From Puzzles and Paradoxes to Concepts in Probability

Manfred Borovcnik and Ramesh Kapadia

Abstract This chapter focuses on how puzzles and paradoxes in probability developed into mathematical concepts. After an introduction to background ideas, we present each paradox, discuss why it is paradoxical, and give a normative solution as well as links to further ideas and teaching; a similar approach is taken to puzzles. After discussing the role of paradoxes, the paradoxes are grouped in topics: equal likelihood, expectation, relative frequencies, and personal probabilities. These cover the usual approaches of the a priori theory (APT), the frequentist theory (FQT), and the subjectivist theory (SJT). From our discussion it should become clear that a restriction to only one philosophical position towards probability—either objectivist or subjectivist—restricts understanding and fails to develop good applications. A section on the central mathematical ideas of probability is included to give an overview for educators to plan a coherent and consistent probability curriculum and conclusions are drawn.

1 How Paradoxes Highlight Conceptual Conflicts

What makes a paradox? Progress in the development of mathematical concepts is accompanied by controversies, ruptures, and new beginnings. The struggle for truth reveals interesting breaks highlighted by paradoxes that mark a situation, which reflects a contradiction to the current base of knowledge. Yet, there is an opportunity to renew the basis and proceed to wider concepts, which can embrace and dissolve the paradox. A puzzle, however, is a situation in which the current concept yields a solution that seems intuitively unacceptable. Such a puzzle shows that the intuitive basis of the concept has to be improved or that the concept is contrary to the expectation of the solver. Examples from other areas of mathematics include negative numbers (for younger children) or complex numbers (for most adults). From puzzles and paradoxes one can learn about crucial properties of the theory involved. The situations are challenging and can also lead experts to err; the purpose of the

M. Borovcnik (✉) · R. Kapadia
Institut für Statistik, Alpen-Adria Universität Klagenfurt, Universitätsstraße 65, 9020 Klagenfurt, Austria
e-mail: manfred.borovcnik@uni-klu.ac.at

E.J. Chernoff, B. Sriraman (eds.), *Probabilistic Thinking*,
Advances in Mathematics Education, DOI 10.1007/978-94-007-7155-0_3,

concepts can also be better understood than by a sequential exposition of theory and examples.

Székely (1986) is the standard exposition which classifies puzzles and paradoxes and highlights the crucial contradictions that have contributed to clarify the basis of probability and mathematical statistics. He discusses each paradox in five parts: history, its formulation, explanation, remarks, and references: the reader is encouraged to study his approach for an extensive range of puzzles and paradoxes, which include some presented in this chapter.

Our approach is to link more closely to the underlying ideas and their application in teaching and learning. This vision is directed at the multi-faceted concepts entrenched with philosophical principles, especially with the two grand schools and conceptions of probability which are linked to the frequentist and Bayesian (subjectivist) interpretation of probability. Our exposition also encompasses the personal thoughts of pupils and students who are encountering these ideas for the first time. The underlying ideas are complex, overlapping, and interwoven. Steps to explain and overcome the traps are needed. For learners, the concepts are still emerging and change their character. This affects the subject of probability, which itself affects other areas of study. Think, for example, of the drastic change of the paradigm in physics which currently has changed completely from a causal to a random ideal.

A major purpose is to explain what is paradoxical and to link the solution to mathematical concepts and their historical perception. We assert that the present dominant position of objectivist probability narrows the flexibility not only of the models but also of the conceptions of the learner with the consequence that the exclusion of subjectivist notions hinders understanding.

We discuss 15 paradoxes and puzzles, only five of these feature in Székely (1986). The paradoxes are grouped alongside mathematical concepts which sometimes deviate from the historical development. Equal likelihood is followed by a review of the principle of insufficient reason, which regulates how and when equally likely cases are present (Sect. 2). A discourse on expected values follows to cover the central competing idea to probability since olden times (Sect. 3). The frequentist conception of probability has emerged as almost the sole interpretation, despite some key puzzles on randomness (Sect. 4). The concept of conditional probability is connected to subjectivist interpretations of probability (Sect. 5). We present mathematical theory from school to university (Sect. 6), although in a curtailed format. While the paradoxes highlight isolated developments in concepts, this section is intended to reveal central ideas that *link* the concepts coherently. This requires more technical and mathematical detail, which sometimes is prone to be avoided. However, a deeper understanding of the fundamental ideas is vital and may get lost if one strives to simplify the mathematics too much. The final section (Sect. 7) presents our conclusions.

Our presentation shows that the mathematical context of the concepts has to be accompanied by philosophical aspects, otherwise a comprehension of the theory will be biased, resulting in difficulties not only to understand but—more importantly—for learners to *accept* the concepts and apply them sensibly.

2 Equal Likelihood

The classical a priori theory (APT) starts with an assumption of equal likelihood, often in games of chance, and uses combinatorial methods to find the probability of various events. It is still, rightly, the initial approach to introduce probability to children. These ideas originally arose from a variety of puzzles and paradoxes.

Combinatorial multiplicity is linked to the possible outcomes which are considered to be equally likely. Moreover, links are soon made to relative frequencies as otherwise the concept of probability would lack an orientation about what will happen in repeated experiments. The emergence of these ideas is interwoven. To investigate equally likely cases and to apply a rule of favourable to possible cases draws heavily on counting the possibilities correctly, which proves to be harder than one might imagine, as evident from the difficulties shown by children and students. The conceptual confusion was aggravated by a competing concept, the expected value: at times, some like Huygens (1657) regarded it as more basic than probability with his value of an enterprise, corresponding to today's expected value. This was shortly after the great leap forward by Pascal and Fermat in 1654 (Fermat and Pascal 1679) when they specified a suitable set of outcomes and counted the possibilities correctly. An explicit, if contested definition of probability had to wait till Laplace (1812/1951). His approach is characterized by an intermixture of sample space—the mathematical part—and the intuitive part of an idea of symmetry. Probability is defined by assuming the equal likelihood of all possible results (APT).

In modern theory, the sample space is separated from probability, which is a function defined on a specific class of subsets of the sample space. This gives the freedom to view any specific probability as a model for a real situation. Until this final step to separate the levels of the model and the real problem, probability was considered as a property of the real world like length or weight. Thus—within bounds of measurement errors—there is *one* probability, a unique value for a problem. Accordingly, the task of a probabilist was either to find a sample space suitable to fit the equally likely conditions, or, since the empiricism of the nineteenth century, to find a suitable random experiment, repeat it often enough, and substitute the unknown probability by the relative frequency (FQT). Laplace recognized the difficulty related to judging cases to be equally likely and modified a principle going back to Jakob Bernoulli, which was later re-evaluated by Bayes: If one is ignorant of the ways an experiment ends up and there is no reason to believe that one case will occur preferentially compared to another, the cases *are* equally likely. This *principle of insufficient reason* underpins the application of Laplace's probability.

The first subsection below includes the historically famous puzzles arising from the struggle to sharpen the conception of possible cases, the rule of favourable to unfavourable cases, and expected value. Subsequently, problems arise from the principle of Laplace about equally likely cases. These provide excellent and motivating starting points to introduce probability in the classroom.

2.1 Early Notions of Probability

When concepts emerge and are not yet well-defined, confusion between closely related terms occurs. This may be confirmed by the early endeavours to find tools to describe and solve problems with uncertainty. As no other embodiments of the pre-concepts were available, it is no wonder that the context of games of chance was used extensively. Furthermore, the old idea of fairness and its close connection to games of chance was used as a "model" to a situation that had no relation to chance before.

The first puzzle, 9 or 10, deals with the sum of three dice. It marks a definite step towards the Cartesian product for counting the possibilities of repeated experiments, thus viewing the result 1.2.6 as different from 2.1.6, which gives a total multiplicity of 6 instead of 1. Lacking a theoretical argument for respecting order, the choice was justified by empirical frequencies. De Méré's problem with sixes marks a step in clarifying what can count as possibility. The confusion about the rule involved might be traced back to an overlap between the concepts of probability and expected value. The third puzzle is remarkable insofar as a counter-intuitive step was needed to sharpen the concept of possible cases: the solution is based on *hypothetical* cases, which are extended against the rules and are thus *impossible*. However, the greatest progress by Pascal and Fermat was to model a situation without a link to probabilities by a hypothetical game of chance to mimic the progress of a competition and use the resulting relative winning probabilities for a fair division of the stakes.

***P_1*: Problem of the Grand Duke of Tuscany** Three dice are thrown. The possibilities to get a sum of 9 or 10 are counted in the following way (Galilei 1613–1623; cited from David 1962, p. 192):

> [...] 9 and 10 can be made up by an equal diversity of numbers (and this is also true of 12 and 11): since 9 is made up of 1.2.6, 1.3.5, 1.4.4, 2.2.5, 2.3.4, 3.3.3, which are six triple numbers, and 10 of 1.3.6, 1.4.5, 2.2.6, 2.3.5, 2.4.4, 3.3.4, and in no other ways, and these also are six combinations.

This theoretical argument is confronted with experience:

> Nevertheless, although 9 and 12 can be made up in as many ways as 10 and 11 respectively, and therefore they should be considered as being of equal utility to these, yet it is known that long observation has made dice-players consider 10 and 11 to be more advantageous than 9 and 12.

Galilei ordered the results of the ways of getting 9 as follows: there are 6 different orderings of 1.2.6 but only 3 out of 2.2.5 and only one from 3.3.3. His table is well worth reproducing and studying by pupils (see Table 1), as its structure conveys a hierarchical process of ordering, first the results of the dice and then ordering them using the symmetry that a sum of 3 and 18 has the same multiplicity, as have 4 and 17, up to 10 and 11.

What is the Paradox? Two counting procedures lead to different numbers and yield different probabilities. As the probabilities were communicated in odds, the

Table 1 Galilei's protocol (from David 1962, p. 194)—slightly modified

10		9		8		7		6		5		4		3		
6.3.1	6	6.2.1	6	6.1.1	3	5.1.1	3	4.1.1	3	3.1.1	3	2.1.1	3	1.1.1	1	
6.2.2	3	5.3.1	6	5.2.1	6	4.2.1	6	3.2.1	6	2.2.1	3					
5.4.1	6	5.2.2	3	4.3.1	6	3.3.1	3	2.2.2	1							
5.3.2	6	4.4.1	3	4.2.2	3	3.2.2	3									
4.4.2	3	4.3.2	6	3.3.2	3											
4.3.3	3	3.3.3	1													
	27		25		21		15		10		6		3		1	108
																108
																216

first count leads to 1: 1 while the second yields 25: 27 for 9 against 10, which is advocated as correct by Galilei; today, we read the result as a probability of 0.1157 for 9 and of 0.1250 for 10. With the modern concept of repeated experiments, *independence* is a theoretical argument in favour of the second way to count. Without such a concept, the result was certainly puzzling. Another strange feature is how they could find such a small difference by playing. For Székely (1986, p. 3) the paradoxical feature of 9 or 10 lies in the fact that for two dice 9 is more probable while for three dice 10 is more probable.

Further Ideas The argument to find which solution is right is interesting. It signifies that since olden times counting the possibilities in order to calculate the relative probabilities was linked to what actually happens in games. As the difference in probabilities for 9 and 10 is very small (0.0093), it is hard to believe that this has really been detected by playing as this would require roughly 10,000 trials.

Such an argument as a substitute for theoretical reasoning is used at several places in the history of probability. It signifies that the theoretical argument alone was too weak to convince and that the writers considered a strong connection from their chances to relative frequencies. If an argument for a way to count contradicts the experience of relative frequencies, it is useless.

From a teaching perspective, there are valuable lessons which can be drawn and used. It is difficult for children to find all the possibilities in throwing three dice and then calculate the probabilities. Nevertheless, the approach here can help to develop combinatorial skills. It also confronts the difference between a theoretical (APT) probability and a frequentist (FQT) interpretation.

***P_2*: De Méré's Problem** In this famous problem, a simple proportional argument suggests that it is equally likely to get (at least) a six in throwing a die four times as to get (at least) a double six in throwing two dice in 24 trials. De Méré posed the problem to Fermat as to why, apocryphally, he won a fortune betting on a six with one die and lost it betting on a double six in 24 trials (cf. David 1962, p. 235). Fermat listed the cases correctly and calculated the winning probabilities as $1 - (5/6)^4 = 671/1296 = 0.518$ for the six and as $1 - (35/36)^{24} = 0.491$ for the double six game. He concluded that, in fact, to bet on the single six game is favourable (higher than $1/2$) and betting on the double six game is unfavourable (lower than $1/2$).

What is the Paradox? In evaluating the chances, the following rule was emerging but far from clear: compare the number of favourable to unfavourable cases, or determine the ratio of favourable to possible cases. In a careless application of the emergent rule of "favourable to possible", the argument might have been as follows: with one die, 4 throws make 4 chances (i.e. 4 favourable cases to get a six). The 6 faces of the die mark the 6 possible cases. The ratio of favourable to possible yields $4/6$. With two dice, the 24 throws establish 24 chances (favourable cases to get a double six) with 36 possible cases and the same rule yields $24/36$. As the ratios are equal, the probabilities should be equal.

This argument is confronted by data in actually playing, whereby the game with one die is favourable while with two dice it is unfavourable. The line between favourable and unfavourable was drawn by the winning probability of $1/2$. If probabilities are linked to relative frequencies then there is definitely a problem with the original solution, and it is difficult to see what is wrong with counting the cases.

Another paradoxical feature lies in the difference between the concepts of probability and expected value, which are often confused in the discussion. The expected number of sixes in four trials with one die equals to $4 \cdot \frac{1}{6} = \frac{4}{6}$ and for double sixes with two dice it equals $24 \cdot \frac{1}{36} = \frac{24}{36}$. In this respect, a correct application of expected value leads to the same result for both games and the question is why this fails to predict the relative frequencies in games.

Further Ideas The classical random experiment is an independent binary 0–1 experiment with probability of p for the result 1. The expected value for one trial is p, for n repetitions it is $n \cdot p$. The reader may note that this yields another rule of favourable (the favourable cases are the n trials) to possible cases (with equal chances $1/p$ is the same as the number of all possible cases):

Expected value $= n \cdot p = n \cdot \frac{1}{1/p}$, which equals the fraction of n chances *to* $(1/p)$ possible cases.

Though the rules are identical they bear a different meaning, which can be recognized only if the difference between the concepts of probability and expectation is discussed in teaching.

This overlap may also be traced in Huygens' method to derive probabilities by calculating the corresponding expected values. However, in de Méré's problem, the two solutions differ. With the games above, the amount to win is 1 if the winning figure occurs (and 0 else); note that the payment is the same irrespective of the actual count and so the double six game is unfavourable. If the amount to win were exactly the *number* of sixes (double sixes), the games, in fact, would be equal. Yet, the double six game has a greater variance and bears more risk to lose but also gives more chances to win a higher amount.

Historically it was difficult to separate the concepts of probability and expectation and arrive at a clear vision of what probability can achieve and how to interpret or evaluate specific probabilities. The situation is usually blurred by the fact that an outcome is related to an impact, especially in games of chance. The perceived impact differs if it is a win or loss—even if the expected amounts are the same, as experiments by Kahneman and Tversky (1979) have shown.

P_3: Division of Stakes

> A and B are playing a fair game of balla. They agree to continue until one has won six rounds. The game actually stops when A has won four and B three. How should the stakes be divided?

Pacioli suggested the stakes should be split as 4 to 3 (Pacioli 1494; see David 1962, p. 37). We changed his original data to the situation Pascal and Fermat deal with in their famous exchange of letters 1654. They assess the possible ways to win the whole series:

> Since A needs 2 points and B needs 3 points the game will be decided in a maximum of four throws. The possibilities are: [see Table 2]. In this enumeration, every case where A has 2, 3, or 4 successes is a case favourable to A, and every case where B has 3 or 4 [successes] represents a case favourable to B. There are 11 for A and 5 for B, so that the odds are 11: 5 in favour of A. (Pascal in Fermat and Pascal 1679, referenced by David 1962, p. 91)

What is the Paradox? The solution is paradoxical from the standpoint of counting the *possibilities*, rather than dividing the stake by the current score of 4:3. What may be viewed as a possibility as the game has been interrupted? A great step forward is marked by introducing a hypothetical continuation of the game on the basis of what could happen if the game is continued. Of course, the series is decided if A wins two more games in this scenario. Thus, there are only 10 actual possibilities and 6 are favourable to A, which would split the stakes as 6 to 4 or 3 to 2. At this point, it is important that Fermat recognized that it makes no sense to consider these “real” cases as equally likely. To assign equal weights to them, he extended the “real” cases by imagined further rounds to *make* them of equal length. Interestingly, this conflicts with the rules of the game as one of the players would already have won and the series finished. To introduce a hypothetical continuation was the first step

Table 2 Pascal’s hypothetical cases and real cases compared

[Counting hypothetical cases by Pascal					Counting “real” cases]				
AAAA	*AAAB*	*AABB*	*ABBB*	*BBBB*	*AA*	*ABA*	*BAA*	*BBAA*	*BBB*
	AABA	*ABAB*	*BABB*			*ABBA*	*BABA*	*BBAB*	
	ABAA	*BAAB*	*BBAB*			*ABBB*	*BABB*		
	BAAA	*ABBA*	*BBBA*						
		BABA							
		BBAA							

to *find* possibilities, to introduce an extension beyond the rules was to *invent* cases, with the aim to make them equally likely.

There are no signs that the stakes puzzle has anything to do with chance. The basic paradox is, however, Pascal and Fermat's view of the situation *as if* it were random. Originally the problem seems to be devoid of probability. The series has been interrupted and it is a problem awaiting a resolution. Probabilities are introduced only in the sense of a scenario. The focus lies on dividing the stakes fairly instead of finding a good model to describe the continuation of the game.

Further Ideas Székely (1986, p. 10) exclaims that "this [...] is considered [...] to be the birth of probability theory [...]". In earlier times, (fair) decisions were made by a game of chance if sufficient knowledge about the situation, or expertise (or trust) was missing. Somehow, probability introduces a sort of higher capacity (knowledge) beyond god (according to Laplace) who can always predict the result. Instead of exploring something like god's decision by a chancy game, the probability model is utilized to advocate a split of the stakes as fair. For teaching purposes, such a theoretical extension requires careful discussion in the classroom; it is instructive to explore the difference between a current proportion (4:3 here) and a fair division (11:5).

2.2 Conceptual Developments in Probability

The use of probability in the eighteenth and nineteenth centuries is signified by a diversity of conceptions. Two fundamental theorems concerned the relation between relative frequencies and probabilities: Bernoulli's law of large numbers (1713/1987) and Bayes' theorem including a corollary (1763).

Bernoulli's *direct probability* approach used the unknown probability p as a constant and the binomial distribution to derive the convergence of the relative frequencies towards this probability. With his theorem, Bernoulli provided the basis for relative frequencies (FQT) as an input to evaluate the probability of arguments. However, he was well aware that one might need more data to reach a reliable result.

Bayes' *inverse probability* method took the opposite approach: the weights on the unknown p converge to the relative frequencies of repeated experiments. Here, the unknown probability p is different from a constant (but unknown) number: in fact, one has to express a distribution upon it, which represents the status of knowledge on this parameter.

Bayes (1763) derived his theorem within an embodiment, in which the uniform prior distribution on p was obvious. Furthermore, he argued that if one lacks any knowledge about the value of the probability p of an event then one should accept equal stakes in betting on $0, 1, \ldots, n$ events in n repeated trials. On this assumption, he *derived* the uniform distribution on p *mathematically*. His argument was quite complex so that following writers abbreviated it to "if one lacks any knowledge about p then it is uniformly distributed" (similar to an older argument by

Bernoulli), which has become known as Bayes' postulate. Bayes' rule of succession is a corollary to his theorem and proves—on the basis of a uniform distribution on the parameter—that the posterior weights (distribution) on p after k successes in n trials have an expected value of $(k+1)/(n+2)$ and a variance that converges to zero (in modern terms he derived a beta distribution for the parameter p). Thus, the posterior distribution restricts itself to a point, which corresponds to the limit of the relative frequencies. Instead of using variance, Bayes calculated the probability of intervals around $(k+1)/(n+2)$, which was correctly interpreted as the probability of an event occurring in the next trial to follow (Price 1764–1765).

De Moivre gave a specification of probability as the number of favourable divided by the number of possible cases based on equally likely cases, which comes close to the generally accepted definition by Laplace (1812/1951). Laplace reproduced Bayes' theorem and used his "postulate" extensively to check and justify whether equally likely cases are appropriate in a problem. His approach ("if we are equally undecided about") was later named as the principle of insufficient reason. This was rejected by the empirical critique of Venn (1888). There is no way to transform complete ignorance into probabilities, which represent a form of knowledge. Venn asked for an *empirical* basis of probability as an idealized relative frequency. The difficulty of the principle of insufficient reason, and of independence is highlighted by the following problem where SJT is contrasted to APT.

***P*₄: D'Alembert's Problem** Two coins are flipped. What is the probability to obtain heads twice?

(a) Applying Laplace's principle on the combinatorial product space of *HH*, *HT*, *TH*, *TT* (thus respecting order) yields the answer 1/4 (APT).
(b) D'Alembert (1754) refers to the fundamental probability set {*no head*, *one head*, *two heads*} (neglecting order) and applied equi-probability to the three cases, giving an answer of 1/3 (will be linked to SJT below).

Since Pascal and Fermat it had been well acknowledged that, for repeated experiments, respecting order helps to find equally likely cases. Therefore, d'Alembert's approach was rejected as mere error illustrating how experts can err with probability (see Székely 1986, p. 3, or Maistrov 1974, p. 123).

What is the Paradox? The problem is a paradox in the history of probability as it has been overlooked that d'Alembert's solution *is* correct if only the *same* principle of insufficient reason is applied. This principle uses the uniform prior distribution on the unknown probability p of the coin to land heads up. With Bayes' rule of succession (a correct mathematical theorem) and the multiplication rule (a correct theorem) the following holds.

Before any data ($k=0$, $n=0$), one can conclude that $P(H)=\frac{0+1}{0+2}=\frac{1}{2}$; after seeing $H(k=1, n=1)$, the "conditional" probability to see heads again is $P(H|H)=\frac{1+1}{1+2}=\frac{2}{3}$. This yields the following probabilities for the three basic cases:

$$P(\textit{two heads}) = P(H)\cdot P(H|H) = \frac{1}{2}\cdot\frac{2}{3} = \frac{1}{3},$$

$$P(no\ head) = P(T) \cdot P(T|T) = \frac{1}{2} \cdot \frac{2}{3} = \frac{1}{3},$$

and for the last it has to be

$$P(head\ tail\ mixed) = 1 - \frac{2}{3} = \frac{1}{3}.$$

The paradox leads to two probabilities (one linked to classical and the other to Bayes' conception of probability), which are irreconcilable.

Further Ideas There can be a strong justification to neglect the order of the results. D'Alembert's argument is theoretically consistent but neglects long-term experimental data on coins, which strongly supports the independence of the throws and corroborates the equi-probability of the four cases with the order distinguished. For a biased coin with probability p for head, the result would be p^2, $p(1-p)$, $(1-p)p$, $(1-p)^2$.

The underlying ideas are certainly worthy of careful discussion in the classroom as, in modern physics, the description of various particles in real physical systems can be described by d'Alembert's approach: photons, nuclei and atoms containing an even number of elementary particles are essentially indistinguishable and may be best described by the Bose–Einstein statistics, which *neglects* order. It is quite startling that the world is found to work in this way, experimentally.

3 Expectation

Historically, the concepts of probability and expected value developed in parallel with overlap and confusion. The first formalization by Huygens (1657) used the expected value in a financial framework based on a situation of implicitly equally likely cases. While the value of an enterprise is unambiguous, probability has been laden with personal conceptions and philosophical difficulties. Of course, both concepts are closely connected and Huygens used this economic value to calculate probabilities in problems, which were discussed at the time. The shift to probabilities was completed by Jakob Bernoulli (1713/1987) with his efficient combinatorial methods, which were faster than the rather lengthy recursive approach of Huygens.

Probability at the same time became strongly connected to relative frequencies by the law of large numbers by Bernoulli. As probability took the lead and expected value became a derived concept, the latter was engulfed by the philosophical "burden" of probability. Already in the publication of Huygens' treatise, the wording *expected value* appeared but this was mainly due to a bad translation to Latin and missed Huygens' intention. Such a change in terminology shifted away from the original frugal meaning of an economic exchange price between risk and certainty to wishes, desires, and similar vague conceptions.

Expected value lacked a strong connection to relative frequencies even if today it is motivated as the average amount paid after a long series of random experiments.

It is important to remember that expected value plays a basic role for the subjectivist position as probabilities get a wider interpretation, which integrates relative frequencies (if available) *and* qualitative knowledge beyond that. By such a weighting process, the subjectively accepted equivalence of a price (an expected value) in a simple 0–1 bet with probability p for 1 "measures" the personal probability of an individual.

3.1 Expectation and Probability

From a modern perspective, it has been forgotten that one may base the whole theory of probability either on the concept of probability or on the concept of expected value (despite the fact that there are mathematical approaches to reconstruct or replace Kolmogorov's axiomatic theory on the basis of expectation). It is no wonder that the two concepts were intertwined in the early stages. Huygens marks a special point in history—a temporary shift away from probability to expectation. He defined the term in an *economic* context as a price one would accept to switch between an uncertain (risky) situation and a situation without uncertainty—which resembles the features of taking out an insurance policy and amounts to a basic paradigm in decision theory.

The fundamental concept for Huygens is his value of an enterprise. Based on equal cases, he states

> to have p chances of obtaining a and q of obtaining b, chances being equal, is worth $\frac{pa+qb}{p+q}$.
> (Huygens 1657, cited from David 1962, p. 116)

Huygens circumvents the need to calculate probabilities or proportions; instead he solves the problems by his economic approach. The wording expected value is unfortunate as it associates hope, fear, and many other emotions related to the potential outcomes while Huygens used the term as a purely financial concept. In an analogy to determine the net present value of a future amount by a discount rate in financial mathematics, the present value of an uncertain enterprise equals the various amounts to gain or lose, *discounted* by their chances. Such a value is vital for any insurance policy. Future potential amounts have to be discounted to a value that is paid today.

The St Petersburg paradox unexpectedly (no pun intended) produced an infinite expected value, which is absurd if the concept is interpreted as an economic notion. The startling situation was resolved by amendments to probabilities, which are still disputed.

***P_5*: St Petersburg Paradox** Two players A and B toss a coin until it shows 'head' for the first time. If this occurs at the nth trial, then player B pays £ 2^{n-1} to player A. What amount should A pay to B before this game starts in order to make it fair? The expected value is infinite as the series diverges (cf. Székely 1986, p. 27):

$$2^0 \cdot \frac{1}{2} + 2^1 \cdot \left(\frac{1}{2}\right)^2 + 2^2 \cdot \left(\frac{1}{2}\right)^3 + \cdots = \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \cdots = \infty.$$

What is the Paradox? As an economic value of exchanging risk and certainty, an infinite value is completely unacceptable. No one could pay an infinite amount of money in advance. Nor would the player ever have a positive balance as all payments from the game will always be finite. And no human could witness the theoretical never-ending of tossing to get this infinite amount. A revision of the concept of expected value or of probability was urgently required. As an equivalent for an uncertain situation, expected value was regarded as a property of the situation. The time was not yet ripe to see this as artificial. However, it is still disputed whether a distribution makes sense if its expected value is infinite.

Further Ideas The paradox was put forward in 1738 by Daniel Bernoulli. Either the concept of expected value has to be revised or probability has to be conceptualized in a different way. The ways to resolve the paradox were twofold.

One way is to introduce *utilities* instead of money. Bernoulli (1738/1954) suggested replacing the payments by their logarithms arguing that the more money one has the less it is of importance to the person. In fact, he brought the expected utility down to a finite value but failed to provide a comprehensive solution to the paradox as, with a slightly different payment table, an infinite (expected) value would still result. The second way is to introduce a new entity such as a moral probability, which can be neglected if sufficiently small. The ensuing dilemma is to specify the size where probabilities lower than this benchmark could be neglected. The suggestions varied from 10^{-4} to 10^{-15}.

The concept of utility has been taken up by various approaches to applied probability. For a Bayesian probabilist, the way to evaluate an unknown probability is first by the subjective degree of credibility. Combinatorial multiplicity is a substantial factor as well as the information on past relative frequencies from similar experiments, but there are also personal and qualitative ingredients. All factors are prone to utility of the outcomes as—in measuring them—Bayesians would use the idea of equivalent bets that are accepted. Nowadays, utility has been revived by the discussion of teaching approaches based on risk which focus not only on probability but also on the impact associated to the possible outcomes.

The probabilist community has still not solved the problem of small probabilities. On the one hand, for events with small probabilities, the related impact may bias the personal perception of the magnitude, and data is missing as the probabilities are so small. On the other hand, small probabilities play a vital role as inherent properties of statistical procedures as the size of a significance test or the confidence level of a confidence interval show. Both reveal a lack of interpretation of small probabilities in the frequentist sense despite widespread endeavour to simulate the underlying assumptions in scenarios of the real situation.

3.2 Independence and Expectation

Probabilities have to be recalculated when games of chance are dependent. However, for expectation, it is irrelevant whether games are dependent or independent. In this sense, expected value is a functional analytic and not a stochastic property while variance is a genuine stochastic concept. In the first example below, two simple independent spinners are introduced and then changed to become dependent. In the second example, samples from a bag of coins are dealt with to illustrate the consequences of replacing or not replacing the drawn coins. The latter example highlights the advantages of the economic concept of a value.

***P_6*: Dependent Spinners** Two simple spinners are spun and the shaded area gives an amount of 1 to the player while the white sector leads to a 0 payment. The small spinner has a winning probability of p, i.e. $P(X=1)=p$, and the big spinner of q, i.e. $P(Y=1)=q$. The expected amounts to win are $E(X)=p$ and $E(Y)=q$. If played independently, one after the other, the fair price is p for the small and q for the big spinner (Fig. 1(a)). The price could also be paid in advance to play the game of $X+Y$ with an expected value of $E(X+Y)=p+q$. Putting the spinners one over the other, the game can be decided in one turn; when the spinner lands in the overlapping sector, the player wins both from the small and the big spinner, that means the win is 2 (Fig. 1(b)). With this variation the expected value of $X+Y$ remains the same. Whether the games are independent or dependent, for expected values the following additivity holds:

$$E(X+Y)=E(X)+E(Y).$$

What is the Paradox? Despite the close connection of expected value to probability, which remained confused for quite a long period, some basic properties differ. Expected values can be calculated from dependent random variables as if they were independent and represent—in this sense—*not a stochastic property*. Yet, the concept is used to find the fair price of a game of chance. In the special case of binary variables with 0 and 1, expected values do actually coincide with the probabilities.

Further Ideas With an overlap of x for the winning sectors in the spinners, the probabilities for the single payments are easily read off Table 3. The exact value of p_0 is not required for further calculations.

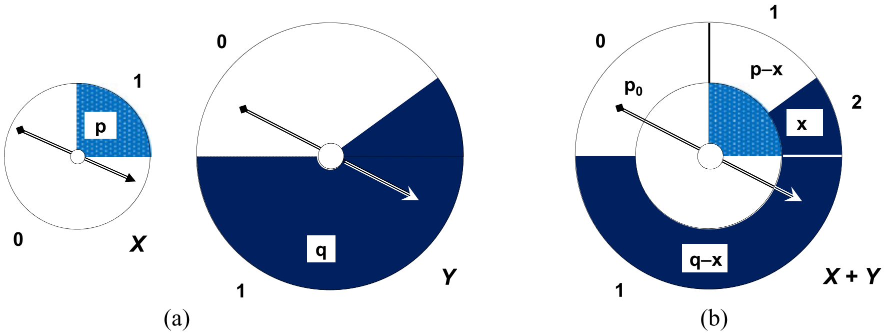


Fig. 1 (**a**) Two independent spinners. (**b**) Dependent spinners

Table 3 Payments for dependent spinners

$X+Y$	Probability
0	p_0
1	$p+q-2x$
2	x

The terms involving x cancel, in fact. Interestingly enough, for the variance an additivity relation holds only in case of independence of the games, i.e.

$$\mathrm{var}(X+Y)=\mathrm{var}(X)+\mathrm{var}(Y) \text{ iff } P(X{=}1 \text{ and } Y{=}1)=P(X{=}1)\cdot P(Y{=}1),$$
$$\text{i.e. } x=p\cdot q.$$

***P*$_7$: Dependent Coins** Another example relates to drawing three coins (values 10p, 10p, 10p, 50p, 50p, 50p, 100p) from a bag with and without replacement (Borovcnik et al. 1991, p. 62). Calculations show that expected value is 3 × 40p, which is the same irrespective of replacement.

What is the Paradox? To calculate the probabilities, one has to know the result of the first draw in case the coins are not put back—the single draws turn to dependent random variables. Yet, for the calculation of the mean of the second draw one can neglect the result of the first. This is counter-intuitive. Furthermore, the design of the game without putting coins back prompts many people to reconstruct the situation personally. Speculating that the 100p coin can be drawn at first and decrease the net gain for the second is quite frequent resembling the sayings "the first wins" or "who dares to start wins".

Further Ideas This surprising result characterizes a substantial difference between probability and expectation. From the perspective of probability, the simplifying relation of linearity seems intuitively unacceptable if the random variables are dependent. However, for expectation there is symmetry; the probability of a specific coin being drawn at the first draw is the same as for the second or third try, even if coins are not replaced. Consistently, its individual contribution to each of the draws is the *same*. As this holds for all coins, the values are equal for all draws,

$$E(X_1)=E(X_2)=E(X_3)=40\text{p},$$

whether coins are replaced or not. Thus the required expectation is 120p. Straightforward mathematical arguments support this reasoning. However, the historic situation was obscured by lengthy calculations, which outweighed its conceptual simplicity. The spinners and coins puzzles indicate that intuitions about expectation need to be discussed in teaching probability; these examples provide good stimuli.

In recent endeavours to enrich curricula and shift probability away from games of chance, risk is being taught in schools: situations are dealt with including the different outcomes *with* their related impact *and* their probabilities. Different options are then compared by their expected impact. For an individual the perception and evaluation of the probability of an outcome is influenced by its related impact. This interdependence is much increased if the probabilities are (very) small but the impact is enormous—as with screening programmes for preventing diseases like cancer. While risk extends the scope of teaching to an important field of application, it is inappropriate to introduce it right from the beginning. As the historical development has shown—the perception of probabilities, and the appreciation of a probability statement takes time to clarify; the notion of impact might obliterate the process of calibrating the feeling what a probability of e.g., 1/4 (or the proportion of 1 to 3) signifies for an event.

4 Relative Frequencies

The early attempts to explore probability were accompanied by the idea of frequency. The right way to count multiplicity was supported by the similarity of derived probabilities to the frequency of occurrence in repeated trials. Bernoulli (1713/1987) proved his law of large numbers and provided a justification to interpret probabilities as relative frequencies, which paved the way to many applications from life-tables to the behaviour of particles in physics.

Another move forward was Laplace's derivation of the central limit theorem (going back to preliminary results of de Moivre 1738/1967), which promoted the normal not only as a limiting distribution to the binomial but also as an element to formulate laws: laws in physics to describe the behaviour of entities at the microscopic level, but also laws to extract an estimation of unknown parameters from data. The focus of applications definitely turned towards empirical probabilities. The basis laid by Laplace was equi-probability and the principle of insufficient reason, which was criticized by empiricists like Venn. The time was ripe for probability as something like idealized relative frequencies (FQT).

The most serious attempt to classify relevant properties of relative frequencies in series of experiments axiomatically was made by von von Mises (1919), though it was dismissed as not sufficiently rigorous. The basic entities of his approach were *infinite series* of (theoretical) relative frequencies and some quite vague properties. One counter-argument to this approach was its complexity, while some contradictions were only repaired by Schnorr (1971). Kolmogorov's (1933) probability used the fundamental probability space and idealized relative frequencies for events instead of series of outcomes. It was universally acknowledged as a sound basis *and* also justified the interpretation of probability as relative frequency though there was an ongoing debate on repairing the direct approach by von Mises at the famous Geneva conference in 1937 (see the proceedings edited by Wavre 1938–1939).

Relative frequencies are based on independently repeating a random experiment, which is not always easy to define as shown by the exemplar paradoxes of the library problem and Bertrand's chord. It is startling to note that the conditions of randomness have to be operationalized when one would initially think that the experiment under scrutiny is random and its description is unambiguous. In both problems, the random selection of an object is operationalized in different ways thus confusing those who may think that random selection and the resulting relative frequencies must lead to a *unique* probability.

There are many idiosyncratic perceptions about how randomness manifests itself in repeated trials, which need to be addressed in the classroom. These perceptions underline the complexity of the concept of probability as the limit of relative frequencies. One irritating aspect refers to the patterns of a finite sequence of trials, which attract many to build up their own structure to continue the short-term behaviour of the frequencies—see Shaughnessy (2003), Konold (1989), or Borovcnik and Bentz (1991) for empirical teaching studies on such phenomena and strategies used. While the examples are puzzles in the sense of confusing problems, the features are also counter-intuitive with respect to personal thought and not due to a mathematical conception of probability as relative frequency.

Intuitively one might think that an experiment which is random has a unique formulation automatically, yet this is far from being true. This is made worse if the conception of probability is seen as a unique and almost physical property, like weight or length. So, if probability *is* nothing but the relative frequencies *in the long run*, how can it be that a problem gives rise to several experiments, which all depict the situation but lead to different relative frequencies and thus to different probabilities? If probability *were* only relative frequencies (of real objects), the situation would amount to a paradox, as illustrated by the library problem and the chord of Bertrand (1888).

***P_8*: Library Problem** A book is selected randomly from the library. Determine the probability that it is written in English if there are 500 English out of 1,000 books in the library. A student has performed the experiments more than 2,000 times in going to the library randomly selecting a book; he reports a relative frequency close to 0.67. The librarian, on the other hand, has randomly chosen the book's index card from the search catalogue and got a relative frequency of 0.5. Why? (cf. Borovcnik et al. 1991, p. 60).

What is the Paradox? To choose randomly an element from a sample space seems to be unambiguous. Thus, following the steps the result should be the same for both. The paradox is that randomness has to be operationalized. There is no unique randomness, as the concept is bound to a *model* of the real situation. According to the model used, the relative frequencies differ and give different probabilities. The probabilities refer to the model rather than to the real situation while one may be surprised that there are various models to represent "random choice". One cannot "act" in the real situation ("behave" randomly) without using a model. This gives a clear hint that probability is a model entity rather than a physical property as it is related genuinely to a *model* of the world.

Assume that the library has a big and a small room, with a corridor in between. The student selects in two random steps: (i) throwing a coin to decide which room to enter; (ii) when in the room, selecting the book from the shelves from left to right according to a random number. Assume that there are $|E_1| = 410$ English books of a total of $|T_1| = 900$ books in the big room while for the small room the corresponding numbers are $|E_2| = 90$ and $|T_2| = 100$. A short calculation will show that, in fact, the student's random selection to get an English book by this selection is $61/90$ while using the card index yields 0.5.

If picking randomly is to be conceptualized by Laplace's approach then all the possible ways of selecting a book should yield the same probability. As various feasible (random) selection processes lead to different answers, the approach fails as long as one refers to the real situation instead of a model of it. The model is determined by the operational steps of selection.

***P*₉: Bertrand's Chord** An equilateral triangle is drawn in a circle with radius R and a line randomly drawn through the circle (Fig. 2). What is the probability that the segment s of the line in the circle is longer than the side a of the triangle? (Bertrand 1888, or Székely 1986, p. 43). Three possible solutions are given here (Figs. 3(a)–(c); cf. Borovcnik et al. 1991, p. 59).

(a) As the segment is uniquely determined by its mid-point M, we may focus on the position of M. If M is contained in the inner circle with radius R_1 with $R_1 = R/2$, we have $s > a$, otherwise $s \le a$ (Fig. 3(a)). Hence

$$P(s > a) = \frac{R_1^2 \pi}{R^2 \pi} = \frac{1}{4}.$$

(b) We may compare the position of s on the diameter d perpendicular to s. If s falls within the interval I (see Fig. 3(b)), its length is greater than the length of a. As $|I| = R$, this yields

$$P(s > a) = \frac{|I|}{d} = \frac{1}{2}.$$

(c) As each segment s cuts the circle in P and Q, we may consider the angle β between s and the tangent t at Q in order to express the position of P in terms of β, which can lie in the range of $(0, 180)$. If $60 < \beta < 120$, we have $s > a$ (Fig. 3(c)), thus

$$P(s > a) = \frac{60}{180} = \frac{1}{3}.$$

What is the Paradox? There should be a unique way to draw a chord in the plane. The puzzling issue is that the experiment is hard to perform in reality and the steps

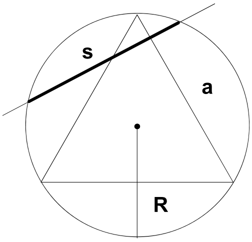


Fig. 2 Bertrand's chord: Line segment s and side a of the triangle to be compared

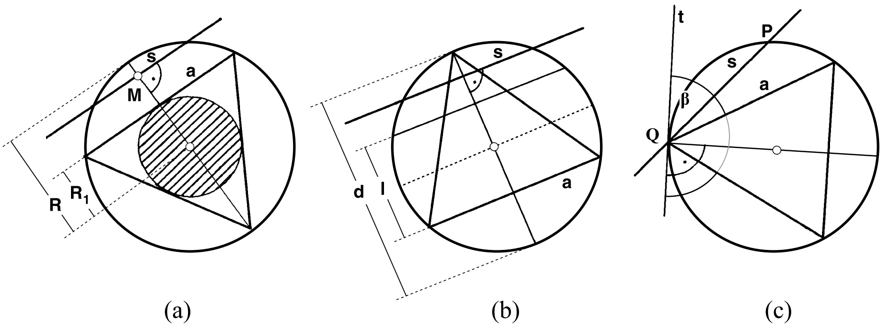


Fig. 3 (**a**)–(**c**): Bertrand's chord: Different possibilities to draw a random line

on how it is done have to be operationalized. In fact, the way the experiment is performed influences the result. Randomly drawing, in whatever way this may be defined, requires the use of Laplacean equi-probability on the possibilities which are open. Equi-probability refers only to the model and there is no guarantee that the right model is chosen. That implies that each of the three solutions represents chance and inherent equi-probability via its particular random generator. Is there more than one way of randomly determining a line? This reflects an intuitive conflict and yields a contradiction to the basic assumption of Laplace's definition; the word *randomly* is neither fully covered by this approach nor is it meaningful without reference to an actual generator of the events.

Further Ideas From today's perspective, there is no paradox. Probability is mathematically defined via the axioms and a stochastic experiment is described by different *models*, which may, of course, lead to different answers. The only question is which of the models in a real experiment delivers the better predictions. Only solution (b) fulfils the requirement of invariance to translations and rotations of the plane (see Székely 1986, p. 45 for a hint, or Palm 1983, for a full explanation), which is better in certain systems in statistical mechanics and gas physics. Nevertheless, the example shows that there is a problem worthy to discuss in the classroom situation. If not, then false intuitions may remain with children, and progress in learning probabilistic ideas and their application is hindered.

5 Personal Probabilities

There is little room (in APT and FQT philosophies) for the idea of probability as a judgement of a person about a statement or an event (SJT). This would make probability personal and subjective, rather than an objective concept. This view of probability is legitimatised by axioms on rational behaviour since de Finetti (1937) who states that probability does not exist, except as a personal idea: this approach is rejected by many as non-scientific. Where the focus is on the measurement of probabilities and the use of random experiments (which is an *idealization*), the critique against subjectivist probability is justified. The problem, however, is that with the exclusion of subjectivist probability other merits of this latter position are abandoned. The recurrent difficulties with conditional probabilities within a closed objectivist probability theory are a convincing argument that the idea of subjectivist probability is integral not only to people's intuitive reconstructions of mathematical concepts but also that a wider conception of probability is needed; this is especially true where events of low probability are concerned.

Despite the eminent role of independence, relevant intuitions are hard to clarify within mathematics. Some vague, nearly mystic arguments about "lack of causal influence" are used to back it up. It is interesting to note that subjectivists avoid these difficulties by replacing independence by *exchangeability*, an intuitively more accessible concept. The difficulties increase even with dependence, which is more than a simple complement to the notion of independence as there is a whole range of dependencies. Dependence is formalized by conditional probability, which is simply the "old" distribution restricted to the subspace determined by the conditioning event.

It should be no surprise that this mathematical approach gives rise to many difficulties in understanding. For example, for the dominant situation with equally likely cases, a reduction to a subspace cannot affect the equal probabilities—can it? On the contrary, for subjectivists, probability is a degree of belief and conditional probability is a basic notion which covers the intuitive idea of revising judgements as new information becomes available.

The Bayesian formula is a key tool and it is clear that for the final (posterior) probability judgement two ingredients have to be integrated, namely the prior probability of the states and the likelihood of the new information under the various states. The formula is so important that its clumsy appearance within Kolmogorov probability is changed into a more suitable and elegant mode for quicker re-evaluation of probabilities, which allows more direct insight on the relative importance of the influence factors (the prior probabilities and the likelihoods) and their impact on the final judgement. For this purpose, subjectivists often speak about probabilities in the form of odds or relative probabilities.

Carranza and Kuzniak (2009) analyse examples included in the curriculum with the conclusion that many deal with conditional probabilities and do not match the rest of the curriculum, which is oriented to the objectivist paradigm. According to objectivist paradigms, probability is strictly a property of objects which can be modelled differently mainly on the basis of available frequencies. The subjectivist paradigm relates probability to a judgement by a person who has some information available, which includes frequencies and qualitative information.

As the subjectivist position is criticized for being subjective (!), it was rejected as solution for a mathematical concept of probability. However, the Bayesian approach (SJT) is much closer to how many people think and can thus much better explain the part of conditional probabilities.

5.1 Inverse Probabilities

The following paradoxes show the difficulties in assimilating information in calculating probabilities. Intuitively many people do not believe that new information can change a probability. The application of Bayes theorem to calculate posterior probabilities is certainly complicated as shown by the furious international discussion of the Monty Hall problem (see Gigerenzer 2002). Here we present Bertrand's paradox and one relating to Father Smith.

***P*₁₀: Bertrand's Paradox** A cabinet has three boxes each with two drawers. Three gold and three silver coins are put into the drawers so that two boxes contain coins of the same kind and one the mixture. Randomly choose a box, then a drawer and open it; it is assumed that it contains a gold coin, which is denoted as event "G". Of interest is whether the box drawn first has coins of same type, which will be denoted as ST. After choosing the box but *before* the drawer is opened, there are 2 of 3 equally likely boxes, which yields $P(ST) = 2/3$ (Bertrand 1888; for an easy-to-play card version of the game, see Gardner 2006, p. 93; Gardner named it the "Three Card Swindle").

After seeing a gold coin, there remains 1 of 2 equally likely boxes, thus $P(ST|\text{“}G\text{”}) = 1/2$. *After* a silver coin is seen in the opened drawer, for symmetry, it holds that $P(ST|\text{“}S\text{”}) = 1/2$. If either a gold or a silver coin is in the opened drawer, the new probability is $1/2$. There is no need to look into the drawer, any result will decrease the probability to $1/2$, thus it *is* $1/2$. But probability of ST cannot be $2/3$ and $1/2$ at the same time (Figs. 4(a)–(b)).

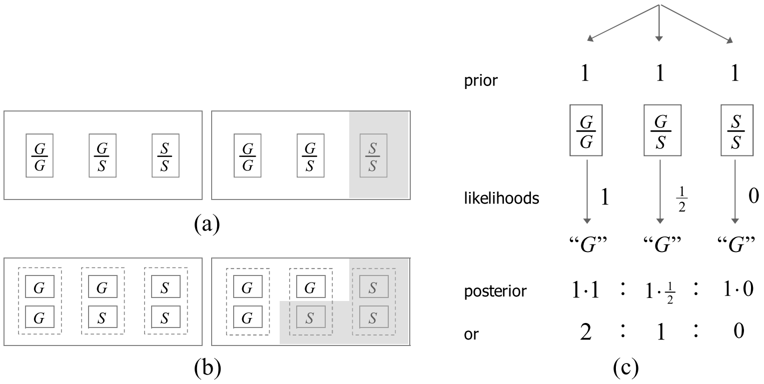


Fig. 4 Bertrand's paradox of drawers:
(**a**) Perceived random selection: 2 of 3 for *same type* before and 1 of 2 after seeing gold.
(**b**) Hidden random selection: 4 of 6 for *same type* before and 2 of 3 after seeing gold.
(**c**) Tree to combine priors and likelihoods

What is the Paradox? The focus on *equi-probable* objects (the boxes) leads to the trap. Logically, information "G" is used to reduce the possibilities—the $\boxed{\frac{S}{S}}$ box is eliminated. With equi-probability on the remaining boxes this yields $1/2$ for $\boxed{\frac{G}{G}}$ and $1/2$ for $\boxed{\frac{G}{S}}$; all in all this yields $1/2$ for same type and for mixed boxes. This argument leads to the paradox.

Seeing gold reduces the space, in fact, to the pure gold and the mixed box. However, the two-stage random selection leads to a hidden selection of the remaining coins as seen in Fig. 4(b), i.e. 2 out 3 gold coins lead to the pure gold box and thus for ST while the third leads to the mixed box. Thus the conditional probability remains at $2/3$ and the paradox is solved. This result can be confirmed by Bayes' formula. With equi-probable cases it is hard to think that this property of equal probabilities can be changed by new information. As, by definition, the conditional probability is simply a reduction of the present to a smaller space, how can this procedure change equal probabilities? The crux is that information "G" is used only on this logic base to reduce the space. Many people forget to discriminate between the other two boxes.

Further Ideas The mixed box has a conditional probability of $1/3$ confirming the hidden lottery argument. The situation can be generalized as it will enhance the *structure* of such situations.

The three boxes are perceived as hypotheses H_i and the evidence A as the result of the opened drawer. The hypotheses have a *prior* probability of $1/3$ each. The new or updated probabilities are calculated using Bayes' formula (the details are omitted):

$$P(H_i|A) = \frac{P(H_i \cap A)}{P(A)} = \frac{P(H_i) \cdot P(A|H_i)}{P(A)}.$$

Bayesians frequently use relative probabilities, so-called odds. For $P(E) = \frac{1}{6}$, the odds of E against its complement $\bar{E}$ are as $\frac{1}{6} : \frac{5}{6}$, or $1 : 5$. Odds are proportions but can freely be read as fractions. From odds of $1 : 5$, the probability is calculated back by $P(E) = \frac{1}{1+5} = \frac{1}{6}$, generally with odds of $a : b$, a probability of $P(E) = \frac{a}{a+b}$ is associated.

Comparing the updated probabilities of the hypotheses by odds yields (Fig. 4(c)):

$$\underbrace{\frac{P(H_i|A)}{P(H_j|A)}}_{\textit{posterior odds}} = \underbrace{\frac{P(H_i)}{P(H_j)}}_{\textit{prior odds}} \cdot \underbrace{\frac{P(A|H_i)}{P(A|H_j)}}_{\textit{likelihood ratio}} .$$

For Bertrand's cabinet this delivers a probability of $2/3$ from the posterior odds of $2 : 1$.

$$\frac{P(\boxed{\frac{G}{G}}|\text{“}G\text{”})}{P(\boxed{\frac{G}{S}}|\text{“}G\text{”})} = \underbrace{\frac{1}{1}}_{\textit{prior}} \cdot \underbrace{\frac{2}{1}}_{\textit{likelihoods}} = \frac{2}{1} = 2 : 1.$$

It is a deep-seated fallacy that the given information about a gold coin will leave the equi-distribution on boxes intact and that equal probabilities can be applied to the reduced sample space with the two remaining boxes—Bertrand (1888) favoured the wrong equi-probability of the two remaining boxes ending up with a paradox. There is also a reluctance to accept the results of Bayes' formula. Various didactical strategies to overcome this problem have been designed.

Freudenthal (1973) uses the technique of *implicit lotteries* (p. 590); the lottery on the boxes is symmetric, but the choice of the drawer is, by no means, symmetric. Falk and Bar-Hillel (1983) and Borovcnik (1987) suggest the *favour concept* which could intuitively clarify the higher estimate of the probability of the pure gold box as gold in the open division is circumstantial evidence for the box with two gold coins. Borovcnik and Peard (1996) suggest adapting mathematical formalism to fit better. The resulting view on Bayes' formula with odds connects objectivist and subjectivist conceptions. It gives a clearer view on the *structure* of the problem with prior possible states and evidence that leads to a new judgement of the probabilities. The value of an indication is represented by the relative likelihoods. The best is: to have evidence that has a high probability under one state and very small probabilities under the other states. Such an indication gives a clear new judgement. However, such situations are rare.

***P_{11}*: Father Smith and Son** Mr. Smith is known to have two children and various items of information may be analysed, leading to different posterior probabilities that he has two sons (this is the Two Children Problem of Gardner 1959, p. 51; cf. also Borovcnik et al. 1991, p. 64). The information is set out in Table 4 ((a) seen in town with a son, (b) visiting his home and randomly see a boy, (c) told eldest child is a boy, (d) told he has at least one boy, (e) told he prefers to go out with his son, (f) told he prefers to take his eldest child out, (g) told there are different probabilities for a boy or a girl to be at home).

What is the Paradox? It is confusing that information that seems to be similar or equivalent has a different impact on the probabilities. One has to judge how the information has been gathered before one can start to solve the problem.

Table 4 Different impact of the evidence on the posterior odds

Item	Information	Posterior odds							Solution	Primitive result
	"B"	*BB*		*BG*		*GB*		*GG*	$P(BB \mid \text{“}B\text{”})$	
(a)	See in town	1	:	$\frac{1}{2}$	:	$\frac{1}{2}$	:	0	$\frac{1}{2}$	$\frac{1}{3}$
(b)	See at home	1	:	$\frac{1}{2}$	:	$\frac{1}{2}$	:	0	$\frac{1}{2}$	$\frac{1}{3}$
(c)	Eldest is boy	1	:	1	:	0	:	0	$\frac{1}{2}$	$\frac{1}{2}$
(d)	At least one boy	1	:	1	:	1	:	0	$\frac{1}{3}$	$\frac{1}{3}$
(e)	p prefers boys	1	:	p	:	p	:	0	$\frac{1}{1+2p}$	$\frac{1}{3}$
(f)	q prefers first	1	:	q	:	$1-q$	:	0	$\frac{1}{2}$	$\frac{1}{3}$
(g)	p_G, p_B at home								$\frac{2-p_B}{4-p_B-p_G}$	

Further Ideas In a structural view of the puzzle, the given information has to be linked to the possible states to re-evaluate them. The method of comparing prior and posterior probabilities of states can be applied to the information; but, as noted above, Bayes' theorem is complicated.

5.2 *Conflicts with Logic*

The theory of probability has a mathematical foundation, derived by logic. It is startling that reasoning with probabilities reveals a structure that appears to conflict with some of the rules of ordinary logic. This amounts to a puzzling situation as users erroneously expect all conclusions with probabilities to be in line with the following logical laws.

(a) *Transitivity* of logical reasoning. If A is bigger than B and B bigger then C, then one can conclude that A is bigger than C. Such a property signifies logical implication: If A implies B and B implies C, then—by transitivity—A implies C. This method establishes an important technique of mathematical proof.

(b) *Poof by exhaustion* or proof by cases. If a logical statement is true in either of two cases and these amount to all possibilities (and are disjoint), then the statement is true in all cases. That principle may be extended to 3 or more (countably infinite) cases, for example, if an equation $q(x) = 0$ holds for $x > 0$ (case 1), $x < 0$ (case 2) and $x = 0$ (case 3), it is true (for all real numbers x).

Probability statements conflict with these two principles, as shown by the following puzzles. As the logical relations seem quite natural, the clash is between properties of probability statements and intuitions. Nothing is wrong with probabilities thereby and nothing can be changed about these properties.

***P*$_{12}$: Intransitive Spinners** Suppose there are three spinners (Fig. 5). Which is the best to choose if two players compete and the higher number wins?

Player 1 chooses a spinner; player 2 chooses a spinner from the two remaining. There is no best choice for player 1; a short calculation shows that

$$P(S_1 > S_2) = 0.52, \qquad P(S_2 > S_3) = 0.61, \quad \text{and} \quad P(S_1 > S_3) = 0.25.$$

The second player can always find an alternative that is better. Player 1 is doomed to lose in this game. Recognizing this, it gets even more confusing that there *is* an

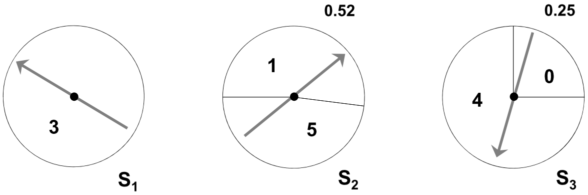


Fig. 5 Intransitive spinners

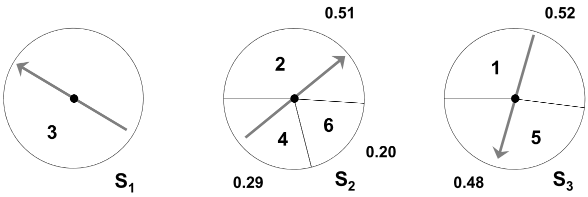


Fig. 6 Spinners from Blyth's paradox

order of the options with respect to losing (which is the complement to winning): To avoid losing too often, player 1 has a ranking on the spinners:

$$S_2 \succ S_3 \succ S_1$$

according to losing probabilities 0.52, 0.61 and 0.75 (against the best alternative).

What is the Paradox? If $P(S_i > S_j) > 0.5$ is interpreted as $S_i \succ S_j$ (and $\succ$ is used in the sense of "is better") then the properties of the spinners read as:

$$S_1 \succ S_2 \text{ and } S_2 \succ S_3 \text{ but } \textit{not}\text{—as transitivity would imply—}S_1 \succ S_3.$$

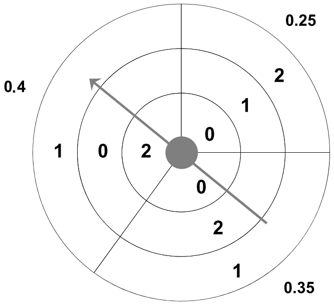
This is also puzzling in an everyday context: if A is better than B (in whatever respect) and B is better than C then, of course, A is better than C. A preference system that contradicts transitivity is counter-intuitive. Another puzzling feature is that with respect to winning there is no ranking for the spinners but to avoid losing too often there is a definite ranking. Is losing complementary to winning or not?

There is a high expectation that possibilities can be ranked according to some criterion and that ranking fulfils transitivity. The lack of transitivity in the choice illustrates that stochastics is different rather than a weak form of logic, as expressed by saying '... is true with probability p' instead of '... is definitely true'.

***P*₁₃: Blyth's Intransitive Spinners** Blyth (1972) varies the situation (see Fig. 6): two or three players might enter the game. In fact, the optimal choice depends on the number of players. The calculations are left to the reader to explore because of restrictions of space, and similarly for the next puzzle.

P_{14}: Reinhardt's Single Spinner Reinhardt (1981) has a single spinner with several wheels with a similar puzzling result (Fig. 7).

Fig. 7 Reinhardt's spinners



For a more detailed discussion of these spinners, see Borovcnik et al. (1991, p. 65). Puzzles like these can be motivating in the classroom situation. Usually they lead to a lively discussion of the applicability of such ideas; sporting situations such as in a football league are a suitable context. It certainly does happen that team A can beat team B but lose to team C, who are also beaten by team B. There are many intransitive situations in the real world that can cause confusion. A third player reversing the ranking of choice does occur.

***P_{15}*: Simpson's Paradox of Proportions** In 1973, the following phenomenon was observed at the University of California at Berkeley: the overall admission rate of female applicants of 35 % was lower than that of their male colleagues, which amounted to 44 %. Females seemed to be less likely to gain admission. Searching for the reason for this sexual 'discrimination', it turned out that in some departments women actually had higher admission rates than men while most of the departments had similar rates for both (see Bickel et al. 1977). The following setting shows that admission rates can be higher for women than for men in *all* departments, and yet be lower for the whole university.

The situation is simplified by assuming there are only two departments. Green marbles represent admission, red marbles rejection, so, in Table 5, for example, 2 out of 5 females and 1 out of 3 males are accepted by department 1.

In both departments, the proportion of green marbles (admitted) is higher for females than for males, $2/5 > 1/3$ and $3/4 > 5/7$. But for the university as a whole, the reverse holds: $5/9$ for females admitted is lower than $6/10$ for males.

What is the Paradox? In department 1, the statement "women have a higher admission rate" is true. This holds also for department 2. Since, the two disjoint cases exhaust all the possibilities, why is the complementary statement true for the whole university? In what follows, a probabilistic framework for the situation will be established, first to show it deals with a *probabilistic* puzzle, second, to extend the analogy to logic.

Let F, M, G, R be the events female, male, green (admitted), red (rejected). Two urns are filled for the departments according to Table 5 and the experiment is drawing a ball within each department urn. Then the experiment is repeated with

Table 5 Illustrative data for Simpson's paradox

	Females		Males		All	
	Green	Red	Green	Red	Green	Red
Department 1	2	3	1	2	3	5
Department 2	3	1	5	2	8	3
University	5	4	6	4	11	8

Table 6 Some probabilities of interest within the departments and overall

	$P(F)$	$P(M)$	$P(R)$	$P(G)$	$P(G\|F)$
Department 1	0.625	0.375	0.625	0.375	0.400
Department 2	0.364	0.636	0.273	0.727	0.750
University	0.474	0.526	0.421	0.579	0.556

a university urn that is filled with all these balls. Some probabilities of interest are listed in Table 6.

In both departments, it holds that $P(G|F) > P(G)$, yet for the university the reverse relation holds as $P(G|F) < P(G)$. The reason for the intuitive clash lies in the *application rates* of males and females; females tend to apply to departments with low admission rates.

Further Ideas For logical implication $A \Rightarrow B$ the truth of statement A implies the truth of B. In analogy to this, a new (and weaker) relation between events is introduced. If one event increases the (conditional) probability of the other, i.e. if $P(B|A) > P(B)$, this is defined as $A \uparrow B$, in words, A *favours* B. Disfavouring, denoted as $A \downarrow B$ means that the conditional probability is smaller. With this concept, reasoning with probabilities is shown to differ from logical conclusions (Table 7).

Conditional probabilities are a subsidiary concept in the usual approach towards probability (either APT or FQT) and within the axiomatic approach there is no room for an investigation about the order or the direction of change of conditional probabilities. Such operations are, however, at the core of subjectivist theory (SJT) of probability, and it is important to integrate some elements from this position into teaching in order to enhance the underlying concepts, as advocated by Carranza and Kuzniak (2009) who supported their view by an analysis of teaching approaches.

The Simpson effect occurs in various contexts. Vancsó (2009) discusses an example with higher mortality in Mexico than in Sweden within all age classes (0–10, 10–20, etc.) but overall mortality being higher for Sweden than for Mexico. Underlying this version of the Simpson paradox is the fact that the Swedish population is much older while Mexico is a young country.

Table 7 Comparing the structure of logical reasoning and favouring

	Favouring	Logical implication
In case 1 it holds	$F \uparrow G$ and	$A \Rightarrow B$
In case 2 it holds	$F \uparrow G$	$A \Rightarrow B$
In all cases it holds	$F \downarrow G$ (in the example)	$A \Rightarrow B$ (generally)

6 Central Ideas of Probability Theory

In line with the intention of Kapadia and Borovcnik (1991), we present some central ideas in the theory of probability to provide a coherent treatment. Paradoxes indicate where a specific conception comes to an end and a reformulation of the terms is required to resolve the conflict. Puzzles show a divergence between official interpretations and private conceptions. Both paradoxes and puzzles though have limited implications. To develop a stable conception of notions, the mathematical *structure* is vital. What are the central ideas that link the concepts? We suggest the following set of ideas: independence and random samples; central theorems; standard situations; axiomatization. In order to reveal the value and role of axiomatization of probability, of course, a more detailed exposition of the underlying mathematics is important. To simplify here bears the risk of transmitting a limited picture of the theory and its potential for applications. The stronger mathematical demand in reading will pay off only afterwards by a deeper evaluation of the scope and limitations of probability.

6.1 Independence and Random Samples

The basic paradigm for probability is the experiment which can be repeated under essentially the same conditions and for which the outcome cannot be known beforehand with absolute certainty. This is modelled by a random variable X and the cumulative distribution F of X, i.e. $F(x) = P(X \le x)$. The notion of independence extends naturally from events to random variables. For events A and B, independence means that

$$P(A \cap B) = P(A) \cdot P(B).$$

With the distribution function, the independence of random variables is defined by the condition:

$$F(x, y) = F_X(x) \cdot F_Y(y) = P(X \le x) \cdot P(Y \le y).$$

A sequence of independent random variables $X_1, X_2, \ldots$, each with the same distribution is called a *random sample* from that distribution and is denoted by

$$X_n \overset{iid}{\sim} X \sim F$$

(*iid* stands for independent and identically distributed). Such a series is a useful model for repeated observations, all independent and following the same distribution F.

The notion of a random sample is easily represented by a spinner which is independently spun several times. Another effective representation for sampling from a finite population is drawing balls from an urn, which is thoroughly mixed, prior to each draw.

6.2 Central Theorems

Historically, the development of the sum $H_n = X_1 + X_2 + \cdots + X_n$ has been investigated in special cases of the distribution of the single variables and H_n was binomially distributed. The deviations of H_n from their expected value tend to zero if n goes to infinity—Bernoulli's law of large numbers. Later, the probabilities of such deviations were asymptotically calculated, which led to the normal distribution and the central limit theorem. The assumption of independence was hidden in the binomial distribution.

Laws of large numbers show a convergence of the average of the random variables to a fixed value, while central limit theorems deal with the convergence of the standardized average from a sample to the normal distribution. Variations of the theorems cover different types of convergence of the average, other statistics like the median or the standard deviation derived from the series, specific distributions of single random variables, and restrictions leading to limiting distributions other than the normal distribution. We use modern notation to describe two theorems that became famous in the early mathematical development of the field, Bernoulli's law of large numbers and Laplace's central limit theorem.

Bernoulli's Law of Large Numbers Let A be an event of an experiment with $P(A) = p$, and X_i be a binary variable determined by an occurrence of A in independent repetitions, i.e.

$$X_i = \begin{cases} 1 & \text{if } A \text{ occurs at the } i\text{th trial,} \\ 0 & \text{if } A \text{ fails to occur at the } i\text{th trial,} \end{cases}$$

and let $H_n = X_1 + X_2 + \cdots + X_n$ be the absolute frequency of A in n trials, then for any positive real number ε

$$\lim_{n\to\infty} P\left(\left|\frac{H_n}{n} - p\right| \geq \varepsilon\right) = 0.$$

A generalization refers to the convergence of the mean of samples to the mean of the underlying population (cf. Meyer 1970, p. 246, or Çinlar 2011, p. 118).

If $X_1, X_2, \ldots, X_n$ are independent random variables from a common distribution with finite mean μ and variance σ^2 then, given $\varepsilon > 0$

$$\lim_{n\to\infty} P\left(\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| \geq \varepsilon\right) = 0.$$

The law of large numbers states that the mean of a sample will be close to the (unknown) mean of a distribution from which the sample was drawn, with a high probability provided that the sample size is sufficiently large and the selection process is random.

There is a strong version of both laws of large numbers stating that the set of infinite series which do not converge to the expected value of a single variable (i.e. p or μ) has a probability of zero (cf. Çinlar 2011, p. 122). This strong law of large numbers goes back to Borel (1909); its disputed status was not clarified until Kolmogorov's axiomatic foundation on the basis of measure theory.

Laplace's Central Limit Theorem The variable H_n (the absolute frequency) in Bernoulli's theorem will deviate from the expected value np in n independent trials so that it is a random variable with its own distribution. De Moivre considered a special case and Laplace found H_n to be approximately normal:

$$\lim_{n\to\infty} P\left(\frac{H_n - np}{\sqrt{np(1-p)}} \leq z\right) = \int_{-\infty}^{z} \varphi(t)\, dt$$

with $\varphi(t)$ being the probability density function of the standard normal distribution.

If the single random variables follow independently the same distribution as X (iid) and this distribution has a finite mean μ and a finite variance σ^2, the theorem would still hold, i.e.

$$\lim_{n\to\infty} P\left(\frac{H_n - n\mu}{\sqrt{n\sigma^2}} \leq z\right) = \int_{-\infty}^{z} \varphi(t)\, dt.$$

Thus, the central limit theorem is a natural basis to approximate the distribution of the mean from random samples in order to derive confidence intervals or statistical tests for the (unknown) mean. For a proof, see Meyer (1970, p. 250), or Çinlar (2011, p. 127).

Central Limit Theorem of Poisson Another limiting situation for the sum of variables, i.e. for $H_n = X_1 + X_2 + \cdots + X_n$ which is binomially distributed, was investigated by Poisson. Let the single summands be independent and binary with $P(X_i = 1) = p_n$ and consider a new parameter $\lambda = n \cdot p_n > 0$. For n tending to infinity and $X = \lim_n X_n$ it holds

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad \text{for } k = 0, 1, 2, \ldots,$$

i.e. the binomial distributions with this restriction converge to the Poisson distribution, which appears as the distribution of rare events as p_n tends to 0 as the product with n remains constant. For a proof, see Meyer (1970, p. 160), or Çinlar (2011, p. 137).

An example for modelling with the Poisson distribution is counting the atoms decaying in a specific period of time out of 1 kg of uranium U_{238}. The convergence above may be illustrated by the following ideal situation. If $n = 10$ raisins (points, events) should be distributed independently and randomly to 5 rolls (unit squares, observational unit), the number of raisins within a special roll is to be analysed. This corresponds to the random variable H_{10} with $X_i = 1$ if the ith raisin was attributed to it. Clearly, H_{10} is binomially distributed with parameters 10 and $1/5$ and expected number of raisins of $n \cdot p_n = 10 \cdot \frac{1}{5} = 2$. With $n = 100$ raisins and 50 rolls, the number of raisins in one special roll H_{100} is binomially distributed with 100 and $1/50$ and expected number of raisins of $n \cdot p_n = 100 \cdot \frac{1}{50} = 2$. In the latter situation, H_{100} is well approximated by the Poisson distribution with parameter $\lambda = 2$, which is the initial number of raisins per roll and is called the intensity of the process of distributing.

6.3 Standard Situations

Here we describe a few standard situations. We start with simpler processes such as those of Laplace and Bernoulli. We go on to more complex ideas of Markov chains and Brownian motion, typically studied at a university. It is remarkable that only a few distributions cover a wide range of applications. For modelling it is important to know the key idea behind the basic situation leading to that distribution. This explains the properties of the model and the modelled phenomenon.

Laplacean Experiments These are experiments where the equi-distribution is plausible on the basis of a physical symmetry. Conventional representations are spinners with equal sectors or urns filled with balls. For teaching, such experiments are useful to illustrate numerical probabilities to calibrate uncertainty.

Bernoulli Experiments The special case of a Laplacean experiment with two outcomes is known as a Bernoulli experiment. There are two different ways to explore this situation: to count “successes” in a specified number of trials leading to the binomial, or to wait for the next “success” leading to the geometric distribution.

Poisson Process Poisson experiments may be introduced as Bernoulli series in which the number of trials is high and the probability p is small. The Poisson process, however, describes a *genuine* random phenomenon of ‘producing’ events within time; heuristically, the process has to obey the following rules (see Meyer, p. 165).

(i) It does not matter when the observation of the process actually starts, the probabilities of various counts of events depend only on the *length* of observation.
(ii) For short periods the probability to have exactly one event is essentially the *intensity* λ of the process to produce events multiplied by the length of observation.
(iii) For short periods one may neglect the probability of two or more events.
(iv) Events occur independently in time.

The variable X which counts events in t units of time then follows a Poisson distribution:

$$P(X=k)=\frac{(\lambda t)^k}{k!}e^{-\lambda t},\quad \text{for } k=0,1,2,\ldots;\ \lambda>0.$$

Beside the probabilities of the number of events (Poisson), it is of interest to derive the probability for the waiting time for the next event (exponential distribution).

Elementary Errors Due to the central limit theorem, the distribution of an observed quantity X can be approximated by the normal distribution if it can be thought of as the result of a sum, i.e. $X=X_1+X_2+\cdots+X_n$. The analogy of small errors (the summands) contributing to generate the final quantity accounts for the ubiquity of the normal distribution. Historically, this reading of the central limit theorem has been a driving force and is still given in textbooks (Meyer 1970, p. 251).

Stochastic Processes The independence assumption is at the core of the random sample idea and also at the basis of central theorems like the law of large numbers and central limit theorem. There was another situation emerging from applications in physics that needed growing attention, where a probability measure was needed for an infinite dimensional Cartesian product: processes describing the change of a variable under scrutiny with progressing time. A slight shift in the description of the Poisson process will illustrate the conceptual change:

The variable X_t counts the number of events in the interval $(0, t]$. To be called a *Poisson* process, it has to fulfil the following conditions:

1. $X_0 = 0$. At the beginning the count starts at 0.
2. The process has stationary increments, i.e. the growth during time $(t, t + s]$ depends only on the length and not on the starting point of observation and its distribution depends only on s.
3. The process has independent increments, i.e. the growth in disjoint intervals is stochastically independent, i.e. for $t_0 < t_1 < t_2 < \cdots$ the increments $X_{t_1} - X_{t_0}$, $X_{t_2} - X_{t_1}, \ldots$ are independent.
4. With the exception of a set of zero measure, the trajectories $X_t(\omega)$ jump at most by 1 unit.

From the conditions one can conclude the following: The number of events in $(0, t]$ follows a Poisson distribution with parameter $\lambda \cdot t$, the waiting time for the next event is exponentially distributed with parameter λ, the location of any event in $(0, t]$ is uniformly distributed over this interval. The process is a Markov process with continuous time. Condition 2 corresponds to (i) and part of (ii) (the rest is not even required), 3 corresponds to (iv), while 4 corresponds to (iii).

An important Markov process was used as a tool in physics to describe the drifting of particles suspended in a fluid. A random walk in two dimensions means walking along the lattice of points with equal probabilities of 1/4 for continuing up, down, right, or left. By refining the grid more and more, and the central limit theorem, the following process—a so-called Wiener process—was motivated. Let X_t be the position of a particle at time t (usually this was a point in three-dimensional space).

1. $X_0 = 0$. The particle starts at the origin.
2. The process has stationary increments, i.e. the growth during time $(s, t]$ depends only on the length and not on the starting point of observation and its distribution depends only on the length $t - s$ and is, in fact, a normal distribution with expected value 0 and variance $t - s$.
3. The process has independent increments, i.e. the growth in disjoint intervals is stochastically independent, i.e. for $t_0 < t_1 < t_2 < \cdots$ the increments $X_{t_1} - X_{t_0}$, $X_{t_2} - X_{t_1}, \ldots$ are independent.
4. With the exception of a set of zero measure, the trajectories $X_t(\omega)$ are continuous functions of time.

For a similar formulation of the Poisson and Wiener process, see van Zanten (2010, p. 3). Most modern expositions like Çinlar (2011, p. 171), use the concept of martingales to define stochastic processes and are therefore less accessible. The phenomenon which was described in physics by such a model is Brownian motion. It is striking that at the time when such applications boosted the theory of thermodynamics, the foundations of probability was still not laid and probability was more or less justified by Laplacean equi-probability and interpreted as relative frequencies in long series of independent trials. But the reader should note that with the Markov processes above there was no independence in trials. And, of course, there was no firm foundation of a probability measure on an infinite-dimensional space as the trajectories of the process were the elements of the probability space. It was high time to solve the situation and in his 1900 address, Hilbert included the axiomatic basis of probability and mechanics as among the most urgent mathematical problems.

6.4 Kolmogorov's Axiomatic Foundation of Probability

Instead of the infinite random sequences of von Mises, Kolmogorov returns to a fundamental probability set which describes the potential of the experiment to produce outcomes in one trial. The question is how to define unambiguously the probability on the sample space and still have repeated independent experiments. The solution was to use measure theory to mathematize probability like other measures such as area or weight, and add the concept of independence between different trials subsequently. Repeated trials are modelled by the sample space which is built from the Cartesian product of the sample space of single experiments.

The Axioms Kolmogorov (1933) developed a system of axioms for the special case of a finite sample space $S = \{x_1, x_2, \ldots, x_m\}$. Instead of using the power set of S (which is feasible here), he refers to a field $\mathcal{F}$ of events (an algebra of events) as the domain for a probability function. A field $\mathcal{F}$ thereby is a system of subsets of S, which has the property that the set operations of union, intersection and complementation (finitely) applied on elements of $\mathcal{F}$ always yield a set that belongs to $\mathcal{F}$ (i.e. the field is closed under the usual set operations). His axioms are (p. 2):

(I) $\mathcal{F}$ is a field of sets.
(II) $\mathcal{F}$ contains the set S.
(III) To each set A in $\mathcal{F}$ a non-negative real number $P(A)$, the probability of A, is assigned.
(IV) $P(S)$ equals 1.
(V) If A and B have no element in common, then $P(A \cup B) = P(A) + P(B)$.

The conditional probability is defined for a fixed event A with $P(A) > 0$ as

$$P_A(B) := \frac{P(A \cap B)}{P(A)}$$

with the justification that the function P_A fulfils the axioms. Two events A and B are defined as independent if

$$P(A \cap B) = P(A) \cdot P(B),$$

which is shown as equivalent to the following relations if both events have a positive probability:

$$P_A(B) = P(B) \quad \text{and} \quad P_B(A) = P(A).$$

At this point, Kolmogorov proceeds (p. 14) to define probability measures on *infinite* spaces by adding one more axiom, the so-called continuity of a probability measure:

(VI) For a strictly decreasing sequence of events $A_1 \supset A_2 \supset \cdots \supset A_n \supset \cdots$ of $\mathcal{F}$ with $\bigcap_n A_n = \emptyset$ the following equation holds $\lim_n P(A_n) = 0$ for $n \to \infty$.

Ensuring this axiom, the countable additivity is proved as a theorem, i.e. it holds that

$$P\left(\bigcup_n A_n\right) = \sum_{n=1}^{\infty} P(A_n) \quad \text{for any sequence } A_n \text{ of } \mathcal{F} \text{ with } A_i \cap A_j = \emptyset \text{ for } i \neq j;$$

i.e., the additivity holds for any *sequence* of pairwise disjoint events.

In the rest of his seminal work, Kolmogorov refers to the system $\mathcal{F}$ as a σ-field (σ-algebra), i.e. he requires the field to be closed also under *countably* infinite applications of the usual set operations by the following argument:

> Only in the case of [σ] fields of probability do we obtain full freedom of action, without danger of the occurrence of events having no probability. (p. 16)

Modern representations of Kolmogorov's axioms prefer to relocate his first two axioms into the denotation of a probability measure P as a real function on a σ-algebra $\mathcal{F}$, i.e.

$$P : \mathcal{F} \to \mathbf{R}$$

and refer to the σ-additivity instead of the continuity so that the axioms read as the following conditions on the function P:

(A$_1$) $P(A) \geq 0$ for any event A from $\mathcal{F}$.

(A$_2$) $P(S) = 1$ for the whole sample space S.

(A$_3$) If $A_0, A_1, \ldots$ is a sequence of mutually exclusive events from $\mathcal{F}$, then $P(\bigcup_{n=0}^{\infty} A_n) = \sum_{n=0}^{\infty} P(A_n)$.

The first two conditions mean that probabilities are non-negative and that certainty is characterized by a value of 1. The substantial condition of *σ-additivity* embodied in A$_3$ means that, mathematically, probability is a measure. It may be regarded in some respect as analogous to 'area', 'mass', or 'weight', measures which also share the additivity property.

Kolmogorov (p. 1) believes that

> The theory of probability, as a mathematical discipline, can and should be developed from axioms, in exactly the same way as Geometry and Algebra [...] all further exposition must be based exclusively on these axioms, independent of the usual concrete meaning of these elements and their relations.

The special choice of a set of axioms has deep consequences on semantics. The axioms are the foundation of the theory and are simultaneously at the interface between theory and reality. The basic axioms could be considered as models of intuitive ideas of probability to be sharpened by the theory. This might be thought of as the historical genesis of ideas in general and also in the sense of how ideas settle down in an individual's learning.

Distribution Functions In his paper, Kolmogorov uses the concept of a (cumulative) distribution function extensively (p. 19). This only makes sense if the concept fully characterizes a probability measure. Therefore, before turning to examples of probability functions on the space of real numbers, Kolmogorov (p. 16) proves an abstract extension theorem, which states that a probability measure on a field $\mathcal{F}$ can be uniquely extended to the smallest σ-field $[\mathcal{F}]$, which contains $\mathcal{F}$. The importance of that theorem cannot be overestimated as in the real numbers the domain of a probability function, the so-called Borel σ-algebra $\mathcal{B}$ of events is precluded from any intuitive access. But, luckily, it is possible to focus on a simple generating system of it which is built up of specific intervals $(a, b]$ (a and b could be real numbers or $\pm\infty$) and their finite unions, which form a field $\mathcal{F}$. Even more, a simpler generating system suffices as there are more general extension theorems that do not require the structure of a field on the generator.

$$\mathcal{C}^* = \big\{(a, b] \mid a, b \in \mathbf{R}\big\} \quad \text{or} \quad \mathcal{C} = \big\{(-\infty, x] \mid x \in \mathbf{R}\big\}.$$

That means, of any probability measure P on the Borel σ-algebra $\mathcal{B}$ on $\mathbf{R}$ it suffices to know the values of P on sets of $\mathcal{C}$. Or, conversely, the pre-probabilities fulfilling the axioms on sets of $\mathcal{C}$ uniquely determine a probability measure P on the Borel sets $\mathcal{B}$.

The complicated story with the Borel sets and σ-algebras has its origin in the following theorem. There can be no probability measure P fulfilling all the axioms for *all* subsets of $\mathbf{R}$. A contradiction can be derived if it is assumed that a probability can be attributed to *all* subsets with all the named properties. To avoid this, the domain of the function P has to be restricted to a true subset of the power set of $\mathbf{R}$. The natural structure of all *admissible* sets (for probability) is that of a σ-algebra, which is a system of subsets that is closed under the usual set operations, countably often applied in any order. The Borel sets $\mathcal{B}$ serve this purpose perfectly.

Probability Measures on Infinite-Dimensional Spaces What distinguishes the theory based on this set of axioms from measure theory is the concept of independence, which is part of a fundamental definition, but not, interestingly, part of the axioms. This independence relation is the key assumption of fundamental theorems like Bernoulli's law of large numbers. Such theorems established a link from the structural approach to the frequentist interpretation and thus contributed to the immediate acceptance of Kolmogorov's axioms within the scientific community.

More complications arise in the case of infinite sample spaces for single trials as the sets of the form $E_1 \times E_2$ (which are known as *cylinder sets*) are only a small part of all subsets of $S_1 \times S_2$. In practice, events in the combined experiment may not be of this special form, e.g. in spinning a pointer twice, consider the event 'position of the trials differ by more than $\pi/4$'. This complication is not a conceptual difficulty of probability as a phenomenon to be modelled but is linked to specific aspects of mathematics. For applications it is fortunate that assigning probabilities to cylinder sets is sufficient to uniquely determine an extension of this assignment to probabilities of all events. For infinite-dimensional spaces, the trick was used to start from cylinder sets defined on a finite number of coordinates (cf. Kolmogorov, p. 27).

Lebesgue Integral There is another unifying element in Kolmogorov's fundamental paper, namely the use of integrals for calculating probabilities and expected values. The distribution function F_X uniquely defines a probability measure P_X on the Borel sets. Probabilities and expected values may be written as integrals as follows

$$P_X(A) = \int_A dF_X(x) \quad \text{and} \quad E(X) = \int x\, dF_X(x),$$

which are Lebesgue–Stieltjes integrals. This unified the theory of probability of discrete and continuous distributions. For practical needs these integrals can be evaluated as ordinary sums or—in case of intervals as events, i.e. $A = (a, b)$, as integrals; for the bulk of applications the Lebesgue integral is not even required in evaluating the integrals involved, as the Riemann integral suffices for the most important distributions:

$$P_X(A) = \begin{cases} \sum_{i \in A} p_i & X \text{ with discrete probabilities } p_i = P(X = i), \\ \int_a^b f_X(x)\, dx & X \text{ with absolutely continuous density } f_X = \frac{d}{dx} F_X, \end{cases}$$

$$E(X) = \begin{cases} \sum_i i p_i, \\ \int x f_X(x)\, dx. \end{cases}$$

These ideas are well beyond school level. Yet, it is important to be aware of and remember this theory. It forms the deep foundation on which probability has been built.

It is important to note that the many deep results like the central limit theorem were derived before a sound axiomatic basis had been established and they retained their validity and importance after the axiomatization. What was different is the prestige probability gained as a scientific discipline, which then attracted many young researchers to the field. Furthermore, axiomatization paved the way to probability distributions on infinite-dimensional spaces and the field of stochastic processes, which revolutionized not only physics but many other fields like financial mathematics. For example, the price of an option in the financial market is derived by the solution of a stochastic differential equation of a stochastic process similar to the one described above.

7 Conclusions

Modern expositions of probability such as Çinlar (2011) have reached an elegance of mathematical standard, which is sometimes in sharp contrast to the philosophical framework. The situation resembles somehow the early culmination of development with Laplace working on the central limit theorem but expressing a naïve determinism that probability is only for those ignorant of the causes. The general philosophical debate such as von von Plato (1994) in the context of physics shows dramatically that probability without a firm philosophical footing misrepresents its scope in limits as well as in reach.

The standard paradigm is to interpret probability initially as equal possibilities and then as the limit of relative frequencies—or as relative frequencies from samples large enough. All the concepts of inferential statistics from the objectivist position heavily draw on this paradigm. This way was paved by Kolmogorov's own views on his axioms justifying the frequentist conception of probability. The reaction from the subjectivist position was fierce, led by de Finetti who ironically noted, in capital letters that "PROBABILITY DOES NOT EXIST" (1974, p. x); he sees probability as a way to *think about* the world. But their mathematical exposition—grounded on axioms of rational behaviour and rational updating of probabilistic information via Bayes' formula—normally uses a different terminology and a unified exposition of probability respecting both subjectivist and objectivist ideas has not been published, yet.

Urgent topical problems on the objectivist side are that small probabilities are growing in importance, yet data is missing or is highly unreliable, or originates from qualitative knowledge. Probability is often used more in the sense of a scenario which means that probabilistic models are used as a heuristic to explore reality instead of finding the best-fitting model and determining the "best" solution relative to it for the real problem under scrutiny (see Borovcnik and Kapadia 2011). Another source of confusion is an adequate understanding of statistical methods that reminds one of the historical problems to understand the puzzling examples or paradoxes on the sole basis of an objectivist probability.

What remains of probability if it is deprived of its main interpretation as relative frequencies is hard to tell. Some key properties such as the additivity of expected values or the key idea behind any distribution, which sets the scene for a structure of situations—a structure that goes beyond and behind the fact that the relative frequencies do fit and can be modelled by it, are illustrated in Borovcnik (2011). The link to relative frequencies remains too dominant. The conception of a degree of belief and how to revise it by new data gives guidance to understand the notions and related methods much more easily. The discussion of the paradoxes shows that such ideas enhance understanding.

A well-balanced exposition covering subjectivist and objectivist probability seems unlikely. Barnett (1982) marked a promising step into this direction with a comparative analysis of the positions but this remains an isolated project within the statistical community despite the fact that the theory based on Kolmogorov's axioms could serve as a common language. Barnett (1982, p. 69) notes that these axioms are a "mathematical milestone" in laying firm foundations; they remain, however, a "philosophical irrelevance" in terms of explaining what probability really is.

Returning to paradoxes and fallacies, they can be entertaining. They raise class interest and motivation. Discussion of these ideas can help to

- analyse obscure or complex probabilistic situations properly;
- understand the basic concepts better;
- interpret formulations and results more effectively;
- balance and shift between different interpretations of probability;
- educate probabilistic intuition and reasoning more firmly.

The misconceptions in the examples show that probabilistic intuitions seem to be one of the poorest among our natural and developed senses. Perhaps, this is a reflection of the desire for deterministic explanation. People have great difficulty in grasping the origins and effects of chance and randomness: they search for pattern and order even amongst chaos. Or, it may be due to an education too restricted to one perception of probability. The examples above illustrate the gap between intuition and mathematical theory, particularly because stochastic reasoning has no empirical control to revise inadequate strategies. Paradoxes and puzzles highlight these difficulties as signs of a cognitive conflict between an intuitive level of reasoning and formalized, mathematical arguments. In a paradox, the 'objective' side is inadequate though intuitively straightforward, whereas in a puzzle the objective side is adequate but intuitively inaccessible. Empirical research in using paradoxes in teaching is limited, though a promising start has been made by Vanscó (2009) with trainee teachers; this now needs to be replicated in schools. Our own long and varied experience leads us to assert that planned discussion of paradoxes and puzzles fosters individual conceptual progress of children and students in learning probability.

References

Barnett, V. (1982). *Comparative statistical inference* (1st ed. 1973). New York: Wiley.

Bayes, T. (1763). An essay towards solving a problem in the Doctrine of Chances. *Philosophical Transactions of the Royal Society*, *53*, 370–418. Reprinted in E. S. Pearson, & M. G. Kendall (1970), *Studies in the history of statistics and probability* (Vol. 1, pp. 131–154). London: Griffin.

Bernoulli, J. (1713/1987). *Ars conjectandi*. Basel: Impensis Thurnisiorum, Fratrun. Translation of 4th part by N. Meunier. Rouen: Institut de Recherche sur l'Enseignement Mathematique. German translation by R. Haussner (1899), *Wahrscheinlichkeitsrechnung*. Leipzig: Engelmann.

Bernoulli, D. (1738/1954). Specimen theoriae novae de mensura sortis. *Commentarii Academiae Scientiarum Imperialis Petropolitanae*, *5*, 175–192; English translation: L. Sommer, Exposition of a new theory on the measurement of risk. *Econometrica*, *22*, 23–36.

Bertrand, J. (1888). *Calcul des probabilités*. Paris: Gauthier-Villars.

Bickel, P. J., Hammel, E. A., & O'Conell, W. J. (1977). Sex bias in graduate admissions: data from Berkeley. In W. B. Fairley, & F. M. Mosteller (Eds.), *Statistics and public policy*. Reading: Addison-Wesley.

Blyth, C. R. (1972). Some probability paradoxes in choice from among random alternatives. *Journal of the American Statistical Association*, *67*, 366–381. With comments by Lindley, D. V., Good, I. J., Winkler, R. L., and Pratt, J. W.

Borel, É. (1909). Les probabilités dénombrables et leurs applications arithmétique. *Rendiconti del Circolo Matematico di Palermo*, *27*(1), 247–271.

Borovcnik, M. (1987). Revising probabilities according to new information—a fundamental stochastic intuition. In R. Davidson, & J. Swift (Eds.), *Proceedings of the second international conference on teaching statistics* (pp. 298–302). Victoria: University of Victoria. Online: www.stat.auckland.ac.nz/~iase/publications/icots2/Borovcnik.pdf.

Borovcnik, M., & Bentz, H.-J. (1991). Empirical research in understanding probability. In R. Kapadia, & M. Borovcnik (Eds.), *Mathematics education library: Vol. 12. Chance encounters: probability in education* (pp. 73–105). Dordrecht: Kluwer Academic.

Borovcnik, M., & Kapadia, R. (2011). Modelling in probability and statistics—key ideas and innovative examples. In J. Maaß, & J. O'Donoghue (Eds.), *Real-world problems for secondary school students—case studies* (pp. 1–44). Rotterdam: Sense Publishers.

Borovcnik, M., & Peard, R. (1996). Probability. In A. Bishop, K. Clements, C. Keitel, J. Kilpatrick, & C. Laborde (Eds.), *International handbook of mathematics education* (pp. 239–288). Dordrecht: Kluwer Academic.

Borovcnik, M., Bentz, H.-J., & Kapadia, R. (1991). A probabilistic perspective. In R. Kapadia, & M. Borovcnik (Eds.), *Mathematics education library: Vol. 12. Chance encounters: probability in education.* (pp. 27–71). Dordrecht: Kluwer Academic.

Carranza, P., & Kuzniak, A. (2009). Duality of probability and statistics teaching in French education. In C. Batanero et al. (Eds.), *Joint ICMI/IASE study: teaching statistics in school mathematics* (5 pp.). Online: www.stat.auckland.ac.nz/~iase/publications/rt08/T1P2_Carranza.pdf.

Çinlar, E. (2011). *Probability and stochastics*. Berlin: Springer.

d'Alembert, J. (1754). Croie ou pile. In D. Diderot, & J. d'Alembert (Eds.), *Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers* (Vol. 4, pp. 512–513). Paris: Société des Gens de Lettres.

David, F. N. (1962). *Games, gods and gambling*. London: Griffin.

de Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, *7*, 1–68. Foresight: Its logical laws, its subjective sources (Chaps. 1–4 translated by H.E. Kyburg Jr.). In S. Kotz, & N.L. Johnson (1992), *Breakthroughs in statistics. Vol. I. Foundations and basic theory* (pp. 134–174). New York: Springer.

de Finetti, B. (1974). *Theory of probability*. New York: Wiley. Translated by A. Machi, & A. Smith.

de Moivre, A. (1738/1967). *The doctrine of chances* (2nd ed., fuller, clearer, and more correct than the first). London: Woodfall. Reprint of 3rd ed. 1967. New York: Chelsea. 1st ed. 1718. London: Pearson.

Falk, R., & Bar-Hillel, M. (1983). Probabilistic dependence between events. *Two-Year College Mathematics Journal*, *14*, 240–247.

Fermat, P., & Pascal, B. (1679). Correspondence 1654. In P. Fermat (Ed.), *Varia Opera Mathematica*. Toulouse: Joannem Pech. English translation in F. N. David (1962), *Games, gods and gambling* (pp. 229–253). London: Griffin.

Freudenthal, H. (1973). *Mathematics as an educational task*. Dordrecht: Reidel.

Galilei, G. (ca. 1613–1623, 1962). *Sopra le Scoperte dei Dadi*. Fragment on dice. Published under the title "Considerazione sopra il Giuoco dei Dadi" in A. Favaro (1890–1897) (Ed.), *Le Opere di Galileo Galilei* (Vol. VIII, pp. 591–594). Firenze: Tipografia La Barbera. English translation in David, F. N. (1962), *Games, gods and gambling, Appendix 2* (pp. 192–195, translated by E. H. Thorne).

Gardner, M. (1959). *The Scientific American book of mathematical puzzles & diversions*. New York: Simon & Schuster.

Gardner, M. (2006). *Aha! A two volume collection: aha! Gotcha aha! Insight*. Washington: Math. Assoc. of America.

Gigerenzer, G. (2002). *Calculated risks: how to know when numbers deceive you*. New York: Simon & Schuster.

Huygens, C. (1657). De ratiociniis in ludo aleae. In F. v. Schooten: *Exercitationes matematicae*, Leyden: Elsevirii.

Kahneman, D., & Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica*, *XLVII*, 263–291.

Kapadia, R., & Borovcnik, M. (Eds.) (1991). *Mathematics education library: Vol. 12. Chance encounters: probability in education*. Dordrecht: Kluwer Academic.

Kolmogorov, A. N. (1933). *Ergebnisse der Mathematik, 2. Band, Heft 3. Grundbegriffe der Wahrscheinlichkeitsrechnung*. Berlin: Springer. Reprinted 1977 (62 pp.). English translation 1956: *Foundations of the theory of probability*. New York: Chelsea.

Konold, C. (1989). Informal conceptions of probability. *Cognition and Instruction*, *6*, 59–98.

Laplace, P. S. (1812/1951). Essai philosophique sur les probabilités. *Journal de l'École Polytechnique*, *VII/VIII*, 140–172. English translation: *A philosophical essay on probabilities* (extended version). New York: Dover.

Maistrov, L. E. (1974). *Probability theory. A historical sketch*. New York: Academic Press.

Meyer, P. L. (1970). *Introductory probability and statistical applications* (2nd ed.). Reading: Addison-Wesley.

Pacioli, L. (1494). *Summa de arithmetica, geometria, proportioni et proportionalitá*. Venedig: Paganino de' Paganini.

Palm, G. (1983). Wo kommen die Wahrscheinlichkeiten eigentlich her? *Der Mathematik-Unterricht*, *29*(1), 50–61.

Price, R. (1764–1765). A demonstration of the second rule in Bayes' essay. *Philosophical Transactions of the Royal Society of London*, *54*, 296–297; *55*, 310–325.

Reinhardt, H. E. (1981). Some statistical paradoxes. In A. P. Shulte, & J. R. Smart (Eds.), *Teaching statistics and probability* (pp. 100–108). Reston: National Council of Teachers of Mathematics.

Schnorr, C. P. (1971). *Lecture notes in mathematics: Vol. 218. Zufälligkeit und Wahrscheinlichkeit. Eine algorithmische Begründung der Wahrscheinlichkeitstheorie*. Berlin: Springer.

Shaughnessy, J. M. (2003). Research on students' understanding of probability. In J. Kilpatrick, W. G. Martin, & D. Schifter (Eds.), *A research companion to principles and standards for school mathematics* (pp. 216–226). Reston: National Council of Teachers of Mathematics.

Székely, G. J. (1986). *Paradoxes in probability and mathematical statistics*. Dordrecht: Reidel.

Vanscó, Ö. (2009). Parallel discussion of classical and Bayesian ways as an introduction to statistical inference. *International Electronic Journal in Mathematics Education*, *4*(3), 291–322. Online: www.iejme.com/032009/main.htm.

Venn, J. (1888). *The logic of chance* (3rd ed.). London: Macmillan. 1st ed., published 1866.

von Mises, R. (1919). Grundlagen der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, *5*, 52–99.

von Plato, J. (1994). *Creating modern probability: its mathematics, physics and philosophy in historical perspective*. Cambridge: Cambridge University Press.

Wavre, R. (Ed.) (1938–1939). Colloque consacré à la théorie des probabilités, Proceedings of the Geneva conference 1937; published in 8 Vols. Paris: Hermann. *Actualités Scientifiques et Industrielles*, *734–740*, 766.

Zanten, H. V. (2010). *An introduction to stochastic processes in continuous time*. Leiden: University of Leiden. Online: www.math.leidenuniv.nl/~spieksma/colleges/sp-master/sp-hvz1.pdf.