WILEY SERIES IN PROBABILITY AND STATISTICS

ADVANCED STATISTICS WITH APPLICATIONS IN R

EUGENE DEMIDENKO

WILEY

Advanced Statistics with Applications in R

The ***Wiley Series in Probability and Statistics*** is well established and authoritative. It covers many topics of current research interest in both pure and applied statistics and probability theory. Written by leading statisticians and institutions, the titles span both state-of-the-art developments in the field and classical methods.

Reflecting the wide range of current research in statistics, the series encompasses applied, methodological and theoretical statistics, ranging from applications and new techniques made possible by advances in computerized practice to rigorous treatment of theoretical approaches. This series provides essential and invaluable reading for all statisticians, whether in academia, industry, government, or research.

Advanced Statistics with Applications in R

Eugene Demidenko

Dartmouth College

WILEY

This edition first published 2020

Registered Office
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA

Editorial Office
111 River Street, Hoboken, NJ 07030, USA

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Limit of Liability/Disclaimer of Warranty
While the publisher and authors have used their best efforts in preparing this work, they make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives, written sales materials or promotional statements for this work. The fact that an organization, website, or product is referred to in this work as a citation and/or potential source of further information does not mean that the publisher and authors endorse the information or services the organization, website, or product may provide or recommendations it may make. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for your situation. You should consult with a specialist where appropriate. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

Library of Congress Cataloging-in-Publication Data
Names: Demidenko, Eugene, 1948- author.
Title: Advanced statistics with applications in R / Eugene Demidenko
(Dartmouth College).
Description: Hoboken, NJ : Wiley, 2020. | Series: Wiley series in probability
and statistics | Includes bibliographical references and index. |
Identifiers: LCCN 2019015124 (print) | LCCN 2019019543 (ebook) | ISBN
9781118594131 (Adobe PDF) | ISBN 9781118594612 (ePub) | ISBN 9781118387986
(hardback)
Subjects: LCSH: Mathematical statistics–Data processing–Problems,
exercises, etc. | Statistics–Data processing–Problems, exercises, etc. |
R (Computer program language)
Classification: LCC QA276.45.R3 (ebook) | LCC QA276.45.R3 D4575 2019 (print)
| DDC 519.5–dc23
LC record available at https://lccn.loc.gov/2019015124

Cover design by Wiley
Cover image: Courtesy of Eugene Demidenko

Set in 11/14pt Computer Modern by SPi Global, Chennai, India

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Contents

To my family

Why I Wrote This Book

My favorite part of the recent American Statistical Association (ASA) statement on the p-value [103] is how it starts: "Why do so many people still use $p = 0.05$ as a threshold?" with the answer "Because that's what they were taught in college or grad school." Many problems in understanding and the interpretation of statistical inference, including the central statistical concept of the p-value arise from the shortage of textbooks in statistics where theoretical and practical aspects of statistics fundamentals are put together. On one hand, we have several excellent theoretical textbooks including Casella and Berger [17], Schervish [87], and Shao [94] without single real-life data example. On the other hand, there are numerous recipe-style statistics textbooks where theoretical considerations, assumptions, and explanations are minimized. This book fills that gap.

Statistical software has become so convenient and versatile these days that many use it without understanding the underlying principles. Unfortunately, `R` packages do not explain the algorithms and mathematics behind computations, greatly contributing to a superficial understanding making statistics too easy. Many times, to my question "How did you compute this, what is the algorithm," I hear the answer, "I found a program on the Internet." Hopefully, this book will break the unwanted trend of such statistics consumption.

I have often been confronted with the question comparing statistics with driving a car: "Why do we need to know how the car works?" Well, because statistics is not a car: the chance of the car breaking is slim, but starting with the wrong statistical analysis is almost guaranteed without solid understanding of statistics background and implied limitations. In this book, we look at what is under the hood.

Each term I start my first class in statistics at Dartmouth with the following statement:

"Mathematics is the queen and statistics is the king of all sciences"

Indeed, mathematics is the idealistic model of the world: one line goes through a pair of points, the perimeter of a polygon converges to $2\pi r$ when the number of edges goes to infinity, etc. Statistics fills mathematics with life. Due to an unavoidable measurement error, one point turns into a cloud of points. How does one draw a line through two clouds of points? How does one measure π in real life? This book starts with a motivating example "Who said π?" in which I suggest to measuring π by taking the ratio of the perimeter of the tire to its diameter. To the surprise of many, the average ratio does not converge to π even if the measurement error is very small. The reader will learn how this seemingly easy problem of estimating π turns into a formidable statistical problem. Statistics is where the rubber meets the road. It is difficult to name a science where statistics is not used.

Examples are a big deal in this book (there are 442 examples in the book). I follow the saying: "Examples are the expressway to knowledge." Only examples show how to use theory and how to solve a real-life problem. Too many theories remain unusable.

Today statistics is impossible without programming: that is why `R` is the language statisticians speak. The era of statistics textbooks with tables of distributions in an appendix is gone. Simulations are a big part of probability and statistics: they are used to set up a probabilistic model, test the analytical answer, and help us to study small-sample properties. Although the speed of computations with the `for` loop have improved due to 64-bit computing, vectorized simulations are preferable and many examples use this approach.

Regarding the title of the book, "Advanced Statistics" is not about doing more mathematics, but an advanced understanding of statistical concepts from the perspective of applications. Statistics is an applied science, and this book is about statistics in action. Most theoretical considerations and concepts are either introduced or applied to examples everybody understands, such as mortgage failure, an oil spill in the ocean, gender salary discrimination, the effect of a drug treatment, cancer distribution in New Hampshire, etc.

I again turn the reader's attention to the p-value. This concept falls through the crack of statistical science. I have seen many mathematical statisticians who work in the area of asymptotic expansions and are incapable of explaining the p-value in layman's terms. I have seen many applied statisticians who mostly use existing statistical packages and describe the p-value incorrectly. The goal of this book is to rigorously explain statistical concepts, including the p-value, and illustrate them with concrete examples dependent on the purpose of statistics applications (I suggest an impatient reader jump to Section 7.10 and then Section 8.5). I emphasize the difference between parameter- and individual-based statistical inference. While classical statistics is concerned with parameters, in real-life applications, we are mostly concerned with individual prediction. For example, given a random sample of individual incomes in town, the classical statistics is concerned with estimation of the town mean income (phantom parameter) and the respective confidence interval, but often we are interested in a more practical question. In what range does the income of a randomly asked resident belong with given probability? This distinction is a common theme of the book.

This book is intended for graduate students in statistics, although some sections are accessible for senior undergraduate statistics students with a solid mathematical background in multivariate calculus and linear algebra along with some courses in elementary statistics and probability. I hope that researchers will find this book useful as well to clarify important statistical concepts.

I am indebted to Steve Quigley, former associate publisher at Wiley, for pursuing me to sign the contract on writing a textbook in statistics. Several people read parts of the book and made helpful comments: Senthil Girimurugan; my Dartmouth students James Brofos, Michael Downs, Daniel Kang; and my colleagues, Dan Rockmore, Zhigang Li, James O'Malley and Todd MacKenzie, among others. I am thankful to anonymous reviewers for their thoughts and corrections that improved the book. Finally, I am grateful to John Morris of *Editide* (http://www.editide.us/) for his professional editorial service.

Data sets and `R` codes can be downloaded at my website:

`www.dartmouth.edu/~eugened`

I suggest that they be saved on the hard drive in the directory`C:\StatBook\`. The codes may be freely distributed and modified .

I would like to hear comments, suggestions and opinions from readers. Please e-mail me at `eugened@dartmouth.edu`.

Dartmouth College *Eugene Demidenko*
Hanover, New Hampshire
August 2019

Chapter 1

Discrete random variables

Two types of random variables are distinguished: discrete and continuous. Theoretically, there may be a combination of these two types, but it is rare in practice. This chapter covers discrete distributions and the next chapter will cover continuous distributions.

1.1 Motivating example

In univariate calculus, a variable x takes values on the real line and we write $x \in (-\infty, \infty)$. In probability and statistics, we also deal with variables that take values in $(-\infty, \infty)$. Unlike calculus, we do not know *exactly* what value it takes. Some values are more likely and some values are less likely. These variables are called *random*. The idea that there is uncertainty in what value the variable takes was uncomfortable for mathematicians at the dawn of the theory of probability, and many refused to recognize this theory as a mathematical discipline. To convey information about a random variable, we must specify its distribution and attach a probability or density for each value it takes. This is why the concept of the distribution and the density functions plays a central role in probability theory and statistics. Once the density is specified, calculus turns into the principal tool for treatment.

Throughout the book we use letters in uppercase and lowercase with different meaning: X denotes the random variable and x denotes a value it may take. Thus $X = x$ indicates the event that random variable X takes value x. For example, we may ask what is the chance (probability) that X takes value x. In mathematical terms, $\Pr(X = x)$. For a continuous random variable, we may be interested in the probability that a random variable takes values less or equal to x or takes values from the interval $[x, x + \Delta]$.

A complete coverage of probability theory is beyond the scope of this book –

Advanced Statistics with Applications in R, First Edition. Eugene Demidenko.

rather, we aim to discuss only those features of probability theory that are useful in statistics. Readers interested in a more rigorous and comprehensive account of the theory of probability are referred to classic books by Feller [45] or Ross [83], among many others.

In the following example, we emphasize the difference between calculus, which assumes that the world is deterministic, and probability and statistics, which assume that the world is random. This difference may be striking.

Example 1.1 ***Who said*** π? *The ratio of a circle's circumference to its diameter is* π. *To test this fact, you may measure the circumference of tires and their diameters from different cars and compute the ratios. Does the average ratio approach* π *as the number of measured tires goes to infinity?*

Perhaps to the reader's surprise, even if there is a slight measurement error of the diameter of a tire, the average of empirically calculated π's does not converge to the theoretical value of π; see Examples 3.36 and 6.126. In order to obtain a consistent estimator of π, we have to divide the sum of all circumferences by the sum of all diameters. This method is difficult to justify by standard mathematical reasoning because tires may come from different cars.

This example amplifies the difference between calculus and probability and statistics. The former works in an ideal environment: no measurement error, a unique line goes through two points, etc. However, the world we live in is not perfect: measurements do not produce exactly a theoretically expected result, points do not fall on a straight line, people answer differently to the same question, patients given the same drug recover and some not, etc. All laws of physics including the Newton's free fall formula $S(t) = 0.5gt^2$ (see Example 9.4) do not exactly match the empirical data. To what extent can the mismatch can be ignored? Do measurements confirm the law? Does the Newton's theory hold? These questions cannot be answered without assuming that the measurements made (basically all data) are intrinsically random. That is why statistics is needed every time data are analyzed.

1.2 Bernoulli random variable

The Bernoulli random variable is the simplest random variable with two outcomes, such as *yes* and *no*, but sometimes referred to as *success* and *failure*. Nevertheless, this variable is a building block of all probability theory (this will be explained later when the central limit theorem is introduced).

Generally, we divide discrete random variables into two groups with respect to how we treat the values they take:

- *Cardinal (or numerical).* The variables take numeric values and therefore can be compared (inequality $<$ is meaningful), and the arithmetic is allowed. Examples of cardinal discrete random variables include the number of children in the family and the number of successes in a series of independent Bernoulli experiments (the binomial random variable). If a random variable takes values 0, 1, and 2, then $1 - 0 = 2 - 1$; the arithmetic mean is meaningful for the cardinal random variables.

- *Nominal (or categorical).* These variables take values that are not numeric but merely indicate the name/label or the state (category). For example, if we are talking about three categories, we may use quotes "1," "2," or "3" if names are not provided. An example of a nominal discrete random variable is the preference of a car shopper among car models "Volvo," "Jeep," "VW," etc. Although the probabilities for each category can be specified, the milestone probability concepts such as the mean and the cumulative distribution function make no sense. Typically, we will be dealing with cardinal random variables. Formally, the Bernoulli random variable is nominal, but with only two outcomes, we may safely code *yes* as 1 and *no* as 0. Then the average of Bernoulli outcomes is interpreted as the proportion of having a *yes* outcome. Variables may take finite or infinite number of values. An example of a discrete random variable that may take an infinite number of values is a Poisson random variable, discussed in Section 1.7. Sometimes, it is convenient to assume that a variable takes an infinite number of values even in cases when the number of cases is bounded, such as in the case of the number of children per family.

An example of a binary (or dichotomous) random variable is the answer to a question such as "Do you play tennis?" (it is assumed that there are only two answers, *yes* and *no*). As was noted earlier, without loss of generality, we can encode *yes* as 1 and *no* as 0. If X codes the answer, we cannot predict the answer – that is why X is a random variable. The key property of X is the probability that a randomly asked person plays tennis (clearly, the probability that a randomly asked person does not play tennis is complementary). Mathematically we write $\Pr(X = 1) = p$. The distribution of a binary random variable X is completely specified by p. An immediate application of the probability is that, assuming that a given community consists of N people, we can estimate the number of tennis players as Np.

We refer to this kind of binary variable as a Bernoulli random variable named after the Swiss mathematician Jacob Bernoulli (1654–1705). We often denote $q = 1 - p$ (complementary probability), so that $\Pr(X = 0) = q$. A compact way to write down the Bernoulli probability of possible outcomes is

$$\Pr(X = y) = p^y (1 - p)^{1-y}, \tag{1.1}$$

where y takes fixed values, 1 or 0. This expression is useful for deriving the likelihood function for statistical purposes that will be used later in the statistics part of the book.

The next example applies the Bernoulli random variable to a real-world problem.

Example 1.2 ***Safe driving.*** *Fred is a safe driver: he has a* $1/10$ *chance each year of getting a traffic ticket. Is it true that he will get at least one traffic ticket over 20 years of driving?*

Solution. Many people say yes. Indeed, since the probability for one year is $1/10$, the probability that he will get a traffic ticket over 20 years is more than 1 and some people would conclude that he will definitely get a ticket. First, this naive computation is suspicious: How can a probability be greater than 1? Second, if he is lucky, he may never get a ticket over 20 years because getting a ticket during one year is just a probability, and the event may never occur this year, next year, etc. To find the probability that Fred will get at least one ticket, we use the method of complementary probability and find the probability that Fred gets no ticket over 20 years. Since the probability to get no ticket each year is $1-1/10$ the probability to get no tickets over 20 years is $(1-1/10)^{20}$. Finally, the probability that Fred gets at least one ticket over 20 years is $1-(1-1/10)^{20} = 1-(9/10)^{20} = 0.88$. In other words, the probability to be ticket-free over 20 years is greater than 10%. This is a fun problem and yet reflects an important phenomenon in our life: things may happen at random and scientific experiments may not be reproducible with positive probability.

Problems

1. Check formula (1.1) by examination. [Hint: Evaluate the formula at $y=0$ and $y=1$.]

2. Demonstrate that the naive answer in Example 1.2 can be supported by the approximation formula $1-(1-x)^n \simeq nx$ for small x and $n>1$. (a) Derive this approximation using the L'Hôpital's rule, and (b) apply it to the probability of getting at least one ticket.

3. Provide an argumentation for the infinite monkey theorem: a monkey hitting keys at random on a computer keyboard for an infinite amount of time will almost surely type a given text, such as "Hamlet" by William Shakespeare (make the necessary assumptions). [Hint: The probability of typing the text starting from any hit is the same and positive; then follow Example 1.2.]

1.3 General discrete random variable

Classical probability theory uses cardinal (numeric) variables: these variables take numeric values that can be ordered and manipulated using arithmetic operations such as summation. For a discrete numeric random variable, we must specify the probability for each unique outcome it takes. It is convenient to use a table to specify its distribution as follows.

Value of X	x_1	x_2	x_3	$\cdots$	x_{n-1}	x_n
Probability	p_1	p_2	p_3	$\cdots$	p_{n-1}	p_n

It is assumed that x_i are all different and the n events $\{X = x_i, i = 1, ..., n\}$ are mutually exclusive; sometimes the set $\{x_i\}$ is called the sample space and particular x_i the outcome or elementary event. Without loss of generality, we will always assume that the values are in ascending order, $x_1 < x_2 < \cdots < x_n$. Indeed, if some x are the same, we sum the probabilities. As follows from this table, X may take n values and

$$\Pr(X = x_i) = p_i, \quad i = 1, 2, ..., n,$$

sometimes referred to as the *probability mass function* (pmf). Since p_i are probabilities and $\{x_i\}$ is an exhaustive set of values, we have

$$\sum_{i=1}^{n} p_i = 1, \quad p_i \geq 0.$$

For $n = 2$, a categorical random variable can be interpreted as a Bernoulli random variable. An example of a categorical random variable with a number of outcomes more than two is a voter's choice in an election, assuming that there are three or more candidates. This is not a cardinal random variable: the categories cannot be arranged in a meaningful order and arithmetic operations do not apply.

An example of a discrete random variable that may take any nonnegative integer value, at least hypothetically, is the number of children in a family. Although practically this variable is bounded (for instance, one may say that the number of children is less than 100), it is convenient to assume that the number of children is unlimited. It is customary to prefer convenience over rigor in statistical applications.

Sometimes we want to know the probability that a random variable X takes a value less or equal to x. This leads to the concept of the *cumulative distribution function* (cdf).

Definition 1.3 *The cumulative distribution function is defined as*

$$F_X(x) = \Pr(X \leq x) = \sum_{x_i \leq x} p_i.$$

The cdf is a step-wise increasing function; the steps are at $\{x_i, i = 1, 2, ..., n\}$. The cdf is convenient for finding the probability of an interval event. For example,

$$\Pr(u < X \leq U) = F_X(U) - F_X(u),$$

where u and U are fixed numbers ($u \leq U$). We will discuss computation of the cdf in R for some specific discrete random variables later in this chapter.

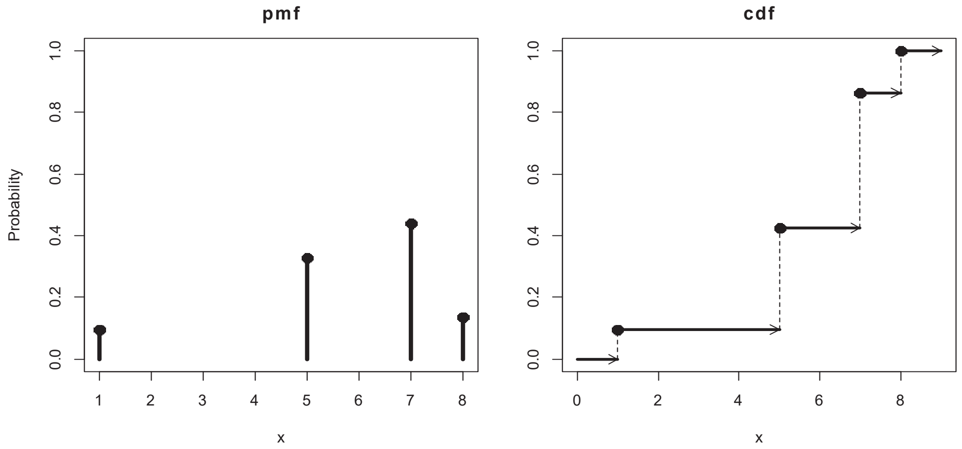


Figure 1.1: *The probability mass function (pmf) and the cumulative distribution function (cdf) of a typical discrete distribution. Note that the cdf is discrete from the left and continuous from the right.*

The pmf and the respective cdf of a typical discrete distribution are shown in Figure 1.1. At each jump the cdf is continuous from the right (indicated by a filled circle) and discrete from the left (indicated by an arrow). This means that $\lim F_X(x) = F(x_i)$ when x approaches x_i from above ($x > x_i$), but $\lim F_X(x) < F(x_i)$ when x approaches x_i from below ($x < x_i$).

Problems

1. (a) Prove that the cdf is a non-decreasing function on $(-\infty, \infty)$. (b) Prove that the cdf approaches 1 when $x \to \infty$ and approaches 0 when $x \to -\infty$.

2. Express p_i in terms of cdf.

3. Express the continuity of a cdf at x_i using notation $\lim_{x\downarrow}$.

1.4 Mean and variance

Expectation or the mean value (mean) is one of the central notions in probability and statistics. Typically, it is difficult to specify the entire distribution of a random variable, but it is informative to know where the center of the distribution lies, the *mean*. The arithmetic average of observations as an estimator of the mean is one of summary statistics characterizing the center of the distribution of a random variable. Another summary statistic, variance, will be discussed in the next section.

We use E to denote the expectation of a random variable. For a discrete random variable X that takes value x_i with probability $p_i = \Pr(X = x_i)$, the mean is defined as

$$E(X) = \sum_{i=1}^{n} x_i \Pr(X = x_i) = \sum_{i=1}^{n} x_i p_i. \tag{1.2}$$

The mean, $E(X)$, can be interpreted as the weighted average of $\{x_i, i = 1, 2, ..., n\}$, where the weights are the probabilities. It is easy to see that for a Bernoulli random variable, the mean equals the probability of occurrence (success), $E(X) = p$. Indeed, as follows from the previous definition of the mean, $E(X) = 1 \times p + 0 \times (1 - p) = p$. This explains why it is convenient to assume that a dichotomous random variable takes the values 0 and 1.

Following standard notation, the Greek letter μ (*mu*) is used for the mean; when several random variables are involved, we may use notation μ_X to indicate that the expectation is of the random variable X.

The mean acts as a linear function: the mean of a linear combination of random variables is the linear combination of the means; in mathematical terms $E(aX) = aE(X)$, $E(X+Y) = E(X)+E(Y)$. The first property is easy to prove; the proof of the second property requires the concept of the bivariate distribution and is deferred to Chapter 3.

1.4.1 Mechanical interpretation of the mean

The mean can be interpreted as the center of mass using the notion of *torque* in physics. Imagine a stick with n masses of weights W_i attached at n locations, $\{x_i,\ i = 1, 2, ..., n\}$. See Figure 1.2 for a geometric illustration. We want to find the support point, μ, where the stick is in balance; thus μ is the center of mass (center of gravity). From physics, we know that a weight W_i located at x_i with respect to μ creates the torque $(x_i - \mu)W_i$. The stick is in balance if the sum of these torques is zero:

$$\sum_{i=1}^{n} (x_i - \mu)W_i = 0. \tag{1.3}$$

This balance condition leads to the solution

$$\mu = \frac{\sum_{i=1}^{n} x_i W_i}{\sum_{i=1}^{n} W_i}. \tag{1.4}$$

For the example depicted in Figure 1.2, the balance point (center of masses) is $\mu = 4.38$,

$$\mu = \frac{1 \times 5 + 2 \times 5 + 3 \times 5 + 4 \times 2 + 5 \times 1 + 6 \times 2 + 7 \times 4 + 8 \times 1 + 9 \times 4}{5 + 5 + 5 + 2 + 1 + 2 + 4 + 1 + 4}.$$

One can interpret solution (1.4) as the weighted average because we may rewrite $\mu = \sum_{i=1}^{n} x_i w_i$, where w_i is the relative weight. As with probability, p_i, $w_i = W_i / \sum_{j=1}^{n} W_j$. In a special case when weights are the same, $W_i = \text{const}$ we arrive at the arithmetic mean, $\overline{x} = \sum_{i=1}^{n} x_i / n$. In summary, the mean is the balance point, or the center of gravity, where probabilities $\{p_i, i = 1, 2, ..., n\}$ act as the relative weights at locations $\{x_i, i = 1, 2, ..., n\}$. A similar interpretation of the mean as a center of the gravity is valid for two and three dimensions.

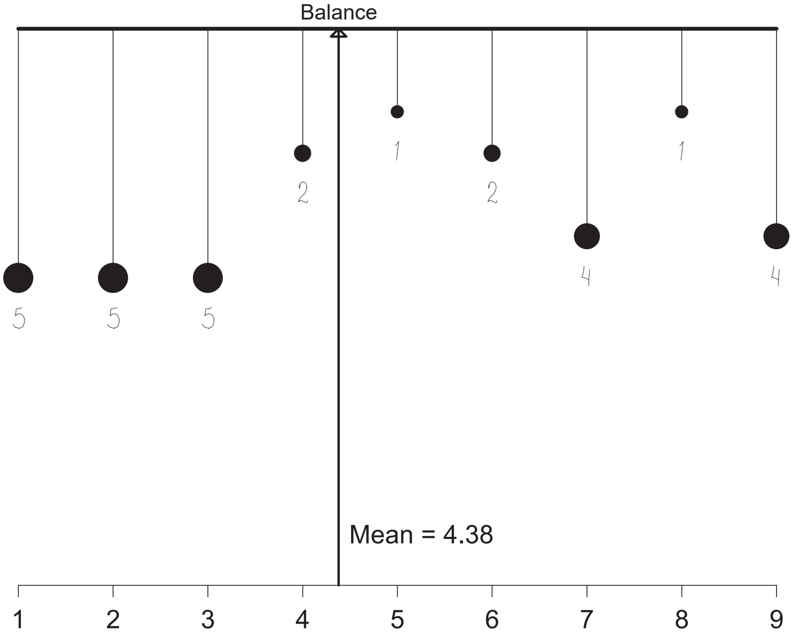


Figure 1.2: *Mechanical interpretation of the mean using hanging weights: the balance is where the resultant torque is zero, $\sum_{i=1}^{n}(x_i - \mu)W_i = 0$. For this example, the support point is $\mu = 4.38$. The mean is the center of gravity.*

The mean is meaningful only for cardinal/numeric random variables, but the mean is not the only way to define the center of the distribution. There are other characteristics of the center, such as mode or median. Sometimes, depending on the subject and the purpose of the study, they may offer a better interpretation than the mean. The mode is defined as the most frequent observation/value: the mode is x_i for which $\Pr(X = x_i) = \max$. Hence one can refer to the mode as the most probable value of the random variable. The mode is applied for categorical random variables where the mean does not make sense, such as when reporting results of the poll among presidential candidates. In another example, we know that the most frequent number of children in American families (mode) is 2 and the average (mean) is 2.2. In other words, 2 is the most probable number of children in the family. If you invite a family with children, you expect to see two kids, not 2.2.

The median is defined as the value x for which the probability of observing $X < x$ is equal to the probability of observing $X > x$. The median does not apply to categorical random variables because it requires ordering.

Sometimes, the weighted mean emerges naturally as a conditional probability, the following is an example.

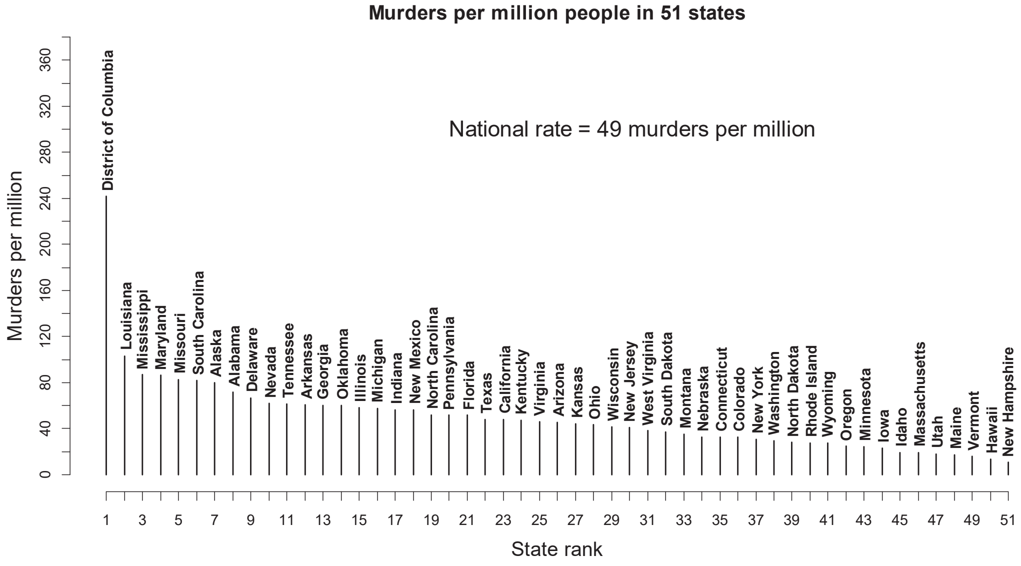


Figure 1.3: *Murder rates in the United States by state. The national murder rate is the weighted mean with the ith weight equal to the number of people living in state i.*

Example 1.4 ***National murder rate.*** *Figure 1.3 shows the number of murders per million people in the District of Columbia and all 50 states of the United States. (a) Express the national murder rate as the weighted mean. (b) Interpret the national murder rate as a conditional probability.*

Solution. (a) By definition, the murder rate, r_i is the number of murders divided by the number of people living in the state (state population), $i = 1, 2, ..., 51$. We associate r_i with the probability that a random person from state i will be murdered. If n_i denotes the state population (in millions), the number of murders is $r_i n_i$ and the total number of murders in the country is $\sum_{i=1}^{51} r_i n_i$. To compute the national murder rate, we need to divide the total number of murders by the total number of people in the United States,

$$r = \frac{\sum_{i=1}^{51} r_i n_i}{\sum_{i=1}^{51} n_i} = \sum_{i=1}^{51} r_i w_i$$

where $w_i = n_i / \sum_{i=1}^{51} n_i$ is the proportion of people living in state i. Using this formula and data presented in Figure 1.3, the national murder rate is $r = 49$

murders per million. Looking back into formula (1.2), one can interpret w_i as probabilities. Therefore one can interpret r as the probability of murdering of a random person in the United States. Note that it would be incorrect to compute the national murder rate as the simple average, $\sum_{i=1}^{51} r_i/51$. If less-populated states have high murder rates their contribution would be the same as large states. (b) Now we formulate the problem as a conditional probability. Define

$$\Pr(B|A_i) = \Pr(\text{random person will be murdered} \mid \text{person lives in state } i) = r_i$$
$$\Pr(A_i) = \Pr(\text{random person from US lives in state } i) = w_i.$$

Using the law of total probability $\Pr(B) = \sum_{i=1}^{n} \Pr(B|A_i)\Pr(A_i)$, where A_i do not overlap and $\sum_{i=1}^{n} \Pr(A_i) = 1$, the probability of being murdered in US is $r = \sum_{i=1}^{51} r_i w_i$, as before. □

The following two examples illustrate that sometimes using the mean as a measure of center of the distribution makes perfect sense, but sometimes it does not.

Example 1.5 ***Town clerk mean.*** *The arithmetic average of house prices is a suitable average characteristic for a town clerk who is concerned with the total amount to collect from the residents.*

Solution. Suppose there are n houses in town with prices $\{P_i, i = 1, 2, ..., n\}$. When reporting the average house price, town officials prefer to use the arithmetic average:

$$\overline{P} = \frac{1}{n}\sum_{i=1}^{n} P_i.$$

Indeed, if r is the property tax rate, the town collects $r\sum_{i=1}^{n} P_i$ dollars and $r\overline{P}$ is the average property tax in the town.

Example 1.6 ***Buyer's house median.*** *A real estate agent shows houses to a potential buyer. What is a suitable average house price for the buyer, the mean or the median?*

Solution. While the mean price makes sense for a town or state official (the property tax is proportional to the mean), it is not useful for the buyer who is thinking of the chance of affording a house he/she likes. Instead, the median means that 50% of the houses he/she saw will have price lower than the median and 50% of the houses will have higher price. In this case, the median has a much more sensible interpretation from the buyer's perspective. □

Another situation where the mean and median (or mode) depends on the subject of application is the salary distribution in a company. For the company's

CEO, the mean, which is the ratio of the personnel cost to the number of employees, is the most meaningful quantity because it directly affects the profit = revenue minus cost (including cost of labor). For an employee, the median (or maybe the mode) is the most informative parameter of the company salary distribution because he or she can assess if he/she is underpaid. In general, mean has a meaningful interpretation if and only if the sum of observations or measurements has an interpretation. This is the case for both examples: the total wealth of properties and the total labor cost are what the town and the CEO are concerned with, respectively. □

In the previous discussion, we compared mean with median. The following example underscores the difference between median and mode.

Example 1.7 ***Mode for the manager of a shoe store and median for a shoe buyer***. *Explain why mode is a more appropriate characteristic of the center of the shoe-size distribution for the manager of a shoe store but median is more appropriate for a shoe buyer.*

Solution. The manager is concerned with the most popular shoe size because it tells him/her about the order to make from a shoe factory. The buyer wants to know if the store has the sufficient stock of the popular shoe size. □

In conclusion, we should not stick with the mean as the most popular and the easiest parameter to characterize the center of the distribution. We must also consider the median or the mode, depending on the application.

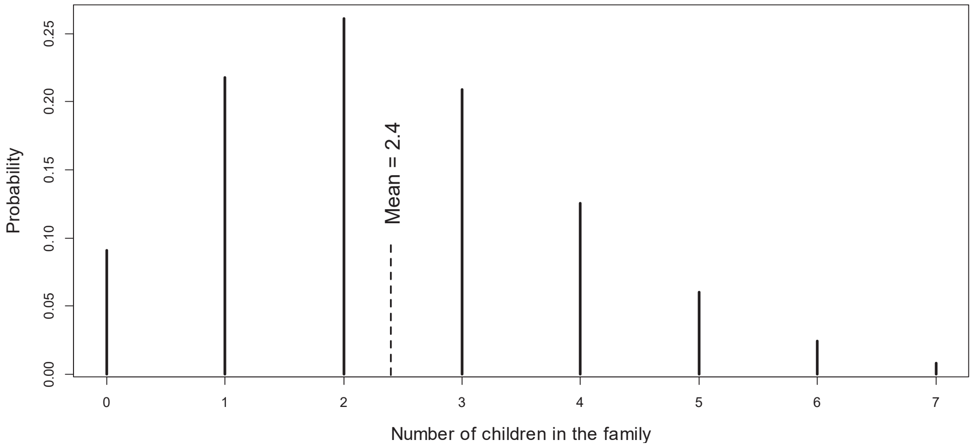


Figure 1.4: *The distribution of the number of children in 100 families. According to the mean you are expected to see* 2.4 *children.*

Example 1.8 ***Number of children in the family***. *You are invited for dinner to a family and you want to bring presents to each child (you do not know the* *number of children). To make an educated guess on how many presents to buy you find on Internet data on the number of children in 100 families, see Figure 1.4. How many presents do you buy?*

Solution. The typical number of children in the family is 2, the mode. According to the mean you expect to see 2.4 children, a somewhat uncomfortable number. See Example 1.17 where the number of toys is solved assuming that the number of children in the family follows a Poisson distribution. □

The following example illustrates that some random variables do not have finite mean.

Example 1.9 ***St. Petersburg paradox.*** *The game involves a single casino player and consists of a series of coin tosses. The pot starts at $1 and casino doubles the pot every time a head appears. When a tail appears, the game ends and the player wins whatever is in the pot. What would be a fair price to pay the casino for entering the game?*

Solution. We define X as the dollar amount paid to the player. With probability 1/2, the player wins one dollar; with probability 1/4 the player wins \$2, with probability 1/8 the player wins \$4, and so on. Thus, $\Pr(X = 2^{k-1}) = 1/2^k$ for the number of tosses, $k = 1, 2, ...$ We calculate the expected value to determine a fair price for a player to pay as

$$E(X) = \frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{8} \times 4 + \cdots = \sum_{k=1}^{\infty} \frac{1}{2^k} 2^{k-1} = \sum_{k=1}^{\infty} \frac{1}{2} = \infty.$$

The player should pay $\$\infty$ to make the game fair to the casino.

1.4.2 Variance

Another milestone concept of probability and statistics is variance. Variance is the expected value of the squared distance from the mean, or symbolically,

$$\mathrm{var}(X) = E(X - \mu)^2.$$

Usually we use the Greek letter σ^2 (*sigma-squared*) to denote the variance. We write $\sigma^2 = \sigma_X^2 = \mathrm{var}$. Variance reflects the spread of the random variable around the mean: the larger the spread/scatter, the larger the variance. The same caution should be used for the variance as for the mean because the variance is the mean of squared distances. Although convenient from the computational standpoint, it may not appropriate for a particular application.

If X is a discrete random variable, $(X - \mu)^2$ can be viewed as another discrete random variable, so its expectation can be computed as

$$\sigma^2 = \sum_{i=1}^{n} (x_i - \mu)^2 p_i. \tag{1.5}$$

In the following theorem we provide an alternative computation of the variance.

Theorem 1.10 *The following formula holds:*

$$\sigma^2 = E(X^2) - \mu^2. \tag{1.6}$$

Proof. Expanding $(x_i - \mu)^2$ in (1.5), we obtain

$$\sum_{i=1}^{n}(x_i - \mu)^2 p_i = \sum_{i=1}^{n}(x_i^2 - 2\mu x_i + \mu^2)p_i = \sum_{i=1}^{n} x_i^2 p_i - 2\mu \sum_{i=1}^{n} x_i p_i + \mu^2 \sum_{i=1}^{n} p_i.$$

But $\sum_{i=1}^{n} x_i^2 p_i = E(X^2)$ and $\sum_{i=1}^{n} x_i p_i = \mu$. Because $\sum_{i=1}^{n} p_i = 1$, we finally obtain $\sigma^2 = E(X^2) - 2\mu \times \mu + \mu^2 = E(X^2) - \mu^2$. □

Sometimes $E(X^2)$ is called the *noncentral* second moment and σ^2 is called the *central* second moment. These moments are connected as shown in formula (1.6). We shall see later that formula (1.6) holds for continuous random variables as well.

To illustrate formula (1.6), we derive the variance of the Bernoulli random variable, we first find $E(X^2) = 0 \times \Pr(X = 0) + 1 \times \Pr(X = 1) = p$. Then, using the fact that $E(X) = p$, the variance of the Bernoulli variable is $\mathrm{var}(X) = p(1-p)$.

Using formula (1.6) we obtain an explicit expression for the variance, as an alternative formula to (1.5),

$$\sigma^2 = \sum_{i=1}^{n} x_i^2 p_i - \left(\sum_{i=1}^{n} x_i p_i\right)^2. \tag{1.7}$$

Variance is always nonnegative and equals to zero if and only if X takes one value, the mean value.

It is easy to prove that $\mathrm{var}(a + bX) = b^2 \mathrm{var}(X)$ where a and b are numbers. In Chapter 3 we prove that if X and Y are independent random variables, then $\mathrm{var}(X + Y) = \mathrm{var}(X) + \mathrm{var}(Y)$.

Standard deviation (SD) is the square root of variance, $\sigma = \sqrt{\mathrm{var}}$. SD also reflects how random variable deviates from the mean. In contrast with the variance, SD does so on the original scale, compared with the variance, which is more convenient for interpretation. For example, if X is measured in feet, variance is measured in square feet but SD is measured in feet as well. For this reason, SD is often reported in applications to specify the scatter of the variable.

The mean and the variance of a discrete random variable are computed by the same formulas (1.2) and (1.5) when the number of outcomes is infinite, $n = \infty$.

The probability distribution completely specifies X. Mean and SD are integral features – where the random variable values concentrate and how wide the spread is.

The expected value is often used in finance to quantify the expected return, and variance is used to quantify the risk (volatility); the following is a typical example. Here we use the fact that the variance of the sum of two independent random variables equals the sum of two variances.

Example 1.11 ***Grant problem.*** *A person applies for a large grant in amount \$100K with the probability of getting funded* 1/5. *There is an alternative: to apply for two small grants in amount of \$50K each with the same probability of funding (it is assumed that the probabilities of funding are independent). What strategy is better?*

Solution. We define the expected return (funding) as the sum of possible amounts weighted with respect to their odds/probabilities. Let X define the binary random variable that takes value 100 with probability 1/5 and 0 with probability 4/5. Then the expected return in the first strategy is $100/5+0\times(4/5)=20$. Similarly, the second strategy leads to the expected return $50/5+50/5=20$. Thus in terms of the expected funding, the two strategies are equivalent. Now we look at these options from the risk perspective (this is typical for finance calculations). Clearly, between two strategies with the same expected return, we choose the strategy with a smaller risk/variance. Using formula (1.5) the variance of the large grant is $(0-20)^2\times(4/5)+(100-20)^2\times(1/5)=1600$. Since getting funded from two small grants is independent, the total variance is twice the variance from each grant, $2\times\left[(0-10)^2\times(4/5)+(50-10)^2\times(1/5)\right]=800$. We conclude that, although the two options are equivalent in terms of the expected return, the second option is less risky. This fact, known as *diversification*, is the pillar of financial decision making and investment risk management analysis. We will return to the problem of diversification when considering the optimal portfolio selection in Section 3.8.

Problems

1. Introduce a function $M(a)=E(X-a)^2$. (a) Express this function through var and μ. (b) Prove that this function takes the minimal value at $a=\mu$. (c) Find minimum of M.

2. Prove that the mean is a linear function of scale, $E(aX)=aE(X)$, where a is a constant. Prove that variance is a quadratic function of scale, $\text{var}(aX)=a^2\text{var}(X)$.

3. Plot the cdf of a Bernoulli random variable.

4. Prove that $F_{aX}(x)=F_X(x/a)$ for a positive a.

5. How would you report the average grade in class: mean, mode, or median? Justify using the concept of the students' standing in the class.

6. (a) What would state officials report in the summary statement: mean, median, or mode of income? (b) If state officials want to attract new business to the area, what would they appeal to: mean, mode, or median of company revenue? Explain.

7. Suppose that in Example 1.9 the pot increases by $q\%$. What would be the fair price to enter the game?

8. Generalize the grant problem to an arbitrary number of grants with equal probability.

1.5 R basics

The R programming language will be used throughout the text, so we introduce it at the very beginning.

R is a public domain statistical package. It is freely available for many computational platforms, such as Windows, Macintosh and Unix at

$$\mathtt{http://www.r-project.org/}$$

The goal of this section is to give the reader the very basics of R programming. More detail will be given throughout the book. A comprehensive yet succinct description of R is found at

`https://cran.r-project.org/doc/manuals/R-intro.pdf`

There are two ways to use R: command line and script (function). We use the command line when a single-formula computation is needed, like a scientific calculator. When computations involve several steps, we combine them in a script. A distinctive feature of R is the assignment operator `<-`. We however prefer the usual `=` symbol.

For example, to compute 2×2, we write after $>$ (the command prompt) `2*2`and press `<Enter>`. We can store the result using an identification (name of a variable), say`four<-2*2`. This means that computer computes 2×2 and stores the result in variable named `four`. Other operations are `+, -, /, ^`. Many standard mathematical functions are available: `log, log10, exp, sin, cos, tan, atan`. Vectors and matrices are easy to handle in R. For example, to create a ten-dimensional vector of ones, we issue `one10<-rep(1,10)`. Then `one10` is a vector with each component 1; to see this we type `one10` and press `<Enter>`. Function `rep` is a special function of R, which is short of *repetition*. It has two arguments and in general form is written as `rep(what,size)`. If you do not want to assign special values to components of a vector, use `NA`, which means that component values are *not available*.

Similarly, one can create a matrix. For example, `mat.pi.20.4<-matrix(pi, nrow=20,ncol=4)` will create a 20×4 matrix with the name `mat.pi.20.4` with all entries π (`.` may be a part of a name). Again, to see the results, you type `mat.pi.20.4` and press `<Enter>`. It is easy to see the history of the commands you issued using arrow keys ↑ or ↓ . Thus, instead of retyping `mat.pi.20.4`, you pick the previous command and remove the unwanted parts.

R distinguishes lowercase and uppercase, thus mat.pi.20.4 and mat.Pi.20.4 mean different things.

If a and b are two one-dimensional arrays (vectors) of the same size, a*b computes the vector whose values are the component-wise products. The same rule holds for matrices of the same dimensions. In contrast to the component-wise multiplication, the symbol %*% is used for vector/matrix multiplication. For example, if $\mathbf{A}$ is a $n \times m$ matrix and $\mathbf{a}$ is an m-dimensional vector (the boldface is used throughout the book to indicate vectors and matrices), then A%*%a computes a vector $\mathbf{Aa}$ of dimension n. If $\mathbf{B}$ is a $m \times k$ matrix, then A%*%B computes an $n \times k$ matrix. Function t is used for vector/matrix transposition. For example, if $\mathbf{A}$ is a $n \times m$ matrix, t(A) is $\mathbf{A}'$. If $\mathbf{a}$ is a vector, then t(a)%*%a, sum(a^2), and sum(a*a) all give the same result. There are a few rules when adding or multiplying matrices and vectors that do not comply with mathematics but are convenient from a computational standpoint. For example, if $\mathbf{A}$ is a $n \times m$ matrix and $\mathbf{a}$ is an m-dimensional vector, then A+a is acceptable in R and computes a matrix with columns that are columns of matrix A plus a. If $\mathbf{a}$ is an m-dimensional vector, then A+a computes a matrix with rows that are rows of A plus a. When A is a square matrix operation is on columns. The same rule works for multiplication (or division). For example, if $\mathbf{A}$ is a $n \times m$ matrix and $\mathbf{a}$ is an n-dimensional vector, then A*a produces a $n \times m$ matrix with the ith row as the product of the ith row of matrix $\mathbf{A}$ times a_i. It is important that a is a vector, not the result of A%*%b that produces a $n \times 1$ matrix. To make A%*%b a vector, use as.vector(A%*%b).

1.5.1 Scripts/functions

Typically statistical computations involve many lines of code you want to keep and edit. In this case, you want to write user-defined scripts (functions). Functions have arguments and a body. For example, if you want to create a function by the name my1, you would type my1<-function(){}. This means that so far this function has no arguments in ( ) and no body in { }. To add this, we type my1<-edit(my1) and a text editor window appears where we can add operators. It is good style to use comments in your program. In R everything after # is a comment up to the next carriage return.

For example, let us say we want to write a program that multiplies two user-defined numbers. We start with program creation twoprod<-function(){} and then add the text using command twoprod<-edit(twoprod).

Here is a version of the R program:

```
twoprod<-function(t1,t2)
{
    prod<-t1*t2
    return(prod)
}
```

Here t1 and t2 are the arguments. The user can use any numbers. For example, if one enters on the command line twoprod(0.3,24), the answer should be 7.2. Another, more explicit way to run the function is to use twoprod(t1=0.3,t2 =24). A convenient feature of R is that one can define default values for the arguments. For example, one may use t1=10 as default. Then the function should be

```
twoprod<-function(t1,t2=10)
{
    prod<-t1*t2
    return(prod)
}
```

To compute 2.4 × 10 it suffices to run twoprod(t1=2.4), but this does not mean that t2=10 always. You still can use any t2 even though the default is specified. Each function should return something, in this case the product of numbers. To run a function, it has to have ( ) even if no arguments are specified; otherwise, it prints out the text of the function.

1.5.2 Text editing in R

You have to use a text editor to edit functions (R codes). The default text editor in R is primitive – it does not even number the lines of the code. For example, when R detects errors in your code you may see a message like this:

```
Error in .External2(C_edit, name, file, title, editor) :
 unexpected ',' occurred on line 521
 use a command like
 x <- edit()
 to recover
```

This means that you have to count 521 lines yourself! There are plenty of public domain Windows text editors used by programmers worldwide, such as notepad++ or sublime. For example, to make notepad++ your R editor, you have to add the line (for Windows PC)

options(editor="c:\\Program Files (x86)\\Notepad++\\notepad++.exe")

to the file Rprofile.site, which, for example, is located in the folder

C:\Program Files\R\R-3.2.2\etc\.

Alternatively, you may issue this command in the R console every time you start a new session. The notepad++ software can be downloaded from the site https://notepad-plus-plus.org/. Here we assume that the program notepad++.exe is saved in the folder c:\\Program Files (x86)\\Notepad++ and you use R-3.2.2 version of R; you have to make slight modifications otherwise. The sublime editor can be downloaded from https://www.sublimetext.com/3.

1.5.3 Saving your `R` code

We strongly recommend saving your code/function/script as a text file every time you run it. For example, assuming that you have a Windows computer and you want to save your function `myfun` in the existing folder/directory, `C:\StatBook`, your first statement in this function may look like this `dump("myfun","C:\\StatBook\\myfun.r")`. Note that double backlashes are used. If you want to read this function, issue `source("C:\\StatBook\\myfun.r")` in the `R` console. Saving as a text file has a double purpose: (i) You can always restore the function in the case you forgot to save the `R` session. (ii) You can keep the `R` code in the same folder where other documents for your project are kept. Throughout the book, all `R` codes are saved in the folder `C:\\StatBook\\`.

If you work on a Mac computer, the syntax is slightly different: single forward slashes are used, and there is no reference to the hard drive letter. For example, to save `myfun`, you use `dump("myfun","/Users/myname/StatBook/myfun.r")` with the `source` command being modified accordingly.

1.5.4 for loop

Repeated computation expressed in a loop is the most important method in computer programming. The simplest example of the loop is

```
for(i in 1:10)
{
    #body of the loop
}
```

which repeats operators between `{` and `}` ten times by letting i=1, i=2, ..., i=10. Here `1:10` creates a sequence of numbers from 1 to 10. There is a more general way to create sequences `seq(from=,to=,by=)` or `seq(from=,to=,length=)`. For example, `s1<-seq(from=1,to=10,by=1)` and `s2<-seq(from=1,to=10,length=10)` both produce the same sequence from 1 to 10. Let us write a program that computes the sum of all even numbers from 2 to n, where n is user-defined. First, we specify the function and edit it: `twonsum<-function(){}; twonsum <-edit(twonsum)`. The code may look like this:

```
twonsum<-function(n)
{
    sn<-0 # initialization
    nseq<-seq(from=2,to=n,by=2)
    for(n in nseq)
        {
            sn<-sn+n # summation
        }
    return(sn)
}
```

If there is only one operator in the loop body, the braces are not required, so a shorter version is

```
twonsum<-function(n)
{
    sn<-0 # initialization
    nseq<-seq(from=2,to=n,by=2)
    for(n in nseq)
        sn<-sn+n # summation
    return(sn)
}
```

This function can be significantly shortened because there is a built-in summation function `sum(x)`, where `x` is a an array. Similarly, `prod(x)` computes the product. Thus, `twonsum` function can be shortened as `sum(seq(from=2,to=n,by=2))`.

1.5.5 Vectorized computations

Many, but not all, loop operations can be vectorized. Then, instead of loops we use operations with vectors. The vectorized versions usually are more compact but sometimes require matrix algebra skills. More importantly, the vectorized approach is more efficient in `R` – loops are slower, but they may need less RAM. Vectorized operations are faster because they are written in `C` or `FORTRAN`. Using the C language, vectorized computations pass the array pointer to C, but the loop communicates with C at every single iteration. Some vectorized operations, like `romSums`, `colMeans`, etc., are already built in.

Compute the mean, variance, and SD of a discrete random variable specified by n-dimensional vectors of values and probabilities.

```
mvsd<-function(x,p)
{
    # x is the vector of values
    # p is the vector of probabilities
    n<-length(x) # recover size
    mux<-sum(x*p) # mean
    ex2<-sum(x^2*p) #E(X^2)
    s2x<-ex2-mux^2 # alternative variance
    sdx<-sqrt(s2x) # SD
    return(c(mux,s2x,sdx)) # return the triple
}
```

For example, if we run `mvsd(x=c(0,1),p=c(.25,.75))` it will give us `0.75000 0.1875000 0.4330127`. Indeed, since we specified a Bernoulli distribution, we should have $E(X) = 0.75$, $\text{var}(X) = 0.25 \times 0.75 = 0.1875$, $\text{SD} = \sqrt{\text{var}(X)} = 0.433$. `R` has `mean`, `var` and `sd` as built-in functions with a vector as the argument. For example, if `a<-c(0.1,0.4,0.5)` then `mean(a)` and `sd(a)` return `0.33333` and `0.2081666`, respectively. Summation, subtraction, and multiplication of vectors with the same length are component-wise and produce a vector of the same length.

To illustrate the built-in vectorized function consider the problem of computing the SD for each row of a $n \times m$ matrix X, where n is big, say, $n = 500000$ as in genetic applications. Of course, we could do a loop over $500,000$ rows and even use the `mvsd` function. The following is a much faster version.

```
SDbig=function(X)
{
    mrow=rowMeans(X) #averaging for each row across columns
    mrow2=rowMeans((X-mrow)^2)
    SDs=sqrt(mrow2)
    return(SDs)
}
```

We make two comments: (i) `rowMeans` returns a vector with length equal to the number of rows in matrix `X`, and each component of this vector is the mean in the row across columns. (ii) Although `X` is a matrix and `mrow` is a vector, `R` does not complain and subtracts `mrow` from each row when we compute `X-mrow`.

Five vectorized computations for double integral approximation are compared in the following example.

Example 1.12 ***Comparison of five vectorized computations.*** *Use vectorized computations for numerical approximation of the double integral*

$$A = \int\int_{x^2+y^2<1} \mathrm{e}^{-(3x^2-4xy+2y^2)} dxdy.$$

Solution. The exact integral value can be obtained by rewriting the integral in a form suitable for symbolic algebra software, such as *Maple* or *Mathematica*:

$$A = \int_{-1}^{1} \left(\int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \mathrm{e}^{-(3x^2-4xy+2y^2)} dy \right) dx = 1.374.$$

To approximate the integral, we replace the integral with the sum of integrand values over the grid for x and y. Let the grid for x and y be an array from -1 to 1 of `length=N` (the rectangular grid must contain the integration domain). First, compute the $N \times N$ matrix $\mathbf{M}$ of values $3x^2 - 4xy + 2y^2$, and second, compute the sum of values $\mathrm{e}^{-(3x^2-4xy+2y^2)}$ multiplied by the step in each grid for which $x^2+y^2 < 1$. The `R` code is found in the file `vecomp.r` and the time of computation by each algorithm is shown in the following table.

Method	Double loop	Single loop	Matrix algebra	`rep`	`expand.grid`	`outer`
`job`	`job=1`	`job=2`	`job=3`	`job=4`	`job=5`	`job=6`
Time (s)	19	4	3	6	7	5

The first method (`job=1`) uses the brute-force double loop matrix computation and fills in the matrix `M` in an element-wise fashion, instead of a vectorized algorithm. The second algorithm (`job=2`) is a vectorized algorithm using a single loop over the x grid and filling in the matrix `M` row by row. The vectorized algorithm in `job=3` uses matrix algebra to compute matrix `M` as $\mathbf{M}=3\mathbf{x}^2\mathbf{1}' - 4\mathbf{x}\mathbf{y}' + 21\mathbf{y}^{2\prime}$, where $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{1}$ are $N \times 1$ vectors ($\mathbf{1}$ is a vector of ones). The fourth algorithm (`job=4`) computes `M` on the grid of values (x_i, y_j) for $i, j = 1, ..., N$ using the `rep` function with two options: (i) to repeat vector `x` with the option `times=N` and (ii) to repeat each element of vector `x` with the option `each=N`. The fifth method (`job=5`) is similar to the previous one, but instead of the `rep` function, another `R` function, `expand.grid`, is used (they produce the same grids for two dimensions). This function is especially convenient with multiple dimension grids, say, for three-dimensional integration. Finally, the sixth method (`job=6`) uses a built-in function `outer`. This function returns a $N \times N$ matrix of values $f(x_i, y_j)$ and is especially convenient for our purposes. Function f should be specified in `R` before computing. The time of computation in seconds is assessed using the `date()` command. A more precise way to find the time of computation is by calling the `Sys.time()` function before and after the operation and taking the difference. Not surprisingly, double loop takes a long time. Although the third method is the fastest, it may be difficult to generalize to nonlinear functions and integrals of higher dimension. The `R` function that implements the five methods can be accessed by issuing `source("c:\\StatBook\\vecomp.r")`. A few remarks on the script: (i) It is easy to lose the `R` code; it is a good idea to save the code as a text file in a safe place on the hard disk. Command `dump("vecomp","c:\\StatBook\\vecomp.r")`saves the `R` code/object in the folder `c:\\StatBook` under the name `vecomp.r` with the full name `c:\\StatBook\\vecomp.r`. (ii) In `job=6` the command `sum(M[x2y2<1])` computes the sum of elements of matrix `M` for which condition `x2y2<1` holds. An advantage of these integral approximations is that any domain of integration may be used if expressed as an inequality condition. The condition should be a matrix with the same dimension as matrix `M`. The double integral A is approximated in Example 3.60 via simulations.

apply

Besides `rowMeans` and the like, such as `colMeans`, `colSums`, `rowSums`, or `pmax`, `R` has a built-in capability to do vectorized computations in general way. Here we illustrate this feature by the `apply` function. This function has three arguments: `X`, `MARGIN`, and `FUN`. The first argument, `X`, specifies a matrix. The second argument, `MARGIN`, specifies row or column, and the third argument, FUN, specifies the function to be performed in the vectorized fashion. For example, instead of `SDbig`, one can use `apply(X,1,sd)`. If the second argument is `2`, it returns SD of the columns. Of course, one could use other functions for `FUN`, such as `sum`,

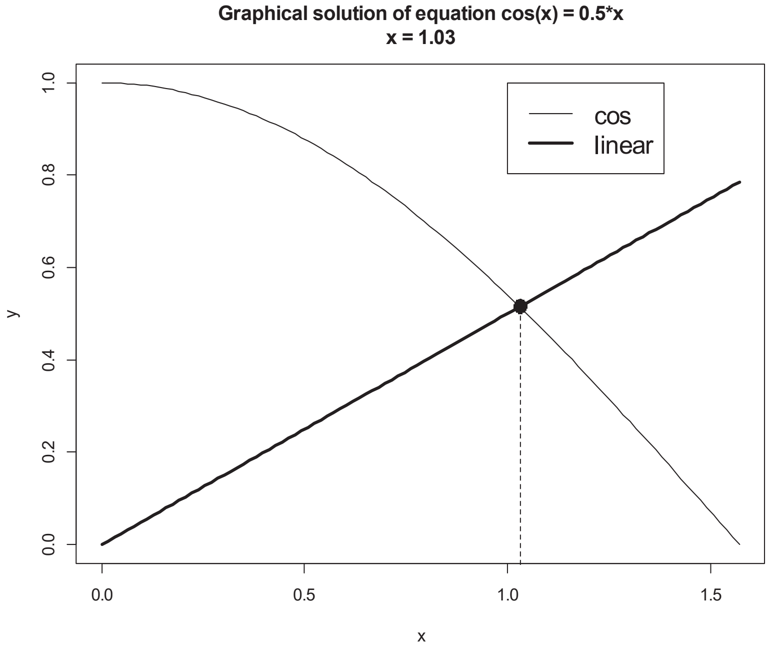


Figure 1.5: *This plot is the result of function* `my1sr()`*. This function returns* `1.031`*, an approximate solution to the equation* $\cos(x) = x/2$*.*

`max`, etc. Note that this is the easiest example of `apply`; there are others, more sophisticated examples of this function as well as other vectorized functions, such as `lapply` and `tapply`. The function specified in `FUN` may be any function of the row data and may return a vector. See Example 6.88 where `apply` is used for vectorized simulations.

Example 1.13 ***Apply command.*** *Fill a 100 by 12 matrix using numbers from 1 to 1200 and compute (a) the normalized matrix (subtract the mean and divide by SD in each row) using* `apply`*, and (b) two numbers in each row such that 25% and 75% of the data is less than those numbers (the first and third quartiles).*

Solution. The matrix is filled with integers from 1 to 1200 using command `X=matrix(1:1200, ncol=12)`. (a) To normalize the matrix, we write a function with a vector argument that may be thought of as a typical row of the matrix: `normal=function(x) (x-mean(x))/sd(x)`. Now the normalized matrix is computed as `apply(X,1,FUN=normal)`. (b) First, the function orders the array of numbers in each row and then picks the $(3n/4)$th element: `tdqurt=function(x) {xo=x[order(x)]; n=length(xo);c(xo[n/4],xo[3*n/4])}`. Note that the call `apply(X,1, FUN=tdqurt)` returns a matrix 2 by `100`.

1.5.6 Graphics

Versatile graphics is one of the most attractive features of R. A picture is worth a thousand words and a good graph may likely become the endpoint of your statistical analysis. Let us start with an example of a graphical solution of a transcendent equation, $\cos x = \alpha x$, where α is a positive user-defined parameter. Specifically, we want to (i) plot two functions, $y_1 = \cos x$ and $y_2 = \alpha x$, on an interval where they intersect and (ii) find the point of intersection. Since y_1 is a decreasing function and y_2 is an increasing function on $(0, \pi/2)$, we plot these functions on this interval. In this function, `alpha` is the used-defined parameter with default value `3`. The graphical output of the `R` function `my1sr` is shown in Figure 1.5. This function can be accessed by issuing `source("c:\\StatBook\\my1sr.r")`. The command `indmin<-which.min(abs(y1-y2))` returns the index for which the absolute difference between function values is minimum.

We make several comments:

1. In the `plot` function, the first argument is an array of x-values, and the second argument is an array of the corresponding y-values; the arrays/vectors should have the same length. The second argument of the function `type` specifies the type of plot. For example, `type="l"` will produce lines and `type="p"` will produce points; `xlab` and `ylab` are the axis labels; `main` specifies the title of the plot.

2. We add the line to the plot using function `lines`. As with the `plot` function, the first and second arguments are the x- and y- values; `lwd` specifies the line width (regular width is 1). The line style can be specified as well (see below), with default `lty=1`.

3. A legend is a must in almost all plots to explain what is plotted. The first argument is the x-coordinate and the second argument is the y-coordinate of the upper-left corner of the legend rectangle; the third argument is the message in the legend, which, in our function, consists of two words; `lty` specifies the line style: `1` is solid, `2` is dotted, and `3` is dashed.

4. In this code, `which.min` returns the index with the minimal value. In our case, this built-in function returns the index where two lines are close to each other.

5. Parameters `cex` and `pch` control the type of the point and its size. The default values are `cex=1`(small) and `pch=0` (empty circle); `cex=2` produces a circle of the size twice large as the default.

6. Title combines text and numbers using the `paste` command. Text and characters need quotes; `\n` forces text to the next line.

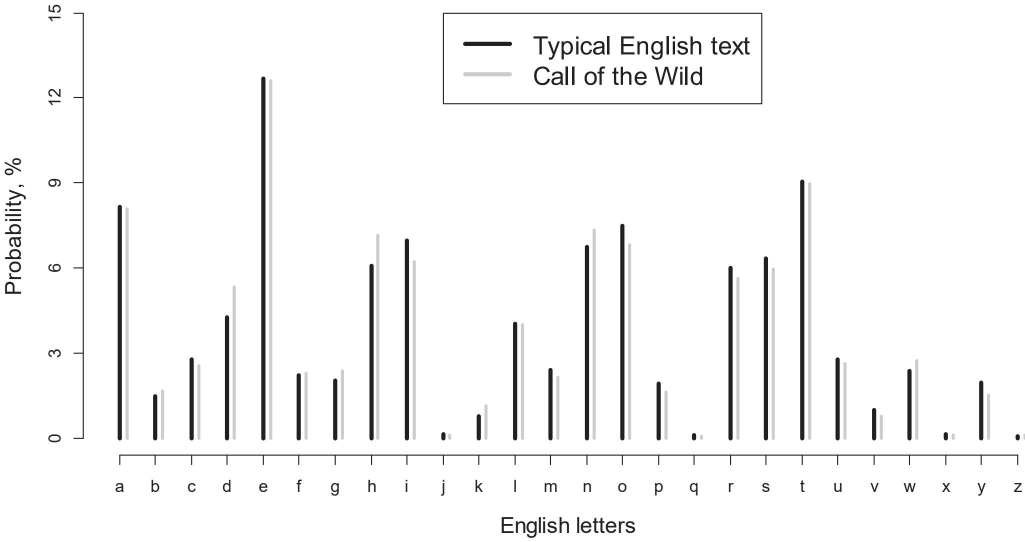


Figure 1.6: *The probability plot of English letters in a typical English text and "Call of the Wild" by Jack London. The distribution of letters in this novel will be further studied in Example 7.66.*

In the following example, we illustrate how two discrete distributions can be visually compared in `R`.

Example 1.14 ***Frequency of English letters in a Jack London novel.*** *Write an* `R` *function that plots the frequency of English letters in a typical text and the frequency of letters in "The Call of the Wild" by Jack London side by side. Do the frequencies look similar?*

Solution. The text file `Jack_London_Call_of_the_Wild_The_f1.char` contains *Call of the Wild* character by character. The frequency of English letters in a typical text, as the average over a large number of English texts, can be found on the Web. The `R` code is found in the file `frlJLET.r`, which creates Figure 1.6. We make a few comments: (i) It is useful to save the `R` code in a text file using the `dump` command every time we run the function. It is easy to restore the function later by issuing `source("c:\\StatBook\\frlJLET.r")`. (ii) The char file is uploaded using the built-in function `scan`; option `what=""` means that letters are characters. (iii) The operator `ch[ch==freqlet[i,1]]` returns a part of array `ch` for which the condition `ch==freqlet[i,1]` holds, or in other words, it returns characters specified by `freqlet[i,1]`.

We conclude that, on average, the frequencies of the letters in a typical English text match those of a famous Jack London novel; however, for some letters such as "d" and "h," the discrepancy is noticeable. Later we shall test that the frequencies of letters in the novel are the same as in a typical English text using the Pearson chi-square test. The entropy of English texts is computed in Section 2.15.

1.5.7 Coding and help in R

Coding in R, as any other programming language, may be a frustrating experience even for a mature programmer. There are two types of error messages: syntax and run-time errors. Syntax errors are displayed after the window with the code is closed. Unfortunately, while R reports the line where the error may be, (i) the actual error may be not in this line but below or above, and (ii) the built-in editor in R does not number lines. More sophisticated text editors can be used to get line numbers. Examples include `Rstudio` (official site `https://www.rstudio.com`) and `Notepad++` (official site `https://notepad-plus-plus.org`). In fact, `Rstudio` provides a different Graphical User Interface (GUI) for R with multiple windows for the command line, graphs, etc.

There is no unique advice to understand why your code does not work when you get run-time errors. Printing out values of variables or arrays may help to figure out what is going on and where exactly the error occurs.

It is easy to lose the code in R. I recommend saving the R code in a text file via the `dump` command as in the code `frlJLET` in Example 1.14. To restore this code, issue `source("c:\\Stat Book\\frlJLET.r")` in the R console.

To get help on a known function, say, `legend` issue `?legend` on the command line. Documentation is the Achilles' heel of R. You are likely to find that the help too concise and sometimes without examples. Typically, options are not explained well. Google it!

Problems

1. Write an R function with argument `x` that computes e^x using Taylor series expansion $1+\sum_{n=1}^{\infty} x^n/n!$ and compare with `exp(x)`. Report the output of your function with `x=-1` and compare it with `exp(-1)`. [Hint: Use `for` loop until the next term is negligible, say, `eps`$<$ `10^-7`.]

2. Write an R function with `n` and `m` as arguments, matrix dimensions. Generate a matrix, say, `matrix(1:(n*m),ncol=m)`. Compute a vector of length `n` as the sum of values in each row using `rowSums`. Do the same using the function `apply`. Then use `sum` with a loop over rows and compare the results.

3. Modify function `vecomp` to approximate integral $\int\int_{x^2+x+y^2<2}(1+2x^2+y^2)^{-1}dxdy$. [Hint: The true value is 2.854; use grid for x from -2 to 1 and for y from $-3/2$ to $3/2$.]

4. Modify Example 1.13 to compare `apply` with the traditional `for` loop in terms of computation time using a large matrix `X` (use `date()` before and after computation. A more accurate time can be computed using the `Sys.time()` call.

5. Demonstrate that `apply` is equivalent to `colMeans`, `rowMeans`, `colSums`, and `rowSums`, but slower. [Hint: Use the `for` loop to obtain the time in the order of minutes.]

6. Write a formula for a straight line that goes through two points and verify this formula graphically using `R` code. The code should have four arguments as coordinate of the first and second point. Use `points` to display the points. [Hint: The equation of the straight line that goes through points (x_1, y_1) and (x_2, y_2) is $(y - y_1)/(y_2 - y_1) = (x - x_1)/(x_2 - x_1)$.]

7. Find the minimum of the function $ax + \mathrm{e}^{-bx}$ graphically where a and b are positive user-defined parameters. Plot the function and numerically find where it takes a minimum using the `which` command. Display the minimum point using `points` and display the minimal value. Find the minimum analytically by taking the derivative, and compare the results.

8. Rearrange the plot in function `my1sr` starting with the most frequent letter 'e'. [Hint: Use `[order(-numfr)]` to rearrange the rows of the `freqlet` matrix.]

9. In the analysis of letter frequency, capital letters were reduced to lowercase using the `tolower` command. Does the conclusion remain the same if only lowercase letters are compared (without using `tolower`)?

10. Find a text on the Internet and save it as a txt file. Use the `for` loop over the words and the loop over the number of characters in the word (`nchar`). Then use `substring` to parse words into characters, compute the frequencies, and plot them as in Figure 1.6 using `green` bars.

11. (a) Confirm by simulation the results of Example 1.11 using the `for` loop. (b) Use vectorized simulations. [Hint: Use `X=runif(n=1)<0.2 to` generate a Bernoulli random variable.]

12. (a) Is the probability of having a boy and a girl in the family the same as having two boys or two girls? (b) Use vectorized simulations to confirm the analytical answer. [Hint: Generate Bernoulli random variables as in the previous problem.]

1.6 Binomial distribution

The binomial distribution is the distribution of successes in a series of independent Bernoulli experiments (trials). Hereafter we use the word *success* just for the occurrence of the binary event and *failure* otherwise. More precisely, let $\{X_i, i = 1, 2, ..., n\}$ be a series of independent identically distributed (iid) Bernoulli random variables with the probability of success in a single experiment p, meaning that

$\Pr(X_i = 1) = p$. Consequently, the probability of failure is $\Pr(X_i = 0) = 1 - p$. Note that sometimes the notation $q = 1 - p$ is used.

Table 1.1. Four `R` functions for binomial distribution (`size=`n, `prob=`p)

`R` function	Formula	Returns	Explan.
`dbinom(x,size,prob)`	$\binom{n}{m}p^m(1-p)^{n-m}$	(1.8)	`x` = m
`pbinom(q,size,prob)`	$\sum_{m=0}^K \binom{n}{m}p^m(1-p)^{n-m}$	(1.10)	`q` = K
`rbinom(n,size,prob)`	$X_1, X_2, ..., X_N \overset{\text{iid}}{\sim} B(n,p)$	rand numb	`n` = N
`qbinom(p,size,prob)`	$\sum_{k=0}^K \binom{n}{m}p^m(1-p)^{n-m} = P$	quantile	`p` = P

We want to find the probability that in n independent Bernoulli experiments, m successes occur. Since $X_i = 1$ encodes success and $X_i = 0$ encodes failure, we can express the number of successes as the sum, $X = \sum_{i=1}^n X_i$. We use the notation $X \sim \mathcal{B}(n,p)$ to indicate that X is a binomial random variable with n trials and the probability of success p in a single trial. Clearly, X can take values $0, 1, ..., m, ..., n$. The celebrated *binomial probability* formula gives the distribution of X, namely,

$$\Pr(X = m) = \binom{n}{m}p^m(1-p)^{n-m}, \quad m = 0, 1, ..., n, \tag{1.8}$$

where the coefficient $\binom{n}{m}$ is called the *binomial coefficient* (we say "n choose m") and can be expressed through the factorial

$$\binom{n}{m} = \frac{n!}{m!(n-m)!}.$$

We let $\binom{n}{0} = 1$; obviously, $\binom{n}{n} = 1$. An algebraic application of the binomial coefficient is the expansion of the nth power of the sum of two numbers:

$$(a+b)^n = \sum_{m=0}^n \binom{n}{m} a^m b^{n-m}. \tag{1.9}$$

In some simple cases, we do not have to use the binomial coefficient. For example, the probability that in n experiments there is no one single success is $\Pr(X = 0) = (1-p)^n$. Similarly, $\Pr(X = n) = p^n$. However, all these probabilities can be derived from the general formula (1.8).

The binomial distribution is built into `R`. There are four different functions for each distribution in `R` including the binomial distribution. The `R` function `pbinom` computes the cdf, the probability that $X \leq K$,

$$F(K) = \Pr(X \leq K) = \sum_{m=0}^K \binom{n}{m} p^m (1-p)^{n-m}, \quad K = 0, 1, ..., n. \tag{1.10}$$

The `R` function `qbinom` is the inverse of `pbinom` and computes K such that the cdf equals the specified probability. All four `R` functions for the binomial distribution are presented in Table 1.1. Arguments may be vectors. For example, `dbinom(0:size,size,prob)` computes probabilities for each outcome $m = 0, ..., n$, where `size` $= n$ and `prob` $= p$.

Note that the quantile of a discrete distribution (K) may be not exactly defined given p due to discreetness of the cdf. The call `qbinom` returns the smallest integer for which the cdf is greater or equal to p. For example, `qbinom(p=0.2,size =5, prob=0.5)` and `qbinom(p=0.4,size=5,prob=0.5)` return the same `2`. In contrast, `pbinom(q=2,size=5,prob=0.5)=0.5`.

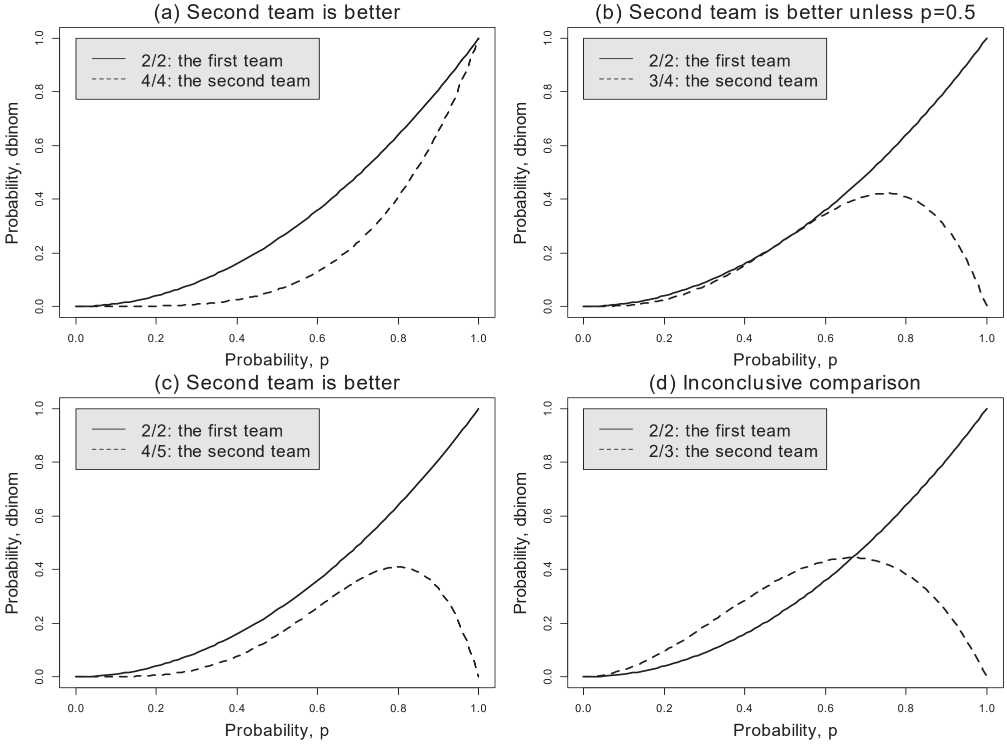


Figure 1.7: *Illustration for Example 1.15. Four answers to the question what team is better based on the number of wins. The second team is better if the probability of the outcome is smaller for all* $0 < p < 1$.

Example 1.15 ***What team is better?*** *Two teams, say, soccer teams, are compared. The first team won two out of two games, and the second team won three out of four games. What team is better? In other words, what team has a better chance/probability of winning in a single game?*

Solution. Before advancing to the solution, we make a couple of assumptions: (i) The teams did not play against each other, that is, 2/2 and 3/4 are scores of the games with other teams. (ii) The probability of winning against other teams is the same for each team under consideration. These assumptions imply that the number of wins follows the binomial distribution with Bernoulli probabilities to win in a single game for team 1 and 2, p_1 and p_2, respectively. Thus, we ask, is $p_1 < p_2$, $p_1 > p_2$, or is the answer inconclusive?

First, consider the case when two teams won all games played (all wins). Let the number of games played by the first team be n_1 and the number of games played by the second team be n_2. Since all games have been won, one may deduce that $p_1 = 1$ and $p_2 = 1$, and the naive answer is that the two teams are the same. However, consider a specific case when $n_1 = 1$ and $n_2 = 100$. Clearly, the second team should be claimed better. In general, under the "all wins" scenario, the second team is better if $n_1 < n_2$. Now we derive this intuitive rule by comparing the probabilities of winning all games. For the first team this probability is p_1^{n1}, and for the second team this probability is p_2^{n2}. How would probabilities p_1 and p_2 be related if the teams won all games they played, n_1 and n_2, respectively? To find p_2 as a function of p_1, we need to solve the equation $p_1^{n1} = p_2^{n2}$ for p_2; that gives $p_1 = p_2^{n_2/n_1} < p_2$. This means that the second team is better, coinciding with our intuition. We do not have to solve the equation for p_1, just plot p^{n_1} and p^{n_2} for $0 < p < 1$ on the same graph; see the upper-left panel in Figure 1.7 with $n_1 = 2$ and $n_2 = 4$, denoted as 2/2 and 4/4, respectively. If the second curve is below, the second team is better.

Second, consider the case where the first team played n_1 games and won all of them, as before, and the second team played n_2 games and won m_2 ($m_2 < n_2$). A naive answer is that the first team is better because $p_1 = n_1/n_1 = 1$ and $p_2 < m_2/n_2 < 1$. But consider the case when $n_1 = 2$ and $m_2 = 99$ and $n_2 = 100$. In this case, the second team is better. As in the previous case, find p_1, which leads to the same result of winning for the second team, $p_1^{n_1} = \binom{n_2}{m_2} p_2^{m_2}(1-p_2)^{n_2-m_2}$. If

$$p_1 = \left[\binom{n_2}{m_2} p_2^{m_2}(1-p_2)^{n_2-m_2}\right]^{1/n_1} < p_2$$

for all $0 < p_2 < 1$, then the second team is better. Again, it is convenient to plot the curves p^{n_1} and $\binom{n_2}{m_2} p^{m_2}(1-p)^{n_2-m_2}$ on the same graph. If the second curve is below the first one for all p the second team is better. If $n_1 = 2$ and $m_2 = 3, n_2 = 4$, the second team is not worse if $\binom{4}{3} p^3(1-p)^{4-3} \leq 1$ for all $0 < p < 1$. But $\binom{4}{3} = \frac{4!}{3!(4-3)!} = \frac{4\times 3!}{3!(4-3)!} = 4$ and indeed $4p^3(1-p) < 1$. Note that the inequality becomes an equality when $p = 0.5$. Hence the second team is better unless the chance of winning in each game is 50/50 for both teams; see panel (b) of Figure 1.7.

Third, in general, we say that the second team is better if its outcome probability is smaller than the outcome of the first team:

$$\binom{n_2}{m_2} p^{m_2}(1-p)^{n_2-m_2} < \binom{n_1}{m_1} p^{m_1}(1-p)^{n_1-m_1}, \quad 0 < p < 1.$$

The two bottom graphs depict $m_2 = 4$ and $n_2 = 5$ and $m_2 = 2$ and $n_2 = 3$. The last outcome leads to an inconclusive comparison. We show in Example 3.39 how to solve the problem of an inconclusive comparison using a noninformative prior for p.

Example 1.16 ***Birthday problem.*** *What is the probability that at least two students in a class of 23 students have the same birthday (assume that there are 365 days/year and birthdays are independent)?*

Solution. It is reasonable to assume that the birth rate is constant over the year. Let the students names be John, Catherine, Bill, etc. Instead of finding the probability that at least two students have the same birthday, we find the probability that all 23 students have birthdays on different days. The probability that John and Catherine do not have the same birthday is $(365-1)/365$. The probability that John, Catherine, and Bill do not have birthday on the same day is $(365-1)(365-2)/365^2$, etc. Finally, the probability that 23 students have at least one shared birthday is

$$1 - \frac{(365-1)(365-2)\cdots(365-22)}{365^{22}}. \tag{1.11}$$

An `R` code that computes this probability is `1-prod(seq(from=364,to=365-22, by=-1))/365^22`, which gives the answer `0.5072972`. Some may find this probability surprisingly high.

Simulations for the birthday problem

It is instructive to do simulations in `R` to verify formula (1.11). Imagine that you can go from class to class with the same number of students and ask if at least two students have the same birthday. Then the empirical probability is the proportion of classes where at least two students have the same birthday.

The following built-in `R` functions are used.

`ceiling(x)` returns the minimum integer, which is greater or equal to `x`. There are two similar functions, `round` and `floor`.

`unique(a)` where `a` is a vector, returns another vector with the vector's distinct (unique) values. For example, `unique(c(1,2,1,2,1))` returns a vector with components 1 and 2.

`length(a)` returns the length (dimension) of the vector.

Function `birthdaysim` simulates this survey. The key to this code is the fact that if a vector has at least two of the same components, the number of unique elements is smaller than the length of the vector. Using the default value Nexp = 100,000, this program estimates the proportion to be 0.507. It is important to understand that each run comes with a different number, but all are around 0.507, a close match with the theoretical value.

Unlike theoretical solution (1.11), this simulation program, after a small modification, allows estimating this probability under the assumption that the distribution of birthdays is not uniform throughout the year. In fact, data from different countries confirm that the distribution is not uniform as discussed by Borja [12]: For many countries in the northern hemisphere, the probability of birth on a specific day looks like a sinusoid with the maximum around September. For example, we can model the probability as $\Pr(\text{birthday} = d) = 0.9 - 0.2\sin(2\pi d/365)$ for $d = 1, 2, ..., 365$. The `R` function is `birthdaysim.sin`. Simulations may take several minutes (use, say, `Nexp=10000` to reduce the time). The probability estimate is around 0.516, slightly higher than under the uniform-birthday distribution. We make a few comments on the code: (i) `cumsum` computes the cumulative sum of an array; this is the shortest way to compute a cdf. (ii) `runif(1)` generates a random number on the interval (0,1); the uniform distribution will be covered in detail in the next chapter.

Problems

1. Derive the formula $\sum_{m=0}^{n}\binom{n}{m} = 2^n$ from the binomial probability (1.8).

2. Use formula (1.8) to prove that $(a+b)^3 = a^3 + 3ab^2 + 3ba^2 + b^3$.

3. Find m for which the binomial probability is maximum.

4. Erica tosses a fair coin n times and Fred tosses $n+1$ times. What is the probability that Fred gets more heads than Erica. Solve the problem theoretically and then write an `R` function with simulations (n is the argument of the function). Compare the results. [Hint: Use the fact that the probability of getting m_F heads in Fred's tosses and m_E heads in Erica's tosses is the product of the probabilities. Apply command `mean(X>Y)` to compute the proportion of elements of array `X` greater than `Y`.]

5. In the birthday example, estimate the probability using simulations that *exactly* two students have the same birthday. (Before computing, what probability is less, *exactly two* or *at least two*?) Plot the theoretical and empirical (from simulations) probabilities that at least two students have the same birthday in the class of n students versus n. Explain the comparison.

6. Sixteen players of two genders sign up for a tournament. What is the probability that there will be eight men and women? Provide a theoretical answer and confirm via simulations.

7. What is the probability that in n fair coin tosses, there will be m head streaks? Give a theoretical formula and verify by simulations. Use two methods to compute the theoretical probability: with or without `dbinom`. Write `R` code to compute the simulated probability with the number of

simulations equal to 1000K. [Hint: Your solution should be one line of code.]

8. The chance that a person will develop lung cancer in his/her lifetime is about 1 in 15. What is the probability that in a small village with 100 people, (a) there are no individuals with lung cancer, and (b) there is only one case? In both cases check your answer with simulations. [Hint: Use `rbinom`.]

9. Early mathematicians believed that in a fair coin tossing after a long streak of heads, a tail is more likely. Here is one of the proofs: make 100 tosses and consider the longest streak of heads. Tails always occurs after the streak. Give pros and cons for this statement. Do you agree with this statement?

10. Conjecture: Any nonuniform distribution may only increase the probability that at least two people have the same birthday. In other words, (1.11) is the lower bound over all possible distributions of birthdays on $1, 2, ..., 365$. [Hint: Modify function `birthdaysim.sin` to check using simulations; run the function with user-defined birthday probabilities within each month.]

1.7 Poisson distribution

Poisson distribution is useful for modeling a distribution of counts. This distribution specifies the probability that a discrete random variable takes value k, where k may be one of $0, 1, 2, ...$ Simeon Denis Poisson (1781–1840), a famous and powerful French mathematician and physicist, introduced this distribution. We say that the random variable X has a Poisson distribution with a positive parameter λ if

$$\Pr(X = k; \lambda) = \frac{1}{k!}\mathrm{e}^{-\lambda}\lambda^k, \quad k = 0, 1, 2, ... \tag{1.12}$$

The presence of the factorial in the denominator (the normalizing coefficient) can be seen from the following calculus formula: $\mathrm{e}^x = 1 + x + \frac{1}{2!}x^2 + \cdots + \frac{1}{k!}x^k + \cdots$. Using this formula it is clear that the probabilities add to one, $\sum_{k=0}^{\infty} \Pr(X = k) = 1$, as in the case with all probability distributions. We will use the symbolic notation $X \sim \mathcal{P}(\lambda)$ to indicate that X has a Poisson distribution with a positive parameter λ. Sometimes λ is called the Poisson rate.

In Table 1.2, we show four `R` functions associated with the Poisson distribution. As is the case with other built-in distributions in `R`, these functions take vector arguments. For example, if `lambda` is a scalar and `x` is a vector, then `dpois(x,lambda)` returns a vector of the same length as `x` keeping `lambda` the same. If `lambda` is a vector of the same length as `x`, this function returns a vector of probabilities computed using the corresponding pairs of `x` and `lambda`. Even though the support of the Poisson distribution is infinite, there are many examples when this distribution may serve as a good probabilistic model despite the fact that a random variable takes a finite set of values when the upper limit is difficult to specify: the number of children in the family, the number of people visiting a specific website, minutes between consecutive telephone calls, the number of earthquakes in 10 years, the number of accidents in town, etc.

Table 1.2. Four `R` functions for Poisson distribution (`lambda=`λ)

`R` function	Formula	Returns	Explan.
`dpois(x,lambda)`	$\frac{1}{k!}\mathrm{e}^{-\lambda}\lambda^k$	(1.12)	`x` = k
`ppois(q,lambda)`	$\sum_{k=0}^{K}\frac{1}{k!}\mathrm{e}^{-\lambda}\lambda^k$	cdf	`q` = K
`rpois(n,lambda)`	$X_1, X_2, ..., X_n \overset{\text{iid}}{\sim} P(\lambda)$	rand. numbers	`n` = n
`qpois(p,lambda)`	$\sum_{k=0}^{K}\frac{1}{k!}\mathrm{e}^{-\lambda}\lambda^k = p$	quantile, K	`p` = p

A continuous analog of the Poisson distribution is the gamma distribution; the two distributions take a similar shape; see Section 2.6. The following is a continuation of Example 1.8.

Example 1.17 ***How many toys to buy?*** *Assuming that the number of children in the family follows a Poisson distribution with* $\lambda = 2.4$, *how many toys should one buy so that every child gets a present.*

Solution. Strictly speaking, whatever number of toys you buy, there is a chance that at least one child will not get a toy because theoretically the number of children is unbounded. As is customary in probability and statistics, we define a probability, close to 1, that each child gets a present. Let this probability be $p = 0.9$. Then the minimum number of toys to buy is the quantile of the Poisson distribution with $\lambda = 2.4$ and is computed in `R` as `qpois(p=0.9,lambda=2.4)=4`.

The Poisson distribution possesses a very unique property: the mean and the variance are the same,

$$E(X) = \mathrm{var}(X) = \lambda.$$

To prove this fact, we refer to the definition of the mean of a discrete random variable:

$$E(X) = \sum_{k=0}^{\infty} k \times \Pr(X = k; \lambda).$$

Using the expression for the probability (1.12) and the fact $\sum_{k=0}^{\infty}\frac{\mathrm{e}^{-\lambda}}{k!}\lambda^k = 1$, we obtain

$$E(X) = \lambda \sum_{k=1}^{\infty} \frac{\mathrm{e}^{-\lambda}}{(k-1)!}\lambda^{k-1} = \lambda \sum_{k=0}^{\infty} \frac{\mathrm{e}^{-\lambda}}{k!}\lambda^k = \lambda.$$

Analogously, one can prove that $E(X^2) = \lambda + \lambda^2$, which, in conjunction with formula (1.6), gives $\mathrm{var}(X) = \lambda$.

The binomial and Poisson distributions are relatives. Loosely speaking, the Poisson distribution is a limiting case of the binomial distribution. This fact is formulated rigorously as follows.

Theorem 1.18 *A binomial distribution with an increasing number of trials, n, and decreasing probability of successes, $p = p(n)$, converges to a Poisson distribution with parameter, $\lambda = \lim_{n\to\infty} np(n)$.*

Proof. Expressing p through λ and n and substituting it into the formula for binomial probability (1.8), we obtain

$$\begin{aligned}\Pr(X = k) &= \frac{n!}{k!(n-k)!}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k}\\ &= \frac{\lambda^k n!}{k!(n-k)!}\frac{1}{n^k}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-k}.\end{aligned}$$

From the famous limit $\lim_{n\to\infty}\left(1+\frac{a}{n}\right)^n = \mathrm{e}^a$ for any fixed a, we have

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n = \mathrm{e}^{-\lambda}.$$

But when $n \to \infty$, we have $\lambda/n \to 0$. In addition, using Stirling's formula, it is possible to prove that $\frac{n!}{(n-k)!}\frac{1}{n^k} \to 1$. Therefore,

$$\lim_{n\to\infty} \Pr(X = k) = \frac{1}{k!}\lambda^k \mathrm{e}^{-\lambda},$$

the Poisson probability. □

As an example, consider the distribution of the number of stolen credit cards a credit company experiences over the year. The probability that a credit card will be stolen $m \geq 1$ times over the year for a specific customer (binomial variable) is very small, say, p. But the credit company may have many customers, n. Then the average number of stolen cards is $\lambda = pn$. Thus, from the Theorem 1.18, one may infer that the number of stolen cards follows a Poisson distribution. An advantage of using the Poisson distribution over the binomial distribution is that the upper limit of stolen cards is not specified (technically speaking, it is infinity). It is important to estimate λ with the assumption that the data on the number of stolen cards is available for several years – a common statistical problem. As we shall learn later, λ may be estimated simply as an average of the number stolen cards over the years.

Example 1.19 ***Raisin in the cookie.*** *n raisins are well mixed in the dough and m cookies are baked. What is the probability that a particular cookie has at least one raisin? Provide the exact formula using the binomial distribution and the Poisson approximation.*

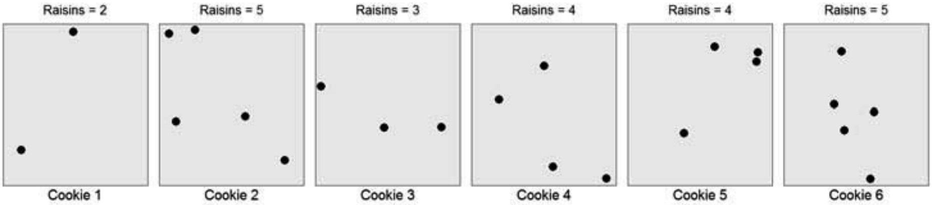


Figure 1.8: *Illustration of the distribution of the number of raisins in a cookie*, $m = 6$, $n = 30$.

Solution. Let X denote the number of raisins in a cookie. See Figure 1.8 for an illustration; since the shape of the dough and cookies does not matter, we depict cookies to have a rectangular shape for simplicity of display. Assuming that raisins are well mixed and the volumes of cookies are equal there will be n/m raisins in the cookie on average with the probability of one raisin $p = 1/m$. The exact distribution of the number of raisins in the cookie X follows a binomial distribution with probability $\Pr(X = k) = \binom{n}{k}p^k(1-p)^{n-k}$. Thus, the exact probability that the cookie has at least one raisin is

$$\Pr(X \geq 1) = 1 - \Pr(X = 0) = 1 - \left(1 - \frac{1}{m}\right)^n. \tag{1.13}$$

If m and n are large, we can approximate the binomial distribution with the Poisson distribution letting $\lambda = n/m$. For large m, the probability, $1/m$, is small, and X can be interpreted as the repetition of n experiments with the mean $\lambda = pn = n/m$. Thus, the probability that a particular cookie has k raisins can be approximated as

$$\Pr(X = k) = \frac{1}{k!}\mathrm{e}^{-n/m}\left(\frac{n}{m}\right)^k. \tag{1.14}$$

Then, the probability that at least one raisin will be in a cookie is $\Pr(X > 0) = 1 - \Pr(X = 0) = 1 - \mathrm{e}^{-n/m}$. We shall show that (1.14) and (1.13) are close for large m. Indeed, using the approximation $(1 - 1/m)^n \simeq \mathrm{e}^{-n/m}$ and letting $\lambda = n/m$, we obtain $\lim_{m\to\infty}\left(1 - \frac{1}{m}\right)^n = \mathrm{e}^{-\lambda} \simeq \mathrm{e}^{-n/m}$. Thus, for large m, both formulas give close answers. □

In applied probability and statistics, we often need to translate a vaguely formulated real-life question to a rigorously defined problem. Typically, we need to make some assumptions – the next example illustrates this point.

Example 1.20 ***Probability of a safe turn.*** *Consider turning from a side street onto a busy avenue. There are, on average, 10 cars per minute passing your side street. Assuming that the time between cars follows a Poisson distribution and it takes five seconds to enter the traffic stream safely, what is the probability that you will be able to enter without waiting for a break in traffic?*

Solution. Let X denote the time in seconds between two passing cars. On average, the number of seconds between two passing cars is $60/10 = 6$. Since X

follows a Poisson distribution, $X \sim \mathcal{P}(6)$, in order to safely turn on the avenue right after you come to the intersection, the time between two passing cars should be $X > 5$. Thus, the requested probability is computed as complementary to the cdf:

$$\Pr(X > 5) = 1 - \Pr(X \leq 5) = 1 - \texttt{ppois(q} = \texttt{5,lambda} = \texttt{6)} = \texttt{0.5543204}.$$

Do not make an instant turn unless you are chased by a person who wants to kill you. See Example 2.17, where we compute the wait required for a safe turn. □

An important property of the Poisson distribution is that the sum of independent Poisson variables follows a Poisson distribution with the rate equal the sum of individual rates. Namely, if $X_i \sim \mathcal{P}(\lambda_i)$, $i = 1, 2, ..., n$ and X_i are independent, then

$$\sum_{i=1}^{n} X_i \sim \mathcal{P}\left(\sum_{i=1}^{n} \lambda_i\right). \tag{1.15}$$

We will prove this property using the moment generating function in Section 2.5 of the next chapter. The next example uses this property.

Example 1.21 ***No typos.*** *The distribution of the number of typographical errors per 100 pages of a document follows a Poisson distribution with a mean value 4. (a) What is the probability that a 300-page book will have no typos? (b) What is the probability that it will have more than 5 typos? (c) What is the probability that a book of 157 pages has no typos? (d) Run simulations that confirm your analytic answer with L as an argument in the* `R` *code.*

Solution. (a) Let X_1 be the number of typographical errors on the first 100 pages of the book. We know that $X_1 \sim \mathcal{P}(4)$. Analogously, let X_2 and X_3 be the number of errors on the next 100 pages and the last 100 pages, respectively. Assuming that the locations of errors are independent, the number of errors in 300 pages, $X = X_1 + X_2 + X_3$, has a Poisson distribution with $\lambda = 3 \times 4 = 12$. Now, to compute the probability that there are no errors on 300 pages, we simply let $k = 0$, so the answer is $\Pr(X = 0) = \frac{1}{0!\mathrm{e}^{12}} 12^0 = \mathrm{e}^{-12} = 6.1 \times 10^{-6}$. Alternatively, we can compute this probability in `R` as `dpois(0,lambda=12)`. In the previous solution, we assume that the distribution of typos follows a Poisson distribution. Alternatively, one can assume that the distribution of typos follows a binomial distribution with $n = 300$ and the probability of typo per page $p = 4/100 = 1/25$. Then the probability that there will be no typos in 300 pages is $(1 - 1/25)^{300} = 4.8 \times 10^{-6}$. (b) The probability that a 300-page book has 5 typos or less is the cumulative probability and can be computed using `ppois` function. The probability that the book has more than 5 typos is the complementary probability, `1-ppois(5,lambda=12)`. The answer is `0.9872594`. (c) The previous solution used in (a) and (b) only works when the number of pages in the book is multiple of 100. What if the number of pages in the book is not multiple of 100, like

157? Then the probability that there are $X = k$ typos on L pages can be modeled using a Poisson distribution with $\lambda = 4L/100 = 0.04L$. Following the previous argument, the probability that there are no typos in a 157-page book is $(0!\mathrm{e}^{0.04\times 157})^{-1}(0.04\times 157)^0 = \mathrm{e}^{-0.04\times 157} = 1.87\times 10^{-3}$. (d) Imagine a large number of books with L pages with the number typos distributed according to the Poisson distribution $\mathcal{P}(0.04L)$. If X counts the number of typos in each book, the probability that there are no typos is estimated as the proportion of books with $X = 0$. The probability of this setup can be estimated using the following line of code, `mean(rpois(n=10000000,lambda=0.04*157)==0)`, where `n` is the number of books/simulations and $L = 157$ (of course, `n` and `lambda` may take any values). This command gives the result `0.0018742`, close to our analytic answer.

Problems

1. (a) Prove that, for $\lambda < 1$, probabilities of the Poisson distribution decrease with k. (b) Prove that for $\lambda > 1$ the maximum probability occurs around $\lambda - 1$. [Hint: Consider the ratio $\Pr(X = k+1)$ to $\Pr(X = k)$.]

2. Explain why the distribution of rare diseases, such as cancer, follows a Poisson distribution.

3. Find the minimum number of toys in Example 1.17 that ensures each child gets a toy with probability 99%.

4. (a) Write an `R` program that computes and plots Poisson distributions of the fertility rates (births per woman) in the United States and India. See https://data.worldbank.org/indicator/sp.dyn.tfrt.in. [Hint: Use `plot` with option `type="h"` and depict the two bars with different color slightly shifted for better visibility, use `legend`.]

5. Prove that for the Poisson distribution, $\mathrm{var}(X) = \lambda$ using formula (1.6).

6. What is the probability that one cookie will have all m raisins?

7. It is well known that the interval mean ±2× SD covers about 95% of the distribution. Test this statement for the Poisson distribution by generating observations using the `ppois` function in `R`: plot the Poisson probabilities on the y-axis versus a grid of values for λ on the x-axis, say, `lambda=seq(from=.1,to=3,by=.1)`, and plot the 0.95 horizontal line.

8. Illustrate Theorem 1.18 by plotting Poisson and binomial probabilities side by side for large n and small p (make them arguments of your `R` function).

9. 10% of families have no children. Assuming that the number of children in the family follows a Poisson distribution, estimate the average number of children in the family.

10. What is the probability that a randomly chosen family with the number of children two or more has children of the same gender? Run simulations to check the answer. [Hint: Assume a Poisson distribution and use conditional probability.]

11. Referring to problem 3 from Section 1.2, how many monkeys is required so that at least one monkey types the novel with probability 0.9? Use an approximation similar to that from Example 1.19.

12. (a) Write `R` code that simulates Example 1.20. (b) Confirm the probability of a safe turn using simulations.

13. The number of children in the family follows a Poisson distribution with $\lambda = 1.8$. Six families with kids are invited to a birthday party. What is the probability that more than 15 kids come to the party. Give the answer under two scenarios: (a) family brings all kids they have, and (b) the probability that they bring a child is 0.8. Write a simulation program to check your answer.

1.8 Random number generation using `sample`

1.8.1 Generation of a discrete random variable

A general finite-value discrete random variable, X, takes values $\mathbf{x} = \{x_i, i = 1, 2, ..., n\}$ with probabilities $\mathbf{p} = \{p_i, i = 1, 2, ..., n\}$. In this section, we address the problem of generating a random sample from this distribution. In `R`, this can be done using a built-in function `sample`: to generate a random sample of size 10K we issue `sample(x=x, size=10000, replace = T, prob = p)`. Note that arrays `x` and `p` must have the same length; `replace` means that values are drawn with replacement (otherwise, if `size` is greater than n, the sample cannot be generated).

In the following example, we use the function `sample` to generate 10,000 Poisson random numbers and test the sample by matching the empirical and theoretical probabilities. Note that since Poisson random variable is unbounded, we must truncate the probabilities. Of course, `sample` is used here solely for illustrative purposes. A better way to generate Poisson numbers is to use `rpois`.

Example 1.22 `R` ***sample***. *Generate 10,000 observations from a Poisson distribution with parameter λ using* `sample` *and compare with theoretical probabilities by plotting them on the same graph.*

Solution. We need to set an upper bound on the outcomes x_i since the Poisson distribution is unbounded. For example, we may set $\max(x) = \lambda + 5\sqrt{\lambda}$; this guarantees that the right-tail probability is very small. The `R` code is found in file `sampP.r`. Option `replace=T` implies that some observations may repeat. We may use `replace=F` (the default option) if distinct observations are required. This option is used when a subsample with no repeat values is needed (`size` $< n$). For example, if one needs a subsample from a survey of 1,000 families, option `replace=F` must be used because otherwise one may get repeated observations as if the same family was asked twice. Note that the empirical probabilities do not exactly match the theoretical probabilities (bars) even for a fairly large number of simulated values, `nSim=10000.` However `sampP(nSim=100000)` produces points landing almost on the top of the bars. □

One can sample from a character vector `x`. For example, `sample(x=c("John", "Tom","Rebecca"),size=100,rep=T,prob=c(.25,.5,.25))` produces an array of size 100 with components `"John","Tom"`,and `"Rebecca"`. The function `sample` is very useful for resampling such as bootstrap – see Example 5.7.

1.8.2 Random Sudoku

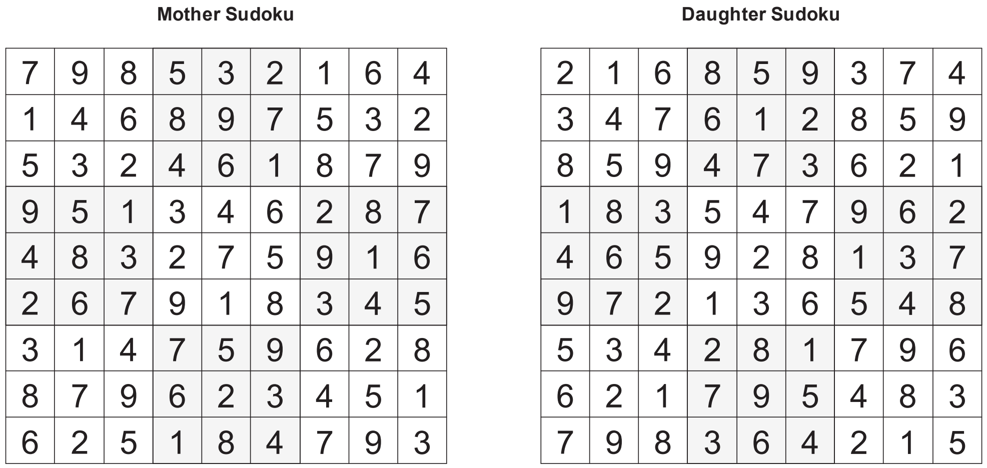
Sudoku is a popular mathematical puzzle that originated in France almost two hundred years ago. The objective is to fill blank cells of a 9×9 grid with digits so that each column, each row, and each of the nine 3×3 subsquares contains all of the digits from 1 to 9. We say Sudoku is complete (solved) if there are no blanks; see Figure 1.9 for some *complete Sudoku* puzzles. By contrast, a Sudoku with blank cells are called *puzzle Sudoku.* The goal of this section is to illustrate how `sample` command can be employed to generate and display Sudoku puzzles.

The key to the following theorem is permutation of indices $\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$. For example, a permutation vector $\mathbf{p} = (3, 9, 5, 4, 8, 7, 2, 6, 1)$ means that 1 is replaced with 3, 2 is replaced with 9, etc. In words, this operation may be viewed as index relabeling.

Theorem 1.23 *Let* $\mathbf{A}$ *be a complete Sudoku and* $\mathbf{p}$ *be any permutation of* $\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$. *Then* $\mathbf{B} = \mathbf{A}(\mathbf{p})$ *is also complete Sudoku.*

According to our definition, $\mathbf{A}(\mathbf{p})$ leads to another complete Sudoku where 1 is replaced with p_2, 2 is replaced with p_2, etc. We refer the reader again to Figure 1.9 where the daughter Sudoku ($\mathbf{B}$) is derived from the mother Sudoku ($\mathbf{A}$) by the permutation vector $\mathbf{p} = (3, 9, 5, 4, 8, 7, 2, 6, 1)$. $9! = 362880$ daughter Sudokus can be derived from a mother Sudoku. The connection to `sample` is that a random permutation vector can be obtained as `sample(1:9,size=9,prob=rep(1/9,9))`. This means that from one mother Sudoku, we can create many other puzzles (daughter Sudokus) using the `sample` command.

As was mentioned earlier, in a puzzle Sudoku, some cells are blank, and Sudoku solver needs to fill in the blanks to arrive at a complete Sudoku; see Figure 1.10. The number of blanks determines how difficult the Sudoku puzzle is. For example, with only few blanks the Sudoku is easy. Similarly, a Sudoku with almost all its cells blank is not difficult to solve as well. To create a puzzle



Mother Sudoku

7	9	8	5	3	2	1	6	4
1	4	6	8	9	7	5	3	2
5	3	2	4	6	1	8	7	9
9	5	1	3	4	6	2	8	7
4	8	3	2	7	5	9	1	6
2	6	7	9	1	8	3	4	5
3	1	4	7	5	9	6	2	8
8	7	9	6	2	3	4	5	1
6	2	5	1	8	4	7	9	3

Daughter Sudoku

2	1	6	8	5	9	3	7	4
3	4	7	6	1	2	8	5	9
8	5	9	4	7	3	6	2	1
1	8	3	5	4	7	9	6	2
4	6	5	9	2	8	1	3	7
9	7	2	1	3	6	5	4	8
5	3	4	2	8	1	7	9	6
6	2	1	7	9	5	4	8	3
7	9	8	3	6	4	2	1	5

Figure 1.9: *Mother and daughter Sudokus. The daughter Sudoku is created from the mother Sudoku upon permutation vector* $\mathbf{p} = (3, 9, 5, 4, 8, 7, 2, 6, 1)$.

Sudoku from a complete Sudoku, one needs to blank out some cells, for example, at random locations. Such a puzzle Sudoku is called a random Sudoku. If k is the number of blank cells, then the number of combinations of empty cells out of 81 is $\binom{81}{k}$. Therefore the total number of possible empty cells is

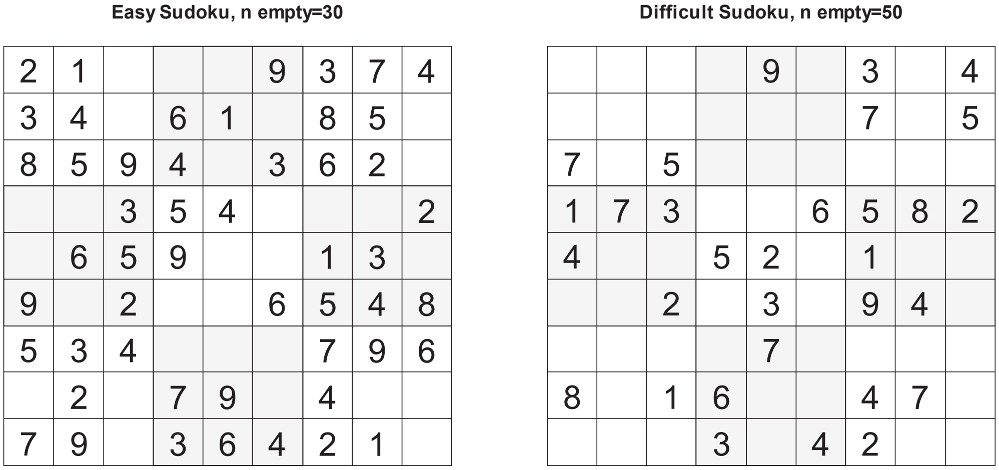
$$\sum_{k=1}^{81} \binom{81}{k} = 2^{81} = 2\,417\,851\,639\,229\,258\,349\,412\,351,$$

enough to keep a Sudoku lover busy!

The `R` code for displaying, testing, and generating a random puzzle Sudoku is found in file `sudoku.r`. The internal function `test.sudoku` tests whether a 9×9 matrix composed of digits from 1 to 9 is a complete Sudoku. It returns 1 if the Sudoku is complete and 0 otherwise. In the latter case, the number of unique values in each row or column is less than 9, or the number of unique values in each small square is less than 9. The internal function `display.sudoku` plots the 9×9 square and the digit in each cell. When the cell is blank, it is specified as missing (`NA`), and therefore is not displayed.

The `sudoku` function does three jobs: The option `job=1` plots the mother Sudoku and tests whether it is complete (of course other mother Sudoku may be used). The option `job=2` creates Figure 1.9. To create a random daughter Sudoku, a random permutation is computed using the `sample` command (the result is vector `i9`). The random number generation is controlled by `setrand`; different values of `setrand` will produce different daughter Sudokus.

The option `job=3` produces a puzzle Sudoku. First, it generates a random daughter Sudoku from the mother Sudoku using a random permutation, and second, the specified number of cells are blanked. In the Sudoku at left in Figure



Easy Sudoku, n empty=30

2	1				9	3	7	4
3	4		6	1		8	5	
8	5	9	4		3	6	2	
		3	5	4				2
	6	5	9			1	3	
9		2			6	5	4	8
5	3	4				7	9	6
	2		7	9		4		
7	9		3	6	4	2	1	

Difficult Sudoku, n empty=50

				9		3		4
						7		5
7		5						
1	7	3			6	5	8	2
4			5	2		1		
		2		3		9	4	
				7				
8		1	6			4	7	
			3		4	2		

Figure 1.10: *Two puzzle Sudokus with different number of blank cells (n empty). The Sudoku at left is easier than Sudoku at right. This figure is generated by issuing* `sudoku(job=3)`.

1.10, this number is `n.blank=30`, and in the Sudoku at right, this number is `n.blank=50`. Once the number of blank cells is given, the random cells get blank/missing again using the `sample` command applied to 81 pairs of digits (i, j) where $i = 1, 2, ..., 9$ and $j = 1, 2, ..., 9$. These 81 pairs are generated using the `rep` command with two options `times` and `each`. Using this function, random puzzle Sudokus of various levels of difficulty can be generated.

Problems

1. Use `sample` to generate binomial random numbers by modifying the aforementioned function `sampP`. Test the sample by plotting the empirical and exact probabilities side by side.

2. Generate $N = 100000$ of $m = 5$-element arrays $(X_1, X_2, ...X_m)$ of Bernoulli variables with probability p using `sample` and compute the proportion of simulated samples when $\sum_{i=1}^{m}(2X_i^2 - X_i) = 1$. Plot this proportion as a function of $p = 0.1, 0.2, ..., 0.9$. Explain the result. [Hint: Generate Nm binary variables and form an $N \times m$ matrix `X`; then use `rowSums`.]

3. Randomly select characters from Example 1.14 and plot the letter frequency in the two samples side by side as in Figure 1.6. Do the frequencies look alike? [Hint: Generate random characters using random index arrays `ir1=sample(1:N,size= N/2,prob=rep(1/N,N))` and to exclude those in `ir1` as `ir2=(1:N)[-ir1]` where `N=length(ch)`.]

4. Write an `R` function that, given two complete Sudokus, tests whether one Sudoku can be obtained as a permutation of another.

5. Is it true that reflections and transpositions of a complete Sudoku can be expressed via permutation?

6. (a) Demonstrate that when the number of blank cells is small or close to 81, Sudoku is easy to solve (by solving yourself). (b) Take any mother Sudoku, create a random Sudoku with number of blank cells equal 40, and try to solve it. Is it harder than in (a)?

7. Create a new mother Sudoku by reflection and transposition of 3×3 squares.

Chapter 2

Continuous random variables

Unlike discrete random variables, continuous random variables may take *any* value on a specified interval. Usually, we deal with continuous random variables that take any value on an infinite or semi-infinite interval, denoted as $(-\infty, \infty)$ or $(0, \infty)$, respectively. Some continuous random variables, such as uniform or beta-distributed random variables, belong to an interval. A characteristic property of a continuous random variable is that $\Pr(X = x) = 0$ for any value x. Although it is possible that a random variable is a combination of a discrete and a continuous variable, we ignore this possibility. Thus, only continuous random variables are considered in this chapter, with calculus as the major mathematical tool for investigating of distributions of these variables. Regarding the notation, typically, random variables are denoted as uppercase, like X, and the values it takes or the argument of the density function as lowercase, x. This notation rule will be followed throughout the book.

2.1 Distribution and density functions

While the cumulative distribution function (cdf) is defined for both types of variables, the probability density function (pdf), or density function, is defined only for the continuous type.

2.1.1 Cumulative distribution function

The cumulative distribution function (cdf), or distribution function of a random variable X, is defined as

$$F_X(x) = \Pr(X \leq x), \quad -\infty < x < \infty. \tag{2.1}$$

In words, the cdf is the probability that the random variable X takes a value less than or equal to x, where x may be any number. Since the distribution function

Advanced Statistics with Applications in R, First Edition. Eugene Demidenko.

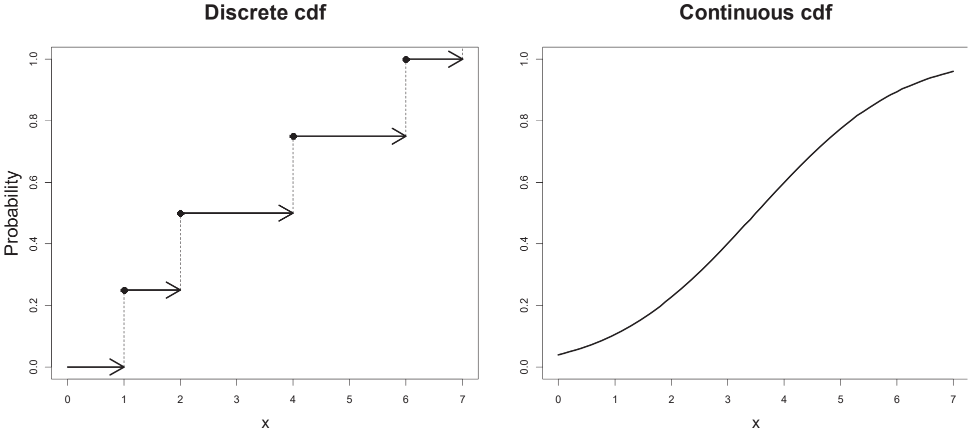


Figure 2.1: *Typical distribution functions for discrete and continuous random variable. Discrete cdf is discontinuous and takes the step p_i and x_i. On the other hand, the cdf of a continuous random variable is a continuous function.*

is a cumulative probability, it is an increasing function of x and takes nonnegative values in the interval [0,1].

A cdf has the following properties:

1. The cdf is an increasing function: $x_1 \leq x_2$ implies $F(x_1) \leq F(x_2)$.

2. The upper limit of the cdf is 1: $F(+\infty) = 1$, or, more precisely, $\lim_{x\to\infty} F(x) = 1$.

3. The lower limit of cdf is 0: $F(-\infty) = 0$, or, more precisely, $\lim_{x\to-\infty} F(x) = 0$.

The first property follows directly from the definition of the distribution function; the other two properties are obvious. A function F is a distribution function if it satisfies properties 1, 2, and 3. For any $a \leq b$ the probability that the random variable belongs to the interval $(a, b]$ can be expressed via a cdf as $\Pr(a < X \leq b) = F(b) - F(a)$. The cdf is defined for discrete or continuous random variable.

If a discrete random variable takes n distinct values $\{x_i, i = 1, 2, ..., n\}$ with respective probabilities $\{p_i, i = 1, 2, ..., n\}$, the distribution function takes a step of height p_i at x_i. This function is continuous from the right but is discontinuous from the left. For a continuous random variable, the distribution function is continuous on the entire real line. See Figure 2.1.

The cdf completely defines the distribution of a continuous random variable. The distribution function is a mathematical concept – in applications, we have to estimate it by the empirical cdf.

2.1.2 Empirical cdf

Let random variable X with unknown cdf, F, take values $x_1, x_2, ..., x_n$. We say that these values are the realization (or observations) of X or a random sample from a general population specified by cdf F.

The procedure to compute the empirical cdf has two steps:

1. List observations in ascending order, $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$, so that $x_{(1)} = \min x_i$ and $x_{(n)} = \max_i x$. $\{x_{(i)}\}$ are called order statistics.

2. Compute the empirical cdf treating the sample as a discrete random variable that takes values $x_{(i)}$ with probability $1/n$. This means that the cdf is computed as

$$\widehat{F}(x) = \frac{1}{n}\#(x_i \leq x), \tag{2.2}$$

where sign $\#$ means the number of observations equal to or less than x. Sometimes the notation $I(x_i \leq x)$ is used, where I is an indicator function with values 0 or 1; then $\#(x_i \leq x) = \sum_{i=1}^n I(x_i \leq x)$. We use "hat" $\widehat{()}$ to indicate that this is an *estimator* of F, a common notation is statistics. It is easy to see that (2.2) can be expressed as a step-wise function with step i/n at $x_{(i)}$. This means that the empirical cdf can plotted as i/n on the y-axis versus $x_{(i)}$ on the x-axis.

For each x, one can treat the numerator in $\widehat{F}(x)$ as the outcome of the binomial random variable with Bernoulli probability $F(x)$ in n trials, where F is the true cdf. Since the expected number of successes is $nF(x)$, we have $E(\widehat{F}(x)) = F(x)$. We say that the empirical cdf is an unbiased estimator of the true cdf. Moreover, since $\text{var}(\widehat{F}(x)) = F(x)(1-F(x)/n$, the empirical cdf approaches F when $n \to \infty$. In statistics, theses two properties of an estimator are called unbiasedness and consistency, respectively.

It is easy to plot the empirical distribution function in `R` using the command `plot()` with option `type="s"` after the original sample is ordered using the `order()` or `sort()` command. This method is illustrated in the following example.

Example 2.1 ***Web hits cdf.*** *An Internet company analyzes the distribution of the number of visitor hits during the day. The dataset* `compwebhits.dat` *contains the times recorded by a computer server for 100 hits during a typical day. Plot the empirical cdf and interpret its pattern.*

Solution. The `R` code below produces Figure 2.2. First, we order observations. Second, we create values for the cdf $\{1/n, 2/n, ...1\}$, and finally plot. Sometimes, we use a continuity correction and plot $(i - 0.5)/n$ on the y-axis,

```
webhits=function()
{
    dump("webhits","c:\\statbook\\webhits.r")
    x<-scan("c:\\statbook\\comwebhits.dat") # read the data
    x<-x[order(x)];n<-length(x) # order observations
    Fx<-(1:n)/n # values for cdf
    plot(x,Fx,type="s",xlab="Time of the website hit, h",
                                  ylab="Probability, cdf")
    text(17,.2,paste("Number of hits during 24 hours =",n))
}
```

Looking at the figure, one may notice that $\widehat{F}(x) = 0$ for $x < 5$ a.m. – people are sleeping. The intensity of visitor hits picks up at around 8 a.m. as people wake up. Hits have a steady rate until 3 p.m. and then slow down. Better insights into the intensity of the hits can be drawn from plotting a histogram; see the next section.

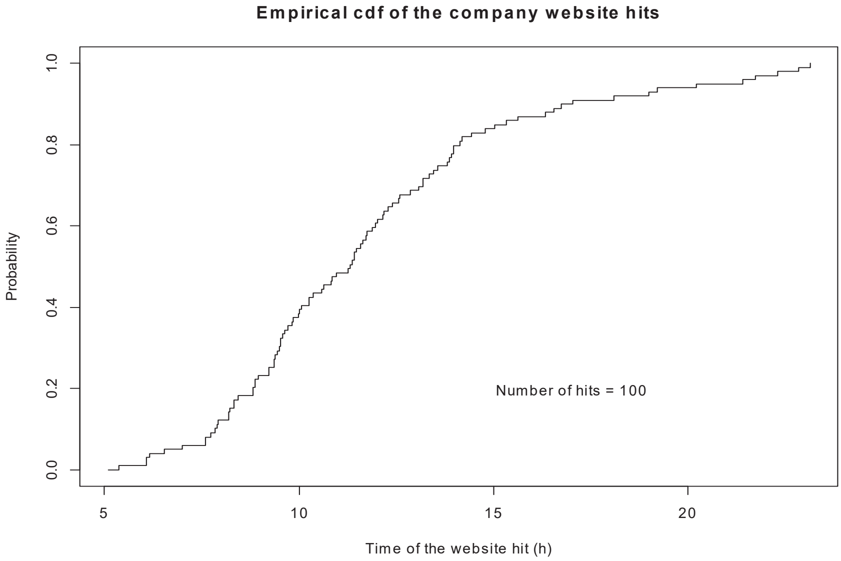


Figure 2.2: *Cumulative distribution function of 100 website hits during a typical day.*

2.1.3 Density function

As we mentioned before, $\Pr(X = x) = 0$ for a continuous random variable. In words, the probability that a continuous random variable takes a specific value is zero. Therefore instead of the point probability, we consider the infinitesimal probability when the length of the interval goes to zero:

$$\frac{\Pr(x < X \leq x + h)}{h}. \tag{2.3}$$

Since $\Pr(x < X \leq x+h) = F(x+h)-F(x)$, we define the infinitesimal probability as

$$f(x) = \lim_{h \to 0} \frac{F(x+h) - F(x)}{h}, \tag{2.4}$$

called the probability density function (pdf), or shortly the density. As follows from the Fundamental Theorem of Calculus, the pdf is the first derivative of the distribution function,

$$f(x) = F'(x),$$

and conversely

$$F(x) = \int_{-\infty}^{x} f(t)dt.$$

Density has the following properties:

1. $f(x) \geq 0$ for all x.

2. $\int_{-\infty}^{\infty} f(x)dx = 1$.

The first property follows from the fact that cdf is an increasing function. Hence pdf cannot be negative. The second property follows from the fact that $F(\infty) - F(-\infty) = 1$. It is easy to see that $f(\pm\infty) = 0$ because otherwise the second property would not hold.

The density reaches its maximum at the *mode*. It is fair to refer to the mode as the most probable value of the random variable because the density at the mode is maximum. We say that the distribution is *unimodal* if there is one mode, i.e. $f(x)$ is (strictly) increasing to the left of the mode and (strictly) decreasing to the right of the mode. We say that the random variable is symmetric if its density is symmetric around the mode. The support of the density is where it is positive. A continuous random variable may be defined on a finite or semi-infinite interval; we call this interval the *support* of the density.

Example 2.2 ***Pdf=cdf.*** *Can the same function be the cdf and the pdf at the same time?*

Solution. Yes. Let $g(x)$ be a cdf and pdf at the same time. Then this function satisfies the ordinary differential equation (ODE) $g' = g$ with a solution $g(x) = \mathrm{e}^x$. Of course, e^x cannot be a cdf for all x because its values goes beyond 1, but if $x \in (-\infty, 0)$, function $g(x) = \mathrm{e}^x$ satisfies the properties of a cdf. Moreover, since the general solution to ODE $g' = g$ is of the form e^{x+c}, where c is any constant, we conclude that the only function that is both a cdf and a density is of the form $g(x) = \mathrm{e}^{x+c}$ on the interval $-\infty < x < -c$. This distribution emerges in Section 2.10 as the limiting distribution of the maximum of observations.

Problems

1. Why can one treat the numerator in (2.2) as the outcome of the binomial random variable with Bernoulli probability $F(x)$ in n trials? [Hint: Use the indicator function $1(X \leq x)$ as a Bernoulli random variable.]

2. A linear combination of distributions is called a mixture distribution; see Section 3.3.2. (a) Prove that if F and G are two cdfs and $0 \leq \lambda \leq 1$ is a fixed number, then $\lambda F + (1-\lambda)G$ is a cdf (mixture). (b) Prove that a similar statement holds for two densities. (c) In the general case, prove that $\sum_{i=1}^n \lambda_i F_i(x)$ and $\sum_{i=1}^n \lambda_i f_i(x)$ are mixture cdfs and densities, where F_i and f_i are component cdfs and densities, $\lambda_i \geq 0$ and $\sum_{i=1}^n \lambda_i = 1$.

3. Let X be a continuous random variable and F be its strictly increasing cdf. Prove that random variables $F(X)$ and $1-F(X)$ have the same distribution. [Hint: Prove that the cdf of $F(X)$ is x.]

4. (a) Plot the cdf in Example 2.1 with the continuity correction using a different color on the same plot. Does it make any difference? (b) Plot $(F(x_{(i+1)})-F(x_{(i)}))/(x_{(i+1)}-x_{(i)})$ versus $x_{(i)}$. Make a connection to density.

5. Let random variable X have cdf $F_X(x)$ and density $f_X(x)$. Find the cdf and density of $Y = a + bX$. [Hint: Consider cases when $b > 0$ and $b < 0$ separately.]

6. F is a cdf. Is F^p a cdf?

7. Let F and G be cdfs. Is their product a cdf? Does an analogous statement hold for densities? [Hint: If the answer is positive prove it, otherwise provide a counterexample.]

8. (a) Is it possible that for two cdfs $F(x) < G(x)$ for all x? (b) Is it possible that that for two densities, $f(x) < g(x)$? (c) Is it true that two cdfs always intersect? (d) Is it true that two pdfs always intersect?

9. If $f(x)$ is a density with support on the entire line, is $g(x) = f(x^2)$ a density? The same question but support of f is positive numbers.

10. The density of a random variable is defined as $c \times \sin x$ for $0 < x < \pi$ and 0 elsewhere. Find c and the cdf.

2.2 Mean, variance, and other moments

The mean and variance of a continuous random variable are defined as for a discrete distribution with the sum replaced by an integral. The mean of a continuous distribution with density $f(x)$ is defined as

$$\mu = \int_{-\infty}^{\infty} x f(x) dx, \tag{2.5}$$

and the variance is defined as

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx.$$

We often shorten the notation for mean and variance to $E(X)$ and $\text{var}(X)$, respectively. Sometimes, we use the notation for the standard deviation, $\text{SD}(X) = \sigma$. The following formula describes the relationship between the mean and variance,

$$\sigma^2 = E(X^2) - \mu^2, \tag{2.6}$$

as the counterpart of (1.7). Since variance is nonnegative, we conclude

$$|\mu| \leq \sqrt{E(X^2)}. \tag{2.7}$$

The mean, μ, can be interpreted as the center of gravity of a stick with mass density $f(x)$, similar to the discrete case as discussed in Section 1.4.1. By definition, μ is the center of gravity defined as the point at which the resultant torque is zero,

$$\int_{-\infty}^{\infty} (x-\mu) f(x) dx = 0,$$

which leads to definition (2.5).

Example 2.3 ***Expectation via cdf.*** *Prove that*

$$E(X) = -\int_{-\infty}^{0} F(x) dx + \int_{0}^{\infty} (1 - F(x)) dx \tag{2.8}$$

where F is the cdf of X.

*Proof***.** We have

$$\int_{-\infty}^{\infty} x f(x) dx = \int_{-\infty}^{0} x f(x) dx + \int_{0}^{\infty} x f(x) dx.$$

Using integration by parts in the first integral, define $u = x, dv = f(x)dx$, which implies $du = dx$ and $v = F(x)$. Consequently, the first integral can be expressed as

$$\int_{-\infty}^{0} x f(x) dx = xF(x)|_{-\infty}^{0} - \int_{-\infty}^{0} F(x) dx = -\int_{-\infty}^{0} F(x) dx.$$

In the second integral, let $y = -x$. That implies

$$\int_{0}^{\infty} x f(x) dx = -\int_{-\infty}^{0} y f(-y) dy.$$

Use integration by parts again by letting $u = y, dv = f(-y)dy$, which implies $du = dy$ and $v = 1 - F(-y)$. Consequently,

$$\int_{-\infty}^{0} yf(-y)dy = y(1-F(-y))|_{-\infty}^{0} - \int_{-\infty}^{0}(1-F(-y))dy = -\int_{0}^{\infty}(1-F(x))dx.$$

Combining the two integrals yields

$$\int_{-\infty}^{\infty} xf(x)dx = -\int_{-\infty}^{0} F(x)dx + \int_{0}^{\infty}(1-F(x))dx.$$

The identity (2.8) is proved. □

When two random variables are measured on different scales, their variation is easier to compare using the coefficient of variation (CV):

$$\mathrm{CV} = \frac{\sigma}{\mu}.$$

Sometimes CV is expressed as a percent to eliminate units. Indeed, CV does not change when applying the transformation $X \to \alpha X$, where α is a positive coefficient. This coefficient naturally emerges in the lognormal distribution; see Section 2.11.

The mean and variance may not exist for certain distributions. For example, they do not exist for the Cauchy distribution defined by the density

$$f(x) = \frac{1}{\pi}\frac{1}{1+x^2}, \quad -\infty < x < \infty. \tag{2.9}$$

Indeed, the integral $\int_{-\infty}^{\infty} x/(1+x^2)dx$ does not exist.

The kth noncentral (ν_k) and central (μ_k) moments are defined as

$$\nu_k = \int_{-\infty}^{\infty} x^k f(x)dx, \quad \mu_k = \int_{-\infty}^{\infty}(x-\mu)^k f(x)dx.$$

In another notation,

$$\nu_k = E(X^k), \quad \mu_k = E((X-\mu)^k).$$

The first noncentral moment is the mean and the second central moment is the variance. We say that the distribution is symmetric if the density is an even function around μ, i.e. $f(x-\mu) = f(\mu - x)$. For symmetric distributions all odd central moments are zero and the mean and mode coincide.

To characterize how skewed a distribution is, we use the *skewness* coefficient:

$$\text{Skewness} = \frac{\mu_3}{\sigma^3}. \tag{2.10}$$

This coefficient is scale independent, so random variables can be compared on the relative scale. If the skewness coefficient is less than zero, we say that the

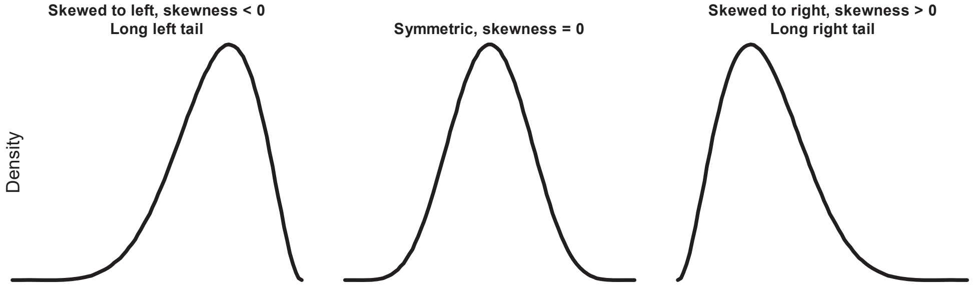


Figure 2.3: *Three types of distributions with respect to the skewness.*

distribution is left skewed (long left tail); if the skewness coefficients greater than zero, we say that the distribution is right skewed (long right tail; see Figure 2.3). The skewness coefficient is zero for a symmetric distribution, but skewness = 0 does not imply that the distribution is symmetric.

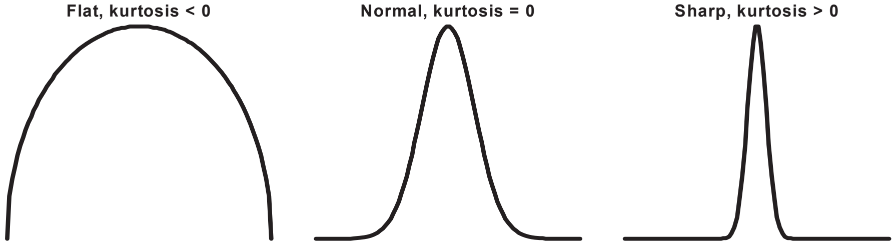


Figure 2.4: *Three types of distributions with respect to kurtosis.*

Kurtosis is used to characterize how flat the distribution is and is computed by the formula

$$\text{kurtosis} = \frac{\mu_4}{\sigma^4} - 3. \tag{2.11}$$

Alternatively, we may say that kurtosis characterizes the sharpness of the density. This coefficient is scale independent as well. The 3 is subtracted to make the normal distribution the reference with kurtosis = 0. For densities with a sharp peak, the kurtosis is positive; for flat densities, the kurtosis is negative. All three cases are illustrated in Figure 2.4.

We calculate skewness and kurtosis for some continuous distributions later in this chapter.

The expectation is defined for any function g of the random variable as

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

Of course, the expectation may not exist.

Example 2.4 ***Jensen's inequality.*** *If f is a convex function ($f'' \geq 0$), then*

$$E(f(X)) \geq f(E(X)). \tag{2.12}$$

If f is a concave function ($f'' \leq 0$), then

$$E(f(X)) \leq f(E(X)).$$

Solution. Since f is convex, we have $f(x) \geq f(x_0) + (x - x_0)f'(x_0)$ for all x. Replace x with X and let $x_0 = \mu = E(X)$:

$$f(X) \geq f(\mu) + (X - \mu)f'(\mu). \tag{2.13}$$

Taking the expectation of both sides, we obtain

$$E(f(X)) \geq E(f(\mu)) + E(X - \mu)f'(\mu) = f(E(X)).$$

The inequality is proved. We prove the inequality similarly for the concave function.

Remark 2.5 *If function f is strictly convex, i.e. $f'' > 0$ and $\Pr(X = \mu) < 1$ (random variable is not a constant), then (2.12) turns into a strict inequality. This follows from the fact that (2.13) turns into a strict inequality for $X \neq \mu$.*

Examples (all functions f are strictly convex):

1. $f(x) = x^2 : E(X^2) > E^2(X)$.

2. $f(x) = \ln x : E(\ln X) < \ln E(X)$ if $X > 0$.

3. $f(x) = 1/x : E(1/X) > 1/E(X)$ if $X > 0$.

Example 2.6 ***Cauchy inequality.*** *Prove that for any random variables X and Y we have*

$$E^2(XY) \leq E(X^2)E(Y^2), \tag{2.14}$$

and the equality holds if and only if $Y = \alpha X$ where α is a constant (all expectations are finite).

Solution. Define $Z = Y - \alpha X$ and express the second noncentral moment through α as

$$E(Z^2) = E(Y^2) - 2\alpha E(XY) + \alpha^2 E(X^2).$$

The right-hand side is a quadratic function of α. Since $E(Z^2) \geq 0$, the discriminant should be nonpositive, $E^2(XY) - E(X^2)E(Y^2) \leq 0$. The equality is true if and only if the discriminant is zero, which happens if $Y = \alpha X$ with probability 1. □

We make several comments: The Cauchy inequality is sometimes called the Cauchy–Schwarz inequality. The Cauchy inequality holds for discrete random variables as well. Inequality (2.14) can be expressed in terms of any functions f and g:

$$E^2(f(X)g(Y)) \leq E(f^2(X))E(g^2(Y)). \tag{2.15}$$

One may just think of $f(X)$ and $g(Y)$ as new random variables. Letting $g(Y) = 1$, we get

$$|E(f(X))| \leq \sqrt{E(f^2(X))},$$

an analog of inequality (2.7). Also, it is easy to apply the Cauchy inequality to random variables with the means subtracted:

$$E^2[(X - \mu_X)(Y - \mu_Y)] \leq \sigma_X^2 \sigma_Y^2. \tag{2.16}$$

This inequality proves the correlation coefficient takes values within the interval $[-1, 1]$. Note that this inequality turns into an equality if and only if random variables are proportionally related, $Y - \mu_Y = \beta(X - \mu_X)$.

The Cauchy inequality can be rewritten in discrete or integral fashion. The discrete version is well known in linear algebra as the inequality between the scalar product and the squared norm,

$$(\mathbf{x}'\mathbf{y})^2 \leq \|\mathbf{x}\|^2 \|\mathbf{y}\|^2,$$

where $\mathbf{x}'\mathbf{y}$ is the scalar (inner, dot) product, or in the index form

$$\left(\sum_{i=1}^n x_i y_i\right)^2 \leq \left(\sum_{i=1}^n x_i^2\right)\left(\sum_{i=1}^n y_i^2\right). \tag{2.17}$$

The integral version of the inequality is obtained by expressing the expected values in (2.15) via integrals,

$$\left(\int_{-\infty}^{\infty} f(x)g(x)dx\right)^2 \leq \int_{-\infty}^{\infty} f^2(x)dx \times \int_{-\infty}^{\infty} g^2(x)dx$$

provided the integrals on the right-hand side exist. This inequality can be proved using the same method as in Example 2.6.

Under mild conditions, it can be proven that a distribution is uniquely defined by all its moments. We can demonstrate this fact on a discrete random variable: If a random variable takes n distinct values x_i, the noncentral moments ν_k for $k = 1, 2, ..., n - 1$ uniquely define probabilities p_i. We want to prove that the system of linear equations given by $n - 1$ equations,

$$\sum_{i=1}^n p_i x_i^k = \nu_k,$$

has a unique solution for p_i. Note that we have $n-1$ equations because $\sum_{i=1}^n p_i = 1$. This system of equations has a unique solution for p_i because its determinant, called the Vandermonde determinant, is not zero.

The following result illustrates how instrumental calculus is in probability and statistics.

Example 2.7 ***Stein's identity.*** *Let f be a differentiable density of a continuous random variable X (without loss of generality it is assumed that the support is the entire line) and g be a differentiable function on $(-\infty, \infty)$ such that $E(g(X))$ is finite. Prove that*

$$E_X\left[g(X)(\ln f(X))' + g'(X)\right] = 0,$$

where $'$ means derivative.

Solution. Represent the expectation as an integral. Since $(\ln f(x))' = f'(x)/f(x)$, we obtain

$$\int_{-\infty}^{\infty} [g(x)(\ln f(x))'dx + g'(x)]f(x)dx = \int_{-\infty}^{\infty} g(x)f'(x)dx + \int_{-\infty}^{\infty} g'(x)f(x)dx.$$

Now we apply integration by parts, $\int_{-\infty}^{\infty} udv = uv|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} vdu$, by letting $u = g(x)$ and $dv = f'(x)dx$, so that $du = g'(x)dx$ and $v = f(x)$. This implies

$$\int_{-\infty}^{\infty} g(x)f'(x)dx = g(x)f(x)|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} g'(x)f(x)dx.$$

Since $E(g(X)) < \infty$, we have $g(x)f(x)|_{-\infty}^{\infty} = 0$. Therefore, the Stein's identity follows.

2.2.1 Quantiles, quartiles, and the median

The pth quantile, q_p, is the solution to the equation

$$F(q_p) = p, \quad 0 < p < 1,$$

where F is the cdf. In the language of the inverse cdf, we can define the pth quantile as $q_p = F^{-1}(p)$. The $\frac{1}{4}$ quantile is called the lower (or the first) quartile and is denoted as $q_{1/4}$ and the $\frac{3}{4}$th quantile is called the upper (or the third) quartile and is denoted as $q_{3/4}$. The median is the $\frac{1}{2}$ quantile (or the second quartile):

$$F(\text{median}) = \frac{1}{2}. \tag{2.18}$$

Percentile is a quantile expressed as a percent: 75th percentile, 25th percentile, etc. Quantiles and quartiles can be used to characterize the range of the distribution. We may characterize the range of the distribution using $q_{1/4}$ and $q_{3/4}$, also called the interquartile range, which indicates that 50% of values fall between them. In other words, the interval $[q_{1/4}, q_{3/4}]$ contains 50% of the random variable values.

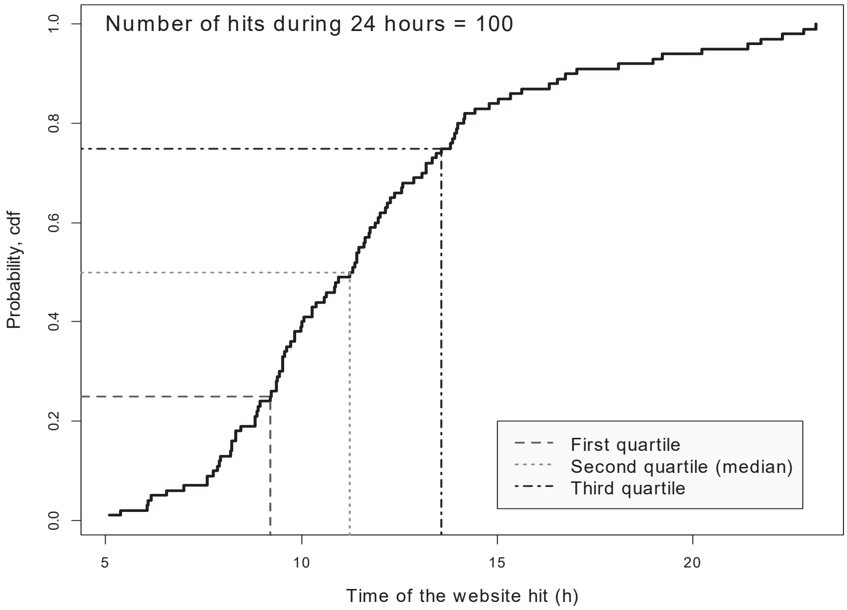


Figure 2.5: *Quartiles for the webhits data.*

Example 2.8 ***Web hits.*** *Compute and display $q_{1/4}, q_{1/2}$, and $q_{3/4}$ for the web hits example. Find the interquartile range.*

Solution. The requested quantities are simply computed as `x[n/4]`,`x[n/2]`, and `x[3*n/4]`, the `R` function is in the file `webhitsQ.r`. The interquartile range interval is `[1] 9.21461 13.56446.` About 50% of the web hits happen between 9:30 a.m. and 2 p.m. The `R` function `webhitsQ` produces Figure 2.5.

2.2.2 The tight confidence range

Quantification of the range of a random variable that represents the general population is an important task of applied statistics. In many applications, the standard deviation, σ, serves as a characteristic of the range, and sometimes authors use $\mu \pm \sigma$ to report the scatter of the random variable. Using the interval $\mu \pm k\sigma$, where k is a constant, silently assumes that data are symmetric around the mean. When a distribution is not symmetric, a symmetric interval around the mean is not optimal because the same coverage probability can be obtained using an interval with a smaller width. Thus, we arrive at the concept of the *tight confidence range*. We emphasize the difference between confidence range and confidence interval in statistics that covers the true unknown parameter, see Section 7.8.

Definition 2.9 *The pth tight confidence range for a random variable X is the interval $[t,T]$ such that (a)* $\Pr(t < X \leq T) = p$, *and (b)* $T - t = \min$.

The pth tight confidence range is the shortest interval that contains values of the random variable with probability p. For example, the interquartile interval contains 50% of the values but may be not optimal. On the other hand, the 50% tight confidence range will be optimal because among all intervals which contain 50% of the data it is the narrowest. The following theorem states that the density values at the ends of the tight confidence range must be the same.

Theorem 2.10 *(a) If X has a continuous and differentiable unimodal density f (the mode belongs to the open support) and $[t,T]$ is the pth tight confidence range, then*

$$f(t) = f(T). \tag{2.19}$$

(b) The pth tight confidence range contains the mode, and when $p \to 0$, this interval shrinks to the mode.

Proof. (a) By the definition of the pth tight range interval ($t < T$), we must have

$$\int_{-\infty}^{T} f(x)dx - \int_{-\infty}^{t} f(x)dx = p. \tag{2.20}$$

Find t and T by solving the optimization problem $T - t = \min$ under constraint (2.20). Introduce the Lagrange function

$$\mathcal{L}(t,T;\lambda) = T - t - \lambda\left(\int_{-\infty}^{T} f(x)dx - \int_{-\infty}^{t} f(x)dx - p\right),$$

where λ is the Lagrange multiplier. The necessary conditions for the minimum are

$$\frac{\partial\mathcal{L}}{\partial t} = -1 + \lambda f(t) = 0, \quad \frac{\partial\mathcal{L}}{\partial T} = 1 - \lambda f(T) = 0.$$

These equations imply (2.19). Since f is a unimodal equation $f(t) = c$ has two solutions, t and T, for each $c \in (0, \max f(x))$. (b) Interval $[t,T]$ contains the mode for each $p > 0$ and it shrinks to the mode because f is a continuous function. □

The above theorem helps compute the pth tight confidence range. To find t and T, we need to solve a system of equations:

$$F(T) - F(t) = p, \quad f(t) = f(T), \tag{2.21}$$

where F is the cdf. We will illustrate computation of the tight confidence range later in the chapter. For symmetric distributions, this interval takes the regular form $\mu \pm k\sigma$. The tight confidence range is especially advantageous for asymmetric distributions, such as the gamma distribution in Section 2.6 or the lognormal distribution in Section 2.11.

The pth tight confidence range can be defined for discrete distributions as well as (t, T) such that $\Pr(t < X \leq T) \geq p$ and $T - t = \min$. Note that we use the inequality sign here because one cannot find t and T such that $\Pr(t < X \leq T) = p$ due to discreteness of the distribution. Since the probability is a discrete function, we cannot rely on calculus. Instead, we use direct computation to find t and T, as in the following example.

Example 2.11 ***Confidence range for the binomial distribution.*** *Write an* `R` *code to compute the* $100\lambda\%$ *tight confidence range for the binomial random variable.*

Solution. The `R` code is found in the text file `tr.binom.r`; see Section 1.5.3 to read the file. To find the optimal range, we use double `for` loop over `m` and `M`. If, for certain m and M, we have $\Pr(m < X \leq M) \geq \lambda$, we check if the length of the interval is smaller than the current one. Specifically, if `minDIF`. If `M-m<minDIF`, we save m and M as `m.opt` and `M.opt`, respectively. For example, the call

```
> tr.binom(n=100,p=.3,lambda=.75)
[1] 23.0000000 34.0000000 0.7616109
```

gives the 75% range $(23, 34)$, which has the minimal width among all intervals that contain the binomial random variable with probability equal to or greater than λ. In fact, the range $(23, 34)$ gives the probability $0.762 > 0.75$. □

In applied statistics, we shall use the confidence range to determine the range of individual values as opposed to the traditional confidence interval that is intended to determine the range of an unknown parameter, such as the mean, μ. Example 2.29 illustrates the computation of the tight confidence range for a distribution of house prices on the real estate market.

Problems

1. Prove that $\text{var}(X) \leq E(X^2)$. Is it true that $E(X - \mu)^k \leq E(X^k)$ for any even positive integer k? Specify distributions for which this is true.

2. (a) Prove that $|E(X)| \leq E^{1/p}(|X|^p)$ for $p \geq 1$. (b) Prove that $E(X^3) \geq E^3(X)$ if $X > 0$.

3. (a) Prove formula (2.6). (b) Derive (2.6) from the Jensen's inequality.

4. The central moments can be expressed via noncentral moments using polynomial expansion of $(x-\mu)^k$. Provide a recursive formula for the coefficients of $E(X - \mu)^{k+1}$ as a linear combination of $\nu_1, ..., \nu_{k+1}$. [Hint: Use formula (1.9).]

5. Prove that skewness and kurtosis are scale independent. In other words, skewness and kurtosis are the same for X and aX, where $a > 0$.

6. Prove that the cdf of the Cauchy distribution with density (2.9) is $\frac{1}{\pi}\arctan x + \frac{1}{2}$. Find the median and the $\frac{1}{4}$ and the $\frac{3}{4}$ quantiles. Show that they are symmetric around zero. Define the *shifted* Cauchy distribution and find its cdf.

7. The density of a random variable is defined as $c(1-x^2)$ for $|x| \leq 1$ and 0 elsewhere. Find c, the cdf, the mean, and the median. Plot the density and cdf in `R` using `mfrow=c(1,2)`.

8. Express the fourth central moment as a linear combination of noncentral moments. Express the fourth noncentral moment as a linear combination of central moments. [Hint: Use formula (1.9).]

9. The support of X is $(-1,1)$. Is it true that the kth central moments vanish when $k \to \infty$? Is it true for noncentral moments?

10. Prove that the Poisson distribution is skewed to the right. [Hint: Prove that the third central moment is positive.]

11. Prove that $E\left(g^p(X)\right)$ is a convex function of $p > 0$ where g is a nonnegative function. Derive the respective Jensen's inequality.

12. $p > 0$ and f is a pdf. Is f^p a pdf? [Hint: Use Jensen's inequality.]

13. Plot in `R` two densities for which median is greater that mode and mode is greater than median. Describe the distributions for which this is true using the language of *tail*.

14. Prove that $E(F(X)) = 1/2$ where F is the cdf of X. [Hint: Assume continuous distribution and apply change of variable.]

15. Prove that the minimum of $\int_{-\infty}^{a} F(x)dx + \int_{a}^{\infty}(1-F(x))dx$ attains when a is the median.

16. Let a continuous random variable have a positive continuously differentiable density with unique maximum, or specifically, $(\ln f)' = 0$ has unique solution, the mode. Is it true that the following inequality holds: mode $\leq$ median $\leq$ mean? Prove that this inequality turns into equality for symmetric distributions. Is it true that this inequality holds for densities such that $(\ln f)'' < 0$? More about this inequality can be learned from Abadir [1].

17. Prove that, for a symmetric distribution, the pth tight confidence range takes the form $\mu \pm k\sigma$.

18. Write an `R` program to compute the $100\lambda\%$ tight population confidence range for the Poisson distribution similarly to Example 2.11.

2.3 Uniform distribution

The uniform (or rectangular) distribution is the simplest continuous distribution. Following its name, the uniform distribution is categorized by the fact that the probability of falling within an interval is proportional to its length. We write $X \sim \mathcal{R}(a, b)$ to indicate that random variable X has a uniform distribution on the finite interval (a, b). Sometimes, one can encounter the notation U instead of $\mathcal{R}$. The density for this distribution is constant and the cdf is a linear function (see Figure 2.6), namely,

$$f(x) = \begin{cases} \frac{1}{b-a} \text{ if } a \leq x \leq b \\ 0 \text{ elsewhere} \end{cases}, \quad F(x) = \begin{cases} \frac{x-a}{b-a} \text{ if } a \leq x \leq b \\ 0 \text{ if } x < a \\ 1 \text{ if } x > b \end{cases}.$$

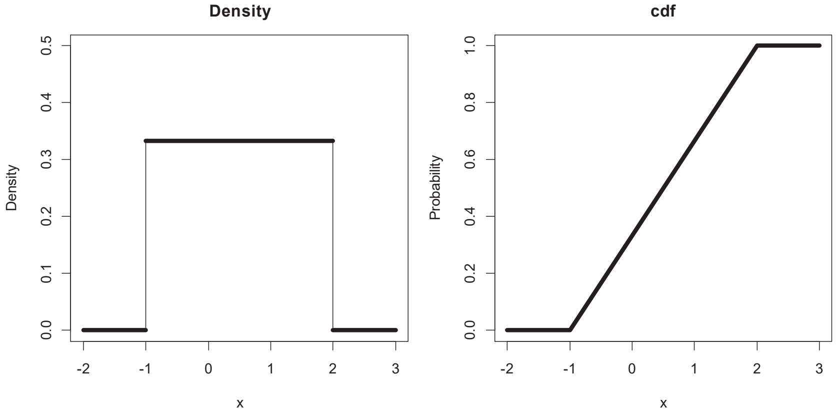


Figure 2.6: *Density and cdf for $X \sim \mathcal{R}(-1, 2)$. Sometimes this distribution is called* ***rectangular*** *because of the shape of the density.*

The mean is $E(X) = (a+b)/2$. Although the result seems obvious, we derive it formally via integration as follows:

$$E(X) = \frac{1}{b-a}\int_a^b x dx = \frac{1}{b-a}\frac{1}{2}x^2\bigg|_a^b = \frac{1}{b-a}\frac{1}{2}(b^2 - a^2) = \frac{a+b}{2}.$$

To be specific, we indicate the distribution of X as a subscript in the mathematical expectation:

$$E_{X\sim\mathcal{R}(a,b)}(X) = \frac{1}{2}(a+b). \tag{2.22}$$

By integration over x^2 and using formula (2.6), we obtain the variance and SD:

$$\text{var}(X) = \frac{(b-a)^2}{12}, \quad \text{SD}(X) = \frac{b-a}{2\sqrt{3}}.$$

The uniform distribution on (a, b) can be derived from the standard uniform distribution on $(0, 1)$ by simple linear transformation. Symbolically we write $\mathcal{R}(a, b) = a+(b-a)\mathcal{R}(0, 1)$. To generate n independent observations from $\mathcal{R}(a, b)$, use `runif(n,min=a,max=b)` where the default values are `a=0` and `b=1`.

Example 2.12 ***Waiting time.*** *The shuttle bus comes every 30 minutes. What is the expected waiting time if you come to the bus stop at random time? Use simulations to confirm your answer.*

Solution. Clearly, the wait time is less than 30 minutes. The arrival time can be modeled as a random variable X distributed as $\mathcal{R}(0, 30)$. As follows from (2.22),

$$\text{the expected waiting time} = E(X) = 30/2 = 15 \text{ minutes.}$$

To simulate we imagine that the bus comes at 0:30, 1, 1:30,..,24:00 or on the minute scale $m = \{30, 60, 90, ..., 60 \times 24\}$ and your arrival can be modeled as the uniform distribution, $X \sim \mathcal{R}(0, 1440)$. Then the average arrival time is the average distance to the next point in m to the right, which can be computed using `ceiling` command: `X=runif(100000,min=0,max=1440);mean(30 *ceiling((X/30))-X)` with the output `15.00843`.

Example 2.13 ***Broken stick.*** *A stick of unit length is broken at random. What is the probability that a longer piece is more than twice the length of the shorter piece. Provide a theoretical answer and confirm it by simulations.*

Solution. The point where the stick is broken is uniformly distributed along its length, $X \sim \mathcal{R}(0, 1)$. The length of the longer piece, $X_L > 0.5$, and therefore $X_L \sim \mathcal{R}(0.5, 1)$. Since the shorter piece has length $1 - X_L$, the length must be such that $X_L > 2(1 - X_L)$, i.e. $X_L > 2/3$. Using the fact that $X_L \sim \mathcal{R}(0.5, 1)$, the asked probability is $(1 - 2/3) \times 2 = 2/3$.

The `R` code that estimates the probability via simulations is shown below.

```
longpiece=function(nExp=100000)
{
 dump("longpiece","c:\\StatBook\\longpiece.r")
 X=runif(nExp)
 X.long=X #initialization
 X.long[X<0.5]=1-X[X<0.5] # if X is short take the other part
 X.short=1-X.long # by definition
 pr=mean(X.long>2*X.short) # proportion of experiments
 pr
}
```

This function gives the probability `0.66858`, very close to the theoretical answer. We make several comments on the code: (i) The default number of experiments/simulations is 100K. (ii) It is a good idea to save the code as a text file in a safe place. In fact, it is easy to lose code in `R`: you may forget to save the workspace on exit, your code may be overwritten by another program, etc.. The `dump` command saves the program as a text file under the name `c:\\StatBook\\longpiece.r`. (iii) `X.long[X<0.5]` means that only components of `X.long` for which `X<0.5` are used. The same is true for the right-hand side of the respective line. (iv) Command `X.long>2*X.short` creates a logical vector with `TRUE` if the inequality is true and `FALSE` otherwise. When a numeric operator such as `mean` is applied, `TRUE` is replaced with `1` and `FALSE` is replaced with `0`, so that `mean` returns the proportion when the inequality is true.

Alternatively, one can compute the longest and the shortest piece using commands `pmax()` and `pmin()`. These commands compute maximum and minimum in a vectorized fashion, so no loop is required. For example, if `v1`, `v2`, and `v3` are vectors of the same length, `pmax(v1,v2,v3)` returns a vector of the same length with the ith component equal to the maximum of ith components of these vectors. Then the longest and the shortest piece are computed as `X.long=pmax(X,1-X)` and `X.short=pmin(X,1-X)`, respectively.

Example 2.14 ***Kurtosis for the uniform distribution.*** *Find the kurtosis of the uniform distribution on* $(0,1)$.

Solution. The formula for the kurtosis is given by (2.11). Find the fourth central moment:

$$\mu_4 = \int_0^1 (x-0.5)^4 dx = \int_{-1/2}^{1/2} z^4 dz = 2\int_0^{1/2} z^4 dz = \frac{1}{5\times 2^4}.$$

But $\sigma^2 = 1/12$ so that kurtosis $= 12^2/5 \times 2^4 - 3 = -1.2$, meaning that the distribution is flatter than normal, as expected. □

Below we use the uniform distribution for raison cookies from Example 1.19.

Example 2.15 ***Raisins in a cookie simulations.*** *Use simulations to confirm that the probability that one cookie has all* n *raisins is* $m^{-(n-1)}$.

Solution. Since the probability that there is one raisin in the cookie is $1/m$ the probability that a particular cookie has all the raisins is m^{-n}. Then, the probability that one cookie has all the raisins is $m \times m^{-n} = m^{-(n-1)}$. The simulations presented in the `R` code `simCookie` confirm this probability. We put m cookies side by side and replace them with unit length segments $[0,1], [1,2], .., [m-1,m]$. Then we throw n raisins on the segment $[0,m]$ at random and observe if all raisins fell into one unit segment; repeat this process `nSim` times (see function `simCookie`). The variable `pr` counts the number of simulations when one cookie has n raisins. The call `simCookie()` with default values `n=3,m=4` gives `0.06304` and the analytical answer is $4^{-(3-1)} = 0.0625$.

Problems

1. Derive a formula for the variance of the uniform distribution. [Hint: Derive the noncentral second moment and then use formula (2.6).]

2. Find the median and the first and third quartiles of the uniform distribution. Use `qunif` function in `R` to test your answers using $a = -2$ and $b = 3$.

3. Use `pmax` and `pmin` in function `longpiece` to compute the longest and the shortest piece, as suggested above. Test that the new version yields the same answer.

4. Demonstrate that the distribution of the long and short pieces in the previous problem is uniform by plotting the empirical cdfs on one graph side by side using `par(mfrow=c(1,2))`. Prove that the distributions are uniform. [Hint: Use the fact that $\max(X, 1-X) \leq y$ is equivalent to $X \leq y$ and $1 - X \leq y$.]

5. Two ways to get the shortest piece of a stick with two random breaks are suggested: (1) break the first time at random and brake the shortest piece at random again, taking the shortest piece and (2) break the stick at two random places and take the shortest piece. What way produces a shorter piece on average? Modify the `longpiece` function to get the answer via simulations. [Hint: Use `pmin`; find the shortest piece in the second method as $\min(x_1, x_2, 1-x_1, 1-x_2, |x_1 - x_2|)$.]

6. (a) Plot the empirical cdf of the shortest piece using the two methods in the previous problem on the same graph. Use different colors and `legend`. Are the distributions uniform? (b) Prove that $X_L \sim \mathcal{R}(0.5, 1)$ by finding the theoretical cdf.

7. Find the kurtosis for $\mathcal{R}(a, b)$.

8. Two friends come to the bus stop at random times between 1 and 2 p.m. The bus arrives every 15 minutes. Use simulations to estimate the probability that the friends end up on the same bus.

9. Modify function `simCookie` to estimate that a particular cookie (your cookie) gets all the raisins, derive an analytical answer and compare the results. [Hint: Use `cookie.id` as an argument in the `R` function to specify your cookie.]

10. Besides m raisins, l chocolate chips are added to the dough. Use simulations to answer the following questions: (a) What is the probability that at least one cookie gets all raisins and all chocolate chips? (b) What is the probability that a particular cookie gets all the chips but no raisins?

2.4 Exponential distribution

The exponential distribution is commonly used for modeling waiting time and survival analysis. It describes the distribution of a positive continuous random variable with the cdf defined as $F(x;\lambda) = 1 - \mathrm{e}^{-\lambda x}$ for $x \geq 0$ and 0 for $x < 0$, where λ is a positive parameter, typically referred to as the *rate*. See Example 2.18 below for why λ is called the rate. The density of the exponential distribution is derived through differentiation:

$$f(x;\lambda) = \lambda \mathrm{e}^{-\lambda x}, \quad x \geq 0. \tag{2.23}$$

We use notation $X \sim \mathcal{E}(\lambda)$ to indicate that X has an exponential distribution with parameter λ. We bring the reader's attention to the notation: parameter λ is separated from the argument by a semicolon. The indication that the distribution depends on λ is especially important in statistics where we estimate λ using a sample of observations from this distribution. The density is a decreasing function of x; see Figure 2.7.

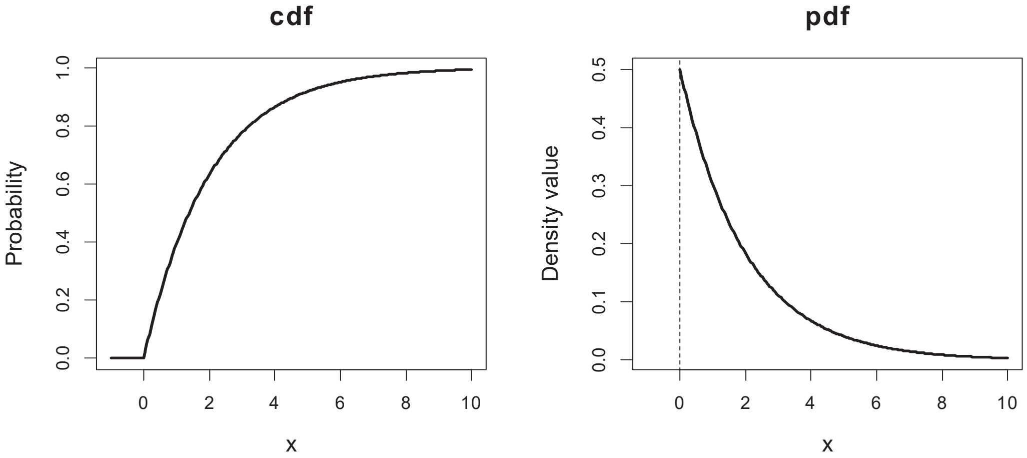


Figure 2.7: *The cdf and pdf of an exponential random variable with the rate parameter $\lambda = 1/2$. The former is computed in* `R` *as* `pexp` *and the latter as* `dexp`*.*

The exponential distribution frequently emerges in survival analysis. Then we interpret $F(x)$ as the proportion of dead by time x, and the complementary function $S(x) = 1 - F(x)$ is interpreted as the survival function, so that $F(x + \Delta x) - F(x)$ is the proportion who died between x and $x + \Delta x$. The mortality rate is defined as $(F(x + \Delta x) - F(x))/\Delta x$, the death rate per unit of time. On the scale of survivors, this rate takes the form

$$\frac{(F(x + \Delta x) - F(x))/\Delta x}{1 - F(x)},$$

the death rate per unit of time per proportion of alive at time x. Letting Δx go to zero we arrive at the definition of the *hazard function,*

$$H(x) = \frac{F'(x)}{1 - F(x)} = \frac{-S'(x)}{S(x)},$$

as the instantaneous relative mortality rate: the proportion of dead among survivors. For the exponential distribution, the hazard function is constant $H(x) = \lambda \mathrm{e}^{-\lambda x}/\mathrm{e}^{-\lambda x} = \lambda$. In words, exponential distribution yields a constant hazard function.

To find the mean of the exponential distribution, we apply integration by parts, $\int u dv = uv - \int v du$, which after letting $u = x$ and $dv = \mathrm{e}^{-\lambda x} dx$ gives

$$E(X) = \lambda \int_0^\infty x\mathrm{e}^{-\lambda x} dx = -\lambda \times \frac{1}{\lambda}\mathrm{e}^{-\lambda x}\bigg|_0^\infty + \frac{\lambda}{\lambda}\int_0^\infty \mathrm{e}^{-\lambda x} dx = \frac{1}{\lambda}.$$

An alternative parametrization uses $\theta = 1/\lambda$, so that the density takes the form $f(x;\theta) = \theta^{-1}\mathrm{e}^{-x/\theta}$. Parameter θ is called the *scale* parameter. An advantage of this parametrization is that θ has the same scale unit as X, and, moreover, $E(X) = \theta$. The rate parametrization is used in `R` by default.

Repeated integration by parts yields the variance

$$\mathrm{var}(X) = \frac{1}{\lambda^2}. \tag{2.24}$$

The exponential distribution is the simplest distribution for positive random variables and may be applied to model the occurrence of random events for which the probability drops monotonically. In particular, it may be applied to describe the time to an event after a meeting arrival, appointment, announcement, or project launch as in the following example.

Example 2.16 ***Telephone call.*** *Bill said that he will call after 10 a.m. Assuming that the time of his call follows an exponential distribution with parameter* $\lambda = 1/10$, *which decision maximizes your probability of talking with Bill: (a) wait the first 10 minutes, or (b) wait for the call from 10:10 to 11:00?*

Solution. Let X denote the time elapsed before Bill calls. It is believable that the density of calls is decreasing with time such that the maximum density occurs right after 10 a.m. ($X = 0$). Consequently, the exponential distribution is a good candidate in this case. Specifically, the cdf takes the form

$$F(x) = \Pr(\text{Bill calls within } (10, 10+x) \text{ time interval}) = \Pr(X \le x) = 1 - \mathrm{e}^{-x/10}.$$

The first probability (a) is $\Pr(X \le 10) = F(10) = 1 - \mathrm{e}^{-10\times 0.1} = 1 - 1/\mathrm{e} = 0.632$. The second probability (b) is $\Pr(10 < X \le 60) = F(60) - F(10) = (1 - \mathrm{e}^{-0.1\times 60}) - (1 - \mathrm{e}^{-0.1\times 10}) = 0.365$. Therefore, it is better to wait the call first 10 minutes after 10 a.m. than to wait the 50 minutes after 10:10 a.m. □

We prove in Section 3.3 that the exponential distribution is memoryless. The following example illustrates how to use simulations with the exponential distribution.

Example 2.17 ***Safe turn wait time.*** *A truck requires 10 seconds to turn at an intersection onto a busy street. Assuming that the time between two passing cars follows an exponential distribution with the average time three seconds, what is the median time required to make a safe turn? Use simulations to find the answer.*

Solution. The `R` code is realized in function `truck.turn`. The arguments of the function with default values are `tr.tu=10`, `car.time=3`, `Nmax=1000`, `nSim=10000`. Simulations are carried out using the loop over `isim`. For each `isim` we generate an array of times between two passing cars using command `X=rexp(Nmax, rate=1/ car.time)` because the rate, λ, is the reciprocal of the mean; `Nmax` is the maximum number of cars (it should be big enough but is irrelevant). The command `cX=cumsum(X)` computes the cumulative time, the time elapsed after the truck comes to the intersection. Since the truck turns at the first 10-second gap in traffic, we compute this time as `nsec[isim]=min(cX[X>tr.tu])`. After all simulations are done, we plot the cdf and compute the median wait time using `median` command. With the default parameters, the truck must wait about one minute to make a safe turn. □

The following example shows the connection between the Poisson and exponential distributions.

Example 2.18 ***Bathroom break.*** *The number of customers arriving at a bank per time t follows the Poisson distribution $\mathcal{P}(\lambda t)$, where λ is the rate of customer flow. If the bank teller just served a customer and needs to take a bathroom break, (a) what is the distribution of time until the next customer walks in? (b) Does he/she have enough time to go to the bathroom if he/she needs five minutes and two customers arrive every 10 minutes on average?*

Solution. First of all, we explain why parameter λ can be interpreted as the rate of customer flow. Because if X denotes the number of customers walking in during time units t, we have $E(X) = \lambda t$. Therefore, $\lambda = E(X)/t$ is the number of expected customers per minute. (a) The probability that no customers arrive within t time units can be modeled as $\Pr(X = 0) = \frac{(\lambda t)^0}{0!}\mathrm{e}^{-\lambda t} = \mathrm{e}^{-\lambda t}$. Therefore, the cdf of time until the next customer walks in within t time units is the complementary probability, $F(t) = 1-\mathrm{e}^{-\lambda t}$, the cdf of the exponential distribution. (b) In our case, the time unit = 10 minutes, so the probability that no customer walks in within 5 minute bathroom break is computed by formula above using $\lambda = 2$ and $t = 1/2$: $\mathrm{e}^{-2/2} = \mathrm{e}^{-1} = 0.37$. □

The exponential distribution has an important connection with the chi-square distribution to be discussed later in Section 4.2.

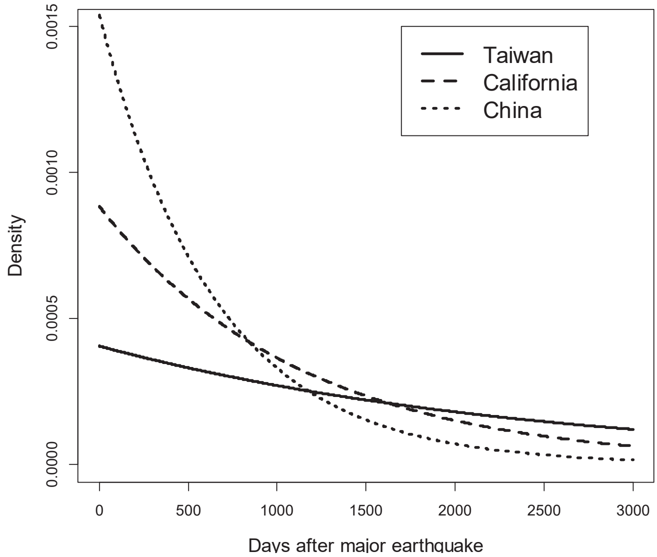


Figure 2.8: *The densities of time between two consecutive earthquakes in three regions.*

Example 2.19 ***Earthquake occurrence.*** *Three seismological active regions, namely, Taiwan, China, and California, have different patterns of earthquake occurrence. In particular, Wang and Kuo [101] confirmed that the interoccurrence (time between two consecutive earthquakes) nearly follows an exponential distribution with a scale parameter for the three regions of 2465, 649, and 1131, respectively. (a) Depict the three densities, (b) find the median and the third quartile (in years), and (c) compute the probability that no earthquake occurs within one year after a past quake.*

Solution. (a) The three densities are depicted in Figure 2.8. (b) The median is found from the equation $1 - \mathrm{e}^{T/\theta} = 0.5$ and the third quartile is found from equation $1 - \mathrm{e}^{T/\theta} = 0.75$. This quantile means the time between consecutive earthquake with probability 0.75. (c) The probability that no earthquakes occur within a year is $\mathrm{e}^{-365/\theta}$. The results are presented in the following table.

	Taiwan	China	California
Scale, θ	2465 days	649 days	1131 days
Rate, λ	0.148 years	0.562 years	0.322 years
Median	4.68 years	1.23 years	2.15 years
3d quartile	7.42 years	1.95 years	3.40 years
Pr(no earthquake within year)	0.862	0.570	0.724

2.4.1 Laplace or double-exponential distribution

The density of the *Laplace* or *double-exponential* (sometimes called biexponential) distribution is defined as

$$f(x;\lambda)=\frac{\lambda}{2}\mathrm{e}^{-\lambda|x|},\quad -\infty<x<\infty, \tag{2.25}$$

with notation $X\sim\mathcal{L}(\lambda)$.

The cdf is easy to obtain by integration:

$$F(x;\lambda)=\begin{cases}\frac{1}{2}\mathrm{e}^{\lambda x}\text{ for }x<0\\ 1-\frac{1}{2}\mathrm{e}^{-\lambda x}\text{ for }x\geq 0\end{cases},$$

see Figure 2.9. This distribution is symmetric, and therefore the mean, mode, and median are the same, 0. The density does not have a derivative at $x=0$. The variance is $2/\lambda^2$. The density of the shifted Laplace distribution is defined as $(\lambda/2)\mathrm{e}^{-\lambda|x-\mu|}$. This distribution is a good model for studying robust statistical inference; as we shall learn later (Section 6.11), this distribution gives rise to the median. See Example 6.161 where statistical properties of the mean and median are compared.

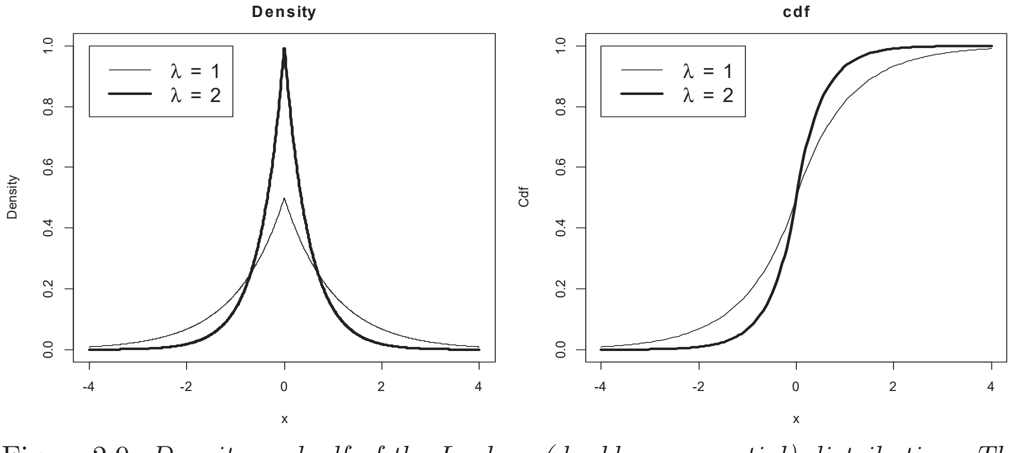


Figure 2.9: *Density and cdf of the Laplace (double-exponential) distribution. The greater* λ*, the higher the density peak, and the steeper the slope of the cdf at zero.*

2.4.2 `R` functions

The exponential distribution has a built-in `R` function. For example, `rexp(n=100, rate=0.5)` will produce 100 random numbers with density $\lambda\mathrm{e}^{-\lambda x}$ where $\lambda=0.5$. There are three other functions associated with the exponential distribution: `qexp`, `pexp`, and `dexp`. For instance, quantiles can be computed as `-log(1-p)/lambda`.

The Laplace distribution is symmetric; we can use this property to generate random numbers using the following two lines of code: `s=sample(x=c(-1,1), replace=T,size=n,prob=c(.5,.5)); x=s*rexp(n=n,rate=lambda)`.

Problems

1. Prove that if $X \sim \mathcal{R}(0,1)$, then $Y = -\ln X$ has an exponential distribution. Write `R` code to check your answer: simulate 100K uniformly distributed random numbers and plot the theoretical and empirical cdfs of Y on the same graph.

2. Prove that the only survival function with a constant hazard function is the exponential survival function, $S(x) = \mathrm{e}^{-\lambda x}$.

3. Justify why θ is called the scale parameter. [Hint: How does θ change when the scale of x changes from days to weeks, or from months to years?]

4. Show that the 0.5 tight confidence range for the exponential distribution is (0,median). [Hint: The conditions of Theorem 2.10 do not hold.]

5. Write an `R` program to do the following: (a) generate 10K observations from $\mathcal{E}(2)$ using the function `rexp`; (b) plot the empirical cdf and superimpose with the theoretical cdf using the `pexp` function; (c) compute and compare the first, second, and third quartiles using empirical and theoretical cdfs.

6. Use simulations to reject the conjecture that, for $X \sim \mathcal{E}(\lambda)$, $E\left|X - \lambda^{-1}\right|^k = \lambda^{-k}$. [Hint: Plot the "theoretical" and empirical moments on the same graph for $k = 2, 3, 4, 5$.]

7. In the telephone example, what is a better probability to talk with Bill: (a) wait the first 5 minutes or (b) wait for the call from 10:10 to 10:40?

8. Derive variance (2.24) using formula (2.6).

9. In Example 2.17, (a) estimate the average wait time and explain why it is greater than the median time, (b) estimate the probability that it takes less than 30 seconds for a safe turn, and (c) estimate the median wait time if the truck driver misses the first gap and turns on the second.

10. Compute the probability that two earthquakes happen within a week in each three regions.

11. Prove that if $X \sim \mathcal{L}(\lambda)$, then $E(X) = 0$ and $\mathrm{var}(X) = 2/\lambda^2$.

12. (a) Generate 10K random observations from a shifted Laplace distribution with $\lambda = 2$ and $m = -1$. (b) Plot the empirical and theoretical cdfs on the same graph.