $[s, t)$. We shall review this later in Chap. 7.

3.4 Problems and Supplements

3.1 We assume a very large population setting so that each individual has equal chance to make contacts with everyone else. The contact rate is λ, that is, from the perspective of an individual, the time to the next contact with another individual is exponentially distributed with rate λ.

(a) Let $(t, t + x]$ be a time interval of length x. Let $N(t, t + x)$ be the total number of contacts made by a typical individual. Show that the distribution of $N(t, t + x)$ only depends on the length x and is identical to the distribution of $N(0, x) \equiv N(x)$; and show that $N(x)$ follows a Poisson distribution given by (3.9) with rate $\theta = \lambda x$.

(b) Let $p \in (0, 1]$ be the probability that the contact is between a pair of susceptible–infectious individuals and during the contact, a transmission occurs. Such a contact is called an infectious contact. Show that, the total number of infectious contacts produced by a typical infected individual during $(t, t+x]$ also follows a Poisson distribution (3.9) with rate $\theta = \lambda px$.

3.2 The set of conditions given in 3.1 is called "homogeneous mixing." Assuming that an infected individual is seeded into an infinitely large susceptible population at time zero. This individual is associated with a random infectious period T_I. Under homogeneous mixing, the number of infectious contacts produced by this individual during the time interval $(0, t]$, denoted by $K(t)$, follows a Poisson distribution with rate $\theta = \lambda pt$, provided that $t < T_I$. The number of infectious contacts produced by this individual during the time interval $(0, t]$ in general is denoted by $N(t)$ such that

$$N(t) = \begin{cases} K(t), & 0 < t < T_I \\ K(T_I), & T_I \leq t. \end{cases} \tag{3.42}$$

We let $N = N(\infty)$ be the cumulative number infectious contacts generated by this infectious individual throughout the entire infectious period.

(a) Show that, p.g.f. of N, $G_N(s)$, is given by (3.23), where $f_I(x)$ is the p.d.f. of the infectious period.
(b) Let $\mu_I = E[T_I]$ be the mean infectious period, show that $E[N] = \lambda p \mu_I$, regardless of the exact distribution of the infectious period as long as μ_I exists.
(c) If T_I is exponentially distributed with hazard rate γ, show that N follows a geometric distribution

$$\Pr(N = n) = \left(\frac{\lambda p}{\lambda p + \gamma}\right)^n \frac{\gamma}{\lambda p + \gamma}.$$

(d) If T_I is gamma distributed with $E[T_I] = \mu_I$ and $var[T_I] = \mu_I^2/\kappa$, what is the distribution of N? Write down its variance.

3.3 Consider the probability generating functions defined by (3.3).

(a) Cross validate that, if the distribution of N is not degenerated to a point mass, that is, there is no $n \geq 0$ such that $\Pr(N = n) = 1$, then the p.g.f. is strictly increasing for $s \in [0, 1]$ and is strictly convex, satisfying (3.4).
(b) Cross validate that the mean and the variance of N, if exist, can be expressed by (3.6).
(c) Express the p.g.f. of the negative binomial distribution as a function of μ and κ, where $\mu = E[N]$ corresponds to (3.20) as well as the p.g.f. when $\kappa \to \infty$. Assuming $\mu = 3$, plot the p.d.f. with $\kappa = 0.5, 1, 2$, and the limiting case $\kappa \to \infty$.
(d) Keeping μ fixed, comment on how κ ranks the variance, the p.g.f., and the probability $f_0 = \Pr(N = 0)$.

3.4 Using the convex order to compare variability in Chap. 2 also applies to random counts. Keeping $\mu = E[N]$ fixed, the variability of N can be ranked according to the convex order. We say N_2 is *more dispersed than* N_1 if $E[N_1] = E[N_2]$ and $E[\Psi(N_1)] \leq E[\Psi(N_2)]$ for all convex functions $\Psi(x)$ for which these expectations exist.

(a) How does this variability order with respect to N rank the p.d.f. $G_N(s)$ for $s \in [0, 1]$?
(b) Consider the p.g.f. given by (3.23), how does the variability of the infectious period T_I, according to convex order, rank the p.d.f. $G_N(s)$ for $s \in [0, 1]$?

3.5 (The order statistics structure of a counting process) Let us consider a counting process $\{K(t)\}_{t=0}^{\infty}$ so that conditioning on $\{K(t) = k\}$, the k arrival times $y_1, \ldots, y_k$ are distributed like order statistics of independent samples arising from an identical distribution $G(y|t)$. Such a counting process is said to have an *order statistics structure*. It generates new events independently and with identical distribution for the time to events.

(a) Show that the time-homogeneous Poisson process in Definition 11 has an order statistics structure with the joint p.d.f. as

$$f(y_1, y_2, \cdots y_k | K(t) = k) = k! \prod_{i=1}^{k} g(y_i|t) = \frac{k!}{t^k}, \ 0 < y_1 < y_2 < \cdots < y_k \leq t,$$

where $g(y_i|t) = t^{-1}$. In other words, given that $K(t) = k$, the k arrival times of the events have the same distribution as the order statistics corresponding to k independent random variables uniformly distributed on the time interval $(0, t)$.

(b) (The non-homogeneous Poisson process) The process $\{K(t) : t \geq 0\}$ is a non-homogeneous Poisson process of intensity $\beta(t)$, if for small $h > 0$, all $t > 0$, and , the following conditions hold:

(i) $K(0) = 0$;
(ii) $\Pr\{K(t) - K(t-h) = 1 | \mathcal{H}_{t-h}\} = \beta(t)h + o(h)$; (Markov property)
(iii) $\Pr\{K(t) - K(t-h) > 1 | \mathcal{H}_{t-h}\} = o(h)$.

(c) Show that the cumulative number of events in the time interval $(s, s+t]$ follows the Poisson distribution

$$\Pr\{K(s, s+t] = k\} = \frac{\left(\int_s^{s+t} \beta(u)du\right)^k}{k!} e^{-\int_s^{s+t} \beta(u)du}.$$

(d) Show that the non-homogeneous Poisson process also have order statistics structure.

$$f(y_1, y_2, \cdots y_k | K(t) = k) = k! \prod_{i=1}^{k} \frac{\beta(y_i)}{B(t)}, \ 0 < y_1 < \cdots < y_k \leq \tau,$$

where $B(t) = \int_0^t \beta(u)du$. This is the distribution of the order statistic in a sample of size k from the density $g(y|t) = \frac{\beta(y)}{B(t)}$, $y \leq \tau$.

(e) If a process has the order statistic structure, such a process must be a Markov chain (Crump 1975). Now consider the counting process $\{N(t)\}_{t=0}^{\infty}$, where $N(t)$ is given in (3.42) in Problem 3.2. Let $t_m < s$ be the time when the mth infection took place and no infection in (t_m, s), what factors does the probability that, *this infected individual is able to infect others at time s*, depend on? Is the process $\{N(t)\}_{t=0}^{\infty}$ a Markov chain?

(f) For a typical infected individual, if the numbers of infectious contacts $\{K(t)\}_{t=0}^{\infty}$ follow a Poisson process (homogeneous mixing), but there is also a random infectious period T_I acting as a stopping time that gives rise to (3.42), is it appropriate to consider that the times to transmission of the pathogen to its infectees as an i.i.d. sample of a lifetime distribution?

3.6 (Incidence, prevalence, and convolutions) Recall the intensity process (3.37) associated with a counting process $\{C(t) : t \geq 0\}$. When $C(t)$ represents the cumulative number of infections in the population by time t, we call $\lambda(t)$ as the *incidence intensity*. If $\lambda(t)$ is a deterministic function, $\int_s^t \lambda(u)du$ represents the expected number of new infectious in the population during the time interval $(s, t]$. We assume that there is a random duration X, with p.d.f. $f_X(x)$ and survival function $\overline{F}_X(x)$, and each infected individual must go through such a duration that leads to a subsequent event (e.g., diagnosis, becoming infectious, death, etc.). The following two convolutions are valid:

$$a(t) = \int_0^t \lambda(u) f_X(t-u)du,$$
$$\Pi(t) = \int_0^t \lambda(u)\overline{F}_X(t-u)du.$$

(a) Give the meanings for $a(t)$ and $\Pi(t)$ in the following situations:

(i) the subsequent event is the onset of symptoms;
(ii) the subsequent event is the diagnosis of the infection;
(iii) the subsequent event is that the infected individual becomes infectious;
(iv) the subsequent event is death.

(b) Let $\mu = E[X] = \int_0^{\infty} \overline{F}_X(t)dt$, under what conditions that $\Pi(t)$ is constant and proportional to μ? (and comment on the statement in many introductory epidemiology textbooks that "prevalence = incidence × duration.")

(c) Assume $\lambda(t)$ is a deterministic function and for practical purposes, we consider discrete time points $t = 1, 2, \ldots$ and $i(t) = \int_{t-1}^t \lambda(u)du$ is the expected number of new infections in the time interval $(t-1, t]$. The duration X is the incubation period, defined as the time from infection to clinical onset. Let $f_X(x) = \overline{F}_X(x-1) - \overline{F}_X(x)$, $x = 1, 2, \ldots$ and we approximate $a(t) = \int_0^t \lambda(u) f_X(t-u)du$ by a discrete time model $a(t) = \sum_{u=0}^t i(u) f_X(t-u)$.

(i) What is the meaning of $a(t)$ in this convolution?
(ii) Let Y_t be the number of clinical onsets during the time interval $(t-1, t]$ and follows a Poisson distribution. Write probability $\Pr(Y_t = y_t)$.

(iii) Assume that $\lambda(t)$ is specified through an unknown parameter θ so that $i(t)$ is expressed as $i(t;\theta)$, whereas $f_X(x)$ does not involve any unknown parameters. We assume that $(Y_t,\ t = 1, 2, \ldots, l)$ are independently distributed and data are observed as $(y_t,\ t = 1, 2, \ldots, l)$. Write the log-likelihood function of θ based on the observations.

(d) In some infectious disease models, $\lambda(t)$ is a stochastic process given by $\lambda(t) = \beta(t)I(t)$, where $\beta(t)$ is a deterministic function of time t and $I(t)$ is the (stochastic) number of individuals who are infectious at time t; meanwhile, X is the infectious period. Examine if the following expression is true:

$$E[I(t)] = \int_0^t \beta(t-x)E[I(t-x)]\overline{F}_X(x)dx.$$

Chapter 4
Behaviors of a Disease Outbreak During the Initial Phase and the Branching Process Approximation

We consider that at the beginning, $t = 0$, there is no disease. We call the system at this condition *the disease-free equilibrium*. We assume that the entire population is susceptible. The size of the susceptible population is denoted by m.

We then seed an initial number of i_0 infectious individuals into this population. When transmission starts, we say that the system is moving away from the equilibrium condition.

We assume that there exists a period of time during which the depletion of the number of susceptible individuals in the population is negligible. We call this period the initial phase of an outbreak. This assumption requires m to be very large and i_0 to be very small. For mathematical convenience, the susceptible population size is approximated by $m \to \infty$.

This chapter studies the behaviors of a disease outbreak during the initial phase. The first part of the chapter shows that regardless of how contagious the disease is, there is always a positive probability δ that the outbreak becomes extinct in a few generations and the system returns to disease-free equilibrium. The expected final number of infected individuals is independent of the population size m. The second part focuses on the condition that, if $1 - \delta > 0$, the expected number of cumulatively infected individuals grows very large, scaled by the size of the susceptible population m. Should that happen, we discuss the properties of the initial growth of the epidemic over generations and in real time during the initial phase.

4.1 The Branching Process Approximation

Discrete and continuous time branching processes are used in infectious disease epidemiology to characterize the initial stage of an outbreak.

P. Yan, G. Chowell, *Quantitative Methods for Investigating Infectious Disease Outbreaks*, Texts in Applied Mathematics 70,
https://doi.org/10.1007/978-3-030-21923-9_4

4.1.1 The Galton-Watson Branching Process

In an outbreak investigation, data from contact tracing give a branching process representation. The basic reproduction number, R_0, is a key quantity in the field of infectious disease epidemiology (Anderson and May 1991; Diekmann and Heesterbeek 2000; Brauer 2006; Castillo-Chavez et al. 2002). It characterizes the early spread of an outbreak in terms of the average number of secondary infections produced by a typical infected individual during the initial phase. Figure 4.1 displays transmission trees for historic outbreaks of smallpox, Ebola, SARS, and MERS and at the same time highlights the role of confined settings (e.g., hospitals) in the spread of infectious diseases. Because it is often difficult to reconstruct the transmission chains of infectious epidemics, we draw connections between the branching process that dictates the generation of secondary cases (Nishiura et al. 2012) and the continuous time models of epidemic growth with a particular attention at non-exponential growth processes and models that support them.

The Galton-Watson branching process is a discrete time branching process. Consider a population consisting of individuals that are able to produce offsprings. Let X_0 be the number of individuals initially present and form Generation Zero. A typical individual i produces a random number N_i of new offsprings. We assume that N_i are independently and identically distributed as N with mean value $R_0 = E[N]$.

The assumption of independency allows us to simplify the assumption that $i_0 = X_0 = 1$ without losing generality.

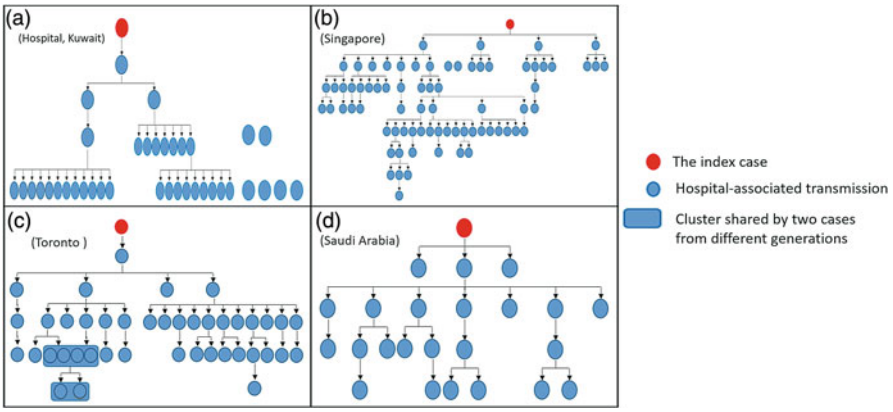


Fig. 4.1 Transmission trees for four infectious disease outbreaks: (**a**) Smallpox in a hospital in Kuwait in 1967 (Arita et al. 1970), (**b**) Ebola outbreak in Nigeria, 2014, (**c**) SARS in Toronto, 2003 (Varia et al. 2003; Chowell et al. 2015), and (**d**) MERS in Al-Hasa, Saudi Arabia, 2013 (Assiri et al. 2013)

All offsprings of Generation Zero constitute the first generation and their number is denoted by X_1. Let X_g denote the size of the gth generation. Suppose that $X_0 = 1$, we can calculate

$$X_g = \sum_{i=0}^{X_{g-1}} N_i. \tag{4.1}$$

We say a branching process becomes extinct at generation g, if $X_{g-1} > 0$ but $X_g = 0$. The final size of the branching process upon extinction is $\sum_g X_g$.

When used to approximate the behavior of an outbreak at its initial phase, it is not only assumed that N_i are i.i.d. but also that the distribution of N does not change over generations. Under such assumptions, a typical infected individual produces on average $R_0 = E[N]$ new infections to make its next generation. The parameter R_0, defined at Generation Zero when the system is at (disease-free) equilibrium, is called the basic reproduction number. It is applied to the first few generations during the initial phase.

This process $\{X_g : g = 0, 1, 2, \cdots\}$ has the Markov property, that is, the distribution of X_g only depends on X_{g-1}. All the variables in this process are random counts. Using terminology regarding disease transmission, they are:

1. N : the number of secondary infections produced by a typical infected individual during the initial phase.
2. M_g : the number of generations until extinction.
3. X_g : the size of infected individuals of the gth generation.
4. Z_s : the final size of an outbreak upon extinction.

The Probability Generation Function for N

The probability generating function for N, denoted by $G_N(s) = E[s^N] = \sum_{j=0}^{\infty} s^j \Pr\{N = j\}$, will be used extensively in the study of the distributions for the random counts 1.–4. listed above. This is largely due to the fact that most of the random counts represented by the Galton-Watson branching process are random sums so that recursive formulas can be easily established based on $G_N(s)$. For instance, it is well known in probability theory (Exercise 4.1) that, for i.i.d. random counts N_i with p.g.f. $G_N(s)$, the p.g.f. of the random sum $N_1 + N_2 + \cdots + N_X$ is $G_X(G_N(s))$, where $G_X(s)$ is the p.g.f. of the random integer X. Therefore the p.g.f. for the random sum given by (4.1) is $G_{X_g}(s) = G_{X_{g-1}}(G_N(s))$.

By definition, $G_N(0) = \Pr\{N = 0\}$, $G_N(1) = \sum_{j=0}^{\infty} \Pr\{N = j\} = 1$. In general, if the offspring distribution is not a point mass, $G_N(s)$ is an increasing and strictly convex function in $s \in [0, 1]$. When the first moment of N exists, the basic reproduction number R_0 is the expected value

$$R_0 = E[N] = G'_N(1). \tag{4.2}$$

We assume that the second moment exists, and call $var[N]$ the basic reproduction variance (Haccou et al. 2005). It can be expressed as

$$var[N] = G''_N(1) + G'_N(1) - \left(G'_N(1)\right)^2.$$

4.1.2 *Embedding the Galton-Watson Branching Process into a Continuous Time Framework*

In the continuous time framework, a counting process $\{K(x) : x \geq 0\}$ is defined to track the cumulative number of new infections of a typical infected individual where $x = 0$ is the time at infection. Each individual is associated with an infectious period $T_I > 0$, which is a continuous random variable. Infections produced by such an individual only occur during T_I. The number of infectious contacts produced by this individual during the entire infectious period is $N = K(T_I)$.

The Galton-Watson process is a discrete time branching process characterized by the marginal distribution of N, along with the distributions of other random counts such as M_g, X_g, and Z_s, without any attention on the stochastic mechanisms regarding the counting process $\{K(x) : x \geq 0\}$ and the distributions of the latent period (during which the infected individual is not able to transmit to other susceptible individuals through contact) and the infectious period. All branching processes in the continuous framework have an embedded Galton-Watson branching process defined by the marginal distributions of the random counts to track the generations. This chapter will demonstrate that some behaviors of the early phase of the outbreak are determined by the properties of these marginal random count distributions, whereas other behaviors will be dependent on the properties of the counting process $\{K(x) : x \geq 0\}$ and the distributions of the latent and the infectious periods.

The Crump-Mode-Jagers (CMJ) Branching Process

The Crump-Mode-Jagers (CMJ) branching process has been used to approximate the early phase of the SIR models (Barttlet 1961; Mode and Sleeman 2000). It assumes that each infected individual is immediately infectious upon infection (i.e., without latent period). It also assumes that $\{K(x) : x \geq 0\}$ and the infectious T_I are independent, and that infected individuals have i.i.d. T_I specified by an arbitrary distribution.

A special case of the CMJ branching process used to approximate the early phase of the epidemic is the assumption that $\{K(x) : x \geq 0\}$ is a stationary Poisson process with intensity β for the infectious contact process. The infectious period $T_I > 0$ is a continuous random variable serving as a stopping time of the counting process. In many simple disease transmission models, T_I is often assumed to be exponentially distributed.

The Probability Generation Function for N in a CMJ Process

We considered the CMJ branching process in which $\{K(x) : x \geq 0\}$ is a stationary Poisson process with intensity β for the infectious contact process. During any time interval of length x within the infectious period, $\Pr\{K(x) = k\} = \frac{(\beta x)^k}{k!} e^{-\beta x}$ with p.g.f.

$$G_K(s; x) = e^{-\beta x(1-s)}. \tag{4.3}$$

It has been shown (Mode and Sleeman 2000) that the p.g.f. for N can be expressed by

$$G_N(s) = \int_0^\infty G_K(s, x) dF_I(x) = \int_0^\infty e^{-\beta x(1-s)} dF_I(x) = L[f_I](\beta(1-s)), \tag{4.4}$$

where $F_I(x)$ is the cumulative distribution of the infectious period and $L[f_I]$ is the Laplace transform of the infectious period T_I.

Extensions of the CMJ Branching Process

The CMJ branching process can be extended in several ways. For instance, one way is illustrated in (b) of Fig. 4.2 in which a random latent period T_E is added to the CMJ process. The usual assumption is that the latent period T_E is independent from the infectious period T_I and both distributions can be arbitrary but specified.

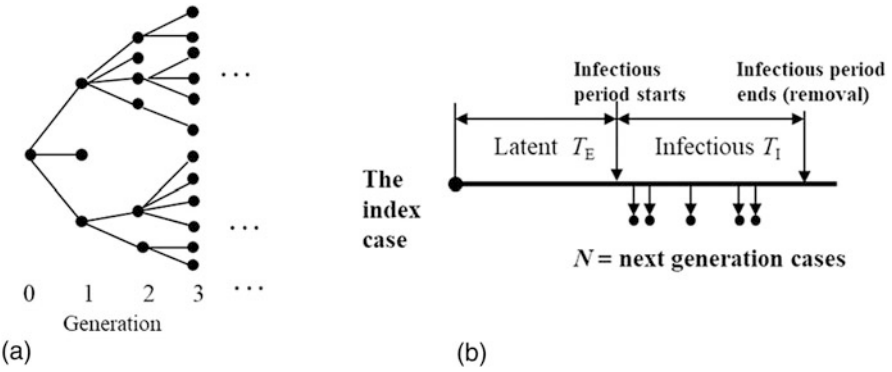


Fig. 4.2 Illustration of the Galton-Watson branching process and the continuous time branching process starting with a single individual. (**a**) The embedded Galton-Watson process in three generations. (**b**) The continuous time branching process in one generation

Another extension is to allow $\{K(x) : x \geq 0\}$ to arise from a counting process other than the Poisson process. For example, infected individuals may have different infectious contact rates. As a result, at the cohort level, the infectious contact process arises as a mixed-Poisson process with marginal distributions given by the mixed-Poisson distribution

$$Poisson(\beta\mu_I) \underset{\beta}{\wedge} u(\beta).$$

Alternatively, $\{K(x) : x \geq 0\}$ may arise as a linear pure birth process (Bartlett 1955; Bhattacharya and Waymire 1990; Allen 2010).

The counting process may be replaced by dynamic social contact networks. In infectious disease models, individuals are represented by vertices, and contacts are represented by edges. Social contacts made over a fixed period of time may be modeled by a random graph. As a function of time, social contacts can be regarded as a random graph process. An infectious contact is a contact at which a transmission of infection takes place (Dietz 1995). All infectious contacts during the same period make a subgraph. The geometry of this subgraph is different from the graph that represents social contacts. If three individuals $\{a, b, c\}$ are friends forming a triangle relationship and if individual a infects both individuals $\{b, c\}$, then b and c do not infect each other. This description determines that the subgraph is directional, grows along a *tree* that resembles a realization of an embedded Galton-Watson branching process.

4.2 Extinction and the Invasion Probability

Starting from $i_0 = X_0 = 1$, except for the degenerated distribution $\Pr\{N = 1\} = 1$, the invasion probability is $1 - \delta$ and can be calculated such that δ is the smallest root of the fixed point equation $G_N(s) = s$ in $s \in (0, 1]$. This is a well-established result in probability theory. For a rigorous proof, we refer to classic textbooks such as Chap. 8 of Karlin and Taylor (1975), Jagers (1975), Haccou et al. (2005) among others. Instead, we jointly describe the probability distribution of generations to extinction and the invasion probability as illustrated in Fig. 4.3.

The generations to extinction M_g is defined by the event $\{M_g = g\}$, referring to {no infected case at generation $g + 1$ and at least one infected case at generation g}. We refer to the initially seeded infectious individuals as generation zero. M_g takes values $g = 0, 1, 2, \ldots$. The number of initially seeded infectious individuals is i_0. Given i_0, the distribution of M_g is determined by the distribution of N, the number of secondary infections produced by a typical infected individual, under the assumption that the distribution of N does not change over generations and that infected individuals produce secondary infections independently. For mathematical convenience, we assume $i_0 = 1$.

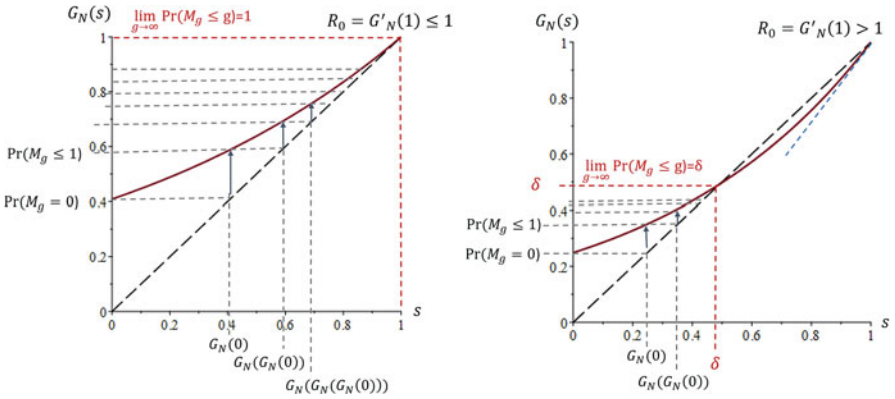


Fig. 4.3 Graphic illustration of the recursive formulae (4.5)

First, $\Pr\{M_g = 0\} = \Pr\{N = 0\} = G_N(0)$. This is the case when there is no secondary transmission. Assuming that the seeded individual in Generation Zero gives $j = 1, 2, \cdots$ new infected individuals for the first generation, and these individuals produce infections independently, then the event $\{M_g \le 1\}$ takes place only if none of the individuals in the first generation produces new infections to form another generation. The probability is $[G_N(0)]^j \Pr\{N = j\}$ which is the jth term of $G_N(G_N(0))$. Therefore $\Pr\{M_g \le 1\} = G_N(G_N(0))$. Continuing by induction, we arrive at a recursive formula for the event $\{M_g \le g\}$ given by

$$\Pr\{M_g \le g\} = G_N^g(0) \overset{\text{def.}}{=} \underbrace{G_N(G_N(\cdots G_N}_{g+1 \text{ times}}(0) \underbrace{\cdots)}_{g+1 \text{ times}}, \quad g = 0, 1, 2, \ldots \tag{4.5}$$

Because $G_N(s)$ is a convex function in $s \in (0, 1]$ (Fig. 4.3), it is clear that the jump

$$\Pr\{M_g = 1\} = G_N(G_N(0)) - G_N(0) < \Pr\{M_g = 0\}.$$

Once more by induction, the sequence $\Pr\{M_g = g\}$, $g = 0, 1, 2, \ldots$ is decreasing and approaching zero. However, these probabilities may not properly define a distribution. The cumulative probability $\Pr\{M_g \le g\}$ is a non-decreasing function of g. Starting from $\Pr\{M_g = 1\} = G_N(0)$, $\Pr\{M_g \le g\}$ has the limit

$$\lim_{g\to\infty} \Pr\{M_g \le g\} = \delta \le 1,$$

where δ is the smallest root of the fixed point equation $G_N(s) = s$ in $s \in (0, 1]$ (Fig. 4.3).

4.2.1 The Effects of Variability of N on the Invasion Probability $1-\delta$ and Generations Toward Extinction

The description of dispersion of two lifetime X_1 and X_2 with equal mean value (should they exist), as shown in Fig. 2.10 in Chap. 2, also applies to random counts to rank their variability. Let f and g be the p.m.f. for N_1 and N_2 with corresponding survival functions $\overline{F}_n$ and $\overline{G}_n$, and $E[N_1] = E[N_2]$. We say that N_2 *is more dispersed (spread out) than* N_1 if

1. $g_n - f_n$ has two sign changes and the sign sequence is: $+, -, +$.
2. $\overline{F}_n - \overline{G}_n$ has one sign change and the sign sequence is: $+, -$.

These two statements are equivalent to say that N_2 is *more dispersed than* N_1 if $E[N_1] = E[N_2]$ and $E[\Psi(N_1)] \leq E[\Psi(N_2)]$ for all convex functions $\Psi(x)$ for which these expectations exist. Applying to the number of secondary infections produced by a typical infected individual, keeping the mean value $R_0 = E[N]$ fixed, the convex order gives order for the probability generating function $G_N(s) = E[s^N]$ since s^N is a convex function of N. It also implies the ordering according to variance $var(N) = E[(N - R_0)^2]$ because x^2 is a convex function. Therefore, the more dispersed the N, the larger the variance $var[N]$ and also larger the value of $G_N(s)$ for any $s \in [0, 1]$. Figure 4.4 illustrates that, among the distributions with equal mean $R_0 = 3$, the geometric distribution ($\kappa = 1$) is more dispersed than the Poisson distribution ($\kappa \to \infty$) and the shape parameter κ of the negative binomial distribution κ ranks the probability generating function. The shape parameter κ also ranks the variance: $var[N] = R_0 + R_0^2/\kappa$. The left panel in Fig. 4.4 shows the two sign changes of the p.m.f. f_n, and the right panel shows the ranks of the p.g.f. $G_N(s)$.

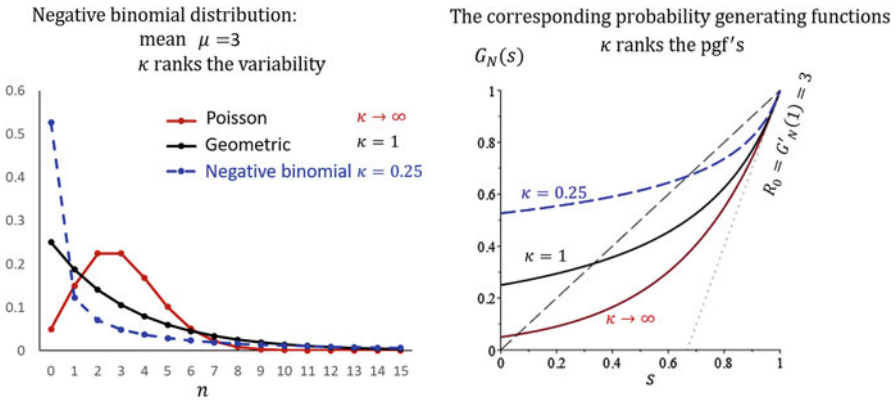


Fig. 4.4 Illustration of the dispersion of the negative binomial distribution given $R_0 = 3$ by comparing the probability mass function and $G_N(s)$

Summary Statements on Invasion Probability and Generations to Extinction

The threshold condition is $R_0 = 1$.

1. When $R_0 = G'_N(1) \leq 1$, the smallest root is $\delta = 1$ and the invasion probability is $1 - \delta = 0$. $\Pr\{M_g \leq g\}$ is a properly defined probability distribution. The mode of the distribution $\Pr\{M_g = g\}$ is $g = 0$. Since convex order gives order for the probability generating function $G_N(s)$,the more dispersed the distribution, the larger that value $\Pr\{M_g = 0\} = G_N(0)$. Figure 4.5 implies that the more dispersed N, the more likely extinction will happen quickly within fewer generations.
2. When $R_0 < 1$, the distribution for M_g has an exponentially decaying tail, that is, there exists a positive number C such that

$$\Pr\{M_g > g\} = 1 - G_N^g(0) \sim C R_0^g, \text{ as } g \to \infty,$$

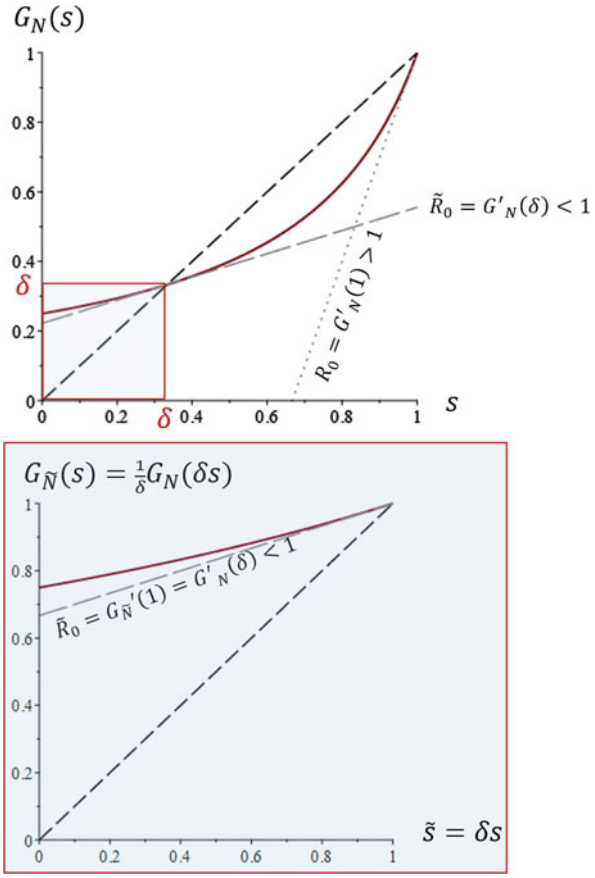


Fig. 4.5 Illustration of $G'_N(\delta) < 1$ when $R_0 = G'_N(1) > 1$

where $G_N^g(0)$ is defined by (4.5) and the expectation $E[M_g]$ is finite. Regarding the precise properties of this distribution, there has been little attention in the literature. Some approximate but useful results are available. If the number of initial seeded infectious individuals i_0 is large, the expected generation number to extinction can be approximated by Haccou et al. (2005)

$$E[M_g] \sim \frac{\log i_0}{|\log R_0|}, \ i_0 \to \infty.$$

3. If $R_0 = 1$, $\Pr\{M_g \leq g\}$ is a properly defined distribution. However,

$$\Pr\{M_g > g\} \sim \frac{2}{\sigma^2 g}, \text{ as } g \to \infty,$$

where $\sigma^2 = var[N]$, assuming $var[N] < \infty$. In this case, M_g has infinite mean.

4. When $R_0 = G'_N(1) > 1$, there is a unique solution δ in the open interval $(0, 1)$ and the invasion probability is $1 - \delta > 0$. In this case, $\Pr\{M_g \leq g\}$ is not a properly defined distribution. However, the normalized distribution $\Pr\{\widetilde{M}_g \leq g\} = \frac{1}{\delta} G_N^g(0)$ is meaningful. We denote $\widetilde{N}$ as the random variable that is distributed according to the conditional distribution of N, given the outcome is not a large outbreak (with probability $\delta > 0$). Waugh (1958) shows that the probability generating function of $\widetilde{N}$ is given by

$$G_{\widetilde{N}}(s) = \frac{1}{\delta} G_N(\delta s). \tag{4.6}$$

It turns out that $E[\widetilde{N}] = G'_N(\delta)$, $var[\widetilde{N}] = \delta G''_N(\delta) + G'_N(\delta) - G'_N(\delta)^2$. It is always true that $\widetilde{R}_0 = G'_{\widetilde{N}}(1) = G'_N(\delta) < 1$ (see Fig. 4.5). The notation $\widetilde{M}_g$ corresponds to the time to extinction, conditioning on the outcome being a small outbreak. The probability $\widetilde{M}_g > g$ is

$$1 - \frac{1}{\delta} G_N^g(0) = \frac{\delta - G_N^g(0)}{\delta}.$$

It can be shown that for a suitable positive constant C,

$$\delta - G_N^g(0) \sim C \left[G'_N(\delta)\right]^g, \text{ as } g \to \infty.$$

and $0 < G'_N(\delta) < 1$. $\widetilde{M}_g$ also has an exponentially decaying tail.

4.2.2 When N Follows the Power Series Distributions

The power series distributions given by (3.7) in Chap. 3 have p.g.f. with the form $G_N(s) = A(s\theta)/A(\theta)$ where θ is the canonical parameter. These distributions include the Poisson distribution with $A(\theta) = e^{\theta}$, $\theta > 0$; the geometric distribution with $A(\theta) = (1-\theta)^{-1}$, $0 < \theta < 1$; the negative binomial distribution with $A(\theta) = (1-\theta)^{-\kappa}$, $\kappa > 0$ and $0 < \theta < 1$; the logarithmic distribution with $A(\theta) = -\log(1-\theta)$, $0 < \theta < 1$. The fixed point equation $G_N(\delta) = \delta$ becomes

$$A(\delta\theta) = \delta A(\theta). \tag{4.7}$$

Conditioning on the final outcome not being a large outbreak, data arise from the distribution of $\widetilde{N}$ with p.g.f.

$$G_{\widetilde{N}}(s) = G_N(\delta s)/\delta = A(\delta s\theta)/\delta A(\theta) = \frac{A(s\theta^*)}{A(\theta^*)} \tag{4.8}$$

where $\theta^* = \delta\theta$. This result directly follows (4.6), $G_N(s) = A(s\theta)/A(\theta)$ and (4.7). It implies that $\widetilde{N}$ and N belong to the same class of distributions within the power series distribution family with a re-scaled canonical parameter θ^*.

The Geometric Distribution for *N*

The geometric distribution has a very distinct role in infectious disease models. Many compartmental models associated with exponentially distributed infectious period give rise to a geometrically distributed random number N representing the number of secondary infections produced by a typical infected individual with $R_0 = E[N]$. This distribution is a member of the power series distribution family with $A(\theta) = (1-\theta)^{-1}$. The p.g.f. is $G_N^{\text{Geo}}(s) = \frac{1-\theta}{1-s\theta}$ and $R_0 = \frac{\theta}{1-\theta}$. The variance of the geometric distribution is $var[N] = R_0 + R_0^2$. The p.g.f. can be alternatively written as

$$G_N^{\text{Geo}}(s) = \frac{1}{1 + R_0(1-s)}. \tag{4.9}$$

The smallest root of the fixed-point equation $G_N^{\text{Geo}}(s) = s$ for $s \in (0, 1]$ is $\delta = \min\{1, 1/R_0\}$. If $R_0 > 1$, the conditional distribution of N given the outcome being a small outbreak is also a geometric distribution with mean value $\widetilde{R}_0 = G'_N(\delta) = 1/R_0$ and variance $var[\widetilde{N}] = \widetilde{R}_0 + \widetilde{R}_0^2 = (R_0+1)/R_0^2$ along with p.g.f.

$$G_{\widetilde{N}}^{\text{Geo}}(s) = \frac{R_0}{R_0 + 1 - s},$$

which could be either viewed as $\frac{1}{1+\widetilde{R}_0\ (1-s)}$ or as $\frac{1}{\delta[1+R_0(1-\delta s)]}$. With respect to M_g, it can be shown that $\Pr\{M_g = 0\} = \Pr\{N = 0\} = (1 + R_0)^{-1}$

$$\Pr\{M_g \le g\} = \frac{R_0^{g+1} - 1}{R_0^{g+2} - 1} \to \begin{cases} 1, & R_0 < 1 \\ \delta = \frac{1}{R_0}, & R_0 > 1 \end{cases}, \quad \text{as } g \to \infty. \tag{4.10}$$

In the case $R_0 > 1$,

$$\Pr\{\widetilde{M}_g \le g\} = \frac{1}{\delta}\Pr\{M_g \le g\} = \frac{R_0(R_0^{g+1} - 1)}{R_0^{g+2} - 1} \tag{4.11}$$

$$= \frac{\widetilde{R}_0^{g+1} - 1}{\widetilde{R}_0^{g+2} - 1} \to 1, \quad \text{as } g \to \infty.$$

Therefore, in an observed small outbreak, the distribution of the time-to-extinction $\widetilde{M}_g$ follows the form $(\mu^{g+1}-1)/(\mu^{g+2}-1)$ with $\mu < 1$. In the absence of additional information, one cannot distinguish $R_0 = \mu$ from $R_0 = \mu^{-1}$.

The Poisson Distribution for N

The Poisson distribution is a member of the power series distribution family with $A(\theta) = e^{\theta}$. The number of secondary infections produced by a typical infected individual N may arise from a Poisson distribution when the infectious contact process is a Poisson process and there is no variation in the infectious period. The probability generating function is $G_N^{\text{Pois}}(s) = \exp\{-\theta(1 - s)\}$ and mean $R_0 = \theta$. Therefore it is commonly written as

$$G_N^{\text{Pois}}(s) = \exp(-R_0(1 - s)). \tag{4.12}$$

Unlike the geometric distribution, there is no analytic close form for the smallest root of the fixed-point equation $G_N^{\text{Pois}}(s) = s$ for $s \in (0, 1]$ when $R_0 > 1$. In fact, δ can be expressed as $\delta = -R_0^{-1}\,\text{LambertW}\left(-R_0 e^{-R_0}\right)$, where the Lambert W function is a special function (Corless et al. 1996) defined as the inverse function of ze^z, satisfying $W(z) = ze^{-W(z)}$. Letting $z = -R_0 e^{-R_0}$, $\text{LambertW}\left(-R_0 e^{-R_0}\right)$ is defined for $R_0 \ge 1$. It is an increasing function of R_0, starts from $\text{LambertW}\left(-e^{-1}\right) = -1$ and converges to zero as $R_0 \to \infty$. Numerical computation for $W(z)$ can be done through many commercially available software, such as Maple. It can be shown that $\delta < 1/R_0$ because

$$\exp(-R_0(1 - s)) < \frac{1}{1 + R_0(1 - s)}, 0 < s < 1.$$

From (4.8), it follows that

$$G_{\widetilde{N}}^{\text{Pois}}(s) = \exp\left\{-\theta^*(1-s)\right\} = \exp(-\delta R_0(1-s))$$

and hence if $R_0 > 1$, the conditional distribution of N given the outcome being a small outbreak is also a Poisson distribution with mean value

$$\widetilde{R}_0 = \delta R_0 = -\,\text{LambertW}\left(-R_0 e^{-R_0}\right) < 1.$$

With respect to M_g, we get the following recursive calculation:

$$\begin{aligned}
&\Pr\{M_g = 0\} = \Pr\{N = 0\} = e^{-R_0},\\
&\Pr\{M_g \le 1\} = e^{-R_0} e^{R_0 e^{-R_0}}.\\
&\Pr\{M_g \le 2\} = e^{-R_0}\left(e^{R_0 e^{-R_0}}\right)^{e^{R_0 e^{-R_0}}},\\
&\Pr\{M_g \le 3\} = e^{-R_0}\left(e^{R_0 e^{-R_0}}\right)^{\left(e^{R_0 e^{-R_0}}\right)^{e^{R_0 e^{-R_0}}}},\\
&\qquad\vdots\\
&\Pr\{M_g \le g\} = e^{-R_0}\Psi_g(e^{R_0 e^{-R_0}})
\end{aligned} \tag{4.13}$$

where $\Psi_g(x)$ denotes the iterated exponential function with $\Psi_1(x) = x$ and $\Psi_n(x) = x^{x^{\cdots^x}}$ for n times. When $R_0 < 1$, (4.13) gives a properly defined distribution with $\lim_{g\to\infty}\Pr\{M_g \le g\} = 1$. If $R_0 > 1$, we need to numerically calculate δ and replace R_0 in (4.13) by $\widetilde{R}_0 = \delta R_0$ to calculate the conditional distribution $\Pr\{\widetilde{M}_g \le g\}$ for the number of generations at extinction, conditioning on the outcome being a small outbreak.

The Negative Binomial Distribution for N

There are many stochastic mechanisms that may result in the phenomenon that the number of secondary infections produced by a typical infected individual N arises from a negative binomial distribution (Chap. 3). This distribution includes the geometric distribution as a special case and the Poisson distribution as a limiting case, and belongs to the power series distribution family with $A(\theta) = (1-\theta)^{-\kappa}$, $\kappa > 0$. The probability generating function is

$$G_N^{\text{NB}}(s) = \left(\frac{1-s\theta}{1-\theta}\right)^{-\kappa} = \left(1 + \frac{R_0(1-s)}{\kappa}\right)^{-\kappa} \tag{4.14}$$

where $R_0 = E[N] = G'_N(1) = \frac{\theta\kappa}{1-\theta}$. The variance is $var[N] = \frac{\theta\kappa}{(1-\theta)^2} = R_0 + R_0^2/\kappa$, a decreasing function of κ. The parameter κ is commonly called the shape parameter. The geometric distribution corresponds to $\kappa = 1$. The Poisson distribution corresponds to $\kappa \to \infty$. $G_N^{\mathrm{NB}}(s)$ is ranked separately by (κ, R_0). If κ is fixed, the larger the mean value R_0, the smaller is $G_N^{\mathrm{NB}}(s)$ and hence δ is a decreasing function of R_0. On the other hand, if R_0 is fixed, the larger the κ, the smaller is $G_N^{\mathrm{NB}}(s)$ and hence δ is a decreasing function of κ and the smallest value of δ when $R_0 > 1$ is $\delta = -R_0^{-1}\,\mathrm{LambertW}\left(-R_0 e^{-R_0}\right)$ as $\kappa \to \infty$. Using (4.8), one needs to write $\delta = \delta^*(\kappa)$ as a decreasing function of κ so that

$$G_{\widetilde{N}}^{\mathrm{NB}}(s) = \left(\frac{1 - s\delta^*(\kappa)\theta}{1 - \delta^*(\kappa)\theta}\right)^{-\kappa} = \left(\frac{\kappa + R_0 - s\delta^*(\kappa)R_0}{\kappa + R_0 - \delta^*(\kappa)R_0}\right)^{-\kappa},$$

and

$$\widetilde{R}_0 = \frac{R_0\kappa\delta^*(\kappa)}{\kappa + (1 - \delta^*(\kappa))\,R_0}. \tag{4.15}$$

In these expressions, $\delta^*(\kappa)$ is the smallest root of $G_N^{\mathrm{NB}}(s) = s$ in $s \in (0, 1]$ for a given κ.

The distribution of the time to extinction M_g can be calculated in the following manner: the probability that the initially seeded individual does not infect anyone, $\Pr\{M_g = 0\} = \Pr\{N = 0\} = \left(1 + \frac{R_0}{\kappa}\right)^{-\kappa}$ is a decreasing function of κ. Starting from this point, we get the following recursive calculation:

$$\Pr\{M_g \le g\} = \left(1 + \frac{R_0}{\kappa}\left(1 - \Pr\{M_g \le g - 1\}\right)\right)^{-\kappa}.$$

If $R_0 \le 1$, this is a properly defined cumulative distribution. If $R_0 > 1$, Conditioning on the final outcome being a small outbreak, then $\widetilde{M}_g$ has the following distribution

$$\begin{aligned}\Pr\{\widetilde{M}_g \le g\} &= \frac{1}{\delta}\Pr\{M_g \le g\} \\ &= \left(1 + \frac{\widetilde{R}_0}{\kappa}\left(1 - \Pr\{\widetilde{M}_g \le g - 1\}\right)\right)^{-\kappa},\end{aligned}$$

where $\widetilde{R}_0$ is given by (4.15).

4.2.3 Final Size Distributions for Small Outbreaks

Diekmann and Heesterbeek (2000) provide a distinction between a *small* outbreak from a *large* outbreak. The cumulative number of infected individuals at time t

is denoted by $C(t)$ and let $C(\infty) = \lim_{t\to\infty} C(t)$. $C(\infty)$ is a random count and $E[C(\infty)]$ is the expected final size of the outbreak.

Small outbreak: As $m \to \infty$, the outbreak becomes extinct after a handful cases such that the expected number $E[C(\infty)]$ is finite. The expected outbreak size as a proportion, $E[C(\infty)]/m$, is concentrated at zero.

Large outbreak: As $m \to \infty$, $E[C(\infty)] \to \infty$ but $E[C(\infty)]/m \to \eta$ where η is a positive quantity, $0 < \eta < 1$. The expected final outbreak size number scales linearly with the size of the susceptible population. In this case, it may be suitable to write $E[C(\infty)] = m\eta$ as proportional to the population size for some $0 < \eta < 1$ so that the final size scales linearly with m.

The original concept of small vs. large outbreaks was described in Kendall (1956), where the author considered the shape of the continuous random variable $Y_m = C(\infty)/m$ defined on $[0, 1]$ as having one of the two shapes: the J-shape and the U-shape. The term J-shape refers to a distribution that is monotonically decreasing so that it has a mode at zero. The distribution is said to have U-shape if it is bimodal.

Nåsell (1995) defines the invasion threshold as a value at which the distribution of Y_m makes a transition from J-shape to U-shape. When the transmission parameter is smaller than this threshold, then the distribution of Y_m has a J-shape with certainty. When it is larger than the threshold, then there exists a value $0 < \delta < 1$ such that, with probability $1-\delta$, the distribution takes the bell-shaped distribution. In the latter case, by the end of the outbreak, a positive proportion η of the population will be eventually infected. In the previous sub-section, we have seen that the threshold condition is $R_0 = 1$ and the probability δ is the smallest root of $G_N(s) = s$ in $s \in (0, 1]$.

A U-shaped distribution can be thought as a binary mixture of a J-shaped distribution and a uni-modal bell-shaped distribution with a probability δ assigned to the J-shaped distribution. Figure 4.6 shows frequencies of the final sizes in various population sizes using a stochastic SIR model corresponding to $R_0 = 1.8$ and an exponentially distributed infectious period. The number of initially seeded infectious individuals is $i_0 = 1$. This model gives (in theory) geometric distribution for N and calculated $\delta = 1/1.8 = 0.55556$. Each simulation is repeated 500 times by initially seeding one infected individual. Results give the frequency of the final sizes with two modes, with one mode at zero and another mode around $\eta = 0.73$ of the size of the susceptible population. In four simulations with populations sizes $n = 50, 200, 1000$, and 10,000, empirically observed δ lies between 0.55 and 0.60.

The condition $R_0 \leq 1$ leads to the probability $\delta = 1$, which gives a J-shaped distribution of Y_m. However, it does not always imply that the final outcome will be a small outbreak defined by $E[C(\infty)] < \infty$. A small outbreak leads to $E[Y_\infty] = \lim_{m\to\infty} E[C(\infty)]/m = 0$, but $E[Y_\infty] = 0$ does not necessarily lead to $E[C(\infty)] < \infty$. This is the case when $R_0 \leq 1$ but lies within the vicinity of 1 of the order of $O(m^{-1/3})$. Rather than scaling $C(\infty)$ by the population size m, Martin-Löf (1988) studied the asymptotic distribution of $Y_n^* = C(\infty)/m^{2/3}$ and showed that it has a limit distribution as $m \to \infty$ which may have different shapes, from J-shape

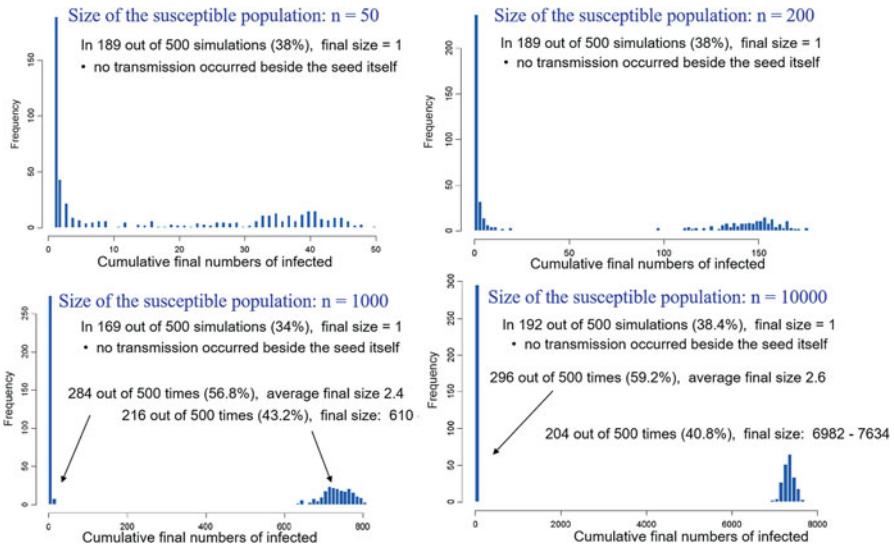


Fig. 4.6 U-shaped distributions generated by simulations by repeatedly seeding $i_0 = 1$ initially infectious individuals 500 times at identical initial conditions (with $R_0 = 1.8$) from susceptible population sizes $n = 50, 20, 1000$ and $10{,}000$

to a bimodal U-shape, to uni-modal with mode not at zero, or with a shape that is rather flat. This makes the final size unpredictable.

Since a small outbreak is defined by $E[C(\infty)] < \infty$, we used the random counts use $Z_s = 0, 1, \ldots,$ for the final size, and its distribution is not dependent on the population size.

Let $G_Z(s) = \sum_{z=1}^{\infty} s^z \Pr\{Z_s = z\}$. It has been shown (ref. Mode and Sleeman 2000, page 193) that

$$G_Z(s) = s \cdot G_N(G_Z(s)). \tag{4.16}$$

Thus $G_Z(1) = \delta \leq 1$ which satisfies $\delta = G_N(\delta)$, hence

$$\delta = \sum_{z=1}^{\infty} \Pr\{Z_s = z\} = \begin{cases} = 1, \text{ if } R_0 \leq 1 \\ < 1, \text{ if } R_0 > 1 \end{cases}.$$

If $R_0 > 1$, $\Pr\{Z_s = z\}$ does not define the complete probability distribution. When $R_0 \leq 1$, $G_Z(s)$ is the p.g.f. for Z_s. Large variability of N gives large values of $G_N(s)$ for any $s \in [0, 1]$ and hence large values of $G_Z(s)$.

Mean and Variances for Z_s

The first two derivatives with respect to s are

$$G'_Z(s) = G_N(G_Z(s)) + sG'_N(G_Z(s))G'_Z(s), \tag{4.17}$$

$$G''_Z(s) = 2G'_N(G_Z(s))G'_Z(s) + sG''_N(G_Z(s))\left(G'_Z(s)\right)^2 + sG'_N(G_Z(s))G''_Z(s). \tag{4.18}$$

When $R_0 < 1, \delta = 1$, $\Pr\{Z_s = z\}$, $z = 0, 1, \ldots$, is a complete probability distribution. In addition, $G_N(1) = G_Z(1) = 1$. Since $G'_N(1) = R_0 < 1$, let $s = 1$ in (4.17), one gets $G'_Z(1) = 1 + R_0 G'_Z(1)$. Hence

$$E[Z_s] = G'_Z(1) = \frac{1}{1 - G'_N(1)} = \frac{1}{1 - R_0}. \tag{4.19}$$

Similarly, let $s = 1$ in (4.18) and use $var[Z_s] = G''_Z(1) + G'_Z(1) - \left(G'_Z(1)\right)^2$, one gets

$$var[Z_s] = \frac{G''_N(1) + G'_N(1) - G'_N(1)^2}{\left(1 - G'_N(1)\right)^3} = \frac{var[N]}{(1 - R_0)^3}. \tag{4.20}$$

Both the mean final size $E[Z_s]$ and its variance $var[Z_s]$ increase with R_0 towards infinity as $R_0 \uparrow 1$. Therefore when R_0 is in the vicinity of 1, the final size is unpredictable. Limiting the discussion for the distribution of N given fixed $R_0 < 1$, the more heterogeneous is N, the larger is $var[N]$ and the larger is $var[Z_s]$.

When $R_0 > 1$, conditioning of the outcome being a small outbreak, the observed phenomenon arises from the distribution of $\widetilde{N}$ with the p.g.f. (4.6). Thus,

$$E[Z_s] = \frac{1}{1 - \widetilde{R}_0}, \text{ where } \widetilde{R}_0 = G'_N(\delta) < 1; \tag{4.21}$$

$$var[Z_s] = \frac{\delta G''_N(\delta) + G'_N(\delta) - G'_N(\delta)^2}{\left(1 - G'_N(\delta)\right)^3} = \frac{var[\widetilde{N}]}{\left(1 - \widetilde{R}_0\right)^3}. \tag{4.22}$$

These are extensions of (4.19) and (4.20) in the sense that they are valid for both $R_0 > 1$ and $R_0 < 1$. For the latter, $\delta = 1$ and $G'_N(\delta) = R_0$.

The Distribution of Z_s

Since $G_Z(s)$ as given by (4.16) is only a properly defined p.g.f. for Z_s when $R_0 \leq 1$, we extend (4.16) as

$$G_Z(s) = s \cdot G_{\widetilde{N}}(G_Z(s)) \tag{4.23}$$

where $\widetilde{N}$ is uniquely defined by the p.g.f. $G_{\widetilde{N}}(s) = \frac{1}{\delta} G_N(\delta s)$. In this case $\widetilde{N}$ is identically distributed as N if $R_0 \leq 1$.

We first define

$$\Pr\{Z_s = z\} = \frac{1}{z!} G_Z^{(z)}(0) \tag{4.24}$$

where $G_Z^{(z)}(0) = \frac{d^z}{ds^z} G_Z(s)\Big|_{s=0}$ satisfying $\sum_{z=1}^{\infty} \Pr\{Z_s = z\} = \delta \leq 1$. When $G_Z(s)$ is defined by (4.23), (4.24) always gives a properly defined probability distribution for Z_s.

For any given p.g.f. $G_{\widetilde{N}}(\cdot)$, a close analytic form of $G_Z(s)$ may not always exist by solving $G_Z(s) = sG_{\widetilde{N}}(G_Z(s))$. However, $G_Z^{(z)}(0)$ can be sometimes solved recursively starting from $\Pr\{Z_s = 1\} = G'_Z(0)$. We use the convention that

$$\Pr\{Z_s = 1\} = G'_Z(0) = G_{\widetilde{N}}(0) = \Pr\{\widetilde{N} = 0\}.$$

The event $\{Z_s = 1\}$ corresponds to the case that there is no secondary transmission in the population and the final size is the number of initially seeded individuals. There is also a convention that $G_Z(0) = \Pr\{Z_s = 0\} = 0$ as there must be at least one infective individual to start an outbreak.

The recursive procedure can be demonstrated for $\Pr\{Z_s = 2\}$ and $\Pr\{Z_s = 3\}$. From (4.18), $G''_Z(0) = 2G'_{\widetilde{N}}(G_Z(0))G'_Z(0) = 2G'_{\widetilde{N}}(0)G_{\widetilde{N}}(0)$, which gives

$$\Pr\{Z_s = 2\} = \frac{1}{2} G''_Z(0) = \Pr\{\widetilde{N} = 1\} \Pr\{\widetilde{N} = 0\}.$$

It is the probability that the index case can transmit to one individual with probability $\Pr\{\widetilde{N} = 1\}$, and the second individual does not transmit with probability $\Pr\{\widetilde{N} = 0\}$.

With a bit more calculus, $G_Z^{(3)}(0) = 3G''_{\widetilde{N}}(0)\left[G_{\widetilde{N}}(0)\right]^2 + 6\left[G'_{\widetilde{N}}(0)\right]^2 G_{\widetilde{N}}(0)$ so that $\Pr\{Z_s = 3\} = \frac{1}{3!} G_Z^{(3)}(0)$ can be expressed as

$$\begin{aligned}\Pr\{Z_s = 3\} &= \frac{1}{2} G''_{\widetilde{N}}(0)\left[G_{\widetilde{N}}(0)\right]^2 + \left[G'_{\widetilde{N}}(0)\right]^2 G_{\widetilde{N}}(0) \\ &= \Pr\{\widetilde{N} = 2\}\left(\Pr\{\widetilde{N} = 0\}\right)^2 + \left(\Pr\{\widetilde{N} = 1\}\right)^2 \Pr\{\widetilde{N} = 0\}.\end{aligned}$$

It implies that either the index case produces two secondary cases with probability $\Pr\{\widetilde{N} = 2\}$ and neither of the secondary cases produces further transmission with probability $\left(\Pr\{\widetilde{N} = 0\}\right)^2$; or the index case produces one transmission and the secondary case produces one transmission with joint probability $\left(\Pr\{\widetilde{N} = 1\}\right)^2$, and the third case does not transmit with probability $\Pr\{\widetilde{N} = 0\}$.

Some Special Cases when N is Distributed Within the Negative Binomial Family

The Negative Binomial Distribution We assume that $R_0 \neq 1$ and N follows the negative binomial distribution with $R_0 = E[N]$ and variance $var[N] = R_0 + R_0^2/\kappa$. Given the outcome being a small outbreak, the observation arises from $\widetilde{N}$. If $R_0 < 1$, $\widetilde{N}$ is identical to N. If $R_0 > 1$, $\widetilde{N}$ also follows a negative binomial distribution by replacing R_0 with $\widetilde{R}_0 < 1$ which is defined by (4.15). The p.g.f. of the negative binomial distribution is $G_{\widetilde{N}}(s) = \left(1 + \widetilde{R}_0(1-s)/\kappa\right)^{-\kappa}$ and (4.23) can be written as:

$$G_Z(s) = s\left(1 + \frac{\widetilde{R}_0(1 - G_Z(s))}{\kappa}\right)^{-\kappa}.$$

First, $\Pr\{Z_s = 1\} = G'_Z(0) = G_{\widetilde{N}}(0) = \left(1 + \widetilde{R}_0/\kappa\right)^{-\kappa}$. For $z \geq 2$, we calculate recursively $G_Z^{(z)}(0) = \prod_{j=0}^{z-2}\left(\frac{j}{\kappa} + z\right)\widetilde{R}_0^{z-1}\left(1 + \frac{\widetilde{R}_0}{\kappa}\right)^{1-z(\kappa+1)}$. From (4.24),

$$\begin{aligned}\Pr\{Z_s = z\} &= \frac{1}{z!}\prod_{j=0}^{z-2}\left(\frac{j}{\kappa} + z\right)\widetilde{R}_0^{z-1}\left(1 + \frac{\widetilde{R}_0}{\kappa}\right)^{1-z(\kappa+1)} \qquad (4.25)\\ &= \frac{\Gamma(z\kappa + z - 1)}{\Gamma(z\kappa)\Gamma(z-1)}\left(\frac{\widetilde{R}_0^{z-1}(1+\widetilde{R}_0)^{1-2z}}{\kappa}\right)^{z-1}, \; z = 2, 3, \ldots.\end{aligned}$$

The mean and variance are

$$E[Z_s] = \frac{1}{1 - \widetilde{R}_0}, \; var[Z_s] = \frac{\widetilde{R}_0 + \widetilde{R}_0^2/\kappa}{(1 - \widetilde{R}_0)^3}.$$

The Geometric Distribution $\kappa = 1$ As a special case, $\Pr\{Z_s = 1\} = \left(1 + \widetilde{R}_0\right)^{-1}$ and for $z \geq 2$

$$\begin{aligned}\Pr\{Z_s = z\} &= \frac{1}{z!}\prod_{j=0}^{z-2}(j + z)\widetilde{R}_0^{z-1}\left(1 + \widetilde{R}_0\right)^{1-2z} \qquad (4.26)\\ &= \frac{\Gamma(2z-1)}{\Gamma(z)\Gamma(z-1)}\left(\widetilde{R}_0^{z-1}(1+\widetilde{R}_0)^{1-2z}\right)^{z-1}, \; z = 2, 3, \ldots..\end{aligned}$$

When $R_0 < 1$, this gives the probability for $z \geq 2$:with mean value

$$E[Z_s] = \frac{1}{1 - R_0}, \; var[Z_s] = \frac{R_0(R_0+1)}{(1-R_0)^3}. \qquad (4.27)$$

When $R_0 > 1$, if the outcome happens to be a small outbreak, then $\widetilde{R}_0 = R_0^{-1}$ and

$$E[Z_s] = \frac{R_0}{R_0 - 1},\ var[Z_s] = \frac{R_0\,(R_0 + 1)}{(R_0 - 1)^3}. \tag{4.28}$$

The Poisson Distribution $\kappa \to \infty$ The limiting case of (4.25) is

$$\Pr\{Z_s = z\} = \frac{(\widetilde{R}_0 z)^{z-1}}{z!} e^{-\widetilde{R}_0 z},\ \ z = 1, 2, 3, \ldots \tag{4.29}$$

where $\widetilde{R}_0 = R_0$ if $R_0 < 1$ and $\widetilde{R}_0 = -\,\text{LambertW}\left(-R_0 e^{-R_0}\right) < 1$ if $R_0 > 1$. This distribution is the Borel-Tanner distribution first discovered by Borel (1942). Its mean and variance are

$$E[Z_s] = \frac{1}{1 - \widetilde{R}_0},\ var[Z_s] = \frac{\widetilde{R}_0}{(1 - \widetilde{R}_0)^3}.$$

4.2.4 Examples

We compare results in two examples. In Example 13, $R_0 < 1$ and extinction is certain. Regarding the three distributions (Poisson, geometric, and negative binomial) examined, the probability distributions for M_g and Z_s are calculated based on $R_0 = 0.5$.

Example 13 Suppose that we conduct a virtual experiment by repeatedly seeding $i_0 = 1$ infected individual in an infinitely large susceptible population. We assume that this individual produces new infections according to a negative binomial distribution with a shape parameter $\kappa > 0$ and on average this individual produces $R_0 = 0.5$ new infections. Each new infection produces new infections in their next generation independently and the number of new infections also follows the same distribution. Given R_0, κ ranks both the p.g.f. $G_N^{\text{NB}}(s)$ and $var[N] = R_0 + R_0^2/\kappa$. The parameter κ gives a ranking of variability: the smaller the value of κ, the larger the variability. It also ranks the probabilities $\Pr\{N = 0\} = \Pr\{M_g = 0\} = \Pr\{Z_s = 1\}$. In the case of the Poisson distribution ($\kappa \to \infty$), $\Pr\{M_g = 0\} = \Pr\{Z_s = 1\} = 0.6065$; the probability of extinction within three generations is $\Pr\{M_g \le 3\} = 0.9582$ and the probability of more than two total individuals infected (including the originally seeded individual) is $\Pr\{Z_s > 2\} = 0.21$. If the variability of N is larger than that of the Poisson distribution (i.e., as κ decreases), the probability $\Pr\{M_g = 0\} = \Pr\{Z_s = 1\}$ increases and it takes fewer generations to extinction. At $\kappa = 0.25$, $\Pr\{M_g = 0\} = \Pr\{Z_s = 1\} = 0.7598$; the probability of extinction within three generations is $\Pr\{M_g \le 3\} = 0.9801$ and the probability of

more than two total individuals infected (including the originally seeded individual) is $\Pr\{Z_s > 2\} = 0.09$.

When $R_0 > 1$, extinction is uncertain. For the same three distributions examined in Example 13, the probabilities of extinction are calculated along with a different value $\widetilde{R}_0 < 1$. In this case, the parameter κ does not rank the probability $\Pr\{\widetilde{M}_g = 0\} = \Pr\{Z_s = 1\}$. This is shown in the following example.

Example 14 Suppose that we conduct the same experiment as in Example 13 but setting $R_0 = 3$. Given R_0, κ ranks both the p.g.f. $G_N^{\mathrm{NB}}(s)$ and $var[N] = R_0 + R_0^2/\kappa$ as well as the probability of extinction δ and the value of $\widetilde{R}_0$. Conditioning on the event of extinction, the distribution of generations to extinction and the final size of the small outbreak are calculated based on different values of $\widetilde{R}_0$.

1. The Poisson distribution ($\kappa \to \infty$) gives the smallest value $\delta \approx 0.06$. It is expected that in 6% of the repeated experiments the transmission will not sustain and data arise as if manifested from a different Poisson distribution with $\widetilde{R}_0 = -\,\mathrm{LambertW}\left(-3e^{-3}\right) = 0.17856$. With respect to generations to extinction, from 4.13, we get $\Pr\{\widetilde{M}_g = 0\} = 0.8365$, $\Pr\{\widetilde{M}_g \le 1\} = 0.9712$, $\Pr\{\widetilde{M}_g \le 2\} = 0.9949$, $\Pr\{\widetilde{M}_g \le 3\} = 0.9991$, $\Pr\{\widetilde{M}_g \le 4\} = 0.9998$, $\cdots$. The average number of total individuals infected (including the originally seeded individual) is 1.217, and the probability of more than two total individuals infected is $\Pr\{Z_s > 2\} = 0.0386$.
2. The geometric distribution ($\kappa = 1$), yields $\delta = 1/3$. It is expected that about in one third of the experiments transmission will not sustain. In this case, data follow a geometric distribution but cannot inform us whether they arise from a conditional distribution of the geometric distribution with mean value $R_0 = 3$, conditioning on extinction; or they arise from an unconditional geometric distribution, the mean value $\widetilde{R}_0 = 1/3$. From 4.10, we get $\Pr\{\widetilde{M}_g = 0\} = 0.75$, $\Pr\{\widetilde{M}_g \le 1\} = 0.923$, $\Pr\{\widetilde{M}_g \le 2\} = 0.975$, $\Pr\{\widetilde{M}_g \le 3\} = 0.992$, $\Pr\{\widetilde{M}_g \le 4\} = 0.997$, $\cdots$. The average number of total individuals infected (including the originally seeded individual) is 1.5 and the probability of more than two total individuals infected is $\Pr\{Z_s > 2\} = 0.109$.
3. The negative binomial distribution with $\kappa = 0.25$ yields $\delta = 0.67$, which is the smallest root of the equation $\left(1 + \frac{3(1-s)}{0.25}\right)^{-0.25} = s$ for $s \in (0, 1]$. In this case, it is expected that about in two thirds of the experiments the transmission will not sustain. Meanwhile, $\widetilde{R}_0 = 0.40546$ calculated from (4.15). With respect to generations to extinction, we get $\Pr\{\widetilde{M}_g = 0\} = 0.785$, $\Pr\{\widetilde{M}_g \le 1\} = 0.9282$, $\Pr\{\widetilde{M}_g \le 2\} = 0.9728$, $\Pr\{\widetilde{M}_g \le 3\} = 0.9893$, $\Pr\{\widetilde{M}_g \le 4\} = 0.996$, $\cdots$. The average of total number of infected individuals (including the originally seeded individual) is 1.682 and the probability of more than two total individuals infected is $\Pr\{Z_s > 2\} = 0.068$.

4.2.5 *Estimation for R_0 Based on the Galton-Watson Branching Process*

Given the roles of $R_0 = E[N]$ in the properties of the branching process, it is important to estimate this parameter based on observations of $(X_0, X_1, \ldots, X_G)$ for the first G generations, assuming that N_i are independently and identically distributed for all individuals.

From the definition (4.1), X_g/X_{g-1} is an unbiased estimator for R_0 for each $g = 1, \ldots, G$, without specifying the family of distributions for N. These estimators can be pooled to provide a single, more efficient estimator. One way of doing this is to take the weighted average so that

$$\widetilde{R}_0 = \sum_{g=1}^{G} w_g \left(X_g/X_{g-1}\right), \; w_1 + w_2 + \cdots + w_g = 1.$$

Harris (1948) introduced the weight

$$w_g = \frac{X_{g-1}}{\sum_{g=1}^{G} X_{g-1}}$$

because the conditional variance $var\left[\frac{X_g}{X_{g-1}}|X_{g-1}\right] = \frac{var[N]}{X_{g-1}}$ and it is appropriate to choose the weight w_g inversely proportional to the conditional variance. This leads to the Harris estimator

$$\widehat{R}_0 = \frac{\sum_{g=1}^{G} X_g}{\sum_{g=1}^{G} X_{g-1}}. \tag{4.30}$$

The standard error estimation for $\widehat{R}_0$ is (Becker 1989)

$$s.e.(\widehat{R}_0) = \left(\frac{\widehat{var[N]}}{\sum_{g=1}^{G} X_{g-1}}\right)^{1/2} \tag{4.31}$$

where the basic reproduction variance $var[N]$ needs to be estimated separately. One may make assumptions on the distributions of N. For example, if N follows a Poisson distribution, then one can use $\widehat{var[N]} = \widehat{R}_0$ in (4.31) to get

$$s.e.(\widehat{R}_0) = \frac{\left(\sum_{g=1}^{G} X_g\right)^{1/2}}{\sum_{g=1}^{G} X_{g-1}}.$$

On the other hand, Heyde (1974) and Dion (1975) independently proposed the following general maximum likelihood estimate

$$\widehat{var[N]} = \frac{1}{G}\sum_{g=1}^{G} X_{g-1}\left(\frac{X_g}{X_{g-1}} - \widehat{R}_0\right)^2 \tag{4.32}$$

based on the central limit result from martingale theory. However, as pointed out by Becker (1989), because G will be generally small in practice, it casts doubts on the precision of the estimator $\widehat{var[N]}$ in (4.32).

Modified Harris Estimator Based on Surveillance Data

Data like $(X_0, X_1, \ldots, X_G)$ track infections through generations and require detailed information of the transmission tree, in terms of who infects whom. Such data are rarely available in practice. Most disease surveillance data are collected at chronological time $t = 0, 1, 2, \ldots, C$, denoted by $(Y_0, Y_1, \ldots, Y_C)$, where Y_t may represent the number of observed events at time t, such as the onset of clinical symptoms.

The data $(Y_0, Y_1, \ldots, Y_C)$ nevertheless correspond to an embedded Galton-Watson branching process. However, generations overlap over time. All that we need is a book keeping device to assign data Y_t to correct generations.

A (clinical) serial interval is defined by Hope Simpson (1948) as the period from the observation of symptoms in one case to the observation of symptoms in a second case directly infected from the first case.

White and Pagano (2008) suggested that, as long as the observation of symptoms associated with the serial interval is in agreement with the case definition of the surveillance system that generates data $(Y_0, Y_1, \ldots, Y_C)$, the distribution of the serial interval can be a good book keeping device.

We assume that the serial interval distribution, represented in discrete time scale, is $\{p_j,\ j = 1, 2, \ldots, k\}$ such that $\sum_{j=1}^{k} p_j = 1$. The expected value $E[Y_t]$ can be expressed by the convolution (a special case of Exercise 3.4 (c)):

$$E[Y_t] = R_0 \sum_{j=1}^{\min(k,t)} p_j Y_{t-j}.$$

White and Pagano (2008) showed that, if $(Y_0, Y_1, \ldots, Y_C)$ arise as independent Poisson distributed random counts with $E[Y_t]$ given by the above convolution, the maximum likelihood estimate for R_0 is

$$\widehat{R}_0 = \frac{\sum_{t=1}^{C} Y_t}{\sum_{t=1}^{C}\sum_{j=1}^{\min(k,t)} p_j Y_{t-j}}. \tag{4.33}$$

The Harris estimator (4.30) is a special case when the serial interval distribution degenerates to a single point.

There are several limitations to apply (4.33).

1. To use the serial interval distribution as a device to estimate R_0 , it is important that the measurements for the serial interval must be based on the same case definition as that for surveillance that generates the time-series. This principle may not always be easy to adhere in practice.
2. The clinical serial intervals may be difficult to observe accurately. We shall discuss this further in Chap. 7.
3. The estimator (4.33) is sensitive to the serial interval distribution which needs to be estimated from additional data.

From Theory to Practice

In addition to the assumptions of independency (that N_i are i.i.d.) and stationarity of the distribution of N over generations, the branching process approximation for the outbreak at its initial phase also carries implicit assumptions that the susceptible population is infinitely large and homogeneous mixing. With these assumptions, the branching process serves as a function of the idealization of the world that provides conceptual clarity. On the other hand, none of the transmission trees presented in Fig. 4.1 meets the assumptions of the branching process approximation.

Building a transmission tree in an outbreak in a large community or in the general population is extremely difficult, even in the very initial phase of an outbreak investigation. The illustrations in Fig. 4.1 are all based on data in special settings.

In many of the outbreaks occurring in the twentieth century, the roles of hospitals have been particularly significant in amplifying outbreaks of importation or emerging infections, for undiagnosed or misdiagnosed patients are often admitted into general wards. Most of the infected people during such outbreaks are found among persons associated with hospitals, either occupationally or as patients or visitors.

Fenner et al. (1988) give many examples for imported smallpox outbreaks in hospitals around the world from 1950 to 1974, with further references. Figure 4.1a is taken from the 1967 smallpox outbreak in Kuwait in which two hospitals were the principal foci of the infection. It started with a misdiagnosis as chickenpox in the Fever Hospital and the transmission continued for four generations in this hospital. Then one of the patients in the third generation, unrecognized, was transferred to a different hospital where a few more individuals were infected (Arita et al. 1970).

The Morbidity and Mortality Weekly Report (MMWR) published by Centers for Disease Control and Prevention (2003) reported detailed case histories of the 5 "super-spreading events" of the SARS outbreak in Singapore. Goh et al. (2006) published the transmission tree, reproduced here as Fig. 4.7.

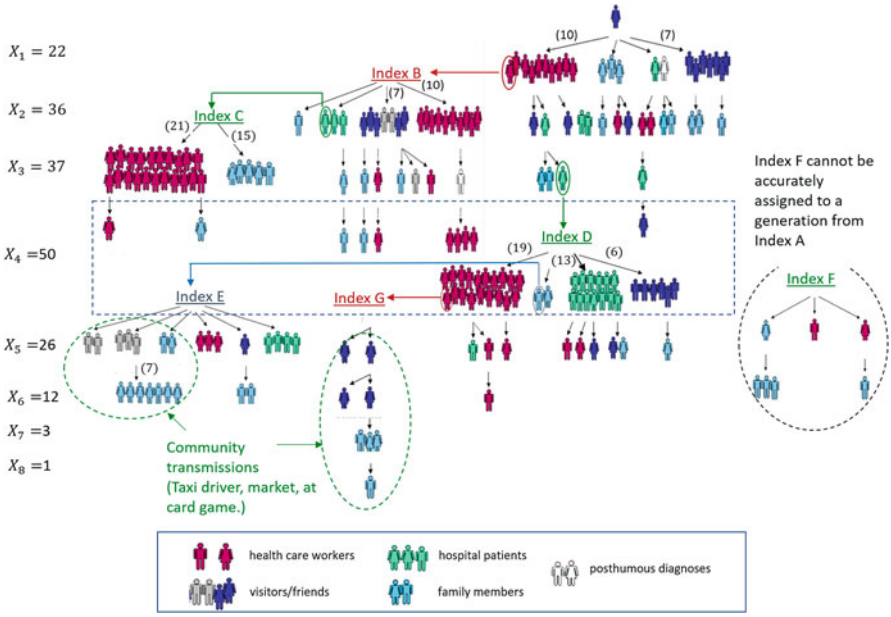


Fig. 4.7 The transmission tree of the 2003 SARS outbreak in Singapore showing strong clustering, adopted from Goh et al. (2006)

1. Index A was believed to be infected in Hong Kong and returned to Singapore with symptoms. This patient entered Ward 5A of the Tan Tock Seng Hospital (TTSH) on March 1, 2003 and was transferred to Ward 8A across the corridor. Epidemiologic investigation suggested that she directly infected 22 people, including 10 health care workers, 7 visitors, 3 family members, and 2 patients.
2. Index B was a nurse who attended Index A in Ward 5A. She became symptomatic on March 7 and was admitted to Ward 8A (and isolated on March 13), was believed to have infected 21 other people.
3. Index C was a patient admitted on March 10 at TTSH, stayed in the same 6-bed Ward 8A with Index B and became symptomatic for SARS on March 12 and was isolated on March 20. This patient was believed to have directly infected 26 other people, including 21 health care workers and 5 family members.
4. Index D was a patient who stayed in Ward 5A of TTSH between March 5 and March 20 for chronic kidney disease. She was transferred to Singapore General Hospital (SGH) Ward 57 to be treated for gastro-intestinal bleeding. She developed high fever on March 29 and transferred to Ward 57 of SGH and isolated on April 4 after confirmation of pneumonia. This patient was believed to be the index case of a cluster of 40 probable SARS cases at SGH.

5. Index E was a sibling of Index D. He visited SGH on March 31, became symptomatic on April 5, and was admitted to National University Hospital (NUH) Ward 64. He was confirmed with SARS and then transferred to TTSH on April 9. A total of 15 SARS cases were epidemiologically linked to this patient.
6. In addition to the above five super-spreading events, there are some additional smaller clusters, including Index F and Index G.

In Fig. 4.7, the distribution of N tends to have a large frequency of zeros and meanwhile a very heavy tail. For example, 15 out of the $X_1 = 22$ individuals in the first generation from Index A produced zero transmission whereas one individual (Index B) produced 21 transmissions; 28 out of the $X_2 = 36$ individuals in the second generation from Index A produced zero transmission whereas one individual (Index C) produced 26 transmissions; 29 out of the $X_3 = 37$ individuals in the second generation from Index A produced zero transmission whereas one individual (Index D) produced 40 transmissions. Of the 196 individuals shown in Fig. 4.7, 152 had zero transmission whereas five individuals had more than 15 transmissions.

In addition, Goh et al. (2006) also documented that Indices A–D mainly transmitted SARS in hospitals, up to the fourth generation from Index A, whereas Indices E and G corresponded to community transmission settings.

The estimates X_g/X_{g-1} and (4.30) are still relevant to outbreak investigation, with careful interpretations of their meanings. There is strong clustering and at the same time, public health intervention such as isolation started as early as March 13 during the first generation. The estimate X_g/X_{g-1} keeps track the temporal effective reproduction number between two successive generations, whereas the Harris estimate (4.30) calculated the weighted average, as the "controlled" reproduction number from Generation Zero to Generation G. Table 4.1 tabulates these crude estimates for the entire transmission tree (except for the cluster from Index F) by generations, the crude estimates for the transmission tree from hospital transmission, as well as for two small community transmission clusters (Indices E and G).

The first four generations from Index A are all attributed to transmissions in hospital settings. The Harris estimate (4.30) is $\widehat{R} = 1.5$, which is the weighted average of X_g/X_{g-1} up to $g = 4$ as the estimate for the mean value $E[N]$. It is interpreted as the controlled reproduction number, taking into consideration the effects of public health interventions. The variance $var[N]$, interpreted as the controlled reproduction variance, estimated by $g = 4$ according to (4.32), is $\widehat{var[N]} = 7.6$.

These are very crude estimates. We shall dedicate Chap. 9 of this book to address spatial structures and behavior change.

Table 4.1 Calculations of $X_g/\ X_{g-1}$ and the Harris estimate (4.30) by transmission clusters

	All transmissions from Index A			Hospital transmissions		
	X_g	X_g/X_{g-1}	$\frac{\sum_{j=1}^{g} X_j}{\sum_{j=1}^{g} X_{j-1}}$	X_g	X_g/X_{g-1}	$\frac{\sum_{j=1}^{g} X_j}{\sum_{j=1}^{g} X_{j-1}}$
$g=0$	1			1		
$g=1$	22	22		22	22	
$g=2$	36	1.64	2.52	36	1.64	2.52
$g=3$	37	1.03	1.61	37	1.03	1.61
$g=4$	50	1.35	1.51	50	1.35	1.51
$g=5$	26	0.52	1.17	17	0.34	1.11
$g=6$	12	0.46	1.06	3	0.18	1.01
$g=7$	3	0.25	1.01	0	0	0.99
$g=8$	1	0.33	1.00			
$g=9$	0	0	0.99			

	Community transmission Index E			Community transmission Index G		
	X_g	X_g/X_{g-1}	$\frac{\sum_{j=1}^{g} X_j}{\sum_{j=1}^{g} X_{j-1}}$	X_g	X_g/X_{g-1}	$\frac{\sum_{j=1}^{g} X_j}{\sum_{j=1}^{g} X_{j-1}}$
$g=0$	1			1		
$g=1$	7	7		2	2	
$g=2$	7	1	1.75	2	1	1.33
$g=3$	0	0	0.93	3	1.5	1.4
$g=4$				1	0.33	1
$g=4$				0	0	0.89

4.3 The Initial Growth Given Non-extinction

Non-extinction (invasion) is only possible when $R_0 > 1$. Should it occur, the outbreak will evolve into complex dynamic processes which may also depend on environmental and demographic changes. In the long run, the outbreak may end with a substantial proportion of the population infected (i.e., large outbreak) or may reach an endemic steady state with sustained on-going transmission persisting for a very long period of time. These issues are investigated in Chap. 5.

In this section, we continue to focus on the initial transmission stage, in which the depletion of the susceptible population is negligible, and the branching process approximation is adequate. Hence, it is reasonable to assume the size of the susceptible population $m \to \infty$.

4.3.1 The Exponential Growth by Generation

In the Galton-Watson branching process, X_g denotes the size of the gth generation and can be represented by the random sum given by (4.1), where the p.g.f. for X_g

is $G_{X_g}(s) = G_{X_{g-1}}(G_N(s))$. When $R_0 > 1$, given $X_0 = 1$ and non-extinction, the expected value and variance are (ref: Theorem 5.1 of Harris 1963):

$$E[X_g] = R_0^g, \quad var[X_g] = \frac{R_0^g(R_0^g - 1)}{R_0^g - R_0} var[N] \tag{4.34}$$

where $g = 0, 1, 2, \ldots$. Thus, the first moment $R_0 = E[N]$ determine the exponential growth of the first moment: $E[X_g] = e^{g \log R_0}$ with growth rate being $r = \log R_0$; the first two moments of N determines the first two moments $E[X_g]$ and $var[X_g]$. If the entire distribution of N is given, then $G_N(s)$ is fully specified and in theory, the distribution for each X_0 can be calculated.

However, one needs to keep in mind that the exponential growth given by (4.34) is only an approximation for the very few generations under the strong assumption that R_0 for all individuals in these generations is the same as $R_0 = E[N]$ corresponding to the initially seeded individuals at Generation Zero.

The distribution of X_g, and the results such as the probability of extinction, the generation to extinction and final outbreak sizes given extinction are properties of the Galton-Watson branching process and are determined by the distribution of N, regardless of the continuous time branching processes in which the Galton-Watson process is embedded.

4.3.2 Growth in Real (Continuous) Time

As the outbreak grows, it becomes increasingly difficult to track who infects whom and keep track of the generation number. The growth in continuous time during the initial phase of an outbreak depends on the properties of the counting process $\{K(x) : x \geq 0\}$, and the distributions of the latent and infectious periods in a continuous time framework where the Galton-Watson branching process is embedded within, such as the CMJ process and its extensions introduced earlier in this chapter.

The Exponential Growth Derived as the Expected Value of a Linear Pure Birth Markov Process

In parallel to the assumption in (4.34) that R_0 does not change over the generations during the initial phase, the underlying assumption in the continuous time is that, when an individual is initially seeded into the large susceptible population, it carries an *intrinsic rate* r. During the initial phase, all infected individuals carry the same intrinsic rate and r is independent of time. Therefore, the cumulatively infected individuals $C(t)$ at time t conforms *homogeneity*. Given $C(t) = n$, the time to produce a new infection in the population is exponentially distributed at rate

nr according to an independent competing risk framework. The parameter r has many names: the exponential growth rate, the intrinsic growth rate, as well as the *Malthusian number*.

This leads to the model introduced in Chap. 3 as (3.41). The cumulative number of infections in the population forms a counting process $\{C(t) : t \geq 0\}$. An instantaneous transmission is specified by the conditional probability $\Pr\{C(t+h) = n+1|C(t) = n\} = nrh + o(h)$. The marginal distribution of $C(t)$ at time t follows a negative binomial distribution, with mean value satisfying the exponential growth $E[C(t)] = i_0 e^{rt}$. The variance also grows exponentially over time: $var[C(t)] = i_0 e^{2rt}(1 - e^{-rt})$.

This model has a deterministic counterpart: $C'_d(t) = rC_d(t)$, where $C_d(t)$ is a non-random function of time.

The Euler-Lotka Equations

The initial growth rate r and the basic reproduction number R_0 are defined in a parallel manner. Both are defined at $t = 0$ when the system is at (disease-free) equilibrium. In both definitions, it is assumed that if the number of initially seeded infected individuals $i_0 > 1$, then these individuals act independently and identically with the same intrinsic parameters. It is further assumed that infected individuals carry the same intrinsic parameters during the initial transmission phase.

Because of these parallel definitions and assumptions, it is natural to investigate their relationship. This is done through embedding the Galton-Watson branching process into a continuous time framework, such as the CMJ process and its extensions, by focusing on the infectious contact processes $\{K(x) : x \geq 0\}$ from the perspective of infected individuals along with the assumed distributions of the latent and infectious periods, where the time $x = 0$ refers to the time at infection of a typical infected individual. This infectious contact process is modeled through an instantaneous intensity $\beta(x)$ which may be a function of time as defined in the general formulation of a counting process in Sect. 3.3.

An infected individual may have a random latent period T_E during which it is not possible to transmit the infection to another individual through contact and a random infectious period T_I. We assume that during the infectious period, the infected individual remains infectious at a constant level of infectivity. The infectious period T_I serves as a stopping time of the infectious contact process given by $\{K(x) : x \geq 0\}$. A new counting process $\{N(x) : x \geq 0\}$ is defined such that $N(x) = 0$ if $0 < x \leq T_E$; $N(x) = K(x - T_E)$, if $T_E < x \leq T_E + T_I$; and $N(x) = K(T_E + T_I)$, if $x \geq T_E + T_I$. We define a binary stochastic process $\{X(x) : x \geq 0\}$ such that $X(x) = 1$ if the infected individual is infectious at time x (i.e., $T_E < x \leq T_E + T_I$) and $X(x) = 0$ otherwise. Let $B(x) = \beta(x)X(x)$. $\{B(x) : x \geq 0\}$ is a stochastic intensity process of the counting process $\{N(x) : x \geq 0\}$. The expected instantaneous rate of producing an infectious contact is $E[B(x)] = \beta(x)\Pr(X(x) = 1)$. We denote $A(x) = \Pr(X(x) = 1)$.

Starting with a single infected individual infected at $t = 0$ which become the index case, at time $t > 0$, the expected number of cumulatively infected individuals in the population includes the original individual plus all the expected cumulatively infected individuals evolving from this individual during $[0, t)$. If an infectious contact is made at time $x \in [0, t)$, the expected cumulative infected individuals at time t evolving from this contact is $E[C(t - x)]$ and the mean number of such infectious contacts in a small time interval containing x is $\beta(x)A(x)$. Thus we obtain the following equation

$$E[C(t)] = 1 + \int_0^t \beta(x)A(x)E[C(t - x)]dx. \tag{4.35}$$

We call (4.35) a renewal-type equation because the classic renewal equation $u(t) = v(t) + \int_0^t f(x)u(t-x)dx$ requires $\int_0^\infty f(x)dx = 1$. In the classic renewal equation, assuming $\int_0^\infty xf(x)dx < \infty$, there is the following asymptotic result (see Feller (1966) or later editions) from renewal theory,

$$u(t) \to \frac{\int_0^\infty v(x)dx}{\int_0^\infty xf(x)dx} \text{ as } t \to \infty.$$

It can be proven that $R_0 = \int_0^\infty \beta(x)A(x)dx$. When $R_0 > 1$, there exists a unique real value $r > 0$ such that $\int_0^\infty e^{-rx}\beta(x)A(x)dx = 1$. Multiplying both sides of (4.35) by e^{-rt}, then

$$e^{-rt}E[C(t)] = e^{-rt} + \int_0^t e^{-rx}\beta(x)A(x)e^{-r(t-x)}E[C(t - x)]dx$$

becomes a classic renewal equation by letting $u(t) = e^{-rt}E[C(t)]$, $v(t) = e^{-rt}$ and $f(x) = e^{-rx}\beta(x)A(x)$. Assuming that $\int_0^\infty xe^{-rt}\beta(x)A(x)dx < \infty$,

$$e^{-rt}E[C(t)] \to \frac{\int_0^\infty e^{-rx}dx}{\int_0^\infty xe^{-rt}\beta(x)A(x)dx} = \text{constant, as } t \to \infty.$$

Therefore, when $\beta(x)A(x)$ is a constant r, (4.35) gives the exponential growth $\frac{d}{dt}E[C(t)] = rE[C(t)]$; when $\beta(x)A(x)$ is not constant, (4.35) corresponds to asymptotic exponential growth $E[C(t)] \propto e^{rt}$ (as $t \to \infty$).

The general linkage between r and R_0 is given by a pair of equations

$$\int_0^\infty \beta(t)A(t)dt = R_0,$$

and

$$\int_0^\infty e^{-rt}\beta(t)A(t)dt = 1. \tag{4.36}$$

The latter is commonly referred to as the Euler-Lotka equation.

4.3.3 The Euler-Lotka Equations Under Models with SEI Structure

In many infectious disease transmission models, it is often assumed:

1. Homogeneous mixing: The (social) contact network is homogeneous as a random graph so that each individual has equal chance to make contact with any other individual in the network which is infinitely large.
2. Homogeneous individuals: All individuals are of the same type. If un-infected, they are equally susceptible. If infected, they are equally infectious.
3. The infectiousness within an infected individual does not change throughout the rest of its lifetime.

Under these assumptions, the infectious contact process $\{K(x) : x \geq 0\}$ is a homogeneous Poisson process (Chap. 3 Definition 11) which has stationary increments (Sect. 3.3) $\beta(x) = \beta$. The Euler-Lotka equation becomes

$$\beta \int_0^\infty e^{-rt}A(t)dt = \beta L[A](r) = 1 \tag{4.37}$$

where $L[A](s) = \int_0^\infty e^{-st}A(t)dt$, $s > 0$ is the Laplace transform of the function $A(t)$.

The function $A(x) = \Pr(X(x) = 1)$ is formulated through the Susceptible-Exposed-Infectious (SEI) structure, corresponding to many compartment transmission models that involve a random latent period T_E and a random infectious period T_I in sequence, including SIR, SEIR, SIS, SEIS, SEIRS, among many others.

Since $A(x) = \Pr(X(x) = 1) = \Pr$(the infected individual is infectious at time x since infection), it can be shown that $\int_0^\infty A(x)dx = \mu_I$, the mean infectious period. Therefore, $R_0 = \beta\mu_I$. When the infectious contact process $\{K(x) : x \geq 0\}$ has stationary increments, the presence of a latent period has no effect on R_0 but does have a strong effect on the initial growth rate r.

When There Is No Latent Period

When there is no latent period, $A(x) = \overline{F}_I(x) = \Pr(T_I > x)$, (4.37) becomes

$$\beta \int_0^\infty e^{-rt}\overline{F}_I(t)dt = \beta L[\overline{F}_I](r) = 1. \tag{4.38}$$

Integration by parts, $L[\overline{F}_I](r) = \frac{1}{r}(1 - L[f_I](r))$ and (4.38) becomes

$$r = \beta \left(1 - \int_0^\infty e^{-rt} f_I(t)dt\right) = \beta (1 - L[f_I](r)). \tag{4.39}$$

1. Given cumulatively infected $C(t)$ individuals by time t, each of them carries an infectious contact rate β but only a proportion of them is infectious at time t. The potential transmission rate r is thus β scaled by the proportion of infected individuals who are infectious at time t, represented by $1 - L[f_I](r)$. Let Y be a "dummy" exponentially distributed random variable with rate r. Assuming Y and T_I are independent, then $1 - L[f_I](r) = \Pr(Y \leq T_I)$ and $r = \beta \Pr(Y \leq T_I)$. Therefore, $r \leq \beta$.
2. The factor $1 - L[f_I](r)$ can be interpreted as a sampling probability. Let us consider a snapshot sample is taken at time t. We introduce a binary indicator $\Delta_i(t)$ such that $\Pr(\Delta_i(t) = 1)$ if an individual i, infected before t, is infectious at time t and 0 otherwise. Thus $1 - L[f_I](r) = \Pr(Y \leq T_I) = \Pr(\Delta_i(t) = 1)$ as

$$r = \beta \Pr(\Delta_i(t) = 1). \tag{4.40}$$

   The $C(t)$ individuals satisfying $\Delta_i(t) = 1$ form a "prevalence cohort" (see Fig. 4.8). Because r is constant, it implies that $\Pr(\Delta_i(t) = 1)$ does not depend on the sampling time. This involves equilibrium conditions to be further discussed in Sect. 4.4.
3. Assuming that the outbreak starts with a single initially seeded infectious individual and the infectious contact process $\{K(x) : x \geq 0\}$ is a homogeneous

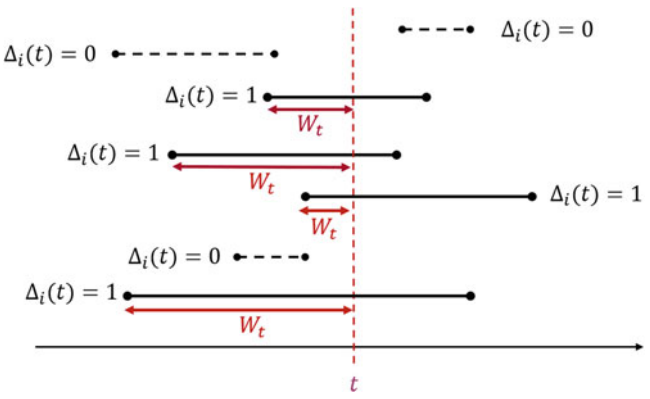


Fig. 4.8 Among the 7 individuals in the figure, 6 individuals are infected by time t (i.e. $C(t) = 6$) and 4 individuals are infectious at time t. These 4 individuals form the prevalence cohort. W_t is the time from infection to the sampling time t for individuals in the prevalence cohort and W_t may depend on t unless suitable equilibrium conditions are met

Poisson process, the probability of extinction δ during the initial phase satisfies the equation $G_N(\delta) = L[f_I](\beta(1-\delta)) = \delta$. When $R_0 > 1$, $\delta < 1$ and (4.39) gives

$$r = \beta\,(1-\delta) \tag{4.41}$$

under suitable equilibrium conditions. In this case, the invasion probability $1 - \delta = \Pr(\Delta_i(t) = 1)$.

We have seen that the marginal distribution of N, which is the number of infectious contacts produced by a typical infected individual during its entire infectious period, uniquely defines R_0 and determines the invasion probability $1-\delta$. However, different stochastic mechanisms may yield the same marginal distribution N but different initial growth rate r, as shown in the following example.

Example 15 The marginal distribution for N follows a geometric distribution with mean value R_0. We have seen that $\delta = 1/R_0$; the distribution of generation numbers to extinction is given by (4.11); the distribution of final size of the small outbreak Z_s given by (4.26), along the associated mean and variance. All these results are uniquely determined by the p.g.f. of the geometric distribution (4.9). The same marginal distribution can arise from

1. an infectious contact process $\{K(x) : x \geq 0\}$ as a homogeneous Poisson process with stationary increment β, combined with exponentially distributed infectious periods with mean value μ_I, such that (4.41) is true and $r = (R_0 - 1)/\mu_I$;
2. a mixed Poisson process for $\{K(x) : x \geq 0\}$ arising from (3.22) in which $u(\beta)$ is exponentially distributed, combined with a constant infectious period $T_I = \mu_I$. In this case (4.41) does not hold. However, it can be derived from (4.39) that r and R_0 are related through $r\mu_I = R_0(1 - e^{-r\mu_I})$.

Since $R_0 = \beta \int_0^\infty \overline{F}_I(t)dt = \beta\mu_I$, another way of re-writing (4.38) is to multiply both sides of $\int_0^\infty e^{-rt}\overline{F}_I(t)dt = 1/\beta$ by μ_I^{-1} (assuming $0 < \mu_I < \infty$) so that

$$L[f_W](r) = \int_0^\infty e^{-rt} f_W(t)dt = 1/R_0. \tag{4.42}$$

This equation gives a direct link between r and R_0, in which $f_W(t) = \overline{F}_I(t)/\mu_I$ is a p.d.f. recognized as the equilibrium distribution of the infectious period. The random variable $W > 0$ is called the backward recurrence time in the theory of renewal processes. In the current context, it is the duration between the time of infection of a typical infected individual in the prevalence cohort and time t. Naturally this duration should depend on time t, denoted as W_t in Fig. 4.8. Only under suitable equilibrium conditions W is independent of t with p.d.f. $f_W(t) = \overline{F}_I(t)/\mu_I$. Assuming Y and W are independent (implied by the assumption that Y and T_I are independent), (4.42) gives $L[f_W](r) = \Pr(Y > W) = 1/R_0$. It may be understood in the following ways.

1. If a typical infected individual in the prevalence cohort at time t produces, on average, $R_0 > 1$ secondary infections, it is expected that a proportion $1 - 1/R_0$ occurred before time t.
2. If the infectious period distribution is completely known and if the system is under equilibrium, then $f_W(t)$ is completely specified. IF, big IF, the parameter r can be estimated, then (4.42) is useful to estimate the basic reproduction number R_0.

Variability of the Infectious Period on the Growth Rate r and on the Invasion Probability $1 - \delta$ Among infectious period distributions with the same mean μ_I, variability among T_I as defined by the convex order (Chap. 2 Definition 6) gives the order of the Laplace transform. From Chap. 2, Definition 7, among infectious periods with equal mean value μ_I, the more variable the infectious period (by convex order), the larger the $L[f_I](s)$, for all $s > 0$ and the smaller the value

$$L[f_W](s) = \frac{1}{\mu_I} L[\overline{F}_I](s) = \frac{1}{s\mu_I} \left(1 - L[f_I](s)\right), \text{ for all } s > 0.$$

Recall (Chap. 2) that $L[f_W](s)$ is a log-convex, monotonically decreasing function of s satisfying $L[f](0) = 1$ and approaches zero as $s \to \infty$, keeping the basic reproduction number R_0 fixed, the more variable the infectious period, the smaller the value of the intrinsic growth rate r. This is an immediate result from (4.42). With respect to the invasion probability, (4.41) is valid if the infectious contact process $\{K(x) : x \geq 0\}$ is a homogeneous Poisson process. In that case, keeping the basic reproduction number R_0 fixed, the more variable the infectious period, the smaller the invasion probability $1 - \delta$.

In the Presence of a Latent Period

We assume that T_E and T_I are sequential and mutually independent. The sample path of the binary process $\{X(x) : x \geq 0\}$ is defined such that $X(0) = 0$ with a sojourn time according to the latent period distribution before making a jump to state 1; then it stays in state 1 according to the infectious period distribution; by the end of the infectious period, it jumps to state 0 and remains in state 0 as $t \to \infty$. Under suitable equilibrium conditions,

$$A(x) = \Pr(X(x) = 1) = \Pr(\{T_E \leq x\} \cap \{T_I > x - T_E\}) = \int_0^x \overline{F}_I(x-u) f_E(u) du$$

and as shown in Yan (2008a,b), (4.38) is extended to

$$\beta \left(\int_0^\infty e^{-rt} f_E(t) dt \right) \left(\int_0^\infty e^{-rt} \overline{F}_I(t) dt \right) = \beta L[f_E](r) L[\overline{F}_I](r) = 1. \tag{4.43}$$

Because $L[\overline{F}_I](r) = \frac{1}{r}\left(1 - L[f_I](r)\right)$, (4.43) can be re-written as

$$r = \beta L[f_E](r)\left(1 - L[f_I](r)\right) \tag{4.44}$$

which is the extension of (4.39).

1. The factor $L[f_E](r)\left(1 - L[f_I](r)\right)$ in (4.44) is still interpreted as the sampling probability. Given cumulatively infected $C(t)$ individuals by time t, each of them carries an infectious contact rate β but only a proportion of them is infectious at time t. The potential transmission rate r is thus β scaled by the proportion of infected individuals who are infectious at a snapshot sample at time t. Let Y be a "dummy" exponentially distributed random variable with rate r. Assuming Y is independent from both T_E and T_I, $L[f_E](r) = \Pr(Y > T_E)$ and $1 - L[f_I](r) = \Pr(Y \leq T_I)$. Because the distribution of Y is memoryless, $\Pr(T_E < Y \leq T_E + T_I | Y > T_E) = \Pr(Y \leq T_I)$. Therefore $L[f_E](r)\left(1 - L[f_I](r)\right) = \Pr(T_E < Y \leq T_E + T_I)$ and

$$r = \beta \Pr(T_E < Y \leq T_E + T_I). \tag{4.45}$$

   Therefore, $r \leq \beta$ and $\Pr(T_E < Y \leq T_E + T_I)$ is the proportion of infected individuals who are infectious at time t.
2. Adding a latent period, Eq. (4.41) is no longer valid, because r depends on both the distribution of the latent period and the infectious period whereas δ depends on the infectious period distribution only. In fact, $r < \beta\left(1 - \delta\right)$.

   The extension of (4.42) is

$$L[f_E](r)L[f_W](r) = 1/R_0, \tag{4.46}$$

where $R_0 = \beta\mu_I > 1$. Since the Laplace transform of the sum of two independent random variables equals the product of the two Laplace transforms, we define

$$T_G = T_E + W$$

where W has p.d.f. $f_W(t) = \overline{F}_I(t)/\mu_I$ and T_G has p.d.f. $f_G(t)$ calculated as the convolution between $f_E(t)$ and $f_W(t)$. Therefore (4.46) can be further written as

$$L[f_G](r) = 1/R_0 \tag{4.47}$$

where $L[f_G](s) = \int_0^\infty e^{-st} f_G(t)dt$, $s > 0$ is the Laplace transform of $f_G(t)$.

Remark Denote $\mu_G = E(T_G)$ and $\mu_E = E(T_E)$, the expected value $E(T_G) = E(T_E) + E(W)$ gives

$$\mu_G = \mu_E + \frac{1}{2}(1 + \phi^2)\mu_I = \mu_E + \frac{\mu_I}{2} + \frac{var[T_I]}{2\mu} \tag{4.48}$$

where $E(W) = \frac{1}{2}(1+\phi^2)\mu_I$ and ϕ is the coefficient of variation of the infectious period T_I defined as the ratio of the standard deviation to the mean. $T_G = T_E + W$ and μ_G is defined with infected individuals along its progression of infectiousness, without involving contacts with other individuals. Many authors in the literature have called μ_G the average generation time. For example:

- Page 17 of Daley and Gani (1999) defined the average generation time as $\mu_G = \mu_E + \mu_I/2$. This expression also appears in Section 2.1 of Roberts and Heesterbeek (2007). This is the case when $\phi = 0$ in (4.48) in which the infectious period is not random.
- Page 14 of Anderson and May (1991) defined the average generation time as $\mu_G = \mu_E + \mu_I$. It also appears in Section 2.4 of Roberts and Heesterbeek (2007). This is the case when $\phi = 1$ in (4.48). This is, but not limited to, the case when the infectious period is exponentially distributed.
- Some models assume that the infectious period follows an Erlang distribution, that can be expressed as a sum of n independently and identically distributed segments following exponential distribution with mean μ_I/n, then $\phi^2 = 1/n$ and $\mu_G = \mu_E + \frac{n+1}{2n}\mu$. This expression can be found in Section 2.3 of Roberts and Heesterbeek (2007).

Variability of the Latent Period on the Growth Rate r We assume that $R_0 = \beta\mu_I$ and the distribution of the infectious periods T_I is all fixed. Among latent periods with the same mean value μ_E, the more variable the latent period T_E (by convex order), the larger the $L[f_E](s)$, for all $s > 0$; therefore, the larger the value of r. However, the distribution of the latent period has no influence on $1 - \delta$.

Relationships Between r and R_0 for Gamma Distributed T_E and T_I

Re-writing (4.44) and (4.46), we have

$$R_0 = \frac{1}{L[f_G](r)} = \frac{r\mu_I}{L[f_E](r)\,(1 - L[f_I](r))}. \tag{4.49}$$

This leads to some specific expressions when the distributions for T_E and T_I have explicit analytical Laplace transforms. If both the latent period and the infectious period are gamma distributed, with $var(T_E) = \mu_E^2/\kappa_E$ and $var(T_I) = \mu_I^2/\kappa_I$, using the corresponding Laplace transforms given by (2.33) in Chap. 2, (4.49) becomes

$$R_0 = \frac{r\mu_I\,(1 + r\mu_E/\kappa_E)^{\kappa_E}}{\left[1 - (1 + r\mu_I/\kappa_I)^{-\kappa_I}\right]} \tag{4.50}$$

which was originally given by Anderson and Watson (1980). When both the latent and the infectious periods are exponentially distributed, with $\mu_E = 1/\alpha$ and $\mu_I = 1/\gamma$, then

$$R_0 = (1 + r\mu_E)(1 + r\mu_I) = \frac{(r + \alpha)(r + \gamma)}{\alpha\gamma}. \tag{4.51}$$

When There Is No Latent Period In this case, (4.50) becomes

$$R_0 = \frac{r\mu_I}{1 - (1 + r\mu_I/\kappa_I)^{-\kappa_I}}.$$

Its special cases include

- $R_0 = r\mu_I/[1 - \exp(-r\mu_I)]$, if the infectious period is constant $T_I = \mu_I$;
- $R_0 = 1 + r\mu_I$, if the infectious period is exponentially distributed.

The variance of the gamma distributed infectious period is $var[T_I] = \mu_I^2/\kappa_I$ and the shape parameter κ_I ranks both the variance and the Laplace transform of the infectious period. The following example shows how κ_I ranks the exponential growth rate r given the same value of R_0.

Example 16 Assuming gamma distributed infectious periods with mean value $\mu_I = 10$ and variance $var[T_I] = 100/\kappa_I$. The basic reproduction number is $R_0 = 3$. When $\kappa_I = 1$, $r = (R_0 - 1)/\mu_I = 0.2$. When $\kappa_I \to \infty$, the infectious period approaches a constant value $T_I = \mu_I$ and r can be numerically solved as $r = 0.28214$. If the infectious period is variable with variance smaller than that of the exponential distribution, then $0.2 < r < 0.28214$. If $var[T_I]$ is larger than that of the exponential distribution, then $r < 0.2$, for instance, $r = 0.084452$ when $\kappa_I = 0.2$. These exponential growth curves are illustrated in Fig. 4.9.

Special Cases with a Latent Period

- If both the latent period and the infectious period are exponentially distributed, $L[f_E](r) = (1 + r\mu_E)^{-1}$ and $L[f_I](r) = (1 + r\mu_I)^{-1}$. In this case, $R_0 = (1 + r\mu_E)(1 + r\mu_I)$.
- If both the latent period and the infectious period are not random, then $L[f_E](r) = e^{-r\mu_E}$, $L[f_I](r) = e^{-r\mu}$ and $R_0 = r\mu_I/[\exp(-r\mu_E)(1 - \exp(-r\mu_I))]$.

The shape parameter κ_E ranks both the variance and the Laplace transform of the latent period. The following example shows how κ_E ranks the exponential growth rate r, given the same value of R_0 and a fully specified infectious period distribution.

Example 17 The basic reproduction number is $R_0 = 3$. The infectious period has mean value $\mu_I = 1$ and variance $var[T_I] = 1$, exponentially distributed. The latent

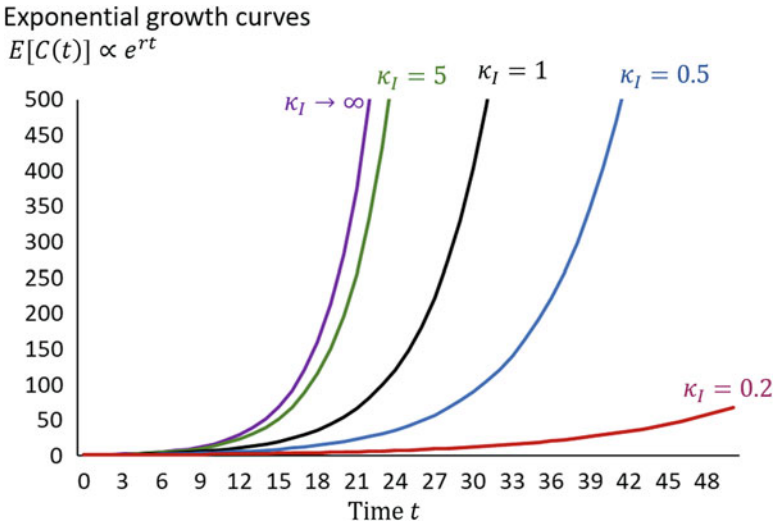


Fig. 4.9 Exponential growth curves when $R_0 = 3$ with gamma distributed infectious periods without latent periods. Special cases are: $\kappa_I \to \infty$ (constant infectious period) and $\kappa_I = 1$ (exponentially distributed infectious period)

period has mean $\mu_E = 7$, gamma distributed with $var[T_E] = 49/\kappa_E$. The Euler-Lotka equation has the special form

$$\frac{3}{1+r}\left(1+\frac{7r}{\kappa_E}\right)^{-\kappa_E} = 1.$$

We show that, as κ_E increases, the less variable is the latent period T_E. Consequently, the value of r is smaller. When $\kappa_E = 1$ (exponential latent period), $var[T_E] = 49$, $r = 0.211$ and $e^{0.211t}$ is illustrated as line 1 in Fig. 4.10. When $\kappa_E = 2$, $r = 0.1715$ shown as line w in Fig. 4.10. When $\kappa_E = 5$, $r = 0.1509$ shown as line 3 in Fig. 4.10. When $\kappa_E \to \infty$, r approaches the limit $r = 0.13842$ which is the solution of $3e^{-7r} = 1 + r$. Note that this is very close to the value $\frac{1}{8}\log(3) = 0.13733$ corresponding to a discrete function given by

$$E[C(t)] = \begin{cases} e^{0.13733t}, & \text{if } t = 8g,\ g = 1, 2, \ldots \\ 0, & \text{otherwise} \end{cases}, \tag{4.52}$$

in which the factor 8 in $t = 8g$, $g = 1, 2, \ldots$ equals $\mu_E + \mu_I$. Thus $E[C(t)]$ in (4.52) is the exponential growth $E[X_g] = R_0^g$ at $R_0 = 3$ in real time separated by the generation time $\mu_E + \mu_I$ in the sense of Anderson and May (1991), illustrated by bars in Fig. 4.10.

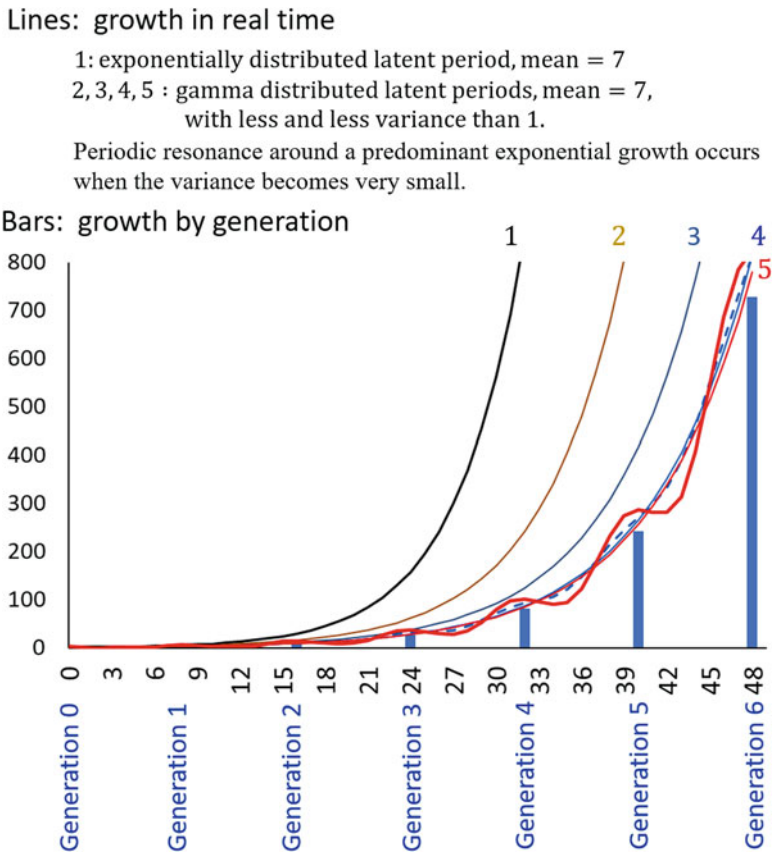


Fig. 4.10 A schematic illustration of the initial growth by generation and in continuous time after seeding $i_0 = 1$ infected individual assuming $R_0 = 3$

Relationships Between r and R_0 When the Exact Distributions for T_E and T_I Are Unknown

Assuming that $T_G = T_E + W$ can be specified to its first and the second moments, by expanding $R_0 = 1/L[f_G](r)$, the following approximation is valid

$$R_0 \approx 1 + r\mu_G + \frac{1}{2}r^2\left(\mu_G^2 - var\,(T_G)\right). \tag{4.53}$$

If both the latent and the infectious periods are exponentially distributed, the above expression is no longer an approximation but an exact result. In this case, W is identically distributed as T_I and T_G is the convolution of two exponentially distributed random variables $T_E + T_I$. The mean value is $\mu_G = \mu_E + \mu_I$ and $var\,(T_G) = \mu_E^2 + \mu_I^2$. Then (4.53) gives $R_0 = 1 + r\,(\mu_E + \mu_I) + r^2\mu_I\mu_E = (1 + r\mu_E)\,(1 + r\mu_I)$.

If the latent period is gamma distributed with mean μ_E and variance μ_E^2/κ_E and the infectious period is gamma distributed with mean μ_I and variance μ_I^2/κ_I, $T_G = T_E + W$ has the mean $\mu_G = \mu_E + \frac{\kappa_I+1}{2\kappa_I}\mu_I$ and variance $var\,(T_G) = \mu_E^2/\kappa_E + \frac{\kappa_I^2+6\kappa_I+5}{12\kappa_I^2}\mu_I^2$. The approximation (4.53) gives

$$R_0 \approx 1 + r\left(\mu_E + \frac{\kappa_I+1}{2\kappa_I}\mu_I\right) + \frac{1}{2}r^2\left(\frac{(\kappa_I^2-1)}{6\kappa_I^2}\mu_I^2 + \mu_E\left(\frac{\kappa_E-1}{\kappa_E}\mu_E + \frac{\kappa_I+1}{\kappa_I}\mu_I\right)\right)$$

which corresponds to the first two terms of the series expansion given by Anderson and Watson (1980).

4.4 On Assumptions and Conditions

The behavior of an outbreak during its initial phase is an approximation based on a list of assumptions and conditions. In the following, we examine some of the important assumptions and conditions. While these assumptions and conditions have provided useful theoretical insights, most of them are questionable in real-life applications.

It is not uncommon in data collected during outbreak investigations that the outbreaks grow with periodic waves along a predominant trend. There are numerous examples in the literature including smallpox outbreaks, SARS in 2003, H1N1 in 2009, and more recently, Ebola outbreaks in West Africa. Numerous historic and contemporary outbreaks have displayed sub-exponential growth patterns in the early ascending phase of infectious disease outbreaks (Viboud et al. 2006; Chowell et al. 2016).

4.4.1 The Initial Phase

The initial phase is based on the premises that the population is extremely large ($m \to \infty$) and during which the depletion of the susceptible population is negligible. It assumes that there exists such an initial phase during which a typical infected individual produces the next generation of secondary infections according to the distribution of N which does not change over generations. Hence, the basic reproduction number R_0 remains the same for the first few generations. Meanwhile, it is assumed that the initial growth rate r is applicable to all infected individuals during the initial phase.

It is understood that both the discrete time and continuous time branching processes are models for the initial phase by approximation. The essential parameters such as R_0 and r, along with equations developed under suitable equilibrium conditions, are used to approximately characterize the early features of the outbreak when the system is moving away from equilibrium in order to obtain important theoretical insight. They may not agree with observed data in practice.

Independency and Homogeneity

A central assumption in all the models discussed in this chapter so far is that infected individuals produce new infections independently. For the Galton-Watson branching process approximation, most of the results are based on a single seeded infected individual $i_0 = X_0 = 1$. The results such as the invasion probability; the distribution of generation numbers to extinction and the distribution of final size if it is a small outbreak can be straightforwardly generalized for $i_0 > 1$ by viewing them as i_0 separate independent branching processes with each starting with a single infected individual. For example, let δ be the smallest root of the fixed point equation $G_N(s) = s$ in $s \in (0, 1]$, the probability of extinction is δ^{i_0} and the invasion probability is $1 - \delta^{i_0}$.

With respect to the linear pure birth process (3.41), all the $C(t) = n$ infected individuals are *homogeneous* in the sense that they carry the same intrinsic rate r to produce a new infection independently from the other $n - 1$ individuals. The time to increase the number of infected individuals from n to $n + 1$ in the population is the first of the n order statistics of the exponential distribution with rate r and the instantaneous rate of the increase is nr. This yields the exponential growth in the expected number $E[C(t)]$.

The Euler-Lotka equation in its general form (4.36) is derived from the renewal-type equation (4.35) which assumes independency. The second term $\int_0^t \beta(x)A(x)E[C(t-x)]dx$ involves the expected number of infected individuals $E[C(t-x)]$ at time $x \in [0, t)$ and each of them independently produces new infections with instantaneous infection rate $\beta(x)A(x)$.

The Euler-Lotka equation (4.37) and its variations (4.38)–(4.46) involve assumptions of the infectious contact process $\{K(x) : x \geq 0\}$ at the individual level. It is often assumed that $\{K(x) : x \geq 0\}$ is a Poisson process which also involves homogeneity among individuals.

The Constant Growth Rate and Asymptotics

The assumption of a constant growth rate r leads to the continuous time exponential growth $\frac{d}{dt}E[C(t)] = rE[C(t)]$ given by the pure birth process (3.41). The exponential growth can be also derived from a renewal-type equation $E[C(t)] = 1 + r\int_0^t E[C(t-x)]dx$. However, the renewal-type equation (4.35) does not lead to exponential growth unless $\beta(x)A(x) = r$, which may be too strong an assumption.

A milder assumption is $\beta(x) = \beta$, that is, the infectious contact process $\{K(x) : x \geq 0\}$ has stationary increments.

The Euler-Lotka equations (4.36) and (4.37) are formulated under the premises that the initial growth is exponential and useful to establish asymptotic (as $t \to \infty$) linkages between the growth rate r and the stochastic mechanisms involving the infectious contact rate $\beta(x)$ and the distributions of the latent and the infectious periods. Especially, (4.37) gives the asymptotic relationships in the form of

$$r = \beta \times \text{sampling probability of being infectious at } t \tag{4.54}$$

with various expressions (4.39)–(4.41) and (4.44)–(4.46).

Equilibrium Conditions

In infectious disease models, the equilibrium condition can be only approximated in several special occasions during an epidemic: (1) at the very beginning when an infected individual is seeded but the system is at disease-free equilibrium; (2) at the end of an outbreak when there is no infectious individual left in the population so that the system returns to disease-free equilibrium; (3) the endemic equilibrium, arising in situations where the depletion of the susceptible population can be replaced by the loss of immunity of individuals who have recovered from infection as described by compartment models SIS, SIRS, SEIRS, etc., or replaced by demography for diseases that are also chronic; or (4) at some transient time when the effective reproduction number is approximately unity.

The distribution of N such that $R_0 = E[N]$ applies to Generation Zero. The initial growth rate r applies to $t = 0$. Both parameters are intrinsic in the initially seeded individuals when the system is at disease-free equilibrium, but are applied to all infected individuals during the initial phase.

According to the asymptotic relationship (4.54), the sampling proportion of those who are "currently" infectious out of all previously infected individuals must be independent of time t. All infected individuals have the same probability of being infectious at any snapshot of time t. Thus, all the $C(t) = n$ infected individuals are homogeneous, carrying the same intrinsic rate r. Consequently, this leads to exponential growth $\frac{d}{dt}E[C(t)] = rE[C(t)]$.

This is the same equilibrium condition required in the Euler-Lotka equations corresponding to $\beta(x) = \beta$ derived from (4.37), as discussed with respect to (4.38)–(4.40), as well as reflected by the equilibrium distribution $f_W(x) = \overline{F}_I(x)/\mu_I$, assuming that the backward recurrence time W_t as illustrated in Fig. 4.8 corresponds to the same random variable W. With respect to $f_W(x)$, it only requires constant incidence whereas the value of the incidence number (or rate) does not play any role.

The equilibrium conditions also play a central theme in mathematical epidemiology beyond the initial phase. In later chapters in this book, we shall see important transcendental relationships in compartment models that connect R_0, defined at

$t = 0$, with parameters in these models through explicit functional forms by setting the system "at equilibrium." We shall also encounter the "final size equation" for large outbreaks which shows that R_0, defined at the disease-free equilibrium at the beginning, transcends to the final proportion of individuals who have "escaped" from infection at the disease-free equilibrium when the outbreak ends. A similar transcendental relationship exists between R_0 and the proportion of susceptible individuals in the population when the system is at endemic equilibrium for diseases in which the depletion of the susceptible population can be replaced by the loss of immunity of individuals who have recovered, or for diseases that are also chronic such that the susceptible population can be replaced by demography.

Between Theory and Practice

Theories are fundamental in scientific understanding. For example, a transcendental relationship between R_0 and $0 < \eta < 1$ in $E[C(\infty)] = m\eta$, as described in Sect. 4.2.3, has been proven in many types of disease transmission models, should a large outbreak end, where η is the final proportion of the population eventually infected. This final size equation will be further discussed in Chap. 5 together with its usage in planning and evaluating effectiveness of public health control measures.

On the other hand, all the assumptions and conditions will be challenged in real life. The basic reproduction number R_0 and the initial phase are theoretical concepts. In practice, for any finite susceptible population, it is not appropriate to assume that the depletion of the susceptible population is negligible. Individuals are more likely made of different types and carry different intrinsic characteristics. The environment in which the transmission occurs is more likely to be highly heterogeneous. The initial exponential growth and its relation to R_0 are established under equilibrium conditions that require constant incidence. Applying the quantities and equations established under such conditions to study the ascending phase of an epidemic seems to be paradoxical. It remains a challenge in developing early assessment tools on key epidemic parameters based on early data, as well as assessment of biases in existing estimation methods in the literature.

4.5 Alternative Initial Growth Curves

4.5.1 Periodic Resonance Around a Predominant Exponential Growth

Under the premises that the initial growth is exponential, the intrinsic growth rate r is connected to the underlying stochastic mechanisms $\beta(x)A(x)$ through (4.36) as the unique real value such that $\int_0^\infty e^{-rx}\beta(x)A(x)dx = 1$.

Figure 4.10 (line 4 and line 5) shows the phenomena of periodic resonance around a predominant exponential growth. They are created under the assumptions in Example 17, in which $\beta(x) = \beta = 3$ but $A(x)$ is periodic over time, by assuming a long latent period T_E with $\mu_E = 7$ and a short infectious period T_I with $\mu_I = 1$. Since the infectious period is assumed to be exponentially distributed, then $\mu_G = \mu_E + \mu_I = 8$ is the generation time (Anderson and May 1991).

Bacaër and Abdurahman (2008) carefully studied the complex roots of the Euler–Lotka equation in its general form (4.36). They found that (4.36) often has a unique positive real root $r > 0$ plus pairs of conjugate complex roots. They postulated that when there is no complex root, or when the real parts of the complex roots are all negative, then the outbreak growth is exponential. If there exist $m \geq 1$ pairs of conjugate complex roots $a_j \pm b_j i$ with $a_j > 0$, the growth curve should be represented as

$$e^{rt} + \sum_{j=1}^{m} e^{a_j t} \cos(b_j t), \tag{4.55}$$

which gives periodic resonance around a predominant exponential growth e^{rt}. The resonance is driven by a pair of conjugate complex roots with the largest real part value a and the resonance becomes more pronounced when a is close to the Malthusian number r.

In Example 17, there is no conjugate complex roots with positive real parts when $\kappa_E < 19$. At $\kappa_E = 19$, there is a positive real root $r = 0.1416$ and a pair of conjugate complex roots $5.2437 \times 10^{-4} \pm 0.82332i$. In this case, the resonance is invisible. However, as κ_E increases, the value a for the pair of the conjugate complex roots with the largest real part value $a \pm bi$ also increases and the resonance becomes visible (e.g., line 4 of Fig. 4.10). When κ_E becomes very large, the variance of the latent period becomes very small. For instance, when $\kappa_E = 200$, $var[T_E] = 7^2/200 = 0.245$. In this case, $r = 0.13872$ which is very close to the limit 0.13842, meanwhile, there are two pairs of conjugate complex roots with real positive parts: $0.10107 \pm 0.81i$ and $0.01472 \pm 1.6522i$. The one with the largest real part has $a = 0.10107$ which is close to r, meanwhile $b = 0.81$ is close to $2\pi/\mu_G = 0.7854$. The resonance becomes very pronounced as indicated by line 5 in Fig. 4.10.

Example 17 has created scenarios that $A(x)$ is periodic due to the relatively long and concentrated (i.e., small variation) latent period under the assumption $\beta(x) = \beta$. The resonance is spaced by the mean generation time. When the latent period has larger variance, the generation spacing is lost, and the growth curve is smooth and tends to rise more rapidly.

Although Example 17 is very artificial, the periodic resonance during the initial growth is quite common in empirical data. Bacaër and Abdurahman (2008) further examined the Euler–Lotka equation in a more general form $\int_0^\infty e^{-\rho x} \beta(x,t) A(x) dx = 1$. In this expression, $\beta(x,t)$ is the intensity function of the infectious contact process that may depend on a fluctuating environment where transmission occurs. There are two time scales: x is the time from infection within

a typical infected individual and t is the calendar time from the initial seeding of an infected individual. The authors provided an extensive list along with references, such as periodic contact rates due to periodic demography, periodic contact patterns (weekdays and weekends) in school and workplace environments, periodic vector or reservoir, just to name a few. These could be amplified by periodic control measures such as pulse vaccination, periodic antiviral treatment, and so on. The authors studied a specific periodic model for $\beta(x,t) = \beta_0(x)(1+\varepsilon\cos\omega t)$. They found that resonance of the growth rate occurs when the Euler–Lotka equation has a complex root with an imaginary part close to ω of the contact rate and a real part not too far from the Malthusian parameter r. They further studied the SEIR and SIR structured models with the modeled $\beta(x,t)$.

Bacaër and Abdurahman (2008) pointed out that (1) For an SIR model with exponentially distributed infectious period, resonance of the initial growth rate is impossible. They emphasized that this model is exceptional in the sense that the initial growth rate is even completely independent of the frequency of the periodic factor. (2) For an SEIR model with exponentially distributed latent and infectious periods, resonance is impossible but the growth rate depends on the periodic factor. We recommend this paper, and emphasize the separate roles of the contact rates $\beta(x,t)$, the latent period distribution and the infectious period distribution in shaping the initial growth of an infection curve.

4.5.2 The Sub-exponential Growth

Although we address the initial growth with respect to the expected value of the cumulative counts $E[C(t)]$, we use the notation $C_d(t)$ for the deterministic function that dominates the underlying trend.

The exponential growth $C_d'(t) = rC_d(t)$ corresponds to the pure birth process given by the conditional probability $\Pr\{C(t+h) = n+1 | C(t) = n\} = nrh + o(h)$. The underlying assumption is that given $C(t) = n$, the instantaneous rate of the next infection arises as the first of the order statistics of i.i.d. exponentially distributed lifetimes of sample size n with constant rate r.

In general, given $C(t) = n$, the instantaneous rate of the next infection arises as the first of the order statistics of i.i.d. lifetimes of sample size n with hazard function $\rho(t)$. The pure birth process is now specified by $\Pr\{C(t+h) = n | C(t) = n\} = 1 - \rho(t)nh + o(h)$, $\rho(t) > 0$ and its deterministic counterpart is $\frac{d}{dt}C_d(t) = \rho(t)C_d(t)$, or equivalently, $\rho(t) = \frac{d}{dt}\log C_d(t)$. Conversely, $C_d(t) = i_0 e^{\int_0^t \rho(x)dx}$. We call $\rho(t)$ the instantaneous growth rate function.

There are different notions of sub-exponential growth in various disciplines of mathematics, such as growth functions that are bounded by exponential functions above it and polynomial functions below it. Supported by a diversity of epidemic growth profiles from empirical data of historic and contemporary outbreaks (Viboud et al. 2016; Chowell et al. 2016), the early ascending phase of infectious disease

outbreaks has shown sub-exponential growth patterns, bounded by the exponential growth and the linear growth.

Definition 18 The sub-exponential growth function $C_d(t)$ is a convex function bounded by the linear growth and the exponential growth, that is, given $C_d(0) = i_0$,

$$i_0\,(1 + rt) \le C_d(t) \le i_0 e^{rt}. \tag{4.56}$$

If $C_d(t)$ is a sub-exponential growth function satisfying (4.56), $\rho(t)$ belongs to a class of completely monotonic functions that are bounded by $\frac{r}{1+rt}$ from below, satisfying

$$\frac{r}{1+rt} \le \rho(t) = \frac{d}{dt}\log C_d(t) \le r, \quad \text{for all } t \ge 0. \tag{4.57}$$

The lower bound $\frac{r}{1+rt}$ is the hazard function of the Pareto-I distribution, which is completely monotonic as defined by (2.32) and log-convex. This class includes, but is not limited to, the following completely monotonic functions:

1. $\rho(t) = \frac{r}{1+rvt}$, $r > 0$, $0 < v \le 1$ that gives

$$C_d(t) = i_0(1 + rvt)^{\frac{1}{v}}, \ r > 0 \text{ and } 0 < v \le 1. \tag{4.58}$$

2. $\rho(t) = \frac{r}{\sqrt{1+2rvt}}$, $r > 0$, $0 < v \le 1$ that gives

$$C_d(t) = i_0 e^{\left(\sqrt{1+2vrt}-1\right)/v}, \ r > 0 \text{ and } 0 < v \le 1.$$

In these expressions, r plays a role as a scale parameter for time t. Figure 4.11 illustrates some of these sub-exponential rate and growth functions assuming $i_0 = 1$.

The Sub-exponential Growth as Approximation for the Convex Increasing Part of the Logistic Growth

The logistic growth function will be discussed in more detail in Chap. 5, which will also show that many dynamic transmission models lead to the logistic growth function. The logistic function has played an important role as a population growth model and applied in epidemiology. It is defined by

$$C_{\text{logis}}(t) = K\frac{i_0}{i_0 + (K - i_0)\,e^{-\rho t}} = \frac{K}{1 + \frac{1}{v}e^{-\rho t}}, \tag{4.59}$$

which is a monotonically increasing function of t when $i_0 < K$ and $\lim_{t\to\infty} C_{\text{logis}}(t) = K$. It is the solution of the logistic differential equation $\frac{d}{dt}C_{\text{logis}}(t) = \rho C_{\text{logis}}(t)\left[1 - C_{\text{logis}}(t)/K\right]$ with the initial condition $C_{\text{logis}}(0) = i_0$

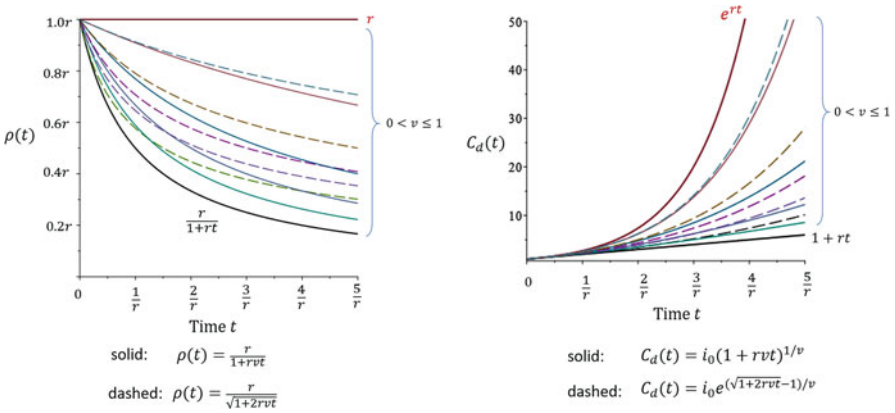


Fig. 4.11 Presentations of $\rho(t)$ and $C_d(t)$ for the sub-exponential growth with $i_0 = 1$ and different values of v: solid lines represent $\rho(t) = r/(1 + rvt)$ and $C_d(t) = (1 + rvt)^{1/v}$; dashed lines represent $\rho(t) = r/\sqrt{(1 + 2rvt)}$ and $C_d(t) = e^{(\sqrt{(1+2rvt)}-1)/v}$

and $\rho > 0$ characterizing the initial growth. It may be re-parameterized via $v = \frac{i_0}{K - i_0}$. We re-scale the initial growth as $r = \rho\left(1 - \frac{i_0}{K}\right) = \frac{\rho}{v+1}$. The logistic function can be written as

$$C_{\text{logis}}(t) = \frac{i_0\,(v+1)}{v + e^{-(v+1)rt}}. \tag{4.60}$$

When $i_0 < K/2$, we have $v < 1$ and $C_{\text{logis}}(t)$ increases as a convex function of t until the inflexion point $t^* = -\frac{1}{(1+v)r}\log v$ and $C_{\text{logis}}(t^*) = K/2$. During the convex increasing phase, $C_{\text{logis}}(t)$ satisfies the following inequalities:

$$i_0\,(1 + rt) < C_{\text{logis}}(t) < i_0(1 + rvt)^{\frac{1}{v}} < i_0 e^{rt}, \text{ for } t > 0.$$

By Definition (4.56), the convex increasing phase of the logistic function is sub-exponential, so is the function $i_0(1 + rvt)^{\frac{1}{v}}$ with $0 < v < 1$. Both are convex functions bounded by a linear function from below and an exponential function from above. Series expansion to the third order of t of these functions reveal that

$$C_{\text{logis}}(t) = i_0\left[1 + rt + \frac{1}{2}(1 - v)\,r^2t^2 + \frac{1}{6}\left(v^2 - 4v + 1\right)r^3t^3 + O\left(t^4\right)\right],$$

$$i_0(1 + rvt)^{\frac{1}{v}} = i_0\left[1 + rt + \frac{1}{2}(1 - v)\,r^2t^2 + \frac{1}{6}\left(2v^2 - 3v + 1\right)r^3t^3 + O\left(t^4\right)\right],$$

$$i_0 \exp(rt) = i_0\left[1 + rt + \frac{1}{2}r^2t^2 + \frac{1}{6}r^3t^3 + O\left(t^4\right)\right].$$

The *intrinsic* exponential growth with respect to the branching process approximation is under the assumption of an infinitely large population corresponding to the case $v = 0$, and in this case, $I_d(t) = i_0 \exp(rt)$ and $r = \rho$.

In finite populations, the exponential growth $i_0 \exp(rt)$ can also be used to approximate the initial growth of the logistic function, but with a discounted growth rate: $r = \frac{\rho}{v+1} = \rho\left(1 - \frac{i_0}{K}\right)$.

A closer approximation is (4.58)

$$C_{\text{sub-exp}}(t) = i_0(1 + rvt)^{\frac{1}{v}}$$

which grows slower than the exponential function $i_0 \exp(rt)$ and is sub-exponential.

The instantaneous growth rate $\rho(t) = \frac{d}{dt} \log C_d(t)$ corresponding to $C_{\text{logis}}(t)$ and $C_{\text{sub-exp}}(t)$ are

$$\rho_{\text{logis}}(t) = \frac{d}{dt} \log C_{\text{logis}}(t) = r \frac{1+v}{1 + ve^{(1+v)rt}},$$
$$\rho_{\text{sub-exp}}(t) = \frac{d}{dt} \log C_{\text{sub-exp}}(t) = r \frac{1}{1 + rvt}.$$

In both cases, $r = \rho_{\text{logis}}(0) = \rho_{\text{sub-exp}}(0)$ takes into account the population size and the initial condition since $r = \frac{K - i_0}{K}\rho$. The decreasing function $\rho_{\text{logis}}(t)$, in the context of the SI and SIS models, takes into account the depletion of the susceptible population. The decreasing function $\rho_{\text{sub-exp}}(t)$ is the close approximation.

Figure 4.12 compares $C_{\text{logis}}(t)$, $C_{\text{sub-exp}}(t)$, and the exponential growth $i_0 e^{rt}$ with $K = 100$, where $r = \rho\left(1 - \frac{i_0}{K}\right)$ and $v = \frac{i_0}{K - i_0}$. Three panels correspond to $i_0 = 1$,

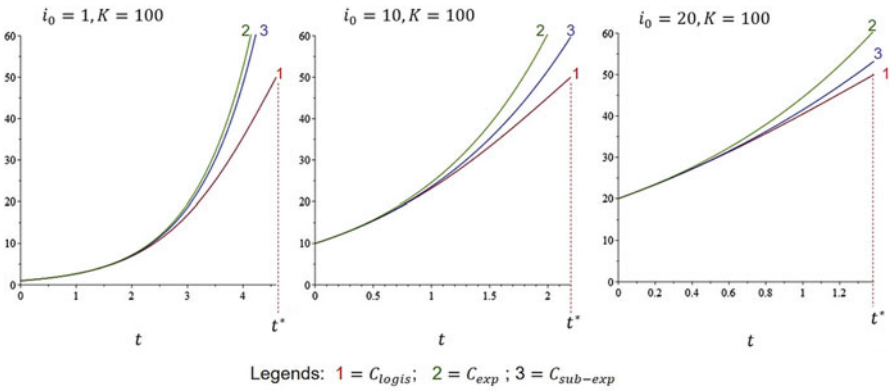


Fig. 4.12 The logistic growth $C_{\text{logis}}(t)$ given by (4.60), plotted up to the infexion point $t^* = -\log v$, is compared with the sub-exponential growth $C_{\text{sub-exp}}(t)$ and the exponential growth $i_0 e^{rt}$, where $r = (1 - i_0/K)$, in a population of size $K = 100$ and initial conditions $i_0 = 1, 10$ and $i_0 = 20$

10, and 20. Since ρ is a scale parameter of time, we set $\rho = 1$. Each panel is plotted for the convex increasing phase of $C_{\text{logis}}(t)$ that ends at the time of the inflexion point $t^* = \log \frac{K-i_0}{i_0} = -\log v$. Figure 4.12 shows that, when i_0/K is very small, the sub-exponential growth $C_{\text{sub-exp}}(t)$ resembles the property of the exponential growth. When i_0/K approaches $1/2$, $C_{\text{sub-exp}}(t)$ approaches linear growth.

An Alternative Formulation for the Sub-exponential Growth Function

Viboud et al. (2016) and Chowell et al. (2016) considered a 2-parameter generalized-growth model (GGM) given by

$$\frac{d}{dt}C_d(t) = rC_d(t)^{1-v},\ 0 < v \le 1. \tag{4.61}$$

This equation is also called the "power law exponential" equation by Banks (1994).

This generalized-growth model creates a sub-exponential growth via a "deceleration of growth" parameter, $p = 1-v$. The authors related this model to the *effective reproduction* over time t. Instead of tracking the depletion of the susceptible population, the authors assume that given cumulatively infected individuals by time t, only a proportion $1 - v$ of $\log C_d(t)$ of them are responsible to produce the next infection. Each of these individuals carries a constant infection rate r.

The sub-exponential function $C_{\text{sub-exp}}(t) = (1 + rvt)^{\frac{1}{v}}$ is the solution of (4.61) given the initial condition $C_d(0) = i_0 = 1$. However, for $C_d(0) = i_0 > 1$, it gives the solution

$$C_d(t) = (i_0^v + rvt)^{\frac{1}{v}},\ 0 < v \le 1. \tag{4.62}$$

with

$$\rho(t) = \frac{d}{dt}\log C_d(t) = \frac{r}{i_0^v + rtv},$$

Unlike that described in (4.57), $\rho(t)$ depends on i_0 and $\rho(0) = r/i_0^v$.

Frailty Interpretation

Section 4.4 pointed out assumptions of independency and homogeneity in the exponential growth model $C_d'(t) = rC_d(t)$. If there are i_0 infected individuals initially seeded at $t = 0$, the rate of an instantaneous increase in the number of infections in the population to i_0+1 is $i_0 r$, arising as the first of the i_0 order statistics of the identical and independent exponential distribution with rate r. There is also an assumption about time-stationarity during the initial phase that the rate of an instantaneous increase in the number of infections in the population given $C(t) = n$

infected individuals at any time t is nr, arising from the first of the n order statistics of the identical and independent exponential distribution with rate r and survival function $\overline{G}(t) = e^{-rt}$. We have the identity $r = -\frac{d}{dt}\log \overline{G}(t) = \frac{d}{dt}\log C_d(t)$.

This can be generalized to the growth function in terms of $\frac{d}{dt}C_d(t) = \rho(t)C_d(t)$, so that given $C_d(t) = n$, the time to the next new infection is the first of the n order statistics of independently distributed Y with hazard function given by $\rho(t)$ and survival function $\overline{G}(t) = \Pr(Y > t)$. We propose the following identity

$$\rho(t) = -\frac{d}{dt}\log \overline{G}(t) = \frac{d}{dt}\log C_d(t) \tag{4.63}$$

which gives $C_d(t) = \overline{G}(t)^{-1}$.

The frailty model, introduced in Sect. 2.6, addresses heterogeneity and yields a class of monotonically decreasing functions for $\rho(t)$ that give rise to sub-exponential growth. There are many sources of heterogeneity:

1. **Intrinsic**: Individuals are not made of the same type and they carry different rates.
2. **Environmental**: The same index case may have different transmission rates depending on where and when it is seeded in a heterogeneous and dynamic complex social network.
3. **"Sampling biases"**: Even if infected individuals are made of the type and the environment in which the disease transmission is homogeneous, during the rapid ascending phase of the epidemic, the probability of being infectious at time t, $\Pr(\Delta_i(t) = 1)$, differs from individual to individual. When the system is not at equilibrium, W_t as illustrated in Fig. 4.8 are not identically distributed as W (with p.d.f. $f_W(x) = \overline{F}_I(x)/\mu_I$). Instead, given time t, each of the $C(t) = n$ individuals has its own distribution W_t depending on its time at infection. This is another source of heterogeneity among infected individuals.

When individual heterogeneity is concerned, we may assume that each infected individual i carries an intrinsic rate $z_i r$ to produce a new infection, where r is a baseline rate and z_i is a frailty variable. If the sources of heterogeneity can be determined and are observable, z_i can be modeled through covariates, such as the proportional hazard regression model. This requires extensive knowledge and understanding of the individuals, the environment, and the disease dynamics.

When the heterogeneity is not observable, a common approach is to consider a frailty model as introduced in Sect. 2.6 by assuming z_i as random variables, i.i.d. with mean value $E(z) = 1$ and probability density function (p.d.f.) $\xi(z)$. If there is no heterogeneity, then $\xi(z)$ degenerates to a point $z \equiv 1$ with no variation.

At the individual level, the survival function given $z_i = z$ is $\overline{G}(t|z) = e^{-rzt}$. When the population is composed of a mixture of heterogeneous individuals, the survival function arises from a mixed distribution:

$$\overline{G}^{(mixed)}(t) = \int_0^\infty e^{-zrt}\xi(z)dz = L[\xi](rt)$$

where $L[\xi](s) = \int_0^\infty e^{-zs}\xi(z)dz$ is the Laplace transform with respect to $\xi(z)$. The hazard function becomes

$$\rho^{(mixed)}(t) = -\frac{d}{dt}\log \overline{G}^{(mixed)}(t) = -\frac{d}{dt}\log L[\xi](rt).$$

It can be shown that, for any non-degenerated p.d.f. $\xi(z)$, $L[\xi](rt) = \int_0^\infty e^{-zrt}\xi(z)dz \geq e^{-rt}$.

As in the exponential growth model, we assume time-stationary over very short periods of time so that it is approximately true that $\frac{d}{dt}C_d(t) = \rho^{(mixed)}(t)C_d(t)$ for the initial phase of the outbreak, that is, $\rho^{(mixed)}(t) = \frac{d}{dt}\log C_d(t)$. This leads to the sub-exponential growth for any mixture distribution $\xi(z)$ given by

$$C_d(t) = [L[\xi](rt)]^{-1} \leq \frac{1}{e^{-rt}} = e^{rt} \text{ for all } t > 0. \tag{4.64}$$

1. $L[\xi](rt)$ is log-convex and $\rho^{(mixed)}(t) = -\frac{d}{dt}\log L[\xi](rt)$ is monotonically decreasing starting from $\rho^{(mixed)}(0) = r$. This has been proven in Marshall and Olkin (2007) and reiterated in Sect. 2.6.1.
2. Chapter 2 also concluded with statements that both $L[\xi](rt)$ and $\rho^{(mixed)}(t)$ are completely monotonic functions.

The Gamma Mixture Let $\xi(z)$ is the p.d.f. of the Gamma distribution with $E(z) = 1$ and variance $var[z] = v > 0$, then

$$\rho^{(mixed)}(t) = r/(1 + rvt), v > 0$$

has the form of the hazard function of the Pareto-I distribution, as given by (2.43) in Chap. 2 (replacing λ with r and $\kappa = v^{-1}$). A sub-exponential growth function is defined by

$$\frac{d}{dt}C_d(t) = \frac{r}{1 + rvt}C_d(t), \ 0 < v \leq 1 \tag{4.65}$$

with the solution (initial condition $C_d(0) = i_0$)

$$C_d(t) = i_0(1 + rvt)^{\frac{1}{v}}, \ 0 < v \leq 1. \tag{4.66}$$

The condition $v \leq 1$ is imposed so that (4.57) holds. The limiting case $\lim_{v\to 0} i_0(1+rvt)^{\frac{1}{v}} = i_0e^{rt}$ corresponds to the exponential growth. When $v = 1$, $C_d(t) = i_0(1 + rt)$ yields the linear growth.

The Inverse-Gaussian Mixture If $\xi(z)$ be the p.d.f. of an inverse-Gaussian distribution with $E(z) = 1$ and variance $var[z] = v$, the Laplace transform is $L[\xi](s) = e^{(1-\sqrt{1+2vs})/v}$. Letting $C_d(0) = i_0$, the expression (4.64) becomes

$$C_d(t) = [L[\xi](rt)]^{-1} = i_0 e^{(\sqrt{1+2vrt}-1)/v},\ 0 < v \leq 1 \tag{4.67}$$

It is also sub-exponential and the limiting case is $\lim_{v\to 0} i_0 e^{(\sqrt{1+2vrt}-1)/v} = i_0 e^{rt}$. It can be shown that $e^{(\sqrt{1+2vrt}-1)/v} \geq (1+vrt)^{1/v}$ for all $s > 0$. However, the difference between $C_d(t)$ given in (4.67) and $C_d(t)$ given in (4.66) is small:

$$e^{(\sqrt{1+2vrt}-1)/v} - (1+vrt)^{1/v} = \frac{1}{6}v^2 r^3 t^3 + O\left(t^4\right).$$

When $v = 1$, $e^{\sqrt{1+2rt}-1} = 1 + rt + \frac{1}{6}r^3t^3 + O\left(t^4\right)$, slightly above the linear growth. Meanwhile,

$$\rho^{(mixed)}(t) = \frac{r}{\sqrt{1+2rvt}},\ 0 < v \leq 1.$$

Discussion on the Sub-exponential Growth $C_d(t) = i_0(1+rvt)^{\frac{1}{v}}$, $0 < v \leq 1$

We have seen that the sub-exponential growth $C_d(t) = i_0(1+rvt)^{\frac{1}{v}}$, $0 < v \leq 1$ may arise as a simple function in a variety of settings. It may be used to approximate the initial growth of an observable process following an initially exponential growth of the number of new infections in the population, with an exponentially distributed delay; or to approximate the initial part of a logistic growth function that typically incorporates the depletion of the susceptible individuals in a finite population, or as a gamma mixture for the exponential growth in the frailty model that addresses individual and environmental heterogeneity. We do not rule out other mechanisms that may also produce sub-exponential phenomena that may be modeled with this simple form.

We also point out that the usefulness of a sub-exponential growth function is that it is able to describe a broad range of phenomena with only two parameters that leads to significant improvements in goodness of fit and short-term forecasting performance when models incorporate the possibility of such early sub-exponential growth, as pointed out by recent studies (e.g., Smirnova et al. 2017) as well as in example in Chap. 8 of this book.

4.6 Problems and Supplements

4.1 Let N_i be i.i.d. random count numbers with p.g.f. $G_N(s)$ and X be a random count number with p.g.f. $G_X(s)$. Show that the p.g.f. of the random sum $N_1 + N_2 + \cdots + N_X$ is $G_X(G_N(s))$.

4.2 We used p.g.f. $G_{X_g}(s) = G_{X_{g-1}}(G_N(s))$ in (4.34) as a tool to calculate $E[X_g] = R_0^g$ and $var[X_g]$ for $g = 1, 2, \ldots$. The goal of this exercise is to develop an alternative approach. Consider the Galton-Watson branching process defined by (4.1) with $X_0 = 1$, and N_i are i.i.d. and its distribution does not change over generations.

(a) Show that given X_{g-1}, $E[X_g|X_{g-1}] = R_0 X_{g-1}$, $g = 1, 2, \ldots$.
(b) Since $E[X_0] = 1$, using $E[X_g] = E[E[X_g|X_{g-1}]]$ recursively to calculate the unconditional expectation $E[X_g]$.
(c) Using the conditional variance formulae

$$var[X_g] = E[var[X_g|X_{g-1}]] + var[E[X_g|X_{g-1}]]$$

to show that, when $R_0 \neq 1$,the variance

$$var[X_g] = R_0^{g-1}\left(\frac{R_0^g - 1}{R_0 - 1}\right) var[N],\ g = 1, 2, \ldots$$

where $var[N] < \infty$.
(d) If $R_0 < 1$, what happens to both $E[X_g]$ and $var[X_g]$ as $g \to \infty$?
(e) Show the following inequality: $E[X_g] \geq \Pr(X_g \geq 1)$. Hence if $R_0 < 1$, $\Pr(X_g = 0) \to 1$, as $g \to \infty$.
(f) Define $\delta = \lim_{g\to\infty} \Pr(X_g = 0|X_0 = 1) = \Pr$(the branching process dies out) and show that

$$\delta = \sum_{j=0}^{\infty} \Pr(\text{the branching process dies out } |X_1 = j)\Pr\{N = j\}$$

(g) Further argue that, Pr(the branching process dies out $|X_1 = j) = \delta^j$, and hence $\delta = \sum_{j=0}^{\infty} \delta^j \Pr\{N = j\}$.
(h) Show that, when $R_0 > 1$, δ is the smallest root of $s = G_N(s)$ in $s \in (0, 1]$.
(i) Show that, if N follows the geometric distribution, $\delta = 1/R_0$.
(j) If we use the Galton-Watson branching process to approximate the initial phase of an outbreak, comment on how variabilities (according to convex order) in N and in the infectious periods T_I affect the probability of a small outbreak, which is δ.

4.3 Define $Z_g = \sum_{j=0}^{g} X_j$ to the total size of the population in the branching process by generation g. The p.d.f. is $G_{Z_g}(s) = E[s^{Z_g}]$.

(a) Since $Z_1 = 1 + X_1$, show that the p.g.f. for Z_1 is $G_{Z_1}(s) = sG_N(s)$, for $s \in [0, 1]$.

(b) Let $\{Z_{gk}, k = 1, 2, \ldots, X_1\}$ be i.i.d. copies of Z_g and $X_1 > 0$. Show that for $g \geq 1$

$$Z_{g+1} = 1 + \sum_{k=1}^{X_1} Z_{gk}.$$

(c) Show that $G_{Z_{g+1}}(s) = E[s^{Z_{g+1}}] = sG_N(G_{Z_g}(s))$, $g \geq 1$.

(d) Define the final size $Z = \sum_{g=0}^{\infty} X_g$ so that $Z_g \uparrow Z$, with p.g.f. given by $G_Z(s) = E[s^Z]$, show that

$$G_Z(s) = sG_N(G_Z(s)).$$

(e) Show that $G_Z(1) = \delta \leq 1$ where $\delta = G_N(\delta)$. Hence if $R_0 > 1$, $G_Z(1) = \sum_{z=1}^{\infty} \Pr\{Z = z\} = \delta < 1$.

4.4 Figure 4.1 illustrated four transmission trees with references in the caption. Select one or more of these transmission trees, write down the numbers X_g, $g = 0, 1, 2, \ldots$, and study the original papers in the references to build the transmission stories as complete as possible for each generation. Calculate the Harris estimate (4.30) for the reproduction number for the first (some chosen) G generations as well as the estimated reproduction variance $\widehat{var}[N]$ using (4.32). Discuss how these estimates might be interpreted, in a similar fashion as the analysis in Sect. 4.2.5, summarized in Table 4.1.

4.5 Let $\overline{F}(x)$ and $f(x)$ be the survival function and the p.d.f. for the lifetime X; $L[\overline{F}](s) = \int_0^{\infty} e^{-sx}\overline{F}(x)dx$ and $L[f](s) = \int_0^{\infty} e^{-sx} f(x)dx$ are the corresponding Laplace transform functions, show that $L[\overline{F}](s) = \frac{1}{s}(1 - L[f](s))$.

4.6 Consider a disease with random infectious periods, corresponding to the p.d.f. $f_I(x)$. Assuming that the outbreak starts with a single initially seeded infectious individual under homogeneous mixing, that is, the infectious contact process $\{K(x) : x \geq 0\}$ is a homogeneous Poisson process with infectious contact rate β,

(a) show that the probability of extinction δ during the initial phase satisfies the equation $L[f_I](\beta(1 - \delta)) = \delta$;

(b) if the infectious period starts immediately upon infection, show that $r = \beta(1 - \delta)$, where r is the intrinsic growth rate satisfying $\beta \int_0^{\infty} e^{-rx}\overline{F}_I(x)dx = 1$;

(c) if the infectious period starts after a random latent period since infection, show that $r < \beta(1 - \delta)$.

(d) In the literature, we frequently see the linear relationship $R_0 = 1 + r\mu_I$ where μ_I is the mean infectious period. Give at least two examples to show that a smaller value of r (i.e., slower initial growth) may correspond to a larger value of R_0, with at least one example assuming that the infectious periods start immediately upon infection and one example assuming that a latent period exists.

Chapter 5
Beyond the Initial Phase: Compartment Models for Disease Transmission

We start with simple models that describe the dynamics of disease transmission over time t in a constant population of size m and investigate the long-term epidemic dynamics as $t \to \infty$. In these simple models, we assume there is no replacement of susceptible individuals due to demographic input of susceptible newborns. The population is partitioned into compartments, with at least one compartment representing the prevalence of individuals who are susceptible to infection and at least one compartment representing the prevalence of individuals who are infectious (at time t).

5.1 The Agent–Host–Environment Relationship and Some Homogeneity Assumptions

We first discuss a key assumption used throughout this chapter with respect to the agent–host–environment relationship for the disease transmission dynamics.

With respect to the *infectious agent*, such as the virus in viral infections, it is assumed that it is not subject to mutations that lead to increased or decreased infectiousness during the study period, so that individuals who have acquired infections at different times possess the same infectiousness as at the time of infection.

With respect to *hosts*, it is assumed that

1. all susceptible individuals are equally susceptible;
2. a typical infected individual remains equally infectious throughout its infectious period;
3. all infectious individuals are equally infectious.

P. Yan, G. Chowell, *Quantitative Methods for Investigating Infectious Disease Outbreaks*, Texts in Applied Mathematics 70,
https://doi.org/10.1007/978-3-030-21923-9_5

With respect to the *environment* in terms of the social contact network, we assume *homogeneous mixing* in the sense that an individual makes contacts with all other individuals in the population with equal probability. If the population size is large ($n \to \infty$), the social network can be approximated by a Bernoulli random graph (Erdös and Rényi 1961). The numbers of vertices adjacent to vertex v_i (the degree of vertex v_i) are identically and independently distributed according to a Poisson distribution. From the perspective of stochastic processes, the social contact network grows in such a way that the number of contacts made by a typical individual in this network follows a stationary Poisson process.

With these assumptions, at any snapshot in time during the epidemic, all infectious individuals are equally infectious regardless of when each individual is infected and how long it has been infected. Each contact between a pair of susceptible-infectious individuals is associated with the same probability that a transmission may occur.

- Denoting the number of individuals who are infectious at time t by $I(t)$, the force of infection onto a typical susceptible individual at any given time t is proportional to the proportion of infectious individuals $I(t)/m$ by a factor β_1.
- Denoting the number of individuals who are susceptible at time t by $S(t)$, the instantaneous transmission rate of a typical infectious individual at any given time t is proportional to the proportion of susceptible individuals $S(t)/m$ by a factor β_2.
- The homogeneity assumptions in the agent–host–environment relationship lead to $\beta_1 = \beta_2 = \beta$, which defines the probability of an instantaneous new infection in the population as $\beta \frac{S(t)I(t)}{m} dt$.

5.2 Susceptible-Infectious-Susceptible Models

5.2.1 *The Birth–Death Markov Process as a Model for the Simple Epidemic and the SIS Epidemic*

In this model, the population is partitioned into two classes of individuals so that at any time t, an individual is either susceptible or infected (and infectious), a birth–death process is used as a model for $\{I(t)\}_0^\infty$. The transition probabilities are modeled by

$$\begin{aligned} \Pr\{I(t+dt) = i+1 | I(t) = i\} &= \beta\left(1 - \tfrac{i}{m}\right) i dt, \\ \Pr\{I(t+dt) = i-1 | I(t) = i\} &= \gamma i dt. \end{aligned} \tag{5.1}$$

If $\gamma > 0$, such a model is a stochastic susceptible-infectious-susceptible (SIS) model. This process has finite state space $\mathfrak{S} = \{0, 1, 2, \ldots, m\}$ with the state $\{0\}$ being an absorbing state and $\{1, 2, \ldots, m\}$ being transient states.

The instantaneous rate $q_{i,i+1} = \beta\left(1 - \frac{i}{m}\right) i$ is a logistic birth rate based on the assumption of homogeneous transmission in the environment that gives $\beta \frac{S(t)I(t)}{m} dt$, where $S(t) = m - I(t)$. The instantaneous rate $q_{i,i-1} = \gamma i$ is linear following an additional assumption that an infected individual spends an exponentially distributed infectious period with mean γ^{-1} and immediately recovers with no conferred immunity.

A value of $\gamma = 0$ implies that the mean infectious period is infinite. The process defined by (5.1) becomes a pure birth process. In the literature, this process is called the *simple epidemic* (Bailey 1975) or an SI model.

The Duration of a Simple Epidemic

A simple epidemic ends when all individuals are infected. The states $\{0, 1, 2, \ldots, m-1\}$ are transient states and the state $\{m\}$ is an absorbing state. The most relevant public health question is the duration of the epidemic, denoted by T_{m,i_0}, as the time to reaching the absorbing state $\{m\}$ given the initial condition $I(0) = i_0$.

According to the model, the sojourn time in state $i \leq m-1$ before moving to state $i+1$ is exponentially distributed with mean value $q_{i,i+1}^{-1}$.The sojourn times in different states are independent. Therefore, given the initial condition $I(0) = i_0$, the distribution of T_{m,i_0} arises as a convolution of $m - i_0$ independently distributed random variables. Without losing generality, we assume $i_0 = 1$.

The mean value of $T_{m,1}$ is the sum

$$\begin{aligned} E[T_{m,1}] &= \sum_{i=1}^{m-1} q_{i,i+1}^{-1} = \frac{1}{\beta} \sum_{i=1}^{m-1} \frac{m}{(m-i)\, i} \\ &= \frac{1}{\beta} \sum_{i=1}^{m-1} \left(\frac{1}{i} + \frac{1}{n-i} \right) = \frac{2}{\beta} \sum_{i=1}^{m-1} \frac{1}{i}. \end{aligned}$$

Because $\sum_{i=1}^{k} \frac{1}{i}$ =Euler constant $+ \log k + \epsilon_k$ where Euler constant ≈ 0.5772 and $\epsilon_k \to 0$ as $k \to \infty$, we have

$$E[T_{m,1}] \approx \frac{2}{\beta} \left[0.5772 + \log(m-1) \right] \tag{5.2}$$

when m is sufficiently large. The expected duration of the epidemic grows with the population size, approximately proportional to $\log(m)$.

The variance for $T_{m,1}$ is the sum

$$Var[T_{m,1}] = \sum_{i=1}^{m-1} q_{i,i+1}^{-2} = \frac{1}{\beta^2} \sum_{i=1}^{m-1} \frac{m^2}{(m-i)^2\, i^2}.$$

Using a partial fraction decomposition $\frac{m^2}{(m-i)^2 i^2} = \frac{2}{mi} + \frac{1}{i^2} + \frac{2}{m(m-i)} + \frac{1}{(m-i)^2}$ (Allen 2010) and $\sum_{i=1}^{\infty} \frac{1}{i^2} = \frac{\pi^2}{6}$, one gets the approximation

$$Var[T_{m,1}] \approx \frac{1}{\beta^2} \left\{ \frac{4}{m} \left[0.5772 + \log(m-1) \right] + \frac{\pi^2}{3} \right\} \approx \frac{\pi^2}{3\beta^2} \tag{5.3}$$

when m is sufficiently large. Unlike (5.2), as m becomes large, the variance of the duration converges to a constant.

For higher moments of $T_{m,1}$, the sojourn time in state $i \leq m-1$ before moving to state $i+1$ is exponentially distributed with the moment generating function (m.g.f.) $\mathcal{M}(s) = \left(1 - \frac{ms}{\beta(m-i)i}\right)^{-1}$. The m.g.f. for $\mathcal{M}_{T_{m,1}}(s)$ can be easily derived as the multiplication

$$\mathcal{M}_{T_{m,1}}(s) = \prod_{i=1}^{m-1} \left(1 - \frac{ms}{\beta\,(m-i)\,i}\right)^{-1}.$$

The Duration of the SIS Model

If $\gamma > 0$, the model is defined by (5.1) and is an SIS model. The states $\{1, 2, \ldots, m\}$ are transient and $\{0\}$ is an absorbing state. The epidemic ends when there are no prevalent infectious individuals in the population.

Based on theories of general birth–death process satisfying $q_{0,1} = 0$ and $q_{i,i+1} > 0$ for $i = 1, \ldots, m-1$; $q_{i,i-1} > 0$ for $i = 1, \ldots, m$, it can be shown that the expected time until extinction (Karlin and Taylor 1975; Ross 1996; Allen 2010; among others), given the current state i, is:

$$\begin{aligned} E[T_{0,1}] &= \tfrac{1}{q_{1,0}} + \textstyle\sum_{j=2}^{m} \frac{q_{1,2}q_{2,3}\cdots q_{j-1,j}}{q_{1,0}q_{2,1}\cdots q_{j,j-1}}, && \text{if } i = 1, \\ E[T_{0,i}] &= E[T_{0,1}] + \textstyle\sum_{s=1}^{i-1} \left[\frac{q_{1,0}q_{2,1}\cdots q_{s,s-1}}{q_{1,2}q_{2,3}\cdots q_{s,s+1}} \sum_{j=s+1}^{m} \frac{q_{1,2}q_{2,3}\cdots q_{j-1,j}}{q_{1,0}q_{2,1}\cdots q_{j,j-1}} \right], && \text{if } i = 2, 3, \ldots, m. \end{aligned}$$

For the SIS model with $q_{i,i+1} = \beta\left(1 - \frac{i}{n}\right) i$, $q_{i,i-1} = \gamma i$, $E[T_{0,1}]$ becomes

$$\gamma\, E[T_{0,1}] = 1 + \sum_{j=2}^{m} \frac{(\beta/\gamma)^{j-1}}{j} \prod_{l=1}^{j-1} (1 - \frac{l}{m}) \tag{5.4}$$

and

$$\gamma\, E[T_{0,i}] = \gamma\, E[T_{0,1}] + \sum_{s=1}^{i-1} \left[\frac{\sum_{j=s+1}^{n} \frac{(\beta/\gamma)^{j-1}}{j} \prod_{l=1}^{j-1} (1 - \frac{l}{n})}{(\beta/\gamma)^{s} \prod_{l=1}^{s} (1 - \frac{l}{n})} \right]. \tag{5.5}$$

By re-scaling time by recovery rate γ, the expected time until extinction $\gamma E[T_{0,1}]$ and $\gamma E[T_{0,i}]$ only depend on the ratio β/γ.

Recall that in the SIS model, the infectious period is exponentially distributed with mean value $\mu_I = \gamma^{-1}$. The ratio β/γ is the basic reproduction number $R_0 = \beta\mu_I$ (4.2) defined at $t = 0$ when the system is at disease-free equilibrium.

If $R_0 < 1$, given the current state $I(t) = 1$, the expected time to the end of the epidemic $\gamma E[T_{0,1}] < \frac{1}{R_0} \log \frac{1}{1-R_0}$. In fact, $R_0^{-1} \log \frac{1}{1-R_0}$ is the limiting case as $m \to \infty$ so that

$$\gamma E[T_{0,1}] \to 1 + \sum_{j=2}^{\infty} \frac{1}{j} \left(\frac{\beta}{\gamma}\right)^{j-1} = R_0^{-1} \log \frac{1}{1 - R_0} < \infty. \tag{5.6}$$

Numerically computed $\gamma E[T_{0,1}]$ values given by (5.4) are presented in Table 5.1. For $R_0 < 1$, the epidemic ends very quickly. This is the feature of a small outbreak as discussed in Chap. 4. It also echoes (4.10) which gives the cumulative probability to extinction by generation g.

If $R_0 = 1$, given the current state $I(t) = 1$,

$$\gamma E[T_{0,1}] = 1 + \sum_{j=2}^{m} \frac{1}{j} \prod_{l=1}^{j-1} (1 - \frac{l}{m}) \approx \frac{1}{2} \log m + o(\log m). \tag{5.7}$$

Although $\gamma E[T_{0,1}] \to \infty$ as $m \to \infty$, $\gamma E[T_{0,1}]$ takes very small values even in very large finite populations. For example, even when $m = 10^5$, $\gamma E[T_{0,1}] = 6.393$.

If $R_0 > 1$, $\gamma E[T_{0,1}] \to \infty$ as $m \to \infty$. By directly calculating from (5.4) for given m, there exists an ε_m such that when $R_0 > 1 + \varepsilon_m$, $\gamma E[T_{0,1}] \to \infty$ very quickly. However, in the range $1 < R_0 \leq \varepsilon_m$, as shown in Table 5.1, $\gamma E[T_{0,1}]$ remains very small. For example, when $R_0 = 1.02$, $\gamma E[T_{0,1}] = 14.95$ at $m = 10^4$.

Table 5.1 $\gamma E[T_{0,1}]$ as calculated using (5.4) for finite m

		m				
		10^2	10^3	10^4	$\approx$	$\to \infty$
R_0	2.3	8.3×10^{10}	1.1×10^{115}	7.2×10^{1160}		∞
	1.4	114.8	2.2×10^{21}	1.7×10^{219}		∞
	1.135	5.372	1.4×10^{3}	4.7×10^{32}		∞
	1.05	3.545	8.550	6×10^{4}		∞
	1.02	3.179	5.043	14.95		∞
	1.0	2.979	4.102	5.245	$\frac{1}{2} \log m + o(\log m)$	∞
	0.98	2.807	3.540	3.897	$R_0^{-1} \log \frac{1}{1-R_0}$	3.992
	0.95	2.591	3.019	3.135		3.153
	0.85	2.098	2.214	2.230		2.232
	0.5	1.377	1.385	1.386		1.386
	0.2	1.114	1.116	1.116		1.116

Andersson and Djehiche (1998, Theorem 1) provided asymptotic properties of the distribution for the time to extinction of the SIS model as $m \to \infty$, considering an initial condition in which $I(0)$ goes to infinity with m. The time to extinction, denoted by T_0, is asymptotically exponentially distributed with a mean that grows exponentially with m if $R_0 > 1$, given by

$$E[T_0] \sim \sqrt{\frac{2\pi}{m}} \frac{R_0}{(R_0-1)^2} e^{m(\log R_0 - 1 + 1/R_0)}. \tag{5.8}$$

There is rich literature in recent years on the distribution for time to extinction of the SIS model, and the above results have been generalized to homogeneous mixing with arbitrarily distributed infectious periods as well as in heterogenous mixing with household structures (Hernádez-Suárez and Castillo-Chavez 1999; Ball et al. 2016). Specifically, (5.6)–(5.8) correspond to the three results in Lemma 3.2 of Ball et al. (2016). For general results of time to extinction for SIS infections in heterogeneous populations, we refer to Clancy (2018).

The Sample Paths of $\{I(t)\}_0^\infty$

The sample paths of $\{I(t)\}_0^\infty$ in (5.1) can be studied through the Kolmogorov differential equations. Given $I(0) = i_0$, let $P_i(t) = \Pr\{I(t) = i | I(0) = i_0\}$, the forward Kolmogorov differential equations are

$$\begin{aligned} P_i'(t) = \beta\left(1 - \tfrac{i-1}{m}\right)(i-1)P_{i-1}(t) + \gamma(i+1)P_{i+1}(t) \\ - \left\{\beta\left(1 - \tfrac{i}{m}\right) + \gamma\right\} i P_i(t) \end{aligned} \tag{5.9}$$

for $i = 1, \ldots m-1$, and $P_0'(t) = \gamma P_1(t)$, $P_m'(t) = \beta\left(1 - \frac{m-1}{n}\right)(m-1)P_{m-1}(t) - \gamma m P_m(t)$. In (5.9), the first two terms with positive signs represent the in-flows to state i.

$$\begin{aligned} i-1 \to i &: q_{i-1,i} = \beta\left(1 - \tfrac{i-1}{m}\right)(i-1) \\ i+1 \to i &: q_{i+1,i} = \gamma(i+1). \end{aligned}$$

The third term, with negative sign, represents the outflows from state i to state $i+1$ and to state $i-1$, with instantaneous rates $\beta\left(1 - \frac{i}{m}\right)$ and γi, respectively.

Multiplying both sides of (5.9) by e^{ui} and summing over $i \in \mathfrak{S}$, one gets

$$\frac{\partial \mathcal{M}(u,t)}{\partial t} = \left(\beta\left(e^u - 1\right) + \gamma\left(e^{-u} - 1\right)\right)\frac{\partial \mathcal{M}(u,t)}{\partial u} - \frac{\beta\left(e^u - 1\right)}{m}\frac{\partial^2 \mathcal{M}(u,t)}{\partial u^2} \tag{5.10}$$

where $\mathcal{M}(u,t) = E[e^{uI(t)}] = \sum_{i\in\mathfrak{S}} e^{ui} P_i(t)$ is the m.g.f. for $I(t)$ at time t.

Proof The left-hand side of (5.10) is $\frac{\partial \mathcal{M}(u,t)}{\partial t} = \sum_{i\in\mathfrak{S}} e^{ui} P_i'(t)$. The right-hand side of (5.10) is

$$\begin{aligned}
&\beta e^u \sum_{i=1}^{n} \left(1 - \tfrac{i-1}{n}\right)(i-1) e^{u(i-1)} P_{i-1}(t) + \gamma e^{-u} \sum_{i=0}^{n-1} (i+1) e^{u(i+1)} P_{i+1}(t) \\
&\qquad - \beta \sum_{i=0}^{n} \left(1 - \tfrac{i}{n}\right) i e^{ui} P_i(t) - \gamma \sum_{i=0}^{n} i e^{ui} P_i(t) \\
&= \beta e^u \sum_{i\in\mathfrak{S}} \left(1 - \tfrac{i}{n}\right) i e^{ui} P_i(t) + \gamma e^{-u} \sum_{i\in\mathfrak{S}} i e^{ui} P_i(t) \\
&\qquad - \beta \sum_{i\in\mathfrak{S}} \left(1 - \tfrac{i}{n}\right) i e^{ui} P_i(t) - \gamma \sum_{i\in\mathfrak{S}} i e^{ui} P_i(t) \\
&= \beta \left(e^u - 1\right) \left(\sum_{i\in\mathfrak{S}} i e^{ui} P_i(t) - \frac{1}{n} \sum_{i\in\mathfrak{S}} i^2 e^{ui} P_i(t) \right) + \gamma \left(e^{-u} - 1\right) \sum_{i\in\mathfrak{S}} i e^{ui} P_i(t),
\end{aligned}$$

of which $\sum_{i\in\mathfrak{S}} i e^{iu} P_i(t) = \frac{\partial \mathcal{M}(u,t)}{\partial u}$ and $\sum_{i\in\mathfrak{S}} i^2 e^{iu} P_i(t) = \frac{\partial^2 \mathcal{M}(u,t)}{\partial u^2}$. ■

Given the initial condition $I(0) = i_0$, the conditional r-th moment for $I(t)$ is

$$E[I^r(t)|I(0) = i_0] = \left.\frac{\partial \mathcal{M}(u,t)}{\partial u}\right|_{u=0} = \sum_{i\in\mathfrak{S}} i^r P_i(t).$$

For simplicity of notation, we write $E[I^r(t)|I(0) = i_0] \triangleq E[I^r(t)]$. Differentiating (5.10) repeatedly with respect to u and setting $u = 0$, we get a system of ordinary differential equations showing that the rate of change of the r-th moment depends on the $(r+1)$-th moment, and an infinite set of coupled differential equations is generated.

$$\begin{aligned}
&\frac{d}{dt} E[I(t)] = (\beta - \gamma) E[I(t)] - \frac{\beta}{m} E[I^2(t)] \qquad (5.11) \\
&\frac{d}{dt} E[I^2(t)] = (\beta + \gamma) E[I(t)] + \left(2(\beta - \gamma) - \frac{\beta}{m}\right) E[I^2(t)] - \frac{2\beta}{m} E[I^3(t)] \\
&\frac{d}{dt} E[I^3(t)] = (\beta - \gamma) E[I(t)] + 3(\beta + \gamma) E[I^2(t)] \\
&\qquad + \left(3(\beta - \gamma) - \frac{\beta}{m}\right) E[I^3(t)] - \frac{3\beta}{m} E[I^4(t)] \\
&\qquad \vdots
\end{aligned}$$

As commonly seen in all nonlinear stochastic processes, moment closure is required. To close the dynamic system, the next moment in the hierarchy must, at some stage, be replaced with an expression containing only lower-order moments. There is rich literature regarding moment closure methods applied to the SIS models. We recommend readers to consult Nåsell (2003), Krishnarajah et al. (2005), Pinto et al. (2009), Clancy and Mendy (2011), among many others.

5.2.2 *The Deterministic SIS Model Represented by an Ordinary Differential Equation*

The first equation of (5.11) can be re-written as

$$\frac{d}{dt}E[I(t)] = \beta\left(1 - \frac{E[I(t)]}{m}\right)E[I(t)] - \gamma E[I(t)] - \frac{\beta var[I(t)]}{m}. \tag{5.12}$$

One of the moment closure procedures is to put assumptions on $var[I(t)]$ to study the dynamics of the expected number of $I(t)$ over time. The term *deterministic* means $var[I(t)] = 0$ for all $t > 0$ and $\{I(t)\}_0^\infty$ is modeled as a non-random, continuous, and differentiable function of t, denoted by $I_d(t)$. The deterministic SIS model is an extreme case of moment closure.

In the deterministic SIS model, we replace $E[I(t)]$ in (5.12) with $I_d(t)$ and get the ordinary differential equation

$$\frac{d}{dt}I_d(t) = \beta\left(1 - \frac{I_d(t)}{m}\right)I_d(t) - \gamma I_d(t). \tag{5.13}$$

Viewing (5.13) as deterministic through moment closure, other stochastic assumptions in the original model (5.1) are carried over. The constant recovery rate $\gamma > 0$ implies that the duration of the infectious period is exponentially distributed with mean value γ^{-1}. The first term of (5.13), $\beta\left(1 - \frac{I_d(t)}{m}\right)I_d(t)$, arises from the homogeneity assumptions in the agent–host–environment relationship, of which it is assumed that an individual makes contacts with all other individuals in the population with equal probability. Therefore, there is an embedded stochastic process governing the social contact network.

The differential equation (5.13) yields the explicit expression for $I_d(t)$ of a logistic function form

$$I_d(t) = \frac{mi_0\,(R_0 - 1)}{i_0 R_0 + (m\,(R_0 - 1) - i_0 R_0)\,e^{-(R_0-1)\gamma t}}. \tag{5.14}$$

It can be re-written as where $R_0 = \beta/\gamma$. It has three parameters (R_0, γ, m) and the initial condition $I_d(0) = i_0$. R_0 is the threshold parameter.

- If $R_0 \leq 1$, for any $i_0 > 0$, $I_d(t)$ decreases monotonically with $\lim_{t\to\infty} I_d(t) = 0$;
- if $R_0 > 1$ and $i_0 > m\,(1 - 1/R_0)$, then $I_d(t)$ decreases monotonically with $\lim_{t\to\infty} I_d(t) = m\,(1 - 1/R_0)$;
- if $R_0 > 1$ and $i_0 < m\,(1 - 1/R_0)$, then $I_d(t)$ increases monotonically with $\lim_{t\to\infty} I_d(t) = m\,(1 - 1/R_0)$;
- if $R_0 > 1$ and $i_0 = m\,(1 - 1/R_0)$, then $I_d(t)$ is constant: $I_d(t) = m\,(1 - 1/R_0)$.

Later in this chapter we shall see that (5.14) can be re-parameterized as the logistic function (4.59) in Chap. 4 with one less parameter.

According to Nåsell (2002), a parameter is called *innocent* if it can be eliminated from the model by re-scaling either the state variable or the time. Otherwise, the parameter is called *essential*. In the deterministic SIS model, the essential parameter is R_0 whereas both γ and m are innocent. This is because $I_d(t)$ depends on t only through γt, γ is a scale parameter and can be eliminated by re-scaling of time $\tau = \gamma t$. The population size m can also be eliminated from the model by re-scaling $y(t) = \frac{I_d(t)}{m}$. The model (5.13) becomes

$$\begin{array}{ll} \frac{d}{dt}y(\tau) = -y(\tau)^2 & \text{if } R_0 = 1 \\ \frac{d}{d\tau}y(\tau) = (R_0 - 1)\left[1 - \frac{R_0}{R_0 - 1}y(\tau)\right]y(\tau) & \text{if } R_0 \neq 1 \end{array} \tag{5.15}$$

which describes the change of the proportion $y(\tau)$ of infectious individuals in a population at time τ (according to a standardized scale $\gamma = 1$) regardless of the population size m.

The standardized form of $I_d(t)$ is

$$y(\tau) = \frac{y_0(1 - R_0^{-1})}{y_0 + \left(\left(1 - R_0^{-1}\right) - y_0\right)e^{-(R_0-1)\tau}}. \tag{5.16}$$

With initial value $y_0 = i_0/m > 0$, if $R_0 \leq 1$, $y(\tau)$ decreases monotonically with $\lim_{\tau\to\infty} y(\tau) = 0$; if $R_0 > 1$, $y(\tau)$ approaches $\lim_{\tau\to\infty} y(\tau) = 1 - R_0^{-1}$, either monotonically decreasing if $y_0 > 1 - R_0^{-1}$ or monotonically increasing if $y_0 < 1 - R_0^{-1}$. In the latter case, $y(\tau)$ increases as a convex function of τ if $\tau < -\frac{1}{(R_0-1)} \log \frac{R_0 y_0}{(R_0-1)-R_0 y_0}$ and as a concave function of τ afterwards.

5.2.3 *Comparing the Stochastic and the Deterministic SIS Models*

We compare the dynamics of $E[I(t)]$ given by (5.12) with its deterministic counterpart $I_d(t)$ given by (5.13).

First, given the same initial condition $I(0) = I_d(0) = i_0$, the mean of the stochastic process as the solution for $E[I(t)]$ in (5.12) is less than the solution for

$I_d(t)$ in (5.13), that is, $E[I(t)] \leq I_d(t)$, for $t \in [0, \infty)$. This is a special case of a general theorem (page 296 of Allen 2010). The difference is due to the variance term in (5.12), in which the dynamic change of $var[I(t)]$ depends both on the first moment $E[I(t)]$ and the third moment $E[I^3(t)]$

$$\begin{aligned}\frac{d}{dt}var[I(t)] = &\left(2(\beta-\gamma) - \frac{\beta}{m} + \frac{2\beta}{m}E[I(t)]\right) var[I(t)] \\ &+ (\beta+\gamma)E[I(t)] - \frac{\beta}{m}E[I(t)]^2 + \frac{2\beta}{m}E[I(t)]^3 \\ &- \frac{2\beta}{m}E[I^3(t)].\end{aligned}$$

On the other hand, the deterministic model assumes that $var[I_d(t)] = 0$ for all $t > 0$.

Second, in the stochastic SIS model, $\{0\}$ is the only absorbing state. When $t \to \infty$, the epidemic will end with certainty. However, the expected time to the end, $\gamma E[T_{0,1}]$, approaches infinity when $R_0 > 1+\varepsilon$, for some ε. In the deterministic SIS model, the epidemic arrives at endemic equilibrium $\lim_{t\to\infty} I_d(t) = m(1 - 1/R_0)$.

Third, the stochastic SIS model is qualitatively different from its deterministic counterpart. In the stochastic SIS model, by re-scaling time $\tau = \gamma t$, the transition probabilities can be written as

$$\begin{aligned}&\Pr\{I(\tau + d\tau) = i+1 | I(\tau) = i\} = \tfrac{\beta}{\gamma}\left(1 - \tfrac{i}{m}\right) i d\tau, \\ &\Pr\{I(\tau + d\tau) = i-1 | I(\tau) = i\} = i d\tau.\end{aligned}$$

Therefore, γ is an innocent parameter for both the deterministic and the stochastic SIS models. The population size m plays an important role in the stochastic model. It defines the state space $\mathfrak{S}$. Therefore it is essential in the stochastic model but innocent in the deterministic model. The parameter $R_0 = \beta/\gamma$ is essential for both models.

At first glance, it seems that the deterministic model gives the mean behavior of the corresponding stochastic system asymptotically with $m \to \infty$, in which case, $\frac{\beta var[I(t)]}{m} \to 0$. Thus for large populations, there is little to be gained from using a stochastic model, which will generally be more difficult to analyze. However, as pointed out in Isham (2005), even in large populations, chance fluctuations do not always average out to have little overall effect. Even when they do, it may be important to take the variability of individual realizations into account, for example in predicting the course of an individual outbreak.

More importantly, the deterministic model as an approximation of the mean behavior of $E[I(t)]$ is in the sense of viewing the occurrence of an epidemic outbreak as a realization of a random event, assuming it can be repeated under identical initial conditions. Each realization has its own sample path. The large population size, approximated by $m \to \infty$, only removes the oscillation within a single sample path, that is, a smooth curve. Since the occurrence of an epidemic outbreak does not arise from a designed experiment and is a single realization of

a stochastic event which cannot be repeated under identical initial conditions, this aspect is better illustrated through stochastic simulation.

5.2.4 Stochastic Simulation of SIS Outbreaks

Simulation provides a virtual experiment assuming that the outbreak can be repeated under identical initial conditions (Allen 2017). The simulation algorithm can be done in two parts:

1. We first simulate the sample paths of the embedded discrete time Markov chain $\{i_0, i_1, i_2, \ldots i_L\}$ in L steps. We start with an initial state $I(0) = i_0$, $1 \leq i_0 \leq m-1$. Then we generate a uniformly distributed random number U between 0 and 1 and assign

$$i_1 = \begin{cases} i_0 + 1, \text{ if } U \leq \frac{\beta(m-i_0)}{\beta(m-i_0)+m\gamma} \\ i_0 - 1, \text{ if } U > \frac{\beta(m-i_0)}{\beta(m-i_0)+m\gamma} \end{cases}. \tag{5.17}$$

   If $i_1 = 0$, then the process reaches the absorbing state and we assign $i_2 = i_3 = \cdots = 0$. Otherwise if $1 \leq i_1 \leq m-1$, we calculate i_2 using (5.17) by replacing i_0 with i_1 and replacing i_1 with i_2 in (5.17). If $i_1 = m$, we assign $i_2 = m-1$. We repeat this algorithm by steps (up to step L) and get discrete sample paths of the embedded Markov chain $\{i_0, i_1, i_2, \ldots i_L\}$.
2. We then simulate the sojourn time in each state in order to determine the time at which each jump occurs. We first generate q random numbers from the standard exponential distribution with p.d.f. $X_0 \sim e^{-x}$. For $j = 0, \ldots, L-1$, the sojourn time of state i_j is distributed according to an exponential distribution with rate $\beta\left(1 - \frac{i_j}{m}\right) i_j + \gamma i_j$, which can be simulated by $X_0 \Big/ \left[\beta\left(1 - \frac{i_j}{n}\right) i_j + \gamma i_j\right]$.

Such a simulation algorithm can be easily implemented by many commonly available mathematical or statistical computing languages, such as MATLAB or R by R Development Core Team (http://www.R-project.org).

Stochastic simulation shows that, when t is sufficiently large, both the mean value $E[I(t)]$ and the variance $var[I(t)]$ appear to be no longer depending on time t. We assume that, at this phase, the marginal distribution $\Pr\{I(t) = i\}$ is stationary. We call this distribution the *quasi-equilibrium distribution* and denote it by $q_i = \Pr\{I(t) = i | I(t) > 0\}$, $i = 1, \ldots, m$ with $\sum_{i=1}^{m} q_i = 1$.

Figure 5.1 demonstrates two of such simulations with population sizes $m = 100$ and $m = 1000$. In both simulations, we choose $\beta = 1.5$, $\gamma = 1$, and the initial value $i_0 = 2$. We repeat each simulation 500 times to generate 500 sample paths. Some typical sample paths are highlighted in color. The deterministic functions $I_d(t)$, given by (5.14), are also plotted against the simulated sample paths in both simulations. Alongside the simulated sample paths in the two populations, normalized histograms for summary statistics of $\{q_i : i = 1, \ldots, m\}$ are also plotted.

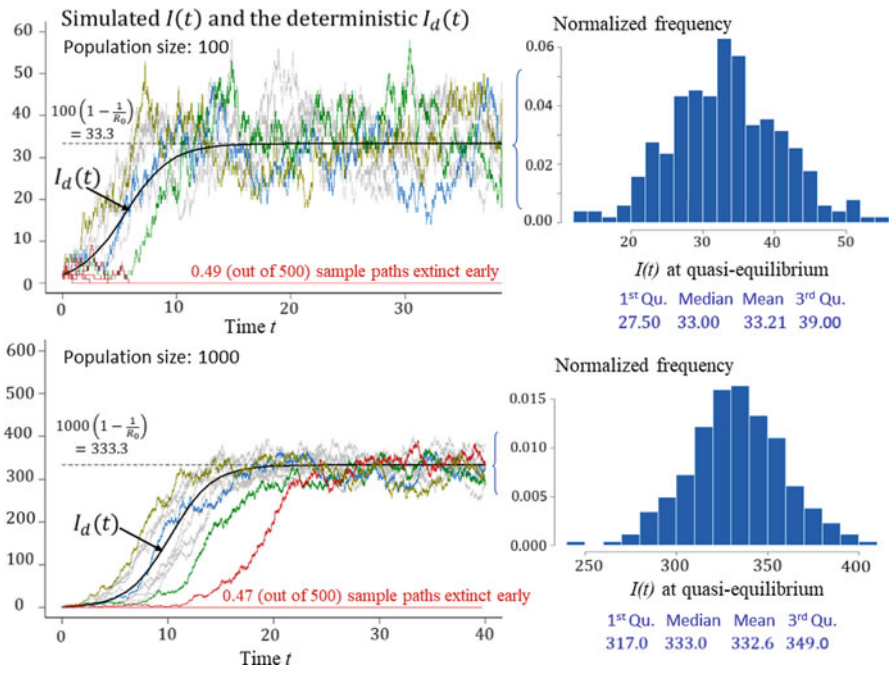


Fig. 5.1 Deterministic functions, $I_d(t)$, given by (5.14) are plotted (solid black lines) against the simulated sample paths in two population sizes ($m = 100$ and 1,000) at $\beta = 1.5$, $\gamma = 1$, and $i_0 = 2$. The histograms are distributions for $I(t)$ at equilibrium and summary statistics of $\{q_i : i = 1, \ldots, n\}$

We summarize the following observations from these simulations.

1. In the stochastic simulations, some realizations of the sample paths become extinct at the very beginning. About 49% of the simulated sample paths had early extinction when $m = 100$, and about 47% of the simulated sample paths had early extinction when $m = 1000$. Comparing with the branching process model in Chap. 4, if the infectious period is exponentially distributed, the probability of such early extinction is $\delta = \min\left\{1, R_0^{-i_0}\right\}$ as $m \to \infty$. In the case $R_0 = 1.5$ and $i_0 = 2$, $\delta = 44.4\%$.
2. The deterministic model suggests $\lim_{t\to\infty} I_d(t) = m\left(1 - R_0^{-1}\right) > 0$ if $R_0 > 1$, which is an endemic equilibrium value. This value can be used to compare with the mean value of the quasi-equilibrium distribution. Figure 5.1 shows that the empirical mean of the quasi-equilibrium distribution $\{q_i : i = 1, \ldots, m\}$ at $m = 100$ is 33.21 and at $m = 1000$ is 332.6. The deterministic endemic equilibrium values are 33.33 at $m = 100$ and 333.33 at $m = 100$, respectively.

3. We have shown theoretically that the stochastic mean $E[I(t)]$ is not necessarily equal to $I_d(t)$ because $var[I(t)] > 0$ in (5.12). In fact, $E[I(t)] \leq I_d(t)$, for $t \in [0, \infty)$. Figure 5.1 seems to confirm this observation, especially taking into consideration that more than 45% of the simulated sample paths have $I(t) = 0$ for some $t > t_0$.
4. Given $R_0 > 1$, large population size m reduces the variance of the quasi-equilibrium distribution $\{q_i : i = 1, \ldots, n\}$ as well as the variation (fluctuation) within a single realization of the sample path. However, it does not reduce the variation among sample paths, especially before the quasi-equilibrium phase has been reached.
5. Under identical initial conditions, a single realization of an epidemic in the stochastic model can be dramatically different from that predicted by its deterministic counterpart. Most of the observational data collected during an outbreak arise from a single realization of such a stochastic event. This leads to a cautionary note about the danger of fitting such data to a deterministic model in order to estimate important parameters.

5.3 Susceptible-Infectious-Recovered Models

In Susceptible-Infectious-Recovered (SIR) models, the population is partitioned into three classes of individuals with $S(t)$, $I(t)$ and $R(t)$, representing the number of susceptible, infectious, and recovered individuals at time t.

(F1) Constant population size: $S(t) + I(t) + R(t) = m$ with no death or emigration, and no birth or immigration.
(F2) There is no replacement of susceptible individuals due to loss of immunity of recovered individuals.
(F3) Individuals immediately become infectious after being infected, without a latent period.
(F4) The infectious period T_I is exponentially distributed with hazard function γ and mean value $\mu_I = \gamma^{-1}$.
(F5) Assume homogeneous transmission in the environment in the sense that:

   (a) the force of infection onto a typical susceptible individual at any given time t is proportional to the proportion of infectious individuals, i.e. $\beta_1 \frac{I(t)}{n}$;
   (b) the instantaneous transmission rate of a typical infectious individual at any given time t is proportional to the proportion of susceptible individuals, i.e. $\beta_2 \frac{S(t)}{n}$;
   (c) bilinearity $\beta = \beta_1 = \beta_2$ that defines the probability of an instantaneous new infection in the population is given by $\beta \frac{S(t)I(t)}{n} dt$.

5.3.1 Representation of the SIR Model as a Bivariate Markov Process

Markov processes can be multivariate, defined in a multidimensional state space. In the study of infectious diseases, very often a population of size m is partitioned into k compartments, and a state is defined by a vector $\underline{i} = (i_1, i_2, \ldots, i_k)$ subject to $i_1 + i_2 + \cdots + i_k = m$. Each element in $\underline{i}$ takes values $\{0, 1, 2, \ldots m\}$. As a result, the state space $\mathfrak{S}$ contains

$$|\mathfrak{S}| = \binom{m+k-1}{k-1}$$

states. A Markov process defined on this space is a multivariate Markov process with $k-1$ independent response variables.

The birth–death process given by (5.1) corresponds to $k = 2$, where the population is partitioned into an infected class modeled by $\{I(t)\}$ and a susceptible class modeled by $\{S(t)\}$ subject to $S(t) = m - I(t)$. In this case, the birth–death process has one independent response variable $I(t)$.

In the SIR model, the population is partitioned into $k = 3$ compartments.

Transitions can only occur as

$$\begin{array}{l}(s, i, r) \to (s-1, i+1, r) \text{ a new infection} \\ (s, i, r) \to (s, i-1, r+1) \text{ a recovery}\end{array}.$$

Because $R(t) = m - S(t) - I(t)$, the SIR model is a bivariate Markov process. The transition probabilities are often modeled by

$$\begin{aligned}\Pr\left\{\begin{pmatrix} S(t+dt) = s-1 \\ I(t+dt) = i+1 \end{pmatrix} \middle| \begin{pmatrix} S(t) = s \\ I(t) = i \end{pmatrix}\right\} &= \beta \tfrac{si}{m} dt, \\ \Pr\left\{\begin{pmatrix} S(t+dt) = s \\ I(t+dt) = i-1 \end{pmatrix} \middle| \begin{pmatrix} S(t) = s \\ I(t) = i \end{pmatrix}\right\} &= \gamma i dt.\end{aligned} \tag{5.18}$$

This bivariate Markov model is time-stationary. One can define the transition probability

$$p_{si}(t) = \Pr\left\{\begin{pmatrix} S(t) = s \\ I(t) = i \end{pmatrix} \middle| \begin{pmatrix} S(0) = s_0 \\ I(0) = i_0 \end{pmatrix}\right\}, \quad s_0 + i_0 = m. \tag{5.19}$$

As before, the initial condition $(S(0), I(0)) = (s_0, i_0)$ is omitted in the notation $P_{si}(t)$ for simplicity. The Kolmogorov forward equations are written as $\frac{d}{dt} P_{s_0 i_0}(t) = -i_0 \left(\frac{\beta s_0}{m} + \gamma\right) P_{s_0 i_0}(t)$ and

$$\begin{aligned}\tfrac{d}{dt} P_{si}(t) = \tfrac{\beta(s+1)(i-1)}{m} P_{s+1,i-1}(t) + \gamma(i+1) P_{s,i+1}(t) \\ - \left(\tfrac{\beta s}{m} + \gamma\right) i P_{si}(t)\end{aligned} \tag{5.20}$$

where $0 \le s + i \le n$, $0 \le s \le s_0$, $0 \le i \le n$. Similar to the interpretation of (5.9), the first two terms of (5.20) with positive signs represent the in-flows

$$\begin{aligned}(s+1, i-1) &\to (s, i) : \text{with rate } \tfrac{\beta(s+1)(i-1)}{m} \\ (s, i+1) &\to (s, i) : \text{with rate } \gamma(i+1)\end{aligned}$$

and the third term, with negative sign, represents the outflows $(s, i) \to (s-1, i+1)$ and $(s, i) \to (s, i-1)$, with instantaneous rates $\frac{\beta si}{m}$ and γi, respectively.

We write the bivariate moment generating function as $\mathcal{M}(u, v|t) = E\left[e^{uS(t)+vI(t)}\right]$. Multiplying both sides of (5.20) by e^{us+vi} and summing over $(s, i) \in \mathfrak{S}$, the correspondence to (5.10) is (Isham 1991)

$$\frac{\partial \mathcal{M}(u,v|t)}{\partial t} = \gamma(e^{-v} - 1)\frac{\partial \mathcal{M}(u,v|t)}{\partial v} + \frac{\beta}{n}\left(e^{v-u} - 1\right)\frac{\partial^2 \mathcal{M}(u,v|t)}{\partial u \partial v}. \tag{5.21}$$

Using the facts that

$$\begin{aligned}E[S(t)] &= \tfrac{\partial \mathcal{M}(u,v|t)}{\partial u}|_{u=v=0}\ E[I(t)] = \tfrac{\partial \mathcal{M}(u,v|t)}{\partial v}|_{u=v=0} \\ E[S^2(t)] &= \tfrac{\partial^2 \mathcal{M}(u,v|t)}{\partial u^2}|_{u=v=0}\ E[I^2(t)] = \tfrac{\partial^2 \mathcal{M}(u,v|t)}{\partial v^2}|_{u=v=0}, \\ E[S(t)I(t)] &= \tfrac{\partial^2 \mathcal{M}(u,v|t)}{\partial u \partial v}|_{u=v=0} \\ E[S^2(t)I(t)] &= \tfrac{\partial^3 \mathcal{M}(u,v|t)}{\partial u^2 \partial v}|_{u=v=0}\ \ E[S(t)I^2(t)] = \tfrac{\partial^3 \mathcal{M}(u,v|t)}{\partial u \partial v^2}|_{u=v=0},\end{aligned}$$

further differentiating (5.21) repeatedly with respect to u or v and letting $u = v = 0$, one gets the equations describing the rate of change of the first moments

$$\begin{cases} \frac{d}{dt}E[S(t)] = -\beta\frac{E[S(t)I(t)]}{m} \\ \frac{d}{dt}E[I(t)] = \beta\frac{E[S(t)I(t)]}{m} - \gamma E[I(t)] \end{cases}. \tag{5.22}$$

The rate of change of the first moments $E[S(t)]$ and $E[I(t)]$ over time depend on the second order cross moment between $S(t)$ and $I(t)$ given by $E[S(t)I(t)]$. Continuing, one gets

$$\begin{aligned}\frac{d}{dt}E[S^2(t)] &= \beta\frac{E[S(t)I(t)]}{m} - 2\beta\frac{E[S^2(t)I(t)]}{m} \\ \frac{d}{dt}E[S(t)I(t)] &= -\left(\frac{\beta}{m} + \gamma\right)E[S(t)I(t)] + \frac{\beta}{m}\left(E[S^2(t)I(t)] - E[S(t)I^2(t)]\right) \\ \frac{d}{dt}E[I^2(t)] &= \gamma E[I(t)] - 2\gamma E[I^2(t)] + \frac{\beta}{m}\left(E[S(t)I(t)] + 2E[S(t)I^2(t)]\right) \\ &\vdots\end{aligned}$$

The equations describing the rate of change of the r-th moments depend on the $(r+1)$-th moments. For moment closure, Isham (1991) assumes that at each time t, $\{S(t), I(t)\}$ follow a bivariate normal distribution so that all the higher order moments can be expressed as functions of the mean and elements in the variance–covariance matrix of the bivariate normal distribution.

Using the covariance expression $cov(X, Y) = E[XY] - E[X]E[Y]$, the change of the first moments for $S(t)$ and $I(t)$ in (5.22) can be re-written as

$$\begin{cases} \frac{d}{dt}E[S(t)] = -\frac{\beta E[S(t)]E[I(t)]}{m} - \frac{\beta}{m}cov\{S(t), I(t)\} \\ \frac{d}{dt}E[I(t)] = \frac{\beta E[S(t)]E[I(t)]}{m} - \gamma E[I(t)] + \frac{\beta}{m}cov\{S(t), I(t)\} \end{cases}. \tag{5.23}$$

The term involving $cov\{S(t), I(t)\}$ becomes negligible as $m \to \infty$.

5.3.2 The Kermack and McKendrick Deterministic SIR Model

Like the discussion for the deterministic SIS model, let $S_d(t)$, $I_d(t)$, and $R_d(t)$ be deterministic (non-random), continuous, and differentiable functions of t, representing the number of susceptible, infectious, and recovered (removed) individuals subject to $S_d(t) + I_d(t) + R_d(t) = m$ for all $t > 0$, the deterministic SIR model (Kermack and McKendrick 1927) is then given by a system of ordinary differential equations by replacing $E[S(t)]$ and $E[I(t)]$ in (5.23) with $S_d(t)$ and $I_d(t)$ and setting $cov\{S(t), I(t)\} = 0$ as moment closure. The model is given by:

$$\begin{cases} \frac{d}{dt}S_d(t) = -\beta\frac{S_d(t)I_d(t)}{m} \\ \frac{d}{dt}I_d(t) = \beta\frac{S_d(t)I_d(t)}{m} - \gamma I_d(t) \end{cases}. \tag{5.24}$$

It implies $\frac{d}{dt}R_d(t) = \gamma I_d(t)$.

Compare Simulated Stochastic SIR Outbreaks with the Deterministic Model

Similar to the comparison between stochastically simulated paths of SIS outbreaks and $I_d(t)$ determined by the ordinary differential equation (5.13), an algorithm can be also implemented in MATLAB or R in two steps: (1) simulate the sample paths of the embedded bivariate discrete time Markov chain; (2) simulate the sojourn times.

Without going into details of the algorithm, we compare results in Fig. 5.2.

1. The deterministic model, given the initial conditions, says "what must happen" as determined by the function $I_d(t)$, such as the precise peak time and magnitude. Under the same initial condition, the stochastic model tells "what might happen." In fact, in both simulations with $m = 1000$ and $m = 10{,}000$, approximately $1/3$ of the times the outbreak did not take place (as the straight line near $I(t) \approx 0$).

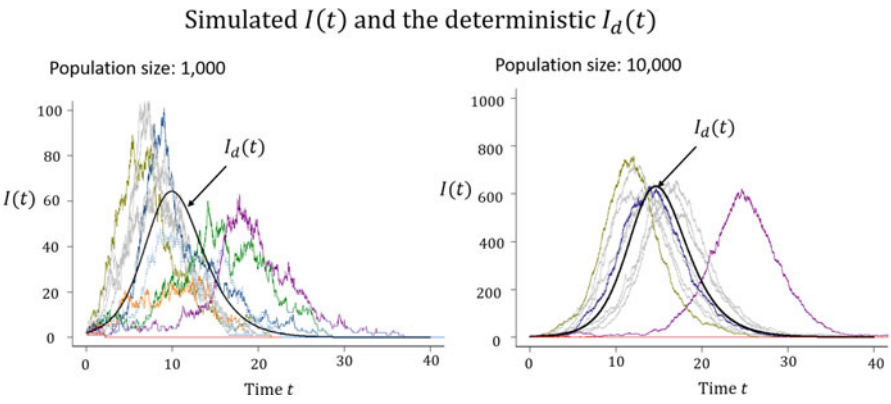


Fig. 5.2 A comparison of stochastically simulated sample paths for $I(t)$ in the SIR model (5.18) against $I_d(t)$ (solid black lines) in the model (5.24) in two population sizes ($m = 1{,}000$ and $m = 10{,}000$) at: $\beta = 1.5$, $\gamma = 1$

2. Large population size m reduces temporal variations within sample paths so that they converge to smooth (random) functions as $m \to \infty$, but not identical to $I_d(t)$.
3. The deterministic model, as approximation to the mean of its stochastic counterpart, should be understood as the average of sample paths assuming a large number of repetitions of outbreaks in the identical population with identical initial conditions, which can only happen in theory or in simulated scenarios.

Important Relationships and Quantities from the Deterministic SIR Model

There is a fundamental difference between the stochastic and the deterministic SIR models. In (5.18), m is an essential parameter that determines the numbers of absorbing and transient states in the model. The state space $\mathfrak{S}$ contains triplets (s, i, r) with

$$|\mathfrak{S}| = \frac{(m+1)(m+2)}{2}$$

states. There are $m+1$ absorbing states: $\mathfrak{S}_1 = \{(s, 0, m-s) : s = 0, 1, \ldots, m\}$. The rest of transient states are in $\mathfrak{S}_2$ with $|\mathfrak{S}_2| = \frac{m(m+1)}{2}$. Table 5.2 illustrates the dimension of the state space for a small population less than or equal to 500.

On the other hand, in (5.24), both γ and m are innocent parameters. Letting $x(t) = \frac{S_d(t)}{m}$, $y(t) = \frac{I_d(t)}{m}$ and $z(t) = \frac{R_d(t)}{m}$ subject to $x(t) + y(t) + z(t) = 1$, and re-scaling time $\tau = \gamma t$, (5.24) becomes

$$\begin{cases} \frac{d}{d\tau}x(\tau) = -\frac{\beta}{\gamma}x(\tau)y(\tau), \\ \frac{d}{d\tau}y(\tau) = \frac{\beta}{\gamma}x(\tau)y(\tau) - y(\tau) \end{cases}. \tag{5.25}$$

Table 5.2 Tabulation of the number of states in the stochastic SIR model with populations sizes up to 500

Pop. size m	10	20	50	100	200	500
Absorbing states $\lvert S_1\rvert$	11	21	51	101	201	501
Transient states $\lvert S_2\rvert$	55	210	1275	5050	20,100	125,250
Total states $\lvert S\rvert$	66	231	1326	5151	20,301	125,751

It implies $\frac{dz(\tau)}{d\tau} = y(\tau)$. The only essential parameter is $R_0 = \beta/\gamma$. The deterministic model is simple to analyze. For the preparation of a list of key results, we put forward the following preserved relationship

$$S_d(t) + I_d(t) - \frac{m}{R_0}\log S_d(t) = S_d(0) + I_d(0) - \frac{m}{R_0}\log S_d(0). \tag{5.26}$$

It is derived from (5.24) using the relationship $\frac{dI_d}{dS_d} = -1 + \frac{\gamma m}{\beta S_d(t)}$, which implies $I_d(t) = -S_d(t) + \frac{m}{R_0}\log S_d(t) + c$, where c is an arbitrary constant. Therefore, $S_d(t) + I_d(t) - \frac{m}{R_0}\log S_d(t)$ defines an orbit for some choice of c, which is determined by the initial values $S_d(0)$ and $I_d(0)$ of $S_d(t)$ and $I_d(t)$. Since $S_d(t) + I_d(t) + R_d(t) = S_d(0) + I_d(0) = m$, (5.26) can also be written as

$$S_d(t) = S_d(0)\exp\left(-\frac{R_0}{m}R_d(t)\right). \tag{5.27}$$

With respect to (5.25), (5.26) becomes

$$1 - x(t) - y(t) + \frac{1}{R_0}\log\frac{x(t)}{x_0} = 0 \tag{5.28}$$

where $x_0 = \frac{S_d(0)}{m}$, and (5.27) can be re-written as $x(t) = x_0\exp(-R_0 z(t))$.

Furthermore, with respect to time $\tau = \gamma t$,

$$\begin{aligned}\frac{dz(\tau)}{d\tau} &= y(\tau) = 1 - z(\tau) - x(\tau)\\ &= 1 - z(\tau) - x_0\exp(-R_0 z(\tau))\end{aligned}$$

which yields, as originally given in (Deakin 1975),

$$\tau = \int_0^{z(\tau)} \frac{1}{1 - x - x_0 e^{-R_0 x}}dx. \tag{5.29}$$

The essential parameter $R_0 = \beta/\gamma$ determines the following important quantities through these preserved relationships.

The Final Size A more important transcendental relationship is the one that links the threshold parameter R_0 defined by (4.2) at the very beginning of an epidemic with the outcome at the very end of an epidemic. They are expressed as the final size equation given by (5.32) below and illustrated in Fig. 6.1 in Chap. 6.

In the deterministic SIR model, given the initial condition $S_d(0) + I_d(0) = m$, the relationship (5.26) leads to

$$m - S_d(\infty) + \frac{m}{R_0} \log\left(\frac{S_d(\infty)}{S_d(0)}\right) = 0, \tag{5.30}$$

where $S_d(\infty)$ is the quantity of interest representing the number of susceptible individuals who eventually escape from infection by the end of the outbreak and $C_d(\infty) = m - S_d(\infty)$ is the *final size*. Because $\frac{d}{dt}\left(S_d(t) + I_d(t)\right) = -\gamma I_d(t)$ and $S_d(t) + I_d(t) + R_d(t) = m$, one also gets

$$\gamma \int_0^\infty I_d(t)dt = C_d(\infty). \tag{5.31}$$

The total area $\int_0^\infty I_d(t)dt = \gamma^{-1} C_d(\infty)$ is sometimes called the value of the epidemic because it is the total infectious person time of the epidemic.

Re-writing the initial condition as $x_0 = S_d(0)/m$, the final size equation, using (5.28), can be written as $1 - x(\infty) = -\frac{1}{R_0} \log \frac{x(\infty)}{x_0} 0$. The left side, denoted by $\eta = 1 - x(\infty)$, is the proportion of the population that will be eventually infected by the end of the epidemic, corresponding to the large outbreak in Sect. 4.2.3 with the notation $C_d(\infty)/m$. Then (5.30) can be written as

$$1 - \eta = x_0 e^{-R_0 \eta}. \tag{5.32}$$

This is the nonlinear monotone relationship (assuming $x_0 \approx 1$) between η and R_0 as shown in Fig. 6.1.

The Peak Prevalence The maximum value of $I_d(t)$, denoted by $I_d^{\max}$, is the peak prevalence and represents the maximum disease burden at a given time. It is attained when $\frac{S_d(t)}{m} = x(t) = \frac{1}{R_0}$, at which $\frac{d}{dt} I_d(t) = 0$. Given the initial condition $x_0 = \frac{S_d(0)}{m}$, $I_d^{\max}$ is derived from (5.26):

$$I_d^{\max} = m\left(1 - \frac{1}{R_0}\left[1 + \log\left(R_0 x_0\right)\right]\right). \tag{5.33}$$

Define the Incidence Quantiles Similar to the definition quantile for the cumulative distribution in probability theory, we defined the *incidence quantile*. The q^{th}-quantile for the cumulative incidence $C_d(t) = m - S_d(t)$ is t_q, satisfying

$$\frac{C_d(t_q)}{m} = q \frac{C_d(\infty)}{m} = q\eta,$$

where $\frac{1-x_0}{\eta} \le q < 1$, $x_0 = \frac{S_d(0)}{m}$ and η satisfies (5.32). For instance, if $q = 0.25$, $t_{0.25}$ is the time from the beginning of the epidemic until it reaches 25% of the final size.

Proposition 19 (Kendall 1956; Deakin 1975) *For the SIR model specified by (5.24), the time when the cumulative incidence $C_d(t)$ reaches $C_d(t)/m = q\eta$ can be calculated by*

$$\tau_q = \gamma t_q = \int_0^{\frac{1}{R_0}\log\frac{x_0}{1-\eta q}} \frac{1}{1 - x - x_0 e^{-R_0 x}} dx. \tag{5.34}$$

Proof From (5.28), $z(\tau) = \frac{1}{R_0}\log\frac{x_0}{x(\tau)}$. The time $\tau_q = \gamma t_q$ when $\frac{C_d(t_q)}{m} = 1 - x(t_q) = q\eta$. Hence $z(\tau_q) = \frac{1}{R_0}\log\frac{x_0}{x(\tau_q)} = \frac{1}{R_0}\log\frac{x_0}{1+\eta q}$. Letting the left-hand side of (5.29) be τ_q, one immediately gets (5.34). ■

We notice that (5.32) can be written as $\frac{1}{R_0}\log\frac{x_0}{1-\eta} = \eta$. Thus

$$\lim_{q\to 1} \frac{1}{R_0}\log\frac{x_0}{1-\eta q} = \eta.$$

Meanwhile, the denominator of the integrand approaches zero as $x \to \eta$. When $q = 100\%$, the integration diverges. In other words, the expected duration, according to the deterministic SIR model, is infinity.

The Timing of the Peak Prevalence and the Peak Incidence The timing of the peak prevalence can be calculated by setting $q = \frac{1}{\eta}\left(1 - \frac{\gamma}{\beta}\right)$ in (5.34) so that $x(t_{\max.I}) = \frac{1}{R_0}$. One gets

$$\gamma t_{\max.I} = \int_0^{\frac{1}{R_0}\log R_0 x_0} \frac{1}{1 - x - x_0 \exp(-R_0 x)} dx. \tag{5.35}$$

The peak incidence for $i_d(t) = \beta\frac{S_d(t)I_d(t)}{m}$ is attained when $\frac{S_d(t)}{m} - \frac{I_d(t)}{m} = x(t) - y(t) = \frac{1}{R_0}$. From (5.28), $y(t) = 1 - x(t) + \frac{1}{R_0}\log\frac{x(t)}{x_0}$. If $i_d(t)$ arrives at its peak at time $t_{\max.i}$, then $2x(t_{\max.i}) - \frac{1}{R_0}\log\frac{x(t_{\max.i})}{x_0} = 1 + \frac{1}{R_0}$, where $x(t_{\max.i}) = \frac{S_d(t_{\max.i})}{m}$. Let $u(t_{\max.i}) = 1 - x(t_{\max.i}) = \frac{C(t_{\max.i})}{m}$, then $u(t_{\max.i})$ can be solved as the root u of the equation $2(1-u) - \frac{1}{R_0}\log\frac{1-u}{x_0} = 1 + \frac{1}{R_0}$, or equivalently

$$2u - \frac{1}{R_0}\log\frac{x_0}{1-u} = 1 - \frac{1}{R_0}. \tag{5.36}$$

To determine $t_{\max.\mathrm{i}}$, we first solve (5.36) for u to obtain $u(t_{\max.\mathrm{i}})$ and then let $q = \frac{u(t_{\max.\mathrm{i}})}{\eta}$ in (5.34) to get

$$\gamma t_{\max.\mathrm{i}} = \int_0^{\frac{1}{R_0}\log\frac{x_0}{1-u(t_{\max.\mathrm{i}})}} \frac{1}{1-x-x_0\exp(-R_0 x)}dx. \tag{5.37}$$

Meanwhile, the peak incidence value is $i_d^{\max} = \beta\frac{S_d(t_{\max.\mathrm{i}})I_d(t_{\max.\mathrm{i}})}{m}$ with $\frac{S_d(t_{\max.\mathrm{i}})}{m} = 1-u(t_{\max.\mathrm{i}})$ and $I_d(t_{\max.\mathrm{i}}) = my(t_{\max.\mathrm{i}})$, so that

$$i_d^{\max} = \beta m\,(1-u(t_{\max.\mathrm{i}}))\left(u(t_{\max.\mathrm{i}}) - \frac{1}{R_0}\log\frac{x_0}{1-u(t_{\max.\mathrm{i}})}\right). \tag{5.38}$$

Since $u(t_{\max.\mathrm{i}})$ satisfies (5.36) and $\beta = R_0/\gamma$, we have

$$\frac{\gamma}{m}i_d^{\max} = R_0\,(1-u(t_{\max.\mathrm{i}}))\left(1-u(t_{\max.\mathrm{i}}) - \frac{1}{R_0}\right). \tag{5.39}$$

Thus, given the population size m, $i_d^{\max}$ is also scaled by γ.

The dynamics of the incidence $i(t)$, the cumulative incidence $C(t)$, the prevalence $I(t)$ and the depletion of the susceptible population $S(t)$ from the deterministic SIR model are schematically illustrated in Fig. 5.3.

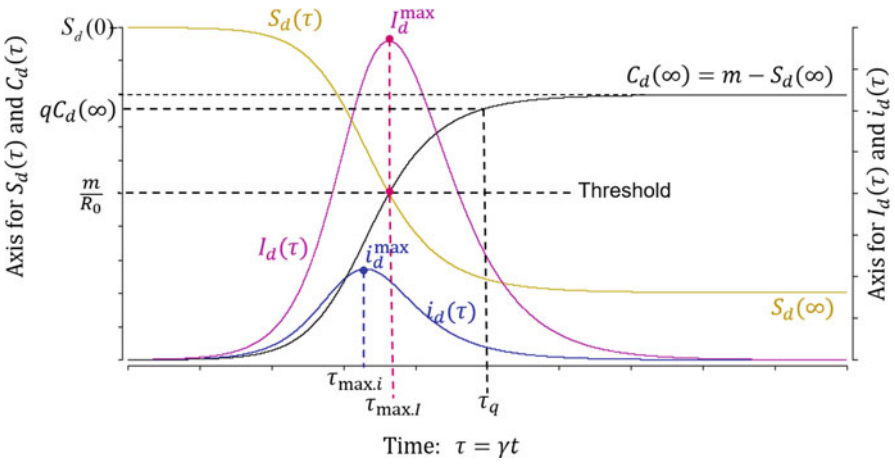


Fig. 5.3 A schematic illustration of all the quantities of the deterministic SIR model at standardized time $\tau = \gamma t$

Summary

1. The final size $C_d(\infty)$ and the peak prevalence $I_d^{\max}$ are scaled by the population size m, independent of the time scale γ.
2. The incidence quantile t_q, the timing of the peak prevalence $t_{\max.\mathrm{I}}$, and the peak incidence $t_{\max.\mathrm{i}}$ are scaled by γ, independent of the population size m.
3. The value of the epidemic $\int_0^\infty I_d(t)dt = \gamma^{-1}C_d(\infty) = m\eta/\gamma$ and the value of the peak incidence $i_d^{\max}$ are scaled by the ratio m/γ.

Example 20 As a numerical example, we consider a fairly large population $m = 10{,}000$ with a single initially infected individual, so that $S_d(0) = 9999$. We fix the value of the basic reproduction number $R_0 = \beta/\gamma = 2$. We compare two scenarios: (a) $\beta = 1$ and $\gamma = 1/2$; (b) $\beta = 0.5$ and $\gamma = 1/4$. In both scenarios, $S_d(\infty) = 2031.5$, the final size $C_d(\infty) = 7968.5$, and the peak prevalence $I_d^{\max} = 1534.8$. The quantile t_q for scenario (a) is half of that for scenario (b). Therefore, the timing of peak prevalences, $t_{\max.\mathrm{I}} = 18.144$ in scenario (a) and $t_{\max.\mathrm{I}} = 36.287$ in scenario (b); the timing of peak incidences, $t_{\max.\mathrm{i}} = 16.503$ in scenario (a) and $t_{\max.\mathrm{i}} = 33.007$ in scenario (b). These are all because the average infectious period in scenario (b) is twice as long as that for scenario (a). Meanwhile, the value of the epidemic $\int_0^\infty I_d(t)dt$ in scenario (a) is half of that in scenario (b), and the peak incidence $i_d^{\max}$ in scenario (a) is twice the value of that in scenario (b), 876.24 vs. 439.12, respectively. There are two quantities that do not depend on m or γ. They are $\eta = 0.7968$ by solving (5.32) and $u(t_{\max.\mathrm{i}}) = 0.36254$ by solving (5.36) (Fig. 5.4).

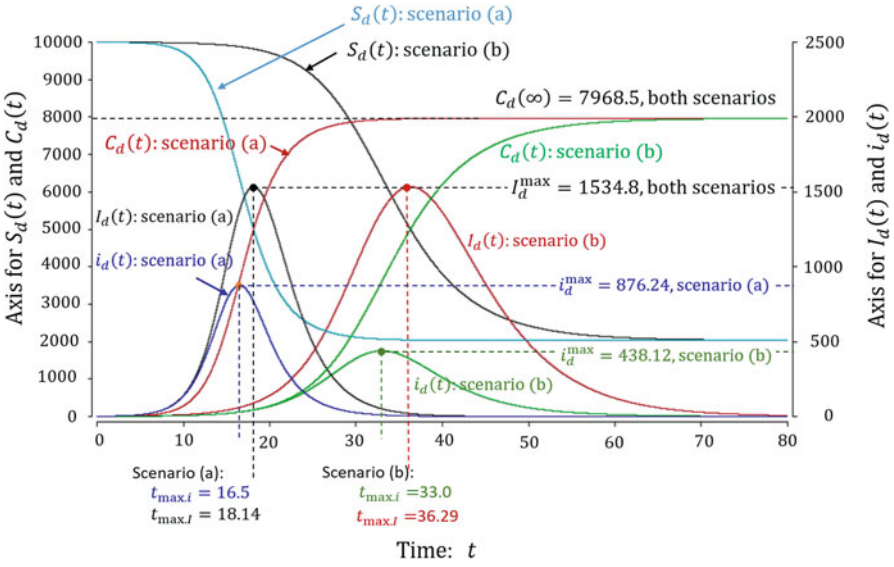


Fig. 5.4 Illustration of the quantities in Example 20

5.3.3 *The Deterministic SIR Model with Non-exponentially Distributed Infectious Periods*

Sir David Cox (2006) wrote:

> It is important to distinguish the primary features from the secondary features of the model. If a primary feature is changed, the research questions of interest are either changed or at least formulated in an importantly different way. If a secondary feature is changed, the research questions are essentially unaltered. Their influence is typically on the method, e.g. simplified, feasible ways to analyze the model or improving the precision of parameter estimates.

We extend the distribution of the infectious period T_I from the exponential distribution to non-exponential distributions. Since the deterministic SIR model (5.24) has yielded a list of important relationships and results, from (5.26) to (5.39), we discuss which of them are generalizable to other infectious period distributions and which depend on the exponential distribution as a primary feature.

Relationships and Quantities Derived from (5.24)–(5.25) in Which the Exponential Distribution of the Infectious Periods Is a Secondary Feature

The Formulation $R_0 = \beta\mu_I$ The essential parameter $R_0 = \beta/\gamma$ in (5.25) is $R_0 = \beta\mu_I$ where $\mu_I = \gamma^{-1}$. The general expression is $R_0 = \beta\mu_I$ in models with SEI structures for arbitrarily distributed infectious periods, as long as μ_I exists. R_0 is the basic reproduction number. In the case without latent periods, the probability generating function (4.4) for N such that $R_0 = E[N]$ is $G_N(s) = \int_0^\infty e^{-\beta x(1-s)} dF_I(x)$. Thus,

$$R_0 = G_N^{'}(1) = \beta \int_0^\infty x dF_I(x) = \beta\mu_I.$$

The Final Size Equation (5.32) The exponential distribution of the infectious periods is a secondary feature with respect to the final size equation. From a stochastic perspective, we first present the following ad hoc arguments under the assumptions that

1. All infectious individuals are equally infectious, regardless when each individual is infected and how long it has been infected. Then the force of infection onto a specific susceptible individual ν_s, $h(t|\nu_s)$, may depend on its susceptibility.
2. All susceptible individuals are also made of the same type, with equal susceptibility, then $h(t|\nu_s) = h(t)$. In this case, the force of infection is $h(t) = \beta \frac{E[I(t)]}{m}$ and the parameter β captures the hazard of becoming infected.

The cumulative force of infection of a typical susceptible individual through its lifetime is $\int_0^\infty h(t)dt = \frac{\beta}{m}\int_0^\infty E[I(t)]dt$. Holding $m = S(t) + I(t) + R(t)$ constant, the probability that a susceptible individual ever gets infected throughout the epidemic is

$$\eta = 1 - \exp\left\{-\int_0^\infty h(t)dt\right\} = 1 - \exp\left\{-\frac{\beta}{m}\int_0^\infty E[I(t)]dt\right\}. \tag{5.40}$$

where $\int_0^\infty E[I(t)]dt$ is the expected total infectious person time and $\mu_I = E[T_I]$ is the average time spent infectious per infection. Thus, $\frac{1}{m}\int_0^\infty E[I(t)]dt = \eta\mu_I$ and

$$\eta = 1 - \exp\{-\eta\beta\mu_I\} = 1 - \exp\{-R_0\eta\}.$$

Note that in the deterministic framework, we approximate $E[I(t)]$ as $I_d(t)$ and $\int_0^\infty I_d(t)dt = m\eta\mu_I = \mu_I C_d(\infty)$.

A formal theory is given by the Proposition 21 below.

Proposition 21 *$\frac{C(\infty)-m\eta}{\sqrt{m}}$ has Gaussian limit distribution $N(0,\sigma^2)$ of which the asymptotic variance is*

$$\sigma^2 = \frac{\eta(1-\eta)}{(1-R_0^2\eta)} + \frac{\eta^2(var[N]-R_0)(1-\eta+\varepsilon)}{(1-R_0\eta)^2}. \tag{5.41}$$

where η is the root of the final size equation $1-\eta = \exp(-R_0(\eta+\varepsilon))$, and N is the random variable corresponding to $R_0 = E[N]$.

This central limit tendency has been studied and proven by many authors under different assumptions regarding disease transmission. We refer readers to von Bahr and Martin-Löf (1980), Ludwig (1975), Scalia-Tomba (1985), Martin-Löf (1988) and Lefèvre and Picard (1995), among many others, for the suitable conditions and proofs.

In Proposition 21, the cumulative number of infections $C(\infty)$ is a discrete random variable taking integer values. When $R_0 > 1$, conditioning on a large outbreak and $\frac{I(0)}{m} = 1 - x_0$, as m becomes large, the random variable $\frac{C(\infty)}{m}$ converges in distribution to a point mass at η. The fluctuations around the limit are Gaussian of order $\frac{1}{\sqrt{m}}$, which become large if the variance $var[N]$ is large. According to this proposition, the exponential distribution of the infectious period is a secondary feature for the mean final size η but is a primary feature for the asymptotic variance of the final size. In this special case, $var[N] = R_0 + R_0^2$.

The Relationship (5.31) The relationship $\int_0^\infty I_d(t)dt = \mu_I C_d(\infty) = m\eta\mu_I$ is general for arbitrarily distributed infectious periods, as long as μ_I exists, and (5.31) is the special case with $\mu_I = \gamma^{-1}$.

Relationships and Quantities Derived from (5.24)–(5.25) in Which the Exponential Distribution of the Infectious Periods Is a Primary Feature

The Peak Prevalence Is Attained When $\frac{S_d(t)}{m} = x(t) = R_0^{-1}$ This relationship only holds when the recovery rate γ is constant. The exponential distribution of the infectious periods is a primary feature. If the recovery rate is a function of time $\gamma_c(t)$, it leads to a nonautonomous differential equation

$$\frac{d}{dt} I_d(t) = \beta \frac{S_d(t) I_d(t)}{m} - \gamma_c(t) I_d(t). \tag{5.42}$$

The threshold condition for $\frac{d}{dt} I_d(t) = 0$ is $\beta \frac{x(t)}{\gamma_c(t)} = 1$. In other words, the peak prevalence is attained when

$$x(t) = \gamma_c(t)/\beta.$$

This condition does not transcend to R_0 through the depletion of susceptible individuals $x(t)$ without an explicit model for $\gamma_c(t)$. Very soon we shall see that, in the case of non-exponentially distributed infectious periods, the recovery rate is an implicit function of time $\gamma_c(t)$, depending on the historical incidence $i_d(s) = \beta \frac{S_d(s) I_d(s)}{m}$, $s \leq t$ and the specific distribution of the infectious period.

The Preserved Relationships and Some of Their Derived Quantities In the deterministic SIR model, we have an important preserved relationship (5.26), alternatively expressed as (5.27). With respect to (5.25), (5.26) becomes (5.28), or equivalently, $x(t) = x_0 \exp(-R_0 z(t))$. The latter leads to the important relationship (5.29). All these relationships are derived from $\frac{dI_d}{dS_d} = -1 + \frac{\gamma m}{\beta S_d(t)}$ and $\frac{dR_d}{dS_d} = -\frac{\gamma m}{\beta S_d(t)}$, where γ is constant. Therefore, the exponential distribution of the infectious periods is a primary feature.

One of the quantities is t_q given by (5.34), derived from (5.29). It further derives quantities such as the timing of the peak prevalence (5.35) and the timing of the peak incidence (5.37). All these results depend on the assumption of an exponentially distributed infectious period. For generally distributed infectious period, Fig. 5.5 illustrates that, for the same R_0, the infectious period distribution has a profound effect on the transmission dynamic over time.

The exponential distribution assumption for the infectious period is also a primary feature for the peak incidence. For the peak incidence (5.38)–(5.39), the quantity $u(t_{\text{max.i}})$ is the root of u of Eq. (5.36), which is derived from (5.28).

The Erlang Distributed Infectious Periods

Of the SIR models with non-exponentially distributed infectious periods, the one corresponding to the Erlang distributed infectious periods can be written as a system of ordinary differential equations.

The Erlang distribution is a subset of the gamma distribution with integer-valued shape parameter $\kappa = 1, 2, \ldots$ including the exponential distribution as a special case $\kappa = 1$. Using the Erlang distribution to model the latent and the infectious periods has been seen in the literature in both deterministic and stochastic frameworks, such as Anderson and Watson (1980), Wearing et al. (2005), Feng et al. (2007), and many others.

For the SIR model, we use μ_I for the mean infectious periods and κ_I for the corresponding shape parameter. The model (5.24) corresponds to $\kappa_I = 1$. When $\kappa_I = 2, 3, \ldots$ the equation $\frac{d}{dt} I_d(t) = \beta \frac{S_d(t) I_d(t)}{m} - \gamma I_d(t)$ in (5.24) is expanded into κ_I separate equations

$$\begin{aligned} \frac{d}{dt} I_{d1}(t) &= \beta \frac{S_d(t) I_d(t)}{m} - \frac{\kappa_I}{\mu_I} I_{d1}(t), \\ \frac{d}{dt} I_{dj}(t) &= \frac{\kappa_I}{\mu_I} I_{dj-1}(t) - \frac{\kappa_I}{\mu_I} I_{dj}(t), \quad j = 2, 3, \cdots, \kappa_I. \end{aligned} \tag{5.43}$$

With these ordinary differential equations, one can numerically calculate the expected cumulative infections $C_d(t) = n - S_d(t)$, the expected prevalence $I_d(t) = \sum_{j=1}^{\kappa_I} I_{dj}(t)$, and other derived quantities. In this model, the infectious period is composed as the sum of κ_I independently and identically distributed periods, with mean durations equal to μ_I/κ_I.

Given the same mean value of infectious period μ_I, the Erlang distribution is more homogeneous than the exponential distribution, both in terms of the variance $var[T_I] = \mu_I^2/\kappa_I$ and the Laplace transform $L[f_I](s) = (1 + s\mu_I/\kappa_I)^{-\kappa_I}$. The Erlang distribution is a subset of the gamma distribution only for $\kappa \geq 1$. Therefore, using the ordinary differential equations given by (5.43) is only suitable for diseases with clear evidence that their infectious periods are less variable than those modeled using the exponential distribution with equal mean values.

Example 22 We compare two deterministic SIR models, both with the mean infectious period $\mu_I = 4$ and $R_0 = 2$ (implying $\beta = 0.5$). The infectious period in Model 1, corresponding to $S_d^{(1)}(t)$ and $I_d^{(1)}(t)$, is exponentially distributed. The infectious period in Model 2, corresponding to $S_d^{(2)}(t)$ and $I_d^{(2)}(t)$, is Erlang distributed with shape parameter $\kappa_I = 7$. For numerical illustration, we choose a population size $m = 10^4$ with $I_d(0) = 1$. Figure 5.5 illustrates that

1. $S_d^{(1)}(\infty) = S_d^{(2)}(\infty) = 2031.5$, satisfying the final size equation $10{,}000 - x + \frac{10{,}000}{2} \log\left(\frac{x}{9999}\right) = 0$.
2. $I_d^{(1)}(t)$ arrives at maximum value $I_{d\,\max}^{(1)}$ at time $t_{\max.\mathrm{I}}^{(1)}$ and $S_d^{(1)}(t_{\max.\mathrm{I}}^{(1)}) = \frac{m}{R_0} = 5000$. On the other hand, $I_d^{(2)}(t)$ arrives at maximum value $I_{d\,\max}^{(2)}$ at time $t_{\max.\mathrm{I}}^{(2)}$ and $S_d^{(2)}(t_{\max.\mathrm{I}}^{(2)}) < \frac{m}{R_0} = 5000$.
3. The final sizes in both models are equal: $C_d(\infty) = 7968.5$. Thus, $\int_0^\infty I_d^{(1)}(t)dt = \int_0^\infty I_d^{(2)}(t)dt = \mu_I C_d(\infty) = 31{,}874$. Because $t_{\max.\mathrm{I}}^{(2)} < t_{\max.\mathrm{I}}^{(1)}$, so $I_{d\,\max}^{(2)} > I_{d\,\max}^{(1)}$.

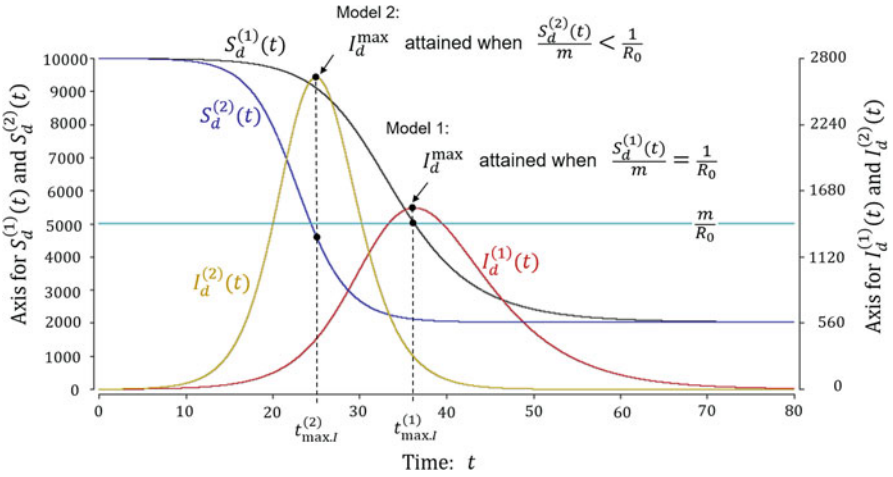


Fig. 5.5 Illustrations corresponding to Example 22

Generally Distributed Infectious Periods

The ordinary differential equations (5.24) correspond to the $f_I(x) = \gamma e^{-\gamma x}$, where γ is a hazard rate. Generalization to an arbitrarily distributed infectious period can be achieved with the p.d.f. and survivor function, $f_I(x)$ and $\overline{F}_I(x)$ respectively. The hazard function is time-dependent $h_I(x) = f_I(x)/\overline{F}_I(x)$, in which x is measured from the time at infection of a typical infected individual. With this generalization, (5.24) becomes a system of integro-differential equations

$$\begin{cases} \frac{d}{dt} S_d(t) = -\beta \frac{S_d(t) I_d(t)}{m} \\ \frac{d}{dt} I_d(t) = i_d(t) - \int_0^t i_d(s) f_I(t-s) ds, \\ \frac{d}{dt} R_d(t) = \int_0^t i_d(s) f_I(t-s) ds. \end{cases} \tag{5.44}$$

where $i_d(t) = \beta \frac{S_d(t) I_d(t)}{m}$. It returns to (5.24) when $f_I(x) = \gamma e^{-\gamma x}$ because

$$\int_0^t i_d(s) f_I(t-s) ds = \gamma \int_0^t i_d(s) e^{-\gamma(t-s)} ds = \gamma I_d(t).$$

The second equation $I_d(t) = \int_0^t i_d(s) e^{-\gamma(t-s)} ds$ is due to the fact that $e^{-\gamma(t-s)}$ is the conditional probability that an individual is still infectious at time t, given the time at infection $s < t$.

When the distribution of T_I is not exponential, not only does the recovery rate depend on time since infection x according to the hazard function $h_I(x)$ from an

individual perspective, but also it depends on chronological time t from the system perspective. The latter is the cohort recovery/removal rate $\gamma_c(t)$ satisfying

$$\gamma_c(t) = \frac{\int_0^t i_d(s) f_I(t-s) ds}{\int_0^t i_d(s) \overline{F}_I(t-s) ds}. \tag{5.45}$$

If the mean infectious period $\mu_I < \infty$, the basic reproduction number is $R_0 = \beta \mu_I$. The expression of the effective reproduction number at time t, as modeled by the depletion of the susceptible population over time, is

$$R_t = R_0 \frac{x(t)}{\mu_I \gamma_c(t)}.$$

It reduces to $R_t = R_0 x(t)$ when $\gamma_c(t) = \gamma = \mu_I^{-1}$. The peak prevalence is attained when $x(t)$ depletes to the level $\mu_I \gamma_c(t) R_0^{-1}$, where $\gamma_c(t)$ is implicitly given by (5.45).

5.3.4 *Depletion of Population by Disease Induced Deaths in a Deterministic SIR Model*

For conceptual clarity and simplicity, we restrict our discussions in the SIR models within the deterministic framework to examine the effect of disease induced deaths. We consider a closed population without births or immigration from outside populations.

The initial population size is $m(0) = S(0) + I(0)$. For a disease that induces deaths among those infected, the probability of an infected individual remaining infectious by time x since infection arises from a competing risk framework.

Applying it to the SIR models, the resulting compartment model is shown in Fig. 5.6, where $\gamma_1(x)$ and $\gamma_2(x)$ are the type specific hazard functions, corresponding to "Recovery" and "Death," respectively. We also restrict the discussions to $\gamma_1(x) =$

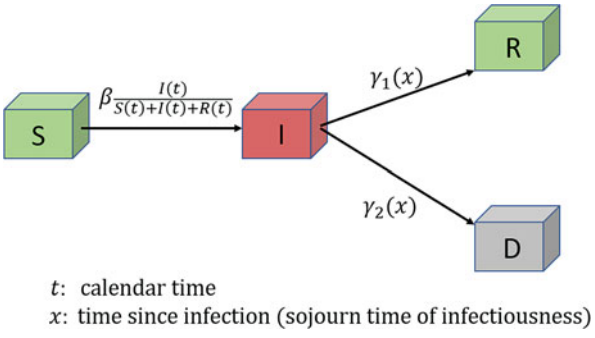


Fig. 5.6 An SIR model with disease induced deaths

γ_1 and $\gamma_2(x) = \gamma_2$, so that the infectious period is exponentially distributed with hazard function $\gamma = \gamma_1 + \gamma_2$, and the mean infectious period $\mu_I = (\gamma_1 + \gamma_2)^{-1}$.

This model introduces a parameter $\varphi = \frac{\gamma_1}{\gamma_1+\gamma_2}$, which is the proportion of infected individuals who survive and recover from the disease. The deterministic model (5.24) is now extended to the following set of ordinary differential equations

$$\begin{cases} \frac{d}{dt} S_d(t) = -\beta \frac{S_d(t) I_d(t)}{m_d(t)} \\ \frac{d}{dt} I_d(t) = \beta \frac{S_d(t) I_d(t)}{m_d(t)} - \gamma I_d(t) \\ \frac{d}{dt} m_d(t) = -(1-\varphi)\gamma I_d(t) \end{cases} \tag{5.46}$$

where $\gamma = \gamma_1 + \gamma_2$, $\varphi = \frac{\gamma_1}{\gamma_1+\gamma_2}$ and $m_d(t) = S_d(t) + I_d(t) + R_d(t)$. The expected number of cumulative infections is $C_d(t) = m(0) - S_d(t)$, and the expected cumulative number of disease induced deaths is $D_d(t) = m(0) - m_d(t)$.

Quantities and Relationships Derived from (5.24) That Are Affected By the Depletion of the Population Size

Key relationships derived from (5.26) and (5.29), such as (5.33), (5.34), (5.30) and (5.32), are all affected by the diminishing population size $m_d(t)$, except for the case $\varphi = 1$.

When $\varphi = 1$, given the initial condition, the basic reproduction number R_0 determines the expected peak prevalence $I_d^{\max}$ and the expected final size $\eta = C_d(\infty)/m$. With additional knowledge of γ, R_0 also determines t_q in (5.34) along with quantities such as $t_{\max .I}$ in (5.35) and $t_{\max .i}$ in (5.37). When $\varphi < 1$, these quantities are not only dependent on R_0, but also on φ.

Example 23 Consider a population with initial size $m(0) = 10{,}000$, along with the initial condition $I(0) = 1$. Let $x_0 = \frac{S_d(0)}{m(0)} = 0.9999$. We choose $\beta = 1.0$ and $\gamma = 1/3$. This gives $R_0 = 3$. We numerically calculate and compare $I_d(t)$ and $C_d(t)$ at $\varphi = 0$, 0.25, 0.5, 0.75, and 1.0. Figure 5.7 shows that, holding β constant over time in model (5.46), depletion of $m_d(t)$ increases the expected instantaneous infection intensity $\beta \frac{S_d(t) I_d(t)}{m_d(t)}$. The case when $\varphi = 1$ corresponds to the simple SIR model with constant $m = m(0)$, which gives the smallest peak incidence (5.33) and the expected final size.

Expressions and Relationships in the Simple SIR Model That Are Not Affected

The expression $R_0 = \beta/\gamma$ is unaltered by recognizing that $\gamma = \gamma_1 + \gamma_2$ in (5.46). The parameter $\varphi = \frac{\gamma_1}{\gamma_1+\gamma_2}$ has no effect.

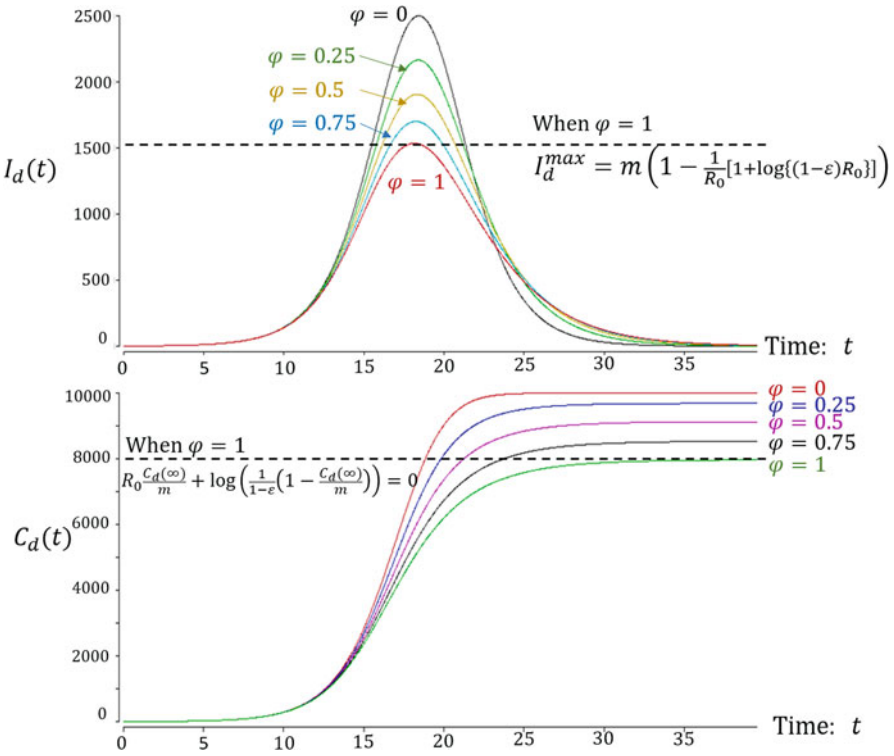


Fig. 5.7 Comparison of $I_d(t)$ and $C_d(t)$ corresponding to Example 23

The relationship (5.31) is also unaltered. However, the parameter φ introduces an additional relationship that $\int_0^\infty I_d(u)du$ is also proportional to $D_d(\infty) = \lim_{t\to\infty} D_d(t)$, the final number of deaths caused by the disease. Together, there is a pair of expressions

$$C_d(\infty) = \gamma \int_0^\infty I_d(u)du, \tag{5.47}$$

$$D_d(\infty) = (1-\varphi)\gamma \int_0^\infty I_d(u)du. \tag{5.48}$$

These can be shown by re-writing (5.46) as

$$\begin{cases} \frac{d}{dt}(S_d(t) + I_d(t)) = -\gamma I_d(t) \\ \frac{d}{dt}m_d(t) = -(1-\varphi)\gamma I_d(t) \end{cases},$$

and $m(0) = S(0) + I(0)$. We refer to Chap. 2 in Brauer et al. (2008) and Diekmann and Heesterbeek (2000) for further readings.

Equations (5.47)–(5.48) with Arbitrary $\gamma_1(x)$ and $\gamma_2(x)$

Equations (5.47)–(5.48) are generalizable when $\gamma_1(x)$ and $\gamma_2(x)$ in Fig. 5.6 are not constants. Viewing $\gamma_1(x)$ and $\gamma_2(x)$ as type-specific hazard functions with respect to arbitrary distributions, the mean infectious period is

$$\mu_I = \int_0^\infty \left(e^{-\int_0^x (\gamma_1(u)+\gamma_2(u))du} \right) dx.$$

Meanwhile, from probability theory associated with independent competing risks, the proportion of infected individuals who survive and recover from the disease φ can be calculated by

$$\varphi = \int_0^\infty \gamma_1(x) e^{-\int_0^x (\gamma_1(u)+\gamma_2(u))du} dx.$$

The expected total infectious person-time is $\int_0^\infty I_d(u)du$. Equations (5.47) and (5.48) are now written as

$$\int_0^\infty I_d(u)du = \mu_I C_d(\infty) = \frac{\mu_I}{1-\varphi} D_d(\infty). \tag{5.49}$$

The Final Size Equation Derived from the Deterministic Model (5.46)

The final size equations (5.30) and (5.32) can be extended to incorporate the parameter φ. The expected total infectious person time $\int_0^\infty I(u)du$ is proportional to both the final number of infected individuals $C_d(\infty)$ and the final number of deaths caused by the disease $D_d(\infty)$, so that $D_d(\infty) = (1-\varphi)\, C_d(\infty)$, as shown by the pair of relationships (5.47)–(5.48). This can be re-written as

$$\log\left(\frac{m_d(\infty)}{m(0)}\right) = \log\left(\varphi + (1-\varphi)\frac{S_d(\infty)}{m(0)}\right). \tag{5.50}$$

On the other hand, the model (5.46) implies

$$\frac{dS_d(t)}{dm_d(t)} = \frac{\beta S_d(t)}{(1-\varphi)\gamma m_d(t)} = \frac{R_0}{1-\varphi}\frac{S_d(t)}{m_d(t)}$$

which gives $\int_0^\infty \frac{dS_d(u)}{S_d(u)}du = \frac{R_0}{1-\varphi}\int_0^\infty \frac{dm_d(u)}{m_d(u)}du$. Therefore

$$\log\left(\frac{S_d(\infty)}{S(0)}\right) = \frac{R_0}{1-\varphi}\log\left(\frac{m_d(\infty)}{m(0)}\right).$$

The initial condition is $x_0 = \frac{S_d(0)}{m(0)}$, and we get

$$\log\left(\frac{m_d(\infty)}{m(0)}\right) = \frac{1-\varphi}{R_0}\log\left(\frac{1}{x_0}\frac{S_d(\infty)}{m(0)}\right). \tag{5.51}$$

Jointly, (5.50) and (5.51) yield a single equation

$$\log\left(\frac{1}{x_0}\frac{S_d(\infty)}{m(0)}\right) = \frac{R_0}{1-\varphi}\log\left(\varphi + (1-\varphi)\frac{S_d(\infty)}{m(0)}\right).$$

Letting $\eta = \frac{C_d(\infty)}{m(0)}$ so that $\frac{S_d(\infty)}{m(0)} = 1-\eta$, the above equation becomes

$$\log\left(\frac{1-\eta}{x_0}\right) = \frac{R_0}{1-\varphi}\log\left(\varphi + (1-\varphi)(1-\eta)\right) \tag{5.52}$$

This is the generalization of (5.32), of which the expected final size η not only depends on R_0 but also depends on φ. By noticing that

$$\lim_{\varphi\to 1}\frac{R_0}{1-\varphi}\log\left(\varphi + (1-\varphi)(1-\eta)\right) = -R_0\eta,$$

it implies that when there is no disease induced mortality, (5.52) returns to (5.32), which is $1-\eta = x_0 e^{-R_0\eta}$.

We further claim that (5.52) is invariant if we relax the conditions $\gamma_1(x) = \gamma_1$ and $\gamma_2(x) = \gamma_2$ in Fig. 5.6. First of all, (5.50) still holds because of (5.49). We leave to the reader to prove (5.51) that leads to the rest of the results.

5.4 The SEIR Models By Adding a Latent Period to the SIR Structure

We restrict our discussion in a closed population without disease induced deaths. The letter E in SEIR stands for "Exposed." Individuals in Compartment E are not only exposed but also infected. However, they are not able to transmit the infection to other susceptible individuals through contacts. It is associated with a duration called the latent period, denoted by T_E.

The stochastic SEIR model as a multivariate Markov process involves $k = 4$ states and 3 independent variables. The size of the state space $\mathfrak{S}$ is $|\mathfrak{S}| = \frac{(m+3)(m+2)(m+1)}{6}$ which increases dramatically with the population size m. For instance, when $m = 10$, $|\mathfrak{S}| = 286$; when $m = 50$, $|\mathfrak{S}| = 23{,}426$; and when $m = 100$, $|\mathfrak{S}| = 176{,}850$. This makes direct analysis of the stochastic SEIR model intractable. For simplicity, we only discuss the deterministic model. We

denote the hazard functions for the latent period T_E and the infectious period T_I as $h_E(x) = \frac{f_E(x)}{\overline{F}_E(x)}$ and $h_I(x) = \frac{f_I(x)}{\overline{F}_I(x)}$, respectively.

5.4.1 *Deterministic SEIR Model with Exponentially Distributed Latent and Infectious Periods*

The deterministic model (5.24) is extended to

$$\begin{cases} \frac{d}{dt}S_d(t) = -\beta\frac{S_d(t)I_d(t)}{m} \\ \frac{d}{dt}E_d(t) = \beta\frac{S_d(t)I_d(t)}{m} - \alpha E_d(t) \\ \frac{d}{dt}I_d(t) = \alpha E_d(t) - \gamma I_d(t) \end{cases} \tag{5.53}$$

which implies $\frac{d}{dt}R_d(t) = \gamma I_d(t)$. Like in (5.24), m is an innocent parameter and all the state variables can be scaled by m and expressed as proportions: $x(t) = \frac{S_d(t)}{m}$, $\epsilon(t) = \frac{E_d(t)}{m}$, $y(t) = \frac{I_d(t)}{m}$ and $z(t) = 1-x(t)-\epsilon(t)-y(t)$. Let $\omega(t) = \epsilon(t)+y(t) = \frac{P_d(t)}{m}$. We can also re-scale the time $\tau = \gamma t$ so that

$$\begin{cases} \frac{d}{d\tau}x(\tau) = -\frac{\beta}{\gamma}x(\tau)y(\tau), \\ \frac{d}{d\tau}\epsilon(\tau) = \frac{\beta}{\gamma}x(\tau)y(\tau) - \frac{\alpha}{\gamma}\epsilon(\tau) \\ \frac{d}{d\tau}y(\tau) = \frac{\alpha}{\gamma}\epsilon(\tau) - y(\tau) \end{cases}. \tag{5.54}$$

It implies $\frac{d}{d\tau}z(\tau) = y(\tau)$. The essential parameters are $R_0 = \beta/\gamma$ and α/γ. Since $\frac{d}{dt}P_d(t) = \frac{d}{dt}E_d(t) + \frac{d}{dt}I_d(t) = \beta\frac{S_d(t)I_d(t)}{m} - \gamma I_d(t)$, it immediately turns out that

$$S_d(t) + P_d(t) - \frac{m}{R_0}\log S_d(t) = S_d(0) + P_d(0) - \frac{m}{R_0}\log S_d(0). \tag{5.55}$$

Given the initial condition $S_d(0) + P_d(0) = m$, it can be written as

$$1 - x(t) - \epsilon(t) - y(t) + \frac{1}{R_0}\log\frac{x(t)}{x_0} = 0. \tag{5.56}$$

These preserved relationships are almost the same as (5.26) and (5.28) except for $P_d(t) = E_d(t) + I_d(t)$.

The relationships (5.55)–(5.56) also lead to the final size equations (5.30) and (5.32). The final size is $C_d(\infty) = m - S_d(\infty)$. Because $\frac{d}{dt}(S_d(t) + E_d(t) + I_d(t)) = -\gamma I_d(t)$ and $S_d(t) + E_d(t) + I_d(t) + R_d(t) = m$, one also gets $\gamma\int_0^\infty I_d(t)dt = C_d(\infty)$. The final size equations and the relationship between the final size and the value of the epidemic $\int_0^\infty I_d(t)dt$ given by (5.31) are not affected by the added latent period.

The peak value $P_d(t)$, the peak prevalence, is attained at $t_{\text{max.P}}$ such that $x(t_{\text{max.P}}) = \frac{S_d(t_{\text{max.P}})}{m} = \frac{1}{R_0}$. From (5.55)

$$\omega(t_{\text{max.P}}) = \frac{P_d^{\max}}{m} = 1 - \frac{1}{R_0}\left[1 + \log(R_0 x_0)\right]. \tag{5.57}$$

It implies that the number of recovered individuals at the time when $P_d^{\max}$ is attained is $z(t_{\text{max.P}}) = \frac{R_d(t_{\text{max.P}})}{m} = \frac{1}{R_0}\log(R_0 x_0)$. Note that (5.57) is identical to (5.33) provided that $P_d(t)$ includes those infected but still in their latent periods. It implies that, comparing the SIR model (5.24) and the SEIR model (5.53) with the same initial condition $x_0 = \frac{S_d(0)}{m}$ and the same R_0, the peak prevalence value remains the same.

Let $\tau = \gamma t$. Since $z(\tau) = 1 - x(\tau) - \epsilon(\tau) - y(\tau)$, (5.56) can be still written as $x(\tau) = x_0 \exp(-R_0 z(\tau))$. However, in the SEIR model,

$$\begin{aligned}\frac{dz(\tau)}{d\tau} &= y(\tau) = 1 - z(\tau) - x(\tau) - \epsilon(\tau)\\ &= 1 - z(\tau) - x_0 \exp(-R_0 z(\tau)) - \epsilon(\tau)\\ &< 1 - z(\tau) - x_0 \exp(-R_0 z(\tau))\end{aligned} \tag{5.58}$$

where the additional term $\epsilon(t)$ makes it difficult to make a simple extension of (5.29). Consequently, for the SEIR model, there is no explicit formula (except for numerical illustration) for quantities such as the time at the peak prevalence. However, the inequality (5.58) implies that, with an added latent period, $z(\tau)$ grows slower than that in the SIR model and takes longer to reach $z(\tau_{\text{max.P}}) = \frac{1}{R_0}\log(R_0 x_0)$, under the same R_0 and initial condition x_0. In other words, $\tau_{\text{max.P}}^{\text{SEIR}} > \tau_{\text{max.I}}^{\text{SIR}}$, where $\tau_{\text{max.I}}^{\text{SIR}} = \gamma t_{\text{max.I}}^{\text{SIR}}$ is given by (5.35), representing the time when $I_d(t)$ in (5.24) reaches the maximum value.

Example 24 Consider a population with $m = 10{,}000$, $x_0 = 0.9999$, $\beta = 0.75$, and $\gamma = 1/3$. Thus $R_0 = 2.25$. Figure 5.8 compares the SIR model with the above parameters with an SEIR model with an added exponentially distributed latent period with mean value $\alpha^{-1} = 3$. Starting with the initial conditions $I_d(0) = 1$ and $S_d(0) = 9999$ in both models, the number of susceptible individuals in the SEIR model, $S_d^{SEIR}(t)$, decreases more slowly than $S_d^{SIR}(t)$, corresponding to the SIR model. Consequently, $S_d^{SEIR}(t)$ arrives at $m/R_0 = 4444.4$ later than $S_d^{SIR}(t)$. This implies that the times of the peak prevalence for the two models satisfy $t_{\text{max.P}}^{\text{SEIR}} > t_{\text{max.P}}^{\text{SIR}}$. However, the following quantities remain the same in both models:

1. the final size $C_d(\infty) = R_d(\infty) = 8534$ and $S_d(\infty) = 1466$, corresponding to the final size equation $1 - \eta = 0.9999e^{-2.25\eta}$ in $(0, 1]$ which gives $\eta = 0.8534$;
2. the relationship between the number of susceptible individuals and recovered individuals: $S_d(t) = 9999e^{-0.000225R_d(t)}$;
3. the value of the peak prevalence: 1951.9.

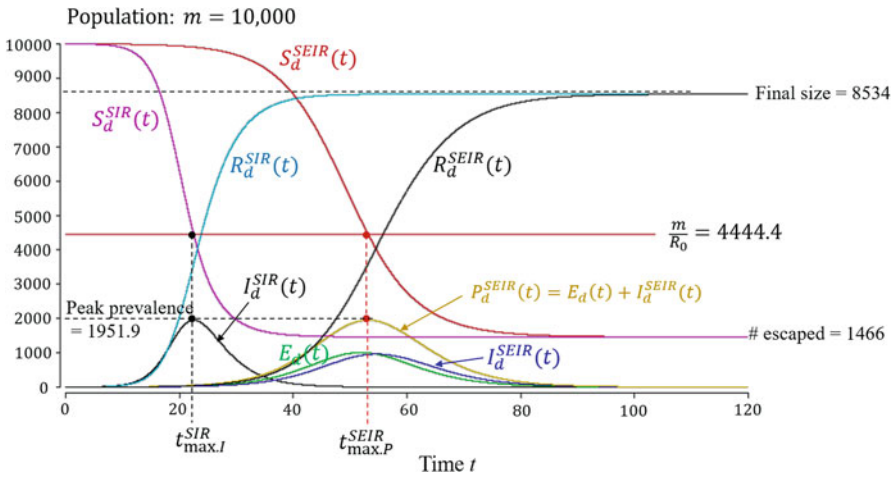


Fig. 5.8 Compare an SIR model and an SEIR model with the same β, γ and initial conditions

5.4.2 *Deterministic SEIR Model with Erlang Distributed Latent and Infectious Periods*

For the SEIR model, we use μ_E and μ_I for the mean latent and the mean infectious periods and κ_E and κ_I for the corresponding shape parameters. We include an Erlang distributed latent period into (5.43). The model (5.53) corresponds to $\kappa_E = \kappa_I = 1$. When $\kappa_E = 2, 3, \ldots$ the equation $\frac{d}{dt}E_d(t) = \beta\frac{S_d(t)I_d(t)}{n} - \alpha E_d(t)$ in (5.53) is expanded into κ_E separate equations

$$\begin{aligned}
\frac{d}{dt}E_{d1}(t) &= \beta\frac{S_d(t)I_d(t)}{m} - \frac{\kappa_E}{\mu_E}E_{d1}(t),\\
\frac{d}{dt}E_{dj}(t) &= \frac{\kappa_E}{\mu_E}E_{dj-1}(t) - \frac{\kappa_E}{\mu_E}E_{dj}(t), \quad j = 2, 3, \cdots, \kappa_E.
\end{aligned} \tag{5.59}$$

Similarly, when $\kappa_I = 2, 3, \ldots$ the equation $\frac{d}{dt}I_d(t) = \alpha E_d(t) - \gamma I_d(t)$ in (5.53) is expanded into κ_I separate equations

$$\begin{aligned}
\frac{d}{dt}I_{d1}(t) &= \frac{\kappa_E}{\mu_E}E_{d\kappa_E}(t) - \frac{\kappa_I}{\mu_I}I_{d1}(t),\\
\frac{d}{dt}I_{dj}(t) &= \frac{\kappa_I}{\mu_I}I_{dj-1}(t) - \frac{\kappa_I}{\mu_I}I_{dj}(t), \quad j = 2, 3, \cdots, \kappa_I.
\end{aligned} \tag{5.60}$$

With these ordinary differential equations, one can numerically calculate the expected cumulative infections $C_d(t) = n - S_d(t)$, the expected prevalence $Prev(t) = \sum_{j=1}^{\kappa_E} E_{dj}(t) + \sum_{j=1}^{\kappa_I} I_{dj}(t)$ and other derived quantities.

5.4.3 *Generally Distributed Latent and Infectious Periods*

For generally distributed latent and infectious periods, let $f_E(x)$ and $h_E(x)$ represent the p.d.f. and the hazard function of the latent periods; and $f_I(x)$ and $h_I(x)$ represent the p.d.f. and the hazard function of the infectious periods, the deterministic SEIR model for non-exponentially distributed latent and infectious periods is a system of integro-differential equations

$$\begin{cases} \frac{d}{dt}S_d(t) &= -\beta\frac{S_d(t)I_d(t)}{m} \\ \frac{d}{dt}E_d(t) &= i_d(t) - \int_0^t i_d(s)f_E(t-s)ds \\ &= i_d(t) - \int_0^t i_d(s)h_E(t-s)\overline{F}_E(t-s)ds \\ \frac{d}{dt}I_d(t) &= i_{d1}(t) - \int_0^t i_{d1}(s)f_I(t-s)ds \\ &= i_{d1}(t) - \int_0^t i_{d1}(s)h_I(t-s)\overline{F}_I(t-s)ds \end{cases}, \tag{5.61}$$

where $i_d(t) = \beta\frac{S_d(t)I_d(t)}{n}$; $i_{d1}(t) = \int_0^t i_d(s)f_E(t-s)ds$ is the expected (instantaneous) number of individuals making a transition from being latent to being infectious; and $\int_0^t i_{d1}(s)f_I(t-s)ds$ is the expected (instantaneous) number of infectious individuals being removed, at time t. This is the extension of (5.44). Under the condition $S_d(t) + E_d(t) + I_d(t) + R_d(t) = m$, we also have

$$\frac{d}{dt}R_d(t) = \int_0^t i_{d1}(s)f_I(t-s)ds = \int_0^t i_{d1}(s)h_I(t-s)\overline{F}_I(t-s)ds.$$

5.5 Endemic Equilibrium When There Is Replacement of the Susceptible Population

One of the mechanisms of susceptible replacement is through the loss of immunity, either immediately after recovery or after a duration of immunity, while the population itself is closed. Adding this to the SEIR model leads to the SEIRS model, with the SIS, SIRS, SEIS models as special cases. Another mechanism is the replacement of the population, where individuals enter and leave the population, and the replacement of the susceptible population is through such a mechanism. These two mechanisms may be combined such that the susceptible population is replaced through both the loss of immunity of recovered individuals and the in-flow of susceptible individuals from outside the population.

We consider a constant population so that the in-flow and the out-flow of the population is balanced. For mathematical simplicity, we assume that all individuals coming into the population are susceptible, i.e. no importation of infected individuals. Natural births and deaths are typical scenarios for such population replacement. In this section, we take a broader perspective so that the in-flow and the out-flow of the population are not limited to natural births and deaths.

Hethcote and van den Driessche (1991) and Li et al. (1999) considered deterministic models of the SEIRS type including natural deaths and showed that, when there is replacement of the susceptible population with both the loss of immunity and through births, under the condition that the overall population remains constant, stability analyses of the dynamic system show that when $R_0 > 1$, there exists an endemic equilibria, in addition to the disease-free equilibria.

There is a rich library of literature regarding such analyses. Chapters 5 and 6 of Brauer (2008) and the collection of papers in Castillo-Chávez et al. (2000) provide nice coverage of these topics, along with many mathematical expressions that will appear in this section.

This section takes a more intuitive approach. Instead of stability analyses of specific dynamic systems, we adopt an independent competing risk approach and use Laplace transforms to present the prevalence of individuals in each class of the SEIRS model, in which the population is constant and the rate of exiting population via movement or natural mortality is constant. The focus is on the asymptotic endemic equilibrium levels for diseases with generally distributed latent and infectious periods and generally distributed durations of immunity after recovery, provided that the system is at endemic equilibrium.

5.5.1 SEIRS Models Without Deaths

We use the term *prevalence* defined by the proportion of individuals in each class of the SEIRS model, denoted by $\Pi(t) = (x(t), \epsilon(t), y(t), z(t))$, using the notations in (5.54), where $\epsilon(t)$ is the proportion of individuals who are infected but not yet infectious (latent); $y(t)$ is the proportion of individuals who are infectious; $z(t)$ is the proportion of individuals who are recovered and immune; $x(t) = 1 - \epsilon(t) - y(t) - z(t)$ is the proportion who are susceptible. We consider the limits $\Pi(\infty) = (x(\infty), \epsilon(\infty), y(\infty), z(\infty))$ as $t \to \infty$. The endemic equilibria is $\Pi(\infty)$ and $y(\infty) > 0$.

We have seen that, in the deterministic SIS model, if $R_0 > 1$, $x(\infty) = \lim_{t\to\infty} S_d(t)/m = 1/R_0$ and $y(\infty) = \lim_{t\to\infty} I_d(t)/m = 1 - 1/R_0$. The population size m is an innocent parameter that can be eliminated by re-scaling.

Well-recognized in epidemiology, a linear relationship exists between the prevalence and the incidence when the system is under equilibrium:

$$\text{prevalence} = \text{incidence} \times \text{average duration}. \tag{5.62}$$

Incidences are instantaneous rates of event occurrences. Suppose that there are two events, an initial event with an incidence function $\lambda(t)$ and a subsequent event with an incidence function $a(t)$. All initial events lead to subsequent events through a random duration X with p.d.f. $f(x)$ and survival function $\overline{F}(x)$. The following pair of convolutions holds

$$a(t) = \int_0^t \lambda(s) f(t-s) ds,$$
$$A(t) = \int_0^t \lambda(s) \overline{F}(t-s) ds$$

where $A(t)$ is the prevalence of individuals who have experienced the initial event but not yet the subsequent event. If $\lim_{t\to\infty} \lambda(t) = \lambda_\infty = \text{constant}$, then $a(\infty) = \lambda_\infty \int_0^\infty f(x)dx = \lambda_\infty$. Meanwhile, $A(\infty) = \lambda_\infty \int_0^\infty \overline{F}(x)dx = \lambda_\infty E[X]$, which is (5.62).

In the current context, the incidence of new infections is modeled by a bilinear relationship $\lambda(t) = \beta x(t) y(t)$. Assuming there is no change in the environment or behavior and there is no control measure throughout the epidemic, under endemic equilibrium, $\lambda(t) \to \lambda_\infty = \beta x(\infty) y(\infty) > 0$. In a closed population without individuals entering or leaving, all infected individuals enter the E-compartment through a latent period and progress to the I-compartment, and then progress to the R-compartment. Therefore the incidences of entering each of these compartments, at equilibrium, equal λ_∞. Meanwhile, $\epsilon(\infty)$, $y(\infty)$, and $z(\infty)$ are prevalences of individuals in each of the compartments, and the average durations for staying in each of these compartments are denoted by μ_E, μ_I, and μ_R, respectively.

Applying (5.62), one gets a system of equations

$$\begin{cases} x(\infty) = 1 - \epsilon(\infty) - y(\infty) - z(\infty) \\ \epsilon(\infty) = \lambda_\infty \mu_E \\ y(\infty) = \lambda_\infty \mu_I \\ z(\infty) = \lambda_\infty \mu_R \end{cases} \tag{5.63}$$

with two sets of solutions: the disease-free equilibrium $[x(\infty) = 1, \epsilon(\infty) = y(\infty) = z(\infty) = 0]$ and the endemic equilibrium (when $R_0 = \beta\mu_I > 1$)

$$\begin{aligned} x(\infty) &= \frac{1}{\beta\mu_I} = \frac{1}{R_0}, \\ \epsilon(\infty) &= \frac{\mu_E}{\mu_E + \mu_I + \mu_R} [1 - x(\infty)], \\ y(\infty) &= \frac{\mu_I}{\mu_E + \mu_I + \mu_R} [1 - x(\infty)], \\ z(\infty) &= \frac{\mu_R}{\mu_E + \mu_I + \mu_R} [1 - x(\infty)]. \end{aligned} \tag{5.64}$$

These are valid for arbitrarily distributed latent period, infectious period, and duration of temporary immunity, with mean values μ_E, μ_I, and μ_R, respectively.

The SIS model corresponds to $\mu_E = \mu_R = 0$. In this case, $y(\infty) = 1 - 1/R_0$ which is the limit of the (5.16) as $\tau \to \infty$.

The SIRS model corresponds to $\mu_E = 0$. The SEIS model corresponds to $\mu_R = 0$. Note that $x(\infty)$ is invariant with respect to the model structure, whereas $\lambda_\infty =$

$\beta x(\infty)y(\infty)$ depends on model structure. The SIS model gives the largest incidence at endemic equilibrium.

The Special Cases Represented By the Ordinary Differential Equations

When the durations in each of the compartments, E, I, and R are exponentially distributed, the SEIRS model is represented by the system of differential equations:

$$\begin{cases} \frac{d}{dt}S_d(t) = -\beta\frac{S_d(t)I_d(t)}{m} + \delta R_d(t) \\ \frac{d}{dt}E_d(t) = \beta\frac{S_d(t)I_d(t)}{m} - \alpha E_d(t) \\ \frac{d}{dt}I_d(t) = \alpha E_d(t) - \gamma I_d(t) \\ \frac{d}{dt}R_d(t) = \gamma I_d(t) - \delta R_d(t) \end{cases}$$

where $\alpha = \mu_E^{-1}$, $\gamma = \mu_I^{-1}$ and $\delta = \mu_R^{-1}$. Letting $x(\infty) = \lim_{t\to\infty} S_d(t)/m$, $\epsilon(\infty) = \lim_{t\to\infty} E_d(t)/m$, $y(\infty) = \lim_{t\to\infty} I_d(t)/m$ and $z(\infty) = \lim_{t\to\infty} R_d(t)/m$, (5.64) becomes

$$\begin{aligned} x(\infty) &= \frac{\gamma}{\beta}, \\ y(\infty) &= \frac{\alpha\delta\,(\beta-\gamma)}{\beta\,(\alpha\gamma+\alpha\delta+\gamma\delta)}, \\ \epsilon(\infty) &= \frac{\gamma}{\alpha}y(\infty),\ z(\infty) = \frac{\gamma}{\delta}y(\infty). \end{aligned} \tag{5.65}$$

The SEIS model corresponds to the case $\delta \to \infty$ and

$$x(\infty) = \frac{\gamma}{\beta},\ y(\infty) = \frac{\alpha\,(\beta-\gamma)}{\beta\,(\alpha+\gamma)},\ \epsilon(\infty) = \frac{\gamma}{\alpha}y(\infty).$$

The SIRS model corresponds to the case $\alpha \to \infty$ and

$$x(\infty) = \frac{\gamma}{\beta},\ y(\infty) = \frac{\delta\,(\beta-\gamma)}{\beta\,(\gamma+\delta)},\ z(\infty) = \frac{\gamma}{\delta}y(\infty).$$

The SIS model corresponds to the case $\alpha \to \infty$ and $\delta \to \infty$, thus $x(\infty) = \gamma/\beta$ and $y(\infty) = 1 - \gamma/\beta$.

5.5.2 SEIRS Models in a Constant Population Where the In-Flow and Out-Flow of Individuals Are Balanced

Using the same notations as in (5.61), the latent periods are generally distributed with p.d.f. $f_E(x)$, survival function $\overline{F}_E(x)$, the hazard function $h_E(x)$, and mean

value $\mu_E = \int_0^\infty \overline{F}_E(x)dx$; the infectious period is generally distributed with p.d.f. $f_I(x)$, survival function $\overline{F}_I(x)$, the hazard function $h_I(x)$, and mean value $\mu_I = \int_0^\infty \overline{F}_I(x)dx$.

We assume that there is an out-flow of individuals from all the compartments with constant rate $\omega > 0$.

When $\omega > 0$, only a fraction of exposed individuals will progress into the infectious stage. This fraction is calculated by the Laplace transform

$$\varphi_1 = \int_0^\infty e^{-\omega x} f_E(x)dx = L[f_E](\omega).$$

The "effective" mean latent period (among those who progress to the infectious stage) is

$$\mu_E^* = \int_0^\infty e^{-\omega x} \overline{F}_E(x)dx = L[\overline{F}_E](\omega).$$

Both of the above expressions are made under the assumption that each individual in the latent period (the E compartment) either leave the population or progress to the next compartment through an independent competing risk framework. At equilibrium, the incidence rate of new infections is $\lambda_\infty = \beta x(\infty)y(\infty)$ and the incidence rate of onset of infectiousness is $\varphi_1\lambda_\infty$. If there is no latent period, we assume that the latent period is degenerated to a single point: $T_E = 0$. In such case, $L[f_E](\omega) = 1$ and $L[\overline{F}_E](\omega) = 0$.

Continuing, only a fraction of infectious individuals will recover. This fraction is calculated by the Laplace transform

$$\varphi_2 = \int_0^\infty e^{-\omega x} f_I(x)dx = L[f_I](\omega).$$

The "effective" mean latent period (among those who recover) is

$$\mu_I^* = \int_0^\infty e^{-\omega x} \overline{F}_I(x)dx = L[\overline{F}_I](\omega).$$

At equilibrium, the incidence rate of onset of infectiousness is $\varphi_1\lambda_\infty$, and the incidence rate of recovery is $\varphi_2\varphi_1\lambda_\infty$.

Assuming there is no change of environment, no change of behavior and no intervention throughout the epidemic, the basic reproduction number can be recovered from the equilibrium state and expressed by

$$R_0 = \beta^*\mu_I^* = \beta L[f_E](\omega)L[\overline{F}_I](\omega) \tag{5.66}$$

where $\beta^* = \beta\varphi_1 = \beta L[f_E](\omega)$.

We further assume that there is a random duration among recovered individuals to lose immunity and become susceptible again. This duration has p.d.f. $f_R(x)$,

survival function $\overline{F}_R(x)$, the hazard function $h_R(x)$, and mean value $\mu_R = \int_0^\infty \overline{F}_R(x)dx$. Therefore, the fraction of recovered individuals who will become susceptible again (before leaving the population) is

$$\varphi_3 = \int_0^\infty e^{-\omega x} f_R(x)dx = L[f_R](\omega).$$

The "effective" mean duration of immunity (among those who become susceptible again) is

$$\mu_R^* = \int_0^\infty e^{-\omega x} \overline{F}_R(x)dx = L[\overline{F}_R](\omega).$$

In models where recovered individuals become susceptible immediately, $L[\overline{F}_R](\omega) = 0$. On the other hand, in models where recovered individuals will remain immune indefinitely, $L[f_R](\omega) = 1 - \omega L[\overline{F}_R](\omega) = 0$, which gives, $L[\overline{F}_R](\omega) = 1/\omega$.

Equations (5.63) are revised as

$$\begin{cases} x(\infty) = 1 - \epsilon(\infty) - y(\infty) - z(\infty) \\ \epsilon(\infty) = \lambda_\infty \mu_E^* = \beta x(\infty) y(\infty) L[\overline{F}_E](\omega) \\ y(\infty) = \varphi_1 \lambda_\infty \mu_I^* = \beta x(\infty) y(\infty) L[f_E](\omega) L[\overline{F}_I](\omega) \\ z(\infty) = \varphi_2 \varphi_1 \lambda_\infty \mu_R = \beta x(\infty) y(\infty) L[f_I](\omega) L[f_E](\omega) L[\overline{F}_R](\omega) \end{cases}, \tag{5.67}$$

which give

$$\begin{aligned} \epsilon(\infty) &= \frac{L[\overline{F}_E](\omega)}{L[f_E](\omega) L[\overline{F}_I](\omega)} y(\infty) = \frac{\mu_E^*}{\varphi_1 \mu_I^*} y(\infty) \\ z(\infty) &= \frac{L[f_I](\omega) L[\overline{F}_R](\omega)}{L[\overline{F}_I](\omega)} y(\infty) = \frac{\varphi_2 \mu_R^*}{\mu_I^*} y(\infty). \end{aligned} \tag{5.68}$$

The following relationships also hold: $L[f_E](\omega) = 1 - \omega L[\overline{F}_E](\omega)$ and $L[f_I](\omega) = 1 - \omega L[\overline{F}_I](\omega)$. The equations (5.67) have a disease-free equilibrium

$$[x(\infty) = 1,\ \epsilon(\infty) = y(\infty) = z(\infty) = 0].$$

When $R_0 = \beta L[f_E](\omega) L[\overline{F}_I](\omega) > 1$, there is also an endemic equilibrium solution

$$x(\infty) = \frac{1}{\beta L[f_E](\omega) L[\overline{F}_I](\omega)} = \frac{1}{\beta \varphi_1 \mu_I^*},$$

$$y(\infty) = \frac{L[f_E](\omega) L[\overline{F}_I](\omega)}{L[\overline{F}_E](\omega) + L[f_E](\omega) \left[L[\overline{F}_I](\omega) + L[f_I](\omega) L[\overline{F}_R](\omega) \right]} [1 - x(\infty)]$$

$$= \frac{\varphi_1 \mu_I^*}{\mu_E^* + \varphi_1 \left(\mu_I^* + \varphi_2 \mu_R^*\right)} [1 - x(\infty)]$$
$$= \frac{\beta \varphi_1 \mu_I^* - 1}{\beta \left[\mu_E^* + \varphi_1 \left(\mu_I^* + \varphi_2 \mu_R^*\right)\right]},$$

and $\epsilon(\infty)$, $z(\infty)$ are calculated through (5.68). The following prevalence gives the disease burden at endemic equilibrium scaled by the population size,

$$\epsilon(\infty) + y(\infty) = \frac{\mu_E^* + \varphi_1 \mu_I^*}{\mu_E^* + \varphi_1 \mu_I^* + \varphi_1 \varphi_2 \mu_R^*} [1 - x(\infty)].$$

Special Cases with Discussions

Only with Loss of Immunity in a Closed Population Without In-Flow and Out-Flow Letting $\omega = 0$, $L[\overline{F}_E](0) = \mu_E$, $L[\overline{F}_I](0) = \mu_I$ and $L[\overline{F}_R](0) = \mu_R$ while $L[f_E](0) = L[f_I](0) = 1$, then

$$x(\infty) = \frac{1}{\beta \mu_I}, \; y(\infty) = \frac{\mu_I}{\mu_E + \mu_I + \mu_R} [1 - x(\infty)],$$
$$\epsilon(\infty) = \frac{\mu_E}{\mu_I} y(\infty), \; z(\infty) = \frac{\mu_R}{\mu_I} y(\infty).$$

which return to (5.64). The values $\epsilon(\infty)$, $y(\infty)$, and $z(\infty)$ are proportions of $[1 - x(\infty)]$. These proportions are the relative average time of individuals spent in each compartment, out of the total time $\mu_E + \mu_I + \mu_R$. Without natural deaths, these values are determined only by the average durations μ_E, μ_I, and μ_R, regardless of the distributions of these durations.

Only with In-Flow and Out-Flow in a Constant Population Without Loss of Immunity When $\omega > 0$, the distributions of the latent periods, the infectious periods, and the durations of immunity after recovery, all play significant roles. We first consider the case without loss of immunity and examine the roles of the distributions of the latent and the infectious periods. First of all, $R_0 = \beta L[f_E](\omega) L[\overline{F}_I](\omega)$. We recall the discussion on variability according to the convex order, Definition 6 in Chap. 2. Holding the mean latent period μ_E and the mean infectious period μ_I constant, at any given $\omega > 0$, the more variable the latent period, the larger the value of $\varphi_1 = L[f_E](\omega)$. In other words, large variability of the latent period increases the probability of individuals progressing to the infectious stage before they leave the population and increases the value of R_0 when β and $L[\overline{F}_I](\omega)$ remain the same. On the other hand, since $L[\overline{F}_I](\omega) = \frac{1}{\omega} [1 - L[f_I](\omega)]$, the more variable the infectious periods, the smaller the value of $\mu_I^* = L[\overline{F}_I](\omega)$. Therefore, large variability of the infectious periods decreases the effective duration of the infection period and decreases the value of R_0 when beta and $L[f_E](\omega)$ remain the same.

When recovered individuals remain immune indefinitely, $\varphi_3 = L[f_R](\omega) = 0$. This gives $\mu_R^* = L[\overline{F}_R](\omega) = 1/\omega$, that is, the average duration of staying in the population. Since $L[f_E](\omega) = 1 - \omega L[\overline{F}_E](\omega)$ and $L[f_I](\omega) = 1 - \omega L[\overline{F}_I](\omega)$, we get

$$\begin{aligned} x(\infty) &= \frac{1}{\beta L[f_E](\omega) L[\overline{F}_I](\omega)} = \frac{1}{\beta \varphi_1 \mu_I^*}, \\ y(\infty) &= L[f_E](\omega)\,[1 - L[f_I](\omega)]\,[1 - x(\infty)] \\ &= \varphi_1(1-\varphi_2)\,[1 - x(\infty)]\,, \\ \epsilon(\infty) &= (1-\varphi_1)\,[1 - x(\infty)]\,, \\ z(\infty) &= \varphi_1\varphi_2\,[1 - x(\infty)]\,. \end{aligned}$$

It implies that variabilities of the latent periods and the infectious periods affect the value of $1 - x(\infty)$. Given the value of $1 - x(\infty)$, larger variability of the latent periods assigns larger value of φ_1 and hence larger proportion of $1 - x(\infty)$ into $y(\infty)$; whereas, larger variability of the infectious periods assigns larger value of φ_2 and hence smaller proportion of $1 - x(\infty)$ into $y(\infty)$. If there is no latent period, $L[f_E](\omega) = 1$ and $L[\overline{F}_E](\omega) = 0$, the above expressions are reduced to

$$\begin{aligned} x(\infty) &= \frac{1}{\beta L[\overline{F}_I](\omega)} = \frac{1}{\beta \mu_I^*}, \\ y(\infty) &= (1-\varphi_2)\,[1 - x(\infty)]\,. \end{aligned}$$

Larger variability of the infectious period assigns smaller proportions of $1 - x(\infty)$ into $y(\infty)$.

SIRS with $\omega > 0$ The absence of the latent period is represented by the Laplace transforms $L[f_E](\omega) = 1$ and $L[\overline{F}_E](\omega) = 0$. This gives

$$\begin{aligned} x(\infty) &= \frac{1}{\beta L[\overline{F}_I](\omega)} = \frac{1}{\beta \mu_I^*}, \\ y(\infty) &= \frac{L[\overline{F}_I](\omega)}{L[\overline{F}_I](\omega) + L[f_I](\omega) L[\overline{F}_R](\omega)}\,[1 - x(\infty)] = \frac{\mu_I^*}{\mu_I^* + \varphi_2 \mu_R^*}\,[1 - x(\infty)]\,, \\ z(\infty) &= \frac{\varphi_2 \mu_R^*}{\mu_I^* + \varphi_2 \mu_R^*}\,[1 - x(\infty)] = \frac{\varphi_2 \mu_R^*}{\mu_I^*} y(\infty). \end{aligned}$$

SEIS with $\omega > 0$ If recovered individuals become susceptible immediately, $L[\overline{F}_R](\omega) = 0$.

$$
\begin{aligned}
x(\infty) &= \frac{1}{\beta L[f_E](\omega) L[\overline{F}_I](\omega)} = \frac{1}{\beta \varphi_1 \mu_I^*}, \\
y(\infty) &= \frac{L[f_E](\omega) L[\overline{F}_I](\omega)}{L[\overline{F}_E](\omega) + L[f_E](\omega) L[\overline{F}_I](\omega)} [1 - x(\infty)] \\
&= \frac{\varphi_1 \mu_I^*}{\mu_E^* + \varphi_1 \mu_I^*} [1 - x(\infty)], \\
\epsilon(\infty) &= \frac{\mu_E^*}{\mu_E^* + \varphi_1 \mu_I^*} [1 - x(\infty)] = \frac{\mu_E^*}{\varphi_1 \mu_I^*} y(\infty).
\end{aligned}
$$

SIS with $\omega > 0$ Letting the Laplace transforms $L[f_E](\omega) = 1$, $L[\overline{F}_E](\omega) = 0$ and $L[\overline{F}_R](\omega) = 0$,

$$
x(\infty) = \frac{1}{\beta L[\overline{F}_I](\omega)} = \frac{1}{\beta \mu_I^*}, \; y(\infty) = 1 - \frac{1}{\beta \mu_I^*}.
$$

Exponentially Distributed Durations Corresponding to Ordinary Differential Equations The ordinary differential equations for the SEIRS model are

$$
\begin{cases}
\frac{d}{dt} S_d(t) = -\beta \frac{S_d(t) I_d(t)}{m} + \omega [m - S_d(t)] + \delta R_d(t) \\
\frac{d}{dt} E_d(t) = \beta \frac{S_d(t) I_d(t)}{m} - (\alpha + \omega) E_d(t) \\
\frac{d}{dt} I_d(t) = \alpha E_d(t) - (\gamma + \omega) I_d(t) \\
\frac{d}{dt} R_d(t) = \gamma I_d(t) - (\delta + \omega) R_d(t)
\end{cases}, \tag{5.69}
$$

as expressed in various papers in the literature, such as Hethcote and van den Driessche (1991), Li et al. (1999), among others. The constant rates (α, γ, δ) correspond to exponentially distributed latent periods, infectious periods, and durations of recovered individuals who remain immune. The Laplace transforms as represented in (5.67) are:

$$
\begin{aligned}
\varphi_1 &= L[f_E](\omega) = \frac{\alpha}{\alpha + \omega}, \; \mu_E^* = L[\overline{F}_E](\omega) = \frac{1}{\alpha + \omega}, \\
\varphi_2 &= L[f_I](\omega) = \frac{\gamma}{\gamma + \omega}, \; \mu_I^* = L[\overline{F}_I](\omega) = \frac{1}{\gamma + \omega}, \\
\mu_R^* &= L[\overline{F}_R](\omega) = \frac{1}{\delta + \omega}.
\end{aligned}
$$

The representation of R_0 is

$$
R_0 = \beta L[f_E](\omega) L[\overline{F}_I](\omega) = \frac{\beta \alpha}{(\alpha + \omega)(\gamma + \omega)}. \tag{5.70}
$$

The prevalences at endemic equilibrium are

$$\begin{aligned} x(\infty) &= \frac{(\alpha+\omega)(\gamma+\omega)}{\beta\alpha}, \\ y(\infty) &= \frac{\alpha(\delta+\omega)}{(\gamma+\omega)(\delta+\omega)+\alpha(\gamma+\delta+\omega)}[1-x(\infty)] \\ &= \frac{(\delta+\omega)}{\beta}\frac{\alpha(\beta-\gamma-\omega)-\omega(\gamma+\omega)}{\alpha(\gamma+\delta+\omega)+(\delta+\omega)(\gamma+\omega)}, \\ \epsilon(\infty) &= \frac{\gamma+\omega}{\alpha}y(\infty),\ z(\infty) = \frac{\gamma}{\delta+\omega}y(\infty). \end{aligned} \tag{5.71}$$

1. SEIRS in a closed population ($\omega = 0$): the above expressions return to (5.65).
2. SEIR with $\omega > 0$ but recovered individuals have permanent immunity $\delta = 0$:

$$\begin{aligned} x(\infty) &= \frac{(\alpha+\omega)(\gamma+\omega)}{\beta\alpha}, \\ y(\infty) &= \frac{\alpha\omega}{(\gamma+\omega)(\alpha+\omega)}[1-x(\infty)] \\ &= \frac{\alpha\omega}{(\gamma+\omega)(\alpha+\omega)} - \frac{\omega}{\beta}, \\ \epsilon(\infty) &= \frac{\gamma+\omega}{\alpha}y(\infty),\ z(\infty) = \frac{\gamma}{\omega}y(\infty). \end{aligned}$$

   This result has been shown in the literature, as a special case of the results on page 175 of Brauer (2008).
3. SIRS with $\omega > 0$, $\delta > 0$ and $\alpha \to \infty$:

$$\begin{aligned} x(\infty) &= \frac{\gamma+\omega}{\beta}, \\ y(\infty) &= \frac{\delta+\omega}{\gamma+\delta+\omega}[1-x(\infty)] = \frac{(\delta+\omega)(\beta-\gamma-\omega)}{\beta(\gamma+\delta+\omega)}, \\ z(\infty) &= \frac{\gamma}{\delta+\omega}y(\infty). \end{aligned}$$

4. SIR with $\omega > 0$, $\delta = 0$ and $\alpha \to \infty$:

$$\begin{aligned} x(\infty) &= \frac{\gamma+\omega}{\beta}, \\ y(\infty) &= \frac{\omega}{\gamma+\omega}[1-x(\infty)] = \frac{\omega}{\gamma+\omega} - \frac{\omega}{\beta}, \\ z(\infty) &= \frac{\gamma}{\omega}y(\infty). \end{aligned}$$

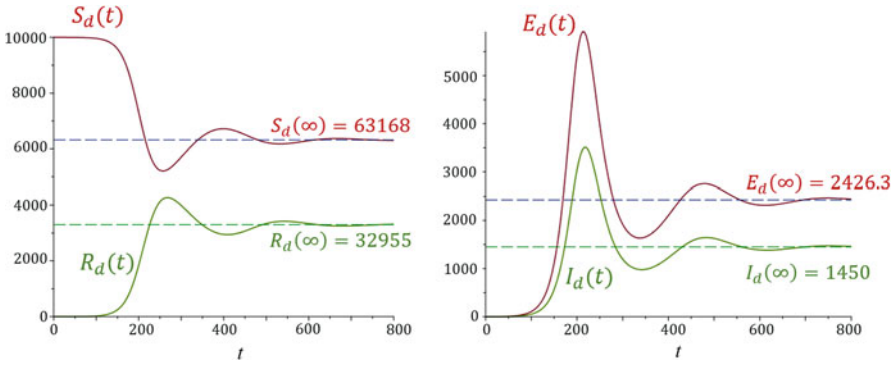


Fig. 5.9 Plots of $S_d(t)$, $I_d(t)$, $E_d(t)$, $R_d(t)$ determined by (5.69) at $\beta = 0.4$, $\alpha = 0.15$, $\gamma = 0.25$, $\delta = 0.01$ and $\omega = 0.001$. The time-scale is measured in days

Example 25 We consider a disease with an average latent period 6.7 days, average infectious period 4 days, and an average duration of immunity after recovery of 100 days. Assuming these durations are all exponentially distributed, they correspond to $\alpha = 0.15$, $\gamma = 0.25$, $\delta = 0.01$. We assume a constant population, and individuals stay in this population on average 1000 days (2.7 years). There is no importation of infected individuals. This gives $\omega = 0.001$. Let $\beta = 0.4$, we calculate $R_0 = 1.5831$ according to (5.70). We also calculate, using (5.71), $x(\infty) = 0.63168$, $y(\infty) = 0.0145$, $\epsilon(\infty) = 0.024263$, and $z(\infty) = 0.32955$. Consider a population with $m = 100{,}000$, the expected prevalence numbers at equilibrium are $S_d(\infty) = 63{,}168$, $I_d(\infty) = 1450$, $E_d(\infty) = 2426.3$, and $R_d(\infty) = 32{,}955$. In a deterministic framework, solving the system of the differential equations (5.69) given the initial values $I_d(0) = 1$ and $S_d(0) = 99{,}999$ yields numerically calculated $[S_d(t), I_d(t), E_d(t), R_d(t)]$ which show epidemic waves in the form of damped oscillations about the steady states $[S_d(\infty), I_d(\infty), E_d(\infty), R_d(\infty)]$. They are presented in Fig. 5.9. Numerical computation in this example is carried out using Maple-2017 (Maplesoft: Waterloo Maple Inc.)

5.6 Problems and Supplements

5.1 Examine the differential equations given by (5.12) and (5.23). The transmission dynamics of the SIS and SIR models with respect to the mean values $E[I(t)]$ and $E[S(t)]$ also depend on the population size m and the second moments, $var[I(t)]$ in the SIS model and $cov\{S(t), I(t)\}$ in the SIR model. There are two different ways to make them in agreement with the deterministic SIS and SIR models given by (5.13) and (5.24): one is assuming $m \to \infty$ in the second terms in (5.12) and (5.23); the other one is assuming $var[I(t)] = 0$ in (5.12) and assuming $cov\{S(t), I(t)\} = 0$ in (5.23).

(a) Which of the above assumptions are true to the meaning of *deterministic*?
(b) In the literature, it is often said that the deterministic models (5.13) and (5.24) are approximations of the mean field of the corresponding stochastic models. Does it mean that these approximations are in the sense of a very large population? Use Figs. 5.1 and 5.2 to facilitate this discussion.
(c) In the SIS and SIR models, stochastic and deterministic alike, the ratio β/γ is the basic reproduction number R_0. From the deterministic point of view, if $R_0 < 1$, there is no epidemic whereas when $R_0 > 1$, there is an epidemic with its course deterministically predicted by the differential equations. What do their stochastic counterparts say?

5.2 Consider the deterministic SIS and SIR models (5.13) and (5.24). The early growth of the epidemic is under the assumption that the depletion of the susceptible population is negligible, that is, $S_d(t)/m = 1$.

(a) Show that $I_d(t)$ has the exponential growth, proportional to e^{rt}, where $r = (R_0 - 1)\gamma$ and $R_0 = \beta/\gamma$. Hence, $r = \beta - \gamma$.
(b) When $I_d(t)$ has the exponential growth with rate r, it implies that the duration Y to the next new infection is exponentially distributed with p.d.f. re^{-ry} and this duration competes with the infection period T_I with p.d.f. $f_I(x) = \gamma e^{-\gamma x}$. Show that

$$\Pr(Y \le T_I) = \frac{r}{r+\gamma}.$$

(c) Verify that when $\beta > \gamma$, $r = \beta - \gamma$ and is the solution of the equation $r = \beta \Pr(Y \le T_I) = \beta\,(1 - L[f_I](r))$.

5.3 According to a deterministic SIR model (5.24), we assume a population of 10,000 individuals, the initial condition $I_d(0) = 1$ and the average infectious periods = 3 days.

(a) What are the values for the transmission rate β when $R_0 = 1.2$ and 2.0?
(b) Generate a graph of the final size $C_d(\infty)$, the total number of infected individuals at the end of the outbreak, against the values of β in the range $0.03 \le \beta \le 1$.
(c) Assuming $R_0 = 1.8$, calculate the following quantities:

(i) the initial growth rate r assuming $S(t)/m = 1$.
(ii) the peak prevalence value $I_d^{\max}$ and the time when the peak prevalence is reached;
(iii) the peak incidence value $i_d^{\max}$ and the time when the peak incidence is reached;
(iv) the final size $C_d(\infty)$ and the value of the epidemic $\int_0^\infty I_d(t)dt$.

5.4 Consider the following SIR model, in which $\frac{d}{dt}S(t) = -\beta\frac{S(t)I(t)}{m}$ but the rate of recovery (from the I-compartment to the R-compartment) depends on the time x since infection, expressed by the hazard function $h_X(x;\gamma) = \frac{4x\gamma^2}{2x\gamma+1}$.

(a) Plot $h_X(x;\gamma)$ in the range $0 < x < 10$ at $\gamma = 1/3$
(b) Show that the p.d.f. of the infectious period T_I is $f_I(x;\gamma) = 4x\gamma^2 e^{-2\gamma x}$. Calculate its mean and compare with the mean of the exponential distribution with p.d.f. $\gamma e^{-\gamma x}$. What is the expression of R_0?
(c) Calculate the variance of T_I and write down the expression of the Laplace transform $L[f_I](s)$. Compare the variance with the variance of the exponential distribution with p.d.f. $\gamma e^{-\gamma x}$ and compare the Laplace transform function with that corresponding to the exponential distribution.
(d) Write the relationship between R_0 and the initial growth rate r for this model. Given the same R_0, is r larger or smaller than $r = (R_0 - 1)\gamma$ as predicted by (5.24)? Make further comments on this finding by referencing Exercise 4.6.
(e) Show that the distribution of T_I can be obtained as the distribution of the sum of two independently distributed random variables $T_I = T_1 + T_2$, where T_1 and T_2 are identically distributed according to the exponential distribution with rate 2γ.
(f) Write a deterministic SIR model expressed by a system of ordinary differential equations with recovery hazard function given by $h_X(x;\gamma) = \frac{4x\gamma^2}{2x\gamma+1}$.
(g) Consider a population of 10,000 individuals, the initial condition $I_d(0) = 1$ and the average infectious periods = 3 days. Assuming $R_0 = 1.8$, calculate the initial growth rate r, the final size $C_d(\infty)$, and the value of the epidemic $\int_0^\infty I_d(t)dt$.
(h) Based on the values given in (g), use numeric differential equation solvers available in commercially available software (e.g., Matlab, Maple, etc.) to produce comparative illustrations for $S(t)$ and $I(t)$ based on the ordinary differential equations developed in (f) against those based on the ordinary differential equations given by (5.24). What do they have in common, and what are the differences?

5.5 For a population of 10,000 individuals, consider the SEIR model (5.53) with an average latent periods = 10 days and an average infectious periods = 3 days. Assuming $R_0 = 1.8$.

(a) Calculate the final size $C_d(\infty) = m - S_d(\infty)$. Does it depend on the presence of latent periods?
(b) Define the peak prevalence as $P(t) = E_d(t) + I_d(t)$ for the SEIR model and calculate the value of the peak prevalence $P_{\max}$. Does the peak prevalence value depend on the presence of latent periods?
(c) Produce a similar figure as Fig. 5.8 that compares the SEIR model with the SIR model and comment what the two models share in common, and what are the differences.

Chapter 6
More Complex Models and Control Measures

We have seen that, under suitable assumptions such as homogeneous mixing, the basic reproduction number R_0, defined at the start of the epidemic and given by (4.2) in Chap. 4, transcends to the asymptotic equilibrium ($t \to \infty$) outcomes such as the final size (5.32) in a closed population or the endemic equilibrium $x(\infty) = \lim_{t\to\infty} S_d(t)/m \to R_0^{-1}$ in a constant population. Meanwhile, we have also seen that, in compartment transmission models of the SEIRS type (in Chap. 5) with exponentially distributed durations, R_0 is expressed as a function of parameters representing rates in these models, such as $R_0 = \beta/\gamma$ in SEIRS models without mortality or other in-flow and out-flow of the population, or (5.70) in SEIRS models with mortality or other in-flow and out-flow of the population.

The SEIRS models in Chap. 5 are building blocks towards models with more complex structures with specific public health questions in mind. Very often these models are designed to evaluate public health measures and treatments.

Before going into these topics, we quickly review the generalizability of the final size equation and expressions of the reproduction number as functions of rate parameters in complex compartment models.

6.1 The Final Size Equation and the Reproduction Number

Proposition 5.32 provides a central limit theorem in a stochastic framework for the distribution of the cumulative number of infections $C(\infty)$ in an infinitely large population. This central limit tendency has been studied and proven by many authors under different assumptions regarding disease transmission. We refer readers to von Bahr and Martin-Löf (1980), Ludwig (1975), Scalia-Tomba (1985), Martin-Löf (1988), and Lefèvre and Picard (1995), among others, for the suitable conditions and proofs.

P. Yan, G. Chowell, *Quantitative Methods for Investigating Infectious Disease Outbreaks*, Texts in Applied Mathematics 70,
https://doi.org/10.1007/978-3-030-21923-9_6

In a deterministic framework, Ma and Earn (2006) show that, in a closed population with homogeneous mixing, if the depletion of the total population by deaths is negligible, the monotonic relationship between R_0 and the final size equation (5.32) is invariant. This includes the existence of latent periods, arbitrarily distributed infectious periods, and any number of distinct infectious stages.

This is an important mathematical relationship. It implies that, in a large outbreak, the essential parameter $R_0 > 1$, which is defined at the very beginning when the system is at disease-free equilibrium, transcends to the final size $0 < \eta < 1$, which is defined at the end when the system is at disease-free equilibrium with no infected individuals left in the population. Figures 5.4, 5.5, 5.6, 5.7, and 5.8 have demonstrated that, under the same initial conditions and R_0, the average lengths and distributions of the infectious periods and the presence of the latent periods affect the time course of the epidemics, but in the long run, the final sizes $C_d(\infty)$, $R_d(\infty)$, and $S_d(\infty)$ do not change. *All roads lead to Rome*.

There is a hidden assumption, however, that no change of contact behavior and no public health intervention, such as isolation, quarantine, vaccination, and other pharmaceutical intervention, ever take place throughout the course of the outbreak. This is unrealistic, as there are always behavior changes and public health interventions during the course of the epidemic (Funk et al. 2009, 2010; Perra et al. 2011). The actual final size, through the final size equation, is conceptually related to a smaller reproduction number $R_c \leq R_0$, as if the epidemic started with R_c at the very beginning of the epidemic.

Figure 6.1 shows the monotonic but nonlinear relationship of the final size equation (5.32) that, given the same resource and effort, the intervention is more

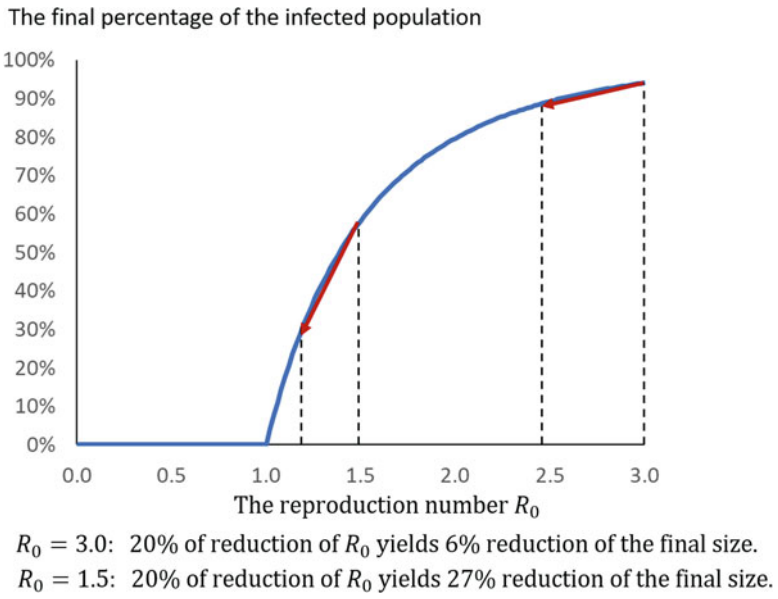


Fig. 6.1 Implicit plot of the final size equation (5.32) and effect of a proportionate reduction of R_0 on the final size

effective when the reproduction number is less than 2 compared to the scenario where the reproduction number is greater than 2.5. This specific insight leads to the critical vaccine coverage needed in order to produce herd immunity (Halloran et al. 2009).

6.2 The Reproduction Number as the Non-negative Eigenvalue of the Next Generation Matrix in Compartmental Disease Transmission Models

Using the notations and framework in pages 160–162 of van den Driessche and Watmough (2008), a general form of compartmental disease transmission models may consist of k disease compartments. A compartment is called a disease compartment if the individuals therein are infected, including both asymptomatic and symptomatic stages of the disease. Let $y \in \mathbb{R}^k$ be the subpopulations in each of these compartments. The linearized infection subsystem can be written in the form of

$$y' = (F - V)\, y, \tag{6.1}$$

where F and V are $k \times k$ matrices. The matrix F corresponds to transmissions. All epidemiological events that lead to new infections are incorporated in the model through F. The element F_{ij} of matrix F is the rate at which infected individuals in the i^{th} disease compartment produce secondary infected individuals in the j^{th} disease compartment. The matrix V corresponds to transitions. The element V_{ij} has the interpretation of the rate of transition from compartment i to compartment j. All these parameters are based on setting the model at disease-free equilibrium. We refer to page 161 of van den Driessche and Watmough (2008) for detailed descriptions. Continuing, the next generation matrix is defined by FV^{-1} and $R_0 = \rho(FV^{-1})$ is the non-negative eigenvalue of FV^{-1}, see pages 163, 173–175 of van den Driessche and Watmough (2008).

For intuitive understanding of (6.1) and $R_0 = \rho(FV^{-1})$, we notice that $y' = (F - V)\, y$ is analogous to $y' = (\beta S - \gamma)\, y$ in the SIS and SIR models and $R_0 = \rho(FV^{-1})$ is analogous to $R_0 = \beta/\gamma$.

In the deterministic SEIRS model given by (5.69), when $\omega = 0$, the presence of a latent period has no effect on the basic reproduction number, which is $R_0 = \beta/\gamma$; when $\omega = 0$, R_0 is expressed as (5.70). The linearized infection subsystem is a system of ordinary differential equations of the matrix form

$$\begin{pmatrix} E' \\ I' \end{pmatrix} = (F - V) \begin{pmatrix} E \\ I \end{pmatrix},$$

where

$$F = \begin{pmatrix} 0 & \beta \\ 0 & 0 \end{pmatrix},\ V = \begin{pmatrix} \alpha + \omega & 0 \\ -\alpha & \gamma + \omega \end{pmatrix} \text{ and } FV^{-1} = \begin{pmatrix} \frac{\beta\alpha}{(\alpha+\omega)(\gamma+\omega)} & \frac{\beta}{\alpha+\omega} \\ 0 & 0 \end{pmatrix}.$$

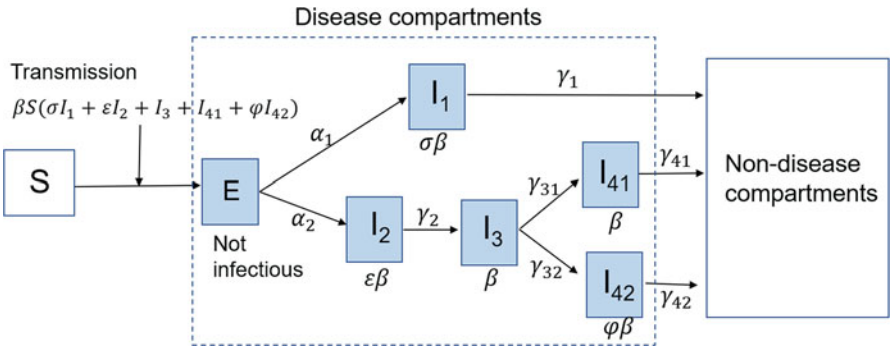


Fig. 6.2 A conceptual expansion of an SEIR model with different infectious stages and intervention parameters

This returns to (5.70):

$$R_0 = \rho(FV^{-1}) = \frac{\beta\alpha}{(\alpha+\omega)(\gamma+\omega)}.$$

Complex compartment models are often used to explicitly model these interventions and evaluate their effectiveness by expanding simple structures. For instance, in the model illustrated in Fig. 6.2, individuals in Compartment E may have a probability $\psi = \frac{\alpha_1}{\alpha_1+\alpha_2}$ to receive public health intervention. When these individuals move into their infectious stages, I_1, their infectious contact rates are reduced to $\sigma\beta$ (assuming the intervention may not be perfectly adhered). For those not receiving the intervention, their infectious periods may be staged into $I_2 \longrightarrow I_3$, where stage I_2 may be more or less infectious than stage I_3. Hence, $\beta_2 = \varepsilon\beta$ and $\beta_3 = \beta$. Individuals in Compartment I_3 may receive treatments, with probability $\phi = \frac{\gamma_{32}}{\gamma_{31}+\gamma_{32}}$. Treated individuals ($I_{42}$) may have reduced infectious contact rates: $\beta_{42} = \varphi\beta$. Assigning these parameters into the model given by (6.2), the reproduction number $\rho(FV^{-1})$ given by (6.3) is explicitly expressed as a function of parameters representing public health interventions, along with other epidemiologic parameters with respect to infectious periods distributions.

We consider a structured model to illustrate (6.1) and $R_c = \rho(FV^{-1})$. We use the notation R_c because the model includes control parameters; and it should be interpreted as the controlled reproduction number R_c rather than the basic reproduction number R_0. The model in Fig. 6.2 includes six disease compartments. Compartment E stands for individuals who are infected but not infectious (i.e., $\beta = 0$). Other disease compartments are all infectious, but with different infectious contact rates, with scale parameters against the baseline parameter β. The rates from E to I_1 and I_2 are denoted by α_1 and α_2, respectively. Other transition parameters are represented by γ with index denoting the current infectious state from which

an infected individual is leaving. The linearized infection subsystem is a system of ordinary differential equations of the matrix form (6.1) as

$$\begin{pmatrix} E' \\ I_1' \\ I_2' \\ I_3' \\ I_{41}' \\ I_{42}' \end{pmatrix} = (F - V) \begin{pmatrix} E \\ I_1 \\ I_2 \\ I_3 \\ I_{41} \\ I_{42} \end{pmatrix}, \tag{6.2}$$

where the transmission matrix

$$F = \begin{pmatrix} 0 & \beta\sigma S_0 & \beta\varepsilon S_0 & \beta S_0 & \beta S_0 & \beta\varphi S_0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

and $S_0 = S(0)$ denotes the number of susceptible individuals at disease-free equilibrium. The transition matrix is

$$V = \begin{pmatrix} \alpha_1 + \alpha_2 & 0 & 0 & 0 & 0 & 0 \\ -\alpha_1 & \gamma_1 & 0 & 0 & 0 & 0 \\ -\alpha_2 & 0 & \gamma_2 & 0 & 0 & 0 \\ 0 & 0 & -\gamma_2 & \gamma_{31} + \gamma_{32} & 0 & 0 \\ 0 & 0 & 0 & -\gamma_{31} & \gamma_{41} & 0 \\ 0 & 0 & 0 & -\gamma_{32} & 0 & \gamma_{42} \end{pmatrix}.$$

Then

$$V^{-1} = \begin{pmatrix} \frac{1}{\alpha_1+\alpha_2} & 0 & 0 & 0 & 0 & 0 \\ \frac{\alpha_1}{\alpha_1+\alpha_2}\frac{1}{\gamma_1} & \frac{1}{\gamma_1} & 0 & 0 & 0 & 0 \\ \frac{\alpha_2}{\alpha_1+\alpha_2}\frac{1}{\gamma_2} & 0 & \frac{1}{\gamma_2} & 0 & 0 & 0 \\ \frac{\alpha_2}{\alpha_1+\alpha_2}\frac{1}{\gamma_{31}+\gamma_{32}} & 0 & \frac{1}{\gamma_{31}+\gamma_{32}} & \frac{1}{\gamma_{31}+\gamma_{32}} & 0 & 0 \\ \frac{\alpha_2}{\alpha_1+\alpha_2}\frac{\gamma_{31}}{\gamma_{31}+\gamma_{32}}\frac{1}{\gamma_{41}} & 0 & \frac{\gamma_{31}}{\gamma_{31}+\gamma_{32}}\frac{1}{\gamma_{41}} & \frac{\gamma_{31}}{\gamma_{31}+\gamma_{32}}\frac{1}{\gamma_{41}} & \frac{1}{\gamma_{41}} & 0 \\ \frac{\alpha_2}{\alpha_1+\alpha_2}\frac{\gamma_{32}}{\gamma_{31}+\gamma_{32}}\frac{1}{\gamma_{42}} & 0 & \frac{\gamma_{32}}{\gamma_{31}+\gamma_{32}}\frac{1}{\gamma_{42}} & \frac{\gamma_{32}}{\gamma_{31}+\gamma_{32}}\frac{1}{\gamma_{42}} & 0 & \frac{1}{\gamma_{42}} \end{pmatrix}.$$

The non-negative eigenvalue of FV^{-1} is $R_c = \rho(FV^{-1})$, expressed as

$$\begin{aligned}
&\beta S_0 \left\{ \frac{\alpha_1}{\alpha_1+\alpha_2}\frac{\sigma}{\gamma_1} \right. \\
&\left. + \frac{\alpha_2}{\alpha_1+\alpha_2}\left[\frac{\varepsilon}{\gamma_2} + \frac{1}{\gamma_{31}+\gamma_{32}} + \frac{\gamma_{31}}{(\gamma_{31}+\gamma_{32})\,\gamma_{41}} + \frac{\gamma_{32}\varphi}{(\gamma_{31}+\gamma_{32})\,\gamma_{42}} \right] \right\} \\
&= \beta S_0 \left\{ \frac{\alpha_1}{\alpha_1+\alpha_2}\frac{\sigma}{\gamma_1} + \frac{\alpha_2}{\alpha_1+\alpha_2}\left[\frac{\varepsilon}{\gamma_2} + \frac{\gamma_{31}\gamma_{42}+\gamma_{41}\gamma_{42}+\varphi\gamma_{32}\gamma_{41}}{\gamma_{41}\gamma_{42}\,(\gamma_{31}+\gamma_{32})} \right] \right\}. \qquad (6.3)
\end{aligned}$$

6.2.1 *An Intuitive Recipe to Express R_c in Complex Compartment Models with Non-exponentially Distributed Sojourn Times in Disease Compartments*

We have noticed that (5.70) is expressed as $R_0 = \beta L[f_E](\omega)L[\overline{F}_I](\omega)$ rather than $R_0 = \rho(FV^{-1})$. The former is also suitable for non-exponentially distributed latent and infectious periods using the independent competing risk approach and Laplace transforms. On the other hand, the linearized infection subsystem as a system of ordinary differential equations (6.1) assumes constant transition rates in matrix V. It implies the sojourn times an infected individual spent in each of the disease compartments is exponentially distributed.

In a similar manner, we use the independent competing risk framework to develop a recipe to generalize $R_0 = \rho(FV^{-1})$ in situations where the sojourn times are distributed according to some arbitrary distributions.

We recall that, in a closed population under homogeneous mixing, the infectious contact process $\{K(x) : x \geq 0\}$ possesses the stationary increment property $\frac{d}{dt}E[K(t)|\mathcal{H}_t] = \beta$ (Sect. 3.3), $R_0 = \beta\mu_I$. This expression is valid for arbitrarily distributed infectious periods as long as the first moment $\mu_I < \infty$. We extend this relationship in the following manner:

1. The stationary increment property can be generalized by staging the infectious period (shown as structure (a)) in Fig. 6.3), so that $\beta(x)$ is piecewise constant defined on different stages. In this serial arrangement, each infected individual passes through a series of k infectious stages $I^{(1)} \to I^{(2)} \to \cdots \to I^{(k)}$. Each stage is random with duration $T_I^{(j)}$ associated with survival function $\overline{F}_I^{(j)}(x)$ and mean μ_j, such that $\mu_I = \sum_{j=1}^{k}\mu_j$. If $\beta(x) = \beta_j$, for $x \in I^{(j)}$, it can be shown that (Mode and Sleeman 2000)

$$R_c = \sum_{j=1}^{k}\int_0^\infty \beta(x)\overline{F}_I^{(j)}(x)dx = \sum_{j=1}^{k}\beta_j\mu_j.$$

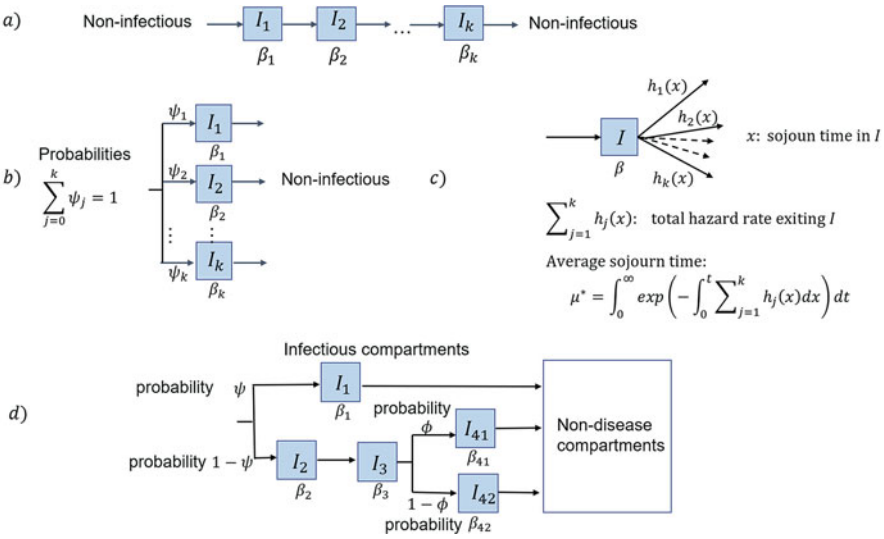


Fig. 6.3 An illustration of (**a**) serial, (**b**) parallel, (**c**) splitting, and (**d**) combined structures of the infectious period

2. Infected individuals are composed of different types. Each type is associated with an infectious period with average duration μ_j and infectious contact rate $\beta_j, = 1, \ldots, l$. A typical susceptible individual has a probability ψ_j to be in contact with infected individuals of type j. In this parallel arrangement, shown as structure (b) in Fig. 6.3, $R_c = \sum_{j=1}^{l} \psi_j \beta_j \mu_j$, which can be proven using $R_0 = G_N^{'}(1)$, in which $G_N(s)$ arises from a mixed distribution with finite mixture $\sum_{j=1}^{l} \psi_j = 1$.
3. Individuals in the infected stage I, with infectious contact rate β, exit to k competing directions in an independent competing risk framework, shown as structure (c) in Fig. 6.3. The average duration of an individual staying in this stage is $\mu_I^* = \int_0^\infty \exp\left(-\int_0^t \sum_{j=1}^{k} h_j(x)dx\right)dt$. In particular, if $h_j(x) = \gamma_j$ for all j, $\mu_I^* = \left(\sum_{j=1}^{k} \gamma_j\right)^{-1}$. We have $R_c = \beta \mu_I^*$.
4. We combine 1. and 2. to make more complex structures involving both serial and parallel structures such as structure (d) in Fig. 6.3.

Figure 6.3 structure (d) is a sub-structure of Fig. 6.2, illustrating a situation where a typical susceptible individual has a probability ψ of coming into contact with infected individuals I_1 with mean infectious period μ_1 and infectious contact rate β_1. It also has a probability $1 - \psi$ of coming into contact with infected individuals whose infectious periods are staged into series I_2 and I_3. After I_3, these infected individuals may also split, with probability ϕ, into two parallel categories of the infectious periods: I_{41} and I_{42}, respectively. Their corresponding mean durations and infectious contact rates while in contact with a susceptible individuals are

labeled as μ_2, μ_3, μ_{41}, μ_{42}, and β_2, β_3, β_{41}, β_{42}, respectively. In this case, one can write

$$R_c = \psi\beta_1\mu_1 + (1-\psi)\left[\beta_2\mu_2 + \beta_3\mu_3^* + \phi\beta_{41}\mu_{41} + (1-\phi)\,\beta_{42}\mu_{42}\right]. \tag{6.4}$$

The expression (6.4) not only generalizes $R_0 = \beta\mu_I$ into a model with complex structure with possible intervention parameters, but also allows the durations in each stage to be arbitrarily distributed, as long as the mean values μ_1, μ_2, μ_3^*, μ_{41}, and μ_{42} are all finite.

In Fig. 6.2, all parameters are expressed by rates. Individuals in Compartment E have probability $\psi = \frac{\alpha_1}{\alpha_1+\alpha_2}$ of making a transition to Compartment I_1 and probability $1-\psi = \frac{\alpha_2}{\alpha_1+\alpha_2}$ of making a transition to Compartment I_2. Similarly, individuals in Compartment I_3 have probability $\phi = \frac{\gamma_{31}}{\gamma_{31}+\gamma_{32}}$ of making a transition to Compartment I_{41} and probability $1-\phi = \frac{\gamma_{32}}{\gamma_{31}+\gamma_{32}}$ of making a transition to Compartment I_{42}. The average sojourn time in I_3 is $\mu_3^* = \frac{1}{\gamma_{31}+\gamma_{32}}$. Let $\beta_1 = \sigma\beta$, $\beta_2 = \varepsilon\beta$, $\beta_3 = \beta_{41} = \beta$, and $\beta_{42} = \varphi\beta$, (6.4) returns to (6.3).

6.3 A Hypothetical Case Study for Preparedness of an Acute Respiratory Infectious Disease

We consider a local outbreak of a typical acute respiratory infectious disease so that the population is approximately constant and closed. The transmission structure is illustrated in Fig. 6.4, which is very similar to Fig. 6.2. There is a proportion of infected individuals who do not progress to symptomatic stages and remain asymptomatic until recovery. We denote the hazard function from Stage L to Stage 0 as $h_{L0}(x)$ and the hazard function from Stage L to Stage A as $h_{LA}(x)$, where x stands for the sojourn time in stage L, then

$$\theta = \int_0^\infty h_{L0}(x)e^{-\int_0^x (h_{LA}(u)+h_{L0}(u))du}dx = \Pr\{\text{ clinically ill} \mid \text{infection }\}.$$

In particular, if $h_{LA}(x) = \alpha_{LA}$ and $h_{L0}(x) = \alpha_{L0}$, then $\theta = \frac{\alpha_{L0}}{\alpha_{LA}+\alpha_{L0}}$. This parameter is called *pathogenicity* in infectious disease literature, which is *the ability of a microbial agent to induce disease* (Nelson et al. 2001).

We assume a treatment is applied to individuals in Phase-I of the symptomatic stage. The time-to-treatment T_X follows a hazard function $h_{1T}(x)$, where x is measured from the symptomatic onset; and the time to progression to Phase-II since symptomatic onset, T_X, follows a hazard function $h_{12}(x)$. Assuming independency between T_X and T_{12}, we define

$$\Phi = \int_0^\infty h_{1T}(x)e^{-\int_0^x (h_{12}(u)+h_{1T}(u))du}dx = \Pr\{T_X \le T_{12}\}, \tag{6.5}$$

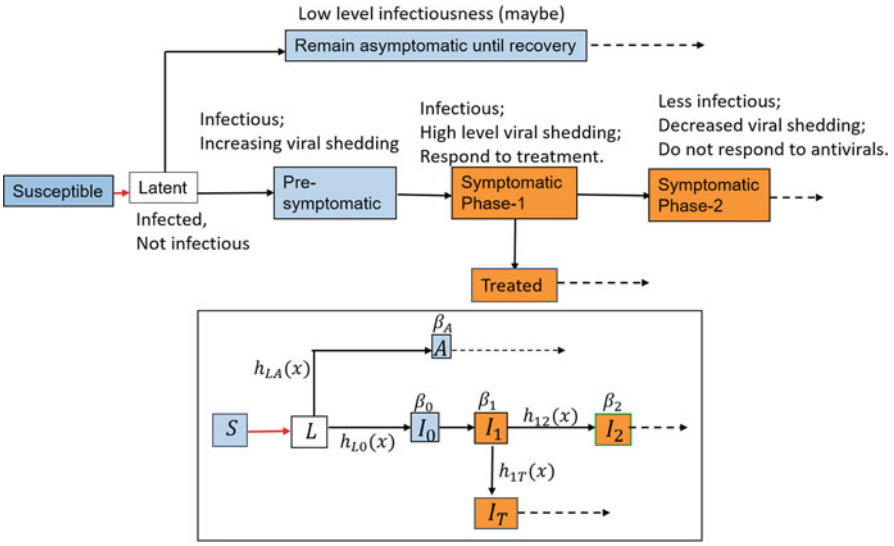


Fig. 6.4 A diagram of the structure of a hypothetical acute respiratory infection

which is the probability of a clinically ill patient being effectively treated, and $\Pr\{T_X \leq T_{12}\}$ symbolizes "race-to-treat."

6.3.1 The Baseline: Without Treatment

Without treatment, the final size is η, corresponding to the final size equation (5.32) and shown in Fig. 6.1. We scale η by pathogenicity to define the clinical attack rate,

$$A_0 = \theta\eta.$$

In the literature, η is also called the *serologic attack rate* (as used in Gani et al. 2005) or the *infection attack rate* (as used in Ferguson et al. 2005). A_0 is linked to the basic reproduction number R_0 via (5.32) and θ. The baseline scenario is without treatment. An explicit expression for R_0 can be obtained using the recipe provided in the proceeding section:

$$R_0 = (1-\theta)\beta_A\mu_A + \theta\left[\beta_0\mu_0 + \beta_1\mu_1 + \beta_2\mu_2\right], \tag{6.6}$$

where μ with subscripts are average duration of individuals staying in each stage according to the numbering in Fig. 6.4.

In the special case when all the transition rates are constants, $h_{LA}(u) = \alpha_{LA}$, $h_{L0}(u) = \alpha_{L0}$, $\mu_A = \gamma_A^{-1}$, $\mu_0 = \gamma_{01}^{-1}$, $\mu_2 = \gamma_{12}^{-1}$, and $\mu_2 = \gamma_2^{-1}$, we can write a system of ordinary differential equations as a deterministic transmission model

$$\begin{aligned}
S' &= -S\left(\beta_A A + \beta_0 I_0 + \beta_1 I_1 + \beta_2 I_2\right)/m \\
L' &= S\left(\beta_A A + \beta_0 I_0 + \beta_1 I_1 + \beta_2 I_2\right)/m - \left(\alpha_{LA} + \alpha_{L0}\right) L \\
A' &= \alpha_{LA} L - \gamma_A A \\
I_0' &= \alpha_{L0} L - \gamma_{01} I_0 \\
I_1' &= \gamma_{01} I_0 - \gamma_{12} I_1 \\
I_2' &= \gamma_{12} I_1 - \gamma_2 I_2 \\
R' &= \gamma_A A + \gamma_2 I_2
\end{aligned},$$

where $\{S, L, A, I_0, I_1, I_2\}$ are states as displayed in Fig. 6.4 satisfying

$$S(t) + L(t) + A(t) + I_0(t) + I_1(t) + I_2(t) + R(t) = m.$$

R_0 can be obtained using the second generation matrix method

$$R_0 = \frac{\alpha_{LA}}{\alpha_{LA} + \alpha_{L0}} \beta_A/\gamma_A + \frac{\alpha_{L0}}{\alpha_{LA} + \alpha_{L0}} \left[\beta_0/\gamma_{01} + \beta_1/\gamma_{12} + \beta_2/\gamma_2\right]$$

which is a special case of (6.6).

6.3.2 With Treatment

The Reproduction Number Reduction

We imagine a treatment, such as an antiviral drug, which is effective both in the reduction of transmissibility (if treated immediately after onset of symptoms) and in the reduction of severe clinical outcomes. Using the recipe provided in the proceeding section,

$$R_c = (1-\theta)\beta_A \mu_A + \theta\left[\beta_0 \mu_0 + \beta_1 \mu_1^* + \Phi \beta_T \mu_T + (1-\Phi)\beta_2 \mu_2\right], \tag{6.7}$$

where μ with subscripts are average duration of individuals staying in each stage according to the numbering in Fig. 6.4 with the exception

$$\mu_1^* = \int_0^\infty e^{-\int_0^x (h_{12}(u) + h_{1T}(u))du} dx.$$

If $h_{1T}(u) = 0$, then $\mu_1^* = \int_0^\infty e^{-\int_0^x h_{12}(u)du} dx = \mu_1$.

We assume that for an effectively treated individual, the mean duration of infectiousness is a reduction of μ_2, so that $\mu_T = \varphi\mu_2$, $\varphi < 1$. Meanwhile, the reduction of infectiousness is $\beta_T = \kappa\beta_2, \kappa < 1$. Then

$$\frac{R_0 - R_c}{\theta} = \beta_1 \left(\mu_1 - \mu_1^*\right) + \Phi\left(1 - \kappa\varphi\right)\beta_2\mu_2.$$

If, without treatment, the natural duration of Phase-I is generally distributed with survival function $\overline{F}_{12}$, and the race-to-treat in Phase-I follows a constant rate $h_{1T}(x) = \gamma_{1T}$, then

$$\mu_1^* = \int_0^\infty e^{-x\gamma_{1T}} e^{-\int_0^x h_{12}(u)du} dx = L[\overline{F}_{12}](\gamma_{1T}),$$

$$\mu_1 = \int_0^\infty e^{-\int_0^x h_{12}(u)du} dx = L[\overline{F}_{12}](0),$$

$$\Phi = \gamma_{1T} \int_0^\infty e^{-x\gamma_{1T}} e^{-\int_0^x h_{12}(u)du} dx = \gamma_{1T} L[\overline{F}_{12}](\gamma_{1T}) = \gamma_{1T}\mu_1^*.$$

The controlled reproduction number is a function of treatment rate γ_{1T}, such that

$$\frac{R_0 - R_c}{\theta} = \beta_1 \left(L[\overline{F}_{12}](0) - L[\overline{F}_{12}](\gamma_{1T})\right) + \gamma_{1T} L[\overline{F}_{12}](\gamma_{1T})\left(1 - \kappa\varphi\right)\beta_2\mu_2.$$

A further special case is when the natural duration of Phase-I is exponentially distributed with mean value $\mu_1 = \gamma_{12}^{-1}$. In this case,

$$\mu_1^* = L[\overline{F}_{12}](\gamma_T) = \frac{1}{\gamma_{12} + \gamma_{1T}} \text{ and } \Phi = \frac{\gamma_{1T}}{\gamma_{12} + \gamma_{1T}}.$$

Thus

$$\Phi = \frac{\gamma_{1T}}{\gamma_{12} + \gamma_{1T}} = \frac{R_0 - R_c}{\theta\left(\beta_1/\gamma_{12} + \left(1 - \kappa\varphi\right)\beta_2\mu_2\right)}. \tag{6.8}$$

It shows that pathogenicity θ plays a very important role. Because the treatment is applied to symptomatic individuals only, if θ is small and there is a large number of asymptomatic infected individuals who also transmit the infection, it becomes more difficult to use an antiviral type treatment alone to achieve the same reduction of the reproduction number.

We write

$$c = \frac{\theta\left(\beta_1/\gamma_{12} + \left(1 - \kappa\varphi\right)\beta_2\mu_2\right)}{(1-\theta)\beta_A\mu_A + \theta\left[\beta_0\mu_0 + \beta_1/\gamma_{12} + \beta_2\mu_2\right]} \tag{6.9}$$

which involves all the parameters about pathogenicity, transmission rates, and average durations without treatment, as well as $\kappa\varphi$ for the treatment. These parameters are pre-assumed. Re-writing (6.8), we get the linear reduction

$$R_c = R_0 (1 - c\Phi) . \tag{6.10}$$

The "Controlled" Clinical Attack Rate and the Calculation for the Maximum Amount of Treatment

The objective is to reduce the clinical attack rate and increase

$$\#\text{ clinical cases averted} = (A_0 - A_c) \times \text{ population size}, \tag{6.11}$$

where A_c is the "controlled" clinical attack rate. A_c is linked to the controlled reproduction number R_c.

Assuming the natural duration of Phase-I is exponentially distributed so that (6.10) holds, we write the controlled clinic attack rate $A_c(\Phi)$ as a function of Φ. The final size equation can be expressed as

$$1 - A_c(\Phi)/\theta = \exp(-R_c A_c(\Phi)/\theta), \ R_c > 1.$$

It can be rewritten as

$$-\log(1 - A_c(\Phi)/\theta) = R_0 (1 - c\Phi) A_c(\Phi)/\theta. \tag{6.12}$$

We define treatment usage as $\Omega(\Phi) = \Phi * A_c(\Phi)$, which can be expressed by

$$\Omega(\Phi) = \frac{\theta}{c}\left[A_c(\Phi)/\theta + \frac{1}{R_0}\log(1 - A_c(\Phi)/\theta)\right]. \tag{6.13}$$

The total number of effectively treated individuals is $m \times \Omega(\Phi)$. This is an important quantity to calculate the amount of treatment by multiplying the appropriate unit of treatment per treated individuals, such as courses, doses, or any other measures.

Both $A_c(\Phi)$ and $\Omega(\Phi)$ are functions of Φ and will be plotted in Fig. 6.5 under specific values of R_0, θ, and c.

To find the maximum treatment usage,

$$\Omega'(\Phi) = \frac{\theta}{c}\left[1 - \frac{1}{R_0}\frac{1}{1 - A_c(\Phi)/\theta}\right] A_c'(\Phi)/\theta.$$

We find $\Omega'(\Phi) = 0$ if $A_c(\Phi)/\theta = 1 - \frac{1}{R_0}$. Therefore, the maximum value is $\Omega_{\max} = \frac{\theta}{c}\left(1 - \frac{1}{R_0} - \frac{1}{R_0}\log R_0\right)$. This gives the following statement.

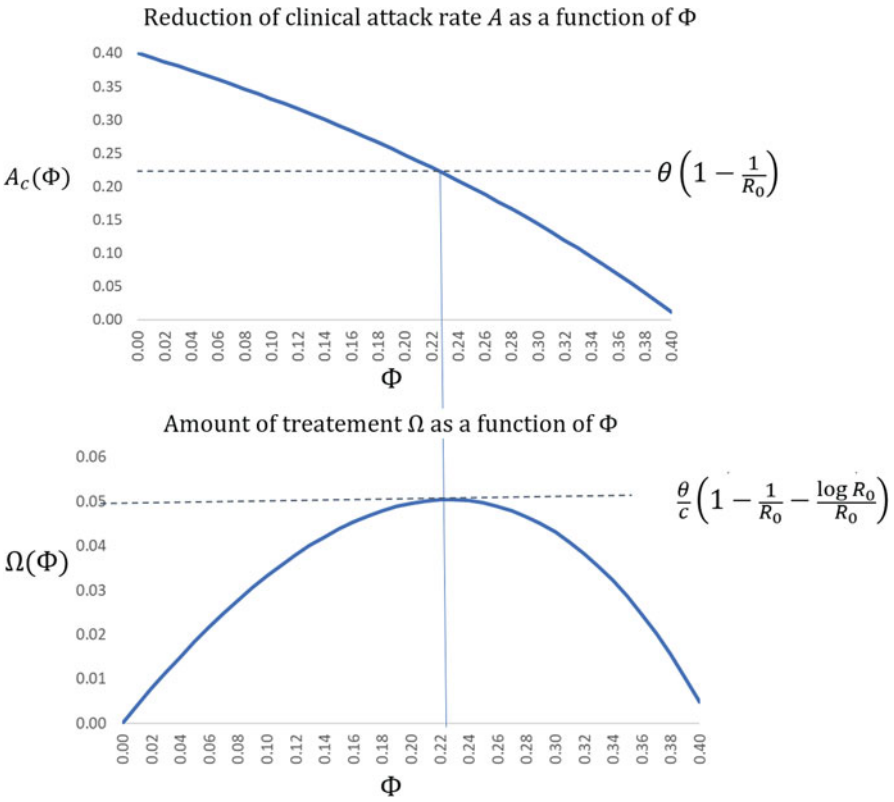


Fig. 6.5 Illustration of (6.12) and (6.13) at $R_0 = 1.386$, $\theta = 0.8$, and $c = 0.68278$

Proposition 26

- *If* $c > 1 - \frac{\log R_0}{R_0 - 1}$ *and* $R_0 > 1$, *the maximum value of* $\Omega(\Phi)$ *is*

$$\Omega_{max} = \frac{\theta}{c}\left(1 - \frac{1}{R_0} - \frac{1}{R_0}\log R_0\right). \tag{6.14}$$

This is when $\Phi = \frac{1}{c}\left(1 - \frac{\log R_0}{R_0 - 1}\right)$. *At this value, the controlled clinical attack rate is*

$$A_c(\Phi) = \theta\left(1 - \frac{1}{R_0}\right).$$

When $\Phi < \frac{1}{c}\left(1 - \frac{\log R_0}{R_0 - 1}\right)$, *an increase of* Φ *implies more treatment amount* $\Omega(\Phi)$. *When* $\Phi > \frac{1}{c}\left(1 - \frac{\log R_0}{R_0 - 1}\right)$, *an increase of* Φ *implies less treatment amount* $\Omega(\Phi)$.

- *If* $c \leq 1 - \frac{\log R_0}{R_0 - 1}$, *the maximum value of* $\Omega(\Phi)$ *is achieved when* 100% *of clinic patients are effectively treated* ($\Phi = 1$).
- *In particular, if* $c = 0$, $A_c(\Phi) = A_0$ *and the maximum value of* $\Omega(\Phi)$ *is the gross attack rate* A_0.

6.3.3 The Deterministic Model

If all the transition rates in Fig. 6.4 are constants, we expand the system of ordinary equations to

$$\begin{aligned}
S' &= -S\left(\beta_A A + \beta_0 I_0 + \beta_1 I_1 + \beta_T I_T + \beta_2 I_2\right)/m \\
L' &= S\left(\beta_A A + \beta_0 I_0 + \beta_1 I_1 + \beta_T I_T + \beta_2 I_2\right)/m - \left(\alpha_{LA} + \alpha_{L0}\right) L \\
A' &= \alpha_{LA} L - \gamma_A A \\
I_0' &= \alpha_{L0} L - \gamma_{01} I_0 \\
I_1' &= \gamma_{01} I_0 - \left(\gamma_{12} + \gamma_{1T}\right) I_1 \\
I_2' &= \gamma_{12} I_1 - \gamma_2 I_2 \\
I_T' &= \gamma_{1T} I_1 - \gamma_T I_T \\
R' &= \gamma_A A + \gamma_2 I_2 + \gamma_T I_T
\end{aligned}, \tag{6.15}$$

where $\{S, L, A, I_0, I_1, I_2, I_T\}$ are states as displayed in Fig. 6.4 satisfying

$$S(t) + L(t) + A(t) + I_0(t) + I_1(t) + I_2(t) + I_T(t) + R(t) = m.$$

R_c can be obtained using the second generation matrix method

$$R_c = \frac{\alpha_{LA}}{\alpha_{LA} + \alpha_{L0}} \frac{\beta_A}{\gamma_A} + \frac{\alpha_{L0}}{\alpha_{LA} + \alpha_{L0}} \left(\frac{\beta_0}{\gamma_{01}} + \frac{\beta_1}{\gamma_{12} + \gamma_{1T}} + \frac{\beta_2 \left(\gamma_{12} + \gamma_{1T} \kappa \varphi\right)}{\gamma_2 \left(\gamma_{12} + \gamma_{1T}\right)} \right)$$

which is a special case of (6.7).

6.3.4 A Numerical Demonstration

Consider an influenza-like disease, in which we assume an average latent period 2 days and average infectious period 5.25 days. The infectious period consists of a short pre-symptomatic period with average 0.25 days; a highly contagious Phase-I of the symptomatic stage, with average 1 day, during which an antiviral treatment is effective. Although the average of Phase-I of the symptomatic stage is assumed 1 day, if it is exponentially distributed, there is still 0.95 probability that treatment is still effective 3 days after symptom onset. Without treatment, Phase-I is followed by a lesser contagious Phase-II of the symptomatic stage with average 4 days. We

Table 6.1 Assumed parameters without treatment

	Infection rates	Average durations (days)
Latent	$\beta_L = 0$	$\mu_L = 2$
Asymptomatic	$\beta_A = \beta_2/2$	$\mu_A = 5.25$
Pre-symptomatic	$\beta_0 = 2\beta_2$	$\mu_0 = 0.25$
Symptomatic-I	$\beta_1 = 7\beta_2$	$\mu_1 = 1$
Symptomatic-II	β_2	$\mu_2 = 4$
Pathogenicity	$\theta = 0.8$	

assume that the pathogenicity parameter $\theta = 0.8$. Of the 20% infected individuals who remain asymptotic, we assume that the average duration of infectiousness is the same as symptomatic individuals, hence, 5.25 days. We denote the infectious contact rate parameter for Phase-II by β_2 and the infectious contact rate in other infectious stages is proportional to β_2.

Assumptions and parameter values are summarized in Table 6.1. Inserting these into (6.6), we get $R_0 = 9.725\beta_2$.

For planning purposes, a baseline clinical attack rate A_0 is often assumed to lie in a range before implementing any public health control measures, such as between 15% and 35% or between 10% and 50%.

For treatment, we assume that the treatment reduces μ_2 by 1 day, that is, $\mu_T = \varphi\mu_2 = 3$ days and $\varphi = 0.75$; also reduces β_2 such that $\beta_T = \kappa\beta_2 = 0.9\beta_2$.

We assume that the sojourn time in Phase-I of the symptomatic stage without race-to-treat follows an exponential distribution. From Table 6.1, $\mu_1 = 1$ gives $\gamma_{12} = 1$. Given the treatment rate γ_{1T}, the proportion of symptomatic individuals effectively treated is $\Phi = \frac{\gamma_{1T}}{1+\gamma_{1T}}$.

Inserting the parameter values from Table 6.1 plus the above additional assumptions into (6.9), we get $c = 0.68278$, which depends on the relative transmission rates β_A/β_2, β_0/β_2, β_1/β_2, but not the value of β_2. Thus,

$$R_c = R_0\,(1 - 0.68278\Phi) = R_0\left(1 - \frac{0.68278\gamma_{1T}}{1+\gamma_{1T}}\right).$$

We now assume, for planning purposes, a baseline clinical attack rate $A_0 = 40\%$. At $\theta = 0.8$ from Table 6.1, it corresponds to $R_0 = 1.386$. The maximum treatment usage is reached when

$$\Phi = \frac{\gamma_{1T}}{1+\gamma_{1T}} = \frac{1}{c}\left(1 - \frac{\log R_0}{R_0 - 1}\right) = 0.22611.$$

Equivalently, $\gamma_{1T} = 0.29217$. In other words, if symptomatic individuals are treated on average 3.4227 days since the time of onset, the treatment usage is at the maximum, at $\Omega_{\max} \approx 0.05$. At this rate, one expects to achieve the reduction of clinic attack rate from $A_0 = 40\%$ to

$$A_c(\Phi) = \theta\left(1 - \frac{1}{R_0}\right) = 0.2229 = 22.3\%.$$

The amount of treatment is $\Omega_{\max}$ multiplying the population size, and multiplying the appropriate unit of treatment. For example, if treatment usage per individual involves ten doses, then one needs to prepare a minimum stockpile in terms of the number of doses covering half of the population.

If the treatment is more timely, faster than on average 3.4227 days since the time of onset, it will require much less treatment at the population level because as Φ increases, $A_c(\Phi)$ decreases in a dramatic fashion and is nonlinear, as demonstrated in Fig. 6.1.

In order to visualize how treatment affects the transmission dynamics over time, we numerically solve the system of ordinary differential equations (6.15). Assuming all the sojourn times in each compartment are exponentially distributed, the parameters in (6.15) are assigned in agreement with those in Table 6.1.

- $\alpha_{LA} = 0.1$ and $\alpha_{L0} = 0.4$, so that $\mu_L = (\alpha_{LA} + \alpha_{L0})^{-1} = 2$ and $\theta = \alpha_{L0}/(\alpha_{LA} + \alpha_{L0}) = 0.8$;
- $\gamma_A = 1/5.25$ so that $\mu_A = 5.25$;
- $\gamma_{01} = 1/0.25$ so that $\mu_0 = 0.25$;
- $\gamma_{12} = \mu_1 = 1$ and $\gamma_2 = 1/\mu_2 = 0.25$;
- $\gamma_T = \frac{1}{\varphi}\gamma_2 = 1/3$, i.e., $\varphi = 0.75$;
- $\beta_T = 0.9\beta_2$, i.e., $\kappa = 0.9$;
- $\beta_2 = 0.14253$, so that $R_0 = 9.725\beta_2 = 1.3861$ which gives the baseline clinic attack rate $A_0 = 40\%$.

The results are shown in Fig. 6.6. The treatment does not only reduce the serologic attack rate (the final size) and the clinic attack rate, but also reduces delays to peak time of the prevalent numbers in all infected compartments and reduces the peak values of these prevalent numbers (i.e., the maximum case load). What is omitted in Fig. 6.6 is the case with race-to-treat more rapidly than on average 3.4227 days. In that scenario, one achieves more reductions of the maximum case load and the clinic attack rate with much less treatment usage.

6.3.5 Potential Extensions

All the discussions and demonstrations in the section are under the assumption that treatment is applied effectively on each treated individual. There are many different situations where this is not the case. We list some of these situations and point out directions of further developments.

Adjustment for Belated Treatment

In practice, treatment is not given to individuals in Phase-I of the symptomatic stage because the transition to Phase-II is not an observable event and the duration of Phase-I is random. Instead, public health organizations may issue a treatment

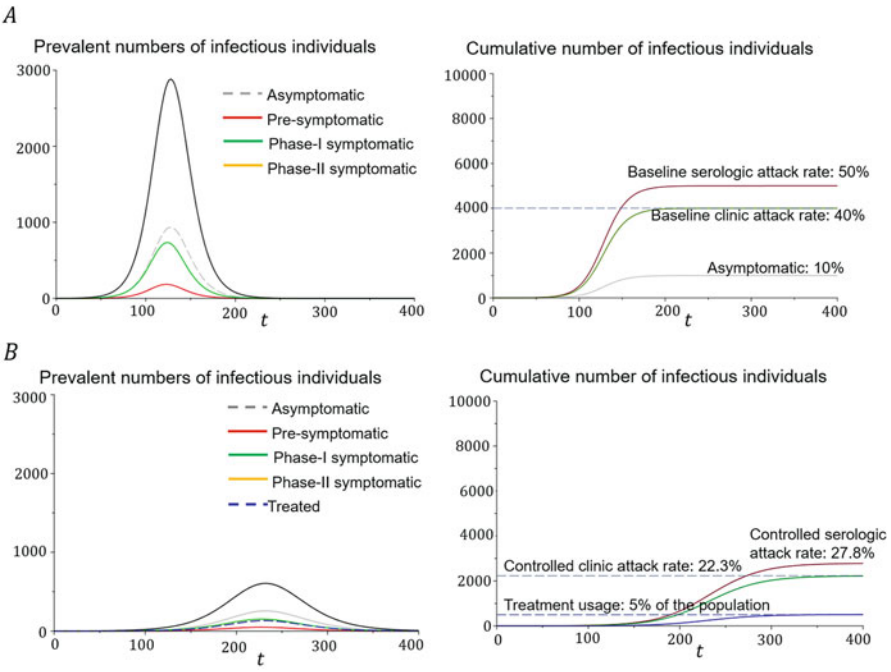


Fig. 6.6 Numeric solutions of (6.15) assuming $m = 100{,}000$, $S(0) = 99{,}999$, and $I_1(0) = 1$ at the baseline and with treatment at a rate corresponding to the maximum usage. (**a**) Baseline: no treatment. (**b**) Race-to-treat average 3.4227 days since the time of onset (maximum treatment usage)

guideline, such as a cut-off time C (e.g., 48 h since symptom onset), so that the treatment is no longer administered.

We have defined Φ as $\Phi = \Pr\{T_X \leq T_{12}\}$, where T_{12} is the time from onset of symptoms to the end of Phase-I (when the window of treatment opportunity expires) and T_X is the time from onset of symptoms to receiving treatment. There will always be a non-zero probability

$$\Pr\{T_X \leq C | T_X > T_{12}\} = \frac{\Pr\{T_{12} < T_X \leq C\}}{\Pr\{T_X > T_{12}\}}. \tag{6.16}$$

This is the probability that although the individual is no longer responsive to treatment, the treatment is administered regardless, following the guideline. The denominator of (6.16) is $\Pr\{T_X > T_{12}\} = 1 - \Phi$. We write the numerator $\Pr\{T_{12} < T_X \leq C\}$ as $\Pr\{T_{12} < T_X \leq C; \Phi\}$ to emphasize that this probability depends on the parameter Φ. If every clinical case with time-to-treatment $T_X \leq C$ receives treatment according to the guideline, one can calculate the "wasted usage" formula

$$A_c(\Phi) \times \Pr\{T_{12} < T_X \leq C; \Phi\}. \tag{6.17}$$

After adjustment for the belated treatment, the total treatment usage as a proportion of the population is $A_\Phi * (\Phi + \Pr\{T_{12} < T_x \le C; \Phi\})$ and better expressed by

$$\begin{aligned}\Omega(\Phi) &= \Phi * A_c(\Phi) * \left[1 + \frac{\Pr\{T_{12} < T_X \le C; \Phi\}}{\Phi}\right] \\ &= \Phi * A_c(\Phi) * \left[1 + \frac{\Pr\{T_{12} < T_X \le C\}}{\Pr\{T_X \le T_{12}\}}\right],\end{aligned} \tag{6.18}$$

where $\frac{\Pr\{T_{12}<T_x\le C\}}{\Pr\{T_x\le T_{12}\}}$ is the odds between those treated within the cut-off time C with an ineffective drug and those treated with an effective drug. These odds are determined by the distributions of T_{12} and T_X. As long as T_{12} and T_X are random, for any cut-off time C, $\Pr\{T_{12} < T_X \le C; \Phi\} > 0$. Therefore, there will always be clinical cases treated within the cut-off time but with an ineffective drug.

We assume that both T_{12} and T_X are exponentially distributed with rates γ_{12} and γ_T, respectively. Under the assumption that T_{12} and T_X are independent, $\Phi = \frac{\gamma_{1T}}{\gamma_{12}+\gamma_{1T}}$, $1 - \Phi = \frac{\gamma_{12}}{\rho+\alpha}$. Provided $\Phi < 1$, we have $\gamma_{1T} = \frac{\Phi}{1-\Phi}\gamma_{12}$, $\gamma_{12} + \gamma_{1T} = \frac{1}{1-\Phi}\gamma_{12}$, and $\frac{\gamma_{1T}}{\gamma_{12}} = \frac{\Phi}{1-\Phi}$. It can be shown that the proportion of clinical cases treated within $T_X \le C$ but not effective is the joint probability

$$\Pr\{T_{12} < T_X \le C\} = 1 - e^{-\frac{\Phi}{1-\Phi}\gamma_{12}C} - \Phi\left(1 - e^{-\frac{1}{1-\Phi}\gamma_{12}C}\right)$$

which is an increasing function of C with $\Pr\{T_{12} < T_X \le C\} \to 1 - \Phi$ as $C \to \infty$.

$$\begin{aligned}\Omega(\Phi) &= \Phi * A_c(\Phi) * \left[1 + \frac{\Pr\{T_{12} < T_X \le C; \Phi\}}{\Phi}\right] \\ &= \Phi * A_c(\Phi) * \left[\frac{1}{\Phi}\left(1 - e^{-\frac{\Phi}{1-\Phi}\gamma_{12}C}\right) + e^{-\frac{1}{1-\Phi}\gamma_{12}C}\right].\end{aligned} \tag{6.19}$$

Adjustment for False Positive Diagnoses

Suppose that a treatment is effective for a specific acute respiratory infectious diseases with influenza-like clinical symptoms. During such an outbreak, individuals may present influenza-like symptoms within the previous few hours and seek treatments. There is a probability that some individuals are not infected with the specific virus. This leads to the probability that some individuals are falsely diagnosed. For these individuals, the treatment is not effective.

We denote σ as the *specificity* of the clinic diagnosis, defined as

$$\sigma = \Pr\{\text{negative diagnosis} \mid \text{not infected with the specific virus}\}.$$

The probability of false positive diagnosis is $1-\sigma$. If one has knowledge of the specificity σ, given the background information of the incidence of non-influenza influenza-like illnesses, one will be able to adjust for the calculation of the usage. We assume that the probability of false negative diagnosis is negligible.

Previously, we used $A_c(\Phi)$ to denote the clinic attack rate for the "true infections" if a proportion Φ of symptomatic individuals is treated. We write it as a probability

$$A_c(\Phi) = \Pr\{\text{true infection}|\Phi\}.$$

Denote:

$$A^{(D)}(\Phi) = \Pr\{\text{positive diagnosis }|\Phi\}$$

which is the "inflated" clinic attack rate due to the presence of false positive diagnoses. Therefore,

$$\begin{aligned} A^{(D)}(\Phi) &= A_c(\Phi) \\ &\quad + (1-A_c(\Phi)) \times \Pr\{\text{positive diagnosis} \mid \text{not infected with the virus}\} \\ &= A_c(\Phi) * \left[1 + \frac{1-A_c(\Phi)}{A_c(\Phi)}(1-\sigma)\right]. \end{aligned} \tag{6.20}$$

The treatment usage is adjusted as

$$\Omega(\Phi) = \Phi \times A_{\Phi}^{(D)} = \Phi * A_c(\Phi) * \left[1 + \frac{1-A_c(\Phi)}{A_c(\Phi)}(1-\sigma)\right]. \tag{6.21}$$

If the diagnosis is 100% specific, then $\Omega(\Phi) = \Phi A_c(\Phi)$. The other extreme is $\sigma \to 0$, that is, all diagnoses are false, then $\Omega(\Phi) = \Phi$.

We assume $\gamma_{12} = 1$ and $C = 2$ (days) and combine (6.19) and (6.21) to update $\Omega(\Phi)$ Fig. 6.1. The results are displayed in Fig. 6.7.

Extending Φ as a Function of Time

We assume $\sigma = 100\%$ and for every treated individual, the treatment is effective.

The time-to-treatment T_X may improve or degenerate due to supply or other logistic issues in the health-care system. For a patient with symptoms at time t, one may extend the definition (6.5) as a function of t by a step-function:

$$\Phi(t) = \begin{cases} \Phi, & \text{if drug available at time } t; \\ 0, & \text{if drug unavailable at time } t. \end{cases} \tag{6.22}$$

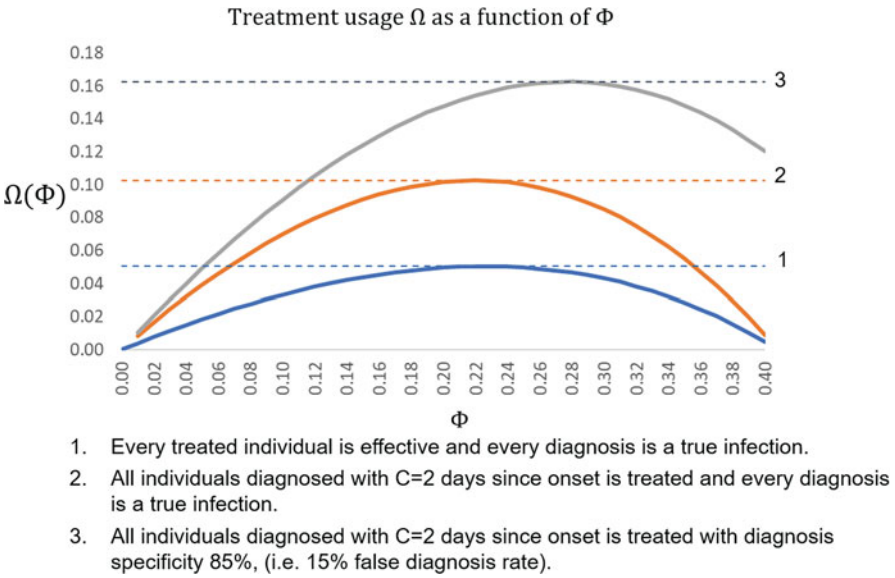


Fig. 6.7 Adjusted treatment usage in consideration of ineffective treatments for those treated within $C = 2$ days since onset and false diagnosis rate $1 - \sigma = 0.15$

1. If all patients are equally accessible to treatment and every patient accesses treatment following the same distribution for T_X at any given time t and the supply is sufficient to cover the entire duration of the outbreak, then Φ is the proportion of effectively treated patients among all patients. In this case, $A_c(\Phi)$ is an implicit function of Φ through (6.12).
2. If the supply is insufficient to cover the entire duration of the outbreak, Φ is the proportion of patients who receive treatment when the supply is available and are effectively treated. If the supply runs out at time t_1, before the end of an outbreak, then $\Phi(t) = \Phi$ if $t \leq t_1$. Let A_1 be the proportion of the population that becomes symptomatic by t_1, and a proportion Φ of them are treated. Let A^* be the proportion of the population that eventually becomes clinically ill. Then $\Phi^* = \frac{A_1}{A^*}\Phi < \Phi$ is the proportion of patients who are effectively treated during the entire outbreak.

For saving the volume of this book, we do not present detailed analyses for this situation, but to point out that (Exercise)

$$\begin{aligned} \Phi A_1 = \Phi^* A^* = & \text{ constant} \\ & = \text{effectively treated patients as proportion of population} \\ & = \text{supply limit of treatment.} \end{aligned} \tag{6.23}$$

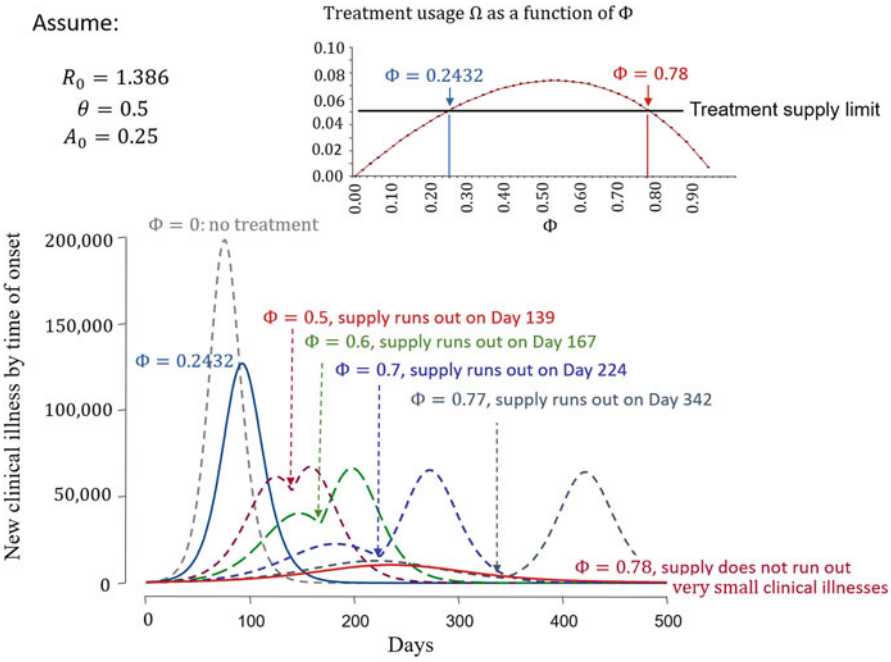


Fig. 6.8 An illustration of treatment supply depletion scenarios and how they affect the time course of the transmission dynamics

Figure 6.8, with slightly different parameter values from that used before, illustrates such a situation that if within the range $0.2423 \leq \Phi < 0.78$, the supply runs out. Using the ordinary differential equations (6.15), $\Phi = \frac{\gamma_{1T}}{\gamma_{12}+\gamma_{1T}}$. It implies that, if the race-to-treatment rate γ_{1T} falls in the range $0.31978\gamma_{12} \leq \gamma_{1T} < 3.5455\gamma_{12}$, the supply runs out at certain point of time. At the population level, the treatment delays the peak of the incidence number of new clinical cases and reduces the peak value. However, at the time when the supply runs out, there is a rebound of the incidence numbers. Figure 6.9 shows that the cumulative number by the end of the outbreak A^* remains the same when $0.2423 \leq \Phi < 0.78$, and $A^* = A_c(\Phi) = 0.21$ when $\Phi = 0.2423$. However, in a race-to-treat scenario that achieves $\Phi >= 0.78\ldots$, not only the supply does not run out, but also the clinic attack rate significantly reduces to $A_c(\Phi) = 0.06$.

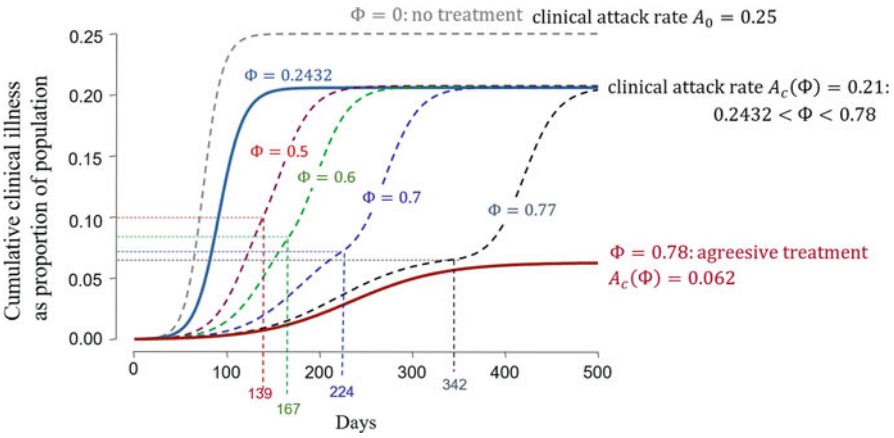


Fig. 6.9 An illustration of treatment supply running out scenarios and how they affect the cumulative number of clinical cases

6.4 Effects of the Variability of the Latent and Infectious Periods on Certain Control Measures

In a simpler structure than the model in Fig. 6.2, we assume that individuals in Compartment E receive a "treatment" with rate ψ. If treated, they are not infectious. A typical treatment is to isolate exposed individuals during their latent periods. Individuals in Compartment I receive a "treatment" with rate ϕ. If treated, they are not infectious. Such a treatment may be pharmaceutical or non-pharmaceutical. One of the treatments is to isolate infected individuals during their infectious periods. On the other hand, we also generalize the model assuming that the latent periods follow a distribution with p.d.f. $f_E(x)$ and hazard function $h_E(x)$; the infectious periods follow a distribution with p.d.f. $f_I(x)$ and hazard function $h_I(x)$. This model is illustrated in Fig. 6.10.

When $\psi = \phi = 0$, this model is an SEIR model represented by the integro-differential equations (5.61). In that case, the basic reproduction number is $R_0 = \beta\mu_I$ and an instantaneous transmission in the population is βSI (scaled by population size).

When $\psi > 0$, the instantaneous rate of an individual in Compartment E either becoming infectious or being isolated is $\psi + h_E(x)$. The survival function for individuals staying in the E-compartment is $\exp\left\{-\int_0^x [\psi + h_E(x)]\,dt\right\}$. The corresponding p.d.f. is

$$\begin{aligned} f(x|\psi) &= [\psi + h_E(x)] \exp\left\{-\int_0^x [\psi + h_E(x)]\,dt\right\} \\ &= \psi e^{-\psi x} \overline{F}_E(x) + e^{-\psi x} f_E(x), \end{aligned}$$

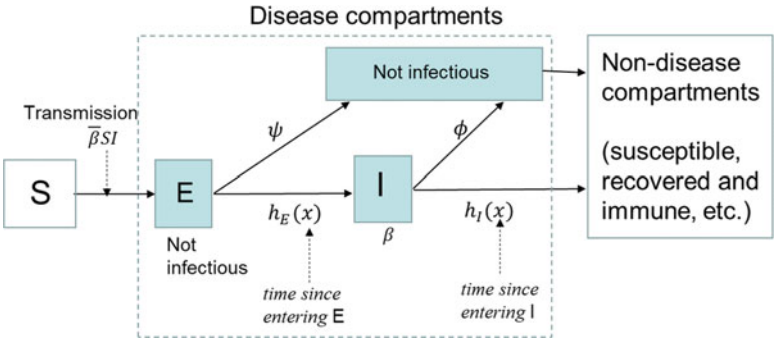


Fig. 6.10 Illustration of isolation of infected but not yet infectious individuals with rate ψ, and isolation of individuals who are infectious with rate ϕ

where $\overline{F}_E(x) = \exp\left\{-\int_0^x h_E(x)dt\right\}$ is the survival function of the latent period. Therefore,

$$\begin{aligned}\int_0^\infty f(x|\psi)dx &= \psi\int_0^\infty e^{-\psi x}\overline{F}_E(x)dx + \int_0^\infty e^{-\psi x}f_E(x)dx \\ &= \psi L[\overline{F}_E](\psi) + L[f_E](\psi) = 1.\end{aligned}$$

The second term $L[f_E](\psi)$ is the proportion of latent individuals that eventually escape from being isolated. These individuals will proceed to Compartment I. Therefore, the instantaneous transmission rate in the population is modified to $\overline{\beta}SI$, where $\overline{\beta} = \beta L[f_E](\psi) < \beta$.

The mean infectious period is $\mu_I = \int_0^\infty \overline{F}_I(x)dx$. When $\phi > 0$, the rate of exiting Compartment I is accelerated to $\phi + h_I(x)$. The survival function of an individual remaining in the infectious stage becomes

$$e^{-\phi x}\overline{F}_I(x) = \exp\left\{-\int_0^x [\phi + h_I(x)]\,dt\right\} < \overline{F}_I(x).$$

Then $L[\overline{F}_I](\phi) = \int_0^\infty e^{-\phi x}\overline{F}_I(x)dx$ has the same meaning as the average sojourn time in Compartment I corresponding to the survival function $e^{-\phi x}\overline{F}_I(x)$. In other words, the average sojourn time in Compartment I has been shortened from μ_I to $\overline{\mu}_I = L[\overline{F}_I](\phi)$ by isolation.

Together, we write the controlled reproduction number as

$$\begin{aligned}R_c(\psi,\phi) &= \overline{\beta}\overline{\mu}_I \\ &= \beta L[f_E](\psi)L[\overline{F}_I](\phi).\end{aligned} \tag{6.24}$$

In the special case when both the latent and infectious periods are exponentially distributed with $h_E(x) = \alpha = \mu_E^{-1}$, $h_I(x) = \gamma = \mu_I^{-1}$, the Laplace transforms $L[f_E](\psi)$ and $L[\overline{F}_I](\phi)$ are explicit. The above can be written as

$$R_c(\psi, \phi) = \frac{\beta\alpha}{(\alpha + \psi)(\gamma + \phi)} = \frac{\beta\mu_I}{(1 + \psi\mu_E)(1 + \phi\mu_I)} = \frac{1}{(1 + \psi\mu_E)(1 + \phi\mu_I)} R_0. \tag{6.25}$$

The expression (6.25) is in agreement with the non-negative eigenvalue of the next generation matrix $\rho(FV^{-1})$, in which,

$$F = \begin{pmatrix} 0 & \beta \\ 0 & 0 \end{pmatrix}, \; V = \begin{pmatrix} \alpha + \psi & 0 \\ -\alpha & \gamma + \phi \end{pmatrix} \text{ and } FV^{-1} = \begin{pmatrix} \frac{\beta\alpha}{(\alpha+\psi)(\gamma+\phi)} & \frac{\beta}{\alpha+\psi} \\ 0 & 0 \end{pmatrix}.$$

As for generally distributed latent and infectious periods, the formulation of $R_c(\psi, \phi)$ given by (6.24), using the arguments presented in Sect. 2.5.2, we arrive at the following two statements.

Proposition 27 *Given the mean latent period μ_E and the treatment (e.g., quarantine) rate $\psi > 0$, the less variable the latent period according to the convex order, the more effective the treatment.*

Proposition 28 *Given the mean infectious period μ_I and the treatment rate ϕ, the less variable the infectious period according to the convex order, the less effective the treatment.*

We further draw parallels with discussions in Sect. 4.3.3, where the variabilities of the latent and infectious periods, characterized by Laplace transforms, are related to the intrinsic (exponential) growth rate r, expressed by (4.43) as $\beta L[f_E](r)L[\overline{F}_I](r) = 1$.

Thus, given the same mean latent period μ_E, small variability of the latent periods is preferable in the sense of a slower initial growth and a more effective treatment measure, such as quarantine to be implemented during the latent period.

On the other hand, given the same mean latent period μ_I, large variability of the infectious periods is preferable in the sense of a slower initial growth and a more effective treatment measure to be implemented during the infectious period.

6.5 Unobservable Heterogeneity in Treatment Rates on Effectiveness

In (6.24), it is assumed that the treatments are 100% effective. Treated individuals are no longer infectious. Yan and Feng (2010) provided some extensions for imperfect quarantine and imperfect isolation. The model in Fig. 6.2 can be also used for

imperfect treatments by modeling the reduction of the infectious contact parameter as $\sigma\beta$ and $\varphi\beta$. These extended models may be used to "mechanically" evaluate how "leakages" of these control measures affect the controlled reproduction number in a quantitative way.

In this section, we discuss unobservable (even un-quantifiable) nonadherence of these treatment measures among individuals and how frailty in treatment rates affects the controlled reproduction number and the final size in a qualitative way. Unobservable (and even un-quantifiable) heterogeneity in treatment could arise in many situations, such as nonadherence (e.g., condom use), "leakage" (e.g., imperfect quarantine or isolation), or, in a prophylactic intervention, individuals may not take the prescribed dose of the medication provided.

We simplify the discussion using a single parameter $\psi = \phi$ so that (6.24) is $R_c(\phi) = \beta L[f_E](\phi)L[\overline{F}_I](\phi)$. We define the p.d.f. for the equilibrium distribution $f_W(x) = \overline{F}_I(x)/\mu_I$ as previously discussed in Sect. 4.3.3, corresponding to a random variable $W > 0$. Assuming independency, the p.d.f. for $G = T_E + W$, $f_G(x)$, is defined by the convolution between $f_E(x)$ and $f_W(x)$. Its Laplace transform is the product $L[f_G](s) = L[f_E](s)L[f_W](s)$, $s > 0$, where $L[f_W](s) = \frac{1}{\mu_I}L[\overline{F}_I](s)$. Thus

$$R_c(\phi) = R_0 \int_0^\infty e^{-\phi x} f_G(x)dx = R_0 L[f_G](\phi). \tag{6.26}$$

In Sect. 2.6.1, the frailty model was introduced for unobservable heterogeneity.

In the current context, the intervention is associated with a rate $\phi > 0$ under the idealized situation. We associate this rate with a baseline hazard function $h_0(x) = \phi$, $H_0(x) = \phi x$, and $\overline{F}_0(x) = e^{-\phi x}$, corresponding to time x at intervention. In the idealized situation, R_c is a fraction of R_0 and this fraction is $L[f_G](\phi) = \int_0^\infty e^{-\phi x} f_G(x)dx$. The threshold condition with respect to (6.26) is

$$R_0 L[f_G](\rho) = 1, \tag{6.27}$$

so that $R_c(\phi) \leq 1$ when $\phi \geq \rho$.

6.5.1 *The Controlled Reproduction Number in the Presence of Frailty*

In the presence of frailty, that is unobservable heterogeneity in implementation and adherence of the treatment measures, we consider the frailty model $h(x|z) = z\phi$, $z > 0$. In this case, z is the *frailty parameter* and is assumed to be random with mean value $E(z) = 1$ and p.d.f. $\xi(z)$. From (2.39) in Chap. 2, we obtain

$$\overline{F}^{(frailty)}(x) = \int_0^\infty e^{-z\phi x}\xi(z)dz = L[\xi](\phi x).$$

The controlled reproduction number is with a non-degenerated p.d.f. $\xi(z)$, replacing $e^{-\phi x}$ with $L[\xi](\phi x) > e^{-\phi x}$, the following inequality is established:

$$\begin{array}{ccccc} R_0 \int_0^\infty e^{-\phi x} f_G(x)dx & < & R_0 \int_0^\infty L[\xi](\phi x) f_G(x)dx & < & R_0 \\ \| & & \| & & \\ R_c(\phi) & & R_v(\phi) & & \end{array}, \tag{6.28}$$

where $R_c(\phi)$ reflects the efficacy in the idealized situation and $R_v(\phi)$ reflects the effectiveness in the presence of frailty when the intervention is applied in a large population.

The following identity, first proven in Goldstein (1932) is useful:

$$\begin{aligned} \int_0^\infty L[\xi](\phi x) f_G(x)dx &= \int_0^\infty \int_0^\infty e^{-\phi x z} \xi(z) f_G(x)dzdx \\ &= \int_0^\infty L[f_G](\phi z)\xi(z)dz. \end{aligned}$$

It leads to

$$R_v(\phi) = R_0 \int_0^\infty L[\xi](\phi x) f_G(x)dx = R_0 \int_0^\infty \mathcal{L}[f_G](\phi z)\xi(z)dz. \tag{6.29}$$

Empirical wisdom has told us that the control measure is most effective if applied homogeneously across all individuals. This is mathematically expressed by (6.28). It has also vaguely and intuitively led to the belief that the more variable the adherence, the less effective the control measure. It can be shown that $L[f_G](z)$ is log-convex with respect to z. Using the convex ordering in Sect. 2.5.2, the more variable the frailty z is according to convex order, the larger the value of $R_v(\phi)$. This insight sharpens the vague intuitive notion.

One of the most illustrative models is the Gamma distribution for $\xi(z)$ with $E[z] = 1, var[z] = v$. The p.d.f. is

$$\xi(z; v) = \frac{1/v}{\Gamma(1/v)} (z/v)^{1/v-1} e^{-z/v}. \tag{6.30}$$

The Laplace transform, from (2.41), is $L[\xi](s) = (1 + sv)^{-1/v}$. Thus, $\overline{F}^{(frailty)}(x) = (1 + \phi x v)^{-1/v}$. For any x, $\overline{F}^{(frailty)}(x)$ is an increasing function of v. The variance v ranks the frailty variable z according to stochastic order. The limiting cases are $\lim_{v\to 0} (1 + \phi x v)^{-1/v} = e^{-\phi x}$ and $\lim_{v\to\infty} (1 + \phi x v)^{-1/v} = 1$.

Figure 6.11 illustrates the shapes of $\xi(z; v)$ and $\overline{F}^{(frailty)}(x)$. The shape of $\xi(z; v)$ changes dramatically at $v = 1$. When $0 < v < 1$, z is "more or less" concentrated at $E[z] = 1$. One may loosely call this "almost homogeneous" in the control measure with some mild variability. When $v > 1$, it is "highly heterogeneous." According to (6.28),

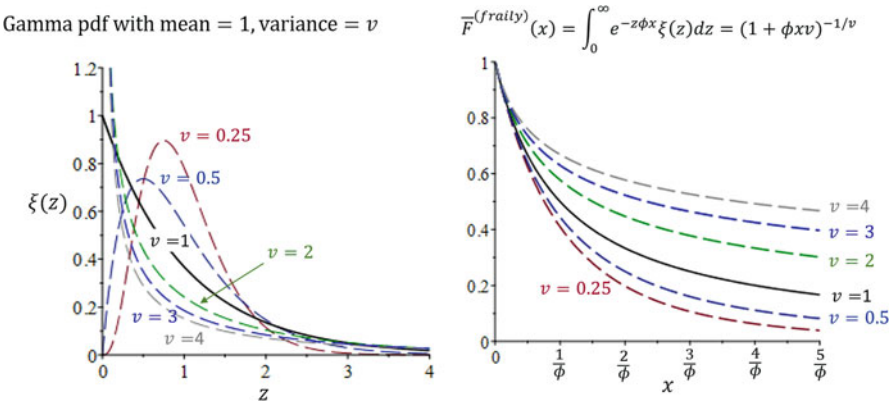


Fig. 6.11 Shapes of $\xi(z)$ and $\overline{F}^{(frailty)}(x) = (1+\phi x v)^{-1/v}$

$$R_v(\phi) = R_0 \int_0^\infty (1+\phi x v)^{-1/v} f_G(x)dx. \tag{6.31}$$

In this case, $\lim_{v\to 0} R_v(\phi) = R_c(\phi)$ and $\lim_{v\to\infty} R_v(\phi) = R_0$. Given the control parameter ϕ, the variance parameter v ranks the values of $R_v(\phi)$ so that the more variable the adherence of the control measure, the less effective it is.

6.5.2 *Invariance to the Time Scale of the Natural History and Robustness to Assumptions in $f_G(x)$*

Assumptions in the p.d.f. $f_G(x)$ include whether there is a latent period, as well as the distributions of the latent and the infectious periods. It usually involves a scale parameter of the time $\lambda > 0$. Since the control parameter ϕ is a rate, it is also scaled by the same parameter λ. We show that the value $R_v(\phi)$ is invariant with respect to λ.

Let $y = \lambda x$ and $f_G(x)$ be represented as $f_G(x;\lambda)$, then $f_G^*(y) = f_G(y;1) = f_G(x;\lambda)/\lambda$ is the standardized p.d.f at $\lambda = 1$. Meanwhile, the Laplace transform of the p.d.f. of a non-negative random variable is a survival function, arising from the mixture of an exponential survival function with the p.d.f. as the mixing distribution (Marshall and Olkin 2007), then $L[\xi](x)$ is a survival function satisfying $L[\xi](y) = L[\xi](\lambda x)$. Therefore,

$$R_v(\phi) = R_0 \int_0^\infty L[\xi](\phi y) f_G^*(y)dy = R_0 \int_0^\infty L[\xi](\lambda\phi x) f_G(x;\lambda)dx, \tag{6.32}$$

$$R_c(\phi) = R_0 \int_0^\infty e^{-\phi y} f_G^*(y)dy = R_0 \int_0^\infty e^{-\lambda\phi x} f_G(x;\lambda)dx. \tag{6.33}$$

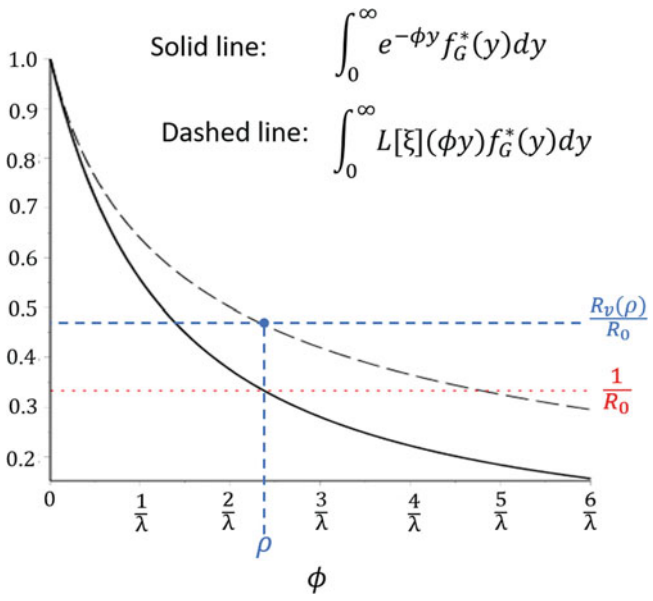


Fig. 6.12 Schematic presentation of $R_v(\phi)/R_0$ and $R_c(\phi)/R_0$ as two survival functions standardized by the scale parameter λ

These values are invariant to the scale parameter λ. Both $R_v(\phi)$ and $R_c(\phi)$, divided by R_0, are plotted in Fig. 6.12, with time scaled by λ.

Let ρ be the control threshold under the idealized condition satisfying (6.27) $R_0 L[f_G](\rho) = 1$. Then $R_v(\phi)$ calculated at $\phi = \rho$ is

$$R_v(\rho) = R_0 \int_0^\infty L[\xi](\rho y) f_G^*(y)dy = R_0 \int_0^\infty L[f_G^*](\rho z)\xi(z)dz,$$

where the second equation is from (6.29). The Laplace transform $L[f_G^*](s)$ can be also regarded as a survival function. The threshold condition is $R_0 L[f_G^*](\rho) = 1$. It is equivalent to say that ρ is the $\left(1 - R_0^{-1}\right)^{th}$ percentile corresponding to the survival function $L[f_G^*](s)$. Together we have

$$R_v(\rho) = \int_0^\infty \frac{L[f_G^*](\rho z)}{L[f_G^*](\rho)}\xi(z)dz, \tag{6.34}$$

where $L[f_G^*](\rho z)$ is $L[f_G^*](z)$ scaled by ρ. The ratio $L[f_G^*](\rho z)/L[f_G^*](\rho)$ is independent of λ. This finding is important. Given the threshold ρ under the idealized situation, the value of $R_v(\rho) > 1$ is invariant with respect to the time scale of disease progression, regardless of whether the disease is acute (e.g., measured in days like influenza) or chronic (e.g., HIV, viral hepatitis). Given the model for

frailty, $R_v(\rho)$ only depends on R_0 and f_G. The latter includes assumptions on the existence of the latent period and the distributions of the latent and the infectious periods.

Yan (2018) further shows, the ratio $L[f_G^*](\rho z)/L[f_G^*](\rho)$ is quite robust with respect to the assumed distributions for f_G^* and can be approximated by a log-convex function of z with a simple Pareto form $R_0/[1+(R_0-1)z]$ that only depends on R_0. This robustness is examined by numerical calculations under the frailty model $\xi(z;v)$ given by (6.30). There are 16 distribution models for f_G. The first four models assume Gamma distributed infectious periods (including exponentially distributed infectious periods and constant infectious periods as special cases) without a latent period. The next 12 models include latent periods. The mean latent period is parameterized as $\mu_E = l\mu_I$, $l \geq 0$ and μ_I is the mean infectious period with the relative length of the average latent periods to the average length of the infectious periods $l = 0.5$, 1, and 2. These are combined with four models for the latent periods and infectious periods convolutions: (1) constant latent period + constant infectious period, (2) exponentially distributed latent period + exponentially distributed infectious period, (3) constant latent period + exponentially distributed infectious period, and (4) exponentially distributed latent period + constant infectious period. Furthermore, numerical calculations are under two levels of R_0. Thus, there are a total of 32 different results.

Without going through the details, we summarize the numeric results in Yan (2018) as Fig. 6.13. In each of these computations, ρ is the threshold control parameter and is a function of (R_0, f_G) and depends on λ. It is calculated separately. Then $R_v(\rho)$ is evaluated at $\phi = \rho$.

In the left panel, $R_v(\rho)$ is calculated at v in the range from 0 to 4 by increments of 0.25. For each v, there are 16 points (in black) corresponding to $R_0 = 3$ and 16 points (in blue) corresponding to $R_0 = 2$. The trends representing $R_v(\rho)$ as functions of v are calculated as average and plotted as lines. It shows that

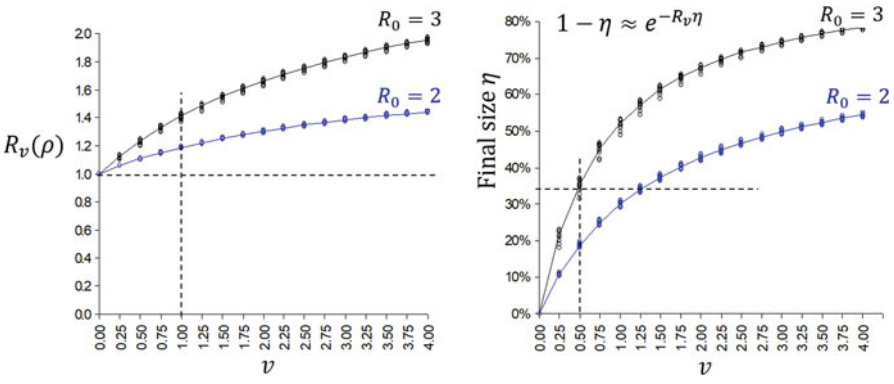


Fig. 6.13 The left panel shows $R_v(\rho)$ and the right panel shows the corresponding final sizes. At each level of R_0, there are 16 points, plotted at each v

assumptions on the structure (e.g., with or without latent period), on the types of distributions of these periods as well as the relative lengths between the latent and the infectious periods have little influence on $R_v(\rho)$. The predominant parameters are R_0 and v. For example: If $R_0 = 3$, a control measure at intervention rate ρ that in theory could have controlled the epidemic (i.e., $R_c(\rho) = 1$) may result in an outbreak as if manifested by a reproduction number $R_v \approx 1.4$.

The right panel of Fig. 6.13 is analogous to the left panel, representing the final size η through the approximate final size equation $1 - \eta = \exp(-R_v\eta)$. The final size, as a measure, is only applicable in certain epidemics, typically without replacement of the susceptible population and recovered individuals have immunity. However, unlike $R_v(\rho)$, which is mainly a theoretical parameter, the final size (where applicable) is also an observable quantity, also known as the *infection attack rate*. Once again, assumptions on the structure (e.g., with or without latent period), on the types of distributions of these periods as well as the relative lengths between the latent and the infectious periods have little influence on η.

Remark It is well known that univariate frailty models are not identifiable from the survival information alone. In the current context, the basic reproduction number R_0 is a theoretical value. The observed (or estimated) parameters during or after the outbreak, such as R_v or η, cannot identify R_0, ϕ, and v. Figure 6.13 shows that, if a control measure with rate at a threshold $\phi = \rho$ has been proven (in theory) to get the epidemic under control, one may still observe an outbreak that ends with final size about 35% of the population. It could arise from an outbreak with $R_0 = 3$ and the control measure is implemented imperfectly but good adherence at $v = 0.5$, or from an outbreak with $R_0 = 2$ and the control measure implemented with poor adherence at $v = 1.25$, or from an outbreak with $R_0 = 1.23$ with no intervention at all. An outbreak with infection attack rate 35% is nonetheless a large outbreak. In the absence of knowledge of R_0 along with unobservable frailty, it leaves impression as if the control measure that looks good on paper "does not work at all" in practice.

The non-identifiability problem poses challenges in the design of intervention studies at the population level (Cobelli and Romanin-Jacur 1976). There are confounding factors that hinder the ability to distinguish the "pure" impact of the intervention without the distorting influence of compliance on the effectiveness, since some level of non-compliance is likely. Both the pure impact (efficacy) and population level effectiveness are important objectives in intervention studies.

6.6 Problems and Supplements

6.1 Prove the expressions of the reproduction numbers in structures (a)–(c) of Fig. 6.3:

(a) Show that, if the infectious period is staged in a serial manner $I^{(1)} \to I^{(2)} \to \cdots \to I^{(k)}$ and the intensity function $\beta(x)$ of the infectious contact

process $\{K(x) : x \geq 0\}$ is defined as piecewise constants $\beta(x) = \beta_j$, for $x \in I^{(j)}$, then the expected total number of secondary infections produced by a typical infected individual during its entire infectious period is the sum $\sum_{j=1}^{k} \beta_j \mu_j$, where μ_j is the average duration in stage $I^{(j)}$.

(b) We assume that there is a single infected stage I with a constant infectious contact rate β, an infected individual may exit the infectious stage in k competing events (e.g., isolation, deaths, recovery, etc.) with $h_j(x)$ being the hazard function of the exiting event of type j, $j = 1, \ldots, k$. We assume that these exiting events are independent from each other. Show that, the expected total number of secondary infections produced by a typical infected individual during its entire infectious period is $\beta \int_0^\infty \exp\left(-\int_0^t \sum_{j=1}^{k} h_j(x)dx\right) dt$. In particular, if $h_j(x) = \gamma_j$, the expected total number of secondary infections is $\beta \left(\sum_{j=1}^{k} \gamma_j\right)^{-1}$.

(c) Infected individuals are composed of different types. Each type is associated with an infectious period with average duration μ_j and infectious contact rate β_j, $= 1, \ldots, k$. A typical susceptible individual has a probability ψ_j to be in contact with infected individuals of type j. Show that, given a typical infected individual (without knowing its type), the expected total number of secondary infections during its entire infectious period is the weighted average $\sum_{j=1}^{k} \psi_j \beta_j \mu_j$.

6.2 You are tasked to design and parameterize the following two models of infectious disease transmission:

(a) Model 1. Starting from the simple SEIR model, incorporate a class of infected individuals that are mildly symptomatic and infectious (Compartment M) before they progress to a fully symptomatic and infectious class. In this model, all latent individuals progress to the M-compartment before they progress to the I-compartment. Mildly symptomatic and infectious individuals have a reduced infectiousness relative to the fully infectious and symptomatic I-individuals. The average time of infected individuals in M-compartment is not necessarily the same as that of infected individuals in the I-compartment.

(i) Draw the diagram of the compartmental model including the rates of the flows.
(ii) Provide a table that includes the model parameter symbols, their definition, and units.
(iii) Provide the expression for the basic reproduction number R_0 for this model.
(iv) Generate and explain a few epidemic simulations (temporal progression of M- and I-individuals) as you vary key parameters governing the contribution of M-individuals.
(v) How does the value of R_0 change as the infectiousness of M-individuals increases? Illustrate with an example.

(b) Model 2. Starting from the simple SEIR model, incorporate a class of hospitalized and infectious individuals (Compartment H). In this model, all latent individuals progress to the symptomatic and infectious Compartment I. I-compartment individuals are either hospitalized or recover without being hospitalized. Hospitalized and infectious individuals recover at a different rate than that of I-individuals. Hospitalized infected individuals have a reduced infectiousness relative to the fully infectious and symptomatic I-individuals because of infection control measures in hospitals.

(i) Draw the diagram of the compartmental model including the rates of the flows.
(ii) Provide a table that includes the model parameter symbols, their definition, and units.
(iii) Because of infection control measures in hospitals, we denote the reproduction number as R_c. Provide the expression for the basic reproduction number R_c for this model.
(iv) Generate and explain a few epidemic simulations (temporal progression of I- and H-individuals) as you vary key parameters governing the contribution of H-individuals.
(v) How does the value of R_c change as the infectiousness of H-individuals decreases? Illustrate with an example.

6.3 Design and parameterize the following model of infectious disease transmission. Starting from the simple SEIR model, incorporate a class of hospitalized and infectious individuals (Compartment H). In this model, all latent individuals progress to the symptomatic and infectious Compartment I. I-compartment individuals are either hospitalized or recover without being hospitalized. Hospitalized and infectious individuals recover at a different rate than that of I-individuals. Hospitalized infected individuals have a reduced infectiousness relative to the fully infectious and symptomatic I-individuals because of infection control measures in hospitals.

(a) Draw the diagram of the compartmental model including the rates of the flows.
(b) Provide a table that includes the model parameter symbols, their definition, and units.
(c) Provide the expression for the reproduction number R_c for this model.
(d) Generate and explain a few epidemic simulations (temporal progression of I- and H-individuals) as you vary key parameters governing the contribution of H-individuals.
(e) How does the value of R_c change as the infectiousness of H-individuals decreases? Illustrate with an example.

6.4 In models with SEI structure where the latent periods are exponentially distributed with rate α and the infectious periods are exponentially distributed with rate γ, (4.51) in Chap. 4 gives the expression $R_0 = \frac{(r+\alpha)(r+\gamma)}{\alpha\gamma}$, where $r > 0$ is the initial growth rate; (5.70) in Chap. 5 gives the expression $R_0 = \frac{\beta\alpha}{(\alpha+\omega)(\gamma+\omega)}$, where $\omega > 0$ is the rate of outflow of individuals in the E-compartment and the I-compartment (while holding the total population size constant); and (6.25) gives the controlled reproduction number $R_c(\psi, \phi) = \frac{\beta\alpha}{(\alpha+\psi)(\gamma+\phi)}$, where $\psi, \phi > 0$ are the rate of isolation of individuals in the E-compartment and the I-compartment, respectively.

(a) Discuss these similarities and their relationships (if there is any).
(b) Compare these expressions when the latent periods and the infectious periods are not exponentially distributed.

6.5 Compare $\overline{F}^{(frailty)}(x) = (1 + \phi x v)^{-1/v}$ in Fig. 6.11 with the sub-exponential growth function $C_d(t) = i_0(1 + rvt)^{\frac{1}{v}}$ given by (4.66) in Chap. 4. Discuss their relationships (if there is any) and connect with the discussions in Problem 6.4.

Chapter 7
Some Statistical Issues

7.1 Models and Parameters

All the models presented in the previous chapters are parametric. They belong to different types and serve different purposes.

Probabilistic models are introduced and discussed in Chaps. 2 and 3. Part of their role in the study of infectious diseases is to formulate assumptions regarding disease progression within an infected host. Examples include the parametric life distribution models characterized by shapes of hazard functions in Chap. 2, as well as the process of infectious contacts from the viewpoint of an infected individual through counting processes in Chap. 3. These probabilistic models, especially the distribution models for random counts and counting processes, are also statistical models that take into account the data-generating process that will be further discussed in this chapter.

Phenomenological population models, both in stochastic and in deterministic frameworks, are the main focus in Chaps. 4–6. They are based on conceptual assumptions regarding the population and the interface among the agent, the host, and the environment. Section 4.4 has provided some detailed discussions regarding these assumptions applied to the initial phase of an outbreak. However, phenomenological population models often carry tacit assumptions at the individual level. For example, deterministic transmission models that can be represented by systems of ordinary differential equations implicitly assume that infected individuals pass each stage of the natural history with exponentially distributed durations. The stochastic models with a Markov structure make the same assumptions.

Hidden assumptions at the level of individuals determine certain crucial epidemiological characteristics as well as the effectiveness of certain disease control measures in a phenomenological way. For example, the relationship between the distributions of the infectious periods and the probabilities of invasion and extinction (Sect. 4.2); the relationship between the distributions of the infectious periods (as

P. Yan, G. Chowell, *Quantitative Methods for Investigating Infectious Disease Outbreaks*, Texts in Applied Mathematics 70,
https://doi.org/10.1007/978-3-030-21923-9_7

well as the latent periods) and the initial growth (4.43); the assumption of the exponentially distributed infectious periods in the SIR model as a primary feature with respect to the peak prevalence of infected individuals and some important preserved quantities (Sect. 5.3.3); the prevalence of individuals in each class of the SEIRS model (5.67) with expressions of $[x(\infty), \epsilon(\infty), y(\infty), z(\infty)]$ under endemic equilibrium and their special cases such as (5.71) in Sect. 5.5.2; the expression of the controlled reproduction number R_c (6.4) in Sect. 6.2.1; and the effects of the distributions of the latent and infectious periods on certain control measures (Sect. 6.4).

There is also a different kind of phenomenological population models with relatively simple forms and without assumptions regarding the agent, the host, the environment, and their interactions in the population. They describe data in a phenomenological way and are often useful to answer some key public health questions during an outbreak investigation. These are the growth curve models. Later in Chap. 8, we shall see some applications of these models to real outbreak data.

The most important function of models are to order our thoughts and to sharpen vague intuitive notions. Whatever their types are, they are connected to the formulation of the research questions and objectives of the subject matter. Different questions and objectives require differently formulated models.

Even the same model can be parameterized differently for different research questions. For example, a simple logistic function has many different expressions such as (4.59) and (5.14). The logistic function may be written as

$$F(t) = \frac{K}{1 + e^{-\rho(t-\alpha)}}.$$

This function can be considered as a *descriptive* model to fit disease incidence data, either cumulatively, or incidence numbers of new occurrences (e.g., daily, weekly, etc.) The three parameters (ρ, α, K) are directly and indirectly connected to important public health questions during a disease outbreak, such as: "when do we expect the outbreak to peak?", "how long do we expect the outbreak to last?" and "how big is the outbreak going to be?" This is because the parameter ρ represents the initial growth of a sigmoid growth function; α represents the inflexion point at which $F'(t)$ arrives at the maximum value as well as when the outbreak is at its midpoint $F(\alpha) = K/2$; and K represents the asymptotic limit $K = \lim_{t\to\infty} F(t)$. However, this model provides little understanding of the process such as the disease transmission process.

On the other hand, the logistic function expressed as

$$F(t) = \frac{mi_0\,(R_0 - 1)}{i_0 R_0 + (m\,(R_0 - 1) - i_0 R_0)\,e^{-(R_0-1)\gamma t}}$$

in (5.14) has four parameters: (R_0, γ, m) and the initial condition $i_0 = F(0)$. In fact, it is a *mechanistic* model because all these parameters are associated with scientific hypotheses about the transmission dynamics with the SIS structure, such as the

basic reproduction number R_0, the mean duration of the (exponentially distributed) infectious period γ^{-1}, and the population size m. This expression is used to describe the prevalence of the number of individuals who are "currently" infectious at time t. Statistically speaking, only three out of the four parameters are identifiable from data, because $\rho = (R_0 - 1)\gamma$ can be regarded as a single scale parameter of time t as the initial growth rate and $K = m\,(1 - 1/R_0)$ is the asymptotic limit. The identifiable parameters are (ρ, i_0, K).

7.1.1 Statistical Models

In their book *Generalized Linear Models,* McCullagh and Nelder (1983) partitioned the model into three components: (1) the random component, (2) the systematic component, and (3) the link function. We adopt the same terminology when combining statistical models with disease transmission models for analyses of outbreak investigation data.

The random component models the data-generating process through probability distributions, denoted here as $f(y; \theta)$. They are the foundation of statistical inference for estimation and testing hypotheses. The discrete distributions and counting processes in Chap. 3 are important statistical models to represent the data-generating process of random counts as realizations of the underlying stochastic processes that generate disease outbreak data that form time-series composed of non-negative integers. The continuous lifetime distributions in Sect. 2.2 are statistical models if the questions under investigation are regarding estimation and testing hypothesis of time-to-event, such as the incubation period defined as the time elapsed from infection to the onset of clinical symptoms, time to recovery, time to death, etc., based on longitudinally observed or retrospectively assessed data. (Under a different context, these continuous lifetime distributions are implicitly built into the phenomenological population models, such as the exponential distribution of the infectious periods.)

The systematic component describes the systematic effects of interest, within the data-generating mechanism. In classic linear regression analysis, this component is formulated through a set of covariates $\underline{x} = \left(x_1, \ldots, x_p\right)$ in a linear form $\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$. When we say a model is linear, we mean linearity in parameters $\underline{\beta}$, not the covariates $\underline{x}$.

If a random sample $(y_1, \ldots, y_n)$ arises from independent observations $y_i \sim f(y_i; \theta_i)$, $i = 1, \ldots, n$, the reduction of dimension of the parameter space where $p \ll n$ is a mapping

$$(\theta_i : i = 1, \ldots, n) \mapsto \left(\beta_j : j = 1, \ldots, p\right)$$

through the covariates $\underline{x}$ by the linear function through the link function $h(\theta)$ so that

$$h(\theta) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p. \tag{7.1}$$

Such a model is a generalized linear model (McCullagh and Nelder 1983). Data arising from random counts are often fitted to one of the discrete distributions in Chap. 3 associated with a positive parameter $\theta > 0$ and a logarithm link function $h(\theta) = \log\theta$. The corresponding generalized linear models are the log-linear models. Binary data are often fitted to the binomial, geometric, or negative-binomial distributions. These distributions are often associated with a proportion parameter $0 < \theta < 1$. A logit link function $h(\theta) = \log \frac{\theta}{1-\theta}$ is often chosen, which gives rise to the logistic regression models. Continuous lifetime data are often fitted with the lifetime distributions in Chap. 2 that may be associated with a log-linear model or a proportional hazard model. The latter, in a broader sense, can also be viewed as a generalized linear model.

The systematic components in models for infectious disease outbreak investigations are typically nonlinear functions with respect to their parameters. These are phenomenological population models. Some of them have explicit analytic forms. We shall see many examples in Chap. 8. Others are implicit, including the transmission dynamic models expressed as a system of differential equations discussed in the proceeding chapters. These nonlinear models create additional challenges in computation algorithms, such as the optimization algorithms in the search for the maximum likelihood estimates or the least square estimation. They are highly sensitive to the initial parameter estimates in those algorithms. In the special case of the generalized linear models, initial estimates are not necessary. Therefore, it is important to carefully evaluate the values of the log-likelihood or the sum of square errors (SSE) upon convergence over a wide range of possible initial estimates.

It is important to recognize that disease transmission is only part of the data generating process, and many of the disease transmission models do not directly predict observable events as reflected by data. Other data generating mechanisms, such as case-definition, how data are organized and reported, length-biasedness, retrospective ascertainment of time of events, reporting delays, among many other issues, also need to be described using statistical models. The link function connects these two components and links them to the distribution of the data.

Lindsey (2001) provides comprehensive discussions on nonlinear models in medical statistics.

7.1.2 Fitting Models to Data and Model Criticism

In fitting a statistical model to data, the information in the data is split into two parts, one to assess the unknown parameters, and the other for model criticism. Both assessment of parameters and model criticism are equally important aspects in statistical inference.

Sprott (2000) points out that the sample information is divided into "Likelihood θ" and "Model f" through the factorization of a likelihood function according to the minimal sufficient division or the maximal ancillary division. A classic example

is the factorization of the Poisson likelihood. Consider an i.i.d. random sample $(y_1, \ldots, y_n)$ from the Poisson distribution with mean value μ, then $t = \sum_{i=1}^{n} y_i$ is Poisson distributed with mean $n\mu$. The joint distribution is

$$f(y_1, \ldots, y_n; \mu) = \prod_{i=1}^{n} \frac{\mu^{y_i} e^{-\mu}}{y_i!} = f(t; \mu) f(y_1, \ldots, y_n | t),$$

where the first factor

$$f(t; \mu) = \frac{e^{-n\mu} (n\mu)^t}{t!}$$

is the likelihood function of μ as represented by the minimal sufficient statistics $t = \sum_{i=1}^{n} y_i$, with one degree of freedom, for the assessment of the parameter μ. The second factor

$$f(y_1, \ldots, y_n | t) = \frac{t!}{\prod_{i=1}^{n} y_i!} \prod_{i=1}^{n} \left(\frac{1}{n}\right)^{y_i}$$

is a multinomial distribution for the residual, with $n - 1$ degrees of freedom. If data cast doubt on this multinomial distribution, they equally cast doubt on the assumed Poisson model.

Residuals are often associated with regression models. If a regression model such as (7.1) involves p unknown parameters, fitting such a model to data of sample size $n \gg p$ yields a residual consisting of $n - p$ degrees of freedom. Residual analyses in the form of goodness-of-fit play a crucial role for model criticism on three levels. At the first level, large residual values indicate a lack of fit. This is often used in conjunction with the testing of hypothesis $H_0 : \beta_j = 0$, $j = 1, \ldots, p$. It is the criticism of a sub-model within a larger model to single out important covariates x_j that are statistically significant. The second level is the testing against some fundamental assumptions in these models. For example, the logistic regression models are often associated with the assumption of proportional odds ratios and the proportional hazard model assumes proportional hazard functions. In good statistical practice, one always needs to take due diligence to test against these assumptions whenever these models are applied. Various testing statistics are available in the literature, such as the Z statistics to test against the proportional hazard assumption in almost every survival analysis textbook. The third level is the testing against the probability distributions, for instance, if data used in a logistic regression model arise from a binomial distribution or if data used in a proportional hazard model arise from a Weibull distribution. This may be optional if the primary interest is in the parameters β_j, $j = 1, \ldots, p$ while treating the underlying distribution as a nuisance parameter problem.

7.1.3 *Fitting Phenomenological Population Models to Time-Series Data*

Fitting phenomenological population models to time-series data collected during an epidemic, often called *curve-fitting*, is commonly practiced for the purposes of parameter estimation and prediction (Smirnova and Chowell 2017). These models can be mechanical disease transmission models with strong assumptions on the transmission process, or other forms of simpler, but nonetheless highly nonlinear descriptive models for the data generating processes. We may regard fitting phenomenological population models to disease outbreak investigation data as nesting a phenomenological population model inside the systematic component of a statistical model in the form of *generalized nonlinear regression*.

In general, we denote the time series of T longitudinal observations by

$$\underline{y} = (y_1, y_2, \ldots, y_T)$$

where $t = 1, 2, \ldots, T$ are discrete or time units, such as daily, weekly, etc., typical in disease outbreak investigations. We regard these data as realizations of random counts $Y(t)$ manifested through a dynamic system, such as in the SEIR system as discussed in Sect. 5.4, appropriately grouped into discrete time units as Y_t.

The systematic component of the model is denoted by $\mu(t; \Theta)$, which is a nonlinear function specified by a set of parameters $\Theta = (\theta_1, \ldots, \theta_m)$. The marginal distribution for Y_t may be only specified to its first moment $E[Y_t] = \mu(t; \Theta)$, or the first two moments, both as functions of Θ, or fully specified such as the Poisson distribution $Poisson(\mu(t; \Theta))$.

Parameter Estimation

We consider the quasi-likelihood estimating equations for the generalized linear models (McCullagh and Nelder 1983)

$$\sum_{t=1}^{T} \frac{\partial \mu(\mu; \Theta)}{\partial \theta_j} \frac{y_t - \mu(t; \Theta)}{V[Y_t; \Theta]} = 0, \ \ j = 1, \ldots, m \tag{7.2}$$

are still valid, where $V[Y_t; \Theta]$ is the variance of Y_t. More generally, the denominator $V[Y_t; \Theta]$ may also involve a correlation matrix, which relaxes the independency assumption among Y_t, $t = 1, 2, \ldots, T$, which are called the generalized estimating equations (Liang and Zeger 1986). The generalized estimating equations can be also applied to zero-mean martingales in Sect. 3.3.2 (Godambe and Heyde 1987).

One of the common choices is assuming $V[Y_t; \Theta] = \alpha\mu(t; \Theta)$, where $\alpha > 0$ is a scalar parameter. This variance form may well approximate variance structures such as $Var[Y_t] = E[Y_t] + E[Y_t]^2/\kappa$ when $E[Y_{t_i}]^2/\kappa$ do not vary greatly with t. The

variance structure of the negative binomial distribution (3.20) follows this form, as well as the mixed Poisson distribution in which the mixing distribution is inverse-Gaussian (see Chap. 3). In this case,

$$\sum_{t=1}^{T} \frac{\partial \mu(\mu;\Theta)}{\partial \theta_j} \frac{y_t - \mu(t;\Theta)}{\alpha \mu(t;\Theta)} = 0, \ \ j = 1, \ldots, m$$

which may be useful to handle data with overdispersion ($\alpha > 1$).

The estimating equations given by (7.2) take the form of the score functions of the likelihood functions of many well-known distributions, such as the Gaussian, Poisson, binomial, among many others, provided that the distributions are correctly specified. Without specifying the distribution, they are unbiased estimating equations that lead to asymptotically unbiased point estimates regardless of any misspecification of the variance–covariance structure. If the variance–covariance structure is correctly specified, they lead to the variance estimation of the parameter estimates. However, the estimated variances of the parameter estimates will be in error with misspecification of the variance–covariance structure.

These estimating equations are usually associated with generalized linear models. In contrast, phenomenological models are nonlinear, and in some cases, are implicitly defined through differential equations without analytic solutions. This poses computational challenges because $\frac{\partial \mu(\mu;\Theta)}{\partial \theta_j}$ is either complicated or prohibitive.

In the following two special cases, optimization algorithms can be employed without the evaluation of $\frac{\partial \mu(\mu;\Theta)}{\partial \theta_j}$.

One is the EE given by

$$\sum_{t=1}^{T} \frac{\partial \mu(\mu;\Theta)}{\partial \theta_j} \frac{y_t - \mu(t;\Theta)}{\mu(t;\Theta)} = 0, \ \ j = 1, \ldots, m. \tag{7.3}$$

It assumes that $Var[Y_t] = E[Y_t] = \mu(t;\Theta)$. As mentioned in Chap. 3, among power series distributions used for random counts, $Var[Y_t] = E[Y_t]$ characterizes the Poisson distribution (Kosambi 1949). In fact, (7.3) is the score function of the likelihood function assuming that $\underline{y} = (y_1, y_2, \ldots, y_T)$ are realizations of independent Poisson random counts. The log-likelihood function is

$$l(\Theta) = \sum_{t=1}^{T} \left[y_t \log \mu(t;\Theta) - \mu(t;\Theta) \right]. \tag{7.4}$$

Therefore, solving (7.3) is equivalent to maximizing (7.4). The maximum likelihood estimate can be expressed as

$$\widehat{\Theta} = \arg\max \sum_{t=1}^{T} \left[y_t \log \mu(t;\Theta) - \mu(t;\Theta) \right]. \tag{7.5}$$

One can use numerical optimization methods in MatLab or R (R Core Team). In R, a general-purpose optimization method based on the downhill simplex method (Nelder-Mead) or the quasi-Newton algorithms are readily available. However, for a nonlinear function $\mu(t;\Theta)$, the optimization algorithms to maximize the log-likelihood are highly sensitive to the initial parameter estimates, which may lead to local maxima. It is important to carefully evaluate the values of the log-likelihood upon convergence over a wide range of possible initial estimates.

An alternative method is the least square estimate, achieved by searching for the set of parameters $\widehat{\Theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_m)$ that minimizes the sum of squared differences between the observed data and the corresponding model solution denoted by$\mu(t;\Theta)$, $t = 1, 2, \ldots, T$. That is, the objective function is given by

$$\widehat{\Theta} = \arg\min \sum_{t=1}^{T} [y_t - \mu(t;\Theta)]^2 . \tag{7.6}$$

This is equivalent to solving the EE

$$\sum_{t=1}^{T} \frac{\partial \mu(\mu;\Theta)}{\partial \theta_j} [y_t - \mu(t;\Theta)] = 0, \;\; j = 1, \ldots, m \tag{7.7}$$

assuming that $Var[Y_t]$ is independent of the mean and does not involve the set of parameters Θ. Although this method does not assume any specific distribution for Y_t except for its first moment $E[Y_t] = \mu(t;\Theta)$, the least square method is equivalent to the maximum likelihood estimation if data Y_t are Gaussian distributed. If the random counts are highly skewed, the least square method may not perform well. In Matlab (The Mathworks, Inc.), two numerical optimization methods are available to solve the nonlinear least squares problem: The trust-region reflective algorithm and the Levenberg-Marquardt algorithm. As with the maximum likelihood estimates, the optimization algorithms to minimize the sum of square errors (SSE) are highly sensitive to the initial parameter estimates, which may lead to local minima. It is important to carefully evaluate the values of the SSE upon convergence over a wide range of possible initial estimates.

Uncertainty in Estimated Parameters

The Likelihood Surface and the Likelihood Ratio Statistics The relative likelihood, $R(\Theta; \underline{y})$, is defined by

$$0 < R(\Theta; \underline{y}) = \frac{L(\Theta; \underline{y})}{\sup_{\Theta} L(\Theta; \underline{y})} = \frac{L(\Theta; \underline{y})}{L(\widehat{\Theta}; \underline{y})} \leq 1$$

where $L(\Theta; \underline{y})$ is the likelihood function of Θ given data $\underline{y}$. It ranks the parameters Θ over the scale from 0 to 1, and the maximum likelihood estimate $\widehat{\Theta}$ is the most plausible value of Θ in that it makes the observed data $\underline{y}$ most probable (Sprott 2000; Kalbfleisch 1985). One could also define the likelihood region such that $R(\Theta; \underline{y}) \geq \varsigma$ where $0 < \varsigma < 1$ as plausible parameter values. By varying the threshold ς, one can define such things as "very plausible," "plausible," or "implausible." These concepts give rise to the likelihood ratio statistics that can be used to construct confidence intervals and test hypotheses. More importantly, visualization of the contour of the likelihood surface in the neighborhood reveals the amount of information the data contain with respect to each parameter. This is feasible when the number of parameters in Θ is less than or equal to 2.

It is more convenient to work on the logarithmic scale. The relative log-likelihood is defined by

$$-\infty < r(\Theta) = l(\Theta) - l(\widehat{\Theta}) \leq 0$$

where $l(\Theta) = \log L(\Theta; \underline{y})$.

A likelihood region is defined on the parameter space such that $R(\Theta; \underline{y}) \geq \varsigma$ for a selected value $0 < \varsigma < 1$. Calculating the $100(1-p)\%$ confidence regions based on the likelihood ratio can be done directly by selecting the value for the likelihood region , that is $r(\Theta) = l(\Theta) - l(\widehat{\Theta}) \geq \log \varsigma$, so that the coverage probability

$$CP \approx \Pr(\chi^2_{\text{df}} \leq -2 \log \varsigma) = 1 - p.$$

The 95% confidence interval for a single parameter θ can be derived by choosing $\varsigma = 0.147$ such that

$$CP \approx \Pr(\chi^2_{(1)} \leq -2 \log 0.147) = 0.95. \tag{7.8}$$

The left panel of Fig. 7.1 shows the 95% confidence interval for the mean value μ of the Poisson distribution a small sample of random count data based on (7.8). The most plausible value is $\widehat{\mu} = 2.2$ and the plausible range is $1.4046 \leq \mu \leq 3.2505$ such that $r(\mu) = l(\mu) - l(\widehat{\mu}) \geq \log 0.147 = -1.9173$.

The 95% confidence region for two parameters (α, β) is the contour of $r(\alpha, \beta)$ by choosing $\varsigma = 0.05$ such that

$$CP \approx \Pr(\chi^2_{(2)} \leq -2 \log 0.05) = 0.95. \tag{7.9}$$

The right panel of Fig. 7.1 shows the 95% joint confidence region of two parameters: the median parameter λ^{-1} and the 95th percentile t_{95} of the log-logistic distribution for the incubation period fitted to a small sample of data collected on people with SARS symptoms, using (7.9).

In nonlinear models, parameters are often inter-related in complex ways. In a two-parameter setting, it is more likely to encounter "banana" log-likelihood

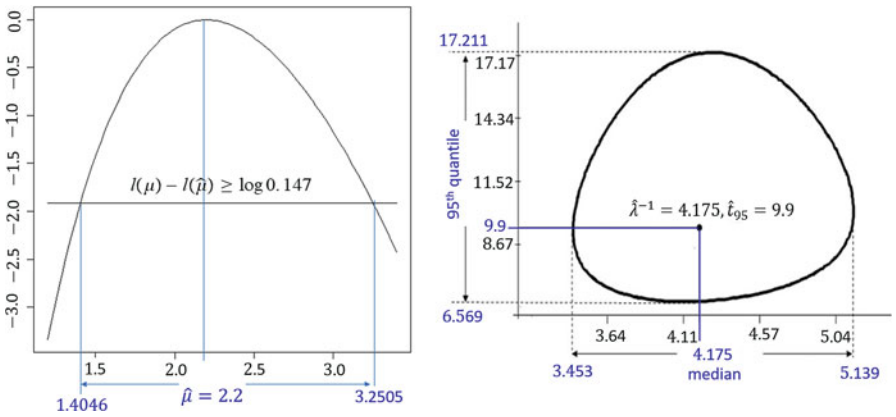


Fig. 7.1 Left: the relative log-likelihood of the Poisson distribution from the sample of random counts: {0, 5, 2, 3, 2, 3, 1, 0, 2, 4} with m.l.e. $\widehat{\mu} = 2.2$ (1.4, 3.25); Right: the joint of likelihood region for the median and the 95th percentile of the log-logistic distribution for the incubation distribution based on a small sample of SARS patients

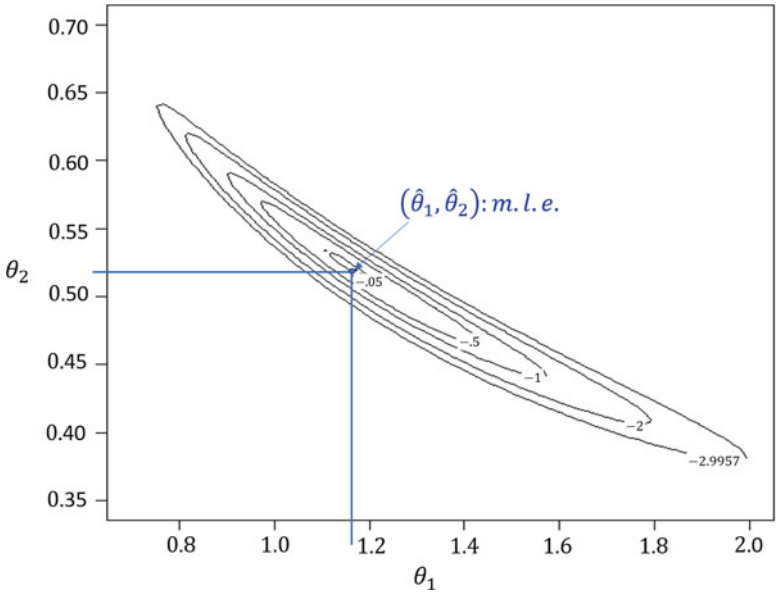


Fig. 7.2 A banana shaped log-likelihood contour showing the correlation of two parameters θ_1 and θ_2 as suggested by data, where $\widehat{\theta_1}$ and $\widehat{\theta_2}$ are the maximum likelihood estimates. The outmost contour line corresponds to the likelihood region with 95% coverage probability

contours as schematically illustrated in Fig. 7.2. The 95% joint likelihood region is the outmost contour, which given the marginal 95% confidence limits for θ_1 between 0.65 and 2.0, and the marginal 95% confidence limits for θ_2 between

0.38 and 0.65. However, it is equally plausible to have these two combinations: ($\theta_1 = 0.7, \theta_2 = 0.65$) or ($\theta_1 = 2.0, \theta_2 = 0.38$). These two pairs may represent very different epidemic scenarios, but they are equally accepted by data. This poses an *identifiability* problem. On the other hand, the likelihood contour also rules out implausible scenarios, such as ($\theta_1 = 1.6, \theta_2 = 0.6$), even though both values are well within their 95% confidence limits.

In the case of more than two parameters, it is still worthwhile to visualize the likelihood surface either as a 3-D function or cross-sectional log-likelihood contours. These will provide more reliable precision intervals than marginal confidence intervals for each parameter, reveal correlation among parameters, and provide better ways to communicate uncertainty. However, these are very time-consuming.

With respect to the testing of the hypothesis $H_0 : \Theta = \Theta_0$, the likelihood ratio statistics is given by

$$D = -2r(\Theta_0) = -2\left[l(\Theta_0) - l(\widehat{\Theta})\right].$$

The significant level is

$$SL = \Pr(D \geq D_{\text{obs}} | H_0 \text{ is true}) \approx \Pr(\chi^2_{\text{df}} \geq D_{\text{obs}}) \tag{7.10}$$

where the degree of freedom, df, is equal to the number of functionally independent parameters in the model. In testing a null hypothesis for a single parameter $H_0 : \theta = \theta_0$, the degree of freedom is df = 1. In testing a null hypothesis for two parameters $H_0 : \alpha = \alpha_0$ and $\beta = \beta_0$, df = 2.

The marginal 95% confidence interval for a single parameter in the presence of many other parameters can be also derived by numerically inverting the testing of null hypothesis $H_0 : \theta = \theta_0$ and calculate the significance level at different θ_0 using the $\chi^2_{(1)}$ approximation in (7.10), until $SL = 0.05$. In the case of m parameters in Θ, it involves two steps:

1. under the null hypothesis, fixing $\theta = \theta_0$, and conduct a maximum likelihood estimation for the remaining $m - 1$ parameters, denoted by Θ^*, and evaluate the value of the log-likelihood $l(\widehat{\Theta^*}|\theta = \theta_0)$;
2. under the alternative hypothesis, conduct a maximum likelihood estimation of all the parameters in Θ.

The likelihood ratio statistics is

$$D = -2\left[l(\widehat{\Theta^*}|\theta = \theta_0) - l(\widehat{\Theta})\right] \tag{7.11}$$

approximated by the $\chi^2_{(1)}$ distribution.

Assessing Uncertainty in the Estimated Parameters Through Bootstrapping The likelihood approach applies only when the joint distribution of $\underline{Y} = (Y_t, Y_{t,} \ldots, Y_T)$ is completely and correctly specified. Other approaches based on the asymptotic properties of the generalized estimating equations are not practical

because of the nonlinear functions employed in $\mu(\mu; \Theta)$ that make the calculations for $\partial\mu(\mu; \Theta)/\partial\theta_j$ prohibitive.

The general bootstrap method (Efron and Tibshirani 1994) based on assumed variance structures to assess uncertainty in the estimated parameters is useful. It is widely applied in quantifying parameter uncertainty and constructing confidence intervals in mathematical modeling studies (see, e.g., Chowell et al. 2006a,b). In this method, multiple observations are repeatedly sampled from the best-fit model by assuming that each point of the time series follows a specific distribution, typically a Poisson or a negative binomial distribution, centered on the estimated mean at that time point. The step-by-step algorithm to quantify parameter uncertainty follows:

1. Derive the parameter estimates $\widehat{\Theta} = (\widehat{\theta_1}, \ldots, \widehat{\theta_m})$ by fitting the model to the time series data $\underline{y} = (y_1, y_2, \ldots, y_T)$ to obtain the best-fitted model $\mu(t; \widehat{\Theta})$, $t = 1, \ldots, T$.
2. Generate replicated simulated datasets through re-sampling. To do so, we first use the best-fit model $\mu(t; \widehat{\Theta})$ to calculate the expected values of the time-series.
3. Each simulated data set is generated by random numbers assuming Poisson or negative binomial error structures based on the expected values. Specifically, for time t, a random number y_t^* with mean value $\mu(t; \widehat{\Theta})$ is drawn from a Poisson or a negative binomial distribution. This forms a simulated time series $\underline{y}^* = (y_1^*, y_2^* \ldots, y_T^*)$. Repeating this simulation s times, we obtain s replicated simulated datasets, denoted by $\underline{y}^*_{(1)}, \underline{y}^*_{(2)}, \ldots, \underline{y}^*_{(s)}$.
4. Re-estimate parameters for each of the s simulated realizations. Estimated parameter sets given by $\widehat{\Theta}_i$, $i = 1, \ldots, s$.
5. Using the set of re-estimated parameters $\widehat{\Theta}_i$, $i = 1, \ldots, s$, it is possible to characterize their empirical distributions, correlations, and construct confidence intervals.

In addition, since infectious disease outbreaks are not repeatable under identical conditions (in the sense of a designed random experiment), the computer-based re-sampling provides a virtual experiment with the resulting uncertainty around the model fit given by $\mu(t; \widehat{\Theta}_1)$, $\mu(t; \widehat{\Theta}_2), \ldots, \mu(t; \widehat{\Theta}_s)$. This is very useful for assessing uncertainty of key disease transmission parameters such as the basic reproduction number (Anderson and May 1991; Diekmann et al. 1990; van den Driessche and Watmough 2002). This parameter is a function of several parameters that characterize the transmission and control process, e.g., transmission rates and infectious periods of the epidemiological classes that contribute to new infections. Uncertainty around this key parameter, estimated through re-sampling, can be viewed through such a perspective.

Residual Analysis

While residuals (differences between model fit and observations), $r_t = y_t - \mu(t; \widehat{\Theta})$, can inform systematic deviations of the model fit to the data, it is also possible to

quantify the error of the model fit to the data using performance metrics (Kuhn and Johnson 2013). These metrics are also useful to quantify the error associated with forecasts. A widely used performance metric is the mean square error (MSE), which is given by

$$MSE = \frac{1}{T}\sum_{t=1}^{T}\left[y_t - \mu(t;\widehat{\Theta})\right]^2. \tag{7.12}$$

Another commonly used residual is the Pearson residual, defined as

$$r_t^{(P)} = \frac{y_t - \mu(t;\widehat{\Theta})}{\widehat{V}_t^{1/2}}$$

where $\mu(t;\widehat{\Theta})$ is the expected value of Y_t and $\widehat{V}_t$ is the estimated variance $Var(Y_t)$. In particular, if the variance structure corresponds to (7.3), then

$$r_t^{(P)} = \frac{y_t - \mu(t;\widehat{\Theta})}{\sqrt{\mu(t;\widehat{\Theta})}}.$$

The performance metric is the weighted mean square error (WMSE)

$$WMSE = \sum_{t=1}^{T}\frac{\left[y_{t_i} - \mu(t;\widehat{\Theta})\right]^2}{\mu(t;\widehat{\Theta})} = \sum_{t=1}^{T}\frac{(O-E)^2}{E} \tag{7.13}$$

where O stands for "Observed" and E stands for "Expected." Although (7.13) assumes the Poisson variance structure, minimizing the weighted sum of squares $\sum_{t=1}^{T}\frac{\left[y_{t_i}-\mu(t;\widehat{\Theta})\right]^2}{\mu(t;\widehat{\Theta})}$ does not correspond to the unbiased estimating equation (7.3) and hence it does not correspond to the maximum likelihood estimation by maximizing (7.4). In fact, the weighted least square estimation by minimizing $\sum_{t=1}^{T}\frac{\left[y_{t_i}-\mu(t;\widehat{\Theta})\right]^2}{\mu(t;\widehat{\Theta})}$ would have yielded the equation

$$\sum_{t=1}^{T}\frac{\partial\mu(\mu;\Theta)}{\partial\theta_j}\frac{y_t-\mu(t;\Theta)}{\mu(t;\Theta)} + \frac{1}{2}\sum_{t=1}^{T}\frac{\partial\mu(\mu;\Theta)}{\partial\theta_j}\left[\frac{y_t-\mu(t;\Theta)}{\mu(t;\Theta)}\right]^2 = 0$$

of which the first term is the left hand of (7.3). The weighted least square estimates are asymptotically biased because (7.3) gives the asymptotically unbiased estimates. In addition to the above arguments, the distributions of the residuals defined by $r_t^{(P)}$ are skewed for non-Gaussian distributions. Consequently, using WMSE as the performance measure for Poisson distributed random counts may not be the best choice.

An improved version of (7.13) is based on the Anscombe residuals (Anscombe 1953). Descriptions for this residual measure can be found in the book by McCullagh and Nelder (1983). For Poisson distributed random counts, the residuals are defined as

$$r_t^{(A)} = \frac{\frac{3}{2}\left[y_t^{2/3} - \mu(t;\Theta)^{2/3}\right]}{\mu(t;\Theta)^{1/6}},$$

and the performance metric is

$$ANSCOMBE = \sum_{t=1}^{T} \left(\frac{\frac{3}{2}\left[y_t^{2/3} - \mu(t;\Theta)^{2/3}\right]}{\mu(t;\Theta)^{1/6}} \right)^2. \tag{7.14}$$

The Anscombe residuals $r_t^{(A)}$ are approximately Gaussian distributed.

Model Criticism

In fitting models to time-series data, one step is to choose an appropriate model for the systematic component. For example, if the time-series exhibits a trend that resembles a sigmoid curve, one may consider a simple growth function, such as a logistic function, and conduct residual analyses and hypothesis tests against some versions of generalized logistic models to examine whether the model captures the general characteristics of the observed time series.

On the other hand, depending on the data fitting methods, there are subtle assumptions on the random component. If the data fitting method is based on a likelihood function, the full specification of the distribution of data must be given. Model criticism will be something like: do observed random counts as a finite time-series arise as an independent sample of, say, a negative-binomial distribution, with its mean values further modelled by a deterministic function of t? Even with empirical curve-fitting, such as the least square method, statistical assumptions such as independency among data points, the relationship between the variance and the mean are still made. These are all subject to criticism in the light of data.

Data usually admit more than one model. Even when a specific model is preferred, for scientific or practical reasons, alternative models also need to be taken into consideration.

In modelling disease outbreak data, it is very common that the available data cannot identify all the parameters involved. What we mean by "not identifiable" is that, in a multiple parameter setting, more than one set of combinations of parameters manifest the same expected value that fits well to data. Since the model is split into random, systematic, and link components, the problem of *identifiability* carries over to the identifiability of these components. This makes model criticism more complex.

7.2 Data

7.2.1 Some Features of Infectious Disease Outbreak Data

A striking feature of data collected during an infectious disease outbreak is that they do not arise from designed experiments, which are either impossible or unethical in the context of epidemics among humans.

Data are not repeatable. Outbreaks of the same disease do not start with identical conditions. Moreover, environmental and behavioral changes occur, and pathogens mutate. Even with data collected as a long sequence of time-series, or data collected from multiple data sources, or even Big Data, one may view them as high-dimensional data based on a single realization of a random event.

Furthermore, outcomes are not independent over disjoint time intervals and between individuals. One example is the phenomenon of *herd immunity* where individuals that are vaccinated indirectly protect those who are not vaccinated.

These data features determine the statistical models and methods that are different from those based on designed experiments with i.i.d. samples.

7.2.2 What Do We Mean by "Large Number"?

In the classic statistics textbooks, "large number" is associated with the *law of large numbers*, the central limit theory, the asymptotic confidence intervals, and asymptotically unbiased estimates, such as the estimates based on unbiased estimating equations. In such context, it is called the sample size, which is understood as the number of repeated independent random experiments under identical conditions. An infectious disease outbreak dataset, no matter how many observations, is considered a small sample.

In a different context, when we say that deterministic models are approximations of the mean field of the corresponding stochastic processes, we do not mean large populations, but large repeated realizations of the same outbreak under identical conditions. To a certain extent, when the population size m in disease transmission models becomes large, the stochastic effects of correlations among the numbers of individuals in different compartments are reduced, even negligible, such as $\frac{\beta}{m} cov\{S(t), I(t)\}$ in (5.23). This may lead to a smooth realized epidemic curve that resembles that predicted by a deterministic model. However, it is not the average of all possible epidemic curves in large numbers of repetitions of the epidemic under identical conditions. This distinction was illustrated in Figs. 5.1 and 5.2 in Chap. 5. The deterministic models can be viewed as approximations of the mean field of the corresponding stochastic processes in the context of a large number of repetitions of the epidemic under identical conditions. The population size is not equivalent to the sample size.

In fitting models to time-series data, increasing the number of observations means more accumulation of data over time to achieve longer time-series. A single time-series is still regarded as sample size = 1. Longer time series also improves the precision of the estimates, but only to a certain extent, and is not equivalent to having a large sample. Increasing observations over time often forces us to change to more complex models whereas increasing the sample size does not.

7.2.3 Lack of Information or Not Identifiable?

In a single parameter setting, the lack of information from data simply means very imprecise estimation. The confidence interval is extremely wide, or one-sided, or unable to yield the point estimate (e.g., the likelihood function is maximized at the boundary of the parameter space). It is often characterized by a very flat likelihood function over the parameter space (Raue et al. 2009; Roosa and Chowell 2019).

In multiple parameter settings data cannot provide accurate estimates for some of the parameters or cannot test against certain hypotheses when fitting the model to a single data source (i.e., single time-series) especially in models that involve sub-models for the random, systematic, and link components. People often say, ambiguously, that data do not have enough information or are not able to identify certain parameters. We would like to point out some subtle differences.

Shared Information in a Multiple Parameter Setting

We start with the classic statistical problem of the i.i.d. sample $\left(x_1, \cdots, x_n\right)$ arising from the Gaussian distribution $N(\mu, \sigma^2)$ with the parameter of interest being the variance σ^2. It is well known that the maximum likelihood estimate of σ^2 is

$$\widehat{\sigma^2} = \begin{cases} \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2, & \text{if } \mu \text{ is known;} \\ \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2, & \text{if } \mu \text{ is unknown and } \widehat{\mu}=\overline{x} \end{cases}.$$

It is also well known that when μ is known, $\widehat{\sigma^2}$ is an unbiased estimator, whereas if μ is unknown $\widehat{\sigma^2}$ is only asymptotically unbiased for σ^2, but biased in any finite population. It is unbiased for $\frac{n-1}{n}\sigma^2$. The unbiased estimator for σ^2 is $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2$. This is because the information for σ^2 in the data is shared with the parameter μ and the minimum sufficient statistics for μ is $\overline{x} = \frac{1}{n}\sum_{i=1}^{n}x_i$. Data can be re-arranged through one-to-one mapping:

$$\left(x_1, \cdots, x_n\right) \mapsto (x_1-\overline{x}, \cdots, x_{n-1}-\overline{x}, \overline{x})$$

in which $\overline{x}$ contains all the information in the data for μ and $(x_1-\overline{x}, \cdots, x_{n-1}-\overline{x})$ are the residuals with $n-1$ degrees of freedom left for the estimation of σ^2.

This argument is formalized in the Fisher-Neyman factorization of the likelihood function

$$f\left(x_{1,}\cdots,x_n\right)=\frac{1}{\left(\sqrt{2\pi}\sigma\right)^n}e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}=\frac{1}{\left(\sqrt{2\pi}\sigma\right)^n}e^{-\frac{n}{2\sigma^2}(\overline{x}-\mu)^2}e^{-\frac{n-1}{2\sigma^2}s^2},$$

where $s^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2$. The pair $(\overline{x},s^2)$ is the minimal sufficient statistics for (μ,σ^2). In other words, data are reduced to the pair $(\overline{x},s^2)$ and the information for variance is summarized as the mean of the square errors $(x_i-\overline{x})^2$ from $n-1$ out of the data of sample size n.

A more revealing example is the Neyman-Scott paradox (Neyman and Scott 1948). Consider $2n$ independent measures $\left(x_{1,}\cdots,x_n\right)$ and $\left(y_{1,}\cdots,y_n\right)$, where $X_i\sim N(\mu_i,\sigma^2)$, $Y_i\sim N(\mu_i,\sigma^2)$, $i=1,\ldots,n$. There are $n+1$ unknown parameters $(\mu_{1,}\cdots,\mu_n,\sigma^2)$. If we use the likelihood function given by the joint distribution

$$f\left(x_{1,}\cdots,x_n,y_{1,}\cdots,y_n\right)=\frac{1}{\left(\sqrt{2\pi}\sigma\right)^{2n}}e^{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(x_i-\mu_i)^2+\sum_{i=1}^{n}(y_i-\mu_i)^2\right)},$$

the maximum likelihood estimates are

$$\widehat{\mu_i}=\frac{1}{2}\left(x_i+y_i\right),\ i=1,\ldots,n$$
$$\widehat{\sigma^2}=\frac{1}{4n}\left(x_i-y_i\right)^2.$$

In this case, it can be shown that $\widehat{\sigma^2}$ is not only biased, but also asymptotically biased:

$$E\left[\widehat{\sigma^2}\right]\to\frac{\sigma^2}{2},\ \text{as } n\to\infty.$$

This analysis under-estimates σ^2 by 50% because half of the information in data about σ^2 is lost in estimating $(\mu_{1,}\cdots,\mu_n)$.

Later in Chap. 8, we shall see an example with discussions, where the model evolves from relatively simple to more complex by adding a shape parameter, adapted to an increasing number of observations of a time-series during a Zika outbreak investigation. In the simpler form, all three parameters directly address three public health questions of interest. Data collected in the early period have little information on these parameters in terms of very wide confidence limits. As data accumulate, the estimation for these parameters becomes more and more precise. Meanwhile, data start to force us to add a shape parameter to the model, which does not directly address any of the questions of public health interest. It does,

however, improve the model's goodness of fit and also correct potential biases in the estimated parameters of interest, which is increasingly apparent as suggested by data. This added parameter needs to be estimated using the information from data at a cost of the precision in the estimation of parameters of interest. Therefore, at some midpoint in the outbreak, discussions are needed to address the pros and cons of expanding the model at that moment when the time-series might not be long enough to accommodate another parameter. It is only after another month of data accumulates that it becomes obvious that the more complex model is necessary and meanwhile the estimation of parameters of interest becomes more precise.

In a different example, Lagakos et al. (1988) presented data based on 258 adults with transfusion-associated AIDS. Data were fitted to a Weibull distribution (2.12) with scale parameter λ and shape parameter $\varsigma > 0$. The purpose is to estimate the incubation period from the time of transfusion and the onset of AIDS illnesses. Data are not i.i.d. in the sense of a random sample from an experiment, but from a different type of observational scheme (to be discussed later). Figure 7.3 illustrates the contours of the surface of the log-likelihood. It shows that data do not have enough information for the scale parameter λ except for $\lambda \leq 0.128$, and the m.l.e. does not exist. Together with estimated $\widehat{\varsigma} \approx 2.1$, at best the data tell us that the incubation period is very long, with the lower bound of the estimated median incubation period ≥ 6.6 years. On the other hand, data are informative about ς, with $1.85 \leq \varsigma \leq 2.38$ based on the approximate 95% confidence limits based on the likelihood ratio statistics. It implies an approximately linear increasing hazard function.

The take home message from the above discussions includes:

1. when there are multiple unknown parameters in the model, information in the data are shared among the parameters;

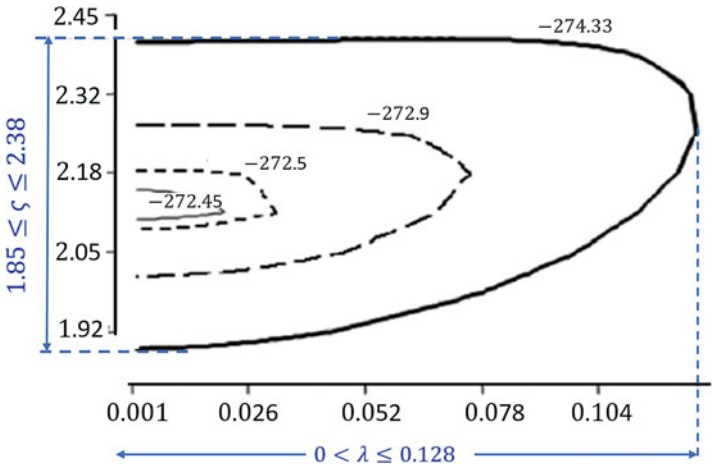


Fig. 7.3 Contours of the surface of the log-likelihood $l(\lambda, \varsigma)$ of the Weibull distribution along with the 95% confidence region using the likelihood ratio statistics for the two parameters based on data from Lagakos et al. (1988)

2. in the presence of *nuisance parameters,* without additional statistical modelling to handle the nuisance parameters, data may not have enough information to estimate the parameter of interest, in the sense of precision as well as potential biases;
3. the amount of information in data with respect to the parameters of interest is also affected by how the data are collected.

Identifiability Among Parameters and Components of Models

For two parameters (α, β), if different combinations (α_1, β_1), (α_2, β_2), ... produce the same expected value of the model that gives equally good fit to data, we say that data are not able to identify them. This problem can arise for different reasons.

One of the sources that can cause the identifiability problem is due to high correlation among parameters in a nonlinear function. The banana shaped likelihood contour in Fig. 7.2 illustrates such a correlation. In Chap. 8, we shall see many banana shaped likelihood contours. In some applications, the paired combination (α, β) corresponds to a specific scenario of scientific or epidemiologic interest. For example, one parameter may represent the number of infected individuals at the beginning of the epidemic and another parameter may represent the rate of growth of the epidemic. Data may not be able to identify whether the observed phenomenon is due to a small number of initially infected individuals combined with a fast growth rate, or a large number of initially infected individuals combined with a slow growth rate. On the one hand, lack of parameter identifiability is not the same concept as the lack of information because, through the contours of the likelihood function, data do provide a great deal of information on how the parameters are correlated and are able to estimate these parameters. On the other hand, as the number of observations increases, more precise estimates can be obtained, which in many cases solves the identifiability problem.

Another source of identifiability issues is the redundancy of parameters. For example, the logistic growth function (5.14), parameterized as

$$\frac{mi_0(\beta - \gamma)}{\beta i_0 + (m\,(\beta - \gamma) - \beta i_0)\, e^{-(\beta-\gamma)t}}$$

has a shape of a sigmoid curve. Data may allow rather precise estimation of the initial growth $\rho = \beta - \gamma$, but cannot identify β or γ separately unless one of them is known. All these parameters have specific scientific interpretations. In the context of the SIS compartment model, the hypothesis $H_0 : \gamma = 0$ distinguishes whether the model is SI or SIS. In this sense, the parameters are not redundant. However, from the data point of view, only three of the four parameters can be estimated. When data admit a logistic functional form, they cannot identify the SIS model. In this case, increasing the number of observations will not solve the identifiability problem of parameters. Re-parameterization of the logistic function will, though it cannot solve the problem of identifying underlying model structures.

The identifiability problem often arises in models that involve sub-models. For instance, it is well known that univariate frailty models (Sect. 2.6) through mixture of distributions are not identifiable from the survival information alone. Similarly, in a model that has random, systematic, and link components, some parameters are specific to the disease transmission process and other parameters are specific to another aspect of the data-generating process beyond disease transmission. A single time-series data usually cannot identify some of the embedded processes or components, especially if two components are linked through convolution.

Now we move on to the *link* component of the model (which has been less discussed so far).

A typical example is back-calculation, in which the disease transmission process is modelled through an intensity function $i(t;\underline{\theta})$, which describes the incidence of new infections over time, where new infections are not directly observable. Data $\underline{y}$ are generated based on the occurrence of the subsequent observable events, such as the onset of clinical symptoms as a consequence of infection. A statistical model may assume the observable events arise from a counting process with the intensity function $\mu(t;\underline{\theta},\underline{\psi})$, which is further modelled as a convolution

$$\mu(t;\underline{\theta},\underline{\psi}) = \int_0^t i(u;\underline{\theta}) f(t-u|u;\underline{\psi}) du \tag{7.15}$$

or $\mu(t;\underline{\theta},\underline{\psi}) = \sum_{u=0}^{t} i(u;\underline{\theta}) f(t-u|u;\underline{\psi})$, depending on whether one takes a continuous time or a discrete time framework. The quasi-likelihood generalized estimating equation (7.2) becomes

$$\sum_{t=1}^{T} \frac{\partial \mu(t;\underline{\theta},\underline{\psi})}{\partial \theta_j} \frac{y_t - \mu(t;\underline{\theta},\underline{\psi})}{V[Y_t;\underline{\theta},\underline{\psi}]} = 0,\ j = 1,\ldots,m. \tag{7.16}$$

In this model, the convolution (7.15) serves the same role as (7.1) in the generalized linear model.

The systematic component is $i(t;\underline{\theta})$ which captures the data-generating process due to disease transmission specified by a vector of parameters $\underline{\theta}$. It may be stochastic or deterministic transmission models that are explicitly linked to the underlying scientific hypotheses regarding the agent–host–environment interface. The number of unknown parameters in $\underline{\theta}$ depends on the complexity of the model. Donnelly and Ferguson (1999) formulated dynamic models for the population biology of the bovine spongiform encephalopathy (BSE) and then embedded this model into a back-calculation framework along with the maximum likelihood estimation. This approach gave estimated annual incidence of animals infected with BSE in Great Britain. In most other cases, it is convenient to adopt some flexible empirical parametric functions for $i(t;\underline{\theta})$ with relatively few parameters such as a generalized logistic function. There are also the flexible step-function models such as

$$i(t;\underline{\theta}) = \theta_j,\ j = 1,\cdots,q \tag{7.17}$$

involving q steps. Each step θ_j is a parameter. The longer the steps, the fewer the number of parameters.

The component $f(t-u|u;\underline{\psi})$ is a model that captures the details of all other data-generating processes since infection, as the conditional probability of being captured in the data at time t after an amount of time $x = t-u$, given infection at time $u < t$, specified by a vector of parameters $\underline{\psi}$. It may include aspects such as the factors that determine diagnoses of infections like the onset of clinical symptoms or external influences such as public health campaigns and screening. It may also include the process of reporting diagnosed infections to a central registry, delays in reporting, whether data are collected prospectively or retrospectively, and so on.

The estimating equation (7.16) incorporates the random components by specifying that $E[Y_t] = \mu(t;\underline{\theta},\underline{\psi})$ and appropriate variance structure $V[Y_t;\underline{\theta},\underline{\psi}]$.

The parameter of interest is the vector $\underline{\theta}$ because the objective is to estimate the incidence of new infections over time. However, data can only identify the convolution $\mu(t;\underline{\theta},\underline{\psi})$ as a whole, but not the systematic component $i(t;\underline{\theta})$ and the link component $f(t-u|u;\underline{\psi})$ separately.

7.2.4 *Observable Data and Unobservable Events*

At the Population Level

The ideal sequence of a scientific investigation is: formulation of research questions—obtaining appropriate data—analysis of data—interpretation of results. However, the investigation based on observational data collected during an infectious disease outbreak is an extreme departure from the ideal sequence, and most of the data are collected for other purposes unrelated to the question under investigation.

For example, a research question may be addressed using an SIR model in Sect. 5.3 with two parameters (β,γ). Using martingales (briefly introduced in Sect. 3.3.2), Becker (1989), Rida (1991), Becker and Hasofer (1997), Becker and Britton (2001), Hohle and Jørgensen (2003), and others have developed estimating equations that yield asymptotically unbiased estimates

$$\widehat{\beta} = \frac{C(t_N)}{\int_0^{t_N} \frac{S(x)I(x)}{m}dx},\ \widehat{\gamma} = \frac{C(t_N)}{\int_0^{t_N} I(x)dx},$$

with standard error estimates

$$s.e.\left(\widehat{\beta}\right) = \frac{\widehat{\beta}}{\sqrt{C(t_N)}},\ s.e.\left(\widehat{\gamma}^{-1}\right) = \frac{\widehat{\gamma}^{-1}}{\sqrt{C(t_N)}},$$

where t_N is the time of the observed end of the epidemic, and $C(t) = m - S(t)$ is the cumulatively infected individuals at time t, and m is the population size. Data

are assumed to arise as continuous and complete observation for $\{[S(t), I(t)] : 0 \leq t \leq t_N\}$. However, such data do not exist in reality, simply because a transmission from an infected individual to a susceptible individual is an unobservable event.

Most data at the population level are random counts based on observable events aggregated into time intervals. They may arise from multiple sources involving multiple agencies regarding the same outbreak. For example, a central public health agency of a country may compile a registry of reported "cases" of a certain reportable disease, which are forwarded from similar disease registry systems in state, provincial, territorial, or local authorities. Meanwhile, a different agency, or institute, or a collaborative sentinel hospital may have a database with variable population coverage regarding hospitalizations, discharges, and other events on the severe end of the same disease. In recent years, syndromic surveillance based on early warning indicators, such as emergency department attendances or emergency telephone calls, have gained much attention for their potential use to detect outbreaks at early stages. In the era of Big Data, the computer algorithm Google Trends aggregates Google Search queries by monitoring millions of users' health tracking behavior online.

The increasing number of data sources comes with pros and cons. On the pros side, multiple data sources, at least conceptually, help to identify high dimensional parameters in a complex model. On the cons side, it becomes increasingly difficult to sort out the data-generating process for each data source and increasingly difficult to develop the corresponding statistical models, not to mention model criticism.

What Is a "Case"?

In workshops on mathematical epidemiology, we have encountered questions from mathematicians such as:

> How do we reconcile differences between the incidence of new infections predicted by our models and the incidence of new cases in surveillance data we try to fit?

In the preceding paragraphs, we have illustrated the gap between the unobservable events predicted by mathematical models and data that are collected at the population level. Here we would like to highlight part of the data collection process, which is the preciseness in definitions and terminologies.

A "case" is one of the most commonly used terms in epidemiology and public health surveillance. A system based on reporting diagnosed diseases to a central registry is often called "case-reporting-surveillance." For each surveillance system, there is a case-definition (which may evolve and change over time).

We would say that a case is a file associated with an individual diagnosed with some "case-defining" illnesses, and this case must exist in some central registry in the system. Inside the "case" (more precisely, in the file), there are multiple observable events associated with different time points. Some events are relevant to the underlying epidemiology, such as the onset of clinical symptoms (but only among those according to the "case-definition"), the diagnosis of such symptoms

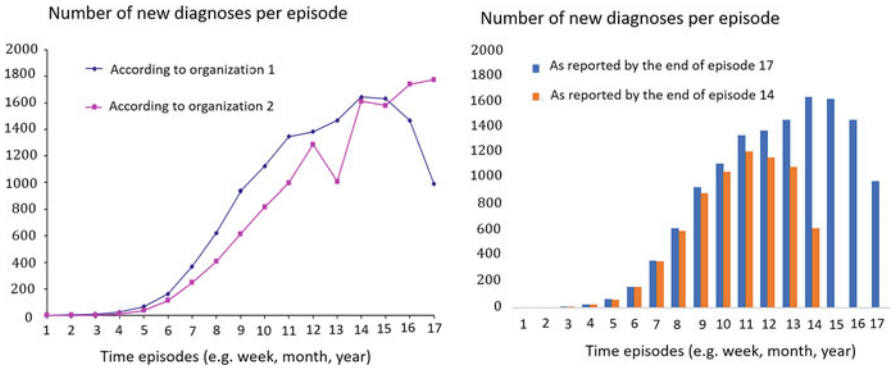


Fig. 7.4 Left: Illustration of two presentations of the "number of cases" per episode for Disease X from the same dataset; Right: Illustration of "number of new diagnoses" per episode being modified as new reports come in, due to reporting delays

(sometime later by a medical doctor among individuals who seek medical attention), and some subsequent clinical events during the follow-up such as morbidity, mortality or recovery. Other events are not directly relevant to epidemiology, but are equally important, not only for bookkeeping of the surveillance system but also serve as a bridge between the unobservable events and the observable data. These events may include the creation of this case (file) with the date, the arrival of the file to a local or a central registry with dates, the entry of the case into the database with date. However, the most important event, the infection and the associated time at infection, is not observable and is not documented in the case.

It is not uncommon in published reports and public health literature that disease trends are presented as "number of cases" over time when the meaning of a "case" is ambiguous. In Fig. 7.4 (for illustration purposes only), the actual disease, the time episodes, the population and the organizations involved are anonymized. The left panel shows two completely different trends about the same disease from the same data disseminated by two organizations and both use the word "case" ambiguously. Upon careful inspection, it turns out that each case-file has two dates associated with two different events: one is the date of a new diagnosis signed by a doctor and the other is the date on the stamp (which could be a computer digital signature) by the system when the file is entered into the registry. Organization 1 presents the trend of "new diagnoses" over time according to the most recent data; whereas, Organization 2 presents the trend of "new reports received by the system" over time. Both organizations call them "numbers of new cases."

The gap between the time of diagnosis and the time when this individual case is reported and entered into the database of a public health registry is called *reporting delay*. If this gap is ignorable (close to zero), then the two trend curves in the left panel of Fig. 7.4 should be nearly identical. If the gap is large but there is little variation from individual to individual, then the two curves should look alike, with a shift of fixed number of periods. If the gap has random variations, then the two curves will not look alike.

When there are substantial reporting delays, Organization 1 would argue that it is the new diagnosis, not the new entry to the system, that is relevant to the trend of the epidemic. However, the numbers of new diagnoses per episode will be modified as newly reported cases come in, especially for the recent episodes, as shown in the right panel of Fig. 7.4. Furthermore, there is usually a declining trend near the end, as cases with most recent time of diagnoses are still not yet reported. Therefore, the trend based on time of diagnoses must be statistically adjusted, especially for the recent past.

Organization 2 would argue that the number of new "cases" defined as new entries of the disease per episode is a static number. By the end of every episode, a new number is added to the database regardless of time at diagnoses and the reporting delay is irrelevant. After all, it presents useful trend information on the case-load seen by the registry. In our opinion, this could be misleading, especially when reporting delays are long and variable (Tariq et al. 2019).

While trends of specific events over time are meaningful, presenting trends of "cases" may not be. A good book-keeping practice in the registration system is to line-list all the events longitudinally for each reported individual whether the event is clinically relevant or not. All the documented events may serve some purposes. The researcher will decide which key event is the most relevant event to the question, but will also use some events and corresponding time lines to adjust biases such as reporting delay. The latter is part of statistical modelling. Naturally this will demand more resources and due diligence both on the system and the researcher.

In analyses involving diverse data sources, many agencies may contribute data originally collected for other purposes with more ambiguity in terminology. It will be more challenging for a researcher to get into the depth of each data source and statistically model the data-generating process. In the era of Big Data, will artificial intelligence be able to model all the data-generating processes and conduct model criticism? Quoting from Cox and Donnelly (2011):

> A large amount of data is in no way synonymous with a large amount of information. In some settings at least, if a modest amount of poor quality data is likely to be modestly misleading, an extremely large amount of poor quality data may be extremely misleading.

At Individual Levels

Phenomenological population models, especially those that mechanically model disease transmission dynamics, carry tacit assumptions at the level of individuals, such as the distributions of the latent periods and the infectious periods, as well as the infectious contact process. All events associated with these assumptions are not observable. One cannot pinpoint the time that an infection, which is the transfer of the infectious agent from an infected individual to a susceptible individual, occurs; nor can one ascertain the time when an infected individual is no longer latent and starts to be infectious. Therefore, there is no ideal data that can be directly used to validate these models.

In many diseases, the onset of clinical symptoms can be ascertained either precisely or within a narrow time interval. If the time of infection can also be ascertained within an acceptable range, for instance, through contact-tracing, then the *incubation period* is defined as the duration from the infection to the onset of symptoms and can be measured with some acceptable uncertainty. In some diseases, symptom onset may be used as a proxy for the beginning of the infectiousness and the *incubation period* may be a proxy for the *latent period*. However, there are diseases where a proportion of infected individuals may remain asymptomatic and are still able to transmit the infection.

The diagnosis of an infection, either due to onset of clinical symptoms or other screening/testing mechanisms, is always observable and ascertained to a specific point in time. This is also the event that generates most of the data. However, this event involves two mechanisms. One is driven by the progression of the disease natural history. The other is influenced by external factors. With respect to Fig. 7.4, Organization 1 is only partially right by saying that the new diagnoses are relevant to the trend of the epidemic. Before the time-series, represented as numbers of new diagnosis over time, become fully informative about the disease spread, statistical models are required to capture the entire data-generating process, incorporating the disease progression, the external factors such as how long since symptom onset and reasons for seeking diagnosis, as well as duration from the initial diagnosis to the entry of the case to the data registry.

For diseases with disease induced mortality, death caused by the disease is an observable event.

For models involving intervention, such as vaccination, treatment, isolation, etc., all of these are also observable.

Serial Interval, Generation Interval, Generation Time, and So On

In recent literature, these "intervals" have been widely cited, measured, and applied to outbreak investigations, especially during the early transmission phase. However, there is a lot of ambiguity. For instance, there are occasions that the same terminology is associated with two different definitions, whereas there are other occasions that the same definition is assigned to different terminologies by different researchers.

To our knowledge, the earliest definition of *serial interval* dates back to Hope Simpson (1948):

> The period from the observation of symptoms in one case to the observation of symptoms in a second case directly infected from the first is the (clinical) serial interval. It is an observable epidemiological unit.

Bailey (1975) wrote that:

> The period from the observation of symptoms in one case to the observation of symptoms in a second case directly infected from the first is the *serial interval*. Thus the serial interval

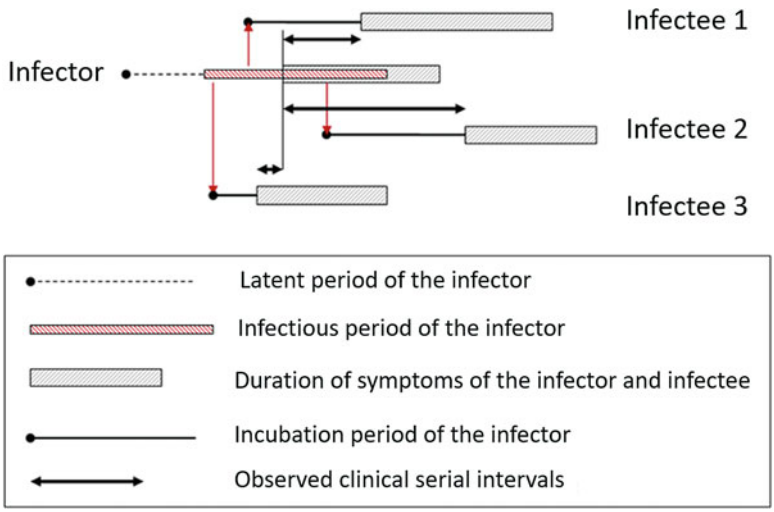


Fig. 7.5 A schematic presentation of three serial intervals produced by the same infector with 3 infectees

is the *observable epidemiological unit*, and it reflects to some extent the life cycle of the infectious organism. Nevertheless, it can not be readily related to the mechanism of transfer.

Decades later, Lipsitch et al. (2003) defined serial interval in the same way as that in Hope Simpson (1948) and Bailey (1975) but also specified that it is defined as the average between the two observed onset of symptoms. Lipsitch et al. (2003) applied such a measure during the outbreak investigation of the transmission of the severe acute respiratory syndrome (SARS) to estimate the basic reproduction R_0.

The features in the above definition of the serial interval include:

1. involving a pair of infected individuals, an infector, and an infectee;
2. observable;
3. book keeping device to track generations, White and Pagano (2008);
4. depending on the latent period and the infectious period (of the infector) as well as the incubation period (of the infectee).

For example, Infectees 1 and 3 in Fig. 7.5 are both infected before the symptom onset of the infector, with Infectee 1 following the natural sequence that its own clinical onset takes place after its infector's onset; whereas Infectee 3 has the reversed sequence, with its own clinical onset taking place before its infector's onset. This can happen in theory, if both the infectious periods and the incubation periods are highly variable.

There has been some confusion in the literature, as various terms have been used to refer to the same concepts. Before Hope Simpson, Pickles (1939) used the term *transmission interval* for what was later defined as serial interval with reference to empirical observations of a hepatitis epidemic in the United Kingdom (Nishiura 2010).

A different measure is the interval between the time of infection and time of transmission by linking two individuals, the infector and the infectee. This is formally named as the *transmission interval* in Fine (2003). Fine (2003) made it clear that (1) the transmission interval (i.e., the interval between successive infections) and (2) the clinical onset serial interval (i.e., the interval between successive clinical onsets) are different both conceptually and quantitatively. This distinction was also made clear in Svensson (2007).

A different name was given to the transmission interval as the *generation interval* in Wallinga and Lipsitch (2007), Roberts and Heesterbeek (2007), described as the duration between "the time of infection of an individual to the time of infection of a secondary case by that individual." Svensson (2007), Nishiura (2010), Kenah et al. (2008), and many others called it the *generation time.* Minor differences in the definitions among these authors are whether this interval is defined as a random variable, or is defined according to its mean value. Svensson (2007), from a sampling point of view, further points out the difference in distribution between the *primary generation time* as measured prospectively from the time of the infection of the infector to the transmission to the infectee, and the *secondary generation time* as measured retrospectively from the time of the transmission to the infectee to the infection of the infector.

The features in the above definition are that

1. involving a pair of infected individuals, an infector and an infectee;
2. both the infection of the infector and the passing of infection to an infectee are unobservable;
3. depending on the latent period and the infectious period.

The generation intervals (or generation times, transmission intervals) are distinguished from the serial intervals by definitions, but sometimes the two are used interchangeably in the literature. Figure 7.6 compares different time periods between the first patient and the second patient, adopted from the Field Epidemiology Manual published by the European Centre for Disease Prevention and Control (ECDC). In this diagram, generation time is synonymous to serial interval, both refer to the interval between successive clinical symptoms.

Anderson and May (1991) defined the *generation time* as the sum of the latent period and the infectious period: $T_E + T_I$. Daley and Gani (1999) defined the *average generation time* as $\mu_E + \mu_I/2$ where $\mu_E = E[T_E]$ and $\mu_I = E[T_I]$. According to strict arguments by Fine (2003), the sum of the latent and (part of) the infectious periods is concerned with the course of a single infection and is different from the interval between successive infections.

In a recent article, Champredon and Dushoff (2015) defined the *intrinsic generation interval* with its distribution defined by the p.d.f.

$$g(x) = \frac{\beta(x)A(x)}{R_0}, \ x > 0 \tag{7.18}$$

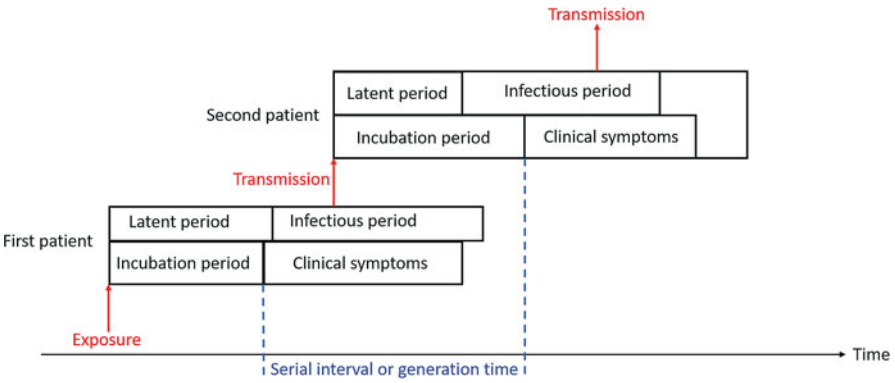


Fig. 7.6 Relationships between time periods

where $R_0 = \int_0^\infty \beta(x)A(x)dx$ is the basic reproduction number as formulated in Chap. 4 following the renewal-type equation (4.35). This concept is behind the developments of the theories in Wallinga and Lipsitch (2007) and Roberts and Heesterbeek (2007) that link the distribution of the generation intervals given by (7.18) to the estimation of R_0 with implicit assumptions that the generation intervals, according to their respective definitions, follow such a distribution. The intrinsic generation interval is defined along the course of a single infection. In particular, if $\beta(x) = \beta$, then $g(x) = A(x)/\mu_I$, $x > 0$. In structured models that involve latent periods T_E followed by infectious periods T_I, assuming independency, $A(x) = \int_0^x \overline{F}_I(x-u)f_E(u)du$, that gives

$$g(x) = \int_0^x f_E(u)\frac{\overline{F}_I(x-u)}{\mu_I}du. \tag{7.19}$$

The mean value is

$$\mu_G = \mu_E + \frac{1}{2}(1+\phi^2)\mu_I = \mu_E + \frac{\mu_I}{2} + \frac{var[T_I]}{2\mu}$$

as previously given as (4.48), where $\mu_E = E(T_E)$ and ϕ is the coefficient of variation of the infectious period T_I defined as the ratio of standard deviation to the mean. If the infectious periods are exponentially distributed with mean μ_I, then $\mu_G = \mu_E + \mu_I$ which is the mean generation time according to Anderson and May (1991). If the infectious periods are the same constant μ_I, then $\mu_G = \mu_E + \mu_I/2$ which is the mean generation time according to Daley and Gani (1999). If the infectious periods can be expressed as the sum of n independently and identically distributed exponential distributions with mean μ_I/n, then $\phi^2 = 1/n$ and $\mu_G = \mu_E + \frac{n+1}{2n}\mu_I$. This expression can be found in Roberts and Heesterbeek (2007).

Champredon and Dushoff (2015) further discussed forward generation interval from the perspective of an infector and backward generation interval from the perspective of infectee. We have discussed in Chap. 4 that the distribution given by (7.19) also has a sampling perspective. It coincides with that based on the convolution of the latent period T_E and the equilibrium distribution given by p.d.f. $f_W(x) = \overline{F}_I(x)/\mu_I$, under suitable assumptions concerning equilibrium conditions of the epidemic at the population level. In this case, it is appropriate to define the (intrinsic) generation time as

$$T_G = T_E + W.$$

It has the sampling property that an arbitrary observer makes a snapshot sample at an arbitrary time. All individuals who are "currently infectious" form a prevalence cohort. The observer looks backward to the time of infection and measures the time from infection to the observation time. The distribution has the p.d.f. given by (7.19).

In summary,

1. Serial intervals are between observable events, subject to observation errors and time-length bias (to be discussed next), but cannot be readily related to the mechanism of transmission. Their distributions, even correctly estimated by data, may not be used to approximate the distributions of the transmission intervals or the intrinsic generation times.
2. Transmission intervals (also known as generation intervals, generation times) are not directly observable. They are further distinguished by forward measuring from the time of an infector to the time of transmission to an infectee, versus backward measuring from the time of infection of an infectee to the time of infection of its infector. These two measurements follow different distributions (Svensson 2007).
3. The intrinsic generation interval includes the definitions given by Anderson and May (1991) and Daley and Gani (1999) as special cases. By definition, it is related to the basic reproduction number R_0 that carries information about disease transmission. This relation is established under equilibrium conditions (implicitly assumed) as R_0 itself is also defined under such conditions. This is more apparent in a structured model with latent and infectious periods, where the intrinsic generation time can be written explicitly as $T_G = T_E + W$, where W corresponds to the equilibrium distribution of the infectious periods. The intrinsic generation intervals do not involve pairs of infectors and infectees and are not observable.

7.3 Time-Length Bias

The time-length bias discussed here is in the same nature of the famous "survivorship bias." During World War II, researchers from the Centre for Naval Analyses conducted a study of the damage made to planes that had returned from missions.

The statistician A. Wald noticed that the study only considered the planes that had survived their missions. Those that had been shot down were not present for the damage assessment. The holes in the returning aircraft represents areas where the aircraft could take damage and return home safely. Wald (1943) proposed that the Navy reinforce areas where the returning planes were unscratched, since those areas, if hit, would cause the plane to be lost. The same type of bias also applies in observational data in the study of disease outbreaks.

Observational data during an infectious disease outbreak are often length-biased with respect to key epidemiological durations, such as the latent periods, infectious periods, incubation periods, generation times, etc. In some cases, individuals associated with longer durations are more likely to be included in the data. In other cases, individuals associated with shorter durations are more likely to be included in the data. At the population level, length biases at individual levels further lead to mis-interpretation of the disease trends. Figure 7.4 has an illustration of disease trends by date of onset, affected by reporting delays.

For a comprehensive review on length-biased sampling and length-biased distribution, we recommend Chapter 1 of Qin (2017).

Disease progression within an infected host involves sequences of events. Each pair of successive events is composed of an *initiating event* that leads to a *subsequent event* over a random duration $X \geq 0$. An individual is denoted by the index i. We used $T_i^{(1)}$ for the time of onset of the initiating event and $T_i^{(2)}$ for the time of onset of the subsequent event. The duration of interest is $X_i = T_i^{(2)} - T_i^{(1)}$.

7.3.1 Prevalence Cohorts and Left-Truncation

The prevalence cohort as illustrated in Fig. 4.8 leads to length-bias that systematically includes individuals with longer durations. Assuming that X_i among individuals are (in theory) i.i.d. with p.d.f. $f_X(x)$, the distribution of X_i as observed in the prevalence cohort is length biased because the arbitrary sampling time t must satisfy $T_i^{(1)} < t \leq T_i^{(2)}$. The length biased duration is denoted by $X_i^{(B)}$. We also write $X_i^{(B)} = W_i + V_i$, where $W_i = t - T_i^{(1)}$ and $V_i = T_i^{(2)} - t$.

We further assume that

1. The occurrence of the initiating events, which is a stochastic process, follows constant incidence rate (i.e., the equilibrium condition).
2. X_i is independent of the occurrence of the initiating event.

Under these conditions, $X_i^{(B)}$, W_i, and V_i have the following equilibrium distributions (Wang 2005) with p.d.f.s

$$X_i^{(B)} \sim \frac{x f_X(x)}{\mu}, \text{ and } W_i \sim V_i \sim \frac{\overline{F}_X(x)}{\mu} \tag{7.20}$$

where $\mu = E[X] = \int_0^\infty x f_X(x)dx = \int_0^\infty \overline{F}_X(x)dx$.

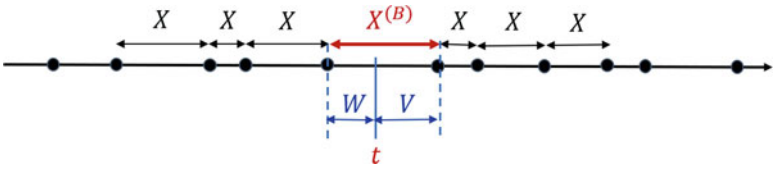


Fig. 7.7 Illustration of a repeated testing scheme on an infectious disease

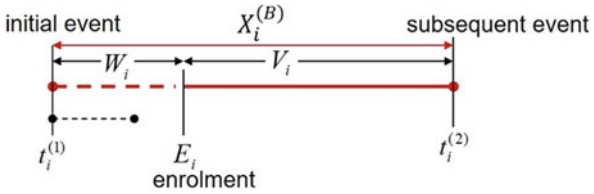


Fig. 7.8 Illustration of the observation scheme with left-truncation. The individual with short duration (dotted line) is not included at the time of enrollment

The same problem is formulated differently as illustrated in Fig. 7.7. Consider a repeat testing scheme for an infectious disease such as HIV. Individuals get tested repeatedly with i.i.d. inter-testing intervals X with p.d.f. $f_X(x)$. Denote $X^{(B)}$ the special interval between the last negative and the first positive tests. Assuming that the epidemic in the population is at equilibrium so that the occurrence time t of the "sero-conversion event" is distributed uniformly between the two tests. This interval is length-biased with equilibrium distributions given by (7.20). These distributions are applied for designing better repeated testing algorithms to reduce the prevalence of not yet diagnosed HIV infections (Yan and Zhang 2018).

A more general setting is the left-truncation in survival analysis. The initial events occur over time t following a random process with intensity $\lambda(t)$. For each individual i, the subsequent event occurs following the initial event after a random duration X_i. Assuming $X_i's$ are i.i.d. as the random variable X, and the objective is to estimate this distribution. However, data are collected by enrollment. The time at enrollment E_i for individual i must satisfy that the initial event has taken place while the subsequent event has not. Thus individuals with longer X_i have more chance to be enrolled. Observed data are length-biased following the distribution $X^{(B)}$. Each individual $X_i^{(B)}$ arises from the conditional distribution of X given $X \geq W_i$. An illustration is given in Fig. 7.8.

The observed part from enrollment until the end point is V_i which has conditional the survival function and the conditional p.d.f.

$$\overline{F}(v_i|w_i) = \frac{\overline{F}(w_i + v_i)}{\overline{F}(w_i)} = \frac{\overline{F}(t_i^{(2)} - t_i^{(1)})}{\overline{F}(E_i - t_i^{(1)})},$$

$$f(v_i|w_i) = \frac{f(w_i + v_i)}{\overline{F}(w_i)} = \frac{f(t_i^{(2)} - t_i^{(1)})}{\overline{F}(E_i - t_i^{(1)})}.$$

This is the residual life distribution discussed in Sect. 2.3 in Chap. 2. If the time of initial event $t_i^{(1)}$ cannot be ascertained but follows a random process with intensity $\lambda(t)$ until enrollment, then the p.d.f. of V_i is

$$f(v_i|w_i) = \frac{\int_{-\infty}^{E_i} \lambda(t) f(t_i^{(2)} - t)dt}{\int_{-\infty}^{E_i} \lambda(t)\overline{F}(E_i - t)dt}.$$

If $\lambda(t) = \lambda$ is constant, then the above becomes

$$\begin{aligned} f(v_i|w_i) &= \frac{\int_{-\infty}^{E_i} f(t_i^{(2)} - t)dt}{\int_{-\infty}^{E_i} \overline{F}(E_i - t)dt} \\ &= \frac{\int_{v_i}^{\infty} f(x)dx}{\int_0^{\infty} \overline{F}(x)dx} = \frac{\overline{F}(v_i)}{\mu} = f_V(v_i), \end{aligned}$$

where $v_i = t_i^{(2)} - E_i$ and $\mu = \int_0^{\infty} \overline{F}(x)dx$. In this case, we have recovered the equilibrium distribution. Data arising from left-truncated data under equilibrium can be used to estimate the distribution $\overline{F}_X(x)$ because the distribution of the observed part $v_i = t_i^{(2)} - E_i$ contains all the information, independent of the truncation time $w_i = E_i - t_i^{(1)}$.

If $\lambda(t)$ is constant but the time $t_i^{(1)}$ can be all ascertained (retrospectively) upon enrollment, then $w_i = E_i - t_i^{(1)}$ is observable and data consist of a pair (x_i, w_i) for each individual where $x_i = t_i^{(2)} - t_i^{(1)}$. Do data have enough information to estimate the distribution of X without modelling $\lambda(t)$?

The good news is that the hazard function under left-truncation is invariant for $x > w$. For each individual, conditioning on $X > w$, the hazard function calculated from the conditional distribution is

$$h_{\text{left-truncation}}(x|w) = \frac{f(x)/\overline{F}(w)}{\overline{F}(x)/\overline{F}(w)} = \frac{f(x)}{\overline{F}(x)} = h(x), \ x > w.$$

All the survival analysis models and methods focusing on hazard rate estimation and comparison (e.g., proportional hazard regression model, other hazard based models, etc.) apply to data with left-truncation. Existing software may be directly applied with minimum modification (e.g., SAS, S-plus, R, etc.). However, the identifiable part of the survival function is the residual survival function

$$\overline{F}(x|w) = \exp\left(-\int_w^x h(u)du\right), \ x > w, \tag{7.21}$$

not the entire distribution $\overline{F}(x) = \exp\left(-\int_0^x h(u)du\right)$.

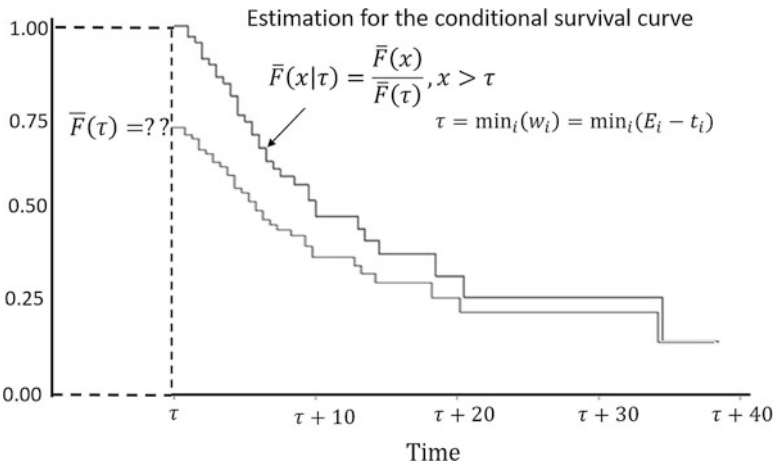


Fig. 7.9 Non-parametrically, left-truncated data are unable to identify the entire survival function. It is only possible to estimate the residual survival function conditioning on $X > \tau$

The above argument can be demonstrated by non-parametric methods. The Kaplan-Meier estimation in survival analysis is based on the empirical estimation of the hazard function at discrete time points. If there is no left-truncation in data, the discrete hazard function yields the empirical survival function as a decreasing step function starting $\overline{F}^{\mathrm{KM}}(0) = 1$. For left-truncated data in pairs (x_i, w_i) and let $\tau = \min(w_i)$, because the hazard function is invariant under left-truncation, the Kaplan-Meier estimation can be still applied (as built into various statistical software packages). The decreasing step function is now understood as the non-parametric estimation for the residual survival function

$$\overline{F}(x|\tau) = \frac{\overline{F}(x)}{\overline{F}(\tau)}, \ x \geq \tau = \min(w_i),$$

starting at $\overline{F}^{\mathrm{KM}}(\tau|\tau) = 1$. The non-identifiable part is $\overline{F}(\tau) \leq 1$. This is illustrated in Fig. 7.9.

There is a rich literature concerning lifetime and life history data with left-truncation. We recommend the books: Lawless (2003) and Cook and Lawless (2007, 2018).

7.3.2 Retrospective Ascertainment and Right-Truncation

As illustrated in Fig. 7.8, in left-truncated data, the inclusion criteria is that, at time of enrollment, the initiating event has occurred but the subsequent event has not. This observation scheme produces time-length bias in favor of longer duration.

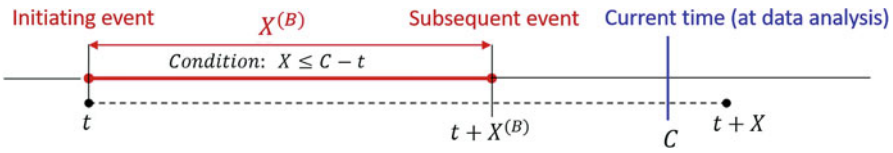


Fig. 7.10 Illustration of the observation scheme with right-truncation. The individual with long duration (dotted line) is not included by time C

The opposite observation scheme is illustrated in Fig. 7.10. The inclusion criteria is based on the occurrence of the subsequent event. Individuals whose initiating events have occurred, but the subsequent events have not, are not included by the time of data analysis. This observation scheme produces time-length bias in favor of shorter duration.

Data arise from the conditional distribution

$$F^*(x|\tau) = \frac{F_X(x)}{F_X(\tau)},\ 0 \le x \le \tau = \max_i \left\{ C - t_i^{(1)} \right\}. \tag{7.22}$$

The probability of inclusion is the cumulative probability $F_X(\tau) = \Pr\{X \le \tau\}$, which itself is the object of the estimation. If τ is sufficiently large, F^* will be a good approximation to F_X. However, sufficiently large τ implies that C is large and one needs to wait for much longer time before starting the analysis. This is not desirable for an emerging infectious disease where one needs information quickly.

Assessment of the Incubation Period Distribution

The incubation period is defined as the duration from the time at infection to the time of onset of clinical symptoms. For acute infectious diseases such as the severe acute respiratory syndrome (SARS) in 2003, knowledge of the incubation period distribution must be generated quickly at the very early stage of the epidemic for guidance to determine the length of quarantine of individuals exposed to infection sources. If a potentially exposed individual has not shown symptoms of the disease after x days of quarantine, the risk of releasing this individual into the susceptible population who subsequently becomes symptomatic (and infectious) is the survivor function of the incubation period. For other infectious diseases that are also chronic, such as HIV/AIDS, the incubation period distribution is the crucial link between an observable event based on clinical presentation such as the diagnosis of AIDS and the unobservable event based on disease transmission such as the infection of HIV.

The incubation period distribution can be estimated using standard survival analysis techniques by following selected cohorts of infected individuals whose dates of exposure to the infection sources are known. However, in emerging new infectious diseases, such as AIDS in the 1980s, SARS in 2003, the pandemic H1N1 (pH1N1) influenza in 2009, knowledge of this distribution must be generated quickly before any formal cohort follow-up studies become feasible.

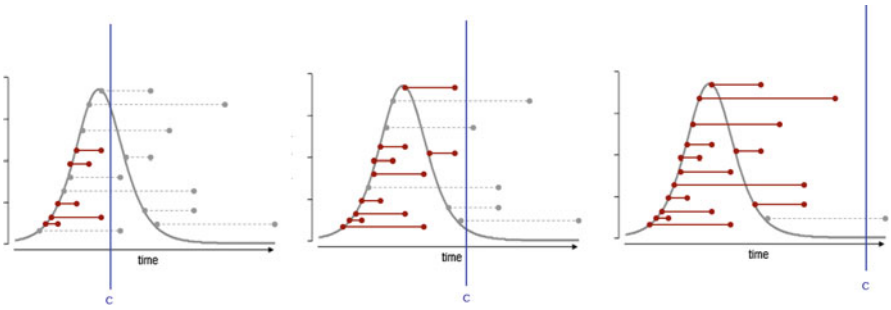


Fig. 7.11 Schematic illustration of retrospectively ascertained incubation periods analyzed during, immediately after and long after a disease outbreak

Early data often arise from selected individuals diagnosed with symptoms with retrospectively ascertained dates of exposure to the infection sources. These early assessments tend to be shorter than assessments made at some later time during the epidemic, which may still be shorter than the actual distribution. This is schematically illustrated in Fig. 7.11. This phenomenon is often misinterpreted by the media and the public to make hypotheses about the mutation of the pathogen, rather than time-length bias in data.

Example 29 This example is well cited in the literature, such as Examples 3.5.3 and 4.3.3 in Lawless (2003). For the incubation period from HIV infection to the development of AIDS illnesses, early studies were based on retrospectively ascertained data for blood transfusion-associated cases, assembled by the U.S. Centers for Disease Control, with transfusion as the only known risk factor. The data were studied by Lui et al. (1986), Medley et al. (1987), Lagakos et al. (1988), Kalbfleisch and Lawless (1989), among many others. A comprehensive survey of various statistical methods of these studies was included in Chap. 4 of Brookmeyer and Gail (1994). Lui et al. (1986) published the earliest results based on data available as of April 1985. The authors acknowledged the bias due to right-truncation and illustrated that the sample average was only 2.6 years based on the naïve approach, whereas based on the conditional distribution (7.22) along with a Weibull distribution model, the estimated mean incubation period was 4.5 years. Kalbfleisch and Lawless (1989) analyzed the data as reported by July 1986 with median estimation approximately 8.5 years.

For right-truncated data regarding the incubation period, uncertainty with respect to the time of infection, $T_i^{(1)}$, is a common problem. Tuite et al. (2010) analyzed 3152 laboratory confirmed pH1N1 cases in Ontario with symptom onset between April 13 and June 20, 2009. A subset of 316 cases containing sufficient information on exposure date and disease onset were used to estimate the incubation period distribution. The dates of exposure were imputed as the midpoint between the earliest and the most recent dates of exposure. Farewell et al. (2005) studied a subset of 128 cases out of a total of 1755 reported cases in a Hong Kong Hospital

Authority database during the 2003 SARS outbreak. The data consist of the date of the appearance of the symptoms of SARS, but the dates of exposure can be only ascertained to an earliest and latest possible date of exposure. The authors explored statistical methodology for retrospective data with the timing of the initiating event being uncertain, except for lying in a given time interval, and what might reasonably be inferred about such a maximum incubation time based on the moderately sized samples that would typically be available in the early course of an epidemic.

Assessment of the Reporting Delay and Estimation of the Number of Occurred But Not Yet Reported Events

In most public health disease surveillance systems, data are compiled upon the *reporting* of the diagnosis of the disease. Official reports often present aggregated counts based on the number of new diagnoses or the disease onset per unit of time. It is thought that these "epicurves" represent, to some degree, the epidemiology of the disease transmission. However, there is a time-length bias in under-reporting: the more recent the diagnosis (or onset), the more severe is the under-reporting. This is reflected by an artificial decline of trend near the end of the time-series. This time-length bias is more profound when data are compiled and analyzed when the outbreak is still ongoing, but it is also more important to present real time trends during the outbreak investigation.

Figure 7.12 illustrates the epicurves of the severe acute respiratory syndrome (SARS) during the spring of 2003 in Canada, Singapore, and Hong Kong, compiled from publicly available information from respective government websites by dates of onset. We immediately see the same phenomenon as illustrated in the right panel of Fig. 7.4. Figure 7.13 illustrates the phenomenon again using epicurves presented by the Ministry of Health of Mexico, by compiling the numbers of onset of symptoms of the H1N1 influenza outbreak in Mexico from April to December of 2009, as officially released by different dates.

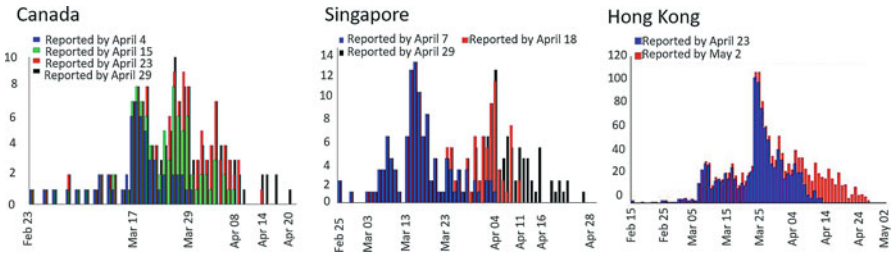


Fig. 7.12 Illustration of SARS Epicurves by dates of onset as reported on different dates during the 2003 SARS outbreak. (Sources: Health Canada; Singapore Ministry of Health; Department of Health, Hong Kong, China)

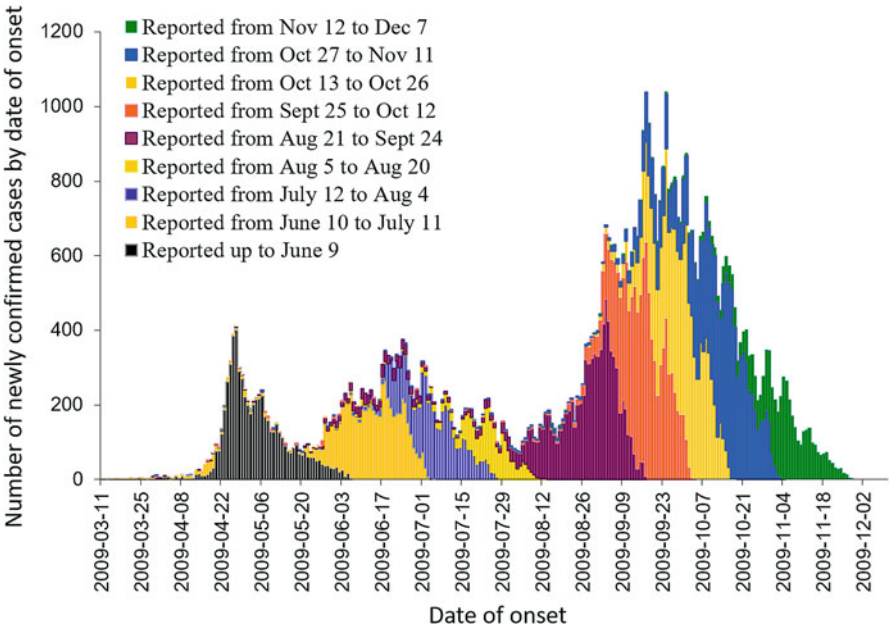


Fig. 7.13 Illustration of the H1N1 influenza outbreak by dates of onset in Mexico as reported on different dates during the 2009 outbreak. (Source: Ministry of Health of Mexico, http://portal.salud.gob.mx/contenidos/noticias/influenza/estatisticas.html)

Such a phenomenon is caused by the reporting delay. Reporting delays are measured at individual levels as the gap between the time of disease onset (or diagnosis) and the time when this individual case is reported and entered into the database of a public health registry.

Let C be the most current time when data are analyzed. We use a generic term "event" for the initiating event, such as disease onset (or diagnosis). The aggregated counts are denoted by

$$N(t; C) = \#\{\text{events occurred at time } t \text{ as reported by time } C\},$$
$$N(t) = N(t; \infty) = \#\{\text{events occurred at time } t\}.$$

$N(t; C)$ is always a proportion of $N(t)$ and the adjustment for reporting delay reduces to the problem of estimating this proportion.

For simplicity, let us assume (for the time being) that the reporting delay can be represented by a random variable X, which is i.i.d. among all individuals. The cumulative distribution is $F(x) = \Pr(X \le x)$. Then

$$F(C - t) = \Pr(X \le C - t)$$

has the same meaning as the probability of events that happened at time $t \leq C$ have been reported by C. Therefore,

$$N(t) = \frac{N(t; C)}{F(C - t)}$$

and the reporting delay adjustment becomes the problem of estimating $F(x)$.

Data on reporting delay is always right-truncated, because only upon reporting can one retrospectively measure the delay (see Fig. 7.10).

There are two levels of time-length biasness involved. Reporting delays produce the time-length bias in aggregated counts over time at the population level with an artificial declining trend near the end of the time-series. The observed reporting delays are also length-biased due to right-truncation. Individuals associated with shorter delays are over-represented in data.

The right-truncation bias in reporting delays is much less recognized than the reporting delay phenomenon at the population level, because frontline workers who make diagnoses, reports, and analyses do not see long delays, even after using naive statistical analyses such as summary statistics directly on observations. Results from formal statistical analysis taking into account right-truncation are often counter-intuitive.

When the first reporting delay adjusted trend of the diagnoses of AIDS in Canada was published in 1993 (Fig. 7.14), it was met with much criticism because it implied a much longer delay most public health workers in the field than felt. It was also dramatically different from a naive analysis based on summary statistics directly calculated from measured delays: median around 1.6 months and only 3% of all individuals were reported after 14 months since diagnoses. The reporting delay

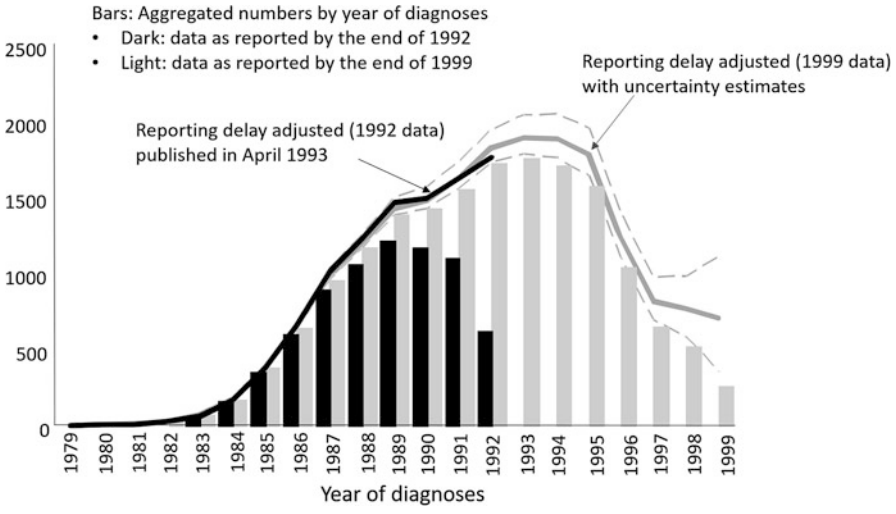


Fig. 7.14 Reporting delay adjusted trends of number of AIDS diagnoses by year from *AIDS in Canada, AIDS Surveillance Report (Health Canada)* in selected years

adjusted trend in Fig. 7.14 based on diagnoses to the end of 1992 suggested a delay with a median of at least 9 months. Despite the doubts and criticisms, history (e.g., reported AIDS incidence by year of diagnosis by the end of 1999) showed that the counter-intuitive delay-adjusted trend was more realistic.

A Simple Method to Estimate Reporting Delay Adjusted Incidence Trends We take a discrete time framework. Let C denote the "current time" which is the time when data are used for analysis. Let $t = 0, 1, 2, \cdots, C$ denote the time of the occurrence of the events under the study where $t = 0$ is the earliest possible time when the events could happen in the population. Let $x = 0, 1, 2, \cdots, C$ denote the report delay and $x = 0$ means that the event is reported within the same time unit. In this setting, data can be grouped into counts

$$n_{tx} = \#\{\text{events occurred at time } t \text{ and reported at time } t + x\},\ x \le C - t.$$

These counts are then arranged into a 2-way contingency table of which the lower triangle remains empty due to right-truncation, as represented in Table 7.1. The row totals are

$$N(t; C) = \sum_{x=0}^{C-t} n_{tx},\ t = 0, 1, \cdots, C.$$

Clearly, as t is getting closer to the current time C, the more likely that $N(t; C)$ under counts the true number of events. The column totals are

$$n_{+x} = \sum_{t=0}^{C-x} n_{tx} = \sum_{i} I(x_i = x).$$

Table 7.1 The upper-triangle table for $\{n_{tx}\}$ with column totals represent the number of events with $X = x$ and row totals represent the number of events over time as reported by C

	Reporting delay x							Row totals
	0	1	$\cdots$	x	$\cdots$	$C-1$	C	
0	n_{00}	n_{01}	$\cdots$	n_{0x}	$\cdots$	$n_{0,C-1}$	n_{0C}	$N(0; C)$
1	n_{10}	n_{11}	$\cdots$	n_{1x}	$\cdots$	$n_{1,C-1}$		$N(1; C)$
$\vdots$	$\vdots$	$\vdots$		$\vdots$				
t	n_{t0}	n_{t1}	$\cdots$	n_{tx}				$N(t; C)$
$\vdots$	$\vdots$	$\vdots$		$\vdots$				
$C-x$	$n_{C-x,0}$	$n_{C-x,1}$	$\cdots$	$n_{C-x,x}$				
$\vdots$	$\vdots$	$\vdots$						
$C-1$	$n_{C-1,0}$	$n_{C-1,1}$						
C	n_{C0}							$N(C; C)$
Col. totals		n_{+1}		n_{+x}			n_{+C}	

representing the total number of events with delay $X = x$ as observed in the data. Let's also denote N_{+x} as the total number of cases with delay $X \leq x$, among events during times $t = 0, 1, \cdots, C - x$, which is the sum of numbers in the 2-way contingency table inside the rectangle area defined by $t = 0, \cdots, C - x$ and $X = 0, \cdots x$:

$$N_{+x} = \sum_{t=0}^{C-x} \sum_{j=0}^{x} n_{tj} = \sum_{i} I(x_i \leq x \leq \tau_i).$$

The ratio

$$g(x) = \frac{n_{+x}}{N_{+x}}, \quad x = 1, \cdots, C$$

gives an estimate for the proportion of events with delay $X = x$ out of those with delay $X \leq x$. Therefore, $1 - g(x)$ gives the estimate for

$$\frac{\#\{X \leq x\} - \#\{X = x\}}{\#\{X \leq x\}} = \frac{\#\{X \leq x - 1\}}{\#\{X \leq x\}}$$

which is the proportion of events with delay $X \leq x - 1$ out of those with delay $X \leq x$. Rewriting $x = C - t + 1$, $1 - g(C - t + 1)$ gives an estimate for

$$\frac{\#\{X \leq C - t\}}{\#\{X \leq [C + 1] - t\}}$$

which is the proportion of events at time t and reported by time C (current), out of those at time t and reported by $C + 1$. This is the estimate for the conditional probability

$$\Pr\{X \leq C - t | X \leq C + 1 - t\} = \frac{F(C - t)}{F([C + 1] - t)}, \quad 1 \leq t \leq C.$$

and hence

$$N(t; C) = N(t; C + 1) \times \frac{F(C - t)}{F([C + 1] - t)} = N(t; C + 1) \times [1 - g(C - t + 1)].$$

Therefore, a one-step prediction for the number of events that occurred at time t as seen in by time $C + 1$, based on current observation $N(t; C)$ is established as

$$\begin{aligned} \widehat{N}(t; C + 1) &= \frac{N(t; C)}{1 - g(C - t + 1)} \\ &= N(t; C) \frac{N_{+,C-t+1}}{N_{+,C-t+1} - n_{+,C-t+1}}, \quad 1 \leq t \leq C. \end{aligned}$$

This also predicts the off-diagonal elements of $\{n_{tx}\}$ in Table 7.1 as

$$\widehat{n}_{t,C-t+1} = N(t;C)\frac{n_{+,C-t+1}}{N_{+,C-t+1} - n_{+,C-t+1}}, \quad 1 \le t \le C.$$

In particular, $\widehat{n}_{1C} = N(1;C)\frac{n_{+C}}{N_{+C}-n_{+C}}$ and $\widehat{n}_{C1} = N(C;C)\frac{n_{+1}}{N_{+1}-n_{+1}}$.

Similarly, it can be shown that

$$\widehat{N}(t;C+2) = \frac{N(t;C)}{[1-g(C-t+1)][1-g(C-t+2)]}, \quad 2 \le t \le C.$$

gives a 2-step prediction for the number of events that occurred at time t as seen by time $C+2$, where the denominator is the estimate for $\Pr\{X \le C-t | X \le C+2-t\} = \frac{F(C-t)}{F([C+2]-t)}$. It further predicts the elements in the lower triangle of Table 7.1 as

$$\widehat{n}_{t,C-t+2} = \widehat{N}(t;C+1)\frac{n_{+,C-t+2}}{N_{+,C-t+2} - n_{+,C-t+2}}, \quad 2 \le t \le C.$$

In particular, $\widehat{n}_{2C} = \widehat{N}(2;C+1)\frac{n_{+C}}{N_{+C}-n_{+C}}$ and $\widehat{n}_{C2} = \widehat{N}(C;C+1)\frac{n_{+2}}{N_{+2}-n_{+2}}$.

Continuing, the maximum is to predict C steps for $t = C$; $C-1$ steps for $t = C-1$; and so on, until all the empty elements in the lower triangle of Table 7.1 are filled by predicted values according to the iterative formulae

$$\widehat{n}_{t,C-t+k} = \widehat{N}(t;C+k-1)\frac{n_{+,C-t+k}}{N_{+,C-t+k} - n_{+,C-t+k}}, \quad k \le t \le C$$

and $k = 1, \ldots, C$. Therefore,

$$\frac{N(t;C)}{[1-g(C-t+1)][1-g(C-t+2)]\cdots[1-g(C)]}, \quad 1 \le t \le C$$

gives the farthest prediction for the number of events that occurred at time $t \le C$ based on current data as seen in the future as data allow, because the longest observable reporting delay is C. The denominator is the estimate for $\Pr\{X \le C-t | X \le C\} = \frac{F(C-t)}{F(C)}$. If it is appropriate to assume $F(C) \approx 1$, the reporting delay adjustment can be written as

$$\widehat{N}(t) = \frac{N(t;C)}{\prod_{x=C-t+1}^{C}[1-g(x)]} \approx \frac{N(t;C)}{\widehat{F}(C-t)}. \tag{7.23}$$

Brookmeyer and Gail (1994) contain a detailed chapter on reporting delays in AIDS surveillance systems. We adopt their example below for illustration of the algorithm.

Table 7.2 The upper-triangle table Table 7.1 filled by numbers from Brookmeyer and Gail (1994) with $C = 4$

Time of diagnosis	Reporting delay x						# diagnoses as reported $N(t; C)$
	0	1	2	3	4		
0	50	20	10	6	2		88
1	100	55	20	12		(88)	187
2	171	115	45		(273)		331
3	207	118		(586)			325
4	220		(836)				220
Col.totals n_{+x} : $x = 1, ..., 4$		308	75	18	2		

Example 30 In Table 7.2, the column totals give $n_{+x} = \sum_{t=0}^{C-x} n_{tx}$:

$$n_{+1} = 308,\ n_{+2} = 75,\ n_{+3} = 18,\ n_{+4} = 2.$$

Numbers in brackets are $N_{+x} = \sum_{t=0}^{C-x} \sum_{j=0}^{x} n_{tj}$:

$$N_{+1} = 836,\ N_{+2} = 586,\ N_{+3} = 273,\ N_{+4} = 88.$$

These quantities yield:

$$g(x) = \left(\frac{308}{836}, \frac{75}{586}, \frac{18}{273}, \frac{2}{88}\right).$$

The reported number diagnoses at the most recent time $t = 4 = C$ is $N(4; 4) = 220$. The probability of being reported by time $C = 4$ is

$$[1 - g(1)]\,[1 - g(2)]\,[1 - g(3)]\,[1 - g(4)] = 0.503.$$

According to (7.23), we adjust this number as $\widehat{N}(4) = 220/0.503 = 437$. Similarly, we get

$$\widehat{N}(1) = \frac{N(1; 4)}{1 - g(4)} = \frac{187}{0.977} = 191$$

$$\widehat{N}(2) = \frac{N(2; 4)}{[1 - g(3)]\,[1 - g(4)]} = \frac{331}{0.913} = 363$$

$$\widehat{N}(3) = \frac{N(3; 4)}{[1 - g(2)]\,[1 - g(3)]\,[1 - g(4)]} = \frac{325}{0.796} = 408$$

$$\widehat{N}(4) = \frac{N(4; 4)}{[1 - g(1)]\,[1 - g(2)]\,[1 - g(3)]\,[1 - g(4)]} = \frac{220}{0.503} = 437.$$

Caveats This simple method has a few caveats.

1. It only provides partial reporting delay adjustment if $F(C) < 1$. The assumption $F(C) \approx 1$ may be suitable for sufficiently large C so that the system can capture very long delays.
2. It assumes that the reporting delays X_i are i.i.d. among all individuals. This is debatable. It also assumes that the distribution is $F(x)$ is stationary, that is, it does not depend on the time t when the events occur. The latter is mostly untrue in practice because the system can improve, deteriorate, or fluctuate over time. The distribution is most likely to be non-stationary, as $F(x|t)$. There is a rich literature on reporting delay adjustments applied to different disease reporting systems, with statistical models designed to handle non-stationary reporting delays distributions. For example, Kalbfleisch and Lawless (1991) and Lawless (1994).

Likelihood Based Approaches for Analyzing Right-Truncated Data

Here we present formal likelihood based approaches for statistical inferences of the distribution $F(x) = \Pr(X \le x)$ when X is right-truncated.

The reverse hazard function is defined by

$$h_X^r(x) = \frac{f_X(x)}{F_X(x)}, \tag{7.24}$$

where $f_X(x) = \Pr\{X = x\}$ when X is discrete and $f_X(x)$ is the p.d.f. of X when X is continuous. The cumulative distribution function $F(x)$ is uniquely determined by $h_X^r(x)$ through

$$F_X(x) = \begin{cases} \prod_{l=x+1}^{\infty}\{1 - h_X^r(l)\}, & X \text{ discrete} \\ \exp\left\{-\int_x^{\infty} h_X^r(u)du\right\}, & X \text{ continuous} \end{cases}, \quad x > 0.$$

These are analogous to the relationships between the survival function and the hazard function, such as $\overline{F}_n = \prod_{j=0}^{n-1}(1 - h_j)$ (3.1) in discrete case, and $\overline{F}_X(x) = \exp\left(-\int_0^x h_X(u)du\right)$ (2.5) in continuous case.

Lagakos et al. (1988) proposed to use (7.24) as the key quantity for statistical inference for right-truncated data. If $h_X^r(x)$ is identifiable for $0 \le x \le \tau$, then the conditional distribution (7.22) is identifiable through

$$\frac{F_X(x)}{F_X(\tau)} = \begin{cases} \prod_{l=x+1}^{\tau}\{1 - h_X^r(l)\}, & X \text{ discrete} \\ \exp\left\{-\int_x^{\tau} h_X^r(u)du\right\}, & X \text{ continuous} \end{cases}, \quad 0 \le x \le \tau.$$

Suppose that a sample of right-truncated data is represented by (x_i, τ_i), $i = 1, \ldots, n$ subject to the condition $x_i \leq \tau_i$ is observed, in which τ_i is the right-truncation time for individual i. This corresponds to Fig. 7.10, $\tau_i = C - t_i$.

If X is continuous, assuming X follows the distribution $F(x; \underline{\theta})$ which is fully specified by a vector of parameters $\underline{\theta}$, one may consider to model the reverse hazard function parametrically as $h^r_X(x; \underline{\theta})$. One may consider maximizing the (conditional) likelihood given by

$$L(\underline{\theta}) \propto \prod_{i=1}^{n} \frac{f(x_i; \underline{\theta})}{F(\tau_i; \underline{\theta})} = \prod_{i=1}^{n} h^r_X(x_i; \underline{\theta}) \exp\left\{ - \int_{x_i}^{\tau_i} h^r_X(u; \underline{\theta}) du \right\}. \tag{7.25}$$

In the discrete time framework, one may consider the likelihood function

$$L \propto \prod_{i=1}^{n} \frac{f(x_i)}{F(\tau_i)} = \prod_{i=1}^{n} h^r_X(x_i) \prod_{l=x_i+1}^{\tau_i} \{1 - h^r_X(l)\} \tag{7.26}$$

and treat each value $h^r_X(x)$ for $x = 0, \ldots, \tau = \max(\tau_i)$ as a "parameter." In this case, nonparametric method can be only used to estimate $F(x|\tau) = F_X(x)/F_X(\tau) = \prod_{l=x+1}^{\tau} \{1 - h^r_X(l)\}$, $0 \leq x \leq \tau$, because $h^r_X(x)$ is only defined up to τ.

The Non-parametric Maximum Likelihood Estimation Lawless (1994) re-wrote (7.26) as

$$\prod_{i=1}^{n} h^r_X(x_i) \prod_{l=x_i+1}^{\tau_i} \{1 - h^r_X(l)\} = \prod_{x=1}^{\tau} \left[h^r_X(x)^{n_{+x}} \{1 - h^r_X(x)\}^{N_{+x} - n_{+x}} \right]. \tag{7.27}$$

The maximum likelihood estimate is

$$\widehat{h^r}_X(x) = \frac{n_{+x}}{N_{+x}} = \frac{\sum_i I(x_i = x)}{\sum_i I(x_i \leq x \leq \tau_i)}, \; x = 0, 1, \cdots, \tau = \max(\tau_i). \tag{7.28}$$

which has been written as $g(x) = \frac{n_{+x}}{N_{+x}}$ previously in the simple reporting delay adjustment algorithm. Standard multinomial large sample theory provides an estimate of the asymptotic covariance matrix

$$diag\left(\frac{\widehat{h^r}_X(x)\{1 - \widehat{h^r}_X(x)\}}{N_{+x}} \right).$$

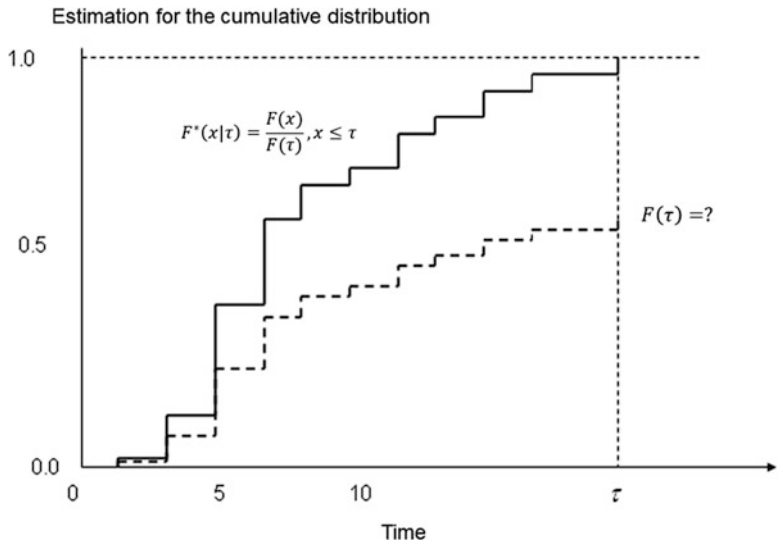


Fig. 7.15 Non-parametrically, right-truncated data are unable to identify the entire cumulative distribution. It is only possible to estimate the conditional distribution conditioning on $X \le \tau$

These yield the estimation for the conditional probability:

$$\widehat{F}^*(x|\tau) = \widehat{\frac{F_X(x)}{F_X(\tau)}} = \prod_{l=x+1}^{\tau}\{1 - \widehat{h^r}_X(l)\} = \prod_{l=x+1}^{\tau}\left(1 - \frac{n_{+l}}{N_{+l}}\right), \quad 1 \le x \le \tau = \max(\tau_i). \tag{7.29}$$

The asymptotic variance estimate (Lawless 1994, 2003) is

$$\widehat{var}\left\{\widehat{F}^*(x|\tau)\right\} = \left\{\widehat{F}^*(x|\tau)\right\}^2 \sum_{x=1}^{\tau} \frac{\widehat{h^r}_X(x)}{N_{+x}\{1 - \widehat{h^r}_X(x)\}}. \tag{7.30}$$

For nonparametric estimation, the identifiable part of $F_X(x)$ is $\{h_X^r(x) : 1 \le x \le \tau\}$. Data do not have information for $\{h_X^r(x) : x = \tau + 1, \cdots, \infty\}$. Therefore, nonparametrically, one cannot fully identify the distribution $F_X(x)$. The best one can achieve is to estimate $F^*(x|\tau)$, $x \le \tau$. Analogous to Fig. 7.9, the non-parametric estimation for $\widehat{F}^*(x|\tau)$ is plotted in Fig. 7.15.

Will a Fully Parametric Model $f(x;\underline{\theta})$ Be Able to Identify the Distribution from Right-Truncated Data? To study this question, we consider a family of the scale-shape distributions where the c.d.f. is defined by $F(x;\lambda,\varsigma) = F_0((\lambda x)^{\varsigma})$, where $F_0(x)$ is a standard distribution not involving unknown parameters with $F_0(0) = 0$ and $F_0(\infty) = 1$, subject to the condition $\lim_{\theta\to 0}\frac{F_0(\theta x^{\varsigma})}{F_0(\theta)} = x^{\varsigma}$ (where $\theta = \lambda^{\varsigma}$). Given the shape parameter ς, if the scale parameter λ is very small, it

approaches a simple power function. In other words, for long underlying durations X, the beginning part of the c.d.f. when $x \in (0, 1]$ behaves like the power function x^{ς}. This family includes the Weibull and the log-logistic distributions.

Now consider a pair of observations (x, τ) subject to $x \leq \tau$. The condition $\lim_{\theta \to 0} \frac{F_0(\theta x^{\varsigma})}{F_0(\theta)} = x^{\varsigma}$ becomes

$$\frac{F(x; \lambda, \varsigma)}{F(\tau; \lambda, \varsigma)} = \frac{F_0\left(\lambda^{\varsigma} x^{\varsigma}\right)}{F_0\left((\lambda\tau)^{\varsigma}\right)} = \frac{F_0\left(\theta\,(x/\tau)^{\varsigma}\right)}{F_0\left(\theta\right)} \to (x/\tau)^{\varsigma}\,, \text{ as } \theta = (\lambda\tau)^{\varsigma} \to 0.$$

Weibull distribution: $F_0(x) = 1 - e^{-x}$.

$$\frac{F_0\left((\lambda x)^{\varsigma}\right)}{F_0\left((\lambda\tau)^{\varsigma}\right)} = \frac{1 - e^{-\theta(x/\tau)^{\varsigma}}}{1 - e^{-\theta}} = \left(\frac{x}{\tau}\right)^{\varsigma} \times$$
$$\left[1 + \frac{\theta}{2}\left(1 - \left(\frac{x}{\tau}\right)^{\varsigma}\right) + \frac{\theta^2}{12}\left(1 - \left(\frac{x}{\tau}\right)^{\varsigma}\right)\left(1 - 2\left(\frac{x}{\tau}\right)^{\varsigma}\right) + O\left(\theta^3\right)\right]$$

Log-logistic distribution: $F_0(x) = \frac{x}{1+x}$

$$\frac{F_0\left((\lambda x)^{\varsigma}\right)}{F_0\left((\lambda\tau)^{\varsigma}\right)} = \left(\frac{x}{\tau}\right)^{\varsigma} \frac{1+\theta}{1+\theta\,(x/\tau)^{\varsigma}}$$
$$= \left(\frac{x}{\tau}\right)^{\varsigma}\left[1 + \theta\left(1 - \left(\frac{x}{\tau}\right)^{\varsigma}\right) + \theta^2\left(\frac{x}{\tau}\right)^{\varsigma}\left(1 - \left(\frac{x}{\tau}\right)^{\varsigma}\right) + O\left(\theta^3\right)\right]$$

In these cases, if $\varsigma > 1$, $\theta = (\lambda\tau)^{\varsigma} \to 0$ implies that for any $\varepsilon > 0$, $\lambda\tau$ must be sufficiently small such that $\lambda\tau < \varepsilon^{1/\varsigma}$. Translating to plain language, it implies that if the truncation time τ is short whereas the underlying distribution of X is long (i.e., small value of the scale parameter λ), the conditional distribution $\frac{F_0((\lambda x)^{\varsigma})}{F_0((\lambda\tau)^{\varsigma})} \approx \left(\frac{x}{\tau}\right)^{\varsigma}$ does not contain λ.

Now we consider a sample of right-truncated data that is represented by (x_i, τ_i), $i = 1, \ldots, n$ subject to the condition $x_i \leq \tau_i$ and let $\tau = \max(\tau_i)$. The above discussion intuitively leads to

1. If the maximum observation window $\tau = \max(\tau_i)$ is relatively short and the underlying distribution for X is long, such that $\lambda\tau \ll 1$, for the above distributions with $\varsigma > 1$, data do not contain enough information about the scale parameter λ.
2. The second and higher order terms of the series expansion of the Weibull and the log-logistic distributions contain the factor $1 - \left(\frac{x}{\tau}\right)^{\varsigma}$. This implies that only a subset of the data such that x_i are neither too close to zero nor too close to $\max(\tau_i)$, but close to $2^{-1/\varsigma} \max(\tau_i)$, may contain some information for λ. This also requires that $\max(\tau_i)$ to be sufficiently large, as well as the specific shape of the underlying distribution.

We demonstrate this through two examples, both related to the estimation of the incubation period based on right truncated data.

Example 31 (Example 29 Continued) Using the subset of 258 adult transfusion associated AIDS by C =June 30, 1986 (Lagakos et al. 1988, Table 1) with $\tau = \max(\tau_i) = 8$ years for the incubation period from HIV infection to the onset of AIDS, data are fitted to a Weibull distribution based on the conditional likelihood (7.25). The estimated shape parameter is $\widehat{\varsigma} \approx 2.1$ with the 95% confidence limits $1.85 \leq \varsigma \leq 2.38$. It implies that the incubation distribution during the first 8 years since infection increases approximately linearly. With respect to λ, data could only provide a one-sided 95% confidence limit $\lambda \leq 0.128$. Together with estimated $\widehat{\varsigma} \approx 2.1$, this gives estimated median incubation ≥ 6.6 years. The contour of the likelihood surface is very flat, which has been shown in earlier discussions (see Fig. 7.3). The upper limit $\lambda^{up} = 0.128$ gives $\lambda^{up}\tau \leq 0.128 * 8 \approx 1$, and hence $\lambda\tau < 1$.

Example 32 This example shows a case $\lambda\tau$ is rather large and both λ and ς are precisely estimated. Dr. Ian Johnson at University of Toronto (personal communication) kindly provided 42 probable SARS cases on April 11, 2003 to assess the incubation distribution. They had been retrospectively ascertained to single exposure dates, ranging from March 6 to March 29, 2003. The longest observable window was $\tau = \max(\tau_i) = 36$ days, from the earliest exposure date March 6 to the time when data are compiled April 11. The maximum observed incubation period in the data was $\max_i\{x_i\} = 10$ days. A log-logistic distribution was used in the conditional likelihood (7.25) which gives the maximum likelihood estimate $\widehat{\lambda} = 0.2395$ and $\widehat{\varsigma} = 3.413$. In this case, $\widehat{\lambda\tau} = 8.622$. We reparameterize the log-logistic distribution in terms of the median λ^{-1} and the 95th quantile. The median was estimated as $\widehat{\lambda}^{-1} = 4.175$ (days) with 95% confidence limits 3.453–5.139 days. The 95th quantile is defined as t_{95} such that $\Pr\{X \leq t_{95}\} = 0.95$. It was estimated as $\widehat{t_{95}} = 9.9$ (days) with 95% confidence limits 6.569–17.211 days. For the goodness-of-fit of the log-logistic model, we compare the cumulative distributions over the parameters ranges of the log-logistic distribution (smooth lines) as well as the non-parametric estimate based on (7.29) as what data suggest. Figure 7.16 illustrates these estimates.

7.4 Some More Discussions About Back-Calculation

Back-calculation has been briefly mentioned twice in this chapter. Once was for the demonstration of the convolution of the systematic component and the link component in a generalized nonlinear model. Another time was for the discussion of the non-identifiability problem.

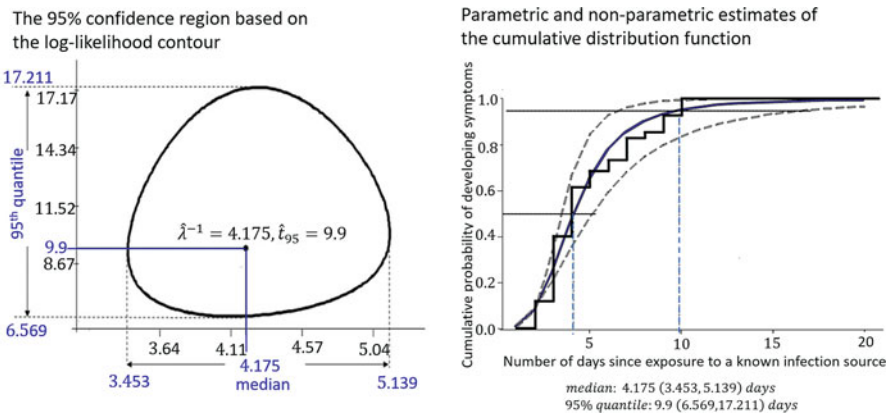


Fig. 7.16 Illustration of fitting the log-logistic distribution to right-truncated incubation times based on a small sample during the 2003 SARS outbreak and comparison with the non-parametric estimate

The convolution, given by

$$\mu(t;\underline{\theta},\underline{\psi}) = \begin{cases} \int_0^t i(u;\underline{\theta}) f(t-u|u;\underline{\psi})du, & \text{continuous time model} \\ \sum_{u=0}^{t} i(u;\underline{\theta}) f(t-u|u;\underline{\psi}), & \text{discrete time model} \end{cases}.$$

is to link the data based on the occurrence of the subsequent observable events from a counting process with mean function $\mu(t;\underline{\theta},\underline{\psi})$ in order to estimate the parameters in the systematic part of the model $i(u;\underline{\theta})$, which is the incidence intensity of the non-observable initial events. The observable events could be onset of clinical symptoms, and the initiating events could be the infection of an agent which then leads to clinical symptoms modelled according to an incubation distribution.

For discussion purposes, let us suppose an ideal situation where for every individual in the data, the time of the initiating event can be back-dated. In this case, the upper triangle matrix as shown in Table 7.1 can be established, in which

$$n_{tx} = \#\{\text{initiating event at time } t, \text{subsequent event at time } t+x\}.$$

where $t = 0, 1, 2, \cdots, C$ denote the time of the initiating event and $x = 0, 1, \cdots, C-t$. In this way, the back-calculation problem is essentially the same as the right-truncation problem as applied in the reporting delay analysis.

The expected values are

$$E[n_{tx}] = \mu(t,x) = i(t) f(x).$$

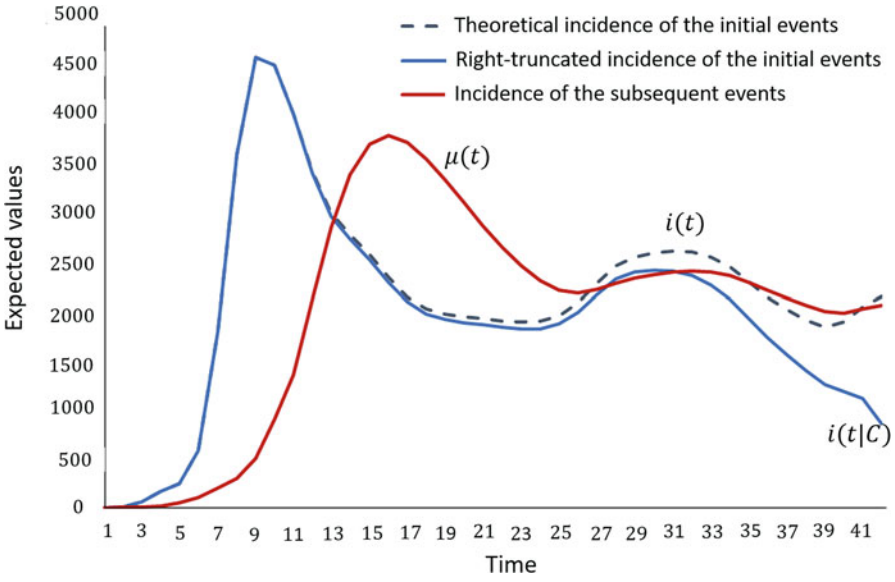


Fig. 7.17 Schematic illustration of the expected counts of the marginal totals $i_1(t|C)$, $i_2(t)$ and the theoretical incidence of the initial events $i_1(t)$ at $C = 41$.

The expected values of the row totals are $E[N(t|C)] \triangleq i(t|C)$, which might be called the right-truncated incidence of the initial events,

$$i(t|C) = i(t) \sum_{x=0}^{C-t} f(x) = i(t)F(C-t), \ t = 0, 1, \cdots, C. \tag{7.31}$$

The expected values of the column totals are

$$\mu(t) = \sum_{s=0}^{t} E[n_{s,t-s}] = \sum_{s=0}^{t} i(s) f(t-s). \tag{7.32}$$

We notice that (7.31) is the model in reporting delay analysis, provided that $F(C-t)$ can be estimated; (7.32) is the model in back-calculation provided that $f(t-s)$ is fully specified. Figure 7.17 conceptually illustrates (7.31), (7.32) and $i(t)$ on the same graph.

Even though one cannot completely identify $i(t)$ from $f(x)$, retrospectively ascertained data still contain some information that one may assess either the trend of the initiating event or the duration distribution with caution.

Since $E[n_{tx}] = \mu(t,x) = [i(t)\,F(C-t)] \times \left[\frac{f(x)}{F(C-t)}\right]$, data (t_i, x_i) are sufficient for the first factor $i(t)\,F(C-t)$. It can be shown that the minimal sufficient statistics for $i(t)F(C-t)$ are the row totals $N(t|C) = \sum_{x=0}^{C-t} n_{tx}$. Conditioning on the row totals, the likelihood function becomes

$$L \propto \prod_{i=1}^{n} \frac{f(x_i)}{F(C-t_i)} = \prod_{i=1}^{n} \frac{f(x_i)}{F(\tau_i)}$$

Therefore the likelihood function (7.26) for analyzing right-truncated data is the conditional likelihood by treating $i(t)$ as the nuisance parameter. As discussed previously, data may not be able to fully identify the distribution $F(x)$ but only up to $\tau = \max(\tau_i)$ as a conditional distribution.

On the other hand, if $F(x)$ is fully specified, then $F(C-t)$ is precisely known. $i(t)$ is estimated through the marginal totals $N(t|C)$ as

$$\widehat{i}(t) = \frac{N(t|C)}{F(C-t)} \tag{7.33}$$

in the same logic as the reporting delay adjustment.

In general, back-calculation methods are developed assuming that each individual in the data only has information on the onset of the subsequent event. Therefore $\{n_{tx}\}$ are not observable and $N(t|C)$ are not observable. The only observable data are the column totals with mean value (7.32).

There is plenty of literature on different back-calculation methods and algorithms, based on continuous time or discrete time models, applied to the studies of HIV/AIDS, viral hepatitis, and many other infectious diseases. We do not intend to write this section about these methods and algorithms, except for the following brief mentioning.

Since the distribution $f(x)$ is fully specified, had the incidence function $i(t)$ been fully specified with all the parameters known, it would have been possible to compute the expected values $E[n_{tx}]$ in each cell of the upper triangle matrix in Table 7.1 based on the observed column totals using a multinomial distribution (Becker et al. 1991). On the other hand, had $\{n_{tx}\}$ been observed, then the back-calculation would have been reduced to a simple algorithm based on (7.33). This is the core of the Expectation-Maximization-Smoothing (EMS) algorithm (Becker et al. 1991) widely used in many back-calculation applications. A generalization of the EMS algorithm based on more than one source of data is given by Yan et al. (2011).

7.5 Problems and Supplements

7.1 Incidence data of the confirmed and probable cases of the Ebola outbreak in the Democratic Republic of Congo (DRC, August 2018–January 2019) are publicly available in the World Health Organization website. Data are manually extracted using WebPlotDigitizer (Rohatgi 2018). In this exercise, data are aggregated as weekly counts and we consider a subset of data with dates of symptom onset starting from August 20, 2018 onwards. We define Week Zero as the week August 20–26, 2018. By the end of Week 9 (i.e., October 28, 2018), there were a total 144 reported cases starting with week of onset at Week Zero. They are cross-tabulated by week of onset and week of report.

Week of onset	Week of report									
	0	1	2	3	4	5	6	7	8	9
0	1	3	2	0	0	0	2	0	0	0
1		3	9	0	0	0	0	0	0	0
2			0	6	0	0	0	0	0	0
3				2	4	1	0	1	0	0
4					0	9	2	0	0	0
5						3	12	2	3	0
6							7	6	13	0
7								0	12	4
8									7	26
9										4

(a) Plot the row totals and the column totals on the same graph and comment on the meaning of these marginal totals.

(b) Reporting delays $x = 0, 1, \ldots$ are calculated by weeks. Diseases that are reported during the same week of the symptoms onset are assigned with $x = 0$. Calculate the frequency of cases with $x = 0, 1, 2, \ldots, 9$ (defined as the numbers of cases with $x = 0, 1, 2, \ldots, 9$ divided by 144) and calculate the cumulative frequency by Week 3. Do you think the reporting delay is that short?

(c) Moving forward, by the end of Week 14 (ending on December 2, 2018) there were a total of 330 reported cases starting with week of onset at Week Zero. The cross-tabulation table is updated. Plot the row totals of the updated table and the row totals of the table ending on Week 9 on the same graph.

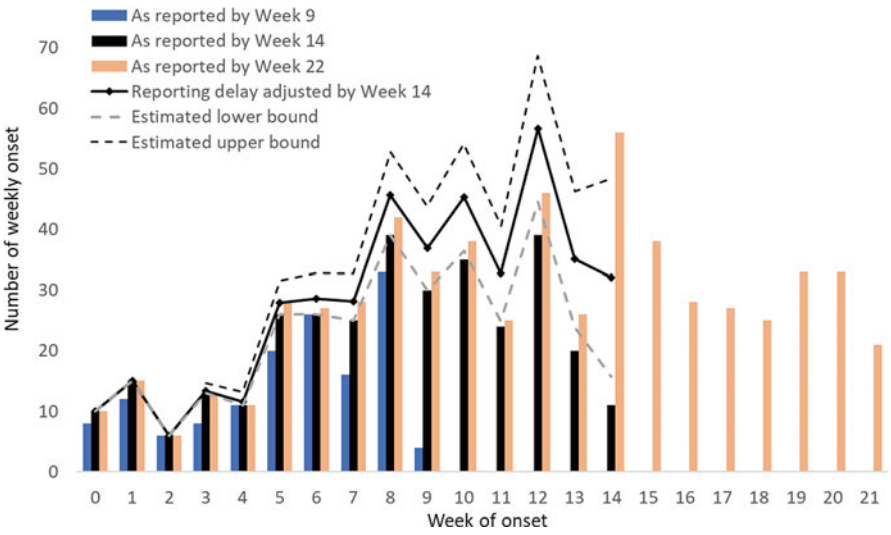
(d) Calculate the frequency cases with $x = 0, 1, 2, \ldots, 14$ (defined as the numbers of cases with $x = 0, 1, 2, \ldots, 14$ divided by 330) and calculate the cumulative frequency by Week 7. Do you think it is because the reporting delay is getting longer or something else?

Wk.of onset	Week of report														
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	1	3	2	0	0	0	2	0	0	0	0	1	0	1	0
1		3	9	0	0	0	0	0	0	0	0	1	2	0	0
2			0	6	0	0	0	0	0	0	0	0	0	0	0
3				2	4	1	0	1	0	0	0	5	0	0	0
4					0	9	2	0	0	0	0	0	0	0	0
5						3	12	2	3	0	0	6	0	0	0
6							7	6	13	0	0	0	0	0	0
7								0	12	4	2	7	0	0	0
8									7	26	1	2	1	2	0
9										4	12	9	3	2	0
10											7	21	3	4	0
11												9	8	7	0
12													22	17	0
13														12	8
14															11

(e) The following table converts the table above to represent n_{tx} as defined in Table 7.1. The column totals represent $n_{+x} = \sum_{t=0}^{C-x} n_{tx} = \sum_i I(x_i = x)$. Use the method in Brookmeyer and Gail (1994) as illustrated in Fig. 7.14 to estimate the weekly number by symptom onset up to the end of Week 9, including cases that were not yet reported.

Week of onset	Reporting delay x														
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	1	3	2	0	0	0	2	0	0	0	0	1	0	1	0
1	3	9	0	0	0	0	0	0	0	1	2	0	0	0	
2	0	6	0	0	0	0	0	0	0	0	0	0	0		
3	2	4	1	0	1	0	0	0	0	0	0	0			
4	0	9	2	0	0	0	0	0	0	0	0				
5	3	12	2	3	0	0	6	0	0	0					
6	7	6	13	0	0	0	0	0	0						
7	0	12	4	2	7	0	0	0							
8	7	26	1	2	1	2	0								
9	4	12	9	3	2	0									
10	7	21	3	4	0										
11	9	8	7	0											
12	22	17	0												
13	12	8													
14	11														
n_{+x}	88	153	44	14	11	2	8	0	5	1	2	1	0	1	0
$N_{+x} = \sum_{t=0}^{C-x} \sum_{j=0}^{x} n_{tj}$															
N_{+x}	88	230	254	229	216	183	161	122	102	77	53	43	30	25	10

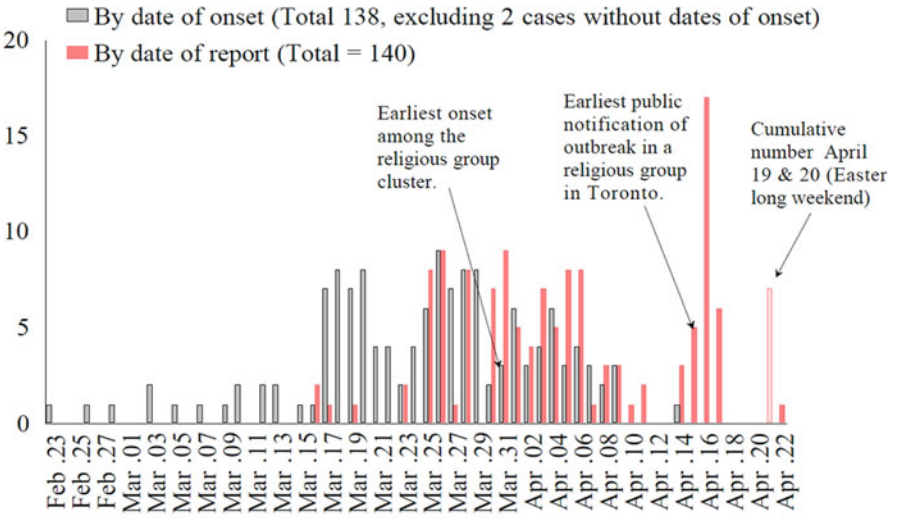
(f) The following figure displays Ebola cases in DRC by week of onset. The bars represent data as reported by Week 9, Week 14, and Week 22. Reporting delays were adjusted using data by the end of Week 14, with point estimates as well as lower and upper 95% confidence limits represented by lines, using the method in Lawless (1994). Consult the original paper and examine the assumptions in the algorithm. Comment on the performance of this method as applied to the Ebola data and discuss potential violations of the assumptions.



7.2 The following figure compares the trends of the 2003 SARS outbreak in Canada based on probable cases reported to Health Canada on April 22, 2003. The dark bars represent the numbers by date of symptom onset whereas the pink bars represent the numbers by date of report. There are several important dates to remember:

- March 13: WHO started worldwide surveillance on atypical pneumonia (later renamed as SARS);
- March 25: SARS became reportable in Canada and surveillance was intensified;
- April 15: health officials made a news release about a new cluster of SARS cases in Toronto related to a religious group;
- April 19–20: Easter long weekend.

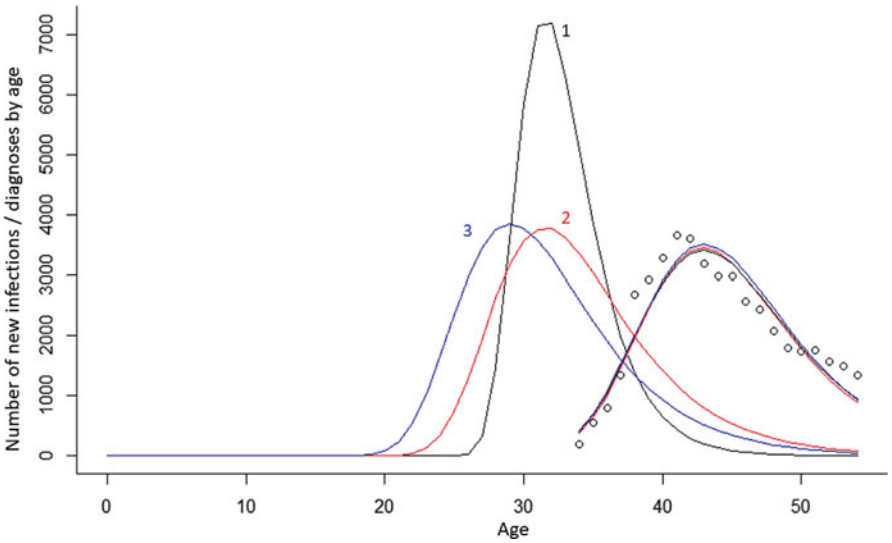
All these events affected the trend based on date of report. Comment on the differences of the trend based on date of onset and date of report. Do you think the trend by date of onset after April 12 was declining? Do you think that the epidemic peaked around April 16? Do you think the simple reporting delay method in Problem 7.5 is suitable?



By date of onset, http://www.hc-sc.gc.ca/pphb-dgspsp/sars-sras/eu-ae/sars20030422_e.html
By date of report, http://www.hc-sc.gc.ca/english/protection/warnings/sars/sars_updates.html

7.3 A case reporting surveillance system is able to document the year of birth as well as the number of new diagnoses of the disease by year with respect to a chronic viral infectious disease. Because the disease natural history is very long (years or decades), the trend of new diagnoses of the disease does not reflect the trend of new infections. There is not much information with respect to the distribution from the time at infection to the time at diagnosis. Empirical evidence has suggested that a log-logistic distribution given by (2.24) is suitable to capture the general shape of such a distribution, with median λ^{-1} and shape parameter ς.

(a) The following figure summarizes some results for a specific birth-cohort. For simplicity, we take year zero to correspond the year of birth and the x-axes in the figure is labelled as age. Surveillance data are shown as circles, starting from age 34 years with peak age in the early 40s. However, auxiliary epidemiologic evidence has shown that the peak of new infections is likely to be in the range between 25 and 34 years of age. The figure shows differently assigned values for λ^{-1} and ς yield differently estimated number of new infections by age/year, but they all provide equally good fit to data. Comment on: in spite of the very different incidence curves (i.e. estimated number of new infections by age/year), is there any feature that is relatively robust with respect to the values of λ^{-1} and ς?



(b) Which of the three incidence curves (labelled as 1, 2 and 3) correspond to which of the following assumptions?

1. a log-logistic distribution with median $\lambda^{-1} = 12$ years and shape parameter $\varsigma = 3.8$;
2. a log-logistic distribution with median $\lambda^{-1} = 12$ years and shape parameter $\varsigma = 10.0$;
3. a log-logistic distribution with median $\lambda^{-1} = 14$ years and shape parameter $\varsigma = 12.0$.

Chapter 8
Characterizing Outbreak Trajectories and the Effective Reproduction Number

8.1 Introduction

Emerging and re-emerging infectious diseases pose major challenges to public health worldwide (Fauci and Morens 2016). Fortunately mathematical and statistical inference and simulation approaches are part of the toolkit for guiding prevention and response plans. As the recent 2013–2016 Ebola epidemic exemplified, an unfolding infectious disease outbreak often forces public health officials to put in place control policies in the context of limited data about the outbreak and in a changing environment where multiple factors positively or negatively impact local disease transmission (Chowell et al. 2017). Hence, the development of public health policies could benefit from mathematically rigorous and computationally efficient approaches that comprehensively assimilate data and model uncertainty in real time in order to (1) estimate transmission rates, (2) assess the impact of control interventions (vaccination campaigns, behavior changes), (3) test hypotheses relating to transmission mechanisms, (4) evaluate how behavior changes affect transmission dynamics, (5) optimize the impact of control strategies, and (6) generate forecasts to guide interventions in the short and long terms.

Mathematical models are quantitative frameworks with which scientists can assess hypotheses on the potential underlying mechanisms that explain patterns in the observed data at different spatial and temporal scales (Chowell 2017). Model

Electronic Supplementary Material The online version of this chapter (https://doi.org/10.1007/978-3-030-21923-9_8) contains supplementary material, which is available to authorized users.

P. Yan, G. Chowell, *Quantitative Methods for Investigating Infectious Disease Outbreaks*, Texts in Applied Mathematics 70,
https://doi.org/10.1007/978-3-030-21923-9_8

complexity can be characterized in terms of the number of variables and parameters that characterize the dynamic states of the system, spatial-temporal resolution (e.g., discrete vs. continuous time), and design (e.g., deterministic or stochastic). While agent-based models, formulated in terms of characteristics and interactions among individual agents, have become increasingly used to model detailed processes often occurring at multiple scales (e.g., within host vs. population level), mean-field models based on systems of ordinary differential equations are widely used in the biological and social sciences. These dynamic models comprise systems of equations and their parameters that together quantify the temporal and spatial states of the system via a set of interrelated dynamic quantities (e.g., susceptibility levels, disease prevalence) (Banks et al. 2009, 2014).

In Sect. 7.1, phenomenological population models refer to stochastic and deterministic models based on conceptual assumptions regarding the population. We used the term phenomenological to distinguish models with respect to assumptions at the level of individuals along the progression of the disease's natural history. We stated that phenomenological population models carry tacit assumptions at the level of individuals.

Deterministic models composed by a system of ordinary differential equations follow this general form:

$$\begin{aligned} x_1^{'}(t) &= f_1(x_1, \cdots, x_h; \Theta) \\ x_2^{'}(t) &= f_2(x_1, \cdots, x_h; \Theta) \\ &\vdots \\ x_h^{'}(t) &= f_h(x_1, \cdots, x_h; \Theta) \end{aligned}$$

where $x_i^{'}(t)$ denotes the rate of change of the system states x_i, $i = 1, \ldots, h$ and $\Theta = (\theta_1, \ldots, \theta_m)$ is the set of model parameters.

In general, the complexity of a model is a function of the parameters that are needed to characterize the states of the system and the spectrum of the dynamics that can be recovered from the model (e.g., number of equilibrium points, oscillations, bifurcations, chaos). A trade-off exists between the level of model complexity and the ability to reliably constrain a model to a specific situation.

8.2 Approximations with Simple Functions

Time-series, loosely called "epi-curves," are widely used in epidemiologic investigation for different purposes. Some make empirical comparisons for spatial and temporal patterns based on data from official surveillance reports. For example, Schanzer et al. (2010) compared epidemic curves on weekly confirmed seasonal influenza-A cases in Canada for multiple influenza seasons as well as with similar

curves in the United States and Europe. Some associate the comparison of spatial patterns with important scientific questions in mind, such as making inferences on transmissibility R_0 and disease impact such as mortality (e.g., Chowell et al. 2007). Others use epidemic curves as information for action. During an infectious disease outbreak, the question of concern is more likely about the current status of the trend, whether it is increasing or decreasing. Sometimes various ad hoc curve fitting techniques are employed to smooth fluctuating data points in order to make short-term projections, forecast health care needs, guide public health decision making, and so on.

In this section, we choose phenomenological models with explicit simple forms such as the sub-exponential function (4.58), the logistic growth (4.59), and various generalized logistic growth functions. As previously discussed in Chaps. 4, 5 and 7, parameters in these models are descriptive by capturing the essence of a time-series data based on a disease outbreak. Although they may not carry any scientific hypotheses regarding the transmission dynamics, they provide an approach to investigate empirical patterns in observed data (Chowell et al. 2016).

In addition, we choose these simple models because

1. most of the transmission dynamic models defined by systems of differential equations do not have explicit solutions;
2. most of the numeric solutions of these equations can be closely approximated by one of the generalized logistic functions;
3. for those models that do have explicit solutions, they are either logistic or generalized logistic functions;
4. time-series data usually do not have sufficient information to identify the "mechanical assumptions" explicitly modeling the transmission dynamics.

8.2.1 The Sub-exponential Growth Function and the Generalized Growth Model (GGM)

This phenomenological model is useful to forecast epidemic growth patterns (Viboud et al. 2016; Chowell and Viboud 2016; Shanafelt et al. 2017; Pell et al. 2018a). In particular, previous analyses highlighted the presence of early sub-exponential growth patterns in infectious disease data across a diversity of disease outbreaks (Viboud et al. 2016).

The sub-exponential growth functions have been previously discussed in Sect. 4.5.2, in which we restricted our definition (4.56) as convex functions $C(t)$ bounded by the linear growth from below and the exponential growth from above, that is, $i_0\,(1 + rt) \le C(t) \le i_0 e^{rt}$. We have pointed out that the classic exponential growth function is associated with a set of strong mathematical assumptions and conditions when the system is at (disease-free) equilibrium, whereas during the initial stage of an outbreak in observed data, sub-exponential growth patterns are

more common. We have previously highlighted several mechanisms that potentially result in such growth patterns.

In particular, we consider the model $C(t) = i_0(1 + rvt)^{1/v}, 0 < v \leq 1$ for the cumulative incidence. If $v \to 0$, this model leads to the well-known exponential growth model, which applies both to the cumulative incidence $C(t)$ and to the instantaneous incidence $C'(t)$, while $v = 1$, corresponds to the linear growth of $C(t)$ and constant incidence per unit of time. $C(t) = i_0(1+rvt)^{1/v}, 0 < v \leq 1$, are illustrated in Fig. 4.11 in Chap. 4, as convex growth functions bounded by the linear growth and the exponential growth.

The generalized-growth model (4.61) in Viboud et al. (2016), which is also called the power law exponential model by Banks (1994), is defined by the differential equation

$$C'(t) = rC(t)^p, \ 0 \leq p \leq 1 \tag{8.1}$$

which allows relaxing the assumption of exponential growth via a "deceleration of growth" or "scaling of growth" parameter, p. $C'(t)$ describes the incidence growth phase over time t; the solution $C(t)$ describes the cumulative number of cases at time t.

When $i_0 = 1$ and letting $p = 1 - v$, the sub-exponential function $C(t) = (1 + rvt)^{1/v}$ is the solution of (4.61).

In semi-logarithmic scale, exponential growth patterns are visually evident when a straight line fits well several consecutive disease generations of epidemic growth, whereas a downward curvature in semi-logarithmic scale indicates early sub-exponential growth dynamics.

8.2.2 The Simple Logistic Function

In Chap. 5, we introduced many types of phenomenological models involving the dynamics of the process of interest (e.g., population or transmission dynamics). These types of models are often formulated in terms of a dynamic system describing the spatial-temporal evolution of a set of variables, and they are useful to evaluate the emergent behavior of the system across the relevant space of parameters (Chowell et al. 2016). In particular, compartmental models are based on systems of ordinary differential equations that focus on the dynamic progression of a population through different epidemiological states (Bailey 1975; Anderson and May 1991; Brauer 2006; Lee et al. 2016). While these models may not be useful for testing scientific hypotheses and formulating theory on disease transmission, they are very useful in practice such as for curve fitting, prediction as well as formulating of some statistical models, such as the back-calculation. One of the most memorable quotes from the wordsmith and former New York Yankees catcher, Yogi Berra, is:

> In theory, theory and practice are the same thing; in practice, they are different.

Several models, such as the SI model, the SIS model, and the model defined by (8.5), produce the logistic epidemiologic curves. Many other models also produce logistic-like epidemiologic curves that can be used to explain patterns in the observed data.

The logistic growth function (4.59) is one of the oldest growth functions with the following equivalent forms

$$C_{\text{logis}}(t) = \frac{i_0 K}{i_0 + (K - i_0)\, e^{-\rho t}} = \frac{K}{1 + \frac{1}{v} e^{-\rho t}} = \frac{K}{1 + e^{-\rho(t-\alpha)}} \tag{8.2}$$

where $v = \frac{i_0}{K - i_0}$ and $\alpha = \frac{1}{\rho} \log \frac{K - i_0}{i_0} = -\frac{1}{\rho} \log v$. In all these representations, there are three functionally independent parameters.

The logistic function was first proposed by Verhulst (1838). For modeling population growth, the logistic model was used and popularized by Pearl (1925), Pearl and Reed (1920), and Yule (1925). The expression

$$C_{\text{logis}}(t) = \frac{K}{1 + e^{-\rho(t-\alpha)}}, \quad -\infty < t, \alpha < \infty, \ \rho, K > 0. \tag{8.3}$$

characterizes the time-series data. The parameters (ρ, α, K) are descriptive about the general shape and are useful to fit to time-series data, of which ρ is the scale parameter associated with the initial growth; α is a location parameter that is also the inflexion point at which the increase of $C_{\text{logis}}(t)$ turns from convex to concave; $K = \lim_{t \to \infty} C(t)$ is the upper limit, referred to as the carrying capacity.

Many infectious disease models lead to the exact logistic growth form or growth functions very closely resembling logistic growth.

The deterministic SIS model produces the logistic function (5.14) for the number of infectious individuals at time t, as

$$I_d(t) = \frac{m i_0 (\beta - \gamma)}{\beta i_0 + (m(\beta - \gamma) - \beta i_0)\, e^{-(\beta - \gamma)t}}. \tag{8.4}$$

If $\beta - \gamma > 0$, $I_d(t)$ increases monotonically and approaches the value $m(1 - \gamma/\beta)$. It is the same as logistic function (4.59) via re-parametrization $K = m(1 - \gamma/\beta)$ and $\rho = \beta - \gamma$. Although in (8.4), the parameters (m, i_0, β, γ) are associated with hypotheses about the transmission dynamic, from the perspective of fitting the model to data, $I_d(t)$ only has three independent parameters. In fact, the time-series data that fit well with the logistic function do not have the information to test the hypothesis H_0: $\gamma = 0$ in the SIS model.

The logistic function can also arise from other deterministic transmission models. Tan (2000) considered the following compartment model

$$\begin{cases} \frac{d}{dt} S_d(t) = -\beta \frac{S_d(t) I_d(t)}{S_d(t) + I_d(t)} \\ \frac{d}{dt} I_d(t) = \beta \frac{S_d(t) I_d(t)}{S_d(t) + I_d(t)} - \gamma I_d(t) \\ \frac{d}{dt} Z_d(t) = \gamma I_d(t) - \delta Z_d(t) \end{cases}, \tag{8.5}$$

where $S_d(t)$ and $I_d(t)$, as in the SIS and SIR models, represent the numbers of susceptible and infected individuals in the population. The main difference from the deterministic models discussed before is that, in this model, all infectious individuals will progress to Compartment Z. Once individuals enter Compartment Z, they no longer make contacts with susceptible individuals. For instance, Compartment Z may represent advanced illness or being isolated. Thus the instantaneous infection function is modified as $\beta \frac{S_d(t) I_d(t)}{S_d(t)+I_d(t)}$.

Let $\psi(t) = \frac{I_d(t)}{S_d(t)+I_d(t)}$ be the proportion of infected individuals before entering Compartment Z, we have

$$\frac{d}{dt}\left[S_d(t) + I_d(t)\right] = -\gamma I_d(t) = -\gamma \psi(t)\left[S_d(t) + I_d(t)\right],$$
$$\frac{1}{S_d(t) + I_d(t)} \frac{d}{dt} I_d(t) = \psi(t)\left\{\left[1 - \psi(t)\right]\beta - \gamma\right\}.$$

It follows that

$$\begin{aligned}\frac{d}{dt}\psi(t) &= \frac{\left[S_d(t) + I_d(t)\right]\frac{d}{dt} I_d(t) - I_d(t)\frac{d}{dt}\left[S_d(t) + I_d(t)\right]}{\left[S_d(t) + I_d(t)\right]^2} \\ &= \frac{1}{S_d(t) + I_d(t)}\left\{\frac{d}{dt} I_d(t) - \psi(t)\frac{d}{dt}\left[S_d(t) + I_d(t)\right]\right\} \\ &= \rho\psi(t)\left[1 - \psi(t)\right],\end{aligned}$$

where $\rho = \beta - \gamma$. Clearly, $\psi(t)$ follows the logistic growth given by

$$\psi(t) = \frac{\psi(0)}{\psi(0) + \left[1 - \psi(0)\right] e^{-\rho t}}$$

and $\psi(t) \to 1$ as $t \to \infty$.

8.2.3 *Generalized Logistic Functions*

The logistic differential equation

$$\frac{d}{dt} C(t) = \rho K \left(\frac{C(t)}{K}\right)\left(1 - \frac{C(t)}{K}\right)$$

assumes that the per capita growth rate decreases linearly with population size or density. Its solution is the logistic function $C_{\text{logis}}(t) = \frac{i_0 K}{i_0 + (K - i_0) e^{-\rho t}}$ which can be also expressed as

$$C_{\text{logis}}(t;\rho,\alpha,K) = \frac{K}{1+e^{-\rho(t-\alpha)}}, \tag{8.6}$$

$$= K\left(1-\frac{1}{1+e^{\rho(t-\alpha)}}\right), \tag{8.7}$$

where $-\infty < t < \infty$ with three parameters (ρ, K, α): $\rho > 0$ is the scale parameter, $-\infty < \alpha < \infty$ is a location parameter and $K = \lim_{t\to\infty} C(t) > 0$ is the carrying capacity. It is symmetric in the sense that α is also the inflexion point at which $C_{\text{logis}}(\alpha) = K/2$. Given K and the initial value $i_0 = C_{\text{logis}}(0)$, $\alpha = \frac{1}{\rho}\log\frac{K-i_0}{i_0}$.

The first derivative $\frac{d}{dt}C_{\text{logis}}(t)$ is

$$I_{\text{logis}}(t) = \frac{d}{dt}C_{\text{logis}}(t) = \frac{ke^{-\rho(t-\alpha)}}{\left(1+e^{-\rho(t-\alpha)}\right)^2},$$

where $k = \rho K$. It reaches the maximum value at $t = \alpha$ such that $I_{\text{logis}}(\alpha) = \frac{k}{4}$. Meanwhile, $\lim_{t\to\infty} I_{\text{logis}}(t) = 0$.

The logistic function may be generalized in two directions: (1) asymmetric function for $I(t)$ by adding a shape parameter $\theta > 0$; (2) $\lim_{t\to\infty} I(t) = c > 0$, where $I(t) = \frac{d}{dt}C(t)$. A further generalization is to combine (1) and (2) to obtain more flexible forms in order to fit empirical data, especially for diseases with apparent endemic equilibrium. These will be discussed below.

Generalization Towards Asymmetry: The Richards Model and Its Variations

The symmetric shape of the logistic function makes it inflexible to fit data suggesting asymmetry. There are different ways to create asymmetric generalized logistic forms, such as

$$I_{\text{Glogis}}(t) = \frac{1}{1+e^{-\rho(t-\alpha)}}\frac{ke^{-\eta(t-\alpha)}}{1+e^{-\eta(t-\alpha)}}$$

where $\eta > 0$ may be different from the initial growth rate ρ. In this generalization, both parameters η and ρ act as scale parameters of time. It is inconvenient to interpret a model representing a time-series with two different scale parameters. In addition, it does not correspond to the generalization of the logistic differential equation.

The Richards growth curve (Richards 1959) is one of the best known generalized logistic functions. It adds a shape parameter $\theta > 0$ to scale the proportion $C(t)/K$ in the logistic differential equation. The result is the theta-logistic equation

$$\frac{d}{dt}C(t) = rK\left(\frac{C(t)}{K}\right)\left(1-\left[\frac{C(t)}{K}\right]^{\theta}\right),\ \theta > 0. \tag{8.8}$$

Causton and Venus (1981) show that, when $r > 0$ and $\theta > 0$, given $C(0) = i_0$,

$$C_{\text{Richards}}(t) = \frac{K}{\left(1 + Qe^{-r\theta t}\right)^{1/\theta}}, \tag{8.9}$$

where $Q = \left(\frac{K}{i_0}\right)^{\theta} - 1$. If we re-parameterize the scale parameter $\rho = r\theta$ and let $Q = e^{\rho\alpha}$, (8.9) becomes the following generalized logistic growth function

$$C_{\text{Richards}}(t; \rho, \alpha, \theta, K) = \frac{K}{\left[1 + e^{-\rho(t-\alpha)}\right]^{1/\theta}}, \tag{8.10}$$

which is directly adding the shape parameter into (8.6). In (8.10), $\rho, \theta, K > 0$ and $-\infty < \alpha < \infty$.

The first derivative $\frac{d}{dt}C_{\text{Richards}}(t)$ is

$$I_{\text{Richards}}(t) = \frac{ke^{-\rho(t-\alpha)}}{\left[1 + e^{-\rho(t-\alpha)}\right]^{\frac{\theta+1}{\theta}}} \tag{8.11}$$

where $k = \frac{\rho}{\theta}K$.

The inflexion point for $C_{\text{Richards}}(t)$ is

$$t^* = \frac{1}{r\theta}\left(\ln\frac{Q}{\theta}\right) = \alpha - \frac{1}{\rho}\log\theta.$$

At the inflexion point, $C_{\text{Richards}}(t^*) = \frac{K}{(1+\theta)^{1/\theta}}$. When $\theta < 1$, $C_{\text{Richards}}(t^*) < K/2$; when $\theta = 1$, $C_{\text{Richards}}(t^*) = K/2$ and when $\theta > 1$, $C_{\text{Richards}}(t^*) > K/2$. At the inflexion point, $I_{\text{Richards}}(t)$ arrives at the peak value $I_{\text{Richards}}(t^*) = \frac{\rho K}{(1+\theta)^{\frac{\theta+1}{\theta}}}$.

The Richards model has been fitted to a range of epidemic curves that exhibit sigmoid cumulative growth patterns (Turner et al. 1976; Ma et al. 2014; Wang et al. 2012; Hsieh and Cheng 2006; Dinh et al. 2016).

A Variation of the Richards Model Instead of adding the shape parameter θ into (8.6), we add it into (8.7) and we get the generalized logistic growth function

$$C_{\text{Richards2}}(t; K, \rho, \alpha, \theta) = K\left(1 - \frac{1}{\left[1 + e^{\rho(t-\alpha)}\right]^{1/\theta}}\right), \quad K, \rho, \alpha, \theta > 0. \tag{8.12}$$

Both generalized logistic functions are related through

$$C_{\text{Richards2}}(t; K, \rho, \alpha, \theta) = K - C_{\text{Richards}}(-t; K, \rho, -\alpha, \theta).$$

It can be easily shown that (8.12) is the solution of the theta-logistic equation

$$\frac{d}{dt}C(t) = rK\left(1 - \left[1 - \frac{C(t)}{K}\right]^{\theta}\right)\left(1 - \frac{C(t)}{K}\right),\ \theta > 0 \tag{8.13}$$

with

$$C_{\text{Richards2}}(t) = K\left(1 - \frac{1}{\left[1 + Q_2 e^{r\theta t}\right]^{1/\theta}}\right) \tag{8.14}$$

where $Q_2 = \left(\frac{K}{K-i_0}\right)^{\theta} - 1$. Clearly, $C_{\text{Richards2}}(t)$ in (8.14) and in (8.12) is the same, via re-parametrization $\rho = r\theta$ and $Q_2 = e^{-\rho\alpha}$.

The first derivative $\frac{d}{dt}C_{\text{Richards2}}(t)$ is

$$I_{\text{Richards2}}(t) = \frac{ke^{\rho(t-\alpha_2)}}{\left[1 + e^{\rho(t-\alpha_2)}\right]^{\frac{\theta+1}{\theta}}} \tag{8.15}$$

where $k = \frac{\rho}{\theta}K$.

If the initial value i_0, the scale parameter ρ, the shape parameter θ, and the carrying capacity K are all the same in both (8.10) and (8.12), the location parameter α in the corresponding solutions (8.10) and (8.12) is different. We denote them separately as α_1 and α_2, respectively. They are

$$\begin{aligned}
\alpha_1 &= \tfrac{1}{\rho}\log\left[\left(\tfrac{K}{i_0}\right)^{\theta} - 1\right], && \text{with respect to (8.9)}\\
\alpha_2 &= -\tfrac{1}{\rho}\log\left[\left(\tfrac{K}{K-i_0}\right)^{\theta} - 1\right], && \text{with respect to (8.14).}
\end{aligned}$$

The inflexion point for $C_{\text{Richards2}}(t)$ is $t_2^* = \alpha_2 + \frac{1}{\rho}\log\theta$ and $C_{\text{Richards2}}(t_2^*) = K\left(1 - \frac{1}{(1+\theta)^{1/\theta}}\right)$. At the inflexion point, $I_{\text{Richards2}}(t)$ reaches the peak value $I_{\text{Richards2}}(t_2^*) = \frac{\rho K}{(1+\theta)^{\frac{\theta+1}{\theta}}}$ which is the same as the peak value of $I_{\text{Richards}}(t)$.

Figure 8.1 compares $C_{\text{Richards}}(t)$ vs. $C_{\text{Richards2}}(t)$, and $I_{\text{Richards}}(t)$ vs. $I_{\text{Richards2}}(t)$, given $K = 1000$, $i_0 = 1$ and $\theta = 0.4$. Since ρ is a scale parameter with respect to time, without losing generality, we let $\rho = 1$. We have

$$\begin{aligned}
\alpha_1 &= 2.6979,\ t_1^* = 2.6979 - \log 0.4 = 3.6142\\
\alpha_2 &= 7.8233,\ t_2^* = 7.8233 + \log 0.4 = 6.907
\end{aligned}$$

and $I_{\text{Glogis1}}(t_1^*) = I_{\text{Glogis2}}(t_2^*) = 308$.

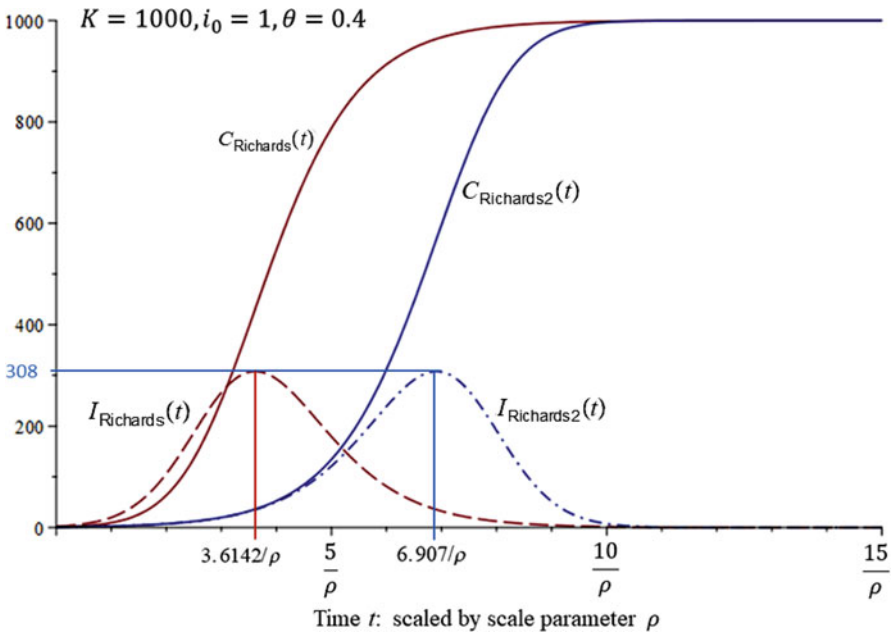


Fig. 8.1 Plots of $C_{\text{Richards}}(t)$, $I_{\text{Richards}}(t)$, $C_{\text{Richards2}}(t)$, and $I_{\text{Richards2}}(t)$ given $K = 1000$, $i_0 = 1$, $\theta = 0.4$. The time scale is standardized according to the scale parameter ρ. The maximum values for $I_{\text{Richards}}(t)$ and $I_{\text{Richards2}}(t)$ are equal: $\rho K\,(1+\theta)^{-\frac{\theta+1}{\theta}} = 308$ at set parameters

The variation of the Richards model (8.12) is closely related to disease transmission model (8.5). In (8.5), $\psi(t) = \frac{I_d(t)}{S_d(t)+I_d(t)}$ is a simple logistic growth function. However, $I_d(t)$ in (8.5) is an asymmetric bell-shaped curve

$$I_d(t) = M(0)\frac{\psi(0)e^{\rho t}}{\{[1-\psi(0)]+\psi(0)e^{\rho t}\}^{1+\gamma/\rho}}, \tag{8.16}$$

where $M(0) = S_d(0) + I_d(0)$. If we re-parameterize $\theta = \rho/\gamma$ and $\psi(0) = \left(\frac{K}{K-i_0}\right)^{\theta} - 1 \equiv Q_2$, then (8.16) becomes

$$I_d(t) = M(0)\frac{Q_2 e^{\rho t}}{\left\{\left(\frac{K}{K-i_0}\right)^{\theta} + Q_2 e^{\rho t}\right\}^{\frac{\theta+1}{\theta}}}.$$

On the other hand, (8.15) with re-parametrization is

$$I_{\text{Richards2}}(t) = k\frac{Q_2 e^{\rho t}}{\{1+Q_2 e^{\rho t}\}^{\frac{\theta+1}{\theta}}}.$$

Thus $I_{\text{Richards2}}(t)$ approximates $I_d(t)$ in (8.16) well when K is large and i_0 is small.

The model given by (8.15) is descriptive and captures the essence of time-series data based on a disease outbreak without any scientific hypothesis regarding the transmission dynamics. It has been used as approximations for some simple compartment models for HIV/AIDS, as discussed in detail in Chap. 9 of Brookmeyer and Gail (1994) and Chap. 1 of Tan (2000).

The following example illustrates the generalized logistic functions (8.12) and (8.15) as good approximations to two different SIR-type models and an SEIR model.

Example 33 In this example, we set $K = 9009$, $\alpha = 53.46$, $\varphi = 0.72$, $\rho = 0.155$ in (8.15). The parameter $\varphi = 0.72 < 1$ gives a slightly skewed incidence function $I_d(t)$ with peak time $t^* = 55.579$. According to this model, the final size is $C(\infty) = 9009$. Both (8.12) and (8.15) approximate very well with the incidence and cumulative incidence functions, which are implicitly determined by the following selected deterministic models (Fig. 8.2):

1. the SIR model given by (5.24) with $m = S(0) + I(0) = 14{,}300$, $\beta = 0.395$, $\gamma = 1/4$, corresponding to $R_0 = 1.58$ and final size 9025;
2. the SEIR model given by (5.53) with $m = S(0) + I(0) = 10{,}150$, $\beta = 0.6$, $\alpha = 1/3.4$, $\gamma = 1/4$, corresponding to $R_0 = 2.4$ and final size 8918;
3. an SIR model governed by the integro-differential equations

$$\begin{cases} \frac{d}{dt}S(t) = -\beta\frac{S(t)I(t)}{n} \\ \frac{d}{dt}I(t) = i(t) - \int_0^t i(s)f_I(t-s)ds, \text{ where } i(t) = \beta\frac{S(t)I(t)}{n} \\ \frac{d}{dt}R(t) = \int_0^t i_1(s)f_I(t-s)ds. \end{cases}$$

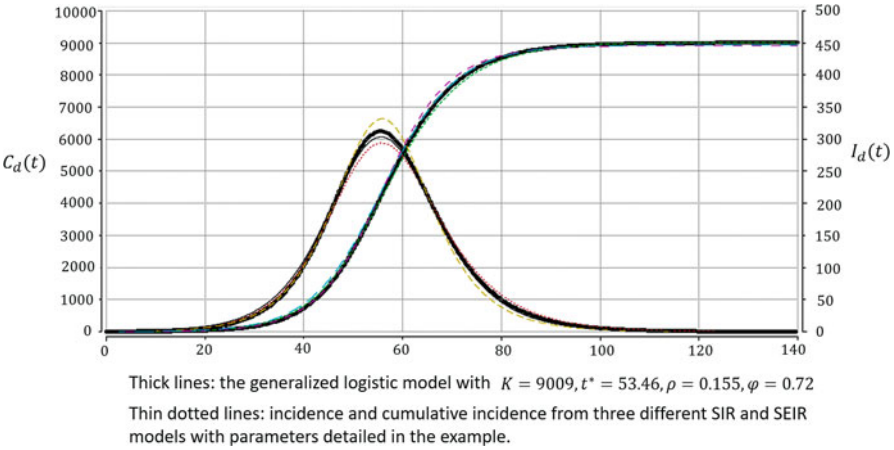


Fig. 8.2 Illustration of generalized logistic models $C_d(t)$ and $i_d(t)$ with comparison from three different transmission models in Example 33

with gamma distributed infectious period $f_I(x)$ with mean value $\mu_I = 4$ and shape parameter $\kappa = 3$, along with $m = S(0) + I(0) = 17{,}500$, $\beta = 0.35$, corresponding to $R_0 = 1.4$ and final size 8945.

Other Variations and Generalizations An alternative to (8.13) is

$$\frac{d}{dt}C(t) = rK\left[\frac{C(t)}{K}\right]^{\theta}\left(1 - \frac{C(t)}{K}\right), \tag{8.17}$$

Although it looks much simpler than (8.13), there is no explicit solution, but $C(t)$ can be solved numerically. It is also a sigmoid growth function with an inflexion point t^* at which $C(t^*) = \frac{p}{p+1}K$.

Other variations include

$$\frac{d}{dt}C(t) = r\left[C(t)\right]^{\theta}\left(1 - \frac{C(t)}{K}\right), \tag{8.18}$$

or the generalized Richards model (Turner et al. 1976) with two shape parameters $\theta_1, \theta_2 > 0$:

$$\frac{d}{dt}C(t) = r\left[C(t)\right]^{\theta_1}\left(1 - \left[\frac{C(t)}{K}\right]^{\theta_2}\right). \tag{8.19}$$

The model (8.19) with $0 < \theta_1 \leq 1$ was used to account for initial sub-exponential growth dynamics (Viboud et al. 2016). In this case, θ_1 is called the "deceleration of growth" parameter. This model has been useful to generate post-peak forecasts of Zika and Ebola epidemics (Pell et al. 2018a; Chowell et al. 2016).

Generalizations of the Logistic and Richards Functions So That $\lim_{t\to\infty} I(t) = c > 0$

It is straightforward to generalize $I_{\text{logis}}(t)$ into

$$I_{\text{logis-c}}(t) = \frac{1}{1 + e^{-\rho(t-\alpha)}}\left(\frac{ke^{-\rho(t-\alpha)}}{1 + e^{-\rho(t-\alpha)}} + c\right). \tag{8.20}$$

It returns to the logistic model when $c = 0$. However, $I_{\text{logis-c}}(\alpha)$ is not the maximum value unless $c = 0$ because $I'_{\text{logis-c}}(\alpha) = \frac{1}{4}c\rho \geq 0$. The maximum value is achieved when $t = t^* = \alpha - \frac{1}{\rho}\log\frac{k-c}{k+c}$ and the maximum value is $I_{\text{logis-c}}(t^*) = \frac{1}{4k}(k+c)^2$.

Similarly, one can generalize the Richards function (8.11) as

$$I_{\text{Richards-c}}(t) = \frac{1}{1 + e^{-\rho(t-\alpha)}}\left(\frac{ke^{-\rho(t-\alpha)}}{\left[1 + e^{-\rho(t-\alpha)}\right]^{\frac{1}{\theta}}} + c\right) \tag{8.21}$$

that includes five parameters. It returns to the Richards model (8.11) when $c = 0$; returns to (8.20) when $\theta = 1$ and returns to the logistic model when $c = 0$ and $\theta = 1$. It satisfies $I_{\text{Richards-c}}(-\infty) = 0$ and $I_{\text{Richards-c}}(\infty) = c \geq 0$. It reaches the peak value when $t = t^*$ that satisfies

$$ke^{-\rho(t-\alpha)} + c\theta\left(e^{-\rho(t-\alpha)} + 1\right)^{\frac{1}{\theta}} = k\theta.$$

For the special cases, $t^* = \alpha$ when $c = 0$ and $\theta = 1$; $t^* = \alpha - \frac{1}{\rho}\log\theta$ when $c = 0$ and $t^* = \alpha - \frac{1}{\rho}\log\frac{k-c}{k+c}$ when $\theta = 1$.

A different generalization of (8.20) is

$$I_{\text{logis-c-2}}(t) = \frac{1}{1 + e^{-\rho(t-\alpha)}}\left[\frac{ke^{-\eta(t-\alpha)}}{1 + e^{-\eta(t-\alpha)}} + c\right] \tag{8.22}$$

where

ρ = rate of increase at the beginning,
η = rate of convergence to the asymptote.
a = a suitable location parameter,
$c = \lim_{t\to\infty} I_{\text{logis-c-2}}(t)$ = asymptote.

The generalized logistic function given by (8.22) has been adopted as one of the parametric models for the incidence of new HIV infections in a computer package, Spectrum (Avenir Health), which is endorsed by the Joint United Nations Programme on HIV/AIDS (UNAIDS) to compile estimates of HIV prevalence in different countries around the world.

Both (8.21) and (8.22) have five parameters. They may be used to approximate SEIRS models in a constant population with demography turn-over. A comparison is shown in Fig. 8.3. However, it is inconvenient to interpret the model (8.22) representing a time-series with two different scale parameters, η and ρ.

These generalized logistic functions can be used to capture the essence of time-series data for prediction purposes. By adding more parameters, one can create models that can capture a broad variety of epidemic curves. For example, the following 6-parameter function

$$I_{twin-peak-c}(t) = \frac{k_1 e^{-\rho(t-\alpha_1)}}{(1 + e^{-\rho(t-\alpha_1)})^2} + \frac{1}{1 + e^{-\rho(t-\alpha_2)}}\left[\frac{k_2 e^{-\rho(t-\alpha_2)}}{1 + e^{-\rho(t-\alpha_2)}} + c\right]$$

is capable of creating a twin peaked curve that approaches an asymptote $c > 0$. However, the time-series data, especially data from a single source, do not have enough information to test hypotheses or make statistical inferences on parameters with specific biological and epidemiological interpretations in transmission models.

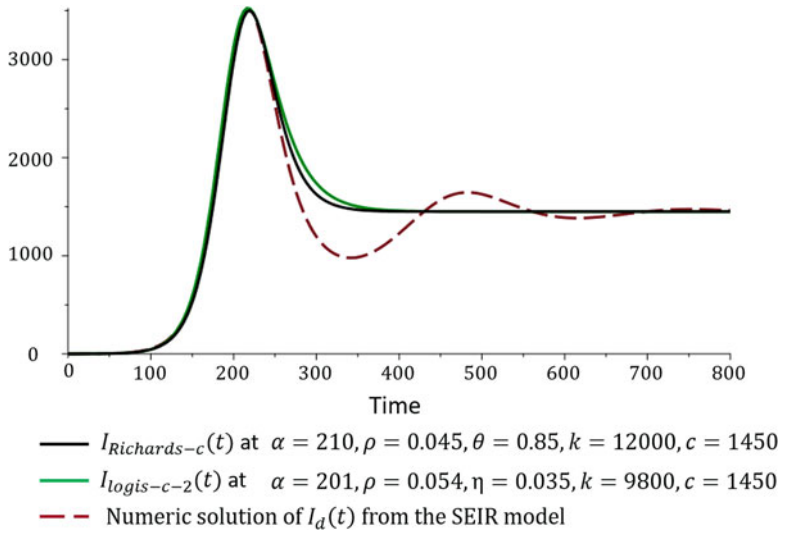


Fig. 8.3 Compare the functions (8.21) and (8.22) with selected parameters against the numbers of infectious individuals over time in an SEIRS model using parameters in Example 25 (dashed line)

8.3 A Comprehensive Demonstration of Curve Fitting Using Nonlinear Phenomenological Models to Outbreak Data from the 2016 Zika Epidemic in Antioquia, Colombia

In the Western Hemisphere, active circulation of ZIKV was first reported in Brazil in May 2015, and the WHO declared the epidemic a Public Health Emergency of International Concern on February 1, 2016. Phylogenetic analyses indicate that the epidemic in the Americas was triggered by an imported case sometime between May and December 2013, a period that coincides with an increase in air travel from ZIKV affected areas in the Pacific to Brazil (Faria et al. 2016).

We analyzed daily counts of Zika cases by date of symptoms onset reported to the Secretary of Health of Antioquia (the time series is available online as an EXTRA MATERIAL). Antioquia is the second largest department in Colombia (with a population size of ~6.3 million people), located in the central northwestern part of the country (Chowell et al. 2016). Because there is still substantial uncertainty on the epidemiology of ZIKV, including the contribution of different modes of transmission (mosquito bites vs. sexual transmission), simple phenomenological models are useful for forecasting epidemic trajectories whereas mechanistic mosquito-borne disease transmission models require more data to appropriately calibrate mosquito reproduction, development, survival, and transmission capacity which are strongly modulated by temperature as well as the transmission rates that dictates the transfer of the virus from mosquitoes to humans and vice versa (e.g., Ross 1911; Focks et al. 1995; Chowell et al. 2007; Gao et al. 2016; Towers et al. 2016; Zhang et al. 2017; Huber et al. 2018).

8.3.1 Fitting Models to Data

The time-series data $\underline{y} = (y_0, y_1, \ldots, y_T)$ represent daily incidence according to the onset of clinical symptoms. The date of the earliest recorded case by date of onset is called Day 0. We model the cumulative number of clinical cases by time t according to a counting process. The random component of the model is the marginal distribution of $C(t)$ grouped into time intervals so that $Y_t = C(t) - C(t-1)$, $-\infty < t < \infty$ is the number of clinical onsets during the time interval $(t-1, t]$. The systematic component of the model is a deterministic growth curve function $C_d(t)$, chosen as one of the growth curve functions introduced in the preceding section, such that the expected value is $E[C(t)] = C_d(t)$, specified by a set of parameters $\Theta = (\theta_1, \ldots, \theta_m)$.

Since we are fitting the model to daily incidence data y_t, we write the systematic component as

$$\mu(t; \Theta) = C_d(t) - C_d(t-1).$$

Most of the growth functions are defined over $-\infty < t < \infty$. Therefore the expected number of new cases by date of symptoms on Day 0 is $E[Y_0] = \mu(0; \Theta) = C_d(0) - C_d(-1)$. In the observed data, $y_0 = 1$.

We use the methods presented and discussed in Sect. 7.1.3 in the following analyses.

8.3.2 Data During the First 20 Days

Exploratory Analysis

Starting from the earliest recorded case by date of onset, denoted as Day 0, the cumulative number of confirmed individuals with clinical symptoms was 183 by Day 20. During the first 2 weeks, daily incidence numbers by symptoms onset were rather sporadic, less than 10 cases per day except for Day 9 and Day 14. From Day 15 to Day 20, the daily incidence numbers were fluctuating between 12 and 20 cases per day.

Exploratory plots of the logarithm of daily incidence data $(y_t,\ t = 0, \ldots, 20)$ against time t and against the logarithm of the cumulative incidence $c_t = \sum_{i=0}^{t} y_i$, $t = 0, \ldots, 20$ (Fig. 8.4) show distinctive sub-exponential growth patterns. In particular, the strong linear relationship between the logarithm of daily incidence data and the logarithm of the cumulative incidence is given by

$$\log y_t = -0.06008 + 0.55466 \log \sum_{i=0}^{t} y_i,$$

which empirically agrees with the relationship $C'(t) = rC(t)^p$, in which, $C'(t)$ is approximated by the daily incidence y_t, $t = 1, \ldots, 20$ and $y_0 = 1$ and $C(t)$ is

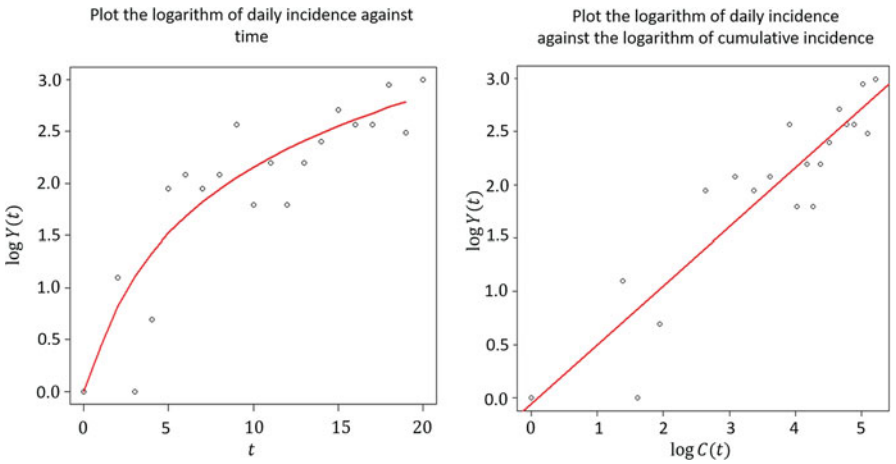


Fig. 8.4 Exploratory plots of the logarithm of daily incidence data $\log y_t$ against time t and against the cumulative incidence $c_t = \sum_{i=0}^{t} y_i$, $t = 0, .., 20$

approximated by the cumulative incidence $c_t = \sum_{i=0}^{t} y_i$. This relationship gives the crude estimates

$$\widetilde{r} = e^{-0.06008} = 0.94169, \ \widetilde{p} = 0.55466.$$

Likelihood Analysis, Estimation and Predictions for the Sub-exponential Model

For formal analysis, we start fitting to daily incidence data using (7.3), assuming that $\underline{y} = (y_0, y_1, \ldots, y_{20})$ are realizations of independent Poisson random counts. The likelihood based approach based on (7.4) is applied with $f(0; \Theta) = i_0$ and $f(t; \Theta) = C(t) - C(t-1)$, $t = 1, \ldots, 20$, where $C(t) = i_0(1 + r(1-p)t)^{\frac{1}{1-p}}$, $0 \leq p < 1$.

We first conduct the likelihood ratio test against the hypothesis $H_0 : i_0 = 1$, which yields a significant level (p-value) of 0.48. There is no evidence from data to reject H_0.

We consider the reduced model $C(t) = (1 + r(1-p)t)^{\frac{1}{1-p}}$, which is the exact solution of $C'(t) = rC(t)^p$ given the initial condition $C(0) = 1$. The parameters are $\Theta = (r, p)$. The maximum likelihood estimates are

$$\widehat{r} = 1.172 \ (0.8173, 1.777)$$

$$\widehat{p} = 0.5189 \ (0.4231, 0.6186)$$

where numbers in brackets are 95% confidence limits calculated using likelihood ratio statistics.

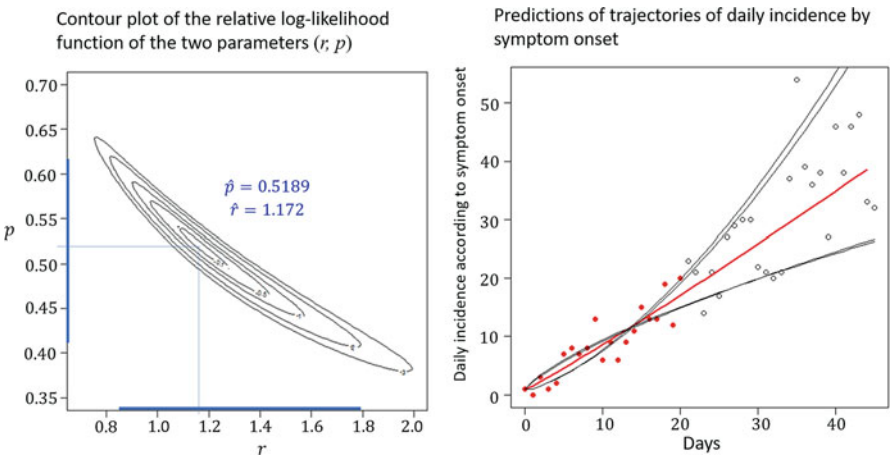


Fig. 8.5 Left: the contours of the relative log-likelihood function for (r, p) in the neighborhood of $(\widehat{r}, \widehat{p})$. Right: five predicted trajectories of daily incidence of onset of symptoms to Day 45. Red dots represent data from the first 20 days and circles are data from Day 21 to Day 45

One of the advantages of using the likelihood based approach is that the likelihood function reveals how much information data contain with respect to each of the parameters as well as correlations among parameters. Figure 8.5 (left) displays the contours of the relative log-likelihood function for (r, p) in the neighborhood of the maximum likelihood estimates. It shows substantial correlation due to the "banana" shape of the log-likelihood contour. Although the 95% confidence interval for each parameter has been calculated marginally for each parameter, not all the combinations of the two parameters within their ranges are plausible. For example, it is very implausible to have the combination $(r_L = 0.8173,\ p_L = 0.4231)$. This leads to the concept of the "profile likelihood," which is to fix one of the parameters at a given value and conduct a likelihood analysis on the rest of the parameters.

- Keeping r fixed at its lower bound at $r_L = 0.8173$, the profile likelihood for p is maximized at $\widehat{p}(r = 0.8173) = 0.6167$.
- Keeping r fixed at its upper bound at $r_U = 1.777$, the profile likelihood for p is maximized at $\widehat{p}(r = 1.777) = 0.411$.
- Keeping p fixed at its lower bound at $p_L = 0.4231$, the profile likelihood for r is maximized at $\widehat{r}(p = 0.4231) = 1.6693$.
- Keeping p fixed at its upper bound at $p_U = 0.6186$, the profile likelihood for r is maximized at $\widehat{r}(p = 0.6186) = 0.8222$.

Short term predictions may be conducted in ad hoc manner by simple extrapolation based on the model $f(t; \Theta)$ and the range of uncertainties in estimated parameters. Figure 8.5 (right) displays five predicted trajectories of daily incidence of onset of symptoms to Day 45 based on observed data up to Day 20. The center red line is the predicted trajectory based on the m.l.e. $(\widehat{r} = 1.172,\ \widehat{p} = 0.5189)$. The four

thin dark lines are predicted trajectories based on the combinations: $r = 0.8173$ and $p = 0.6167$; $r = 1.777$ and $p = 0.411$; $r = 1.6693$ and $p = 0.4231$; $r = 0.8222$ and $p = 0.6186$.

A word of caution is in place. Prediction of future trajectories based on historical data involves two sources of uncertainty: the uncertainty about the parameters and the uncertainty of future data due to randomness for any fixed parameter values in the probability distribution. The predictions in Fig. 8.5 (right) partially take into account the first source uncertainty but fail to take into account the second source. This issue is deeply rooted in the foundations of statistical inferences, and there is a scarcity of literature on "predictive likelihood" that is applicable to predicting trajectories of disease outbreaks.

Least Square Estimation and Predictions for the Sub-exponential Model

The least square method by minimizing (7.6) provides similar estimates

$$\widetilde{r} = 1.2023\ (0.76,\ 1.9)$$
$$\widetilde{p} = 0.5119\ (0.39,\ 0.64)$$

where numbers in brackets are 95% confidence limits based on 500 bootstrap samples, which are slightly wider than, but comparable to, those based on the likelihood ratio statistics.

Both the maximum likelihood estimates based on the Poisson distribution and the least square estimates are based on unbiased estimating equations. They are asymptotically unbiased point estimates regardless of any mis-specification of the variance–covariance structure. The word "asymptotic" is used in the sense of a large number of realizations of the same epidemic assuming the outbreak can be repeated under identical conditions. However, the point estimates in both methods are based on a single realization and are subject to biases.

For assessment of uncertainties, both methods are prone to mis-specification of the variance–covariance structure. In order to compare with the variance estimates based on the likelihood approach, 500 bootstrap replicates are generated for the least square estimation assuming a Poisson variance structure. Prediction intervals are also generated using bootstrapping to predict the distribution of individual future points.

The 95% confidence intervals for p based on both methods show strong significance against the hypothesis of the exponential growth function: $p \to 1$.

With respect to prediction of future trajectories, the bootstrapping method takes into account both sources of uncertainty. The cyan curves in Fig. 8.6 (right) correspond to 500 bootstrap replicates of the epidemic curve assuming a Poisson variance structure. These are predicted random numbers. The uncertainty in predicted trajectories is much larger than that in Fig. 8.5 (right). However, there are several issues worth discussing.

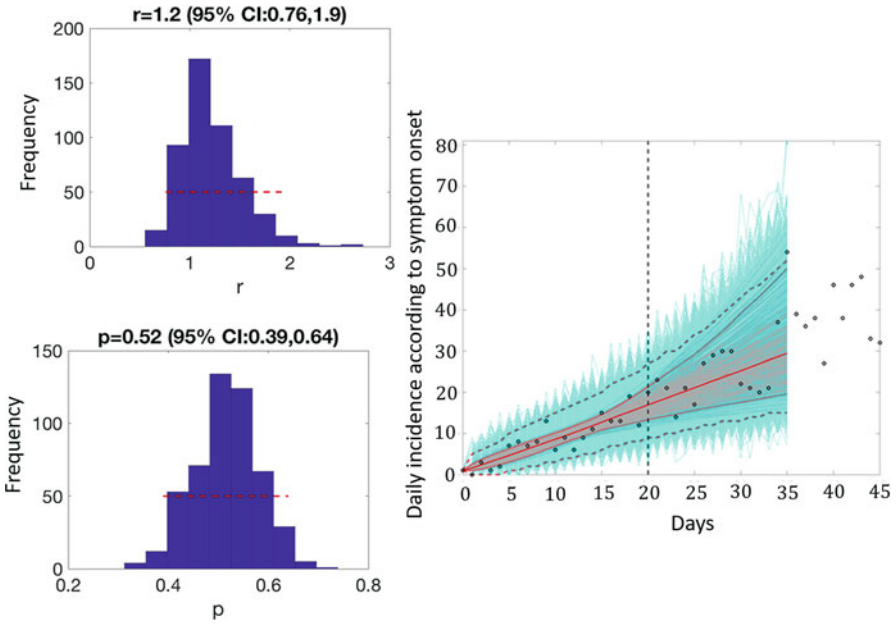


Fig. 8.6 Left: empirical distributions of the estimated parameters based on 500 bootstrap replicates. Right: 15-day forecast when the model is calibrated to the first 20 days. The black circles are the daily incidence data. The cyan curves correspond to 500 bootstrap replicates of the epidemic curve assuming a Poisson error structure. The red solid line corresponds to the asymptotic mean value of these replicates. The gray lines correspond to the fits of the model to each of the 500 bootstrap replicates from which 95% confidence intervals for the mean model fit can be derived (red dotted lines). The vertical line separates the calibration and forecasting periods

First, we note that the marginal distributions of the parameters (r, p) alone do not provide the insight as some of the combination of parameters (r, p) while $r \in (0.76, 1.9)$ and $p \in (0.39, 0.64)$ are highly implausible. Therefore, running simulations in these parameter ranges may produce larger than expected uncertainty. However, one could make use of the raw empirical distributions of the parameters including their correlations which were derived from the bootstrap approach in order to avoid selecting implausible parameter combinations.

Second, the vertical line separates the calibration and forecasting periods in Fig. 8.6 (right). The cyan curves correspond to 500 bootstrap replicates of the epidemic curve assuming a Poisson variance structure. They show large uncertainty in data that have already occurred. This is due to the virtual experiment conducted by the computer simulation assuming the outbreak can be repeated in identical conditions and the uncertainty in data reflects such randomness. However, the disease outbreak only occurs once and data in the past 20 days are given. Given past data, conditional prediction for the future is desirable whereas extrapolating "predictions" made for the past that include large uncertainty into the future is not desirable.

Fitting the Logistic Growth Model to the First 20 Epidemic Days of Data

Choosing the appropriate models and how to parameterize the models depends on what public health questions need to be addressed. For instance, public health officials may be less interested in short term predictions but more interested in questions such as

- Are the daily incidence numbers approaching the peak value?
- If this outbreak is going to be a single wave, how long is this wave expected to last?
- How many cumulative infections do we expect by the end of this wave?

Although sub-exponential growth (8.1) describes the growth pattern for the first 20 days of data and makes short-term predictions, it is unable to answer these questions. The logistic growth function characterizes such single wave phenomenon. The expression

$$C(t) = \frac{K}{1 + e^{-\rho(t-\alpha)}}$$

corresponds to the three questions, where the peak time is the location parameter α; the peak value of daily incidence is $C'(\alpha) = \rho K/4$ and by the end of the outbreak, the cumulative number of infected individuals is K. In addition, the logistic model is symmetric such that, at the peak time α, the cumulative incidence $C(\alpha) = K/2$.

For data $\underline{y} = (y_0, y_1, \ldots, y_{20})$, we specify the mean value $E[Y_t] = f(t; \Theta)$ where $f(t; \Theta) = C(t) - C(t-1)$ is the difference of the two adjacent cumulative values. Since the logistic model is defined for $-\infty < t < \infty$, $f(0; \Theta) = C(0) - C(-1)$ which is the expected value for Y_0.

The Likelihood Analysis The maximum likelihood estimates, assuming Poisson distribution for Y_t, are

$$\widehat{\rho} = 0.171 \ (0.092, \ 0.238)$$

$$\widehat{\alpha} = 19.71 \ (15.5, \ 45.8)$$

$$\widehat{K} = 377.898 \ (246, \ 4250)$$

where numbers in brackets are 95% confidence limits based on the likelihood ratio statistics. The very wide confidence limits show that data have little information about the key parameters of interest. The peak time could be anywhere from Day 15 to Day 46, and the total number of infections by the end of the wave could be anywhere between 246 and 4250.

Figure 8.7 illustrates cross-sectional contour plots for (α, K) according to selected growth rate values of ρ. Figure 8.7 shows that at a slow growth rate $\rho_L = 0.092$, the likelihood function suggests that the most plausible peak time occurs around Day 46, and by the end of the outbreak, the total number of infections would most likely be around 2500. However, there is a great deal of uncertainty

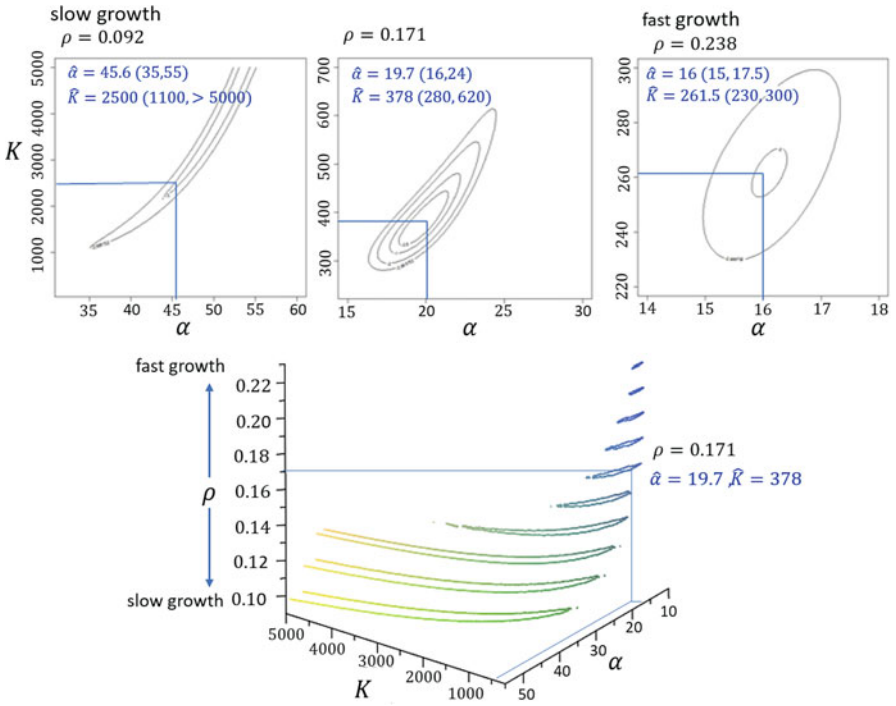


Fig. 8.7 Cross-sectional plots of the contours of the log-likelihood surface for (α, K) at $\rho = 0.092$, 0.171, and 0.238

suggesting that K could be as large as >5000. On the other hand, at a fast growth rate $\rho_H = 0.238$, the most plausible peak time occurs around Day 16, and the likelihood function predicts a small outbreak with the most plausible $K = 262$.

The logistic function can be also parameterized as

$$C(t) = \frac{K}{1 + e^{-\rho(t-\alpha)}} = \frac{i_0(v+1)}{v + e^{-\rho t}}$$

where $i_0 = C(0)$ is the cumulative number of clinical cases at time $t = 0$ and $v = e^{-\rho\alpha} = \frac{i_0}{K - i_0}$. The maximum likelihood estimates are

$$\widehat{\rho} = 0.171\ (0.092,\ 0.238)$$
$$\widehat{v} = 0.034\ (0.00001,\ 0.061)$$
$$\widehat{i_0} = 12.5\ (5.25,\ 39),$$

where numbers in brackets are 95% confidence limits based on the likelihood ratio statistics. Since data have little information about K, they equally have little information about v.

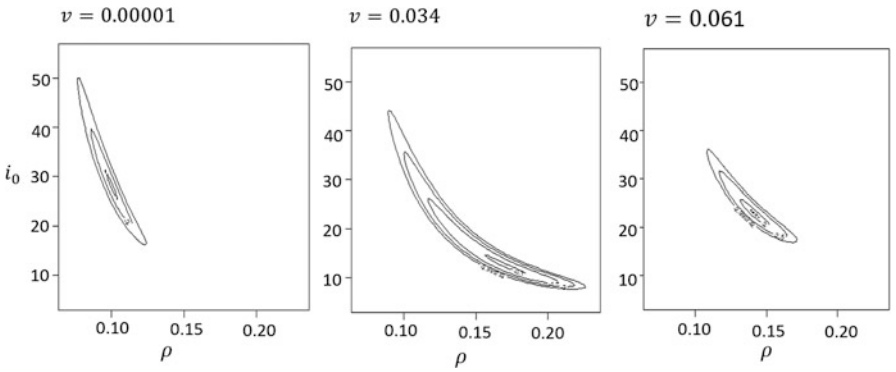


Fig. 8.8 Cross-sectional plots of the contours of the log-likelihood surface for (ρ, i_0) at $v = 0.00001$, 0.034, and 0.061

The logistic model suggests that on Day 0, the cumulative number $i_0 = C(0)$ was likely between 5 and 39, suggesting the outbreak might have started earlier. This estimated range, combined with the uncertainty estimates of the other parameters, suggests that the range of daily incidence on Day 0 might be in the range (1.1, 3.4). This is because the expected value for Y_0 is $f(0) = C(0) - C(-1)$. In the data, $y_0 = 1$.

Based on the contours of the log-likelihood (Fig. 8.8), the growth rate ρ is correlated with the initial cumulative number $i_0 = C(0)$. Although there was little information in the early data, we may tentatively make the following statements regarding the following three scenarios:

1. High cumulative numbers i_0 in the range between 30 and 40 (approximately three new clinical cases on Day 0) combined with slow growth $\rho < 0.1$. Under this scenario, it is likely that the peak time will be rather late, after Day 40. Consequently, the cumulative number by the end of the wave (assuming a single wave) might be above 2000, or even above 4000.
2. A plausible cumulative numbers i_0 around 12 (approximately two new clinical cases on Day 0) combined with a growth rate around 0.171. Although this scenario corresponds to the maximum likelihood, it is also likely to be biased, as it puts the estimated peak time at $\widehat{\alpha} = 19.71$ corresponding to the last data point ($t = 20$). There is no indication in data that the daily incidence is reaching its peak. If this were true, then the cumulative number by the end of the wave would be approximately twice the cumulative number at Day 20, which is 377.898.
3. Low cumulative numbers i_0 around 5 (approximately one new clinical case on Day 0) combined with a fast growth $\rho > 0.23$. This scenario may fit better for the very early part of the data, for instance, $t = 0, \ldots, 5$. However, there is no indication in data suggesting that the peak time has taken place before Day 20.

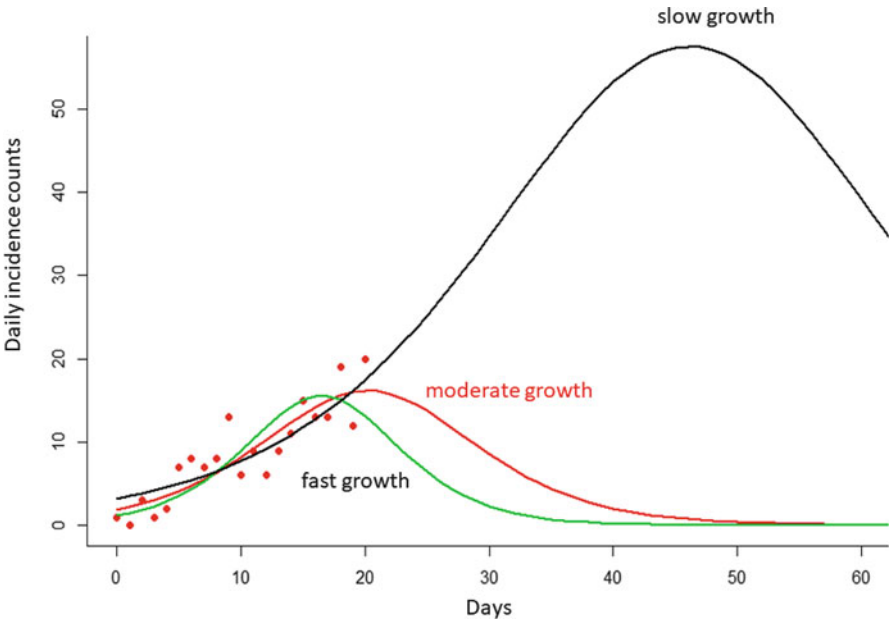


Fig. 8.9 Three predicted trajectories: (1) slow growth at $r = 0.092$, $\alpha = 45.6$ and $K = 2500$; (2) moderate growth at $r = 0.171$, $\alpha = 19.7$ and $K = 370$; (3) fast growth at $r = 0.238$, $\alpha = 16$ and $K = 262$

Figure 8.9 shows predictions to Day 60, with parameters chosen from the slow growth (with high i_0), moderate and fast growth patterns (with lower i_0). Data from the first 20 days cannot distinguish these scenarios.

We remark that the likelihood surfaces are highly asymmetric with respect to the parameters of interest. For example, the m.l.e. $\widehat{K} = 377.898$ is close to its lower bound 246 but away from its upper bound 4250. The likelihood function suggests equal likelihood between $K = 246$ and $K = 4250$. Therefore it is more plausible that the true value of K lies in the region $(378, 4250)$ than in the region $(246, 378]$. Similarly, the m.l.e. for the peak time $\widehat{\alpha} = 19.71$ is also associated with an asymmetric likelihood based confidence interval between $\alpha = 15.5$ and 45.8. Together, they provide asymmetric scenarios in the predicted trajectories in Fig. 8.9. Although these predictions are very imprecise and not very useful, the asymmetric feature may also suggest that it is more plausible that the outbreak has not yet peaked and the final cumulative number could be in thousands. Only future data can tell.

8.3.3 *Data During the First 45 Days*

The Logistic Model: Likelihood Based Analyses

We re-fit the three-parameter logistic model to daily incidence data y_t, $t = 0, \ldots, 45$. The maximum likelihood estimates for the logistic function are

$$\begin{aligned} \widehat{\rho} &= 0.0917\ (0.0787,\ 0.105) \\ \widehat{\alpha} &= 41.29\ (41.21,\ 43.97) \\ \widehat{K} &= 1689.992\ (1422,\ 2088) \end{aligned} \tag{8.23}$$

where numbers in brackets are 95% confidence limits calculated using likelihood ratio statistics. When parameterized as

$$C(t) = \frac{i_0 K}{i_0 + (K - i_0)\, e^{-\rho t}}$$

the parameter $i_0 = C(0) = \frac{K}{1+e^{\rho\alpha}}$ has the epidemiologic meaning as the cumulative number of clinical cases by Day 0. Treating i_0 as a parameter, the maximum likelihood estimate is

$$\widehat{i_0} = 37.5\ (31.95, 43.2).$$

The expected daily incidence on Day 0 is $\widehat{C}(0) - \widehat{C}(-1) = 3.2$, as opposed to a single case on Day 0 as in the reported data.

Compared to the maximum likelihood estimates based on the first 20-day data, the extra information from Day 21 to Day 45 has resulted in remarkably improved precision for all the parameters.

The revised likelihood analysis suggests that Scenario 1 from analyses using the first 20 data points has turned out to be the most likely scenario. It is implausible that the outbreak could have peaked before Day 41. There is also evidence, at significance level 0.05, that the outbreak has peaked by Day 44. Updated data suggest a much narrower range for the uncertainty of K.

The logistic model suggests that the outbreak started approximately 30 days prior to Day 0. Since the logistic model gives a symmetric daily incidence curve, it further suggests that the outbreak will probably end around Day 115.

We update Fig. 8.9 as Fig. 8.10. The three scenarios are: (1) early peak at $\alpha_L = 41.21$, with $\rho = 0.0898$, $K = 1799$; (2) at the maximum likelihood estimates: $\widehat{\rho} = 0.0917$, $\widehat{\alpha} = 41.29$ and $\widehat{K} = 1689.992$; (3) late peak at $\alpha_U = 43.97$, with most plausible values $\rho = 0.08626$, $K = 1799$.

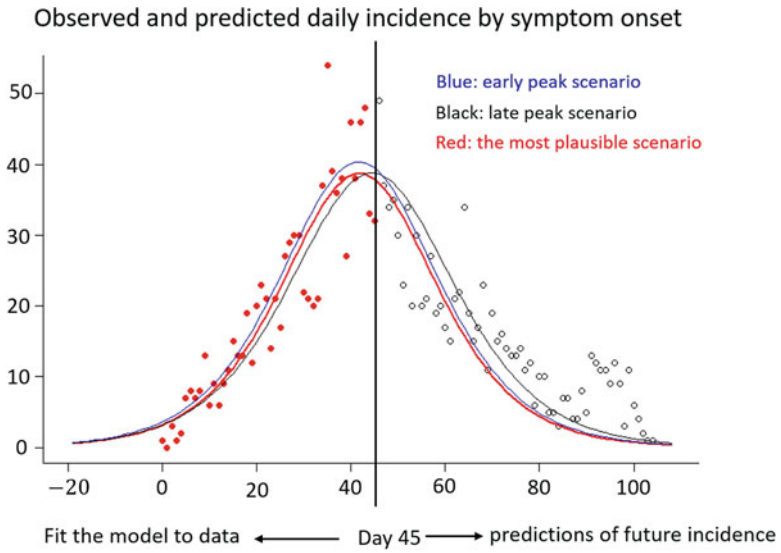


Fig. 8.10 Predictions with information from the first 45 days of data

The Logistic Model: Least Square Analysis

Least square estimates are performed based on the logistic function parameterized as $C(t) = \frac{i_0 K}{i_0 + (K - i_0)e^{-\rho t}}$. The estimated parameter values are

$$\widetilde{\rho} = 0.08922,\ \widetilde{i}_0 = 43.137,\ \widetilde{K} = 1804.4.$$

The least square method suggests that the cumulative number of infections by Day 0 was $\widetilde{C}(0) = \widetilde{i}_0 = 43.137$ and the daily incidence on Day 0 was 3.5. The estimated peak incidence time is $\widetilde{\alpha} = \frac{1}{\widetilde{\rho}} \log \frac{\widetilde{K} - \widetilde{i}_0}{\widetilde{i}_0} = 41.576$. All these estimates are in close agreement with the maximum likelihood estimates.

We compare predicted daily incidence by symptom onset based on the maximum likelihood estimation and the least square estimation from the logistic model against the observed daily incidence data as circles in Fig. 8.11. Each curve in Fig. 8.11 represents the expected values $f(t; \widehat{\Theta})$, $t = 0, \ldots, 45$. These values, together with data $(y_t, t = 0, \ldots, 45)$, are used to compute the summary measures MSE, WMSE, and Anscombe in (7.12), (7.13) and (7.14), respectively.

Residual analyses in Table 8.1 show that the least square estimates give the smaller mean square errors (MSE) by default and also slightly outperform the maximum likelihood estimates based on the sum of the squares of the Pearson residuals (WMSE). The maximum likelihood estimates perform slightly better based on the sum of the squares of the Anscombe residuals.

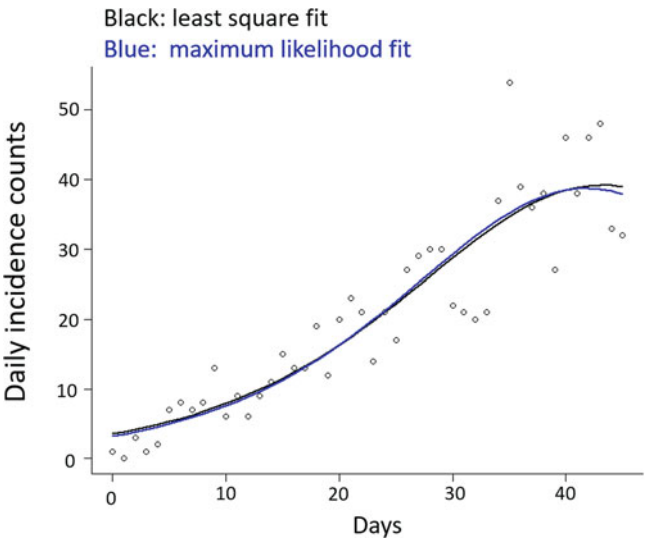


Fig. 8.11 Observed daily incidence for the first 45 days (circles) and two expected daily incidence curves as predicted by the logistic model, using the least square method (black) and the maximum likelihood method (blue)

Table 8.1 Summary residual measures (7.12)–(7.14) comparing the maximum likelihood and the least square estimates

	Maximum likelihood	Least square
MSE	32.37	32.17
WMSE	63.84	63.44
Anscombe	69.6	70.15

It is more informative to plot the residuals

$$r_t = y_t - f(t;\widehat{\Theta}),$$

$$r_t^{(P)} = \frac{y_t - f(t;\widehat{\Theta})}{\sqrt{f(t;\widehat{\Theta})}}, \qquad t = 0,\ldots,45$$

$$r_t^{(A)} = \frac{\frac{3}{2}\left[y_t^{2/3} - f(t;\widehat{\Theta})^{2/3}\right]}{f(t;\widehat{\Theta})^{1/6}}.$$

rather than their sum-of-squares. These residuals are plotted in Fig. 8.12. The plotted Anscombe residuals $r_t^{(A)}$ are approximately Gaussian distributed. Therefore the standard error lines ± 1.96 are also plotted in Fig. 8.12. The Anscombe residual plots in Fig. 8.12 detect two significant outliers for the maximum likelihood estimates.

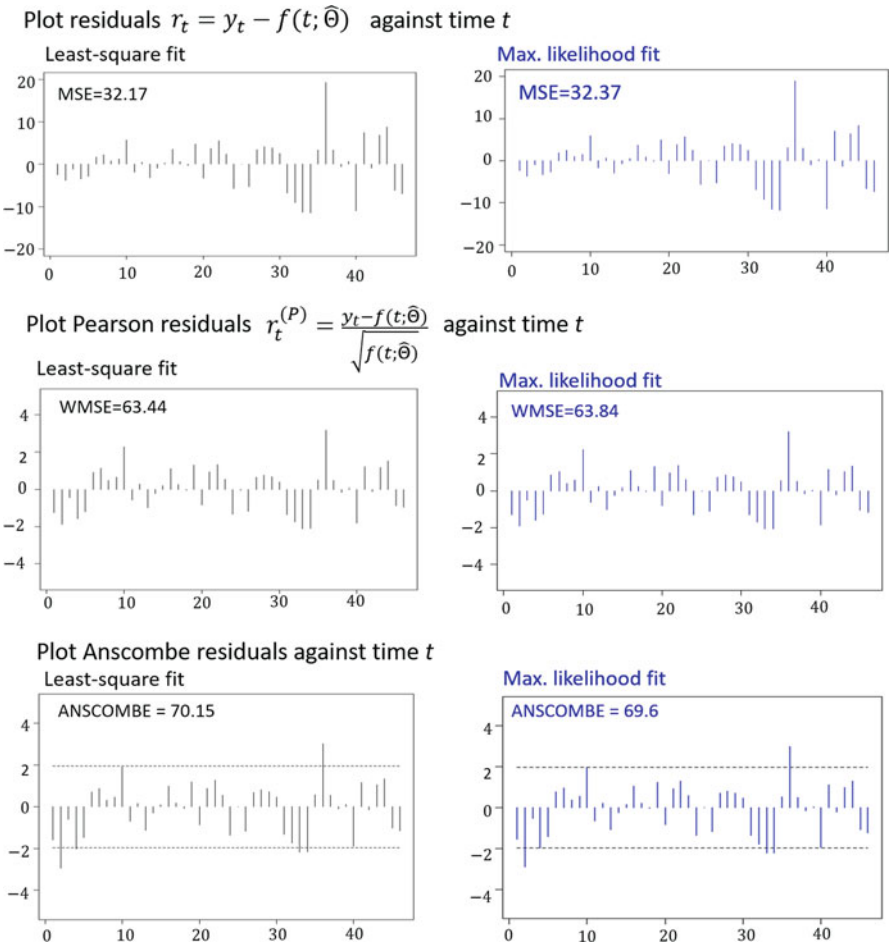


Fig. 8.12 Compare residuals between the least square estimates and the maximum likelihood estimates based on the first 45 epidemic days

Fitting Generalized Logistic Models to the First 45-Day Data with Discussions on Over-parameterization

One of the reasons to fit a generalized logistic model is to test the goodness-of-fit of the logistic model. For example, the Richards model

$$C(t) = \frac{K}{\left(1 + \left[\left(\frac{K}{i_0}\right)^{\theta} - 1\right] e^{-\rho t}\right)^{1/\theta}} \tag{8.24}$$

is the generalization of the logistic model $C(t) = \frac{i_0 K}{i_0+(K-i_0)e^{-rt}}$ when $\theta = 1$. In differential equation forms, the Richard model (as parameterized above) is the solution of

$$\frac{d}{dt}C(t) = \frac{\rho}{\theta}C(t)\left(1 - \left[\frac{C(t)}{K}\right]^{\theta}\right), \ \theta > 0 \tag{8.25}$$

whereas the logistic model is the solution of $\frac{d}{dt}C(t) = rC(t)\left(1 - \frac{C(t)}{K}\right)$, where $r = \rho/\theta$.

To test against the null hypothesis $H_0 : \theta = 1$, we conduct the likelihood ratio test based on (7.11). The value of the test statistics is

$$D = -2\left[l(\widehat{\rho}, \widehat{i_0}, \widehat{K} \,|H_0) - l(\widehat{\rho}, \widehat{i_0}, \widehat{K}, \widehat{\theta})\right] = 1.532 \tag{8.26}$$

where $l(\widehat{\rho}, \widehat{i_0}, \widehat{K} \,|H_0)$ is the value of the log-likelihood at $\widehat{\rho} = 0.0917$, $\widehat{i_0} = 37.5$ and $\widehat{K} = 1689.992$ assuming $\theta = 1$; and $l(\widehat{\rho}, \widehat{\theta}, \widehat{i_0}, \widehat{K})$ is the value of the log-likelihood of the Richards model where the maximum likelihood estimates are

$$\begin{aligned}
\widehat{\rho} &= 0.0637 \ (0.0274, \ 0.086) \\
\widehat{i_0} &= 23.9 \ (8.7, \ 51.8) \\
\widehat{K} &= 1999.76 \ (1502.5, \ 6300) \\
\widehat{\theta} &= 0.476 \ (0.132, \ 1.05)
\end{aligned} \tag{8.27}$$

Numbers in brackets are 95% confidence limits calculated using likelihood ratio statistics. The significance level for $H_0 : \theta = 1$ based on the likelihood ratio test is

$$SL = \Pr(\chi^2_{(1)} \geq 1.532) = 0.2155. \tag{8.28}$$

Discussion Although there is no evidence to reject the logistic model, the question remains whether we should discard the four-parameter Richards model which treats the additional parameter θ as a nuisance parameter, or adopt the Richards model as it may provide more valuable knowledge of public health importance.

From the point of view in favor of treating θ as a nuisance parameter, the focus is on the enormous uncertainty for the parameter of interest, such as $1502 < K < 6300$, compared to the "more precise" estimate $1422 < K < 2088$ by assuming $\theta = 1$. The argument is that information in the limited data from the first 45-days is "wasted" in the estimation of θ, which does not have direct public health interpretations as other parameters do (e.g., growth rate, initially infected number, peak time, final size). The very large uncertainty in the estimation of K, due to the inclusion of θ as a free parameter to be estimated, is not useful for public health decision makers. With this argument, the Richards model is "over-parameterized." In fact, the Richards model can be re-written in a logistic form such that $C(t)^{\theta}$ is a logistic function

$$C(t)^{\theta} = \frac{i_0^{\theta} K^{\theta}}{i_0^{\theta} + \left(K^{\theta} - i_0^{\theta}\right) e^{-\rho t}}.$$

Limited data are not informative in separating θ from K in the combined form K^{θ}. After all, the estimated θ is also associated with large uncertainty. As there is no statistical significance to reject the logistic model based on (8.28), we should discard the four-parameter Richards model.

The opposite point of view is that the significance level based on (8.28) implies that both the logistic and the Richards model fit data equally well up to Day 45 (see Fig. 8.13). It is only about the goodness-of-fit of models with respect to early part of the data, not about the important questions relating to the entire outbreak. Hence the "overly confident" estimates based on the logistic model may fail to acknowledge large uncertainties beyond Day 45. As shown in Fig. 8.13, when $\theta < 1$, the Richards model gives an asymmetric daily incidence curve with a longer tail, which is not only more realistic in most settings, but also suggests that the logistic model might have under estimated K. In fact, comparing (8.23) and (8.27), the logistic model might have under predicted approximately 300 clinical cases by the end of the outbreak. Meanwhile, the wide confidence intervals in (8.27) should be appreciated and emphasized.

At this moment in time (assuming we were on Day 45), we take notes on both arguments and move to the next phase when more data are available.

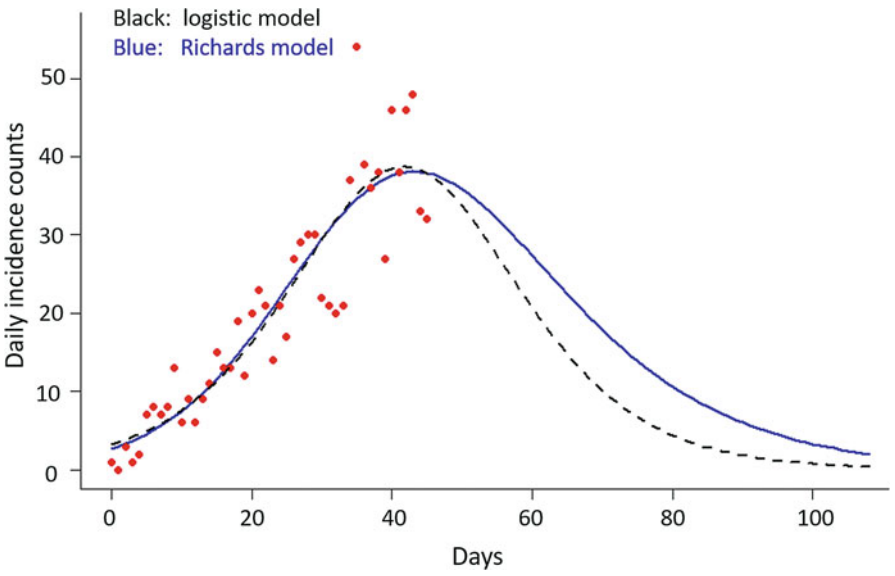


Fig. 8.13 Compare model predicted daily incidence by symptom onset based on the logistic and the Richards models, when their parameters are set at the maximum likelihood estimates. Red dots are observed data by Day 45

8.3.4 Data by Day 75

The maximum likelihood estimates for the parameters in the logistic and the Richards models are shown below:

$$\begin{array}{ll} \text{Logistic} & \text{Richards} \\ \widehat{\rho} = 0.082\ (0.0766,\ 0.0875) & \widehat{\rho} = 0.0617\ (0.0574,\ 0.0733) \\ \widehat{i_0} = 51.9\ (41.75,\ 63.75) & \widehat{i_0} = 17.36\ (7.5,\ 34.8) \\ \widehat{K} = 1805\ (1713,\ 1902) & \widehat{K} = 1852\ (1750,\ 1967) \\ H_0 : \theta = 1 & \widehat{\theta} = 0.355\ (0.152,\ 0.645) \end{array} \tag{8.29}$$

where numbers in brackets are 95% confidence limits based on the likelihood ratio statistics. The extra 30-day data, from Day 45 to Day 75, have yielded more precise estimates for all three parameters in the Richards model.

The Richards model can be also parameterized as

$$C(t) = \frac{K}{\left(1 + \theta e^{-\rho(t-\alpha)}\right)^{1/\theta}}$$

where $\alpha = \frac{1}{\rho} \log \frac{\left(\frac{K}{i_0}\right)^{\theta} - 1}{\theta}$ is the inflexion point at which $C'(t)$ is maximized. The maximum likelihood estimate for the inflexion point is

$$\widehat{\alpha} = 40.227 \text{ days } (38.36,\ 42.1).$$

The logistic model under $H_0 : \theta = 1$ gives $\widehat{\alpha} = 42.95$ days (41.8, 44.1).

Except for the estimation of K, there are significant differences in the estimation of the initial cumulative numbers $i_0 = C(0)$ and the peak time for the daily incidence α. The logistic model over predicts $i_0 = C(0)$, which also implies approximately 4 individuals developed onset on Day 0 and an earlier start of the outbreak approximately around Day-20, whereas the Richards model suggests a much smaller value for $C(0)$, approximately two individuals developed onset on Day 0 and the start of the outbreak approximately around Day-15. These are shown in Fig. 8.14, between the blue solid line and the red broken line.

By Day 75, data have shown statistical significance to reject the logistic model, corresponding to the hypothesis $H_0 : \theta = 1$. The value of the likelihood ratio statistic in (8.26) has been updated to $D = 12.902$ and $SL = \Pr(\chi^2_{(1)} \geq 12.91) = 0.0003$.

It is not appropriate to directly compare $\widehat{\rho}$ in the two models given by (8.29), because according to (8.25), the initial growth rate in the Richards model as parameterized above is the ratio $\widehat{r} = \widehat{\rho}/\widehat{\theta} = 0.1738$, in par with the m.l.e. for ρ based on the logistic model using the initial data by Day 20. This is also shown in Fig. 8.14, comparing the expected numbers of daily incidence based on the Richards model fitted to data by Day 75 with that based on the logistic model fitted to data

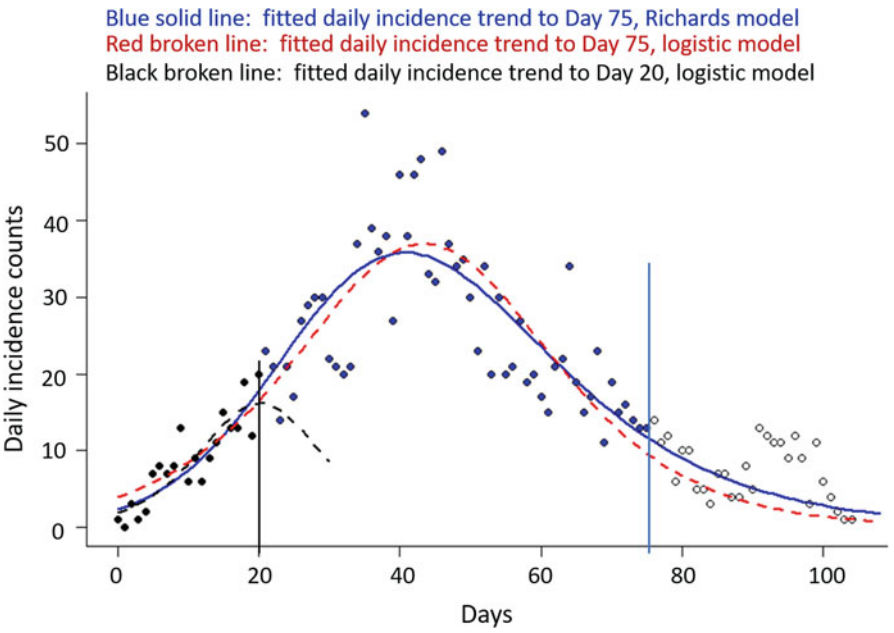


Fig. 8.14 Fitted Richards models to daily incidence counts during the first 75 days (blue dots) with comparisons with the fitted logistic model

by Day 20 (the black broken line). This is due to the asymmetry supported by the Richards model that allows for better fits on both extremes of the time-series data, whereas the logistic model is limited by its symmetric shape.

The Richards model requires a relatively large number of data points. It is not suitable as the initial model during the early phase, definitely not for data collected during the first 20 days and questionable for data collected during the first 45 days. However, as the number of data points increases, simpler models will start to misrepresent data. This will force us to adopt more complex models, not only for better prediction purposes, but also for capturing the data generating process.

We also compare the maximum likelihood estimates with the least-square estimates for the Richards model. The least square estimation yields very similar results:

$$\widetilde{\rho} = 0.0695\ (0.062,\ 0.083)$$

$$\widetilde{i_0} = 22.3\ (13,\ 31)$$

$$\widetilde{K} = 1801.9\ (1700,\ 1900)$$

$$\widetilde{\theta} = 0.52\ (0.38,\ 0.83)$$

where numbers in brackets are estimated 95% confidence limits based on 500 bootstrap samples.

Table 8.2 Summary residual measures (7.12)–(7.14) comparing the maximum likelihood and the least square estimates

	Maximum likelihood	Least square
MSE	34.08	33.6
WMSE	106.6	108.9
Anscombe	108.3	110.4

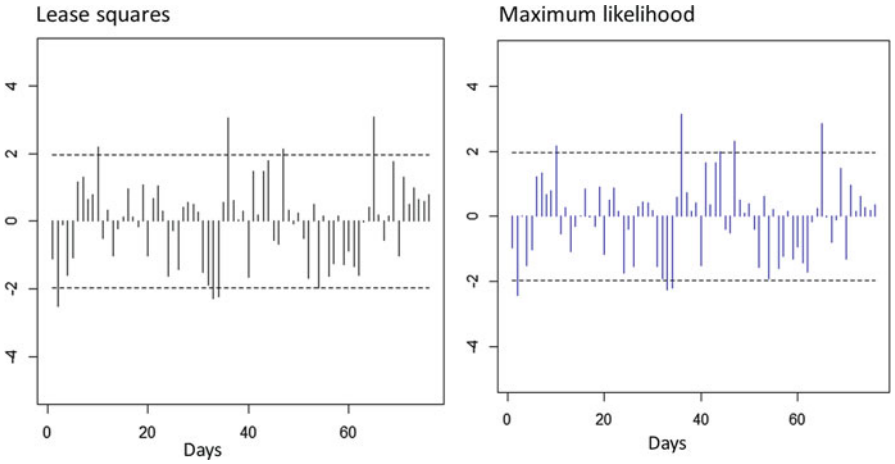


Fig. 8.15 Plots of the Ancombe residuals $r_t^{(A)} = \frac{\frac{3}{2}\left[y_t^{2/3} - f(t;\widehat{\Theta})^{2/3}\right]}{f(t;\widehat{\Theta})^{1/6}}$ for both the least square and the maximum likelihood estimates. The standard error lines ± 1.96 are based on the approximate Gaussian distribution of the Anscombe residuals

Residual analyses in Table 8.2 show that, although the least square estimates give the smaller mean square errors (MSE) by default, the maximum likelihood estimates perform slightly better based on the two other measures: the weighted mean square errors (WMSE) based on the sum of the squares of the Pearson residuals and the sum of the squares of the Anscombe residuals.

Plots of the Anscombe residuals, Fig. 8.15, reveals that both estimation methods fit data equally well. There are a few outliers in data, noticeably, on $t = 1$, 32, 33, 35, 46, 64 (days), that are either due to the inadequacy of the assumed Poisson model (i.e., over-dispersion) or the assumed Richards model. The standard errors ± 1.96 in Fig. 8.15 are based on the approximate Gaussian distribution of the Anscombe residuals.

Figure 8.16 is based on 500 bootstrap replicates of the epidemic curve based on the least square estimates, assuming the outbreak were repeated under identical conditions with Poisson error structure. The outliers on $t = 1$, 32, 33, 35, 46, 64 (days) are shown as points outside the dashed red lines indicating the 95% prediction intervals.

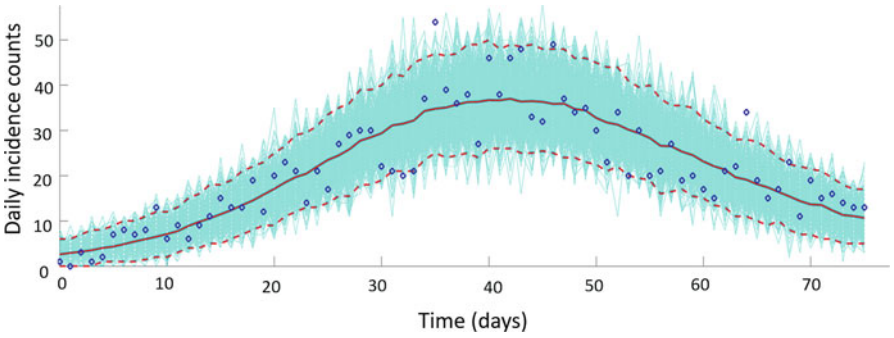


Fig. 8.16 The cyan lines correspond to 500 bootstrap replicates of the epidemic curve assuming a Poisson error structure based on the least square estimates. The solid red line corresponds to the mean, while the dashed red lines indicate the 95% prediction intervals

An additional technical note is that the least square method has been used in two different ways:

1. fitting directly to the explicit expression of the Richards model (8.24) with the four parameters (ρ, i_0, K, θ);
2. fitting the numerical solution of the differential equation (8.25) with three explicit parameters (ρ, K, θ) plus the 4th parameter $i_0 = C(0)$ as the initial condition that is also estimated.

We report back that, after careful sensitivity analyses of parameter estimates (Arriola and Hyman 2009), with respect to the initial values of the parameters in the search algorithms, and careful evaluation of the calculated values of the sum of square errors (SSE), the two fitting methods yield nearly identical numerical estimates. This comparison gives us more confidence in the least square estimates obtained directly from the numerical solution of the differential equations for other generalized logistic models in which explicit solutions do not exist.

Least Square Estimates for Other Generalized Logistic Models

We consider fitting variations of the generalized Richards model (Turner et al. 1976) with two shape parameters $\theta_1, \theta_2 > 0$:

$$\frac{d}{dt}C(t) = rC(t)^{\theta_1}\left(1 - \left[\frac{C(t)}{K}\right]^{\theta_2}\right). \tag{8.30}$$

Except for the sub-exponential model ($\theta_1 = p, \theta_2 \to \infty$), the logistic model ($\theta_1 = \theta_2 = 1$), and the Richards model ($\theta_1 = 1, \theta_2 > 0$), explicit solutions do not exist in general.

Letting $\theta_2 = 1$ and $\theta_1 = p$, we get the model defined by

$$\frac{d}{dt}C(t) = rC(t)^p\left(1 - \frac{C(t)}{K}\right), \tag{8.31}$$

which also includes a hidden parameter $i_0 = C(0)$. The least square estimation for the parameters are

$$\widetilde{r} = 0.28\ (0.21, 0.36),\ \ \widetilde{p} = 0.82\ (0.78, 0.87),$$
$$\widetilde{i_0} = 12\ (7.2, 17),\ \ \widetilde{K} = 1800\ (1700, 1900).$$

The confidence limits of the estimated parameters and the goodness-of-fit of the model are illustrated in Fig. 8.17.

Alternatively, we modify the above so that the scaling parameter p is applied to the proportion $C(t)/K$, which is more in line with the Richards model,

$$\frac{d}{dt}C(t) = rK\left[\frac{C(t)}{K}\right]^p\left(1 - \frac{C(t)}{K}\right).$$

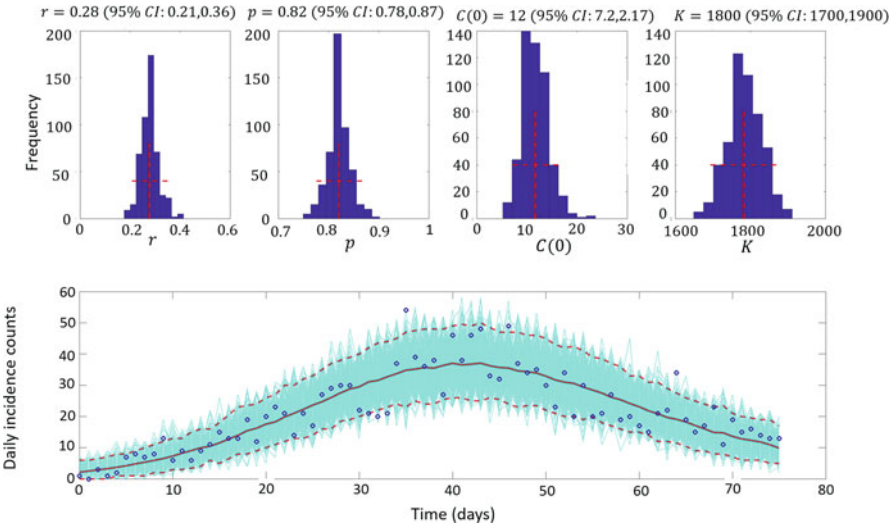


Fig. 8.17 The histograms display the empirical distributions of the parameter estimates using 500 bootstrap replicates generated assuming a Poisson error structure. The horizontal red dashed lines indicate the 95% confidence intervals of the parameter estimates. The bottom panel shows the fit of the model to the data. The blue circles are the daily incidence data. The cyan lines correspond to 500 bootstrap replicates of the epidemic curve assuming a Poisson error structure. The solid red line corresponds to mean values of the simulated sample while the red dashed lines indicate the 95% prediction intervals

Table 8.3 Summary residual measures (7.12)–(7.14) comparing three generalized logistic models with equal number of parameters

	Richards $C' = rC\left(1 - \left[\frac{C}{K}\right]^{\theta}\right)$	Gen logistic 1 $C' = rC^p\left(1 - \frac{C}{K}\right)$	Gen logistic 2 $C' = rK\left(\frac{C}{K}\right)^p\left(1 - \frac{C}{K}\right)$
MSE	33.6	34.5	33.9
WMSE	108.9	112.3	109.4
Anscombe	110.4	112.5	110.7

The least square estimates for the parameters are

$$\widetilde{r} = 0.073\ (0.065, 0.083),\ \ \widetilde{p} = 0.83\ (0.77, 0.9),$$
$$\widetilde{i_0} = 13\ (6.4, 21),\ \ \widetilde{K} = 1800\ (1700, 1900).$$

These two generalized logistic models have the same number of parameters as the Richards model with very similar estimated key parameters of epidemiologic interest, i_0 and K and almost equally good fit to data (Table 8.3). They do not offer more insight than the Richards model at least for this outbreak.

We also conducted the LS estimation with respect to the generalized Richards model (8.30) with parameter estimates.

$$\widetilde{r} = 0.22\ (0.18, 0.26),\ \widetilde{\theta_1} = 0.86\ (0.83, 0.89),\ \widetilde{\theta_2} = 0.95\ (0.73, 1.12),$$
$$\widetilde{i_0} = 18\ (12, 27),\ \widetilde{K} = 1800\ (1700, 1900),$$

where numbers in brackets are estimated 95% confidence limits based on 500 bootstrap samples. These estimates are very close to those based on the model (8.31) because $\widetilde{\theta_2} = 0.95\ (0.73, 0.12)$. Therefore, based on data by Day 75, it is not advisable to recommend more complex models with five parameters.

We stop this analysis at this point. In hindsight, the outbreak stopped on Day 104. The cumulative number was $K = 1852$, consistent with the m.l.e. using the Richards model fitted to data of the first 75 days. However, the Richards model could not forecast a small cluster or "a second wave" as shown in Fig. 8.14. The recorded first case remains as a single case on Day 0, which could not be captured by the models considered here.

8.3.5 Lessons Learned

1. When a new infectious disease emerges, exploratory analyses and simple phenomenological models are useful for forecasting epidemic trajectories.
2. When the number of observations is less, the initial model should be simple enough with as few parameters as possible. Over- parameterization results in undesirable large uncertainties in key parameters of interest. Among highly correlated parameters, it also leads to identifiability problems among parameters.

That is, two or more sets of parameter values yield the same expected values for data. For example, in Fig. 8.8 when $v = 0.034$, the first 20-day data equally admit the pairs of parameters: $i_0 = 40$ and $\rho = 0.1$; as well as $i_0 = 10$ and $\rho = 0.2$, corresponding to two distinct scenarios: high initial numbers with slow growth rate versus low initial numbers with fast growth rate.

3. Increasing the number of observations will, to some extent, improve the precision and identifiability among parameters in the simple model. However, beyond a certain limit, this gain will be diminished and off-set by biased estimates and lack-of-fit to data. This will force us to shift to a more complex model and closer to the data generating process. We have demonstrated this adaptive approach in the discussions while fitting models to data accumulated by Days 20, 45, and 75.
4. Even though the parameters in the simple growth curve models do not have any physical meaning (unlike those in transmission dynamic models), these simple models still need to be carefully selected and parameterized, so they can be useful in addressing key public health questions.
5. The curve models that we have employed here are highly nonlinear. The optimization algorithms to maximize the log-likelihood or to minimize the sum of square errors (SSE) are highly sensitive to the initial parameter estimates, which may lead to a local maximum or minimum. It is important to carefully evaluate the values of the log-likelihood or SSE upon convergence over a wide range of possible initial estimates.
6. Transformation of parameters does not affect assumptions of a model, but it may make interpretations more or less easy. Different ways of parametrizing the same growth function should be explored. One of the reasons is to have the parameters interpretable and aligned with public health questions. Another reason may be associated with the parameter searching algorithms. For example, parameter transformations to make the log-likelihood contours more like symmetric ellipsoids will generally facilitate numerical optimization.
7. Correlation among parameters: The "banana shaped" log-likelihood contours are typical signatures of correlation among parameters. The cross-sectional bivariate log-likelihood contour plots (e.g., Figs. 8.5 and 8.8) yield important information about correlations between pairs of important parameters.
8. When possible, graphical presentation of the likelihood surface is worthwhile, either as a 3-D function or cross-sectional log-likelihood contours. These will provide more reliable precision intervals than marginal confidence intervals for each parameter, reveal correlation among parameters, and provide better ways to communicate uncertainty. However, these are very time-consuming.
9. Approximate confidence intervals based on the likelihood ratio statistics are in agreement with the contours of the likelihood surface, as opposed to those based on standard errors which rely on a quadratic approximation. A very common feature is that these confidence intervals are highly asymmetric around their point estimates. When extremely asymmetric, the emphasis should be on the plausibility range towards the wider side of the interval rather than the point estimate. This is very important in communicating uncertainty. We demonstrated this while analyzing the Zika data during the first 20 days, with a very wide

plausibility region in favor of the slow growth pattern. This was confirmed when more data were collected by Day 45.

8.4 The Effective Reproduction Number, R_t, with Quantified Uncertainty

The basic reproduction number, commonly denoted by R_0, quantifies transmission potential in a fully susceptible population during the early epidemic take off (Anderson and May 1982). According to the classical theory of epidemics, largely based on compartmental modeling (e.g., Anderson and May 1991; Diekmann and Heesterbeek 2000; van den Driessche 2017; van den Driessche and Watmough 2002; Diekmann et al. 2010), R_0 is expected to remain invariant during the early phase of an epidemic that grows exponentially and as long as susceptible depletion remains negligible (Diekmann and Heesterbeek 2000).

In Chap. 4, Eq. (4.47): $L[g](r) = \int_0^\infty e^{-rx} g(x)dx = R_0^{-1}$ can be re-written in the renewal form

$$i(t) = R_0 \int_0^\infty g(x) i(t-x) dx \tag{8.32}$$

where $i(t) \approx e^{rt}$ is the instantaneous density of infected individuals at the very beginning of the outbreak approximated by exponential growth, and $g(x)$ is the probability density function of the (intrinsic) generation time T_G associated with the Lotka equations in Sect. 4.3.3, formally defined and further discussed as (7.18) in Chap. 7. This approximation is suitable when t is extremely small, near the disease-free equilibrium. Wallinga and Lipsitch (2007) suggested ways of estimating the basic reproduction number based on the initial growth rate r through fitting the exponential growth to early outbreak data, provided that the generation time distribution $g(x)$ is fully specified so that

$$\widehat{R}_0 = L[g](\widehat{r})^{-1}$$

where $\widehat{r}$ is the fitted initial growth rate to data.

In contrast, the effective reproduction number R_t captures changes in transmission potential over time when the system starts to move away from the equilibrium condition (Chowell et al. 2016; Nishiura and Chowell 2009). The effective reproduction number R_t is given by

$$R_t = \frac{S(t)}{S(0)} R_0$$

where $S(t)$ is the expected number of susceptible individuals in the population at time t. It is understood as the expected number of secondary infections transmitted by a typical infectious individual at calendar time t.

Nishiura and Chowell (2009) generalized the above renewal type equation through analysis of an infection-age structure model, The term "infection-age" refers to the time elapsed since infection. Define $A(t, x)$ as the rate at which an infectious individual at calendar time t and infection age τ produces secondary infections so that

$$i(t) = \int_0^\infty A(t, x) i(t - x) dx,$$

under the assumption that the relative infectiousness to infection-age is independent of calendar time (Fraser 2007), Nishiura and Chowell (2009) argue that $A(t, x)$ can be decomposed as $A(t, x) = R_t g(x)$, where $g(x)$ is the same generating time distribution as in (8.32). This leads to

$$i(t) = R_t \int_0^\infty g(x) i(t - x) dx.$$

To fit to data observed in discrete (grouped) time units over a finite period $t = 0, \ldots, T$, the following approximation

$$i(t) = R_t \sum_{x=0}^{T} g(x) i(t - x), \ t = 0, \ldots,$$

has been considered (The World Health Organization Emergency Response Team 2014, Supplementary Appendix 1; Chowell et al. 2016), where $i(t)$ is the expected number for the incidence data during the time unit t, such as daily incidence Y_t so that $E[Y_t] = i(t)$.

Assuming that the incidence up to time $t - 1$ is Poisson distributed, the daily incidence Y_t is

$$Y_t \sim Poisson \left(R_t \sum_{x=0}^{T} g(x) i(t - x) \right),$$

Then, given the incidence data as a longitudinal series denoted by $\underline{y} = (y_1, y_2, \ldots, y_T)$, the maximum likelihood estimate for R_t is

$$\widehat{R}_t = \frac{i(t; \widehat{\Theta})}{\sum_{x=0}^{T} g(x) i(t - x; \widehat{\Theta})}, \ t = 0, \ldots, T, \tag{8.33}$$

where $i(t; \Theta)$ is a suitable phenomenological model to describe the data generating process $\underline{y} = (y_1, y_2, \ldots, y_T)$, and $i(t; \widehat{\Theta})$ is the fitted incidence, provided that the generation time distribution $g(x)$ is fully specified.

Although uncertainties in parameter estimates $\widehat{\Theta}$ can be derived using the likelihood ratio statistics, establishing variance estimation for $\widehat{R}_t$ is more complex. Therefore, computer based re-sampling methods, such as bootstrapping, are preferred.

Next, we assume the sub-exponential model $i'(t) = r\,[i(t)]^p$ starting with a single individual, with the explicit form

$$i(t) = (1 + r(1-p)t)^{\frac{1}{1-p}}.$$

This model can reproduce a range of growth dynamics from constant incidence ($p = 0$) to exponential growth ($p = 1$) (Viboud et al. 2016).

We denote $(\widehat{r}^{(i)}, \widehat{p}^{(i)})$ as the estimated parameters based on the ith bootstrap sample in a re-sampling regime. Then (8.33) gives

$$\widehat{R}_t^{(i)} = \frac{i(t; \widehat{r}^{(i)}, \widehat{p}^{(i)})}{\sum_{x=0}^{T} g(x) i(t-x; \widehat{r}^{(i)}, \widehat{p}^{(i)})}, \; t = 0, \ldots, T.$$

Based on the maximum likelihood estimate (8.33) from the incidence data, a large number of bootstrap realizations create a virtual experiment with repetitions of the outbreak under identical conditions, which produce the average of $\widehat{R}_t^{(i)}$ as well as the plausible ranges for uncertainty.

8.4.1 *Example Based on the 2016 Epidemic of Yellow Fever in Two Areas of Angola: Luanda and Huambo*

For illustration, we estimated the effective reproduction number during the early phase of a yellow fever epidemic. The epidemic spread between December 2015 and August 2016 in Angola, mostly affecting the provinces of Luanda (the capital) and Huambo. Numbers of confirmed and probable reported cases are grouped into discrete time intervals on a weekly basis and assembled by the World Health Organization (The World Health Organization 2016). The corresponding time series data are available online as EXTRA MATERIALS.

For the goal of estimating R_t, we assumed a gamma distribution for the generation interval of yellow fever with a mean of 15 days (2.143 weeks) and variance of 36 days (5.143 weeks). We fitted the generalized growth model (4.61) to the growth phase of the epidemics.

The yellow fever epidemic in Luanda followed an initial growth phase consistent with exponential growth dynamics (Fig. 8.18) with the scaling of growth parameter p very close to 1.0 and our most recent estimate of the effective reproduction number at 3.3 (95%CI: 2.6, 3.6). The corresponding curves of the effective reproduction number are shown in the bottom panel of Fig. 8.18. In contrast, for Huambo, the effective reproduction number was most recently estimated at 1.2, 95% CI: 1.1, 1.4) with a relatively low scaling of growth parameter (0.36, 95% CI: 0.17, 0.55) as shown in Fig. 8.19. The curves of the effective reproduction number are shown in the bottom panel of Fig. 8.19.

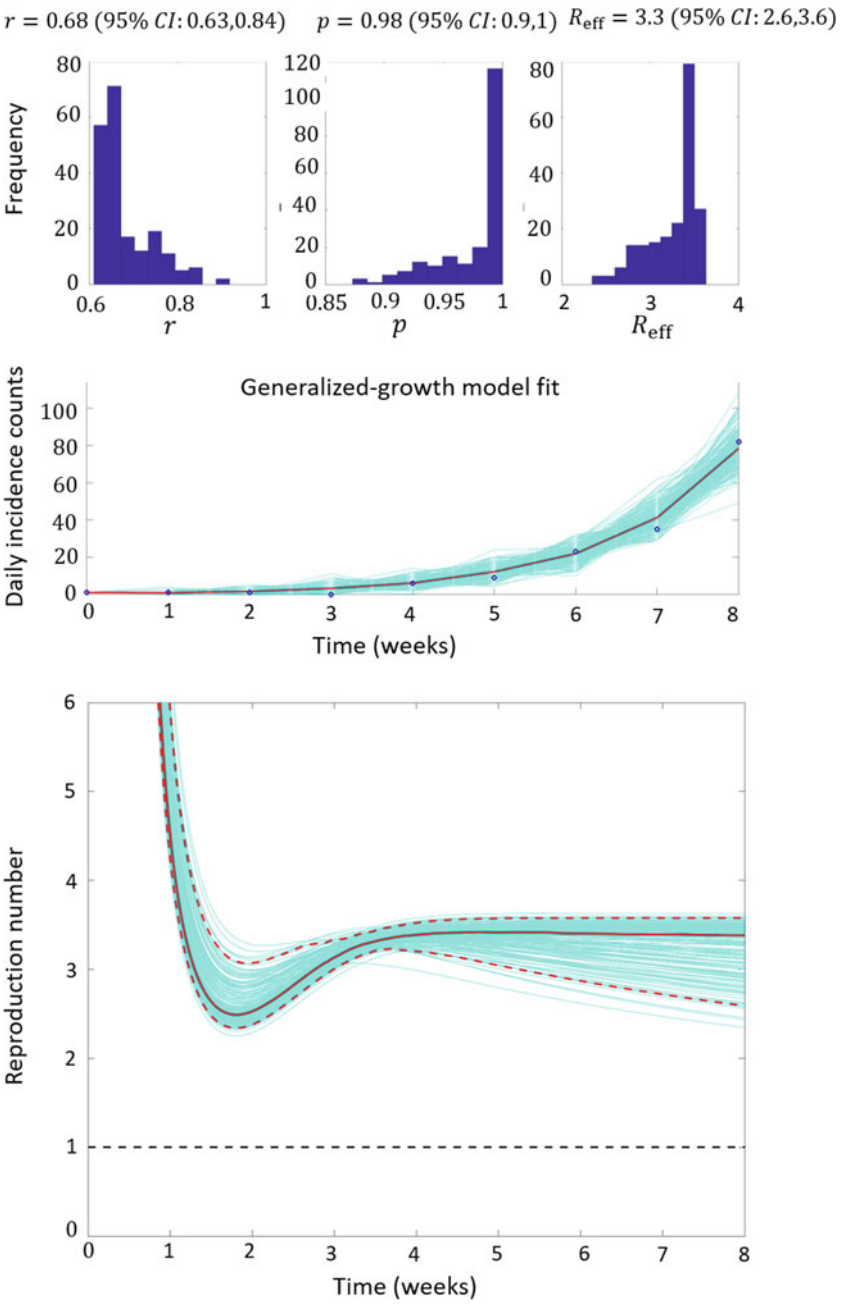


Fig. 8.18 Top panels display the empirical distributions of the growth rate, the scaling parameter, and the effective reproduction number based on fitting (4.61) to the yellow fever epidemic in Luanda, Angola. The middle panel shows the fit to the epidemic growth phase. Circles correspond to the data while the solid red line corresponds to the best fit obtained using the generalized-growth model. The blue lines correspond to the uncertainty around the model fit. The bottom panel is the weekly effective reproduction number estimated during the epidemic growth phase assuming a gamma distribution for the generation interval of yellow fever with a mean of 15 days and variance of 36

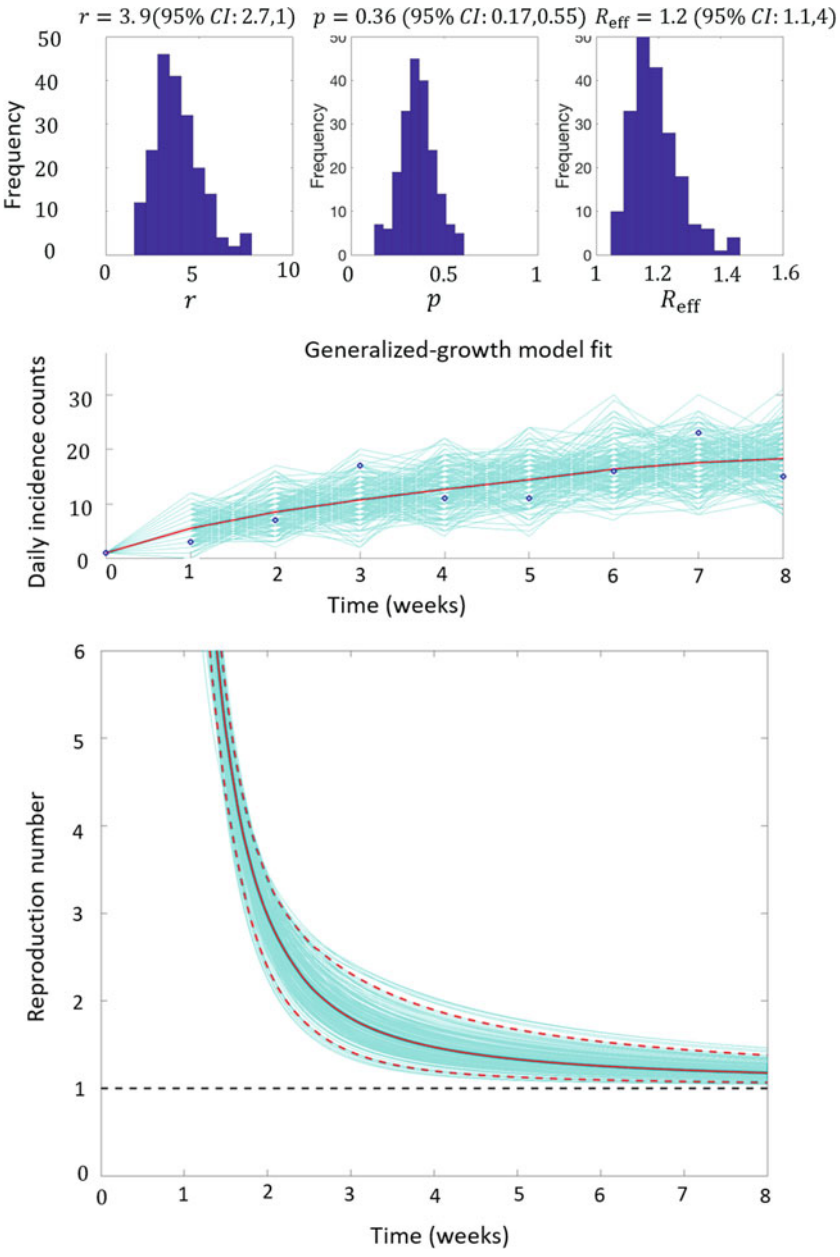


Fig. 8.19 Top panels display the empirical distributions of the growth rate, the scaling parameter, and the effective reproduction number based on fitting (4.61) to the yellow fever epidemic in Huambo, Angola. The middle panel shows the fit to the epidemic growth phase. Circles correspond to the data while the solid red line corresponds to the best fit obtained using the generalized-growth model. The blue lines correspond to the uncertainty around the model fit. The bottom panel is the weekly effective reproduction number estimated during the epidemic growth phase assuming a gamma distribution for the generation interval of yellow fever with a mean of 15 days and variance of 36

In conclusion, in this final section we have demonstrated how phenomenological models, such as the generalized-growth model, along with parameter uncertainty derived from the parametric bootstrap least-square fitting approach can be exploited to characterize transmission dynamics and their uncertainty such as the effective reproduction number through the renewal equation. Indeed, with additional data of the outbreak trajectory, we could have considered other phenomenological models such as logistic-type models and/or more elaborate error structures of the random component to account for observation correlations or data overdispersion.

8.5 Problems and Supplements

8.1 In this exercise, the reader will use phenomenological models (e.g., GGM and Richards models) to analyze the trajectory of the 2001 foot-and-mouth disease epidemic in the UK using the daily curve of the number of new infected premises. The daily number of new, real-time notifications of infected premises during the 2001 foot-and-mouth disease epidemic in the UK was obtained from the Department of Environmental and Rural Affairs (DEFRA) and is available online as an EXTRA MATERIAL. Answer the following questions:

(a) Using the GGM, what are your estimates of the growth rate (r) and the deceleration of growth parameter (p) using the first 20 epidemic days? Use maximum-likelihood estimation with a Poisson error structure.
(b) Based on your analysis in (a), assess the Anscombe residuals and compute the value of the Anscombe performance metric
(c) Based on your analysis in (a), what can you conclude from your estimate of the deceleration of growth parameter (p)?
(d) How do your parameter estimates in (a) compare with those obtained using the least-square fitting approach with a parametric bootstrap Poisson error structure?
(e) Based on your analysis in (d), assess the 95% prediction intervals around the model fit.
(f) Calibrate the Richards model to the first 10, 20, 30, or 40 epidemic days. Discuss parameter identifiability and lack of information when the model is fitted to an increasing number of observations.
(g) Using 30 and 40 epidemic days, what are the point estimates of the epidemic size? and how do these estimates compare with the actual epidemic size? What are the corresponding estimates of the epidemic peak and duration?

8.2 In this exercise, you will generate estimates of transmission potential of the 1918 influenza pandemic in San Francisco, California. The daily number of reported cases is available online as an EXTRA MATERIAL. Answer the following questions:

(a) Using the simple SEIR model without demographic factors and assuming a mean latent period of 2 days, a mean infectious period of 4 days and a population size of 550,000 provide the mean estimate and 95% confidence intervals of the basic reproduction number R_0 using 16, 18, and 20 days of the initial growth phase. For parameter estimation you can use the least square fitting approach with the Poisson parametric bootstrap which is described in Chap. 7 and illustrated with examples in Chap. 8. Note that you only need to estimate the transmission rate using your favorite technical computing language while keeping the initial number of infectious individuals $I(0)$ fixed according to the first data point. Are the R_0 estimates relatively stable during the study period?
(b) What are the corresponding values of the RMSE from your analysis in (a)?
(c) Assess the residuals and the 95% prediction intervals around the model fit and discuss your observations.
(d) Using the GGM, what are your estimates of the growth rate (r) and the deceleration of growth parameter (p) when the model is fitted to the study periods in (a)?
(e) What can you conclude from your estimate of the deceleration of growth parameter (p)? Is this parameter stable as you use 16, 18, and 20 epidemics days of data?
(f) Using your calibrated GGM based on your analysis in (d) and the approach described in Sect. 8.4, estimate the effective reproduction number R_t during the first 20 epidemic days. Compare your estimates of R_0 derived in (a) with your estimates of R_t.

8.3 Using the generalized-growth model, characterize the early ascending phase of the HIV/AIDS epidemic using monthly or annual case incidence data from any area, region, or country of the world. Answer the following questions:

(a) Using the GGM, what are your estimates of the growth rate (r) and the deceleration of growth parameter (p) when the model is fitted to the first 10 years of the epidemic?
(b) What can you conclude from your estimate of the deceleration of growth parameter (p)?
(c) Document in detail the source of your data (e.g., publication reference, website, etc.).

Chapter 9
Mechanistic Models with Spatial Structures and Reactive Behavior Change

As we have emphasized in Chaps. 4 and 5, simple homogeneous models of transmission or growth dynamics often yield an early exponential epidemic growth phase even when the population is stratified into different groups (e.g., age, gender, regions). However, recent work has highlighted the presence of early sub-exponential growth patterns in case incidence from empirical outbreak data (Chowell et al. 2016; Viboud et al. 2016). This suggests that integrating detailed and often unobserved heterogeneity into simple mechanistic models (Yan 2018) could open the door to a new and exciting research area to better understand the role of heterogeneity on key transmission parameters, epidemic size, stochastic extinction, the effects of interventions, and disease forecasts.

The diversity of infectious disease dynamics can be shaped by multiple and often unobservable factors including the characteristics of the contact network structure, individual-level heterogeneity in infection risk, and behavior changes (Chowell et al. 2016). For instance, in the simplest setting when disease spreads assuming homogeneous mixing, it is well-known that the incidence curve grows exponentially in the absence of susceptible depletion, behavior changes, and interventions (Diekmann and Heesterbeek 2000). It is worth noting that exponential growth can only unfold in the presence of a constant growth rate (as highlighted in Chaps. 4 and 5). By contrast, an early transmission phase characterized by slower than exponential growth (sub-exponential) can result from spatial constraints in contact-network structures over which disease spreads or the early onset of behavior changes or control interventions. Therefore, predictions of final epidemic size based on models that assume early exponential growth will tend to overestimate epidemic size whenever the early dynamics of disease transmission are governed by mechanisms that induce slower transmission patterns. In turn, public health authorities could get better estimates of the effectiveness of control interventions.

P. Yan, G. Chowell, *Quantitative Methods for Investigating Infectious Disease Outbreaks*, Texts in Applied Mathematics 70,
https://doi.org/10.1007/978-3-030-21923-9_9

We devote this chapter to review mechanistic transmission models that incorporate spatial details or realistic population mixing structures, including metapopulation models, individual-based network models as well as simple SIR-type models that incorporate the effects of reactive behavior changes or inhomogeneous mixing (Fenichel et al. 2011). We argue that designing mechanistic models and statistical approaches that capture a diversity of disease dynamics could lead to enhanced model fit, improved estimates of key transmission parameters, and more realistic epidemic forecasts (Chowell et al. 2016).

Structured population models can be traced back to the 1940s (Wilson and Worcester 1945) and 1950s (Rushton and Mautner 1955). The number of infectious disease spatial modeling studies has been increasing during the last couple of decades with a research production of less than five articles per year in 1997 to more than 120 articles per (Chowell and Rothenberg 2018). Models of the spread of infectious diseases can be formulated at the subpopulation (metapopulation) and individual levels. In metapopulation models the population is divided in a set of interacting population groups according to spatial or demographic characteristics. On the other hand, individual-level network models rely on individual-level contact matrix to define interactions which could be static or dynamic.

9.1 Metapopulation Spatial Models

Metapopulation formulations offer a popular mathematical framework to study the spatial spread of human infectious diseases (Arino et al. 2005; Chowell et al. 2006; Hethcote 2000; Hadeler and Castillo-Chavez 1995; Jacquez 1996; Keeling and Rohani 2008; Sattenspiel 2009). Metapopulation models can be represented as networks with the subpopulations represented by nodes and the interactions among groups represented as the weighted network links (Riley 2007). The subpopulations being modeled using a metapopulation approach are assumed to be discrete groups that are connected in some fashion. Usually subpopulations are considered to be well mixed and homogeneous, while the interaction between groups may be either explicit or implicit, leading to the development of two general classes of spatial metapopulation models: (a) cross-coupled models and (b) mobility models (Sattenspiel 2009). Cross-coupled models simplify the analysis by modeling the strength of the interactions (i.e., coupling) between groups. In mobility models, the modeler mechanistically incorporates the movement of individuals between groups.

Cross-coupled metapopulation models (early examples include Wilson and Worcester 1945; Rushton and Mautner 1955; Murray and Cliff 1977) only model the influence of one group over the others via a contact matrix that represents the strength or sum total of those contacts only. The elements of this matrix capture the strength of the interactions between any two subpopulations, which modulates the transmission risk. A simple SIR deterministic cross-coupled epidemic model can be written as follows:

$$\begin{aligned}
\frac{dS_i}{dt} &= \mu N_i - \mu S_i - S_i \sum_{j=1}^{n} \frac{\phi_{ij} I_j}{N_i} \\
\frac{dI_i}{dt} &= S_i \sum_{j=1}^{n} \frac{\phi_{ij} I_j}{N_i} - (\mu + \gamma) I_i \\
\frac{dR_i}{dt} &= \gamma I_i - \mu R_i,
\end{aligned} \tag{9.1}$$

where S_i, I_i, and R_i are the numbers of susceptible, infectious, and recovered individuals, respectively, N_i is the total population size in subpopulation i, γ is the recovery rate, and μ is the rate of birth (and death) under the assumption of a non-growing population (total births = total deaths). ϕ_{ij} is the rate of effective contact between subpopulation i and subpopulation j; the set of ϕ_{ij} characterizes the WAIFW matrix. The ϕ_{ij} implicitly include both the rate of contact and the probability of transmission.

For illustration, Fig. 9.1 displays the impact of increasing transmission rates of the 4-nearest neighbors on local epidemic simulations using a cross-coupled metapopulation model where 100 local populations each of size 100,000 are spatially arranged in a 10×10 square lattice structure. Perhaps not surprisingly, one can observe how the early local epidemic growth dynamics during the first few generation intervals corresponds well to the epidemic growth derived from a simple SEIR transmission model in a homogenously mixed population. Temporal snapshots of the spatial distribution of disease prevalence using contour plots are shown in Fig. 9.2.

The gravity contact matrix assumes that the rate of contact between two groups is directly proportional to their population size and inversely proportional to their geographic distance (Xia et al. 2004; Viboud et al. 2006; Weinberger et al. 2012). A generalized gravity model takes the form

$$m_{jk} = \frac{N_j^a N_k^b}{d_{jk}^c},$$

where m_{jk} represents the contact between groups j and k, N_j and N_k are the population sizes of the groups, d_{jk} is the distance between the two groups, and a, b, and c are parameters typically estimated from data relating the interactions between the groups.

Recently, Simini et al. (2012) proposed a radiation mobility model. Their model is intended to represent commuting behavior, and they assume that the destinations are determined only by job selection, which is a decision that depends on the size of the location of a specific job opportunity as well as the benefits (e.g., income, working hours, conditions, and other characteristics) of the potential opportunity. Individuals choose the closest job to their home region that has higher benefits than those within the home region. The assigned work locations of all members of a

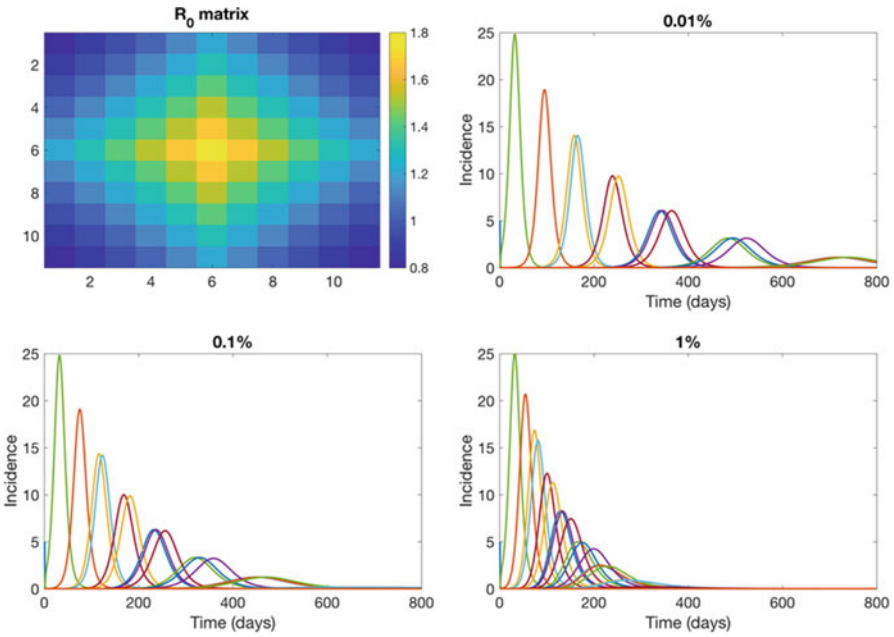


Fig. 9.1 Local epidemics generated using a cross-coupled metapopulation model where 100 local populations are spatially arranged in a 10 × 10 square lattice with periodic boundary conditions. The local dynamics across all patches follow a simple SEIR (susceptible–exposed–infectious–removed) transmission model with a mean latent period of 2 days, a mean infectious period of 3 days, a local basic reproduction number, R0 at 1.5, and a local population size in each patch of 100,000 individuals. A constant transmission between the 4-nearest neighbors is modeled as a fraction of the local transmission rate, which takes values of (a) 0.1%, (b) 0.5%, (c) 1%, and (d) 5%. For reference, the red dotted line corresponds to the curve of total incidence, while the dashed black line corresponds to the solution of the homogenous-mixing SEIR model considering the total homogenously mixed population in a single patch

region determine the daily commuter fluxes. The average flux, T_{ij}, from region i to region j at a distance r_{ij} apart is given by

$$\langle T_{ij} \rangle = T_i \frac{N_i N_j}{(N_i + s_{ij})(N_i + N_j + s_{ij})},$$

where N_i and N_j are the population sizes of regions i and j, respectively, and s_{ij} is the total population in a circle of radius r_{ij} centered on region i but excluding both the source and destination populations. $T_i = \sum_{j \neq i} T_{ij}$ is the total number of commuters who begin their commute in region i. Population distribution is the only required input for this model.

Mobility metapopulation models mechanistically aim to describe the actual movement of individuals across subpopulations (e.g., Arino et al. 2007; Belik et al. 2011; Kenah et al. 2011; Vincenot and Moriya 2011; Xiao et al. 2011; Tizzoni et al.

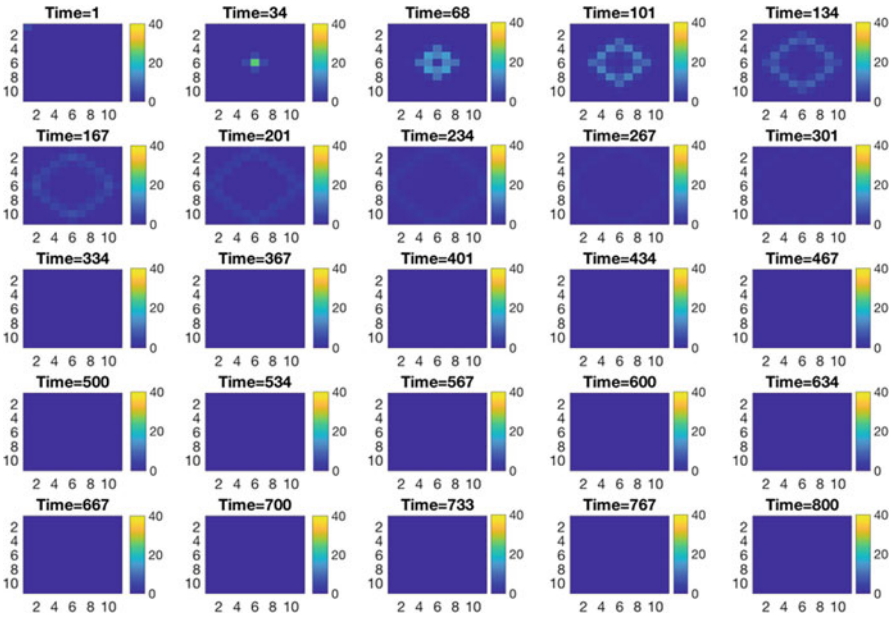


Fig. 9.2 Spatial spread of SEIR metapopulation model in a 10×10 lattice

2012; Appoloni et al. 2013; Apolloni et al. 2014; Marguta and Parisi 2015). Hence, transmission of the pathogen occurs within subpopulations considering the local and visitor populations. This process can also be modeled by considering first the rates at which individuals leave groups to visit other locations and then the possible destinations and average durations of those trips (Sattenspiel and Dietz 1995). An example of a deterministic SIR mobility metapopulation model is the following set of equations:

$$\begin{aligned}
\frac{dS_i}{dt} &= \mu N_i - \frac{\beta_i S_i I_i}{N_i} - \mu S_i + \sum_{j=1}^{n} \theta_{ij} S_j \\
\frac{dI_i}{dt} &= \frac{\beta_i S_i I_i}{N_i} - (\mu + \gamma) I_i + \sum_{j=1}^{n} \theta_{ij} I_j \\
\frac{dR_i}{dt} &= \gamma I_i - \mu R_i + \sum_{j=1}^{n} \theta_{ij} R_j,
\end{aligned} \tag{9.2}$$

where S_i, I_i, and R_i are the numbers of susceptible, infectious, and recovered individuals, respectively, and N_i is the total population size of subpopulation i, μ is the rate of birth (and death) where total births = total deaths, β_i is the transmission parameter in subpopulation i, and θ_{ij} is the rate of movement to subpopulation i

from subpopulation j. Moreover, rates of movement are assumed to be the same for all disease states in this simple model.

9.2 Individual-Based Network Models

Individual-level network models are being increasingly used to study infectious disease dynamics where contacts (links) can be either static or dynamic (reviewed in Halloran et al. 2002; Keeling and Eames 2005; Bansal et al. 2007; Capaldi et al. 2012; Danon et al. 2011). A contact-network model explicitly represents host interactions that dictate disease transmission. A node in a contact network represents an individual host, and an edge between two nodes represents an interaction through which infection is possible. Network-based models are then useful for investigating the impact of individual-level characteristics and their disease-relevant interactions on the transmission dynamics observed at the population level. A number of network models have been proposed in the literature ranging from random, small-world, to scale-free networks (Watts and Strogatz 1998; Barabási and Albert 1999; Albert and Barabasi 2002). One of the most popular and parsimonious contact-network models is the "small-world" network model as it allows for tuning the average degree of the nodes, the average connectivity (path length), and the clustering that quantifies the extent to which contacts of a node are also contacts of each other (Watts and Strogatz 1998).

Figure 9.3 shows two small-world networks with two different rewiring probabilities. While the original Watts–Strogatz model starts from a ring network structure, the idea can be extended to other regular networks. For instance, Fig. 9.4 displays examples of small world networks based on two dimensional lattices where

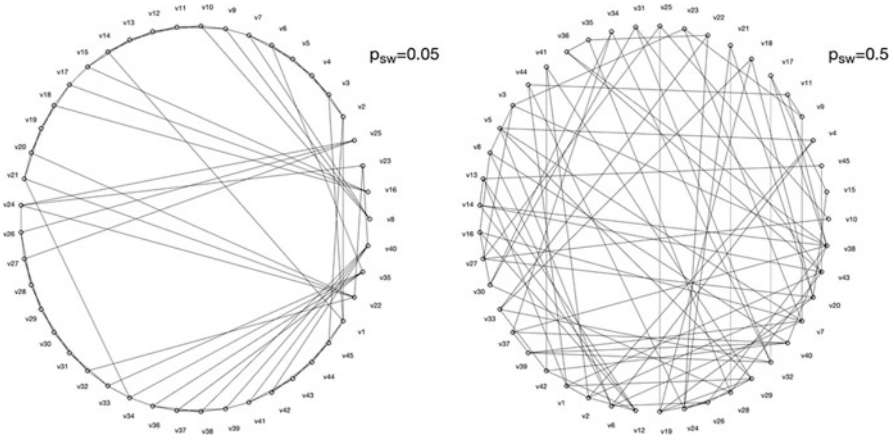


Fig. 9.3 Small-world networks with two rewiring probabilities

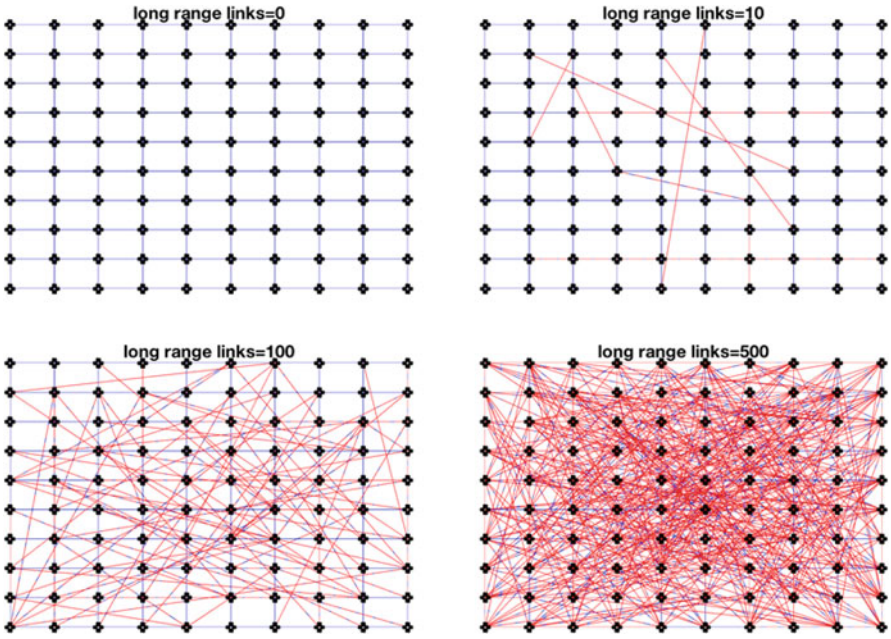


Fig. 9.4 Schematic representation of 2D square lattices where each node is connected to its 4-nearest neighbors with periodic boundary conditions and with the addition of a few random long-range links

each node is connected to its 4 nearest neighbors, and the small-world feature is incorporated by adding a fixed number of random links.

For illustration, we simulated SIR (susceptible–infectious–removed) dynamics on small-world networks using networks of size $N = 90{,}000$ and node connectivity to the 4 nearest neighbors, and we increased the edge rewiring probability parameter (p_{sw}) from 0.001 to 0.01 of the small-world network model of Watts and Strogatz (1998). For each value of p_{sw}, we analyzed the early epidemic growth profile comprising 35 days of disease transmission from 200 stochastic realizations. The transmission rate per contact per unit of time was set at 2 and the infectious period was assumed to be exponentially distributed with mean $1/\gamma$ which is set at 3 days. Each simulation started with one infectious individual selected at random from the network. For reference, the baseline SIR transmission dynamics on the regular network with node connectivity to the 4 nearest neighbors and without long-range links correspond to a wave of steady case incidence at about 4 cases per day.

9.2.1 An Individual-Level Network Model with Household-Community Structure

One of the putative mechanisms leading to early polynomial growth dynamics of transmission is clustering (Szendroi and Csányi 2004; Chowell et al. 2015; Merler et al. 2015; Viboud et al. 2016; Chowell et al. 2017), a network property that quantifies the extent to which the contacts of one individual are also contacts of each other (Watts and Strogatz 1998). Social contact networks are particularly useful to explore the impact of clustering and play an important role in the dissemination of infectious diseases at the community level.

Several authors have put forward relatively simple mathematical models that incorporate household and other social structures such as schools and workplaces (Longini and Koopman 1982; Longini et al. 2007; Ball et al. 2009, 2015; Fraser 2007; Goldstein et al. 2009; Pellis et al. 2009, 2012, 2015; Blythe and Castillo-Chavez 1989). For instance, a network-based transmission model with household structure embedded in a structure of overlapping communities has been previously applied to study the transmission dynamics of Ebola (Kiskowski 2014; Kiskowski and Chowell 2015). In this model, individuals are organized within households of size H (each household contains H individuals) and households are organized within communities of size C households (each community contains $\times$ H individuals) (see Fig. 9.5). Network connectivity is identical for every individual. The transmission potential is characterized by the household reproduction number and the community reproduction number. For a given household size H, prior studies have investigated

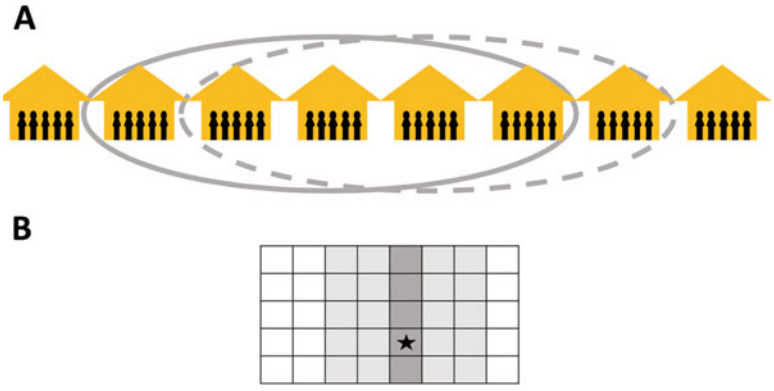


Fig. 9.5 Schematic representation of the household-community mixing structure with overlapping communities. In this model, individuals are organized within households of size H (each household contains H individuals) and households are organized within communities of size C households (each community contains $\times H$ individuals) (panel (**a**)). Network connectivity is identical for every individual. The transmission potential is characterized by the household reproduction number and the community reproduction number. The matrix-level representation of the model is shown in panel (**b**)

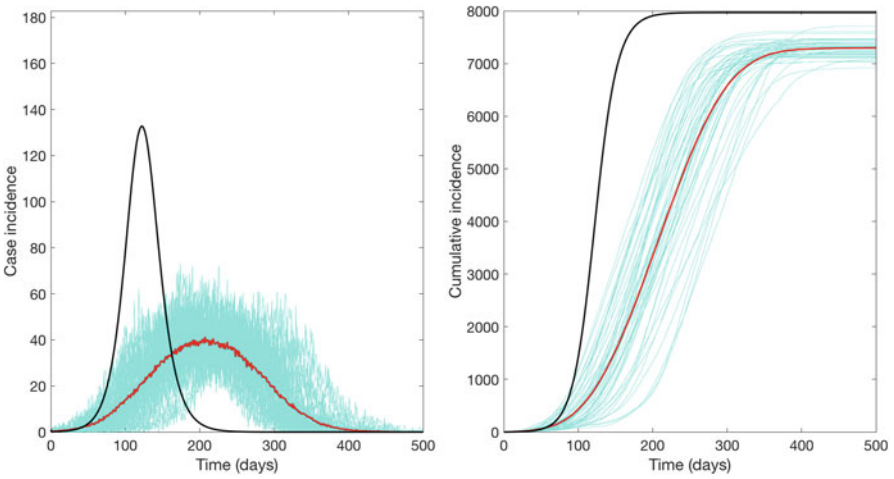


Fig. 9.6 Stochastic SEIR simulations (cyan lines) of the household-community model with $H = 6$, $C = 165$, and $R_{0H} = 1.5$ and $R_{0H} = 0.5$. The mean of the ensemble of stochastic realizations is the red solid line. The corresponding solution of the deterministic SEIR model under homogenous mixing with $R_0 = 2$ corresponds to the solid black line. Baseline epidemiological parameters were set according to the epidemiology of Ebola (i.e., incubation period of 5 days (Eichner et al. 2011; The World Health Organization Emergency Response Team 2014) and infectious period of 7 days (Chowell et al. 2004; The World Health Organization Emergency Response Team 2014)). The population size at 10,000

the impact of varying the community size parameter C on the early transmission phase. As the community size increases, the scaling of epidemic growth approaches the exponential growth regime (Kiskowski and Chowell 2015; Chowell et al. 2016). Figure 9.6 contrasts simulations derived from the household-community model with the deterministic solution of the SEIR model under homogenous mixing with the same R_0. In particular, outbreaks not only spread more slowly in the spatial household-community model, but the size of those epidemics is smaller compared to the homogenous mixing SEIR model using baseline epidemiological parameters (mean latent and infectious periods) in line with the epidemiology of Ebola.

9.3 Capture Dynamic Reactive Behavior Changes Through a Generalized-Growth SEIR Model

The generalized-growth SEIR model (GG-SEIR) is a novel modeling framework (Chowell et al. 2016) that builds on the well-known SEIR (susceptible–exposed–infectious–recovered) transmission model (Anderson and May 1991) by incorporating flexible early epidemic growth profiles, e.g., sub-exponential and exponential growth dynamics. This is achieved by allowing a dynamic nature of the effective

reproduction number R_t in the context of early sub-exponential (e.g., polynomial) growth dynamics.

The standard deterministic SEIR epidemic model represents the simplest and most popular mechanistic compartmental model for describing the spread of an infectious agent in a well-mixed population. As explained before, the force of infection per unit of time is simply given by the product of three quantities: a constant transmission rate (β), the number of susceptible individuals in the population ($S(t)$), and the probability that a susceptible individual encounters an infectious individual ($I(t)/m$). Moreover, infected individuals experience a mean latent and a mean infectious period given by α^{-1} and γ^{-1}, respectively. The model is based on a system of ordinary differential equations that keep track of the temporal progression in the number of susceptible, exposed, infectious, and removed individuals (see Eq. (5.53)).

In a completely susceptible population, e.g., $S(0) = m$, the average number of secondary cases generated per primary case, $R_0 = \beta/\gamma$. However, as the number of susceptible individuals in the population declines due to a growing number of infections, the effective reproduction number over time, R_t, is given by the product of and the proportion of susceptible individuals in the population:

$$R_t = \frac{\beta}{\gamma}\frac{S(t)}{m}. \tag{9.3}$$

During the first few generations of disease transmission when $S(t) \approx m$, in the absence of control interventions or reactive population behavior changes, the standard SEIR model supports a reproduction number that is essentially invariant, i.e., $R_t \approx R_0$. By contrast, in the context of epidemics characterized by early sub-exponential growth dynamics, we have shown that the reproduction number is a dynamic quantity that declines over disease generations towards 1.0 (Chowell et al. 2016). Here we introduce the generalized-growth modeling framework based on the well-known SEIR model (GGM-SEIR) that incorporates the possibility of early sub-exponential growth dynamics by explicitly modeling the dynamic behavior of the effective reproduction number via a time-dependent transmission rate $\beta(t)$ such that the force of infection becomes: $\beta(t)S(t)I(t)/m$. Specifically, we consider a transmission rate function $\beta(t)$ of the form:

$$\beta(t) = \beta_0\left[(1-\phi)\, f(t;\Theta) + \phi\right],$$

where $f(t;\Theta)$ is a function that declines over time from 1 towards zero so that the transmission rate $\beta(t)$ declines from an initial value β_0 towards $\phi\beta_0$. The quantity $(1-\phi)$ models the proportionate reduction in β_0 that is needed to reach an effective stationary reproduction number at 1.0, in line with early sub-exponential growth dynamics (Chowell et al. 2016). For the standard SEIR model, ϕ can be simply estimated as γ/β_0 since $R_0 = \beta/\gamma$ during the early growth phase when $S(t) \approx m$.

Here we employ an exponential decline function for the transmission rate, which is given by

$$\beta(t) = \beta_0 \left[(1-\phi)\, e^{-qt} + \phi\right],\ 0 < q \leq 1 \text{ and } \phi > 1.$$

Alternatively, harmonic and hyperbolic functions could be used to model the decline in the transmission rate as follows:

$$\beta(t) = \beta_0 \left[(1-\phi)\,(1+qvt)^{-1} + \phi\right],$$

$$\beta(t) = \beta_0 \left[(1-\phi)\,(1+qvt)^{-1/v} + \phi\right].$$

This modeling framework allows to capture early sub-exponential growth dynamics whenever $R_0 > 1$ and $q > 0$. If $q = 0$, the transmission rate $\beta(t) = \beta_0$ remains at the baseline value, and we recover the classic SEIR transmission model with exponential growth dynamics and $R_0 = \beta/\gamma$. In general, the higher the value of q, the faster the decline of the reproduction number from $R_0 > 1$ to a stationary reproduction number at 1.0. We can interpret the parameters q and v through the half time value or the average time elapsed to achieve a transmission rate $\frac{1}{2}\beta_0\,(1-\phi)$. The half time value is given by: $\log(2)/q$.

Importantly, in the context of early sub-exponential (e.g., polynomial) epidemic growth for which $q > 0$, the basic reproduction number is no longer the product of the initial transmission rate β_0 and the mean infectious period γ^{-1} because the transmission rate $\beta(t)$ is no longer constant, but declines during the duration of the infectious period of primary cases at the onset of the epidemic, yielding a lower R_0. For this situation, R_0 can be estimated numerically using the following integral equation (Bacaër and Ait Dads el 2011):

$$R_0 = \int_0^\infty \beta(t) e^{-\gamma\tau} d\tau = \int_0^\infty \beta_0 \left[(1-\phi)\, e^{-q\tau} + \phi\right] e^{-\gamma\tau} d\tau.$$

For a given value of β_0 and γ, the basic reproduction number is R_0 expected to decline from β/γ as parameter q increases above 0. More generally, the effective reproduction number, R_t, during the early epidemic growth phase comprising the first few disease generations of transmission when $S(t) \approx m$ can be numerically computed as follows:

$$R_t = \int_t^\infty \beta(t) e^{-\gamma(\tau-t)} d\tau = \int_t^\infty \beta_0 \left[(1-\phi)\, e^{-q\tau} + \phi\right] e^{-\gamma(\tau-t)} d\tau.$$

For illustration, Fig. 9.7 displays temporal profiles of the transmission rate, the effective reproduction number, and the corresponding simulations of the early epidemic growth phase.

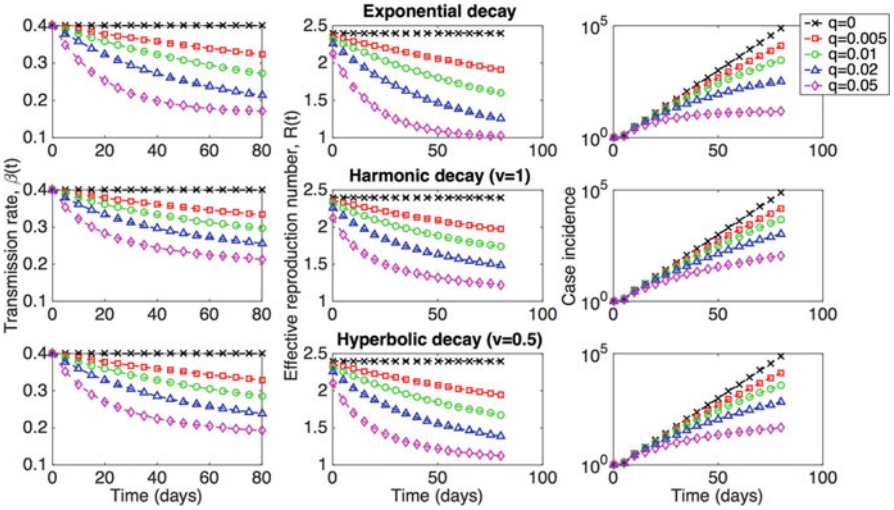


Fig. 9.7 Representative profiles of the transmission rate $\beta(t)$, the effective reproduction number R_t, and corresponding simulations of the early epidemic growth phase derived from the generalized-growth SEIR model (GG-SEIR) for different values of the decline rate parameter q and $\beta_0 = 0.4$,$\alpha = 1/5$, and $\gamma = 1/6$ in a large population size ($N = 10{,}000{,}000$). The epidemic simulations start with one infectious individual. In semi-logarithmic scale, exponential growth is evident if a straight line fits well several consecutive disease generations of the epidemic curve, whereas a strong downward curvature in semi-logarithmic scale is indicative of sub-exponential growth. Our simulations show that case incidence curves display early sub-exponential growth dynamics even for very low values of q

9.4 Case Study: Modeling the Effectiveness of Contact Tracing During Ebola Epidemics

In Mali, two Ebola cases were imported from neighboring Guinea in two different instances, the first resulting in one death and no local secondary cases, and the second resulting in two generations of transmission with a total of eight cases and six deaths in the capital city of Bamako (Breakwell et al. 2016). Both Ebola importations occurred in the fall of 2014 as the epidemic was still unabated in Guinea (2015, 2016), about a month after an Ebola case was imported to Senegal from Guinea, and 4 months after an Ebola case was imported to Nigeria from Liberia (Abdoulaye et al. 2014; The World Health Organization 2014). No further Ebola importations were reported from highly affected countries to neighboring, high-risk countries during the 2014–2016 West Africa Ebola epidemic.

The second Ebola importation in Mali occurred in a grand Imam who traveled from Guinea to Mali and sought care at a private clinic on October 25, 2014, in Bamako where he was treated for kidney failure and was not suspected with Ebola. He died on October 27th and had an unsecured burial on October 28th. Control measures in Mali, including contact tracing, began on November 8th 2014 (Breakwell et al. 2016).

Contact tracing was the primary intervention in response to the second Ebola importation into Mali. Briefly, contact tracing is a method used to prevent further cases of an infectious disease that involves contacting and routinely following up with individuals who have been identified as being exposed to a patient or other vector of a disease for the duration of the maximum observed incubation period of the disease (21 days for Ebola (Shrivastava et al. 2014)). Through effective contact tracing, secondary cases are quickly isolated to prevent further transmission (Eames and Keeling 2003). Although contact tracing is a critical piece of a response to Ebola outbreaks, it was implemented with varying levels of effectiveness across all three of the most affected countries in West Africa (Pandey et al. 2014; Martín et al. 2016; Olu et al. 2016). The success of contact tracing is tightly linked to behavioral interventions, training in infection prevention and control practices in healthcare settings, and initiation of surveillance protocols (Breakwell et al. 2016).

In this case study, we analyze the relation between contact tracing activities and the decline in disease transmission during the Ebola epidemic in Mali. For this purpose, we carried out a comprehensive analysis of contact tracing trees and modeled the relationship between the time-dependent effects of contact tracing and the trajectory of the Ebola outbreak in Bamako assuming two different population structures: (1) a standard homogenous mixing model and (2) a spatially structured model. We illustrate the effect of the rapid and effective implementation of contact tracing activities on outbreak trajectory and size using stochastic simulations.

9.4.1 Model 1: Homogenous-Mixing SEIR Transmission Model

The main features of this model have been described in previous chapters (see Sect. 5.4). A similar model has been previously used to model transmission and control of the Ebola outbreak in Nigeria in 2014 (Chowell et al. 2004; Fasina et al. 2014). For the sake of simplicity, we only model a single infectious compartment while adjusting the time-specific transmission rate according to data of the time-dependent effectiveness in contact tracing activities conducted during the Ebola outbreak in Bamako. Hence, the modeled population was divided into five categories: susceptible individuals (S); exposed individuals (E); infectious and symptomatic individuals (I); and recovered or dead individuals (R). Susceptible individuals infected through contact with infectious individuals enter the latent stage at mean rate $\beta f(t) I(t)/N(t)$, where β is the baseline mean human-to-human transmission rate per day in the absence of interventions, $f(t)$ quantifies the time-dependent effectiveness of contact tracing activities, and $N(t)$ is the total population size at time t. Thus, $f(t)$ ranges from 0 (fully complete contact tracing activities are in place) to 1 (contact tracing efforts are yet to start) to quantify the effectiveness of the isolation of infectious individuals that decrease Ebola transmission through contact tracing efforts. Values of $f(t)$ close to 0 illustrate "near-perfect" contact

tracing, while values closer to 1 illustrate "imperfect" contact tracing efforts. Symptomatic infectious individuals $I(t)$ recover at the mean rate γ. Individuals in the "removed" category do not contribute to the transmission process. Thus, the time-dependent contact tracing effectiveness, $f(t)$, remains at 1.0 before the start of contact tracing activities. Baseline epidemiological parameters were set according to the epidemiology of Ebola (i.e., incubation period of 5 days (Eichner et al. 2011; The World Health Organization 2014) and infectious period of 7 days (Chowell et al. 2004; The World Health Organization 2014)). We set the effective population size at 2,400,000 based on the population size of Bamako. For this model, R_0 is given by the product of the transmission rate β and the mean infectious period $1/\gamma$. Hence, specific values of R_0 (range: 1.6–2.0 based on estimates of the Western African outbreak (Althaus 2014; Nishiura and Chowell 2014)) were calibrated by tuning β. Once interventions are put in place, the effective reproduction number declines according to the formula

$$R_t = R_0 \frac{S(t)}{m} f(t),$$

where $S(t)/m$ quantifies the proportion of susceptible individuals at time t.

9.4.2 *Model 2: Spatially Structured Ebola Transmission Model*

One of the putative mechanisms leading to early polynomial growth dynamics of Ebola transmission is clustering (Szendroi and Csányi 2004; Chowell et al. 2015; Merler et al. 2015; Viboud et al. 2016; Chowell et al. 2017), a network property that quantifies the extent to which the contacts of one individual are also contacts of each other (Watts and Strogatz 1998). Social contact networks are particularly useful to explore the impact of clustering and play an important in the dissemination of Ebola at the community level. We employ a network-based transmission model with household-community structure, which has been previously applied to study the transmission dynamics of Ebola (Kiskowski 2014; Kiskowski and Chowell 2015).

In this model, individuals are organized within households of size H (each household contains H individuals) and households are organized within communities of size C households (each community contains $C \times H$ individuals). Network connectivity is identical for every individual. The household reproduction number R_{0H} was varied between 1.6 and 2.0 and the community reproduction number R_{0C} was set at 0.7 based on previous study (Kiskowski and Chowell 2015). For a fixed household size at $H = 6$, which is in line with the average household size for Bamako in 2014 and various values of the community size parameter (range: 25–65 households per community), we analyze the resulting outbreak size distribution.

9.4.3 *Modeling the Time-Dependent Effectiveness of Contact Tracing Efforts in Bamako, Mali*

After the start of the interventions at time t_0, the function $f(t)$ modulates a decline in transmission rate according to the time-dependent completeness of contact tracing efforts. The functional form for $f(t)$ was assumed to follow an exponential decline after the start of contact tracing activities. That is,

$$f(t) = \begin{cases} 1, & 0 < t < t_0 \\ 1 - (1 - e^{-q(t-t_0)}), & t \geq t_0 \end{cases}.$$

Parameters q and t_0 could be estimated by fitting $f(t)$ to the daily contact tracing completeness calculated as the daily proportion of contact persons that were monitored out of the total number of registered contact persons at risk. For illustration purposes, we set $q = 0.14$ while the start of contact tracing efforts is fixed at $t_0 = 21$, which is in line with the outbreak response in Mali. The corresponding estimates of the effective reproduction number are shown in Fig. 9.9b.

Stochastic Simulations

To assess the temporal and size distribution of outbreaks, we generated 200 stochastic epidemic simulations that start with the introduction of the index case (i.e., $I(0) = 1$). Simulation code in Matlab is available upon request from the authors.

In the absence of interventions, the spatial and non-spatial models exhibit strikingly different epidemic trajectories as shown in Fig. 9.8.

The resulting curves of the effective reproduction number, R_t, capturing the time-dependent effects of contact tracing efforts for three different values of R_0 are shown in Fig. 9.9 based on the homogeneous mixing model. R_t declined below the epidemic threshold of 1.0 between November 10th and November 13th, 2014. The illustrated effect of control interventions on the transmission of Ebola in Mali is shown with an ensemble of stochastic epidemic realizations in Fig. 9.9a, which shows the relative reduction in the transmission rate as a function of the time-dependent effectiveness of contact tracing activities. After the start of the interventions, the function modulates a decline in transmission rate according to the time-dependent effectiveness of contact tracing efforts as explained in the text. This time-dependent function was assumed to follow an exponential decline after the start of contact tracing activities. Figure 9.9b shows the effective reproduction number over time reflecting the impact of contact tracing activities for three different

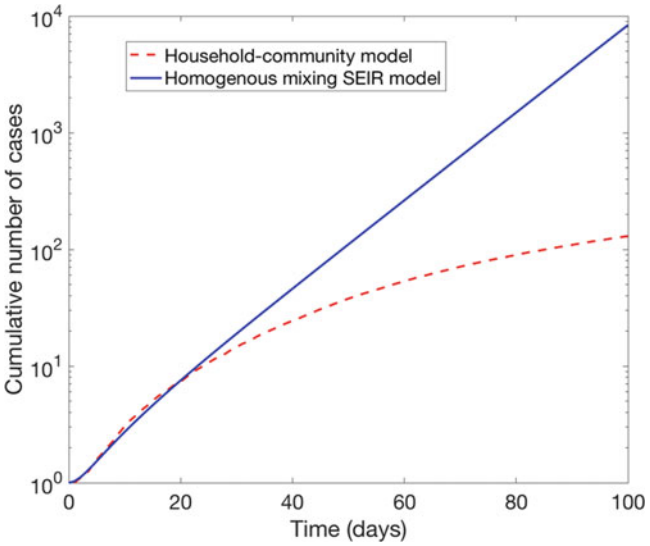


Fig. 9.8 The mean epidemic trajectories derived from the spatial and non-spatial models during the first 100 days of the Ebola epidemic in the absence of interventions

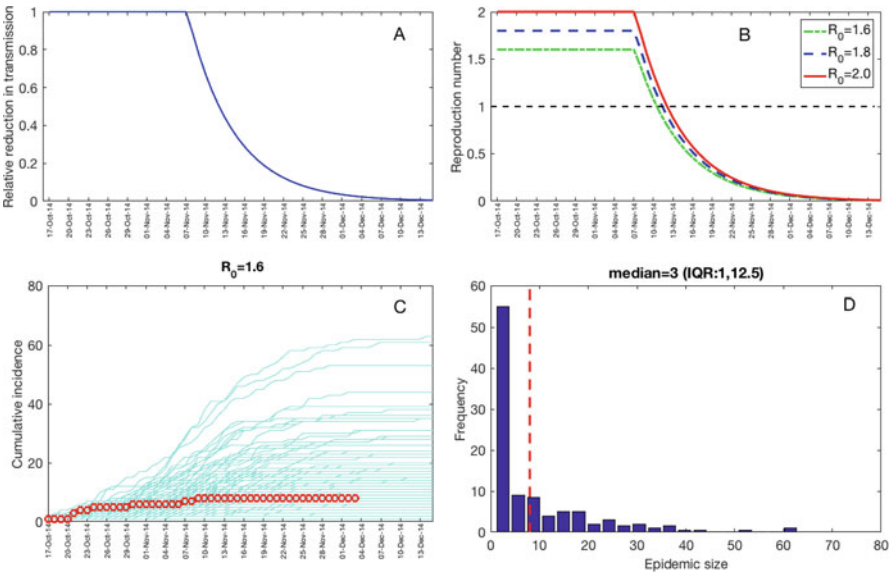


Fig. 9.9 (**a**) The relative reduction in the transmission rate as a function of the time-dependent effectiveness of contact tracing activities. (**b**) The effective reproduction number over time reflecting the impact of contact tracing. (**c**) Stochastic epidemic realizations using the homogenous-mixing SEIR model (Model 1) at $R_0 = 1.6$. The red circles correspond to the actual outbreak trajectory and the cyan blue lines correspond to 200 stochastic realizations. (**d**) The corresponding distribution of outbreak sizes using the homogenous-mixing SEIR model (Model 1) with an R_0 set at 1.6. The vertical dashed line indicates the actual Ebola outbreak size in Mali

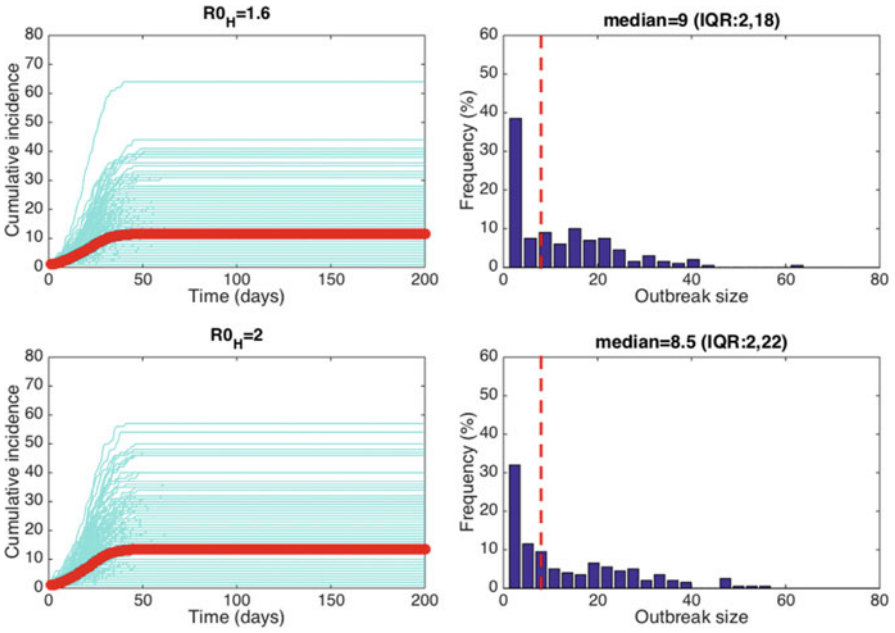


Fig. 9.10 Stochastic epidemic realizations using the household-community SEIR model (Model 2) with a community size of 25 households and household size of 6 which is in line with the average household size for Bamako in 2014

values of R_0 in the range 1.6–2.0. Figure 9.9c illustrates stochastic epidemic realizations using the homogenous-mixing SEIR model (Model 1) with an R_0 set at 1.6. Figure 9.9d shows the corresponding distribution of outbreak sizes using the homogenous-mixing SEIR model (Model 1) with an R_0 set at 1.6. The corresponding results based on the spatially structured model are shown in Fig. 9.10 assuming a community size $C = 25$.

Our modeling analysis demonstrates that the decline in transmission and subsequent halting of the Ebola outbreak in Mali coincided with the implementation of contact tracing activities that improved over the course of the outbreak. The results suggest that contact tracing done completely during an outbreak could minimize the size of future outbreaks. While the spatial and non-spatial models yield significantly different epidemic trajectories in the absence of interventions (Fig. 9.8), it is perhaps not surprising that the spatial and non-spatial transmission models yielded similar outbreak size distributions because the virus was contained before it could spread beyond a few generations of disease transmission. In the absence of comprehensive contact tracing efforts, person-to-person transmission of Ebola could have increased rapidly, ensuing in a sizable urban epidemic.

9.5 Problems and Supplements

9.1 Consider a simple SIR model with an $R_0 = 1.8$, a mean infectious period of 3 days and a population size of 100,000 people that incorporates the effects of behavior changes that mitigate the transmission rate as follows: After the first 30 days of the epidemic, the transmission rate decreases exponentially fast with a half-life of 10 days. Answer the following questions:

(a) Compare the size of the epidemics obtained with and without the effects of behavior changes.
(b) Explore how the epidemic size changes as you vary the timing of the start of the behavior change and the half-life of the transmission rate decay associated with the behavior change.

9.2 Consider a simple two-patch SEIR model with local $R_0 = 1.5$, mean latent period of 7 days, mean infectious period of 4 days, and a population size of 10,000 people in each patch. Further, transmission can occur in two different ways: (1) local transmission within each patch and (2) directed transmission from the first patch to the second patch (but not from the second to the first patch) where this patch-to-patch transmission rate is a fraction ρ relative to the local transmission rate. Answer the following questions:

(a) Using the simple SEIR model without demographic factors and assuming a mean latent period of 2 days, a mean infectious period of 4 days, and a population size of 550,000, provide the mean estimate and 95% confidence intervals of the basic reproduction number R_0 using 16, 18, and 20 days of the initial growth phase. For parameter estimation you can use the least square fitting approach with the Poisson parametric bootstrap which is described in Chap. 7 and illustrated with examples in Chap. 8. Note that you only need to estimate the transmission rate using your favorite technical computing language while keeping the initial number of infectious individuals $I(0)$ fixed according to the first data point. Are the R_0 estimates relatively stable during the study period?
(b) Describe the dynamics of the epidemics as the parameter rho is increased from 0.00001 to 0.01. In particular, how many peaks do the total incidence curve exhibit as this parameter is varied?
(c) Describe the epidemic duration and size that result from (a).
(d) Repeat the analyses in (a) using a system of 4 patches connected in a linear fashion where patch-to-patch transmission only occurs from patch j to patch $j + 1$.

References

Abdoulaye, B., Moussa, S., Daye, K., Boubakar, B. S., Cor, S. S., Idrissa, T., et al. (2014). Experience on the management of the first imported Ebola virus disease case in Senegal. *The Pan African Medical Journal, 22*(Suppl. 1), 6.

Albert, R., & Barabasi, A. L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics, 74*(1), 47–97.

Allen, L. J. (2010). *An introduction to stochastic processes with applications to biology*. Boca Raton, FL: CRC Press.

Allen, L. J. (2017). A primer on stochastic epidemic models: Formulation, numerical simulation, and analysis. *Infectious Disease Modelling, 2*(2), 128–142.

Alonso, W. J., Nascimento, F. C., Chowell, G., & Schuck-Paim, C. (2018). We could learn much more from 1918 pandemic-the (mis)fortune of research relying on original death certificates. *Annals of Epidemiology, 28*(5), 289–292.

Althaus, C. L. (2014). Estimating the reproduction number of Ebola virus (EBOV) during the 2014 outbreak in West Africa. arXiv preprint. arXiv:1408.3505.

Andersen, P. K., Borgan, O., Gill, R. D., & Keiding, N. (1993). *Statistical models based on counting processes*. New York, NY: Springer.

Anderson, D., & Watson, R. (1980). On the spread of a disease with gamma distributed latent and infectious periods. *Biometrika, 67*(1), 191–198.

Anderson, R. M., & May, R. M. (1982). Directly transmitted infectious diseases: Control by vaccination. *Science, 215*, 1053–1060.

Anderson, R. M., & May, R. M. (1991) *Infectious diseases of humans, dynamics and control*. Oxford: Oxford University Press.

Andersson, H., & Britton, T. (2012). *Stochastic epidemic models and their statistical analysis* (Vol. 151). New York, NY: Springer.

Andersson, H., & Djehiche, B. (1998). A threshold limit theorem for the stochastic logistic epidemic. *Journal of Applied Probability, 35*(3), 662–670.

Anscombe, F. J. (1953) Contribution to the discussion of H. Hotelling's paper. *Journal of Royal Statistics Society (B), 15*, 229–230.

Apolloni, A., Poletto, C., Ramasco, J. J., Jensen P., & Colizza, V. (2014). Metapopulation epidemic models with heterogeneous mixing and travel behaviour. *Theoretical Biology and Medical Modelling, 11*, 3.

Appoloni, A., Poletto, C., & Colizza, V. (2013). Age-specific contacts and travel patterns in the spatial spread of 2009 H1N1 influenza pandemic. *BMC Infectious Diseases, 13*, 176.

P. Yan, G. Chowell, *Quantitative Methods for Investigating Infectious Disease Outbreaks*, Texts in Applied Mathematics 70,
https://doi.org/10.1007/978-3-030-21923-9

Arino, J., Davis, J. R., Hartley, D., Jordan, R., Miller, J. M., & van den Driessche, P. (2005). A multi-species epidemic model with spatial dynamics. *Mathematical Medicine and Biology, 22*(2), 129–142.

Arino, J., Jordan, R., & van den Driessche, P. (2007). Quarantine in a multi-species epidemic model with spatial dynamics. *Mathematical Biosciences, 206*(1), 46–60.

Arita, I., Shafa, E., & Kader, A. (1970). Role of hospital in smallpox outbreak in Kuwait. *American Journal of Public Health and the Nations Health, 60*(10), 1960–1966.

Arnold, B. C. (1983). *Pareto distributions*. Fairland, MD: International Co-operative Publishing House.

Arriola, L., & Hyman, J. M. (2009). Sensitivity analysis for uncertainty quantification in mathematical models. In G. Chowell, J. M. Hyman, L. M. Bettencourt, & C. Castillo-Chavez (Eds.), *Mathematical and statistical estimation approaches in epidemiology*. Dordrecht: Springer.

Assiri, A., McGeer, A., Perl, T. M., Price, C. S., Al Rabeeah, A. A., Cummings, D. A., et al. (2013). Hospital outbreak of Middle East respiratory syndrome coronavirus. *New England Journal of Medicine, 369*(5), 407–416.

Bacaër, N., & Abdurahman, X. (2008). Resonance of the epidemic threshold in a periodic environment. *Journal of Mathematical Biology, 57*, 649–673.

Bacaër, N., & Ait Dads el, H. (2011). Genealogy with seasonality, the basic reproduction number, and the influenza pandemic. *Journal of Mathematical Biology, 62*(5), 741–762.

Bacchetti, P. B., & Moss, A. R. (1989). Incubation period of AIDS in San Francisco, *Nature, 338*, 251–253.

Bailey, N. T. J. (1975). *The mathematical theory of infectious diseases and its applications* (2nd ed.). London: The Griffin & Company Ltd.

Ball, F. G., Britton, T., House, T., Isham, V., Mollison, D., Pellis, L., et al. (2015). Seven challenges for metapopulation models of epidemics, including households models. *Epidemics, 10*, 63–67.

Ball, F. G., Britton, T., & Neal, P. (2016). On expected durations of birth-death processes, with applications to branching processes and SIS epidemics. *Journal of Applied Probability, 53*, 203–215.

Ball, F. G., Sirl, D., & Trapman, P. (2009). Threshold behaviour and final outcome of an epidemic on a random network with household structure. *Advances in Applied Probability, 41*, 765–796.

Banks, H. T., Davidian, M., Samuels, J. R., & Sutton, K. L. (2009). An inverse problem statistical methodology summary. In G. Chowell, J. M. Hyman, L. M. Bettencourt, & C. Castillo-Chavez (Eds.), *Mathematical and statistical estimation approaches in epidemiology*. Dordrecht: Springer.

Banks, H. T., Hu, S., & Thompson, W. C. (2014). *Modeling and inverse problems in the presence of uncertainty*. Boca Raton, FL: CRC Press.

Banks, R. B. (1994). *Growth and diffusion phenomena: Mathematical frameworks and applications*. Berlin: Springer.

Bansal, S., Grenfell, B. T., & Meyers, L. A. (2007). When individual behaviour matters: Homogeneous and network models in epidemiology. *Journal of the Royal Society Interface, 4*(16), 879–891.

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random network. *Science, 286*, 509–512.

Bartlett, M. S. (1955). *An introduction to stochastic processes*. London: Cambridge University Press.

Barttlet, M. S. (1961). *Stochastic population models in ecology and epidemiology*. London: Methuen and Co. Ltd.

Becker, N. G. (1989). *Analysis of infectious disease data*. London: Chapman and Hall/CRC.

Becker, N. G. (2015). *Modeling to inform infectious disease control*. London: Chapman and Hall/CRC.

Becker, N. G., & Britton, T. (2001). Design issues of studies of infectious diseases. *Journal of Statistical Planning and Inference, 96*, 41–66.

Becker, N. G., & Hasofer, A. M. (1997). Estimation in epidemics with incomplete observations. *Journal of the Royal Statistical Society: Series B, 59*(2), 415–429.

Becker, N. G., Watson, L. F., & Carlin, J. B. (1991). A method of non-parametric back-projection and its application to AIDS data. *Statistics in Medicine, 10*, 1527–1542.

Belik, V., Geisel, T., & Brockmann, D. (2011). Natural human mobility patterns and spatial spread of infectious diseases. *Physical Review X, 1*, 011001.

Bellman, R. E., & Roth, R. S. (1984). *The Laplace transform*. Singapore: World Scientific.

Bhattacharya, R. N., & Waymire, E. C. (1990). *Stochastic processes with applications*. New York, NY: Wiley.

Biggerstaff, M., Alper, D., Dredze, M., Fox, S., Fung, I. C., Hickmann, K. S., et al. (2016). Results from the centers for disease control and prevention's predict the 2013–2014 Influenza Season Challenge. *BMC Infectious Diseases, 16*, 357.

Blythe, S. P., & Castillo-Chavez, C. (1989). Like-with-like preference and sexual mixing models. *Mathematical Biosciences, 96*, 221–238.

Bondesson, L. (1979). On generalized gamma and generalized negative binomial convolutions, part II. *Scandinavian Actuarial Journal, 2–3*, 147–166.

Borel, É. (1942). Sur l'emploi du théorème de Bernoulli pour faciliter le calcul d'une infinité de coefficients. Application au problème de l'attente à un guichet. *Comptes rendus de l'Académie des Sciences, 214*, 452–456.

Brauer, F. (2006). Some simple epidemic models. *Mathematical Biosciences and Engineering, 3*, 1–15.

Brauer, F. (2008). Compartmental models in epidemiology. In F. Brauer, P. van den Driessche, & J. Wu (Eds.), *Mathematical epidemiology* (Chapter 2). Berlin: Springer.

Brauer, F., & Castillo-Chávez, C. (2001). *Mathematical models in population biology and epidemiology*. New York, NY: Springer.

Brauer, F., van den Driessche, P., & Wu, J. (Eds.). (2008). *Mathematical epidemiology*. Berlin: Springer.

Breakwell, L., Gerber, A. R., Greiner, A. L., Hastings, D. L., Mirkovic, K., Paczkowski, M. M., et al. (2016). Early identification and prevention of the spread of Ebola in high-risk African countries. *MMWR Supplements, 65*(3), 21–27.

Britton, T., & Scalia Tomba, G. (2019). Estimation in emerging epidemics: Biases and remedies. *Journal of the Royal Society Interface, 16*(150), 20180670.

Brookmeyer, R., & Gail, M. H. (1994). *AIDS epidemiology: A quantitative approach*. New York, NY: Oxford University Press.

Brookmeyer, R., & Goedert, J. J. (1989). Censoring in an epidemic with an application to hemophilia-associated AIDS. *Biometrics, 45*, 325–335.

Capaldi, A., Behrend, S., Berman, B., Smith, J., Wright, J., & Lloyd, A. L. (2012). Parameter estimation and uncertainty quantification for an epidemic model. *Mathematical Biosciences & Engineering, 9*, 553–576.

Castillo-Chávez, C., Blower, S., van den Driessche, P., Kirschner D., & Yakubu, A. A. (2000). *Mathematical approaches for emerging and reemerging infectious diseases*. New York, NY: Springer.

Castillo-Chavez C., Feng, Z., & Huang, W. (2002). On the computation Ro and its role on global stability. In C. Castillo-Chavez, P. van den Driessche, D. Kirschner, & A.-A. Yakubu (Eds.), *Mathematical approaches for emerging and reemerging infectious diseases: An introduction, IMA* (Vol. 125, pp. 229–250). Berlin: Springer.

Causton, D. R., & Venus, J. C. (1981). *The biometry of plant growth*. London: Edward Arnold.

Champredon, D., & Dushoff, J. (2015). Intrinsic and realized generation intervals in infectious-disease transmission. *Proceedings of the Royal Society B, 282*(1821), 2015–2026.

Chowell, G. (2017). Fitting dynamic models to epidemic outbreaks with quantified uncertainty: A primer for parameter uncertainty, identifiability, and forecasts. *Infectious Disease Modelling, 2*, 379–398.

Chowell, G., Ammon, C. E., Hengartner, N. W., & Hyman, J. M. (2006). Transmission dynamics of the great influenza pandemic of 1918 in Geneva, Switzerland: Assessing the effects of hypothetical interventions. *Journal of Theoretical Biology, 241*, 193–204.

Chowell, G., Hengartner, N. W., Castillo-Chavez, C., Fenimore, P. W., & Hyman, J. M. (2004). The basic reproductive number of Ebola and the effects of public health measures: The cases of Congo and Uganda. *Journal of Theoretical Biology, 229*(1), 119–126.

Chowell, G., Hincapie-Palacio, D., Ospina, J., Pell, B., Tariq, A., Dahal, S., et al. (2016). Using phenomenological models to characterize transmissibility and forecast patterns and final burden of Zika epidemics. *PLOS Currents Outbreaks, 8*. https://doi.org/10.1371/currents.outbreaks.f14b2217c902f453d9320a43a35b9583

Chowell, G., & Hyman, J. M. (Eds.). (2016). *Mathematical and statistical modeling for emerging and re-emerging infectious diseases*. Cham: Springer.

Chowell, G., Nishiura, H., & Bettencourt, L. M. (2007). Comparative estimation of the reproduction number for pandemic influenza from daily case notification data. *Journal of the Royal Society Interface, 4*(12), 155–166.

Chowell, G., Rivas, A. L., Hengartner, N. W., Hyman, J. M., & Castillo-Chavez, C. (2006). The role of spatial mixing in the spread of foot-and-mouth disease. *Preventive Veterinary Medicine, 73*(4), 297–314.

Chowell, G., & Rothenberg, R. (2018). Spatial infectious disease epidemiology: On the cusp. *BMC Medicine, 16*(1), 192.

Chowell, G., Sattenspiel, L., Bansal, S., & Viboud, C. (2016). Mathematical models to characterize early epidemic growth: A review. *Physics of Life Reviews, 18*, 66–97.

Chowell, G., Shim, E., Brauer, F., Diaz-Duenas, P., Hyman, J. M., & Castillo-Chavez, C. (2006). Modelling the transmission dynamics of acute haemorrhagic conjunctivitis: Application to the 2003 outbreak in Mexico. *Statistics in Medicine, 25*, 1840–1857.

Chowell, G., & Viboud, C. (2016). Is it growing exponentially fast? – Impact of assuming exponential growth for characterizing and forecasting epidemics with initial near-exponential growth dynamics. *Infectious Disease Modelling, 1*, 71–78.

Chowell, G., Viboud, C., Hyman, J. M., & Simonsen, L. (2015). The Western Africa ebola virus disease epidemic exhibits both global exponential and local polynomial growth rates. *PLoS Currents, 7*. https://doi.org/10.1371/currents.outbreaks.8b55f4bad99ac5c5db3663e916803261

Chowell, G., Viboud, C., Simonsen, L., Merler, S., & Vespignani, A. (2017). Perspectives on model forecasts of the 2014–2015 Ebola epidemic in West Africa: Lessons and the way forward. *BMC Medicine, 15*, 42.

Chowell, G., Viboud, C., Simonsen, L., & Moghadas, S. (2016). Characterizing the reproduction number of epidemics with early sub-exponential growth dynamics. *Journal of the Royal Society Interface, 13*(123). https://doi.org/10.1098/rsif.2016.0659

Chretien, J. P., Swedlow, D., Eckstrand, I., Johansson, M., Huffman, R., & Hebbeler, A.(2015). Advancing epidemic prediction and forecasting: A new US government initiative. *Online Journal of Public Health Informatics, 7*(1), e13.

Clancy, D. (2018). Precise estimates of persistence time for SIS infections in heterogeneous populations. *Bulletin of Mathematical Biology, 80*(11), 2871–2896. https://dol.org/10.1007/s11538-018-0491-6

Clancy, D., & Mendy, S. T. (2011). Approximating the quasi-stationary distribution of the SIS model for endemic infection. *Methodology and Computing in Applied Probability, 12*(3). https://doi.org/10.1007/s11009-010-9177-8

Cobelli, C., & Romanin-Jacur, G. (1976). Controllability, observability and structural identifiability of multi input and multi output biological compartmental systems. *IEEE Transactions on Biomedical Engineering, 23*, 93–100.

Cook, R. J., & Lawless, J. F. (2007). *The statistical analysis of recurrent events*. New York, NY: Springer.

Cook, R. J., & Lawless, J. F. (2018). *Multistate models for the analysis of life history data*. New York, NY: Chapman and Hall/CRC.

Corless, R. M., Gonnet, G. H., Hare, D. G. E., Jeffery, D. J., & Knuth, D. E. (1996). On the Lambert W function. *Advances in Computational Mathematics, 5*, 329–359.

Cox, D. R. (2006). *Principles of statistical inference*. Cambridge: Cambridge Press.

Cox, D. R., & Donnelly, C. A. (2011). *Principles of applied statistics*. Cambridge: Cambridge University Press.

Cox, D. R., & Oaks, D. (1987). *Analysis of survival data*. New York, NY: Champan and Hall/CRC.

Cresswell, W. L., & Froggatt, P. (1963). *The causation of bus driver accidents*. London: Oxford University Press.

Crump, K. (1975). On point processes having an order statistic structure. *Sankhya, Series A, 37*, 396–404.

Dahal, S., Jenner, M., Dinh, L., Mizumoto, K., Viboud, C., & Chowell, G. (2017). Excess mortality patterns during 1918–1921 influenza pandemic in the state of Arizona, USA. *Annals of Epidemiology, 28*(5), 273–280.

Daley, D. J., & Gani, J. (1999). *Epidemic modelling, an introduction*. Cambridge: Cambridge University Press.

Danon, L., Ford, A. P., House, T., Jewell, C. P., Keeling, M. J., Roberts, G. O., et al. (2011). Networks and the epidemiology of infectious disease. *Interdisciplinary Perspectives on Infectious Diseases, 2011*, 284909.

Deakin, M. A. B. (1975). A standard form for the Kermack-McKendrick epidemic equations. *Bulletin of Mathematical Biology, 37*, 91–95.

Devroye, L. (1992). Random variate generation for the digamma and trigamma distributions. *Journal of Statistical Computation and Simulation, 43*(3–4), 197–216.

Diekmann, O., & Heesterbeek, J. A. P. (2000). *Mathematical epidemiology of infectious diseases: Model building, analysis and interpretation*. *Mathematical and computational biology* (Vol. 5). Chichester: Wiley.

Diekmann, O., Heesterbeek, J. A. P., & Metz, J. A. (1990). On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations. *Journal of Mathematical Biology, 28*, 365–382.

Diekmann, O., Heesterbeek, J. A. P., & Roberts, M. G. (2010). The construction of next-generation matrices for compartmental epidemic models. *Journal of the Royal Society Interface, 7*, 873–885.

Dietz, K. (1995). Some problems in the theory of infectious diseases transmission and control. In D. Mollison (Ed.), *Epidemic models: Their structure and relation to data* (pp. 3–16). Cambridge: Cambridge University Press.

Dinh, L., Chowell, G., Mizumoto, K., & Nishiura, H. (2016). Estimating the subcritical transmissibility of the Zika outbreak in the State of Florida, USA, 2016. *Theoretical Biology and Medical Modelling, 13*, 20.

Dion, J. P. (1975). Estimation of the variance of a branching process. *The Annuals of Statistics, 3*(5), 1183–1187.

Donnelly, C. A., & Ferguson, N. M. (1999). *Statistical aspects of BSE and vCJD, models for epidemics*. New York, NY: Chapman and Hall/CRC.

Dubey, S. D. (1968). A compound Weibull distribution. *Naval Research Logistics Quarterly, 15*, 179–188.

Dubey, S. D. (1969). A new derivation of the logistic distribution. *Naval Research Logistics Quarterly, 16*, 37–40.

Eames, K. T., & Keeling, M. J. (2003). Contact tracing and disease control. *Proceedings of the Royal Society of London B: Biological Sciences, 270*(1533), 2565–2571.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. New York, NY: CRC Press.

Eichner, M., Dowell, S. F., & Firese, N. (2011). Incubation period of ebola hemorrhagic virus subtype zaire. *Osong Public Health and Research Perspectives, 2*(1), 3–7.

Erdös, P., & Rényi, A. (1961). On the evolution of random graphs. *Bulletin of the International Statistical Institute, 38*, 343–347.

Farewell, V. T., Herzberg, A. M., James, K. W., Ho, L. M., & Leung, G. M. (2005). SARS incubation and quarantine times: When is an exposed individual known to be disease free? *Statistics in Medicine, 24*, 3431–3445.

Faria, N. R., da Silva Azevedo, R. D. S., Kraemer, M. U., Souza, R., Cunha, M. S., Hill, S. C., Thézé, J., Bonsall, M. B., Bowden, T. A., Rissanen, I., & Rocco, I. M. (2016). Zika virus in the Americas: Early epidemiological and genetic findings. *Science, 352*(6283), 345–349.

Farrington, C. P., & Grant, A. D. (1999). The distribution of time to extinction in subcritical branching processes: Applications to outbreaks of infectious disease. *Journal of Applied Probability, 36*, 771–779

Farrington, C. P., Kanaan, M., & Gay, J. N. (2003). Branching process models for surveillance of infectious diseases controlled by mass vaccination. *Biostatistics, 4*, 279–295.

Fasina, F., Shittu, A., Lazarus, D., Tomori, O., Simonsen, L., Viboud, C., et al. (2014). Transmission dynamics and control of Ebola virus disease outbreak in Nigeria, July to September 2014. *Eurosurveillance, 19*(40), 20920.

Fauci, A. S., & Morens, D. M. (2016). Zika virus in the Americas–Yet another arbovirus threat. *New England Journal of Medicine, 374*, 601–604.

Feller, W. (1943). On a generalized class of contagious distributions. *Annuals of Mathematical Statistics, 14*, 389–400.

Feller, W. (1966). *An introduction to probability theory and its applications*. New York, NY: Wiley.

Feng, Z., Xu, D., & Zhao, H. (2007). Epidemiological models with non-exponentially distributed disease stages and applications to disease control. *Bulletin of Mathematical Biology*. https://doi.org/10.1007/s11538-006-9174-9

Fenichel, E. P., Castillo-Chavez, C., Ceddia, M. G., Chowell, G., Parra, P. A. G., Hickling, G. J., et al. (2011). Adaptive human behavior in epidemiological models. *Proceedings of the National Academy of Sciences, 108*(15), 6306–6311.

Fenner, F., Henderson, D. A., Arita, I., Jezěk, Z., & Ladnyi, I. D. (1988). *Smallpox and its eradication*. Geneva: World Health Organization.

Ferguson, N. M., Cummings, D. A. T., Cauchemez, S., Fraser, C., Riley, S., Meeyai, A., et al. (2005). Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature, 437*(7056), 209.

Fine, P. E. M. (2003). The interval between successive cases of an infectious disease. *American Journal of Epidemiology, 158*(11), 1039–1047.

Fleming, T. R., & Harrington, D. P. (1991). *Counting processes and survival analysis*. New York, NY: Wiley.

Focks, D. A., Daniels, E., Haile, D. G., & Keesling, J. E. (1995). A simulation model of the epidemiology of urban dengue fever: Literature analysis, model development, preliminary validation, and samples of simulation results. *The American Journal of Tropical Medicine and Hygiene, 53*(5), 489–506.

Fraser, C. (2007). Estimating individual and household reproduction numbers in an emerging epidemic. *PLoS One, 2*(8), e758.

Funk, S., Gilad, E., Watkins, C., & Jansen, V. A. (2009). The spread of awareness and its impact on epidemic outbreaks. *Proceedings of the National Academy of Sciences, 106*(16), 6872–6877.

Funk, S., Salathé, M., & Jansen, V. A. (2010). Modelling the influence of human behaviour on the spread of infectious diseases: A review. *Journal of the Royal Society Interface, 7*(50), 1247–1256.

Gani, R., Hughes, H., Fleming, D., Griffin, T., Medlock, J., & Leach, S. (2005). Potential impact of antiviral drug use during influenza pandemic. *Emerging Infectious Diseases, 11*(9), 1355–1362.

Gao, D., Lou, Y., He, D., Porco, T. C., Kuang, Y., Chowell, G., et al. (2016). Prevention and control of Zika as a mosquito-borne and sexually transmitted disease: A mathematical modeling analysis. *Scientific Reports, 6*, 28070.

Gauvreau, K., Degruttola, V., & Pagano, M. (1994). The effect of covariates on the induction time of AIDS using improved imputation of exact seroconversion times. *Statistics in Medicine, 13*, 2021–2130.

Gleser, L. J. (1989). The gamma distribution as a mixture of exponential distributions. *The American Statistician, 43*, 115–117.

Godambe, V. P., & Heyde, C. C. (1987). Quasi-likelihood and optimal estimation. *International Statistical Review, 55*, 231–244.

Goedert, J. J., Kessler, C. M., Aledort, L. M., Biggar, R. J., Andes, W. A., White, G. C., et al. (1989). A perspective study of human immunodeficiency virus type 1 infection and the development of AIDS in subjects with hemophilia. *The New England Journal of Medicine, 321*, 1141–1148.

Goh, K. T., Cutter, J., Heng, B. H., Ma, S., Koh, B. K., Kwok, C., et al. (2006). Epidemiology and control of SARS in Singapore. *Annals-Academy of Medicine Singapore, 35*(5), 301.

Goldstein, E., Paur, K., Fraser, C., Kenah, E., Wallinga, J., & Lipsitch, M. (2009). Reproductive numbers, epidemic spread and control in a community of households. *Mathematical Biosciences, 221*, 11–25.

Goldstein, S. (1932). Operational representation of Whittaker's confluent hypergeometric function and Weber's parabolic cylinder function. *Proceedings of the London Mathematical Society, 2*, 103–125.

Greenwood, M., & Yule, G. (1920). An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to occurrence of multiple attacks of diseases or of repeated accidents. *Journal of the Royal Statistical Society A, 83*, 255–279.

Haccou, P., Jagers, P., & Vatutin, V. (2005). *Branching processes: Variation, growth, and extinction of populations*. Cambridge: Cambridge University Press.

Hadeler, K. P., & Castillo-Chavez, C. (1995). A core group model for disease transmission. *Mathematical Biosciences, 128*, 41–55.

Halloran, M. E., Longini, I. M., Nizam, A., & Yang, Y. (2002). Containing bioterrorist smallpox. *Science, 298*(5597), 1428–1432.

Halloran, M. E., Longini, I. M., & Struchiner, C. J. (2009). *Design and analysis of vaccine studies*. New York, NY: Springer.

Harris, T. (1948). Branching processes. *Annals of Mathematical Statistics, 19*, 474–494.

Harris, T. (1963). *The Theory of Branching Processes*. Berlin: Springer.

Hernádez-Suárez, C. M., & Castillo-Chavex, C. (1999). A basic result on the integral for birth-death Markov processes. *Mathematical Biosciences, 161*, 95–104.

Hesselager, O., Wang, S., & Willmot, G. E. (1998). Exponential and scale mixtures and equilibrium distributions. *Scandinavian Actuarial Journal, 2*, 125–142.

Hessol, N. A., Lifson, A. R., O'Malley, P. M., Doll, L. S., Jaffe, H. W., Rutherford, G. W. (1989). Prevalence, incidence, and progression of human immunodeficiency virus infection in homosexual and bisexual men in hepatitis B vaccine trials, 1978–1988. *American Journal of Epidemiology, 130*, 1167–1175.

Hethcote, H. W. (2000). The mathematics of infectious diseases. *SIAM Review, 42*(4), 599–653.

Hethcote, H. W., & van den Driessche, P. (1991). Some epidemiological models with nonlinear incidence. *Journal of Mathematical Biology, 29*, 271.

Heyde, C. C. (1974). On estimating the variance of the offspring distribution in a simple branching process. *Advances in Applied Probability, 6*(3), 421–433.

Hohle, M., & Jørgensen, E. (2003). Estimating parameters for stochastic epidemics. *Dina Research Report, 102*. http://staff.math.su.se/hoehle/pubs/dina102.pdf

Hope Simpson, R. E. (1948). The period of transmission in certain epidemic diseases: An observational method for its discovery. *Lancet, 2*, 755–760.

Hougaard, P. (1984). Life table methods for heterogenous populations. Distributions describing the heterogeneity. *Biometrika, 71*, 75–83.

Hsieh, Y. H., & Cheng, Y. S. (2006). Real-time forecast of multiphase outbreak. *Emerging Infectious Diseases, 12*, 122–127.

Huber, J. H., Childs, M. L., Caldwell, J. M. & Mordecai, E. A. (2018). Seasonal temperature variation influences climate suitability for dengue, chikungunya, and Zika transmission. *PLoS Neglected Tropical Diseases, 12*, e0006451. https://doi.org/10.1371/journal.pntd.0006451

Irwin, J. O. (1941). Discussion on chambers and Yule's paper. *Journal of Royal Statistical Society, 7*(2), 101–109.

Isham, V. (1991). Assessing the variability of stochastic epidemics. *Mathematical Biosciences, 107*, 209–224.

Isham, V. (2005). Stochastic models for epidemics. In A. C. Davison, Y. Dodge, & N. Wermuth (Eds.), *Celebrating statistics: papers in honour of Sir David Cox on his 80th birthday. Oxford statistical science series* (Chapter 1, Vol. 33). Oxford: Oxford University Press.

Italian Seroconversion Study. (1992). Disease progression and early predictors of AIDS in HIV-seroconverted injecting drug users. *AIDS, 6*, 421–426.

Jacquez, J. A. (1996). *Compartmental analysis in biology and medicine*. Dexter, MI: Michigan Thompson-Shore Inc.

Jagers, P. (1975). *Branching processes with biological applications*. London: Wiley.

Johnson, N., Kotz, S., & Kemp, A. W. (1993). *Univariate discrete distributions* (2nd ed.). New York, NY: Wiley.

Johnson, N. P., & Mueller, J. (2002). Updating the accounts: Global mortality of the 1918–1920 "Spanish" influenza pandemic. *Bulletin of the History of Medicine, 76*(1), 105–115.

Kalbfleisch, J. G. (1985). *Probability and statistical inference, vol 2: Statistical inference* (2nd ed.) New York: Springer.

Kalbfleisch, J. D., & Lawless, J. F. (1989). Estimating the incubation time distribution and expected number of cases of transfusion-associated acquired immune deficiency syndrome. *Transfusion, 29*, 672–676.

Kalbfleisch, J. D., & Lawless, J. F. (1991). Regression models for right truncated data with applications to AIDS incubation times and reporting lags. *Statistica Sinica, 1*, 19–32.

Kalbfleisch, J. D., & Prentice, R. L. (2002). *Statistical analysis for failure time data* (2nd ed.). New York, NY: Wiley.

Karlin, S., & Taylor, H. M. (1975). *A first course in stochastic processes* (2nd ed.). Cambridge, MA: Academic Press.

Karlis, D., & Xekalaki, E. (2005). Mixed Poisson distributions. *International Statistical Review, 73*(1), 35–58.

Keeling, M. J., & Eames, K. T. (2005). Networks and epidemic models. *Journal of the Royal Society Interface, 2*(4), 295–307.

Keeling, M. J., & Rohani, P. (2008). *Modeling infectious diseases in humans and animals*. Princeton, NJ: Princeton University Press.

Kenah, E., Chao, D. L., Matrajt, L., Halloran, M. E., & Longini, I. M. Jr. (2011). The global transmission and control of influenza. *PLoS One, 6*(5), e19515.

Kenah, E., Lipsitch, M., & Robins, J. M. (2008). Generation interval contraction and epidemic data analysis. *Mathematical Biosciences, 213*, 71–79.

Kendall, D. (1956). Deterministic and stochastic epidemics in closed populations. In *Proceedings of Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 4, pp. 149–165). Berkeley, CA: University of California Press.

Kermack, W. O., & McKendrick, A. G. (1927). Contributions to the mathematical theory of epidemics, part I. *Proceedings of the Royal Society London A, 115*, 700–721.

Kiskowski, M. (2014). Three-scale network model for the early growth dynamics of 2014 West Africa Ebola epidemic. *PLOS Currents Outbreaks*. https://doi.org/10.1371/currents.outbreaks.b4690859d91684da963dc40e00f3da81

Kiskowski, M., & Chowell, G. (2015). Modeling household and community transmission of Ebola virus disease: Epidemic growth, spatial dynamics and insights for epidemic control. *Virulence, 7*(2), 63–73.

Klar, B. (2002). A note on the L-class of life distributions. *Journal of Applied Probability, 39*, 11–19.

Kosambi, D. D. (1949). Characteristic properties of series distributions. *Proceedings of the National Institute for Science, India, 15*, 109–113.

Krishnarajah, I., Cook, A., Marion, G., & Gibson, G. (2005). Novel moment closure approximations in stochastic epidemics. *Bulletin of Mathematical Biology, 67*, 855–873.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. New York, NY: Springer.

Lagakos, S. W., Rarraj, L. M., & De Gruttola, V. (1988). Nonparametric analysis of truncated survival data, with application to AIDS. *Biometrika, 75*(5), 15–23.

Lawless, J. F. (1994). Adjustments for reporting delays and the prediction of occurred but not reported events. *The Canadian Journal of Statistics, 22*(1), 15–31.

Lawless, J. F. (2003). *Statistical models and methods for lifetime data* (2nd ed.). New York, NY: Wiley.

Lee, J., Chowell, G., & Jung, E. (2016). A dynamic compartmental model for the Middle East respiratory syndrome outbreak in the Republic of Korea: A retrospective analysis on control interventions and superspreading events. *Journal of Theoretical Biology, 408*, 118–126.

Lefèvre, C., & Picard, P. (1995). Collective epidemic processes: A general modelling approach to the final outcome of SIR infectious diseases. In D. Mollison (Ed.), *Epidemic models: Their structure and relation to data* (pp. 53–70). Cambridge: Cambridge University Press.

Li, H., Han, D., Hou, Y., Chen, H., & Chen, Z. (2015). Statistical inference methods for two crossing survival curves: A comparison of methods. *PLoS One, 10*(1), e0116774. https://doi.org/10.1371/journal.pone.0116774

Li, M. Y., Muldowney, J. S., & van den Driessche, P. (1999). Global stability of SEIRS models in epidemiology. *Canadian Applied Mathematics Quarterly, 7*(4), 409–425.

Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73*, 13–22.

Liljeros, F., Edling, C. R., & Amaral, L. A. N. (2003). Sexual networks: Implications for the transmission of sexually transmitted infections. *Microbes and Infection, 3*, 189–196.

Lindsey, J. K. (2001). *Nonlinear models in medical statistics*. *Oxford statistical science series* (Vol. 24). Oxford: Oxford University Press.

Lipsitch, M., Cohen, T., Cooper, B., Robins, J. M., Ma, S., James, L., et al. (2003). Transmission dynamics and control of severe acute respiratory syndrome. *Science, 300*, 1966–1970.

Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E., & Getz, W. M. (2005). Superspreading and the effect of individual variation on disease emergence. *Nature, 438*(7066), 355.

Longini, I. M. Jr., & Koopman, J. S. (1982). Household and community transmission parameters from final distributions of infections in households. *Biometrics, 38*, 115–126.

Longini, I. M., Jr., Halloran, M. E., Nizam, A., Yang, Y., Xu, S., Burke, D. S., Cummings, D. A., & Epstein, J. M. (2007). Containing a large bioterrorist smallpox attack: A computer simulation approach. *International Journal of Infectious Diseases, 11*(2), 98–108.

Lord, D., & Geedipally, R. S. (2011). The negative binomial–Lindley distribution as a tool for analyzing crash data characterized by a large amount of zeros. *Accident Analysis & Prevention, 43*, 1738–1742.

Ludwig, D. (1975). Final size distributions for epidemics. *Mathematical Biosciences, 23*, 33–46.

Lui, K. J., Darrow, W. W., & Ruthford, G. W. (1988). A model-based estimate of the mean incubation period for AIDS in homosexual men. *Science, 240*, 1333–1335.

Lui, K. J., Lawrence, D. N., Morgan, W. M., Peterman, T. A., Haverkos, H. W., & Bregrnan, D. J. (1986). A model-based approach for estimating the mean incubation period of transfusion-associated acquired immunodeficiency syndrome. *Proceedings of the National Academy of Sciences, 83*, 3051–3055.

Lynch, J. (1988). Mixtures, generalized convexity and balayages. *Scandinavian Journal of Statistics, 15*, 203–210.

Ma, J., Dushoff, J., Bolker, B. M., & Earn, D. J. (2014). Estimating initial epidemic growth rates. *Bulletin of Mathematical Biology, 76*, 245–260.

Ma, J., & Earn, D. J. (2006). Generality of the final size formula for an epidemic of a newly invading infectious disease. *Bulletin of Mathematical Biology, 68*, 679–702.

Manfredi, P., & D'Onofrio, A. (Eds.). (2013). *Modeling the interplay between human behavior and the spread of infectious diseases*. New York, NY: Springer.

Marguta, R., & Parisi, A. (2015). Impact of human mobility on the periodicities and mechanisms underlying measles dynamics. *Journal of the Royal Society Interface, 12*(104), 20141317.

Marshall, A. W., & Olkin, I. (2007). *Life distributions, structure of nonparametric, semiparametric and parametric families*. New York, NY: Springer.

Martín, A. C., Derrough, T., Honomou, P., Kolie, N., Diallo, B., Konê, M., et al. (2016). Social and cultural factors behind community resistance during an Ebola outbreak in a village of the Guinean Forest region, February 2015: A field experience. *International Health, 8*, 227–229.

Martin-Löf, A. (1988). The final size of a nearly critical epidemic, and the first passage time of a Wienner process to a parabolic barrier. *Journal of Applied Probability, 35*, 671–682.

McCullagh, P., & Nelder, J. A. (1983). *Generalized linear models*. London: Chapman and Hall.

McKendrick, A. G. (1925). The applications of mathematics to medical problems. *Proceedings of Edinburgh Mathematical Society, 44*, 98–130.

Medley, G. F., Anderson, R. M., Cox, D. R., & Billard, I. (1987). Incubation period of AIDS in patients infected via blood transfusion. *Nature, 328*(7), 19–21.

Meeker, W. Q., & Escobar, L. A. (1998). *Statistical methods for reliability data*. *Wiley Series in probability and statistics*. New York, NY: Wiley.

Merler, S., Ajelli, M., Fumanelli, L., Gomes, M. F., Piontti, A. P., Rossi, L., et al. (2015). Spatiotemporal spread of the 2014 outbreak of Ebola virus disease in Liberia and the effectiveness of non-pharmaceutical interventions: A computational modelling analysis. *The Lancet Infectious Diseases, 15*(2), 204–211.

Mills, C. E., Robins, J. M., & Lipsitch, M. (2004). Transmissibility of 1918 pandemic influenza. *Nature, 432*(7019), 904.

Mode, C. J., & Sleeman, C. K. (2000). *Stochastic processes in epidemiology, HIV/AIDS, other infectious diseases and computers*. Singapore: World Scientific.

Murray, G. D., & Cliff, A. D. (1977). A stochastic model for measles epidemics in a multi-region setting. *Transactions of the Institute of British Geographers, 2*, 158–174.

Nåsell, I. (1995). The threshold concept in stochastic and endemic models. In D. Mollison (Ed.), *Epidemic models: Their structure and relation to data* (pp. 71–83). Cambridge: Cambridge University Press.

Nåsell, I. (2002). Stochastic models of some endemic infections. *Mathematical Biosciences, 179*, 1–19.

Nåsell, I. (2003). Moment closure and the stochastic logistic model. *Theoretical Population Biology, 63*(2), 159–168.

Nelson, K. E., Williams, C. M., & Graham, N. M. H. (2001). *Infectious disease epidemiology: Theory and practice*. Gaithersburg, MD: An Aspen Publication.

Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica, 16*, 1–32.

Nishiura, H. (2007). Early efforts in modeling the incubation period of infectious diseases with an acute course of illness. *Emerging Themes in Epidemiology, 4*, 2. https://doi.org/10.1186/1742-7622-4-2

Nishiura, H. (2010). Time variations in the generation time of an infectious diseases: Implications for sampling to appropriately quantify transmission potential. *Mathematical Biosciences & Engineering, 7*(4), 851–869.

Nishiura, H., & Chowell, G. (2009). The effective reproduction number as a prelude to statistical estimation of time-dependent epidemic trends. In G. Chowell, J. M. Hyman, L. M. A. Bettencourt, & C. Castillo-Chavez (Eds.), *Mathematical and statistical estimation approaches in epidemiology*. Dordrecht: Springer.

Nishiura, H., & Chowell, G. (2014). Feedback from modelling to surveillance of Ebola virus disease. *Eurosurveillance, 19*(37), pii=20908.

Nishiura, H., Yan, P., Sleeman, C. K., & Mode, C. J. (2012). Estimating the transmission potential of supercritical processes based on the final size distribution of minor outbreaks. *Journal of Theoretical Biology, 294*, 48–55.

Olu, O. O., Lamunu, M., Nanyunja, M., Dafae, F., Samba, T., Sempiira, N., et al. (2016). Contact tracing during an outbreak of Ebola virus disease in the western area districts of Sierra Leone: Lessons for future Ebola outbreak response. *Frontiers in Public Health, 4*, 130.

Pandey, A., Atkins, K. E., Medlock, J., Wenzel, N., Townsend, J. P., Childs, J. E., et al. (2014). Strategies for containing Ebola in West Africa. *Science, 346*(6212), 991–995.

Panjer, H. H., & Willmot, G. E. (1982). Recursions for compound distributions. *ASTIN Bulletin, 25*(1), 5–17.
Pearl, R. (1925). *The biology of population growth*. New York, NY: Knopf.
Pearl, R., & Reed, L. J. (1920). On the rate of growth of the population of the United States since 1790 and its mathematical representation. *Proceedings of the National Academy of Sciences of the United States of America, 6*, 275–288.
Pell, B., Kuang, Y. Viboud, C., & Chowell, G. (2018a). Using phenomenological models for forecasting the 2015 Ebola challenge. *Epidemics, 22*, 62–70.
Pell, B., Phan, T., Rutter, E. M., Chowell, G., & Kuang, Y. (2018b). Simple multi-scale modeling of the transmission dynamics of the 1905 plague epidemic in Bombay. *Mathematical Biosciences, 301*, 83–92.
Pellis, L., Ball, F., Bansal, S., Eames, K., House, T., Isham, V., et al. (2015). Eight challenges for network epidemic models. *Epidemics, 10*, 58–62.
Pellis, L., Ball, F., & Trapman, P. (2012). Reproduction numbers for epidemic models with households and other social structures. I. Definition and calculation of R0. *Mathematical Biosciences, 235*, 85–97.
Pellis, L., Ferguson, N. M., & Fraser, C. (2009). Threshold parameters for a model of epidemic spread among households and workplaces. *Journal of the Royal Society Interface, 6*, 979–987.
Perra, N., Balcan, D., Gonçalves, B., & Vespignani, A. (2011). Towards a characterization of behavior-disease models. *PloS One, 6*(8), e23084.
Pickles, W. (1939). *Epidemiology in country practice*. Bristol: John Wright and Sons.
Pinto, A., Martins, J., & Stollenwerk, N. (2009). The higher moments dynamic on SIS model. In T. E. Simos, et al. (Eds.), *Numerical Analysis and Applied Mathematics, AIP Conference Proceedings* (Vol. 1168, pp. 1527–1530). College Park, MD: AIP.
Qin, J. (2017). Biased sampling. In *Over-identified parameter problems and beyond* (ICSA book series in statistics). Springer Nature Singapore. https://doi.org/10.1007/978-981-10-4856-2_1.
Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmuller, U., et al. (2009). Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics, 25*, 1923–1929.
Richards, F. (1959). A flexible growth function for empirical use. *Journal of Experimental Botany, 10*, 290–301.
Rida, W. N. (1991). Asymptotic properties of some estimators for the infection rate in the general stochastic epidemic. *Journal of Royal Statistical Society (B), 53*, 209–283.
Riley, S. (2007). Large-scale spatial-transmission models of infectious disease. *Science, 316*(5829), 1298–1301.
Roberts, G. M., & Heesterbeek, J. A. P. (2007). Model-consistent estimation of the basic reproduction number from the incidence of an emerging infection. *Journal of Mathematical Biology, 55*, 803–816.
Rohatgi, A., (2018). WebPlotDigitizer Version: 4.1. Austin, TX.
Roosa, K., & Chowell, G. (2019). Assessing parameter identifiability in compartmental dynamic models using a computational approach: Application to infectious disease transmission models. *Theoretical Biology and Medical Modelling, 16*(1), 1.
Ross, R. (1911). *The prevention of malaria*. London: John Murray.
Ross, R. (1928). *Studies on malaria*. London: John Murray.
Ross, S. M. (1996). *Stochastic processes* (2nd ed.). New York, NY: Wiley.
Ross, S. M. (2019). *Introduction to probability models* (12th ed.). Cambridge, MA: Academic Press.
Rushton, S. P., & Mautner, A. (1955). The deterministic model of a simple epidemic for more than one community. *Biometrika, 42*, 126–132.
Sartwell, P. E. (1966). The incubation period and the dynamics of infectious disease. *American Journal of Epidemiology, 83*(2), 204–216.
Sattenspiel, L. (2009). *The geographic spread of infectious diseases: Models and applications*. Princeton, NJ: Princeton University Press.

Sattenspiel, L., & Dietz, K. (1995). A structured epidemic model incorporating geographic mobility among regions. *Mathematical Biosciences, 128*(1–2), 71–91.
Scalia-Tomba, G. (1985). Asymptotic final size distribution for some chain binomial processes. *Advances in Applied Probability, 17*, 477–495.
Schanzer, D. L., Langley, J. M., Dummer, T., Viboud, C., & Tam, T. W. (2010). A composite epidemic curve for seasonal influenza in Canada with an international comparison. *Influenza and Other Respiratory Viruses, 4*(5), 295–306.
Shaked, M. (1980). On mixtures from exponential families. *Journal of Royal Statistical Society B, 42*, 192–198.
Shaked, M., & Shanthikumar, J. G. (2007). *Stochastic orders*. New York, NY: Springer.
Shanafelt, D. W., Jones, G., Lima, M., Perrings, C., & Chowell, G. (2017). Forecasting the 2001 foot-and-mouth disease epidemic in the UK. *Ecohealth, 15*(2), 338–347.
Shrivastava, S. R., Shrivastava, P. S., & Ramasamy, J. (2014). Utility of contact tracing in reducing the magnitude of Ebola disease. *Germs, 4*(4), 97.
Simini, P., González, M. C., Maritan, A., & Barabási, A. L. (2012). A universal model for mobility and migration patterns. *Nature, 484*, 96–100.
Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika, 42*, 425–440.
Simon, L. J. (1962). An introduction to the negative binomial distribution and its applications. *Proceedings, Casuality Acturial Society* (No. 91, Vol. XLIX, Part 1).
Simonsen, L., Chowell, G., Andreasen, V., Gaffey, R., Barry, J., Olson, D., et al. (2018). A review of the 1918 herald pandemic wave: Importance for contemporary pandemic response strategies. *Annals of Epidemiology, 28*(5), 281–288.
Smirnova, A., & Chowell, G. (2017). A primer on stable parameter estimation and forecasting in epidemiology by a problem-oriented regularized least squares algorithm. *Infectious Disease Modeling, 2*(2), 268–275.
Smirnova, A., deCamp, L., & Chowell, G. (2017). Forecasting epidemics through nonparametric estimation of time-dependent transmission rates using the SEIR model. *Bulletin of Mathematical Biology*. https://doi.org/10.1007/s11538-017-0284-3
Smith, H. (2011). Distributed delay equations and the linear chain trick. In *An introduction to delay differential equations with applications to the life sciences*. *Texts in applied mathematics* (Vol. 57). New York, NY: Springer.
Sprott, D. A. (2000). *Statistical inference in science*. New York, NY: Springer.
Stoyan, D. (1983). *Comparison models for queues and other stochastic models*. New York, NY: Wiley.
Stumpf, M. Wiuf, C., & May, R. (2005). Subset of scale-free networks are not scale free: Sampling properties of networks. *Proceedings of the National Academy of Sciences, 102*, 4221–4224.
Svensson, A. A. (2007). A note on generation times in epidemic models. *Mathematical Biosciences, 208*, 300–311.
Szendroi, B., & Csányi, G. (2004). Polynomial epidemics and clustering in contact networks. *Proceedings of the Royal Society of London. Series B: Biological Sciences, 271*(Suppl. 5), S364–S366.
Tan, W. Y. (2000). *Stochastic modeling of AIDS epidemiology and HIV pathogenesis*. River Edge, NJ: World Scientific.
Tariq, A., Roosa, K., Mizumoto, K., & Chowell, G. (2019). Assessing reporting delays and the effective reproduction number: The 2018–19 Ebola epidemic in DRC, May 2018-January 2019. Epidemics. https://doi.org/10.1016/j.epidem.2019.01.003
The World Health Organization. (2014). *WHO declares end of Ebola outbreak in Nigeria*. World Health Organization Media Statement. Retrieved October 20.
The World Health Organization. (2016). Situation report for 29 July 2016. The 2016 Yellow fever epidemic in Angola. Available from: https://www.who.int/emergencies/yellow-fever/situation-reports/29-july-2016/en/
The World Health Organization Emergency Response Team. (2014). Ebola virus disease in West Africa - The first 9 months of the epidemic and forward projections. *New England Journal of Medicine, 371*(16), 1481–1495.

Tizzoni, M., Bajardi, P., Poletto, C., Ramasco, J. J., Balcan, D., Goncalves, B., et al. (2012). Real-time numerical forecast of global epidemic spreading: Case study of 2009 A/H1N1pdm. *BMC Medicine, 10*, 165.

Towers, S., Brauer, F., Castillo-Chavez, C., Falconar, A. K., Mubayi, A., & Romero-Vivas, C. M. (2016). Estimate of the reproduction number of the 2015 Zika virus outbreak in Barranquilla, Colombia, and estimation of the relative role of sexual transmission. *Epidemics, 17*, 50–55.

Tuite, A. R., Greer, A. L., Whelan, M., Winter, A. L., Yan, P., Wu, J, et al. (2010). Estimated epidemiologic parameters and morbidity associated with pandemic H1N1 influenza. *Canadian Medical Association Journal, 182*, 131–136.

Turner, M. E. Jr., Bradley, E. L. Jr., Kirk, K., & Pruitt, K. M. (1976). A theory of growth. *Mathematical Biosciences, 29*, 367–373.

Valleron, A. J., Cori, A., Valtat, S., Meurisse, S., Carrat, F., & Boelle, P. Y. (2010). Transmissibility and geographic spread of the 1889 influenza pandemic. *Proceedings of the National Academy of Sciences, 107*(19), 8778–8781.

van den Driessche, P. (2017). Reproduction numbers of infectious disease models. *Infectious Disease Modelling, 2*(3), 288–303.

van den Driessche, P., & Watmough, J. (2002). Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. *Mathematical Biosciences, 180*, 29–48.

van den Driessche, P., & Watmough, J. (2008). Further notes on the basic reproduction number. In: F. Brauer, P. van den Driessche, & J. Wu (Eds.) *Mathematical epidemiology*. *Lecture notes in mathematics* (Vol. 1945). Berlin: Springer.

Varia, M., Wilson, S., Sarwal, S., McGeer, A., Gournis, E., & Galanis, E. (2003). Investigation of a nosocomial outbreak of severe acute respiratory syndrome (SARS) in Toronto, Canada. *Canadian Medical Association Journal, 169*(4), 285–292.

Verhulst, P. J. (1838). Notice sur la loi que la population suit dan sons accroissement. *Correspondance mathématique et physique, 10*, 113–121.

Viboud, C., Bjornstad, O. N., Smith, D. L., Simonsen, L., Miller, M. A., & Grenfell, B. T. (2006). Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science, 312*(5772), 447–451.

Viboud, C., Simonsen, L., Chowell, G. (2016). A generalized-growth model to characterize the early ascending phase of infectious disease outbreaks. *Epidemics 15*, 27–37.

Viboud, C., Sun, K., Gaffey, R., Ajelli, M., Fumanelli, L., Merler, S., et al. (2018). The RAPID ebola forecasting challenge: Synthesis and lessons learnt. *Epidemics, 22*, 13–21.

Vincenot, C. E., & Moriya, K. (2011). Impact of the topology of metapopulations on the resurgence of epidemics rendered by a new multiscale hybrid modeling approach. *Ecological Informatics, 6*, 177–186.

von Bahr, B., & Martin-Löf, A. (1980). Threshold limit theorems for some epidemic processes. *Advances in Applied Probability, 12*, 319–349.

Vynnycky, E., & White, R. (2010). *An introduction to infectious disease modelling*. Oxford: Oxford University Press.

Wald, A. (1943). *A method of estimating plane vulnerability based on damage pf survivors*. Statistical Research Group, Columbia University. CRC (Vol. 432). Arlington County, VA: Center for Naval Analyses.

Walker, S. & Stephens, D. (1999). A multivariate family of distributions on $(0, \infty)^p$. *Biometrika, 86*(3), 703–709.

Wallinga, J., & Lipsitch, M. (2007). How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of Royal Society B, 274*, 599–604.

Wang, M. C. (2005) Length bias. In *Encyclopedia of biostatistics*. New York, NY: Wiley. https://doi.org/10.1002/0470011815.b2a11044

Wang, X. S., Wu, J., & Yang, Y. (2012). Richards model revisited: Validation by and application to infection dynamics. *Journal of Theoretical Biology, 313*, 12–19.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature, 393*(6684), 440.

Waugh, W. A. O'N. (1958). Conditioned Markov processes. *Biometrika, 45*(1–2), 241–249.
Wearing, H. J., Rohani, P., & Keeling, M. J. (2005). Appropriate models from the management of infectious diseases. *PLoS Medicine, 7*, 621–627.
Weinberger, D. M., Krause, T. G., Molbak, K., Cliff, A., Briem, H., Viboud, C., et al. (2012). Influenza epidemics in Iceland over 9 decades: Changes in timing and synchrony with the United States and Europe. *American Journal of Epidemiology, 176*(7), 649–655.
White, L. F., & Pagano, M. (2008). A likelihood-based method for real-time estimation of the serial interval and reproductive number of an epidemic. *Statistics in Medicine, 27*, 2999–3016.
Willmot, G. E. (1993). On recursive evaluation of mixed Poisson probabilities and related quantities. *Scandnavian Acturarial Journal, 2*, 114–133.
Wilson, E. B., & Worcester, J. (1945). The spread of an epidemic. *Proceedings of the National Academy of Sciences of the United States of America, 31*, 327–333.
Xekalaki, E. (1983). The univariate generalized Waring distribution in relation to accident theory: Proneness, spells or contagion? *Biometrics, 39*(3), 887–895.
Xekalaki, E. (2014). On the distribution theory of over-dispersion. *Journal of Statistical Distributions and Applications, 1*, 19.
Xekalaki, E., & Zografi, M. (2008). The generalized waring process and its application. *Communications in Statistics—Theory and Methods, 37*(12), 1835–1854.
Xekalaki, E. A., & Panaretos, J. (2006). On the association of the Pareto and the Yule distribution. *Theory of Probability and Its Applications, 33*(1), 191–195.
Xia, Y. C., Bjørnstad, O. N., & Grenfell, B. T. (2004). Measles metapopulation dynamics: A gravity model for epidemiological coupling and dynamics. *The American Naturalist, 164*, 267–281.
Xiao, Y., Zhou, Y., & Tang, S. (2011). Modelling disease spread in dispersal networks at two levels. *Mathematical Medicine and Biology: A Journal of the IMA, 28*, 227–244.
Yan, P. (2008a). Distribution theory, stochastic processes and infectious disease modelling. In F. Brauer, P. van den Driessche, & J. Wu (Eds.), *Mathematical epidemiology*. *Lecture notes in mathematics* (Vol. 1945). Berlin: Springer.
Yan, P. (2008b). Separate roles of the latent and infectious periods in shaping the relation between the basic reproduction number and the intrinsic growth rate of infectious disease outbreaks. *Journal of Theoretical Biology, 251*, 238–252.
Yan, P. (2018). A frailty model for intervention effectiveness against disease transmission when implemented with unobservable heterogeneity. *Mathematical Biosciences & Engineering, 15*(1), 275–298.
Yan, P., & Feng, Z. (2010). Variability order of the latent and the infectious periods in a deterministic SEIR epidemic model and evaluation of control effectiveness. *Mathematical Biosciences, 224*, 43–52.
Yan, P., & Zhang, F. (2018). A case study of nonlinear programming approach for repeated testing of HIV in a population stratified by subpopulations according to different risks of new infections. *Operations Research for Health Care*. https://doi.org/10.1016/j.orhc.2018.03.007
Yan, P., Zhang, F., & Wand, H. (2011). Using HIV diagnostic data to estimate HIV incidence: Method and simulation. *Statistical Communications in Infectious Diseases, 3*, 1.
Yule, G. U. (1925). The growth of population and the factor which controls it. *Journal of the Royal Statistical Society: Series A, 88*, 1–58.
Zhang, Q., Sun, K., Chinazzi, M., y Piontti, A. P., Dean, N. E., Rojas, D. P., et al. (2017). Spread of Zika virus in the Americas. *Proceedings of the National Academy of Sciences, 114*(22), E4334–E4343.
Zipf, G. K. (1949). *Human behaviour and the principle of least effort*. Cambridge, MA: Addison-Wesley.

Index

P. Yan, G. Chowell, *Quantitative Methods for Investigating Infectious Disease Outbreaks*, Texts in Applied Mathematics 70,
https://doi.org/10.1007/978-3-030-21923-9