Understanding Probability

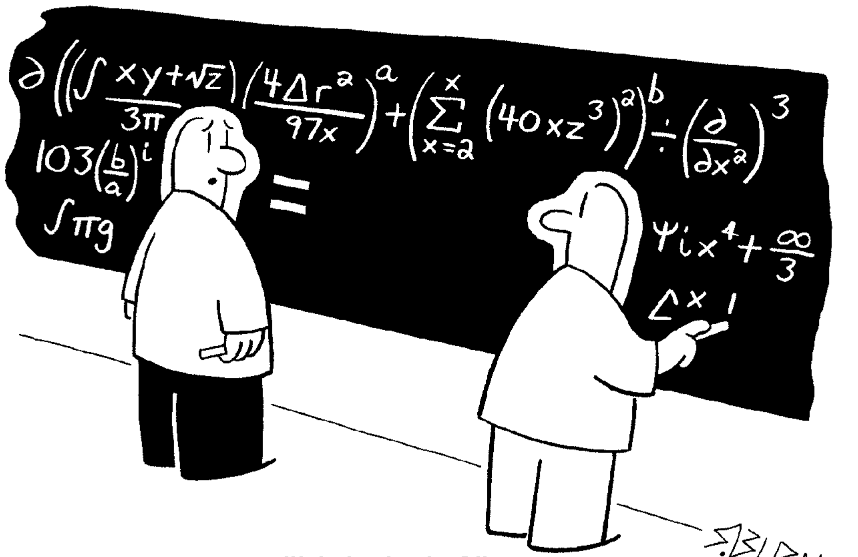
Third edition, Cambridge University Press

Henk Tijms

Vrije University, Amsterdam

First published 2004
Second edition 2007
Third edition 2012
Printed in the United Kingdom at the University Press, Cambridge

PART TWO: ESSENTIALS OF PROBABILITY



"Little help?"

7
Foundations of probability theory

Constructing the mathematical foundations of probability theory has proven to be a long-lasting process of trial and error. The approach consisting of defining probabilities as relative frequencies in cases of repeatable experiments leads to an unsatisfactory theory. The frequency view of probability has a long history that goes back to Aristotle. It was not until 1933 that the great Russian mathematician Andrej Nikolajewitsch Kolmogorov(1903–1987) laid a satisfactory mathematical foundation of probability theory. He did this by taking a number of axioms as his starting point, as had been done in other fields of mathematics. Axioms state a number of minimal requirements that the mathematical objects in question (such as points and lines in geometry) must satisfy. In the axiomatic approach of Kolmogorov, probability figures as a function on subsets of a so-called sample space, where the sample space represents the set of all possible outcomes the experiment. The axioms are the basis for the mathematical theory of probability. As a milestone, the law of large numbers can be deduced from the axioms by logical reasoning. The law of large numbers confirms our intuition that the probability of an event in a repeatable experiment can be estimated by the relative frequency of its occurrence in many repetitions of the experiment. This law is the fundamental link between theory and the real world. Its proof has to be postponed until Chapter 14.

The purpose of this chapter is to discuss the axioms of probability theory and to derive from the axioms a number of basic rules for the calculation of probabilities. These rules include the addition rule and the more general inclusion-exclusion rule. Various examples will be given to illustrate the rules. In a few of the examples counting methods and binomial coefficients will be used. A short but self-contained treatment of these basic tools from combinatorics can be found in the Appendix at the end of this book.

7.1 Probabilistic foundations

A probability model is a mathematical representation of a real-word situation or a random experiment. It consists of a complete description of all possible outcomes of the experiment and an assignment of probabilities to these outcomes. The set of all possible outcomes of the experiment is called the *sample space*. A sample space is always such that one and only one of the possible outcomes occurs if the experiment is performed. Let us give a few examples.

- The experiment is to toss a coin once. The sample space is the set $\{H, T\}$, where H means that the outcome of the toss is a head and T that it is a tail. Each of the two outcomes gets assigned a probability of $\frac{1}{2}$ if the coin is fair.
- The experiment is to roll a die once. The sample space is the set $\{1, 2, \ldots, 6\}$, where the outcome i means that i dots appear on the up face. Each of the six outcomes gets assigned a probability of $\frac{1}{6}$ if the die is fair.
- The experiment is to choose a letter at random from the word statistics. The sample space is the set $\{s, t, a, i, c\}$. The probabilities $\frac{3}{10}$, $\frac{3}{10}$, $\frac{1}{10}$, $\frac{2}{10}$, and $\frac{1}{10}$ are assigned to the five outcomes s, t ,a, i, and c.
- The experiment is to repeatedly roll a fair die until the first six shows up. The sample space is the set $\{1, 2, \ldots\}$ of the positive integers. Outcome k indicates that the first six shows up on the kth roll. The probabilities $\frac{1}{6}, \frac{5}{6} \times \frac{1}{6}, (\frac{5}{6})^2 \times \frac{1}{6}, \ldots$ can be assigned to the outcomes $1, 2, 3, \ldots$.
- The experiment is to measure the time until the first emission of a particle from a radioactive source. The sample space is the set $(0, \infty)$ of the positive real numbers, where the outcome t indicates that it takes a time t until the first emission of a particle. Taking an appropriate unit of time, the probability $\int_a^b e^{-t}\, dt$ can be assigned to each time interval (a, b) on the basis of physical properties of radioactive material, where $e = 2.71828\ldots$ is the base of the natural logarithm.

Various choices for the sample space are sometimes possible. In the experiment of tossing a coin twice, a possible choice for the sample space is the set $\{HH, HT, TH, TT\}$. Another possible choice is the set $\{0, 1, 2\}$, where the outcome indicates the number of heads obtained. The assignment of probabilities to the elements of the sample space differs for the two choices.

In the first three examples above the sample space is a *finite* set. In the fourth example the sample space is a so-called *countably infinite* set, while in the fifth example the sample space is a so-called *uncountable* set. Let us briefly explain these basic concepts from set theory. The set of natural numbers (positive integers) is an infinite set and is the prototype of a countably infinite set. In general, a nonfinite set is called countably infinite if a one to one function exists which maps the elements of the set to the set of natural numbers. In other words, every element of the set can be assigned to a unique natural number and conversely each natural number corresponds to a unique element of the set. For example, the set of squared numbers $1, 4, 9, 16, 25, \ldots$ is countably infinite. Not all sets with an infinite number of elements are countably infinite. The set of all points on a line and the set of all real numbers between 0 and 1 are examples of infinite sets that are not countable. The German mathematician Georg Cantor (1845–1918) proved this result in the nineteenth century. This discovery represented an important milestone in the development of mathematics and logic (the concept of infinity, to which even scholars from ancient Greece had devoted considerable energy, obtained a solid theoretical basis for the first time through Cantor's work). Sets that are neither finite nor countably infinite are called uncountable, whereas sets that are either finite or countably infinite are called countable.

7.1.1 Axioms of probability theory

A probability model consists of a sample space together with the assignment of probability, where probability is a function that assigns numbers between 0 and 1 to subsets of the sample space. The axioms of probability are mathematical rules that the probability function must satisfy. In the informal Section 2.2.2, we discussed already these rules for the case of a finite sample space. The axioms of probability are essentially the same for a chance experiment with a countable or an uncountable sample space. A distinction must be made, however, between the sorts of subsets to which probabilities can be assigned, whether these subsets occur in countable or uncountable sample spaces. In the case of a finite or countably infinite sample space, probabilities can be assigned to each subset of the sample space. In the case of an uncountable sample space, weird subsets can be constructed to which we cannot associate a probability. These technical matters will not be discussed in this introductory book. The reader is asked to accept the fact that, for more fundamental mathematical reasons, probabilities can only be assigned to sufficiently well-behaved subsets of an uncountable sample space. In the case that the sample space is the set of real numbers, then essentially only those subsets consisting of a finite interval, the complement of each finite interval, and the union of each countable number of finite intervals are assigned a probability. These subsets suffice for practical purposes. The probability measure on the sample space is denoted by P. It assigns to each subset A a probability $P(A)$ and must satisfy the following properties:

Axiom 7.1 $P(A) \geq 0$ *for each subset* A.

Axiom 7.2 $P(A) = 1$ *when* A *is equal to the sample space.*

Axiom 7.3 $P\left(\bigcup\limits_{i=1}^{\infty} A_i\right) = \sum\limits_{i=1}^{\infty} P(A_i)$ *for every collection of pairwise disjoint subsets* $A_1, A_2, \ldots$.

The *union* $\bigcup_{i=1}^{\infty} A_i$ of the subsets $A_1, A_2, \ldots$ is defined as the set of all outcomes which belong to at least one of the subsets $A_1, A_2, \ldots$. The subsets $A_1, A_2, \ldots$ are said to be *pairwise disjoint* when any two subsets have no element in common. In probability terms, any subset of the sample space is called an *event*. If the outcome of the chance experiment belongs to A, the event A is said to *occur*. The events $A_1, A_2, \ldots$ are said to be *mutually exclusive* (or disjoint) if the corresponding sets $A_1, A_2, \ldots$ are pairwise disjoint.

The first two axioms simply express a probability as a number between 0 and 1. The crucial third axiom states that, for any sequence of mutually exclusive events, the probability of at least one of these events occurring is the sum of their individual probabilities. Starting with these three axioms and a few definitions, a powerful and beautiful theory of probability can be developed.

The standard notation for the sample space is the symbol Ω. An outcome of the sample space is denoted by ω. A sample space together with a collection of events and an assignment of probabilities to the events is called a *probability space*. For a finite or countably infinite sample space Ω, it is sufficient to assign a probability $p(\omega)$ to each element $\omega \in \Omega$ such that $p(\omega) \geq 0$ and $\sum_{\omega \in \Omega} p(\omega) = 1$. A probability measure P on Ω is then defined by specifying the probability of each subset A of Ω as

$$P(A) = \sum_{\omega \in A} p(\omega).$$

In other words, $P(A)$ is the sum of the individual probabilities of the outcomes ω that belong to the set A. It is left to the reader to verify that P satisfies the Axioms 7.1 to 7.3.

A probability model is constructed with a specific situation or experiment in mind. The assignment of probabilities is part of the translation process from a concrete context into a mathematical model. Probabilities may be assigned to events any way you like, as long the above axioms are satisfied. To make your choice of the probabilities useful, the assignment should result in a "good" model for the real-world situation.

Equally likely outcomes

In many experiments with finitely many outcomes $\omega_1, \ldots, \omega_N$ it is natural to assume that all these outcomes are equally likely to occur. In such a case, $p(\omega_i) = \frac{1}{N}$ for $i = 1, \ldots, N$ and each event A gets assigned the probability

$$P(A) = \frac{N(A)}{N},$$

where $N(A)$ is the number of outcomes in the set A. This model is sometimes called the classical probability model.

Example 7.1 John, Pedro and Rosita each roll one fair die. How do we calculate the probability that the score of Rosita is equal to the sum of the scores of John and Pedro?

Solution. The sample space of the chance experiment is chosen as $\{(i, j, k) : i, j, k = 1, \ldots, 6\}$, where the outcome (i, j, k) occurs if the score of John is i dots, the score of Pedro is j dots, and the score of Rosita is k dots. Each of the 216 possible outcomes is equally probable and thus gets assigned a probability mass of $\frac{1}{216}$. The score of Rosita is equal to the sum of the scores of John and Pedro if one of the 15 outcomes (1,1,2), (1,2,3), (2,1,3), (1,3,4), (3,1,4), (2,2,4), (1,4,5), (4,1,5), (2,3,5), (3,2,5), (1,5,6), (5,1,6), (2,4,6), (4,2,6), (3,3,6) occurs. The probability of this event is thus $\frac{15}{216}$.

Example 7.2 Three players enter a room and are given a red or a blue hat to wear. The color of each hat is determined by a fair coin toss. Players cannot see the color of their own hats, but do see the color of the other two players' hats. The game is won when at least one of the players correctly guesses the color of his own hat and no player gives an incorrect answer. In addition to having the opportunity to guess a color, players may also pass. Communication of any kind between players is not permissible after they have been given hats; however, they may agree on a group strategy beforehand. The players decided upon the following strategy. A player who sees that the other two players wear a hat with the same color guesses the opposite color for his/her own hat; otherwise, the player says nothing. What is the probability of winning the game under this strategy?

Solution. This chance experiment can be seen as tossing a fair coin three times. As sample space, we take the set consisting of the eight elements RRR, RRB, RBR, BRR, BBB, BBR, BRB and RBB, where R stands for a red hat and B for a blue hat. Each element of the sample space is equally probable and gets assigned a probability of $\frac{1}{8}$. The strategy is winning if one the six outcomes RRB, RBR, BRR, BBR, BRB or RBB occurs (verify!). Thus the probability of winning the game under the chosen strategy is $\frac{3}{4}$.

Uncountable sample space

The following two examples illustrate the choice of a probability measure for an uncountable sample space. These two examples deal with so-called geometric probability problems. In the analysis of the geometric probability problems we use the continuous analog of the assignment of probabilities in the probability model having a finite number of equally likely outcomes.

Example 7.3 You randomly throw a dart at a circular dartboard with radius R. It is assumed that the dart is infinitely sharp and lands on a completely random point on the dartboard. How do you calculate the probability of the dart hitting the bull's-eye having radius b?

Solution. The sample space of this experiment consists of the set of pairs of real numbers (x, y) with $x^2 + y^2 \leq R^2$, where (x, y) indicates the point at which the dart hits the dartboard. This sample space is uncountable. We first make the following observation. The probability that the dart lands exactly on a *prespecified* point is zero. It makes only sense to speak of the probability of the dart hitting a given region of the dartboard. This observation expresses a fundamental difference between a probability model with a finite or countably infinite sample space and a probability model with an uncountable sample space.The assumption of the dart hitting the dartboard on a completely random point is translated by assigning the probability

$$P(A) = \frac{\text{the area of the region } A}{\pi R^2}$$

to each subset A of the sample space. Hence the probability of the dart hitting the bull's-eye is $\pi b^2/(\pi R^2) = b^2/R^2$.

Example 7.4 A floor is ruled with equally spaced parallel lines a distance D apart. A needle of length L is dropped at random on the floor. It is assumed that $L \leq D$. What is the probability that the needle will intersect one of the lines? This problem is known as Buffon's needle problem.

Solution. This geometric probability problem can be translated into the picking of a random point in a certain region. Let y be the distance from the center of the needle to the closest line and let x be the angle at which the

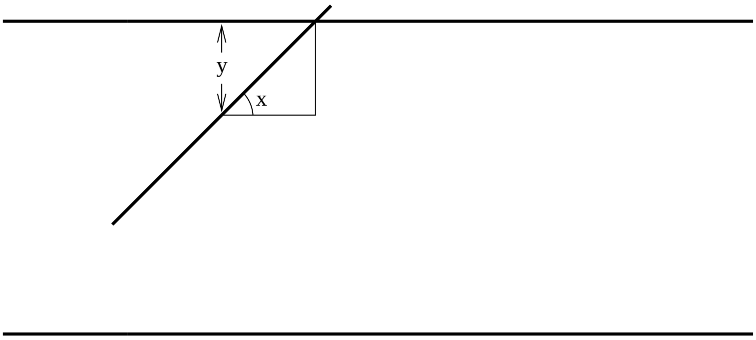


Fig. 7.1. The landing of Buffon's needle.

needle falls, where x is measured against a line parallel to the lines on the floor; see Figure 7.1. The sample space of the experiment can be taken as the rectangle R consisting of the points (x, y) with $0 \leq x \leq \pi$ and $0 \leq y \leq \frac{1}{2}D$. Dropping the needle at random on the floor can be seen to be equivalent to choosing a random point in the rectangle R. The needle will land on a line only if the hypotenuse of the right-angled triangle in Figure 7.1 is less than half of the length L of the needle. That is, we get an intersection only if $\frac{y}{\sin(x)} < \frac{1}{2}L$. Thus, the probability that the needle will intersect one of the lines equals the probability that a point (x, y) chosen at random in the rectangle R satisfies $y < \frac{1}{2}L\sin(x)$. In other words, the area under the curve $y = \frac{1}{2}L\sin(x)$ divided by the total area of the rectangle R gives the probability of an intersection. This ratio is

$$\frac{\int_0^\pi \frac{1}{2}L\sin(x)\,dx}{\frac{1}{2}\pi D} = \left.\frac{-L\cos(x)}{\pi D}\right|_0^\pi$$

and so

$$P(\text{needle intersects one of the lines}) = \frac{2L}{\pi D}.$$

In the following problems you should first specify a sample space before calculating the probabilities.

Problem 7.1 Three fair dice are rolled. What is the probability that the sum of the three numbers shown is an odd number? What is the probability that the product of the three numbers shown is an odd number?

Problem 7.2 In a township, there are two plumbers. On a particular day three residents call village plumbers independently of each other. Each resident randomly chooses one of the two plumbers. What is the probability that all three residents will choose the same plumber?

Problem 7.3 Four black socks and five white socks lie mixed up in a drawer. You grab two socks at random from the drawer. What is the probability of having grabbed one black sock and one white sock?

Problem 7.4 Two letters have fallen out of the word Cincinnati at random places. What is the probability that these two letters are the same?

Problem 7.5 Two players A and B each roll one die. The absolute difference of the outcomes is computed. Player A wins if the difference is 0, 1, or 2; otherwise, player B wins. What is the probability that player A wins?

Problem 7.6 Independently of each other, two people think of a number between 1 and 10. What is the probability that five or more numbers will separate the two numbers chosen at random by the two people?

Problem 7.7 You have four mathematics books, three physics books and two chemistry books. The books are put in random order on a bookshelf. What is the probability of having the books ordered per subject on the bookshelf?

Problem 7.8 Three friends go to the cinema together on a weekly basis. Before buying their tickets, all three friends toss a fair coin into the air once. If one of the three gets a different outcome than the other two, that one pays for all three tickets; otherwise, everyone pays his own way. What is the probability that one of the three friends will have to pay for all three tickets?

Problem 7.9 You choose eleven times a letter at random from the word Mississippi without replacement. What is the probability that you can form the word Mississippi with the eleven chosen letters? *Hint*: it may be helpful to number the eleven letters as $1, 2, \ldots, 11$.

Problem 7.10 You choose ten times a number at random from the numbers 1, $2, \ldots,$ 100. What is the probability of choosing ten distinct numbers? What is the probability that the first number chosen is larger than each of the other nine numbers chosen?

Problem 7.11 The game of franc-carreau was a popular game in eighteenth-century France. In this game, a coin is tossed on a chessboard. The player wins if the coin does not fall on one of the lines of the board. Suppose now that a round coin with a diameter of d is blindly tossed on a large table. The surface of the table is divided into squares whose sides measure a in length, such that $a > d$. Define an appropriate probability space and calculate the probability of the coin falling entirely within the confines of a square. *Hint*: consider the position of the coin's middle point.

8
Conditional probability and Bayes

The concept of conditional probability lies at the heart of probability theory. It is an intuitive concept. To illustrate this, most people reason as follows to find the probability of getting two aces when two cards are selected at random from an ordinary deck of 52 cards. The probability of getting an ace on the first card is $\frac{4}{52}$. Given that one ace is gone from the deck, the probability of getting an ace on the second card is $\frac{3}{51}$. The desired probability is therefore $\frac{4}{52} \times \frac{3}{51}$. Letting A_1 be the event that the first card is an ace and A_2 the event that the second card is an ace, one intuitively applies the fundamental formula $P(A_1A_2) = P(A_1)P(A_2 \mid A_1)$, where $P(A_2 \mid A_1)$ is the notation for the conditional probability that the second card will be an ace given that the first card was an ace.

The purpose of this chapter is to present the basics of conditional probability. You will learn about the multiplication rule for probabilities and the law of conditional probabilities. These results are extremely useful in problem solving. Much attention will be given to Bayes' rule for revising conditional probabilities in light of new information. This rule is inextricably bound up with conditional probabilities. The odds form of Bayes' rule is particularly useful and will be illustrated with several examples. Following on from Bayes' rule, we explain Bayesian inference for discrete models and give several statistical applications.

8.1 Conditional probability

The starting-point for the definition of conditional probability is a chance experiment for which a sample space and a probability measure P are defined. Let A be an event of the experiment. The probability $P(A)$ reflects our knowledge of the occurrence of event A *before* the experiment takes place. Therefore the probability $P(A)$ is sometimes referred to as the *a pri-*

ori probability of A or the *unconditional* probability of A. Suppose now we are told that an event B has occurred in the experiment, but we still do not know the precise outcome in the set B. In light of this added information, the set B replaces the sample space as the set of possible outcomes and consequently the probability of the occurrence of event A changes. A conditional probability now reflects our knowledge of the occurrence of the event A given that event B has occurred. The notation for this new probability is $P(A \mid B)$.

Definition 8.1 *For any two events A and B with $P(B) > 0$, the conditional probability $P(A \mid B)$ is defined as*

$$P(A \mid B) = \frac{P(AB)}{P(B)}.$$

Here AB stands for the occurrence of both event A and event B. This is not an arbitrary definition. It can be intuitively reasoned through a comparable property of the relative frequency. Let's define the relative frequency $f_n(E)$ of the occurrence of event E as $\frac{n(E)}{n}$, where $n(E)$ represents the number of times that E occurs in n repetitions of the experiment. Assume, now, that in n independent repetitions of the experiment, event B occurs r times simultaneously with event A and s times without event A. We can then say that $f_n(AB) = \frac{r}{n}$ and $f_n(B) = \frac{r+s}{n}$. If we divide $f_n(AB)$ by $f_n(B)$, then we find that

$$\frac{f_n(AB)}{f_n(B)} = \frac{r}{r+s}.$$

Now define $f_n(A \mid B)$ as the relative frequency of event A in those repetitions of the experiment in which event B has occurred. From $f_n(A \mid B) = \frac{r}{r+s}$ we now get the following relationship:

$$f_n(A \mid B) = \frac{f_n(AB)}{f_n(B)}.$$

This relationship accounts for the definition of conditional probability $P(A \mid B)$. The relative frequency interpretation also tells us how a conditional probability must be estimated in a simulation study.

Example 8.1 Someone has rolled two dice. You know that one of the dice turned up a face value of six. What is the probability that the other die turned up a six as well?

Solution. Let A be the event that two sixes show up and B the event that at least one six shows up. The sample space of this experiment is the set $\{(i, j) \mid i, j = 1, \ldots, 6\}$, where i and j denote the outcomes of the two dice. A probability of $\frac{1}{36}$ is assigned to each element of the sample space. The event A is given by the set $A = \{(6,6)\}$ and the event B by the set $B = \{(1,6), \ldots, (5,6), (6,6), (6,5), \ldots, (6,1)\}$. The probability of event B occurring is $\frac{11}{36}$ and the probability of both event A and event B occurring is $\frac{1}{36}$. Given that you know that event B has occurred, the probability that event A has also occurred is $P(A \mid B)$. Applying the definition of conditional probability gives

$$P(A \mid B) = \frac{P(AB)}{P(B)} = \frac{1/36}{11/36}.$$

Hence the desired probability is $\frac{1}{11}$ (not $\frac{1}{6}$).

The above example illustrates once again how careful you have to be when you are interpreting the information a problem is conveying. The wording of the problem is crucial: you know that one of the dice turned up a six but you do not know which one. In the case where one of the dice had dropped on the floor and you had seen the outcome six for that die, the probability of the other die turning up a six would have been $\frac{1}{6}$. An intuitive explanation of the difference between the two probabilities is the fact that in the second scenario you have the additional information which one of the two dice (red and blue) shows a six.

Example 8.2 John, Pedro and Rosita are experienced dart players. The probability of John hitting the bull's eye in a single throw is $\frac{1}{3}$. This hitting probability is $\frac{1}{5}$ for Pedro and $\frac{1}{4}$ for Rosita. The three players each throw simultaneously one dart. Two of the darts hit the bull's eye and one of the darts misses the bull's eye. What is the probability that John is the one who missed?

Solution. The sample space of the chance experiment consists of the eight elements (H,H,H), (H,H,M), (H,M,H), (H,M,M), (M,H,H), (M,H,M), (M,M,H), and (M,M,M), where M stands for "miss" and H stands for "hit." The first component of each element of the sample space refers to John's throw, the second component refers to Pedro's throw, and the third component refers to Rosita's throw. By the independence of the outcomes of the individual throws, we assign the probability $\frac{1}{3} \times \frac{1}{5} \times \frac{1}{4} = \frac{1}{60}$ to the outcome (H,H,H), the probability $\frac{1}{3} \times \frac{1}{5} \times \frac{3}{4} = \frac{3}{60}$ to the outcome (H,H,M), the probability $\frac{1}{3} \times \frac{4}{5} \times \frac{1}{4} = \frac{4}{60}$ to the outcome (H,M,H), the probability $\frac{1}{3} \times \frac{4}{5} \times \frac{3}{4} = \frac{12}{60}$ to the outcome (H,M,M), the probability $\frac{2}{3} \times \frac{1}{5} \times \frac{1}{4} = \frac{2}{60}$ to the outcome (M,H,H), the probability $\frac{2}{3} \times \frac{1}{5} \times \frac{3}{4} = \frac{6}{60}$ to the outcome (M,H,M), the probability $\frac{2}{3} \times \frac{4}{5} \times \frac{1}{4} = \frac{8}{60}$ to the outcome (M,M,H), and the probability $\frac{2}{3} \times \frac{4}{5} \times \frac{3}{4} = \frac{24}{60}$ to the outcome

(M,M,M). Let A be the event that John misses and let B be the event that exactly two of the darts hit the target. The event AB occurs if the outcome (M,H,H) occurs and the event B occurs if one of the outcomes $(H,H,M),(H,M,H),(M,H,H)$ occurs. Thus, $P(AB) = \frac{2}{60}$ and $P(B) = \frac{3}{60} + \frac{4}{60} + \frac{2}{60} = \frac{9}{60}$. We are now ready to determine the conditional probability $P(A \mid B)$. Applying the formula $P(A \mid B) = P(AB)/P(B)$, we can conclude that

$$P(\text{John misses} \,|\, \text{exactly two darts hit the target}) = \frac{2}{9}.$$

Problem 8.1 Three fair dice are rolled. What is the conditional probability that the sum of the three faces is 10 given that the three dice are showing different faces?

Problem 8.2 You toss a nickel, a dime and a quarter, independently of each other. What is the conditional probability that the quarter shows up heads given that the coins showing up heads represent an amount of at least 15 cents?

Problem 8.3 Every evening, two weather stations issue a weather forecast for the next day. The weather forecasts of the two stations are independent of each other. On average, the weather forecast of station 1 is correct in 90% of the cases, irrespective of the weather type. This percentage is 80% for station 2. On a given day, station 1 predicts sunny weather for the next day, whereas station 2 predicts rain. What is the probability that the weather forecast of station 1 will be correct?

Problem 8.4 The experiment is to toss a fair coin once and to roll a fair die once. Let A be the event that the coin lands "heads" and B the event that the die lands "six". After the experiment, you are informed that at least one of the two events has occurred. What is the probability that both events have occurred and what is the probability that event A has occurred, given the information?

Problem 8.5 You simultaneously grab two balls at random from an urn containing two red balls, one blue ball and one green ball. What is the probability that you have grabbed two non-red balls given that you have grabbed at least one non-red ball? What is the probability that you have grabbed two non-red balls given that you have grabbed the green ball? Can you give an intuitive explanation of why the second probability is larger than the first one?

Problem 8.6 The following game is played in a particular carnival tent.

The carnival master has two covered beakers, each containing one die. He shakes the beakers thoroughly, removes the lids and peers inside. You have agreed that whenever at least one of the two dice shows an even number of points, you will bet with even odds that the other die will also show an even number of points. Is this a fair bet?

Problem 8.7 Suppose a bridge player's hand of thirteen cards contains an ace. What is the probability that the player has at least one more ace? What is the answer to this question if you know that the player had the ace of hearts? Can you explain why the second probability is larger than the first one?

8.1.1 Assigning probabilities by conditioning

The formula for the conditional probability $P(A \mid B)$ can be rewritten as

$$P(AB) = P(A \mid B)P(B).$$

This phrasing lines up more naturally with the intuitive way people think about probabilities. The result $P(AB) = P(A \mid B)P(B)$ is called the *multiplication rule* for probabilities. In many cases, this rule is used in attributing probabilities to elements of the sample space. In illustration of this, consider the experiment in which two marbles are randomly chosen without replacements from a receptacle holding seven red and three white marbles. One possible choice for the sample space of this experiment is the set consisting of four elements (R, R), (R, W), (W, W), and (W, R), where R stands for red and W for white. The first component of each element indicates the color of the first marble chosen and the second component the color of the second marble chosen. On grounds of the reasoning that $P(1^{\text{st}}$ marble is red$) = \frac{7}{10}$ and $P(2^{\text{nd}}$ marble is white $\mid 1^{\text{st}}$ marble is red$) = \frac{3}{9}$, we attribute the probability of $P(R, W) = \frac{7}{10} \times \frac{3}{9} = \frac{7}{30}$ to the element (R, W). In the same way we attribute the probabilities $P(R, R) = \frac{7}{10} \times \frac{6}{9} = \frac{7}{15}$, $P(W, W) = \frac{3}{10} \times \frac{2}{9} = \frac{1}{15}$, and $P(W, R) = \frac{3}{10} \times \frac{7}{9} = \frac{7}{30}$ to the remaining elements. It is common practice for this type of problem to assign probabilities to the elements of the sample space as a product of probabilities, one marginal and the others conditional. More generally, by a repeated application of the basic formula $P(A) = P(A \mid B)P(B)$, we have the following useful rule:

Rule 8.1 *For any sequence of events $A_1, A_2, \ldots, A_n$,*

$$\begin{aligned}&P(A_1 A_2 \cdots A_n)\\ &= P(A_1) \times P(A_2 \mid A_1) \times P(A_3 \mid A_1 A_2) \times \cdots \times P(A_n \mid A_1 A_2 \cdots A_{n-1}).\end{aligned}$$

Many probability problems with a solution by counting arguments can be more easily tackled by using conditional probabilities. To illustrate this, consider question (a) from Example 7.9. For $i = 1, \ldots, 4$, let A_i be the event that two teams from the same country play against each other in the ith match. The four matches are sequentially determined by lot. The probability asked in question (a) is then given by $P(A_1A_2A_3A_4)$. Obviously, $P(A_1) = \frac{1}{7}$, $P(A_2 \,|\, A_1) = \frac{1}{5}$, $P(A_3 \,|\, A_1, A_2) = \frac{1}{3}$ and $P(A_4 \,|\, A_1, A_2, A_3) = 1$ and so the desired probability $P(A_1A_2A_3A_4)$ equals $\frac{1}{7} \times \frac{1}{5} \times \frac{1}{3} \times 1 = 0.0095$. As another illustration, what is the probability that it takes 10 or more cards before the first ace appears if cards are randomly drawn one by one from an ordinary deck of 52 cards? To answer this question by counting arguments, we take as sample space the set of all possible permutations of the integers $1, 2, \ldots, 52$. Each of the 52! elements is equally likely. Let A be the event that it takes 10 or more cards before the first ace appears. The set A contains $48 \times 47 \times \cdots \times 40 \times 41!$ elements of the sample space and so $P(A) = 48 \times 47 \times \cdots \times 40 \times 41!/52! = 0.4559$. To answer the question with conditional probabilities, define A_i as the event that the ith card drawn is not an ace. Applying Rule 8.1, we find $P(A_1A_2 \cdots A_9) = \frac{48}{52} \times \frac{47}{51} \times \cdots \times \frac{40}{44} = 0.4559$.

The following important remark is made. In using conditional probabilities, you usually perform the probability calculations without explicitly specifying a sample space; an assignment of probabilities to properly chosen events suffices. Solving a probability problem by counting arguments always requires the specification of a sample space.

Example 8.3 A group of fifteen tourists is stranded in a city with four hotels of the same class. Each of the hotels has enough room available to accommodate the fifteen tourists. The group's guide, who has a good working relationship with each of the four hotels, assigns the tourists to the hotels as follows. First, he randomly determines how many are to go to hotel A, then how many of the remaining tourists are to go to hotel B, and then how many are to go to hotel C. All remaining tourists are sent to hotel D. Note that each stage of the assignment the guide draws at random a number between zero and the number of tourists left. What is the probability that all four hotels receive guests from the group?

Solution. Let the outcome (i_A, i_B, i_C, i_D) correspond with the situation in which i_A tourists are sent to hotel A, i_B tourists to hotel B, i_C tourists to hotel C, and i_D tourists to hotel D. The probability

$$\frac{1}{16} \times \frac{1}{16 - i_A} \times \frac{1}{16 - i_A - i_B}$$

is assigned to the outcome (i_A, i_B, i_C, i_D) for $0 \leq i_A, i_B, i_C, i_D \leq 15$ and $i_A + i_B + i_C + i_D = 15$. The probability that all four hotels will receive guests is given by

$$\sum_{i_A=1}^{12} \sum_{i_B=1}^{13-i_A} \sum_{i_C=1}^{14-i_A-i_B} \frac{1}{16} \times \frac{1}{16 - i_A} \times \frac{1}{16 - i_A - i_B} = 0.2856.$$

The next example deals with a famous problem known as the *gambler's ruin problem.*

Example 8.4 John and Pete enter a coin-tossing game and they play until one of them has won all of the money. John starts with a dollars and Pete with b dollars. They flip a fair coin. If the coin lands heads, John gets one dollar from Pete; otherwise, John loses one dollar to Pete. What is the probability that John will win all of the money? Next consider the situation that ten people including John and Pete enter the coin-tossing game. Each player has the same starting capital. The first two players, selected by lot, will play until one wins all the other's money. The winner having doubled his money will next play against another player until one has won all the other's money, and so on. Suppose that John is in the first game and Pete plays the last game against the survivor of the first nine players. Who has the best chance to be the ultimate winner- John or Pete?

Solution. Let $P(a, b)$ denote the probability of John winning all of the money when John starts with a dollars and Pete with b dollars. It will be proved that

$$P(a, b) = \frac{a}{a + b}.$$

Let E be the event that John wins all of the money. A recurrence equation for $P(E) = P(a,\, b)$ is obtained by conditioning on the outcome of the first flip. Let H be the event that the first flip lands heads and let $\overline{H}$ be the event that the first flip lands tails. The event E is the union of the disjoint events EH and $E\overline{H}$ and so $P(E) = P(EH) + P(E\overline{H})$. Applying the formula $P(AB) = P(A \mid B)P(B)$, it next follows that

$$P(E) = P(E \mid H)P(H) + P(E \mid \overline{H})P(\overline{H}).$$

If the first flip lands heads, you get the changed situation that John has $a+1$ dollars and Pete has $b - 1$ dollars. This gives $P(E \mid H) = P(a + 1, b - 1)$. Similarly, $P(E \mid \overline{H}) = P(a - 1, b + 1)$. The coin is fair and so $P(H) = P(\overline{H}) = \frac{1}{2}$. Thus we obtain the recurrence equation

$$P(a, b) = \frac{1}{2}P(a + 1, b - 1) + \frac{1}{2}P(a - 1, b + 1).$$

9
Basic rules for discrete random variables

In performing a chance experiment, one is often not interested in the particular outcome that occurs but in a specific numerical value associated with that outcome. Any function that assigns a real number to each outcome in the sample space of the experiment is called a *random variable*. Intuitively, a random variable can be thought of as a quantity whose value is not fixed. The value of a random variable is determined by the outcome of the experiment and consequently probabilities can be assigned to the possible values of the random variable.

The purpose of this chapter is to familiarize the reader with a number of basic rules for calculating characteristics of random variables such as the expected value and the variance. In addition, we give rules for the expected value and the variance of a sum of random variables, including the square-root rule. The rules for random variables are easiest explained and understood in the context of discrete random variables. These random variables can take on only a finite or countably infinite number of values (the so-called continuous random variables that can take on a continuum of values are treated in the next chapter). To conclude this chapter, we discuss the most important discrete random variables such the binomial, the Poisson and the hypergeometric random variables.

9.1 Random variables

The concept of random variable is always a difficult concept for the beginner. Intuitively, a random variable is a function that takes on its values by chance. A random variable is not a variable in the traditional sense of the word and actually it is a little misleading to call it a variable. The convention is to use capital letters such as X, Y, and Z to denote random variables. Formally, a random variable is defined as a real-valued function on the sample space of a chance experiment. A random variable X assigns a numerical value $X(\omega)$ to each element ω of the sample space. For example, if the random variable X is defined as the smallest of the two numbers rolled in the experiment of rolling a fair die twice, then the random variable X assigns the numerical value $\min(i, j)$ to the outcome (i, j) of the chance experiment. As said before, a random variable X takes on its values by chance. A random variable X gets its value only *after* the underlying chance experiment has been performed. Before the experiment is performed, we can only describe the set of possible values of X. Illustrative examples of random variables are:

- The number of winners in a football pool next week.
- The number of major hurricanes that will hit the United States next year.
- The daily number of claims submitted to an insurance company.
- The amount of rainfall that the city of London will receive next year.
- The lifetime of a newly bought battery.
- The duration of a telephone call.

The first three examples are examples of discrete random variables taking on a discrete number of values and the other three examples describe continuous random variables taking on a continuum of values.

In this chapter we consider only discrete random variables. A random variable X is said to be *discrete* if its set of possible values is finite or countably infinite. The set of possible values of X is called the *range* of X and is denoted by I. The probabilities associated with these possible values are determined by the probability measure P on the sample space of the chance experiment. The *probability mass function* of a discrete random variable X is defined by $P(X = x)$ for $x \in I$, where the notation $P(X = x)$ is shorthand for

$$P(X = x) = P(\{\omega : X(\omega) = x\}).$$

In words, $P(X = x)$ is the probability mass assigned by the probability measure P to the set of all outcomes ω for which $X(\omega) = x$. For example, consider the experiment of rolling a fair die twice and let the random variable X be defined as the smallest of the two numbers rolled. The range of X is the set $I = \{1, 2, \ldots, 6\}$. The random variable X takes on the value 1 if one of the eleven outcomes $(1, 1)$, $(1, 2)$, …, $(1, 6)$, $(2, 1)$, $(3, 1)$, …, $(6, 1)$ occurs and so $P(X = 1) = \frac{11}{36}$. In the same way, $P(X = 2) = \frac{9}{36}$, $P(X = 3) = \frac{7}{36}$, $P(X = 4) = \frac{5}{36}$, $P(X = 5) = \frac{3}{36}$, and $P(X = 6) = \frac{1}{36}$.

Example 9.1 In your pocket you have three dimes (coins of 10 cents) and two quarters (coins of 25 cents). You grab at random two coins from your pocket. What is the probability mass function of the amount you grabbed?

Solution. The sample space of the chance experiment is chosen as $\Omega = \{(D,D),(D,Q),(Q,D),(Q,Q)\}$. The outcome (D,D) occurs if the first coin taken is a dime and the second one is also a dime, the outcome (D,Q) occurs if the first coin taken is a dime and the second one is a quarter, etc. The probability $\frac{3}{5}\times\frac{2}{4}=\frac{3}{10}$ is assigned to the outcome (D,D), the probability $\frac{3}{5}\times\frac{2}{4}=\frac{3}{10}$ to the outcome (D,Q), the probability $\frac{2}{5}\times\frac{3}{4}=\frac{3}{10}$ to the outcome (Q,D), and the probability $\frac{2}{5}\times\frac{1}{4}=\frac{1}{10}$ to the outcome (Q,Q). Let the random variable X denote the total number of cents you have grabbed. The random variable X has 20, 35, and 50 as possible values. The random variable X takes on the value 20 if the outcome (D,D) occurs, the value 35 if either the outcome (D,Q) or (Q,D) occurs, and the value 50 if the outcome (Q,Q) occurs. Thus, the probability mass function of X is given by $P(X=20)=\frac{3}{10}$, $P(X=35)=\frac{3}{10}+\frac{3}{10}=\frac{3}{5}$, and $P(X=50)=\frac{1}{10}$.

Example 9.2 You have a well-shuffled ordinary deck of 52 cards. You remove the cards one at a time until you get an ace. Let the random variable X be the number of cards removed. What is the probability mass function of X?

Solution. The range of the random variable X is the set $\{1,2,\ldots,49\}$. Obviously, $P(X=1)=\frac{4}{52}$. Using Rule 8.1, we find for $i=2,\ldots,49$:

$$P(X=i)=\frac{48}{52}\times\cdots\times\frac{48-(i-2)}{52-(i-2)}\times\frac{4}{52-(i-1)}.$$

The latter example gives rise to the following important observation. Often an explicit listing of the sample space is not necessary to assign a probability distribution to a random variable. Usually the probability distribution of a random variable is modeled without worrying about the assignment of probability to an underlying sample space. In most problems, you will perform probability calculations without explicitly specifying a sample space; an assignment of probabilities to properly chosen events usually suffices.

Problem 9.1 A fair die is tossed two times. Let the random variable X be the largest of the two outcomes. What is the probability mass function of X?

Problem 9.2 Imagine that people enter a room one by one and announce their birthdays. Let the random variable X be the number of people required to have a matching birthday. What is the probability mass function of X?

Problem 9.3 A bag contains three coins. One coin is two-headed and the other two are normal. A coin is chosen at random from the bag and is tossed two times? Let the random variable X denote the number of heads obtained. What is the probability mass function of X?

Problem 9.4 A fair die is rolled until each of the six possible outcomes has occurred. Let the random variable X be the number of rolls required. What is the probability mass function of X? *Hint*: use the answer to Problem 7.49 and the relation $P(X = i) = P(X > i - 1) - P(X > i)$.

9.2 Expected value

The most important characteristic of a random variable is its *expected value.* In Chapter 2 we informally introduced the concept of expected value. The expected value of a discrete random variable is a weighted mean of the values the random variable can take on, the weights being furnished by the probability mass function of the random variable. The nomenclature of expected value may be misleading. The expected value is in general not a typical value that the random variable can take on.

Definition 9.1 *The expected value of the discrete random variable X having I as its set of possible values is defined by*

$$E(X) = \sum_{x \in I} x\, P(X = x).$$

To illustrate this definition, consider again the experiment of rolling a fair die twice and let the random variable X denote the smallest of the two numbers rolled. Then,

$$E(X) = 1 \times \frac{11}{36} + 2 \times \frac{9}{36} + 3 \times \frac{7}{36} + 4 \times \frac{5}{36} + 5 \times \frac{3}{36} + 6 \times \frac{1}{36} = 2.528.$$

Before we give some other examples, note the following remarks. Definition 9.1 is only meaningful if the sum is well-defined. The sum is always well-defined if the range I is finite. However, the sum over countably infinite many terms is not always well-defined when both positive and negative terms are involved. For example, the infinite series $1 - 1 + 1 - 1 + \ldots$ has the sum 0 when you sum the terms according to $(1 - 1) + (1 - 1) + \ldots$, whereas you get the sum 1 when you sum the terms according to $1 + (-1 + 1) + (-1 + 1) + (-1 + 1) + \cdots$. Such abnormalities cannot happen when all terms in the infinite summation are nonnegative. The sum of infinitely many *nonnegative* terms is always well-defined, with ∞ as a possible value for the sum. For a sequence $a_1, a_2, \ldots$ consisting of both positive and negative terms, a basic result from the theory of series states that the infinite series $\sum_{k=1}^{\infty} a_k$ is always well-defined with a finite sum if the series is absolutely convergent, where absolute convergence means that $\sum_{k=1}^{\infty} |a_k| < \infty$. In case the series $\sum_{k=1}^{\infty} a_k$ is absolutely convergent, the sum is uniquely determined and does not depend on the order in which the individual terms are added. For a discrete random variable X with range I, it is said that the expected value $E(X)$ *exists* if X is nonnegative or if $\sum_{x \in I} |x| \, P(X = x) < \infty$. An example of a random variable X for which $E(X)$ does not exist is the random variable X with probability mass function $P(X = k) = \frac{3}{\pi^2 k^2}$ for $k = \pm 1, \pm 2, \ldots$ (by the celebrated result $\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}$, the probabilities sum to 1). The reason that $E(X)$ does not exist is the well-known fact from calculus that $\sum_{k=1}^{\infty} \frac{1}{k} = \infty$.

Example 9.1 (continued) What is the expected value of the amount you grabbed from your pocket?

Solution. Since the probability mass function of the number of cents you grabbed from your pocket is given by $P(X = 20) = \frac{3}{10}$, $P(X = 35) = \frac{3}{5}$, and $P(X = 50) = \frac{1}{10}$, the expected value of the amount you grabbed is equal to

$$E(X) = 20 \times \frac{3}{10} + 35 \times \frac{3}{5} + 50 \times \frac{1}{10} = 32 \text{ cents.}$$

Example 9.3 Joe and his friend make a guess every week whether the Dow Jones index will have risen at the end of the week or not. Both put \$10 in the pot. Joe observes that his friend is just guessing and is making his choice by the toss of a fair coin. Joe asks his friend if he could contribute \$20 to the pot and submit his guess together with that of his brother. The friend agrees. In each week, however, Joe's brother submits a prediction opposite to that of Joe. The person having a correct prediction wins the entire pot. If more than one person has a correct prediction, the pot is split evenly. How favorable is the game to Joe and his brother?

Solution. Let the random variable X denote the payoff to Joe and his brother in any given week. Either Joe or his brother will have a correct prediction. If Joe's friend is wrong he wins nothing, and if he is correct he shares the \$30 pot with either Joe or his brother. Thus, X takes on the values 30 and 15 with equal chances. This gives $E(X) = \frac{1}{2} \times 30 + \frac{1}{2} \times 15 = 22.5$ dollars. Joe and his brother have an expected profit of \$2.5 every week.

Example 9.4 Three friends go to the cinema together every week. Each week, in order to decide which friend will pay for the other two, they all toss a fair coin into the air simultaneously. They continue to toss coins until one of the three gets a different outcome from the other two. What is the expected value of the number of trials required?

Solution. Let the random variable X denote the number of trials until one of the three friends gets a different outcome from the other two. The probability that any given trial does not lead to three equal outcomes is $p = 1 - \frac{1}{8} - \frac{1}{8} = \frac{3}{4}$. Thus,

$$P(X = j) = (1-p)^{j-1}p \qquad \text{for } j = 1, 2, \ldots$$

with $p = \frac{3}{4}$. The expected value of X is given by

$$E(X) = \sum_{j=1}^{\infty} j(1-p)^{j-1}p = p\sum_{j=1}^{\infty} j(1-p)^{j-1} = \frac{p}{[1-(1-p)]^2} = \frac{1}{p},$$

using the fact that $\sum_{j=1}^{\infty} jx^{j-1} = 1/(1-x)^2$ for all $0 < x < 1$ (see the Appendix). Hence the expected value of the number of trials required is $\frac{4}{3}$.

The expected value of a random variable X is also known as *expectation*, or *mean*, or *first moment* and *average value* of X. The term “average value” can be explained as follows. Consider Example 9.3. If the game is played many times in succession, then the average profit per game will approach $E(X)$ as the number of repetitions of the game increases without bound. This result is known as the law of large numbers. This law will be made more precise in Section 14.4 (see also the informal discussion in Section 2.3).

Problem 9.5 You are playing a game in which four fair dice are rolled. A \$1 stake is required. The payoff is \$100 if all four dice show the same number and \$15 if two dice show the same even number and the other two dice show the same odd number. Is this a favorable game? Answer this same question for the following game in which the stake is \$2. A fair coin is tossed no more than five times. The game ends if the coin comes up tails or five straight heads appear, whichever happens first. You get a payoff of \$1 each time heads appears plus a bonus of \$25 if five heads appear in a row.

Problem 9.6 Calculate the expected value of the greater of two numbers when two different numbers are picked at random from the numbers $1, \ldots, n$. What is the expected value of the absolute difference between the two numbers?

Problem 9.7 You throw darts at a circular target on which two concentric circles of radius 1 cm and 3 cm are drawn. The target itself has a radius of 5 cm. You receive 15 points for hitting the target inside the smaller circle, 8 points for hitting the middle annular region, and 5 points for hitting the outer annular region. The probability of hitting the target at all is 0.75. If the dart hits the target, the hitting point is a completely random point on the target. Let the random variable X denote the number of points scored on a single throw of the dart. What is the expected value of X?

Problem 9.8 Three players, A, B, and C each put ten dollars into a pot with a list on which they have predicted the outcome of three successive tosses of a fair coin. The fair coin is then tossed three times. The player having most correctly predicted the three outcomes gets the contents of the pot. The contents are to be divided if multiple players guess the same number of correct outcomes. Suppose that players A and B collaborate, unbeknownst to player C. The collaboration consists of the two players agreeing that the list of player B will always be a mirror image of player A's list (should player A predict an outcome of HTT, for example, then player B would predict TTH). What the expected value of the amount that player C will receive?

Problem 9.9 You spin a game board spinner with 1,000 equal sections numbered as $1, 2, \ldots, 1{,}000$. After your first spin, you have to decide whether to spin the spinner for a second time. Your payoff is the total score of your spins as long as this score does not exceed 1,000; otherwise, your payoff is zero. What strategy maximizes the expected value of your payoff?

Problem 9.10 In a charity lottery, one thousand tickets numbered as $000, 001, \ldots, 999$ are sold. Each contestant buys only one ticket. The prize winners of the lottery are determined by drawing one number at random from the numbers $000, 001, \ldots, 999$. You are a prize winner when the number on your ticket is the same as the number drawn or is a random permutation of the number drawn. What is the probability mass function of the number of prize winners and what is the expected value of the number of prize winners? What is the probability that a randomly picked contestant will be a prize winner?

Problem 9.11 A stick is broken at random into two pieces. You bet on the ratio of the length of the longer piece to the length of the smaller piece. You receive \$$k$ if the ratio is between k and $k+1$ for some k with $1 \leq k \leq m-1$, while you receive \$$m$ if the ratio is larger than m. Here m is a given positive integer. What should be your stake to make this a fair bet? Verify that your stake should be $\$2[1+\frac{1}{2}+\cdots+\frac{1}{m+1}-1]$ (this amount is approximately equal to $\$2[\ln(m+1)+\gamma-1+\frac{1}{2(m+1)}]$ for m large, where $\gamma = 0.57722\ldots$ is Euler's constant).

10
Continuous random variables

In many practical applications of probability, physical situations are better described by random variables that can take on a *continuum* of possible values rather than a *discrete* number of values. Examples are the decay time of a radioactive particle, the time until the occurrence of the next earthquake in a certain region, the lifetime of a battery, the annual rainfall in London, and so on. These examples make clear what the fundamental difference is between discrete random variables taking on a discrete number of values and continuous random variables taking on a continuum of values. Whereas a discrete random variable associates *positive* probabilities to its individual values, any individual value has probability *zero* for a continuous random variable. It is only meaningful to speak of the probability of a continuous random variable taking on a value in some interval. Taking the lifetime of a battery as an example, it will be intuitively clear that the probability of this lifetime taking on a specific value becomes zero when a finer and finer unit of time is used. If you can measure the heights of people with infinite precision, the height of a randomly chosen person is a continuous random variable. In reality, heights cannot be measured with infinite precision, but the mathematical analysis of the distribution of heights of people is greatly simplified when using a mathematical model in which the height of a randomly chosen person is modeled as a continuous random variable. Integral calculus is required to formulate the continuous analog of a probability mass function of a discrete random variable.

The purpose of this chapter is to familiarize the reader with the concept of probability density function of a continuous random variable. This is always a difficult concept for the beginning student. However, integral calculus enables us to give an enlightening interpretation of a probability density. Also, this chapter summarizes the most important probability densities used in practice. In particular, the exponential density and the normal density are treated in depth. Many practical phenomena can be modeled by these distributions which are of fundamental importance. Also, attention is given to the central limit theorem being the most important theorem of probability theory. Finally, the inverse-transformation method for simulating a random observation from a continuous random variable and the important concept of failure rate function will be discussed.

10.1 Concept of probability density

The most simple example of a continuous random variable is the random choice of a number from the interval $(0, 1)$. The probability that the randomly chosen number will take on a prespecified value is zero. It makes only sense to speak of the probability of the randomly chosen number falling in a given subinterval of $(0, 1)$. This probability is equal to the length of that subinterval. For example, if a dart is thrown at random to the interval $(0, 1)$, the probability of the dart hitting exactly the point 0.25 is zero, but the probability of the dart landing somewhere in the interval between 0.2 and 0.3 is 0.1 (assuming that the dart has an infinitely thin point). No matter how small Δx is, any subinterval of the length Δx has probability Δx of containing the point at which the dart will land. You might say that the probability mass associated with the landing point of the dart is smeared out over the interval $(0, 1)$ in such a way that the density is the same everywhere. For the random variable X denoting the point at which the dart will land, we have that the cumulative probability $P(X \leq a) = a$ for $0 \leq a \leq 1$ can be represented as $P(X \leq a) = \int_0^a f(x)dx$ with the density $f(x)$ identically equal to 1 on the interval $(0, 1)$. Before defining the concept of probability density within a general framework, it is instructive to consider the following example.

Example 10.1 A stick of unit length is broken at random into two pieces. What is the probability that the ratio of the length of the shorter piece to that of the longer piece is smaller than or equal to a for any $0 < a < 1$?

Solution. The sample space of the chance experiment is the interval $(0, 1)$, where the outcome $\omega = u$ means that the point at which the stick is broken is a distance u from the beginning of the stick. Let the random variable X denote the ratio of length of the shorter piece to that of the longer piece of the broken stick. Denote by $F(a)$ the probability that the random variable X takes on a value smaller than or equal to a. Fix $0 < a < 1$. The probability that the ratio of the length of the shorter piece to that of the longer piece is smaller than or equal to a is nothing else than the probability that a random number from the interval (0,1) falls either in $(\frac{1}{1+a}, 1)$ or in $(0, 1 - \frac{1}{1+a})$. The latter probability is equal to $2(1 - \frac{1}{1+a}) = \frac{2a}{1+a}$. Thus,

$$F(a) = \frac{2a}{1+a} \qquad \text{for } 0 < a < 1.$$

Obviously, $F(a) = 0$ for $a \leq 0$ and $F(a) = 1$ for $a \geq 1$. Denoting by $f(a) = \frac{2}{(1+a)^2}$ the derivative of $F(a)$ for $0 < a < 1$ and letting $f(a) = 0$ outside the interval (0,1), it follows that

$$F(a) = \int_{-\infty}^{a} f(x)dx \qquad \text{for all } a.$$

In this specific example, we have a continuous analog of the cumulative probability $F(a)$ in the discrete case: if X is a discrete random variable having possible values $a_1, a_2, \ldots$ with associated probabilities $p_1, p_2, \ldots$, then the probability that X takes on a value smaller than or equal to a is represented by

$$F(a) = \sum_{i:\, a_i \leq a} p_i \qquad \text{for all } a.$$

We now come to the definition of a continuous random variable. Let X be a random variable that is defined on a sample space with probability measure P. It is assumed that the set of possible values of X is uncountable and is a finite or infinite interval on the real line.

Definition 10.1 *The random variable X is said to be (absolutely) continuously distributed if a function $f(x)$ exists such that*

$$P\left(X \leq a\right) = \int_{-\infty}^{a} f(x)\, dx \qquad \textit{for each real number } a,$$

where the function $f(x)$ satisfies

$$f(x) \geq 0 \qquad \textit{for all } x \quad \textit{and} \quad \int_{-\infty}^{\infty} f(x)\, dx = 1.$$

The notation $P(X \leq a)$ stands for the probability that is assigned by the probability measure P to the set of all outcomes ω for which $X\left(\omega\right) \leq a$. The function $P(X \leq x)$ is called the *(cumulative) probability distribution function* of the random variable X, and the function $f(x)$ is called the *probability density function* of X. Unlike the probability distribution function of a discrete random variable, the probability distribution function of a continuous random variable has no jumps and is continuous everywhere.

Beginning students often misinterpret the nonnegative number $f(a)$ as a probability, namely as the probability $P(X = a)$. This interpretation is wrong. Nevertheless, it is possible to give an intuitive interpretation of the nonnegative number $f(a)$ in terms of probabilities. Before doing this, we present two examples of a continuous random variable with a probability density function.

Example 10.2 Suppose that the lifetime X of a battery has the cumulative probability distribution function

$$P(X \leq x) = \begin{cases} 0 & \text{for } x < 0, \\ \frac{1}{4}x^2 & \text{for } 0 \leq x \leq 2, \\ 1 & \text{for } x > 2. \end{cases}$$

The probability distribution function $P(X \leq x)$ is continuous and is differentiable at each point x except for the two points $x = 0$ and $x = 2$. Also, the derivative is integrable. We can now conclude from the fundamental theorem of integral calculus that the random variable X has a probability density function. This probability density function is obtained by differentiation of the probability distribution function and is given by

$$f(x) = \begin{cases} \frac{1}{2}x & \text{for } 0 < x < 2, \\ 0 & \text{otherwise.} \end{cases}$$

In each of the finite number of points x at which $P(X \leq x)$ has no derivative, it does not matter what value we give $f(x)$. These values do not affect $\int_{-\infty}^{a} f(x)\,dx$. Usually, we give $f(x)$ the value 0 at any of these exceptional points.

Example 10.3 A continuous random variable X has a probability density $f(x)$ of the form $f(x) = ax + b$ for $0 < x < 1$ and $f(x) = 0$ otherwise. What conditions on the constants a and b must be satisfied? What is the cumulative probability distribution function of X?

Solution. The requirements for $f(x)$ are $ax + b \geq 0$ for $0 < x < 1$ and $\int_0^1 (ax + b)\,dx = 1$. The first requirement gives $a + b \geq 0$ and $b \geq 0$. The second requirement gives $\frac{1}{2}a+b=1$. The cumulative probability distribution function of X is equal to

$$F(x) = \int_0^x (av + b)\,dv = \frac{1}{2}ax^2 + bx \quad \text{for } 0 \leq x \leq 1.$$

Further, $F(x) = 0$ for $x < 0$ and $F(x) = 1$ for $x > 1$.

Problem 10.1 The lifetime of an appliance is a continuous random variable X and has a probability density $f(x)$ of the form $f(x) = c(1 + x)^{-3}$ for $x > 0$ and $f(x) = 0$ otherwise. What is the value of the constant c? Find $P(X \leq 0.5)$, $P(0.5 < X \leq 1.5)$ and $P(0.5 < X \leq 1.5 \mid X > 0.5)$.

Problem 10.2 Let the random variable X be the portion of a flood insurance claim for flooding damage to the house. The probability density of X has the form $f(x) = c(3x^2 - 8x - 5)$ for $0 < x < 1$. What is the value of the constant c? What is the cumulative probability distribution function of X?

Problem 10.3 Sizes of insurance claims can be modeled by a continuous random variable with probability density $f(x) = \frac{1}{50}(10 - x)$ for $0 < x < 10$ and $f(x) = 0$ otherwise. What is the probability that the size of a particular claim is larger than 5 given that the size exceeds 2?

Problem 10.4 The lengths of phone calls (in minutes) made by a travel agent can be modeled as a continuous random variable with probability density $f(x) = 0.25e^{-0.25x}$ for $x > 0$. What is the probability that a particular phone call will take more than 7 minutes?

10.1.1 Interpretation of the probability density

The use of the word "density" originated with the analogy to the distribution of matter in space. In physics, any finite volume, no matter how small, has a positive mass, but there is no mass at a single point. A similar description applies to continuous random variables. To make this more precise, we first express $P(a < X \leq b)$ in terms of the density $f(x)$ for any constants a and b with $a < b$. Noting that the event $\{X \leq b\}$ is the union of the two disjoint events $\{a < X \leq b\}$ and $\{X \leq a\}$, it follows that $P(X \leq b) = P(a < X \leq b) + P(X \leq a)$. Hence,

$$\begin{aligned} P(a < X \leq b) &= P(X \leq b) - P(X \leq a) \\ &= \int_{-\infty}^{b} f(x)\,dx - \int_{-\infty}^{a} f(x)\,dx \qquad \text{for } a < b \end{aligned}$$

and so

$$P(a < X \leq b) = \int_{a}^{b} f(x)\,dx \qquad \text{for } a < b.$$

In other words, the area under the graph of $f(x)$ between the points a and b gives the probability $P(a < X \leq b)$. Next, we find that

$$\begin{aligned} P(X = a) &= \lim_{n\to\infty} P(a - \frac{1}{n} < X \leq a) \\ &= \lim_{n\to\infty} \int_{a-\frac{1}{n}}^{a} f(x)\,dx = \int_{a}^{a} f(x)dx, \end{aligned}$$

using the continuity property of the probability measure P stating that $\lim_{n\to\infty} P(A_n) = P(\lim_{n\to\infty} A_n)$ for any nonincreasing sequence of events A_n (see Section 7.1.3). Hence, we arrive at the conclusion

$$P(X = a) = 0 \qquad \text{for each real number } a.$$

This formally proves that, for a continuous random variable X, it makes no sense to speak of the probability that the random variable X will take on a *prespecified* value. This probability is always zero. It only makes sense to speak of the probability that the continuous random variable X will take on a value in some interval. Incidentally, since $P(X = c) = 0$ for any number c, the probability that X takes on a value in an interval with endpoints a and b is not influenced by whether or not the endpoints are included. In other words, for any two real numbers a and b with $a < b$, we have

$$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b).$$

The fact that the area under the graph of $f(x)$ can be interpreted as a probability leads to an intuitive interpretation of $f(a)$. Let a be a given continuity point of $f(x)$. Consider now a small interval of length Δa around the point a, say $[a - \frac{1}{2}\Delta a, a + \frac{1}{2}\Delta a]$. Since

$$P(a - \frac{1}{2}\Delta a \leq X \leq a + \frac{1}{2}\Delta a) = \int_{a-\frac{1}{2}\Delta a}^{a+\frac{1}{2}\Delta a} f(x)\, dx$$

and

$$\int_{a-\frac{1}{2}\Delta a}^{a+\frac{1}{2}\Delta a} f(x)\, dx \approx f(a)\Delta a \qquad \text{for } \Delta a \text{ small},$$

we obtain that

$$P(a - \frac{1}{2}\Delta a \leq X \leq a + \frac{1}{2}\Delta a) \approx f(a)\Delta a \qquad \text{for } \Delta a \text{ small}.$$

In other words, the probability of random variable X taking on a value in a *small* interval around point a is approximately equal to $f(a)\Delta a$ when Δa is the length of the interval. You see that the number $f(a)$ itself is *not* a probability, but it is a relative measure for the likelihood that random variable X will take on a value in the immediate neighborhood of point a. Stated differently, the probability density function $f(x)$ expresses how densely the probability mass of random variable X is smeared out in the neighborhood of point x. Hence, the name of density function. The probability density function provides the most useful description of a continuous random variable. The graph of the density function provides a good picture of the likelihood of the possible values of the random variable.

10.1.2 Verification of a probability density

In general, how can we verify whether a random variable X has a probability density? In concrete situations, we first determine the cumulative distribution function $F(a) = P(X \leq a)$ and next we verify whether $F(a)$ can be written in the form of $F(a) = \int_{-\infty}^{a} f(x)\,dx$. A sufficient condition is that $F(x)$ is continuous at every point x and is differentiable except for a finite number of points x. The following two examples are given in illustration of this point.

Example 10.4 Let the random variable be given by $X = -\frac{1}{\lambda}\ln(U)$, where U is a random number between 0 and 1 and λ is a given positive number. What is the probability density function of X?

Solution. To answer the question, note first that X is a positive random variable. For any $x > 0$,

$$\begin{aligned} P(X \leq x) &= P(-\frac{1}{\lambda}\ln(U) \leq x) = P(\ln(U) \geq -\lambda x) \\ &= P(U \geq e^{-\lambda x}) = 1 - P(U \leq e^{-\lambda x}), \end{aligned}$$

where the last equality uses the fact that $P(U < u) = P(U \leq u)$ for the continuous random variable U. Since $P(U \leq u) = u$ for $0 < u < 1$, it follows that

$$P(X \leq x) = 1 - e^{-\lambda x}, \qquad x > 0.$$

Obviously, $P(X \leq x) = 0$ for $x \leq 0$. Noting that the expression for $P(X \leq x)$ is continuous at every point x and is differentiable except at $x = 0$, we obtain by differentiation that X has a probability density function $f(x)$ with $f(x) = \lambda e^{-\lambda x}$ for $x > 0$ and $f(x) = 0$ for $x \leq 0$. This density function is the so-called exponential density function. In many situations, it describes adequately the density function of the waiting time until a *rare* event occurs.

Example 10.5 A point is picked at random in the inside of a circular disk with radius r. Let the random variable X denote the distance from the center of the disk to this point. Does the random variable X have a probability density function and, if so, what is its form?

Solution. To answer the question, we first define a sample space with an appropriate probability measure P for the chance experiment. The sample space is taken as the set of all points (x, y) in the two-dimensional plane with $x^2 + y^2 \leq r^2$. Since the point inside the circular disk is chosen at random, we assign to each well-defined subset A of the sample space the probability

$$P(A) = \frac{\text{area of region } A}{\pi r^2}.$$

The cumulative probability distribution function $P(X \leq x)$ is easily calculated. The event $X \leq a$ occurs if and only if the randomly picked point falls in the disk of radius a with area πa^2. Therefore

$$P(X \leq a) = \frac{\pi a^2}{\pi r^2} = \frac{a^2}{r^2} \qquad \text{for } 0 \leq a \leq r.$$

Obviously, $P(X \leq a) = 0$ for $a < 0$ and $P(X \leq a) = 1$ for $a > r$. Since the expression for $P(X \leq x)$ is continuous at every point x and is differentiable except at the point $x = a$, it follows that X has a probability density function which is given by

$$f(x) = \begin{cases} \frac{2x}{r^2} & \text{for } 0 < x < r, \\ 0 & \text{otherwise.} \end{cases}$$

All of the foregoing examples follow the same procedure in order to find the probability density function of a random variable X. The cumulative probability distribution function $P(X \leq x)$ is determined first and this distribution function is then differentiated to obtain the probability density.

As pointed out before, the value of the probability density at any point a is a relative measure for the likelihood that the random variable will take on a value in the immediate neighborhood of the point a. To illustrate this, let us put the following question with regard to the last example. A point will be randomly chosen within the unit disk and you are asked to bet on the value of the distance from the chosen point to the center of the disk. You win the bet if your guess is no more than 5% off from the observed distance. Should your guess be a number close to zero or close to 1? The probability density function of the distance is $f(x) = 2x$ for $0 < x < 1$ and so your guess should be close to 1. The best value of your guess follows by maximizing $\int_{c-0.05c}^{c+0.05c} f(x)\,dx$ with respect to c. This gives $c = \frac{20}{21}$ with a win probability of 0.1814.

Problem 10.5 Let X be a positive random variable with probability density function $f(x)$. Define the random variable Y by $Y = X^2$. What is the probability density function of Y? Also, find the density function of the random variable $W = V^2$ if V is a number chosen at random from the interval $(-a, a)$ with $a > 0$.

Problem 10.6 A point Q is chosen at random inside the unit square. What is the density function of the sum of the coordinates of the point Q? What is the density function of the product of the coordinates of the point Q? Use geometry to find these densities.

Problem 10.7 The number X is chosen at random between 0 and 1. Determine the probability density function of each of the random variables $V = X/(1-X)$ and $W = X(1-X)$.

Problem 10.8 A stick of unit length is broken at random into two pieces. Let the random variable X represent the length of the shorter piece. What is the probability density of X? Also, use the probability distribution function of X to give an alternative derivation of the probability density of the random variable $X/(1-X)$ from Example 10.1.

Problem 10.9 A point is randomly chosen inside the unit square. The random variables V and W be defined as the largest and the smallest of the two coordinates of the point. What are the probability density functions of the random variables V and W?

Problem 10.10 Suppose you decide to take a ride on the ferris wheel at an amusement park. The ferris wheel has a diameter of 30 meters. After several turns, the ferris wheel suddenly stops due to a power outage. What random variable determines your height above the ground when the ferris wheel stops? What is the probability that this height is not more than 22.5 meters? And the probability of no more than 7.5 meters? What is the probability density function of the random variable governing the height above the ground? It is assumed that the bottom of the ferris wheel is level with the ground.

10.2 Expected value of a continuous random variable

The expected value of a continuous random variable X with probability density function $f(x)$ is defined by

$$E(X) = \int_{-\infty}^{\infty} x f(x)\, dx,$$

provided that the integral $\int_{-\infty}^{\infty} |x| f(x)\, dx$ is finite (the latter integral is always well-defined by the nonnegativity of the integrand). It is then said that $E(X)$ exists. In the case that X is a nonnegative random variable, the integral $\int_0^{\infty} x f(x)\, dx$ is always well-defined when allowing ∞ as possible value. The definition of expected value in the continuous case parallels the definition $E(X) = \sum x_i p(x_i)$ for a discrete random variable X with $x_1, x_2, \ldots$ as possible values and $p(x_i) = P(X = x_i)$. For dx small, the quantity $f(x)\, dx$ in a discrete approximation of the continuous case corresponds with $p(x)$ in the discrete case. The summation becomes an integral when dx approaches zero. Results for discrete random variables are typically expressed as sums.

11
Jointly distributed random variables

In experiments, one is often interested not only in individual random variables, but also in relationships between two or more random variables. For example, if the experiment is the testing of a new medicine, the researcher might be interested in cholesterol level, blood pressure, and glucose level of a test person. Similarly, a political scientist investigating the behavior of voters might be interested in the income and level of education of a voter. There are many more examples in the physical sciences, medical sciences, and social sciences. In applications, one often wishes to make inferences about one random variable on the basis of observations of other random variables.

The purpose of this chapter is to familiarize the student with the notations and the techniques relating to experiments whose outcomes are described by two or more real numbers. The discussion is restricted to the case of pairs of random variables. The chapter treats joint and marginal densities, along with covariance and correlation. Also, the transformation rule for jointly distributed random variables and regression to the mean are discussed.

11.1 Joint probability mass function

If X and Y are two discrete random variables defined on a same sample space with probability measure P, the mass function $p(x, y)$ defined by

$$p(x, y) = P(X = x, Y = y)$$

is called the *joint probability mass function* of X and Y. The quantity $P(X = x, Y = y)$ is the probability assigned by P to the intersection of the two sets $A = \{\omega : X(\omega) = x\}$ and $B = \{\omega : Y(\omega) = y\}$, with ω representing an element of the sample space. Define the marginal probability

Table 11.1. *The joint probability mass function $p(x, y)$.*

x \ y	2	3	4	5	6	7	8	9	10	11	12	$p_X(x)$
1	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	0	0	0	0	0	$\frac{11}{36}$
2	0	0	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	0	0	0	0	$\frac{9}{36}$
3	0	0	0	0	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	0	0	0	$\frac{7}{36}$
4	0	0	0	0	0	0	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	0	0	$\frac{5}{36}$
5	0	0	0	0	0	0	0	0	$\frac{1}{36}$	$\frac{2}{36}$	0	$\frac{3}{36}$
6	0	0	0	0	0	0	0	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$
$p_Y(y)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	sum = 1

mass functions of the random variables X and Y by

$$p_X(x) = P(X = x) \quad \text{and} \quad p_Y(y) = P(Y = y).$$

The marginal probability mass functions can be obtained from the joint probability mass function by

$$p_X(x) = \sum_y P(X = x,\, Y = y), \quad p_Y(y) = \sum_x P(X = x,\, Y = y).$$

These relations follow from the result that $P(A) = \sum_{i=1}^{n} P(A_i)$ if the event A is the union of mutually exclusive events $A_1, A_2, \ldots, A_n$.

Example 11.1 Two fair dice are rolled. Let the random variable X represent the smallest of the outcomes of the two rolls, and let Y represent the sum of the outcomes of the two rolls. What is the joint probability mass function of X and Y?

Solution. The random variables X and Y are defined on a same sample space. The sample space is the set of all 36 pairs (i, j) for $i, j = 1, \ldots, 6$, where i and j are the outcomes of the first and second dice. A probability of $\frac{1}{36}$ is assigned to each element of the sample space. In Table 11.1, we give the joint probability mass function $p(x,\, y) = P(X = x,\, Y = y)$. For example, $P(X = 2,\, Y = 5)$ is the probability of the intersection of the sets $A = \{(2,2), (2,3), (2,4), (2,5), (2,6), (3,2), (4,2), (5,2), (6,2)\}$ and $B = \{(1,4), (4,1), (2,3), (3,2)\}$. The set $\{(2,3), (3,2)\}$ is the intersection of these two sets and has probability $\frac{2}{36}$.

Problem 11.1 You roll a pair of dice. What is the joint probability mass function of the low and high points rolled?

Problem 11.2 Let X denote the number of hearts and Y the number of diamonds in a bridge hand. What is the joint probability mass function of X and Y?

Problem 11.3 You choose three different numbers at random from the numbers $1, 2, \ldots, 10$. Let X be the smallest of these three numbers and Y the largest. What is the joint probability mass function of X and Y? What are the marginal distributions of X and Y and what is the probability mass function of $Y - X$?

Problem 11.4 You repeatedly draw a number at random from the numbers $1, 2, \ldots, 10$. Let X be the number of drawings until the number 1 appears and Y the number of drawings until the number 10 appears. What is the joint probability mass function of X and Y? What are the probability mass functions of $\min(X, Y)$ and $\max(X, Y)$?

Problem 11.5 You repeatedly toss two fair coins until both coins show heads. Let X and Y denote the number of heads resulting from the tosses of the first and the second coin respectively. What is the joint probability mass function of X and Y and what are the marginal distributions of X and Y? What is $P(X = Y)$? *Hint*: evaluate $P(X = i, Y = j, N = n)$, where N is the number of tosses until both coins show heads. Use the identity $\sum_{k=0}^{\infty} \binom{m+k}{m} x^k = 1/(1-x)^{m+1}$ for $|x| < 1$.

11.2 Joint probability density function

The following example provides a good starting point for a discussion of joint probability densities.

Example 11.2 A point is picked at random inside a circular disc with radius r. Let the random variable X denote the length of the line segment between the center of the disc and the randomly picked point, and let the random variable Y denote the angle between this line segment and the horizontal axis (Y is measured in radians and so $0 \leq Y < 2\pi$). What is the joint distribution of X and Y?

Solution. The two continuous random variables X and Y are defined on a common sample space. The sample space consists of all points (v, w) in the two-dimensional plane with $v^2 + w^2 \leq r^2$, where the point $(0, 0)$ represents the center of the disc. The probability $P(A)$ assigned to each well-defined subset A of the sample space is taken as the area of region A divided by πr^2. The probability of the event of X taking on a value less than or equal to a and Y taking on a value less than or equal to b is denoted by $P(X \leq a, Y \leq b)$. This event occurs only if the randomly picked point falls inside the disc segment with radius a and angle b. The area of this disc segment is $\frac{b}{2\pi}\pi a^2$. Dividing this by πr^2 gives

$$P\left(X \leq a, Y \leq b\right) = \frac{b}{2\pi}\frac{a^2}{r^2} \qquad \text{for } 0 \leq a \leq r \text{ and } 0 \leq b \leq 2\pi.$$

We are now in a position to introduce the concept of joint density. Let X and Y be two random variables that are defined on a same sample space with probability measure P. The *joint cumulative probability distribution function* of X and Y is defined by $P(X \leq x, Y \leq y)$ for all x, y, where $P(X \leq x, Y \leq y)$ is a shorthand for $P(\{\omega : X(\omega) \leq x \text{ and } Y(\omega) \leq y\})$ and the symbol ω represents an element of the sample space.

Definition 11.1 *The continuous random variables X and Y are said to have a joint probability density function $f(x, y)$ if the joint cumulative probability distribution function $P(X \leq a, Y \leq b)$ allows for the representation*

$$P(X \leq a, Y \leq b) = \int_{x=-\infty}^{a} \int_{y=-\infty}^{b} f(x, y)\, dx\, dy, \qquad -\infty < a, b < \infty,$$

where the function $f(x, y)$ satisfies

$$f(x, y) \geq 0 \qquad \textit{for all } x, y \quad \textit{and} \quad \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dxdy = 1.$$

Just as in the one-dimensional case, $f(a, b)$ allows for the interpretation:

$$\begin{aligned} & f(a, b)\, \Delta a\, \Delta b \\ \approx \quad & P(a - \frac{1}{2}\Delta a \leq X \leq a + \frac{1}{2}\Delta a, b - \frac{1}{2}\Delta b \leq Y \leq b + \frac{1}{2}\Delta b) \end{aligned}$$

for small positive values of Δa and Δb provided that $f(x, y)$ is continuous in the point (a, b). In other words, the probability that the random point (X, Y) falls into a small rectangle with sides of lengths $\Delta a, \Delta b$ around the point (a, b) is approximately given by $f(a, b)\, \Delta a\, \Delta b$.

To obtain the joint probability density function $f(x, y)$ of the random variables X and Y in Example 11.2, we take the partial derivatives of $P(X \leq x, Y \leq y)$ with respect to x and y. It then follows from

$$f(x, y) = \frac{\partial^2}{\partial x \partial y} P(X \leq x, Y \leq y)$$

12
Multivariate normal distribution

Do the one-dimensional normal distribution and the one-dimensional central limit theorem allow for a generalization to dimension two or higher? The answer is yes. Just as the one-dimensional normal density is completely determined by its expected value and variance, the bivariate normal density is completely specified by the expected values and the variances of its marginal densities and by its correlation coefficient. The bivariate normal distribution appears in many applied probability problems. This probability distribution can be extended to the multivariate normal distribution in higher dimensions. The multivariate normal distribution arises when you take the sum of a large number of independent random vectors. To get this distribution, all you have to do is to compute a vector of expected values and a matrix of covariances. The multidimensional central limit theorem explains why so many natural phenomena have the multivariate normal distribution. A nice feature of the multivariate normal distribution is its mathematical tractability. The fact that any linear combination of multivariate normal random variables has a univariate normal distribution makes the multivariate normal distribution very convenient for financial portfolio analysis, among others.

The purpose of this chapter is to give a first introduction to the multivariate normal distribution and the multidimensional central limit theorem. Several practical applications will be discussed, including the drunkard's walk in higher dimensions and the chi-square test.

12.1 Bivariate normal distribution

A random vector (X, Y) is said to have a *standard bivariate normal distribution* with parameter ρ if it has a joint probability density function of the form

$$f(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}}\, e^{-\frac{1}{2}\left(x^2-2\rho xy+y^2\right)/\left(1-\rho^2\right)}, \qquad -\infty < x, y < \infty,$$

where ρ is a constant with $-1 < \rho < 1$. Before showing that ρ can be interpreted as the correlation coefficient of X and Y, we derive the marginal densities of X and Y. Therefore, we first decompose the joint density function $f(x,y)$ as:

$$f(x,y) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2}\ \frac{1}{\sqrt{1-\rho^2}\sqrt{2\pi}}\, e^{-\frac{1}{2}(y-\rho x)^2/(1-\rho^2)}.$$

Next observe that, for *fixed* x,

$$g(y) = \frac{1}{\sqrt{1-\rho^2}\sqrt{2\pi}}\, e^{-\frac{1}{2}(y-\rho x)^2/(1-\rho^2)}$$

is an $N(\rho x, 1-\rho^2)$ density. This implies that $\int_{-\infty}^{\infty} g(y)\,dy = 1$ and so

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2}, \qquad -\infty < x < \infty.$$

In other words, the marginal density $f_X(x)$ of X is the standard normal density. Also, for reasons of symmetry, the marginal density $f_Y(y)$ of Y is the standard normal density. Next, we prove that ρ is the correlation coefficient $\rho(X,Y)$ of X and Y. Since $\mathrm{var}(X) = \mathrm{var}(Y) = 1$, it suffices to verify that $\mathrm{cov}(X,Y) = \rho$. To do so, we use again the above decomposition of the bivariate normal density $f(x,y)$. By $E(X) = E(Y) = 0$, we have $\mathrm{cov}(X,Y) = E(XY)$ and so $\mathrm{cov}(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)\,dx\,dy$. Letting $\tau^2 = (1-\rho^2)$, it now follows that

$$\begin{aligned}\mathrm{cov}(X,Y) &= \int_{x=-\infty}^{\infty} x\frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2}\,dx \int_{y=-\infty}^{\infty} y\frac{1}{\tau\sqrt{2\pi}}\, e^{-\frac{1}{2}(y-\rho x)^2/\tau^2}\,dy \\ &= \int_{-\infty}^{\infty} \rho x^2\frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2}\,dx = \rho,\end{aligned}$$

where the third equality uses the fact that the expected value of an $N(\rho x, \tau^2)$ random variable is ρx and the last equality uses the fact that $E(Z^2) = \sigma^2(Z) = 1$ for a standard normal random variable Z.

A random vector (X,Y) is said to be *bivariate normal* distributed with parameters $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ if the standardized random vector

$$\left(\frac{X-\mu_1}{\sigma_1}, \frac{Y-\mu_2}{\sigma_2}\right)$$

has the standard bivariate normal distribution with parameter ρ. In this case the joint density $f(x, y)$ of the random variables X and Y is given by

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2}\left[(\frac{x-\mu_1}{\sigma_1})^2 - 2\rho(\frac{x-\mu_1}{\sigma_1})(\frac{y-\mu_2}{\sigma_2}) + (\frac{y-\mu_2}{\sigma_2})^2\right]/(1-\rho^2)}.$$

Rule 12.1 *Suppose that the random vector (X, Y) has a bivariate normal distribution with parameters $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. Then,*

(a) *The marginal densities $f_X(x)$ and $f_Y(y)$ of X and Y are the $N(\mu_1, \sigma_1^2)$ density and the $N(\mu_2, \sigma_2^2)$ density.*

(b) *The correlation coefficient of X and Y is given by $\rho(X, Y) = \rho$.*

The result (a) follows directly from the fact that $(X - \mu_1)/\sigma_1$ and $(Y - \mu_2)/\sigma_2$ are $N(0, 1)$ distributed, as was verified above. Also, it was shown above that the covariance of $(X - \mu_1)/\sigma_1$ and $(Y - \mu_2)/\sigma_2$ equals ρ. Using the basic formula cov$(aX + b, cY + d)$=accov(X, Y) for any constants a, b, c, and d, we next find the desired result

$$\rho = \text{cov}\left(\frac{X-\mu_1}{\sigma_1}, \frac{Y-\mu_2}{\sigma_2}\right) = \frac{1}{\sigma_1\sigma_2}\text{cov}(X,Y) = \rho(X,Y).$$

In general, uncorrelatedness is a necessary but not sufficient condition for independence of two random variables. However, for a bivariate normal distribution, uncorrelatedness is a necessary and sufficient condition for independence:

Rule 12.2 *Bivariate normal random variables X and Y are independent if and only if they are uncorrelated.*

This important result is a direct consequence of Rule 11.2, since the above representation of the bivariate normal density $f(x, y)$ reveals that $f(x, y) = f_X(x)f_Y(y)$ if and only if $\rho = 0$.

As already pointed out, the bivariate normal distribution has the important property that its marginal distributions are one-dimensional normal distributions. The following characterization of the bivariate normal distribution can be given.

Rule 12.3 *The random variables X and Y have a bivariate normal distribution if and only if $aX + bY$ is normally distributed for any constants a and b.*†

† To be precise, this result requires the following convention: if X is normally distributed and $Y = aX + b$ for constants a and b, then (X, Y) is said to have a bivariate normal distribution. This is a singular bivariate distribution: the probability mass of the two-dimensional vector (X, Y) is concentrated on the one-dimensional line $y = ax + b$. Also, a random variable X with $P(X = \mu) = 1$ for a constant μ is said to have a degenerate $N(\mu, 0)$ distribution with its mass concentrated at a single point.

13
Conditioning by random variables

In Chapter 8, conditional probabilities are introduced by conditioning upon the occurrence of an event B of nonzero probability. In applications, this event B is often of the form $Y = b$ for a discrete random variable Y. However, when the random variable Y is continuous, the condition $Y = b$ has probability zero for any number b. The purpose of this chapter is to develop techniques for handling a condition provided by the observed value of a continuous random variable. We will see that the conditional probability density function of X given $Y = b$ for continuous random variables is analogous to the conditional probability mass function of X given $Y = b$ for discrete random variables. The conditional distribution of X given $Y = b$ enables us to define the natural concept of conditional expectation of X given $Y = b$. This concept allows for an intuitive understanding and is of utmost importance. In statistical applications, it is often more convenient to work with conditional expectations instead of the correlation coefficient when measuring the strength of the relationship between two dependent random variables. In applied probability problems, the computation of the expected value of a random variable X is often greatly simplified by conditioning on an appropriately chosen random variable Y. Learning the value of Y provides additional information about the random variable X and for that reason the computation of the conditional expectation of X given $Y = b$ is often simple. The law of conditional expectation and several practical applications of this law will be discussed. In the final section we explain Bayesian inference for continuous models and give several statistical applications.

13.1 Conditional distributions

Suppose that the random variables X and Y are defined on the same sample space Ω with probability measure P. A basic question for dependent random variables X and Y is: if the observed value of Y is y, what distribution now describes the distribution of X? We first answer this question for the discrete case. Conditioning on a discrete random variable is nothing else than conditioning on an event having nonzero probability. The analysis for the continuous case involves some technical subtleties, because the probability that a continuous random variable will take on a particular value is always zero.

13.1.1 Conditional probability mass function

Let X and Y be two discrete random variables with joint probability mass function $p(x, y) = P(X = x, Y = y)$. The *conditional probability mass function* of X given that $Y = y$ is denoted and defined by

$$P(X = x \mid Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$$

for any fixed y with $P(Y = y) > 0$. This definition is just $P(A \mid B) = \frac{P(AB)}{P(B)}$ written in terms of random variables, where $A = \{\omega : X(\omega) = x\}$ and $B = \{\omega : Y(\omega) = y\}$ with ω denoting an element of the sample space. Note that, for any *fixed* y, the function $P(X = x \mid Y = y)$ satisfies

$$P(X = x \mid Y = y) \geq 0 \quad \text{for all } x \quad \text{and} \quad \sum_x P(X = x \mid Y = y) = 1,$$

showing that $P(X = x \mid Y = y)$ as function of x is indeed a probability mass function. The notation $p_X(x \,|\, y)$ is often used for $P(X = x \,|\, Y = y)$.

Using the representation

$$P(X = x, Y = y) = P(X = x \mid Y = y)P(Y = y)$$

and the fact that $\sum_y P(X = x, Y = y) = P(X = x)$, the unconditional probability $P(X = x)$ can be calculated from

$$P(X = x) = \sum_y P(X = x \mid Y = y)P(Y = y).$$

This is the law of conditional probability stated in terms of discrete random variables.

Example 13.1 Two fair dice are rolled. Let the random variable X represent the smallest of the outcomes of the two rolls, and let Y represent the sum of the outcomes of the two rolls. What are the conditional probability mass functions of X and Y?

Solution. The joint probability mass function $p(x, y) = P(X = x, Y =$

$y)$ of X and Y is given in Table 11.1. The conditional mass functions follow directly from this table. For example, the conditional mass function $p_X(x \,|\, 7) = P(X = x \,|\, Y = 7)$ is given by

$$\begin{aligned} p_X(1 \,|\, 7) &= \frac{2/36}{6/36} = \frac{1}{3},\ p_X(2 \,|\, 7) = \frac{2/36}{6/36} = \frac{1}{3},\ p_X(3 \,|\, 7) = \frac{2/36}{6/36} = \frac{1}{3}, \\ p_X(x \,|\, 7) &= 0 \qquad \text{for } x = 4, 5, 6. \end{aligned}$$

This conditional distribution is a discrete uniform distribution on $\{1, 2, 3\}$. We also give the conditional mass function $p_Y(y \,|\, 3) = P(Y = y \,|\, X = 3)$:

$$\begin{aligned} p_Y(6 \,|\, 3) &= \frac{1/36}{7/36} = \frac{1}{7},\ p_Y(7 \,|\, 3) = p_Y(8 \,|\, 3) = p_Y(9 \,|\, 3) = \frac{2/36}{7/36} = \frac{2}{7} \\ p_Y(y \,|\, 3) &= 0 \qquad \text{for } y = 2, 3, 4, 5, 10, 11, 12. \end{aligned}$$

Problem 13.1 You repeatedly draw a number at random from the numbers $1, 2, \ldots, 10$ with replacement. Let X be the number of draws until the number 1 appears for the first time and Y the number of draws until the number 10 appears for the first time. What is the conditional mass function $P(X = x \mid Y = y)$?

Problem 13.2 Three different numbers are chosen at random from the numbers $1, 2, \ldots, 10$. Let X be the smallest of these three numbers and Y the largest. What are the conditional mass functions $P(X = x \mid Y = y)$ and $P(Y = y \mid X = x)$?

Problem 13.3 A fair die is rolled until a six appears. Each time the die is rolled a fair coin is tossed. Let the random variables X and Y denote the number of rolls of the die and the number of heads from the tosses of the coin. What is the conditional mass function $P(X = x \mid Y = y)$. *Hint*: use the identity $\sum_{n=r}^{\infty} \binom{n}{r} a^n = a^r/(1-a)^{r+1}$ for $0 < a < 1$ to find $P(Y = j)$ when $j \neq 0$.

Problem 13.4 Two dice are rolled. Let the random variable X be the smallest of the two outcomes and let Y be the largest of the two outcomes. What are the conditional mass functions $P(X = x \mid Y = y)$ and $P(Y = y \mid X = x)$?

Problem 13.5 You simultaneously roll 24 dice. Next you roll only those dice that showed the face value six in the first roll. Let the random variable X denote the number of sixes in the first roll and Y the number of sixes in the second roll. What is the conditional mass function $P(X = x \mid Y = y)$?

Problem 13.6 Let X denote the number of hearts and Y the number of diamonds in a bridge hand. What are the conditional mass functions $P(X = x \mid Y = y)$ and $P(Y = y \mid X = x)$?

13.1.2 Conditional probability density function

What is the continuous analog of the conditional probability mass function when X and Y are continuous random variables with a joint probability density function $f(x, y)$? In this situation, we have the complication that $P(Y = y) = 0$ for each real number y. Nevertheless, this situation also allows for a natural definition of the concept of conditional distribution. Toward this end, we need the probabilistic interpretations of the joint density function $f(x,\ y)$ and the marginal densities $f_X(x)$ and $f_Y(y)$ of the random variables X and Y. For small values of $\Delta a > 0$ and $\Delta b > 0$,

$$\begin{aligned} & P(a - \frac{1}{2}\Delta a \leq X \leq a + \frac{1}{2}\Delta a \,\big|\, b - \frac{1}{2}\Delta b \leq Y \leq b + \frac{1}{2}\Delta b) \\ = \quad & \frac{P(a - \frac{1}{2}\Delta a \leq X \leq a + \frac{1}{2}\Delta a, b - \frac{1}{2}\Delta b \leq Y \leq b + \frac{1}{2}\Delta b)}{P(b - \frac{1}{2}\Delta b \leq Y \leq b + \frac{1}{2}\Delta b)} \\ \approx \quad & \frac{f(a, b)\Delta a\Delta b}{f_Y(b)\Delta b} = \frac{f(a, b)}{f_Y(b)}\Delta a \end{aligned}$$

provided that (a, b) is a continuity point of $f(x, y)$ and $f_Y(b) > 0$. This leads to the following definition.

Definition 13.1 *If X and Y are continuous random variables with joint probability density function $f(x, y)$ and $f_Y(y)$ is the marginal density function of Y, then the conditional probability density function of X given that $Y = y$ is defined by*

$$f_X(x \mid y) = \frac{f(x, y)}{f_Y(y)}, \qquad -\infty < x < \infty$$

for any fixed y with $f_Y(y) > 0$.

Note that, for any *fixed* y, the function $f_X(x \mid y)$ satisfies

$$f_X(x \mid y) \geq 0 \quad \text{for all } x \quad \text{and} \quad \int_{-\infty}^{\infty} f_X(x \mid y)\, dx = 1,$$

showing that $f_X(x \mid y)$ as function of x is indeed a probability density function. Similarly, the conditional probability density function of the random variable Y given that $X = x$ is defined by $f_Y(y \,|\, x) = \frac{f(x,y)}{f_X(x)}$ for any fixed x with $f_X(x) > 0$.

A probabilistic interpretation can be given to $f_X(a \mid b)$: given that the observed value of Y is b, the probability of the other random variable X taking on a value in a small interval of length Δa around the point a is approximately equal to $f_X(a \mid b)\Delta a$ if a is a continuity point of $f_X(x \mid b)$.

The concept of conditional probability distribution function is defined as follows. For any fixed y with $f_Y(y) > 0$, the conditional probability that the random variable X takes on a value smaller than or equal to x given that $Y = y$ is denoted by $P(X \leq x \mid Y = y)$ and is defined by

$$P(X \leq x \mid Y = y) = \int_{-\infty}^{x} f_X(u \mid y)\, du.$$

Before discussing implications of this definition, we illustrate the concept of conditional probability density function with two examples.

Example 13.2 A point (X, Y) is chosen at random inside the unit circle. What is the conditional density of X?

Solution. In Example 11.6, we determined the joint density function $f(x, y)$ of X and Y together with the marginal density function $f_Y(y)$ of Y. This gives for any fixed y with $-1 < y < 1$,

$$f_X(x \,|\, y) = \begin{cases} \frac{1}{2\sqrt{1-y^2}} & \text{for } -\sqrt{1-y^2} < x < \sqrt{1-y^2} \\ 0 & \text{otherwise.} \end{cases}$$

In other words, the conditional distribution of X given that $Y = y$ is the uniform distribution on the interval $(-\sqrt{1-y^2}, \sqrt{1-y^2})$. The same distribution as that of the x-coordinate of a randomly chosen point of the horizontal chord through the point $(0, y)$. This chord has length $2\sqrt{1-y^2}$, by Pythagoras.

Example 13.3 Suppose that the random variables X and Y have a bivariate normal distribution with parameters $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. What are the conditional probability densities of X and Y?

Solution. The joint density function $f(x, y)$ is specified in Section 12.1. Also, in this section we find that the marginal probability densities $f_X(x)$ and $f_Y(y)$ of X and Y are given by the $N(\mu_1, \sigma_1^2)$ density and the $N(\mu_2, \sigma_2^2)$ density. Substituting the expressions for these densities in the formulas for the conditional densities, we find after simple algebra that the conditional probability density $f_X(x \mid y)$ of X given that $Y = y$ is the

$$N\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y - \mu_2), \sigma_1^2(1 - \rho^2)\right)$$

14
Generating functions

Generating functions were introduced by the Swiss genius Leonhard Euler (1707–1783) in the eighteenth century to facilitate calculations in counting problems. However, this important concept is also extremely useful in applied probability, as was first demonstrated by the work of Abraham de Moivre (1667–1754) who discovered the technique of generating functions independently of Euler. In modern probability theory, generating functions are an indispensable tool in combination with methods from numerical analysis.

The purpose of this chapter is to give the basic properties of generating functions and to show the utility of this concept. First, the generating function is defined for a discrete random variable on nonnegative integers. Next, we consider the more general moment-generating function, which is defined for any random variable. The (moment) generating function is a powerful tool for both theoretical and computational purposes. In particular, it can be used to prove the central limit theorem. A sketch of the proof will be given. This chapter also gives a proof of the strong law of large numbers, using moment-generating functions together with so-called Chernoff bounds. Finally, the strong law of large numbers is used to establish the powerful renewal-reward theorem for stochastic processes having the property that the process probabilistically restarts itself at certain points of time.

14.1 Generating functions

We first introduce the concept of generating function for a discrete random variable X whose possible values belong to the set of nonnegative integers.

Definition 14.1 *If X is a nonnegative, integer-valued random variable, then* *the generating function of X is defined by*

$$G_X(z) = \sum_{k=0}^{\infty} z^k P(X = k), \qquad |z| \leq 1.$$

The power series $G_X(z)$ is absolutely convergent for any $|z| \leq 1$ (why?). For any z, we can interpret $G_X(z)$ as

$$G_X(z) = E\left(z^X\right),$$

as follows by applying Rule 9.2. The probability mass function of X is uniquely determined by the generating function of X. To see this, use the fact that the derivative of an infinite series is obtained by differentiating the series term by term. Thus,

$$\frac{d^r}{dz^r} G_X(z) = \sum_{k=r}^{\infty} k(k-1)\cdots(k-r+1) z^{k-r} P(X = k), \qquad r = 1, 2, \ldots .$$

In particular, by taking $z = 0$,

$$P(X = r) = \frac{1}{r!} \frac{d^r}{dz^r} G_X(z)|_{z=0}, \qquad r = 1, 2, \ldots .$$

This proves that the generating function uniquely determines the probability mass function. This basic result explains the importance of the generating function. In many applications, it is relatively easy to obtain the generating function of a random variable X even when the probability mass function is not explicitly given. An example will be given below. Once we know the generating function of a random variable X, it is a simple matter to obtain the factorial moments of the random variable X. The rth factorial moment of the random variable X is defined by $E[X(X-1)\cdots(X-r+1)]$ for $r = 1, 2, \ldots$. In particular, the first factorial moment of X is the expected value of X. The variance of X is determined by the first and the second factorial moment of X. Putting $z = 1$ in the above expression for the rth derivative of $G_X(z)$, we obtain

$$E\left[X(X-1)\cdots(X-r+1)\right] = \frac{d^r}{dz^r} G_X(z)|_{z=1}, \qquad r = 1, 2, \ldots .$$

In particular,

$$E(X) = G'_X(1) \quad \text{and} \quad E\left(X^2\right) = G''_X(1) + G'_X(1).$$

Example 14.1 Suppose that the random variable X has a Poisson distribution with expected value λ. Verify that the generating function of X is given by

$$G_X(z) = e^{-\lambda(1-z)}, \qquad |z| \leq 1.$$

What are the expected value and the standard deviation of X?

Solution. Applying the definition of generating function and using the series expansion $e^x = \sum_{n=0}^{\infty} x^n/n!$, we find

$$\sum_{k=0}^{\infty} z^k e^{-\lambda} \frac{\lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda z)^k}{k!} = e^{-\lambda} e^{\lambda z},$$

as was to be verified. Differentiating $G_X(z)$ gives $G'_X(1) = \lambda$ and $G''_X(1) = \lambda^2$. Hence, $E(X) = \lambda$ and $E(X^2) = \lambda^2 + \lambda$. This implies that both the expected value and the variance of a Poisson-distributed random variable with parameter λ are given by λ, in agreement with earlier results in Example 9.8.

14.1.1 Convolution rule

The importance of the concept of generating function comes up especially when calculating the probability mass function of a sum of independent random variables that are nonnegative and integer-valued.

Rule 14.1 *Let X and Y be two nonnegative, integer-valued random variables. If the random variables X and Y are independent, then*

$$G_{X+Y}(z) = G_X(z) G_Y(z), \qquad |z| \leq 1.$$

Rule 14.1 is known as the *convolution rule* for generating functions and can be directly extended to the case of a finite sum of independent random variables. The proof is simple. If X and Y are independent, then the random variables $U = z^X$ and $V = z^Y$ are independent for any fixed z (see Rule 9.5). Also, by Rule 9.7, $E(UV) = E(U)E(V)$ for independent U and V. Thus

$$E\left(z^{X+Y}\right) = E\left(z^X z^Y\right) = E\left(z^X\right) E\left(z^Y\right),$$

proving that $G_{X+Y}(z) = G_X(z)G_Y(z)$. The converse of the statement in Rule 14.1 is, in general, not true. The random variables X and Y are not necessarily independent if $G_{X+Y}(z) = G_X(z)G_Y(z)$. It is left to the reader to verify that a counterexample is provided by the random vector (X, Y) that takes on the values (1,1), (2,2) and (3,3) each with probability $\frac{1}{9}$ and the values (1,2), (2,3) and (3,1) each with probability $\frac{2}{9}$. This counterexample was communicated to me by Fred Steutel.

Example 14.2 Suppose that X and Y are independent random variables that are Poisson distributed with respective parameters λ and μ. What is the probability mass function of $X + Y$?

15
Discrete-time Markov Chains

In previous chapters we have dealt with sequences of independent random variables. However, many random systems evolving in time involve sequences of dependent random variables. Think of the outside weather temperature on successive days, or the prize of IBM stock at the end of successive trading days. Many such systems have the property that the current state alone contains sufficient information to give the probability distribution of the next state. The probability model with this feature is called a Markov chain. The concepts of state and state transition are at the heart of Markov chain analysis. The line of thinking through the concepts of state and state transition is very useful to analyze many practical problems in applied probability.

Markov chains are named after the Russian mathematician Andrey Markov (1856-1922), who first developed this probability model in order to analyze the alternation of vowels and consonants in Pushkin's poem "Eugine Onegin." His work helped to launch the modern theory of stochastic processes (a *stochastic process* is a collection of random variables, indexed by an ordered time variable). The characteristic property of a Markov chain is that its memory goes back only to the most recent state. Knowledge of the current state only is sufficient to describe the future development of the process. A Markov model is the simplest model for random systems evolving in time when the successive states of the system are not independent. But this model is no exception to the rule that simple models are often the most useful models for analyzing practical problems. The theory of Markov chains has applications to a wide variety of fields, including biology, physics, engineering, and computer science.

In this chapter we only consider Markov chains with a finite number of states. We first present techniques to analyze the time-dependent behavior of Markov chains. In particular, we give much attention to Markov chains with one or more absorbing states. Such Markov chains have interesting applications in the analysis of success runs. Next, we deal with the long-run behavior of Markov chains and give solution methods to answer questions such as: what is the long-run proportion of time that the system will be in any given subset of states. Finally, we discuss the method of Markov chain Monte Carlo simulation which has revolutionized the field of Bayesian statistics and many other areas of science.

15.1 Markov chain model

A Markov chain deals with a collection of random variables, indexed by an ordered time parameter. The Markov model is the simplest conceivable generalization of a sequence of independent random variables. A Markov chain is a sequence of trials having the property that the outcome of each last trial provides enough information to predict the outcome of any future trial. Despite its very simple structure, the Markov model is extremely useful in a wide variety of practical probability problems. The beginning student often has difficulties in grasping the concept of the Markov chain when a formal definition is given. Let's begin with an example that illustrates the essence of what a Markov process is.

Example 15.1 A drunkard wanders about a town square. At each step he no longer remembers anymore the direction of his previous steps. Each step is a unit distance in a randomly chosen direction and has equal probability $\frac{1}{4}$ of going north, south, east or west as long as the drunkard has not reached the edge of the square (see Figure 15.1). The drunkard never leaves the square. Should he reach the boundary of the square, his next step is equally likely to be in one of the three remaining directions if he is not at a corner point, and is equally likely to be in one of the two remaining directions otherwise. The drunkard starts in the middle of the square. What stochastic process describes the drunkard's walk?

Solution. To answer this question, we define the random variable X_n as

$$X_n = \text{the position of the drunkard just after the } n\text{th step}$$

for $n = 0, 1, \ldots$ with the convention $X_0 = (0, 0)$. We say that the drunkard is in state (x, y) when the current position of the drunkard is described by the point (x, y). The collection $\{X_0, X_1, \ldots\}$ of random variables is a stochastic process with discrete time-parameter and finite state space

$$I = \{(x, y) : x, y \text{ integer and } -L \leq x, y \leq L\},$$

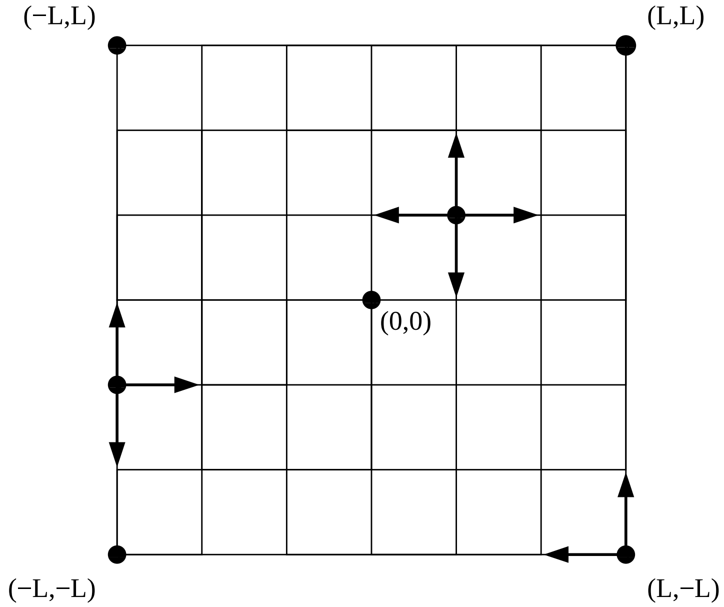


Fig. 15.1. The drunkard's walk.

where L is the distance from the middle of the square to its boundary. The successive states of the drunkard are not independent of each other, but the next position of the drunkard depends only on his current position and is not influenced by the earlier positions in his path. That is, the process $\{X_0, X_1, \ldots\}$ has the so-called *Markovian property*, which says that the state at any given time summarizes everything about the past that is relevant to the future.

Many random systems evolving over time can be modeled to satisfy the Markovian property. Having this property introduced informally, we are now ready to give a formal definition of a Markov chain. Let $X_0, X_1, \ldots$ be a sequence of random variables. It is helpful to think of X_n as the state of a dynamic system at time $t = n$. The sequence $X_0, X_1, \ldots$ is called a discrete-time stochastic process. In the sequel, the set of possible values of the random variables X_n is assumed to be *finite* and is denoted by I. The set I is called the *state space* of the stochastic process $\{X_0, X_1, \ldots\}$.

Definition 15.1 *The stochastic process* $\{X_n,\ n = 0, 1, \ldots\}$ *with state space* I *is said to be a discrete-time Markov chain if it possesses the Markovian property, that is, for each time point* $n = 0, 1, \ldots$ *and all possible values of* *the states* $i_0, i_1, \ldots, i_{n+1} \in I$, *the process has the property*

$$\begin{aligned}&P(X_{n+1} = i_{n+1} \mid X_0 = i_0, X_1 = i_1, \ldots, X_{n-1} = i_{n-1}, X_n = i_n)\\ &= P(X_{n+1} = i_{n+1} \mid X_n = i_n).\end{aligned}$$

The term $P(X_{n+1} = i_{n+1} \mid X_0 = i_0, X_1 = i_1, \ldots, X_{n-1} = i_{n-1}, X_n = i_n)$ should be read as follows: it is the conditional probability that the system will be in state i_{n+1} at the *next* time point $t = n+1$ if the system is in state i_n at the *current* time $t = n$ and has reached the current state i_n via the states $i_0, i_1, \ldots, i_{n-1}$ at the *past* time points $t = 0, 1, \ldots, n-1$. The Markovian property says that this conditional probability depends only the current state i_n and is not altered by knowledge of the past states $i_0, i_1, \ldots, i_{n-1}$. The current state summarizes everything about the past that is relevant to the future. At any time $t = n$ the process essentially forgets how it got into the state X_n. It is not true that the state X_{n+1} at the next time point $t = n + 1$ is independent of $X_0, \ldots, X_{n-1}$, but all of the dependency is captured by X_n.

The Markov chain approach is a very powerful tool in applied probability. Using the concept of state and choosing the state in an appropriate way, numerous probability problems can be solved by Markov chain methods.

In Example 15.1 the Markovian property was satisfied in a natural way by choosing the state of the process as the position of the drunkard on the square. However, in other applications the choice of the state variable(s) may require more thought in order to satisfy the Markovian property. To illustrate this, consider Example 15.1 again and assume now that the drunkard never chooses the same direction as was chosen in the previous step. Then, we need an extra state variable in order to satisfy the Markovian property. Let's say that the drunkard is in state (x, y, N) when the position of the drunkard on the square is (x, y) and he moved northward in his previous step. Similarly, the states $(x, y, E), (x, y, S)$ and (x, y, W) are defined. Letting X_n be the state of the drunkard after the nth step (with the convention $X_0 = (0, 0)$), the stochastic process $\{X_0, X_1, \ldots\}$ satisfies the Markovian property and thus is a Markov chain. The transition probabilities are easy to give. For example, if the current state of the process is (x, y, S) with (x, y) an interior point of the square, the next state of the process is equally likely to be one of the three states $(x+1, y, E), (x-1, y, W)$, and $(x, y+1, N)$. In the drunkard's walk the concepts of *state* and *state transition* come up in a natural way. These concepts are at the heart of Markov chain analysis.

In the following, we will restrict our attention to time-homogeneous Markov chains. For such chains the transition probability $P(X_{n+1} = j \mid$

$X_n = i$) does not depend on the value of the time parameter n and so $P(X_{n+1} = j \mid X_n = i) = P(X_1 = j \mid X_0 = i)$ for all n. We write

$$p_{ij} = P(X_{n+1} = j \mid X_n = i).$$

The probabilities p_{ij} are called the *one-step transition probabilities* of the Markov chain and are the same for all time points n. They satisfy

$$p_{ij} \geq 0 \qquad \text{for } i, j \in I \qquad \text{and} \qquad \sum_{j \in I} p_{ij} = 1 \qquad \text{for all } i \in I.$$

The notation p_{ij} is sometimes confusing for the beginning student: p_{ij} is not a joint probability, but a conditional probability. However, the notation p_{ij} rather than the notation $p(j \mid i)$ has found widespread acceptance.

A Markov chain $\{X_n,\, n = 0, 1, \ldots\}$ is completely determined by the probability distribution of the initial state X_0 and the one-step transition probabilities p_{ij}. In applications of Markov chains the art is:

(a) to choose the state variable(s) such that the Markovian property holds
(b) to determine the one-step transition probabilities p_{ij}.

How to formulate a Markov chain model for a concrete problem is largely an art that is developed with practice. Putting yourselves in the shoes of someone who has to write a simulation program for the problem in question may be helpful in choosing the state variable(s). Once the (difficult) modeling step is done, the rest is simply a matter of applying the theory that will be developed in the next sections. The student cannot be urged strongly enough to try the problems at the end of this section to acquire skills to model new situations. In order to help students develop intuition into how practical situations can be modeled as a Markov chain, we give three examples. The first example deals with the Ehrenfest model for gas diffusion. In physics the Ehrenfest model resolved at the beginning of the twentieth century a seeming contradiction between the second law of thermodynamics and the laws of mechanics.

Example 15.2 Two compartments A and B together contain r particles. With the passage of every time unit, one of the particles is selected at random and is removed from its compartment to the other. What stochastic process describes the contents of the compartments?

Solution. Let us take as state of the system the number of particles in compartment A. If compartment A contains i particles, then compartment B contains $r - i$ particles. Define the random variable X_n as

$$X_n = \text{the number of particles in compartment } A \text{ after the } n\text{th transfer.}$$

By the physical construction of the model with independent selections of a particle, the process $\{X_n\}$ satisfies the Markovian property and thus is a Markov chain. The state space is $I = \{0, 1, \ldots, r\}$. The probability of going from state i to state j in one step is zero unless $|i - j| = 1$. The one-step transition probability $p_{i,i+1}$ translates into the probability that the randomly selected particle belongs to compartment B and $p_{i,i-1}$ translates into the probability that the randomly selected particle belongs to compartment A. Thus, for $1 \leq i \leq r - 1$,

$$p_{i,i+1} = \frac{r-i}{r} \quad \text{and} \quad p_{i,i-1} = \frac{i}{r}.$$

Further, $p_{01} = p_{r,r-1} = 1$. The other p_{ij} are zero.

Example 15.3 An absent-minded professor drives every morning from his home to the office and at the end of the day from the office to home. At any given time, his driver's license is located at his home or at the office. If his driver's licence is at his location of departure, he takes it with him with probability 0.5. What stochastic process describes whether the professor has the driver's license with him when driving his car to home or to the office?

Solution. Your first thought might be to define two states 1 and 0, where state 1 describes the situation that the professor has his driver's licence with him when driving his car and state 0 describes the situation that he does not have his driver's license with him when driving his car. However, these two states do not suffice for a Markov model: state 0 does not provide enough information to predict the state at the next drive. In order to give the probability distribution of this next state, you need information about the current location of the driver's license of the professor. You get a Markov model by simply inserting this information into the state description. Let's say that the system is in state 1 if the professor is driving his car and has his driver's license with him, in state 2 if the professor is driving his car and his driver's license is at the point of departure, and in state 3 if the professor is driving his car and his driver's license is at his destination. Define the random variable X_n as

$$X_n = \text{the state at the } n\text{th drive to home or to the office.}$$

The process $\{X_n\}$ has the property that any present state contains sufficient information for predicting future states. Thus, the process $\{X_n\}$ is a Markov chain with state space $I = \{1, 2, 3\}$. Next, we determine the p_{ij}. For example, p_{32} translates into the probability that the professor will not have his driver's license with him at the next drive given that his driver's license

16

Continuous-time Markov Chains

Many random phenomena happen in continuous time. Examples include occurrence of cell phone calls, spread of epidemic diseases, stock fluctuations, etc. A continuous-time Markov chain is a very useful stochastic process to model such phenomena. It is a process that goes from state to state according to a Markov chain, but the times between state transitions are continuous random variables having an exponential distribution.

The purpose of this chapter is to give am elementary introduction to continuous-time Markov chains. The basic concept of the continuous-time Markov chain model is the so-called transition rate function. Several examples will be given to illustrate this basic concept. Next we discuss the time-dependent behavior of the process and give Kolmogorov's differential equations to compute the time-dependent state probabilities. Finally, we present the flow-rate-equation method to compute the limiting state probabilities and illustrate this powerful method with several examples dealing with queueing systems.

16.1 Markov chain model

A continuous-time stochastic process $\{X(t),\, t \geq 0\}$ is a collection of random variables indexed by a continuous time parameter $t \in [0, \infty)$, where the random variable $X(t)$ is called the state of the process at time t. In an inventory problem $X(t)$ might be the stock on hand at time t and in a queueing problem $X(t)$ might be the number of customers present at time t. The formal definition of a continuous-time Markov chain is a natural extension of the definition of a discrete-time Markov chain.

Definition 16.1 *The stochastic process $\{X(t),\, t \geq 0\}$ with discrete state space I is said to be a continuous-time Markov chain if it possesses the*

Markovian property, that is, for all time points $s, t \geq 0$ *and states* $i, j, x(u)$ *with* $0 \leq u < s$,

$$
\begin{aligned}
&P(X(t+s)=j) \mid X(u)=x(u) \text{ for } 0 \leq u < s,\, X(s)=i) \\
&= P(X(t+s)=j \mid X(s)=i).
\end{aligned}
$$

In words, the Markovian property says that if you know the present state at time s, then all additional information about the states at times prior to time s is irrelevant for the probabilistic development of the process in the future. All that matters for future states is what the present state is. A continuous-time Markov chain is said to be *time-homogeneous* if for any $s, t > 0$ and any states $i, j \in I$,

$$P(X(t+s)=j \mid X(s)=i) = P(X(t)=j \mid X(0)=i).$$

The transition functions $p_{ij}(t)$ are defined by

$$p_{ij}(t) = P(X(t)=j \mid X(0)=i) \quad \text{for } t \geq 0 \text{ and } i, j \in I.$$

In addition to the assumption of time-homogeneity, we now make the assumption that the state space I is *finite*. This assumption is made to avoid technical complications involved with a countably infinite state space. However, under some regularity conditions the results for the case of a finite state space carry over to the case of a countably infinite state space.

Transition rates

In continuous time there are no smallest time steps and hence we cannot speak about one-step transition probabilities as in discrete time. In a continuous-time Markov chain we would like to know, for very small time steps of length h, what the probability $p_{ij}(h)$ is of being in a *different* state j at time $t+h$ if the present state at time t is i. This probability is determined by the so-called transition rates.† These transition rates are to continuous-time Markov chains what the one-step transition probabilities are to discrete-time Markov chains. Formally, the transition rate q_{ij} can be introduced as the derivative of $p_{ij}(t)$ at $t=0$. For the case of a finite state space, the $p_{ij}(t)$ are differentiable for any $t \geq 0$. The reader is asked to take for granted this deep result. In particular, the right-hand derivative of $p_{ij}(t)$ at $t=0$ exists for all i, j. This limit is denoted by q_{ij} for $j \neq i$ and by $-\nu_i$

† Rate is a measure of how quickly something happens. It can be seen as the frequency at which a repeatable event happens per unit time. That is, an arrival rate λ means that the average frequency of arrivals per unit time is λ.

for $j = i$. Using the fact that

$$p_{ij}(0) = \begin{cases} 0 & \text{for } j \neq i \\ 1 & \text{for } j = i, \end{cases}$$

we have $(p_{ij}(h) - p_{ij}(0))/h = p_{ij}(h)/h$ for $j \neq i$ and $(p_{ii}(h) - p_{ii}(0))/h = (p_{ii}(h) - 1)/h$ for $j = i$. The rates q_{ij} and the rates ν_i are now defined by

$$q_{ij} = \lim_{h \to 0} \frac{p_{ij}(h)}{h} \quad \text{for } j \neq i \quad \text{and} \quad \nu_i = \lim_{h \to 0} \frac{1 - p_{ii}(h)}{h}.$$

Using the mathematical symbol $o(h)$ we can write this in a more convenient form. The symbol $o(h)$ is the generic notation for any function $f(h)$ with the property that

$$\lim_{h \to 0} \frac{f(h)}{h} = 0,$$

that is, $o(h)$ represents some unspecified function that is negligibly small compared to h itself as $h \to 0$. For example, any function $f(h) = h^a$ with $a > 1$ is an $o(h)$-function. A useful fact is that both the sum and the product of a finite number of $o(h)$-functions are again $o(h)$-functions. Summarizing,

Rule 16.1 *For any $t \geq 0$ and small h,*

$$P(X(t+h) = j \mid X(t) = i) = \begin{cases} q_{ij}h + o(h) & \text{for } j \neq i \\ 1 - \nu_i h + o(h) & \text{for } j = i. \end{cases}$$

The transition rate q_{ij} gives the rate at which the process tries to enter a different state j when the process is in state i. The exit rate ν_i gives the rate at which the process tries to leave state i for another state when the process is in state i. Since $p_{ii}(h) + \sum_{j \neq i} p_{ij}(h) = 1$, it follows that

$$\nu_i = \sum_{j \neq i} q_{ij} \quad \text{for all } i \in I.$$

It is important to note that the q_{ij} are rates, not probabilities and, as such, while they must be nonnegative, they are not bounded by 1. However, for very small h, you can interpret $q_{ij}h$ as a probability, namely as the probability that in the next time interval h the process will jump to a different state j when the present state is i. Also, it is important to note from Rule 16.1 and the finiteness of the state space that the probability of two or more state changes in such a small time interval is $o(h)$. The transition rates q_{ij} are obtained from the transition functions $p_{ij}(t)$. Conversely, it can be proved that transition rates q_{ij} uniquely determine transition functions $p_{ij}(t)$ when the state space is finite.

How should you find the q_{ij} in practical applications? You will see that the exponential distribution is the building block for the transition rates. Before proceeding with the Markov chain model, we first discuss the most important properties of this distribution and the closely related Poisson process.

Intermezzo: the exponential distribution and the Poisson process

Recall from Chapter 10 that a continuous random variable T is said to have an exponential distribution with rate $\lambda > 0$ if its probability density function $f(t)$ is given by

$$f(t) = \begin{cases} \lambda e^{-\lambda t} & \text{for } t > 0 \\ 0 & \text{otherwise .} \end{cases}$$

The cumulative probability distribution function of T is $P(T \leq t) = 1 - e^{-\lambda t}$ for $t \geq 0$, and T has $1/\lambda$ as expected value. The key property of the exponential distribution is its memoryless property. That is, for any $s, t \geq 0$,

$$P(T > t + s \mid T > s) = P(T > t).$$

Imagining that T represents the lifetime of an item, this property says that the remaining lifetime of the item has the same exponential distribution as the original lifetime, regardless of how long the item has been already in use. This memoryless property has been proved in Section 10.4.3. The exponential distribution is the only continuous distribution possessing this property. For building and understanding continuous-time Markov chains, it is more convenient to express the memoryless property of the exponential distribution as

$$P(T \leq t + h \mid T > t) = \lambda h + o(h) \quad \text{as } h \to 0,$$

no matter what the value of t is. In other words, the exponential distribution has a constant failure rate. The proof is as follows. By the memoryless property, $P(T \leq t + h \mid T > t) = e^{-\lambda h}$. Expanding out $e^{-\lambda h}$ in a Taylor series, we find

$$\begin{aligned} P(T \leq t + h \mid T > t) &= 1 - \left(1 - \frac{\lambda h}{1!} + \frac{(\lambda h)^2}{2!} - \frac{(\lambda h)^3}{3!} + \cdots\right) \\ &= \lambda h + o(h) \quad \text{as } h \to 0. \end{aligned}$$

The Poisson process often appears in applications of continuous-time Markov chains. This process was already discussed in Section 4.2.3, see also Section 10.4.3. It is a continuous-time counting process $\{N(t),\, t \geq 0\}$, where $N(t)$ is the number of events (e.g. arrivals of customers or jobs) that have occurred up to time t. Several equivalent definitions of the Poisson process can be given.

First definition A stochastic process $\{N(t),\, t \geq 0\}$ with $N(0) = 0$ is said to be a Poisson process with rate λ if

(a) The random variable $N(t)$ counts the number of events that have occurred up to time t.
(b) The times between events are independent random variables having a common exponential distribution with expected value $1/\lambda$.

From this definition the following memoryless property can be shown for the Poisson process: at each point of time the waiting time until the next occurrence of an event has the same exponential distribution as the original inter-occurrence times, regardless of how long ago the last event has occurred. For our purposes in continuous-time Markov chains, the following equivalent definition is more appropriate.

Second definition A stochastic process $\{N(t),\, t \geq 0\}$ with $N(0) = 0$ is said to be a Poisson process with rate λ if

(a) Occurrences of events in any time interval $(t, t+h)$ are independent of what happened up to time t.
(b) For any $t \geq 0$, the probability $P(N(t+h) - N(t) = 1)$ of one occurrence of an event in the time interval $(t, t+h)$ is $\lambda h + o(h)$, the probability $P(N(t+h) - N(t) = 0)$ of no occurrence of an event is $1 - \lambda h + o(h)$, and the probability $P(N(t+h) - N(t) \geq 2)$ of more than one is $o(h)$ as $h \to 0$.

The proof of the equivalence of the two definitions will not be given. As pointed out in Chapter 4, the reason for the name of Poisson process is the fact that the number of events occurring in any given time interval of length t has a Poisson distribution with expected value λt.

Alternative construction of a continuous-time Markov chain

In this paragraph we give a revealing way to think about a continuous-time Markov chain. The process can be characterized in another way that leads to an understanding how to simulate the process. When the continuous-time Markov chain $\{X(t)\}$ enters state i at some time, say time 0, the process stays there a random amount of time before making a transition into a different state. Denote this amount of time by T_i. What does Definition 16.1 suggest about this random variable T_i? Could we directly find $P(T_i > t)$?