Preface

Starting with the fundamentals of probability, this text leads readers to computer simulations and Monte Carlo methods, stochastic processes and Markov chains, queuing theory, statistical inference, and regression. These areas are heavily used in modern computer science, computer engineering, software engineering, and related fields.

For whom this book is written

The book is primarily intended for junior undergraduate to beginning graduate level students majoring in computer-related fields. At the same time, it can be used by electrical engineering, mathematics, statistics, actuarial science, and other majors for a standard calculus-based introductory statistics course. Standard topics in probability and statistics are covered in Chapters 1–4 and Chapters 8–10.

Graduate students can use this book to prepare for probability-based courses such as queuing theory, artificial neural networks, computer performance, etc.

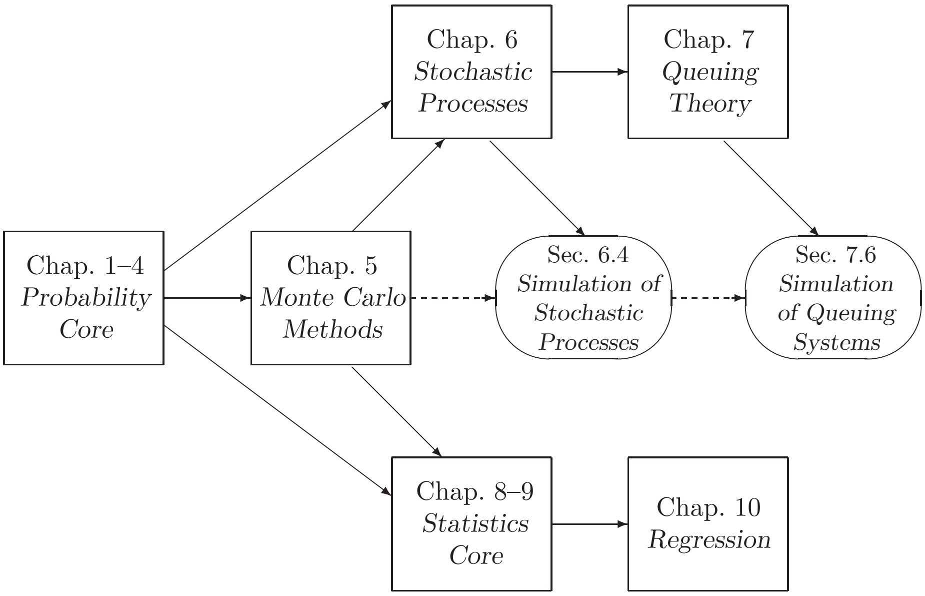
The book can also be used as a standard reference on probability and statistical methods, simulation, and modeling tools.

Recommended courses

The text is recommended for a one-semester course with several open-end options available.

After introducing probability and discrete and continuous distributions in Chapters 1–4, instructors may choose the following continuations.

Probability-oriented course. Proceed to Chapters 6–8 for Stochastic Processes, Markov Chains, and Queuing Theory. Computer science majors will find it attractive to supplement such a course with computer simulations and Monte Carlo methods. Students can learn and practice general simulation techniques in Chapter 5, then advance to the simulation of stochastic processes and rather complex queuing systems in Sections 6.4 and 7.6. Chapter 5 is highly recommended but not required for the rest of the material.



Statistics-emphasized course. Proceed to Chapters 8–10 directly after the probability core. Such a curriculum is more standard, and it is suitable for a wide range of majors. Chapter 5 remains optional but recommended; it discusses statistical methods based on computer simulations.

A course satisfying ABET requirements. Topics covered in this book satisfy ABET (Accreditation Board for Engineering and Technology) requirements for probability and statistics. To meet the requirements, instructors should choose topics from Chapters 1–10. All or some of Chapters 5–7 and 10 may be considered optional, depending on the program's ABET objectives.

Prerequisites, and use of the appendix

Working *differentiation and integration skills* are required starting from Chapter 4. They are usually covered in one semester of university calculus.

As a refresher, appendix has a very brief summary of the minimum calculus techniques required for reading this book (Section 11.3). Certainly, this section cannot be used to *learn* calculus "from scratch." It only serves as a reference and student aid.

Next, Chapters 6–7 and Sections 10.3–10.4 rely on very basic matrix computations. Essentially, readers should be able to multiply matrices, solve linear systems (Chapters 6–7) and compute inverse matrices (Section 10.3). A basic refresher with some examples is in the appendix, Section 11.4.

Style and motivation

The book is written in a lively style and reasonably simple language that students find easy to read and understand. Reading this book, students should feel as if an experienced and enthusiastic lecturer is addressing them in person.

Besides computer science applications and multiple motivating examples, the book contains related interesting facts, paradoxical statements, wide applications to other fields, etc. I expect prepared students to enjoy the course, benefit from it, and find it attractive and useful for their careers.

Computers, demos, illustrations, and MATLAB®

Frequent self-explaining figures help readers understand and visualize concepts, formulas, and even some proofs. Moreover, instructors and students are invited to use included short programs for *computer demonstrations*. Randomness, behavior of random variables, convergence results such as the Central Limit Theorem, and especially simulations can be nicely visualized by animated graphics.

These small computer codes contain very basic and simple MATLAB commands, with commentary. Knowledge of MATLAB is not necessary. Readers can choose another language and use the given commands as a *block-chart*. With this in mind, I intentionally did not include any complicated commands or MATLAB toolboxes.

Thanks and acknowledgments

The author is grateful to Taylor & Francis for their constant professional help, responsiveness, support, and encouragement. Special thanks are due Bob Stern, Barbara Johnson, and Rachael Panthier. Many thanks also go to Professor Ali Hooshyar for the computer and moral support; Elena Zavelskaya for creative illustrations; Kate Pechekhonova for interesting examples; and last, but not least, to Abraham and Anthony Baron for their amazing patience and understanding.

Contents

CHAPTER 1

Introduction and Overview

1.1 Making decisions under uncertainty

This course is about uncertainty, measuring and quantifying uncertainty, and making decisions under uncertainty. Loosely speaking, by *uncertainty* we mean the condition when results, outcomes, the nearest and remote future are not completely determined; their development depends on a number of factors and just on a pure chance.

Simple examples of uncertainty appear when you buy a lottery ticket, turn a wheel of fortune, or toss a coin to make a choice.

Uncertainly appears in virtually all areas of *Computer Science* and *Software Engineering*. Installation of software requires uncertain time and often uncertain disk space. A newly released software contains an uncertain number of defects. When a computer program is executed, the amount of required memory may be uncertain. When a job is sent to a printer, it takes uncertain time to print, and there is always a different number of jobs in a queue ahead of it. Electronic components fail at uncertain times, and the order of their failures cannot be predicted exactly. Viruses attack a system at unpredictable times and affect an unpredictable number of files and directories.

Uncertainty surrounds us in *everyday life*, at home, at work, in business, and in leisure. To take a snapshot, let us listen to the evening news.

Example 1.1. We may find out that the stock market had several ups and downs today which were caused by new contracts being made, financial reports being released, and other events of this sort. Many turns of stock prices remained unexplained. Clearly, nobody would have ever lost a cent in stock trading had the market contained no uncertainty.

We may find out that a launch of a space shuttle was postponed because of weather conditions. Why did not they know it in advance, when the event was scheduled? Forecasting weather precisely, with no error, is not a solvable problem, again, due to uncertainty.

To support these words, a meteorologist predicts, say, a 60% chance of rain. Why cannot she let us know exactly, whether it will rain or not, so we'll know whether or not to take our umbrellas? Yes, because of uncertainty. Because she cannot always know the situation with future precipitation for sure.

We may find out that eruption of an active volcano has suddenly started, and it is not clear which regions will have to evacuate.

We may find out that a heavily favored home team unexpectedly lost to an outsider, and a young tennis player won against expectations. Existence and popularity of totalizators, where participants place bets on sports results, show that uncertainty enters sports, results of each game and even the final standing.

We may also hear reports of traffic accidents, crimes, and convictions. Of course, if that driver knew about the coming accident ahead of time, he would have stayed home. ◇

Certainly, this list can be continued (at least one thing is certain!). Even when you drive to your college tomorrow, you will see an unpredictable number of green lights when you approach them, you will find an uncertain number of vacant parking slots, you will reach the classroom at an uncertain time, and you cannot be certain now about the number of classmates you will find in the classroom when you enter it.

Realizing that many important phenomena around us bear uncertainty, we have to understand it and deal with it. Most of the time, we are forced *to make decisions under uncertainty*. For example, we have to deal with internet and e-mail knowing that we may not be protected against all kinds of viruses. New software has to be released even if its testing probably did not reveal all the defects. Some memory or disk quota has to be allocated for each customer by servers, internet service providers, etc., without knowing exactly what portion of users will be satisfied with these limitations. And so on.

This book is about measuring and dealing with *uncertainty* and *randomness*. Through basic theory and numerous examples, it teaches

- how to evaluate *probabilities*, or chances of different results (when the exact result is uncertain),
- how to select a suitable *model* for a phenomenon containing uncertainty and use it in subsequent decision making,
- how to evaluate performance characteristics and other important *parameters* for new devices and servers,
- how to make optimal decisions under uncertainty.

Summary and conclusion

Uncertainty is a condition when the situation cannot be predetermined or predicted for sure with no error. Uncertainty exists in computer science, software engineering, in many aspects of science, business, and our everyday life. It is an objective reality, and one has to be able to deal with it. We are forced to make decisions under uncertainty.

1.2 Overview of this book

The next chapter introduces a language that we'll use to describe and quantify uncertainty. It is a language of *Probability*. When outcomes are uncertain, one can identify more likely and less likely ones and assign, respectively, high and low probabilities to them. Probabilities are numbers between 0 and 1, with 0 being assigned to an *impossible event* and 1 being the probability of an event that occurs *for sure*.

Next, using the introduced language, we shall discuss *random variables* as quantities that depend on chance. They assume different values with different probabilities. Due to uncertainty, an exact value of a random variable cannot be computed before this variable is actually observed or measured. Then, the best way to describe its behavior is to list all its *possible values* along with the corresponding probabilities.

Such a collection of probabilities is called a *distribution*. Amazingly, many different phenomena of seemingly unrelated nature can be described by the same distribution or by the same *family of distributions*. This allows a rather general approach to the entire class of situations involving uncertainty. As an application, it will be possible to compute probabilities of interest, once a suitable family of distributions is found. Chapters 3 and 4 introduce families of distributions that are most commonly used in computer science and other fields.

In modern practice, however, one often deals with rather complicated random phenomena where computation of probabilities and other quantities of interest is far from being straightforward. In such situations, we will make use of *Monte Carlo methods* (Chapter 5). Instead of direct computation, we shall learn methods of *simulation* or *generation* of random variables. If we are able to write a computer code for simulation of a certain phenomenon, we can immediately put it in a loop and simulate such a phenomenon thousands or millions of times and simply count how many times our event of interest occurred. This is how we shall distinguish more likely and less likely events. We can then *estimate* probability of an event by computing a proportion of simulations that led to the occurrence of this event.

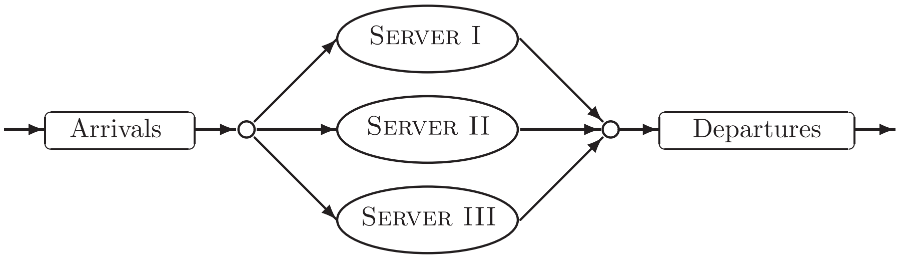


Figure 1.1 *A queuing system with 3 servers.*

As a step up to the next level, we shall realize that many random variables depend not only on a chance but also on *time*. That is, they evolve and develop in time while being random at each particular moment. Examples include the number of concurrent users, the number of jobs in a queue, the system's available capacity, intensity of internet traffic, stock prices, air temperature, etc. A random variable that depends on time is called a *stochastic process*. In Chapter 6, we study some commonly used types of stochastic processes and use these models to compute probabilities of events and other quantities of interest.

An important application of virtually all the material acquired so far lies in *queuing systems* (Chapter 7). These are systems of one or several servers performing certain tasks and serving jobs or customers. There is a lot of uncertainty in such systems. Customers arrive at unpredictable times, spend random time waiting in a queue, get assigned to a server, spend random time receiving service, and depart (Figure 1.1). In simple cases, we shall use our methods of computing probabilities and analyzing stochastic processes to compute such important characteristics of a queuing system as the utilization of a server, average waiting time of customers, average response time (from arrival until departure), average number of jobs in the system at any time, or the proportion of time the server is idle. This is extremely important for planning purposes. Performance characteristics can be recalculated for the next year, when, say, the number of customers is anticipated to increase by 5%. As a result, we'll know whether the system will remain satisfactory or will require an upgrade.

When direct computation is too complicated, resource consuming, too approximate, or simply not feasible, we shall use Monte Carlo methods. The book contains standard examples of computer codes simulating rather complex queuing systems and evaluating their vital characteristics. The codes are written in MATLAB, with detailed explanations of steps.

Next, we turn to *Statistical Inference*. While in Probability, we usually deal with more or less clearly described situations (models), in Statistics, all the analysis is based solely on collected and observed data. Given the data, a suitable model (say, a family of distributions) is fitted, its parameters are estimated, and conclusions are drawn concerning the entire totality of observed and unobserved subjects of interest that should follow the same model.

A typical Probability problem sounds like this.

Example 1.2. A folder contains 50 executable files. When a computer virus attacks a system, each file is affected with probability 0.2. Compute the probability that during a virus attack, more than 15 files get affected. ◇

Notice that the situation is rather clearly described, in terms of the total number of files and the chance of affecting each file. The only uncertain quantity is the number of affected files, which cannot be predicted for sure.

A typical Statistics problem sounds like this.

Example 1.3. A folder contains 50 executable files. When a computer virus attacks a system, each file is affected with the same probability p. It has been observed that during a virus attack, 15 files got affected. Estimate p. Is there a strong indication that p is greater than 0.2? ◇

This is a practical situation. A user only knows the objectively observed data: the number of files in the folder and the number of files that got affected. Based on that, he needs to estimate p, the proportion of *all* the files, including the ones in his system and any similar systems. One may provide a point estimator of p, a real number, or may opt to construct a *confidence interval* of "most probable" values of p. Similarly, a meteorologist may predict, say, a temperature of 70^oF, which, realistically, does not exclude a possibility of 69 or 72 degrees, or she may give us an interval by promising, say, between 68 and 72 degrees.

Most forecasts are being made from a carefully and suitably chosen model that fits the data. A widely used method is *regression* that utilizes the observed data to find a mathematical form of relationship between two variables (Chapter 10). One variable is called *predictor*, the other is *response*. When the relationship between them is established, one can use the predictor to infer about the response. For example, one can more or less accurately estimate the average installation time of a software given the size of its executable files. An even more accurate inference about the response can be made based on *several predictors*. For example, the size of executable files, amount of random access memory (RAM), and type of processor and operation system. This will call for *multivariate regression*.

Each method will be illustrated by numerous practical examples and exercises. As the ultimate target, by the end of this course, students should be able to read a word problem or a corporate report, realize the uncertainty involved in the described situation, select a suitable probability model, estimate and test its parameters based on real data, compute probabilities of interesting events and other vital characteristics, and make meaningful conclusions and forecasts.

Summary

In this course, uncertainty is measured and described on a language of Probability. Using this language, we shall study random variables and stochastic processes, learn the most commonly used types of distributions. In particular, we shall be able to find a suitable stochastic model for the described situation and use it to compute probabilities and other quantities of interest. When direct computation is not feasible, Monte Carlo methods will be used based on a random number generator. We shall then learn how to make decisions under uncertainty based on the observed data, how to estimate parameters of interest, test hypotheses, fit regression models, and make forecasts.

Questions and exercises

1.1. List 20 situations involving uncertainty that happened with you yesterday.

1.2. Name 10 random variables that you observed or dealt with yesterday.

1.3. Name 5 stochastic processes that played a role in your actions yesterday.

1.4. In a famous joke, a rather lazy student tosses a coin and decides what to do next. If it turns up heads, play a computer game. If tails, watch a video. If it stands on its edge, do the homework. If it hangs in the air, prepare for an exam.

(a) Which events should be assigned probability 0, probability 1, and some probability strictly between 0 and 1?

(b) What probability between 0 and 1 would you assign to the event "watch a video," and how does it help you to define "a fair coin"?

1.5. A new software package is being tested by specialists. Every day, a number of defects is found and corrected. It is planned to release this software in 30 days. Is it possible to predict how many defects per day specialists will be finding at the time of the release? What data should be collected for this purpose, what is the predictor, and what is the response?

1.6. Mr. Cheap plans to open a computer store to trade hardware. He would like to stock an optimal number of different hardware products in order to optimize his monthly profit. Data are available on similar computer stores opened in the area. What kind of data should Mr. Cheap collect in order to predict his monthly profit? What should he use as a predictor and as a response in his analysis?

CHAPTER 2

Probability

This chapter introduces the key concept of *probability*, its fundamental rules and properties, and discusses most basic methods of computing probabilities of various events.

2.1 Sample space, events, and probability

Probability: its meaning and definition

The concept of *probability* perfectly agrees with our intuition. In everyday life, probability of an event is understood as a chance that this event will happen.

Example 2.1. If a fair coin is tossed, we say that it has a 50-50 (equal) chance of turning up heads or tails. Hence, the probability of each side equals 1/2. It does not mean that a coin tossed 10 times will always produce exactly 5 heads and 5 tails. If you don't believe, try it! However, if you toss a coin 1 million times, the proportion of heads is anticipated to be very close to 1/2. ◇

This example suggests that in a long run, probability can be viewed as a proportion, or relative frequency. In forecasting, it is common to speak about the probability of an event as a likelihood of this event to happen (say, the company's profit is likely to rise during the next quarter). In gambling and lottery, probability is equivalent to odds. Having the winning odds of 1 to 100 (1:100) means that the probability to win is 0.01. It also means, on a relative-frequency language, that if you play long enough, you will win about 1% of the time.

Example 2.2. If there are 5 communication channels in service, and a channel is selected at random when a telephone call is made, then each channel has a probability of $1/5 = 0.2$ of being selected. $\diamond$

Example 2.3. Two competing software companies are after an important contract. Company A is twice as likely to win this competition as company B. Hence, the probability to win the contract equals $2/3$ for A and $1/3$ for B. $\diamond$

A mathematical definition of *probability* is this.

DEFINITION 2.1

> **Probability** is a finite measure. Being finite means that it has the largest possible value, which is one. Being a measure means first of all that it is *a function*, or a mechanism that takes input, event E, and converts it into output, probability $\boldsymbol{P}\{E\}$.

As any function, probability has its domain and range.

The range of probability, or the set of possible outputs, is the interval $[0, 1]$. That is, probability of any event is a real number between 0 and 1, including 0 and including 1.

The domain of probability, or the set of possible inputs, is not necessarily a set of numbers. Indeed, we consider probabilities *of what*? The answer is, we consider probabilities of events. Thus, *the domain of probability is a set of events*.

Outcomes, events, and the sample space

Probabilities arise when one considers and weighs possible results of some *experiment*. An experiment may be simple such as a coin toss, or very complex such as starting a new business.

DEFINITION 2.2

> A collection of all elementary results, or **outcomes** of an experiment, is called a **sample space**.

DEFINITION 2.3

An **event** is a set of outcomes, and simultaneously, a subset of the sample space.

Example 2.4. A tossed die can produce one of 6 possible outcomes: 1 dot through 6 dots. Each outcome is an event. There are other events: observing an even number of dots, an odd number of dots, a number of dots less than 3, etc. ◇

A sample space of N possible outcomes yields 2^N possible events.

PROOF: To count all possible events, we shall see how many ways an event can be constructed. The first outcome can be included into our event or excluded, so there are two possibilities. Then, every next outcome is either included or excluded, so every time the number of possibilities doubles. Overall, we have

$$\overbrace{2 \cdot 2 \cdot \ldots \cdot 2}^{N \text{ times}} = 2^N \tag{2.1}$$

possibilities, leading to a total of 2^N possible events. □

Example 2.5. Consider a football game between the Dallas Cowboys and the New York Giants. The sample space consists of 3 outcomes,

$$\Omega = \{ \text{ Cowboys win, Giants win, they tie } \}$$

Combining these outcomes in all possible ways, we obtain the following $2^3 = 8$ events: Cowboys win, lose, tie, get at least a tie, get at most a tie, no tie, get *some* result, and get *no result*. The event "some result" is the entire sample space Ω, and by common sense, it should have probability 1. The event "no result" is empty, it does not contain any outcomes, so its probability is 0. ◇

NOTATION	Ω	=	sample space
	$\varnothing$	=	empty event
	$\boldsymbol{P}\{E\}$	=	probability of event E

2.2 Rules of Probability

Set operations

Events are *sets* of outcomes. Therefore, to learn how to compute probabilities of events, we shall discuss some *set theory*. Namely, we shall define unions, intersections, differences, and complements.

DEFINITION 2.4

A **union** of events $A, B, C, \ldots$ is an event consisting of *all* the outcomes in all these events. It occurs if *any* of $A, B, C, \ldots$ occurs, and therefore, corresponds to the word "OR": A or B or C or ... (Figure 2.1a).

Diagrams like Figure 2.1, where events are represented by circles, are called *Venn diagrams*.

DEFINITION 2.5

An **intersection** of events $A, B, C, \ldots$ is an event consisting of outcomes that are *common* in all these events. It occurs if *each* $A, B, C, \ldots$ occurs, and therefore, corresponds to the word "AND": A and B and C and ... (Figure 2.1b).

DEFINITION 2.6

A **complement** of an event A is an event that occurs every time when A does not occur. It consists of outcomes excluded from A, and therefore, corresponds to the word "NOT": not A (Figure 2.1c).

DEFINITION 2.7

A **difference** of events A and B consists of all outcomes included in A but excluded from B. It occurs when A occurs and B does not, and corresponds to "BUT NOT": A but not B (Figure 2.1d).

NOTATION			
	$A \cup B$	$=$	union
	$A \cap B$	$=$	intersection
	$\overline{A}$ or A^c	$=$	complement
	$A \backslash B$	$=$	difference

DEFINITION 2.8

Events $A, B, C, \ldots$ are **disjoint** or **mutually exclusive** if their intersection is empty, i.e.,

$$A \cap B \cap C \cap \ldots = \varnothing.$$

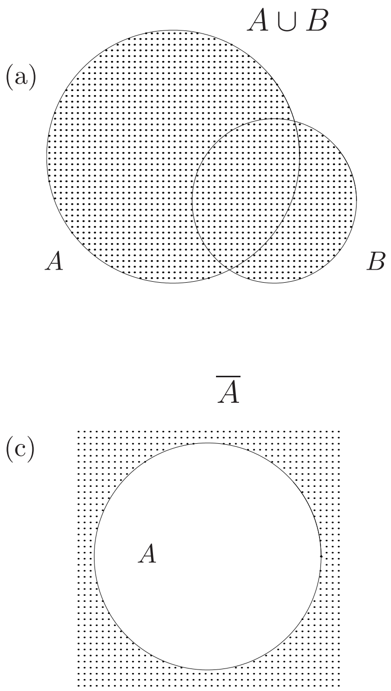


Figure 2.1 *Venn diagrams for (a) union, (b) intersection, (c) complement, and (d) difference of events.*

DEFINITION 2.9

Events $A, B, C, \ldots$ are **exhaustive** if their union equals the whole sample space, i.e.,

$$A \cup B \cup C \cup \ldots = \Omega.$$

Mutually exclusive events will never occur all at the same time. Exhaustive events cover the entire Ω, so that "there is nothing left." In other words, among any collection of exhaustive events, at least one occurs for sure.

Example 2.6. When a card is pooled from a deck at random, the four suits are at the same time disjoint and exhaustive. $\diamond$

Example 2.7. Any event A and its complement $\overline{A}$ represent a classical example of disjoint and exhaustive events. $\diamond$

Example 2.8. Receiving a grade of A, B, or C for some course are mutually exclusive events, but unfortunately, they are not exhaustive. ◇

As seen in the next section, it is often easier to compute probability of an intersection than probability of a union. Taking complements converts a union into an intersection, see (2.2).

$$\boxed{\overline{E_1 \cup \ldots \cup E_n} = \overline{E}_1 \cap \ldots \cap \overline{E}_n, \qquad \overline{E_1 \cap \ldots \cap E_n} = \overline{E}_1 \cup \ldots \cup \overline{E}_n} \tag{2.2}$$

Proof of (2.2): Since the union $E_1 \cup \ldots \cup E_n$ represents the event "at least one event occurs," its complement has the form

$$\begin{aligned} \overline{E_1 \cup \ldots \cup E_n} &= \{ \text{ none of them occurs } \} \\ &= \{ \overline{E_1 \text{ does not occur}} \cap \ldots \cap \overline{E_n \text{ does not occur}} \} \\ &= \overline{E}_1 \cap \ldots \cap \overline{E}_n. \end{aligned}$$

The other equality in (2.2) is left as Exercise 2.34. □

Example 2.9. Graduating with a GPA of 4.0 is an *intersection* of getting an A in *each* course. Its *complement*, graduating with a GPA below 4.0, is a *union* of receiving a grade below A *at least in one* course. ◇

Rephrasing (2.2), a complement to "nothing" is "something," and "not everything" means "at least one missing,"

2.2.1 Basic probability rules

Armed with the fundamentals of set theory, we are now able to compute probabilities of many interesting events.

Extreme cases

A sample space Ω consists of all possible outcomes, therefore, it occurs for sure. On the contrary, an empty event $\varnothing$ never occurs. Hence,

$$\boldsymbol{P}\{\Omega\} = 1 \text{ and } \boldsymbol{P}\{\varnothing\} = 0. \tag{2.3}$$

Union

Any event consists of some number of outcomes,

$$E = \{\omega_1, ..., \omega_n\}.$$

Summing probabilities of these outcomes, we obtain the probability of the entire event,

$$\boxed{P\{E\} = \sum_{\omega_k \in E} P\{\omega_k\} = P\{\omega_1\} + \ldots + P\{\omega_n\}}$$

If events A and B are mutually exclusive, then they have no common outcomes. Their union, $A \cup B$, consists of all the outcomes put together, hence,

$$\boldsymbol{P}\{A\} + \boldsymbol{P}\{B\} = \sum_{\omega_k \in A} P\{\omega_k\} + \sum_{\omega_k \in B} P\{\omega_k\} = \sum_{\omega_k \in A \cup B} P\{\omega_k\} = P\{A \cup B\}.$$

This rule extends to any number of mutually exclusive events.

Example 2.10. If a job sent to a printer appears first in line with probability 60%, and second in line with probability 30%, then with probability 90% it appears either first or second in line. ◇

Example 2.11. During some construction, a network blackout occurs on Monday with probability 0.7, and on Tuesday with probability 0.5. Then, does it appear on Monday *or* Tuesday with probability $0.7 + 0.5 = 1.2$? Obviously not, because probability should always be between 0 and 1! The rule above does not apply here because blackouts on Monday and Tuesday are not mutually exclusive. In other words, it is not impossible to see blackouts on both days. ◇

In Example 2.11, blind application of the rule for the union of mutually exclusive events clearly overestimated the actual probability. The Venn diagram shown in Figure 2.2a explains it. We see that in the sum $\boldsymbol{P}\{A\} + \boldsymbol{P}\{B\}$, all the common outcomes are counted *twice*. Certainly, this caused the overestimation. Each outcome should be counted only once! To correct the formula, subtract probabilities of common outcomes, which is $\boldsymbol{P}\{A \cap B\}$.

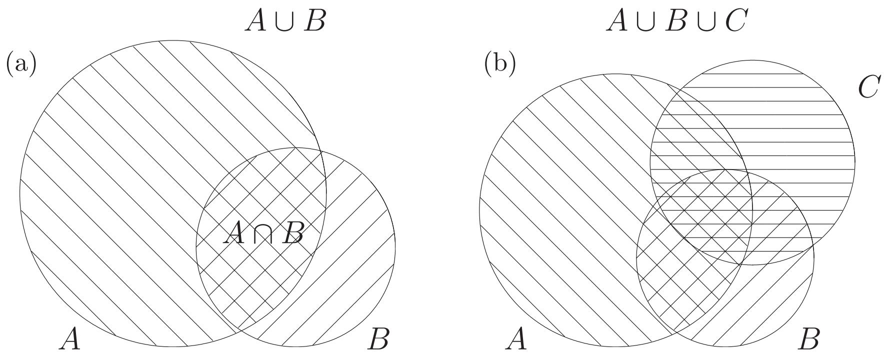


Figure 2.2 *(a) Union of two events. (b) Union of three events.*

Probability of a union

$$\boxed{\begin{array}{l} \boldsymbol{P}\{A \cup B\} = \boldsymbol{P}\{A\} + \boldsymbol{P}\{B\} - \boldsymbol{P}\{A \cap B\} \\ \text{For mutually exclusive events,} \\ \boldsymbol{P}\{A \cup B\} = \boldsymbol{P}\{A\} + \boldsymbol{P}\{B\} \end{array}} \qquad (2.4)$$

Generalization of this formula is not straightforward. For 3 events,

$$\begin{aligned} P\{A \cup B \cup C\} &= \boldsymbol{P}\{A\} + \boldsymbol{P}\{B\} + \boldsymbol{P}\{C\} - \boldsymbol{P}\{A \cap B\} - \boldsymbol{P}\{A \cap C\} \\ &\quad - \boldsymbol{P}\{B \cap C\} + \boldsymbol{P}\{A \cap B \cap C\}. \end{aligned}$$

As seen in Figure 2.2b, when we add probabilities of A, B, and C, each pairwise intersection is counted twice. Therefore, we subtract the probabilities of $P\{A \cap B\}$, etc. Finally, consider the triple intersection $A \cap B \cap C$. Its probability is counted 3 times within each main event, then subtracted 3 times with each pairwise intersection. Thus, it is not counted at all so far! Therefore, we add its probability $\boldsymbol{P}\{A \cap B \cap C\}$ in the end.

For an arbitrary collection of events, see Exercise 2.33.

Example 2.12. In Example 2.11, suppose there is a probability 0.35 of experiencing network blackouts on both Monday and Tuesday. Then the probability of having a blackout on Monday *or* Tuesday equals

$$0.7 + 0.5 - 0.35 = 0.85.$$

$\diamond$

Complement

Recall that events A and $\overline{A}$ are exhaustive, hence $A \cup \overline{A} = \Omega$. Also, they are disjoint, hence

$$\boldsymbol{P}\{A\} + \boldsymbol{P}\left\{\overline{A}\right\} = \boldsymbol{P}\left\{A \cup \overline{A}\right\} = \boldsymbol{P}\{\Omega\} = 1.$$

Solving this for $\boldsymbol{P}\left\{\overline{A}\right\}$, we obtain a rule that perfectly agrees with the common sense,

Complement rule $\boxed{\boldsymbol{P}\left\{\overline{A}\right\} = 1 - \boldsymbol{P}\{A\}}$

Example 2.13. If a system appears protected against a new computer virus with probability 0.7, then it is exposed to it with probability $1 - 0.7 = 0.3$. ◇

Example 2.14. Suppose a computer code has no errors with probability 0.45. Then, it has at least one error with probability 0.55. ◇

Intersection of independent events

DEFINITION 2.10

Events $E_1, \ldots, E_n$ are **independent** if they occur independently of each other, i.e., occurrence of one event does not affect the probabilities of others.

The following basic formula can serve as the criterion of independence.

Independent events $\boxed{\boldsymbol{P}\{E_1 \cap \ldots \cap E_n\} = \boldsymbol{P}\{E_1\} \cdot \ldots \cdot \boldsymbol{P}\{E_n\}}$

We shall defer explanation of this formula until the next section which will also give a rule for intersections of dependent events.

2.2.2 Applications in reliability

Formulas of the previous Section are widely used in *reliability*, when one computes probability for a system of several components to be functional.

Example 2.15 (Reliability of backups). There is a 1% probability for a hard drive to crash. Therefore, it has two backups, each having a 2% probability to crash, and all three components are independent of each other. The stored information is lost only in an unfortunate situation when all three devices crash. What is the probability that the information is saved?

Solution. Organize the data. Denote the events, say,

$$H = \{\text{ hard drive crashes }\},$$

$$B_1 = \{\text{ first backup crashes }\},\ B_2 = \{\text{ second backup crashes }\}.$$

It is given that H, B_1, and B_2 are independent,

$$\boldsymbol{P}\{H\} = 0.01, \text{ and } \boldsymbol{P}\{B_1\} = \boldsymbol{P}\{B_2\} = 0.02.$$

Applying rules for the complement and for the intersection of independent events,

$$\begin{aligned}\boldsymbol{P}\{\text{ saved }\} &= 1 - \boldsymbol{P}\{\text{ lost }\} = 1 - \boldsymbol{P}\{H \cap B_1 \cap B_2\}\\ &= 1 - \boldsymbol{P}\{H\}\boldsymbol{P}\{B_1\}\boldsymbol{P}\{B_2\}\\ &= 1 - (0.01)(0.02)(0.02) = 0.999996.\end{aligned}$$

(This is precisely the reason of having backups, isn't it? Without backups, the probability for information to be saved is only 0.99.) ◇

When the system's components are connected *in parallel*, it is sufficient for at least one component to work in order for the whole system to function. Reliability of such a system is computed as in Example 2.15. Backups can always be considered as devices connected in parallel.

At the other end, consider a system whose components are connected *in sequel*. Failure of one component inevitably causes the whole system to fail. Such a system is more "vulnerable." In order to function with a high probability, it needs each component to be reliable, as in the next example.

Example 2.16. Suppose that a shuttle's launch depends on three key devices that operate independently of each other and malfunction with probabilities 0.01, 0.02, and 0.02, respectively. If any of the key devices malfunctions, the launch will be postponed. Compute the probability for the shuttle to be launched on time, according to its schedule.

Solution. In this case,

$$\begin{aligned}\boldsymbol{P}\{\text{ on time }\} &= \boldsymbol{P}\{\text{ all function }\}\\ &= \boldsymbol{P}\left\{\overline{H}\cap\overline{B}_1\cap\overline{B}_2\right\}\\ &= \boldsymbol{P}\left\{\overline{H}\right\}\boldsymbol{P}\left\{\overline{B}_1\right\}\boldsymbol{P}\left\{\overline{B}_2\right\}\quad(\textit{independence})\\ &= (1-0.01)(1-0.02)(1-0.02)\quad(\textit{complement rule})\\ &= 0.9508.\end{aligned}$$

Notice how with the same probabilities of individual components as in Example 2.15, the system's reliability decreased because the components were connected sequentially. ◇

Many modern systems consist of a great number of devices connected in sequel and in parallel.

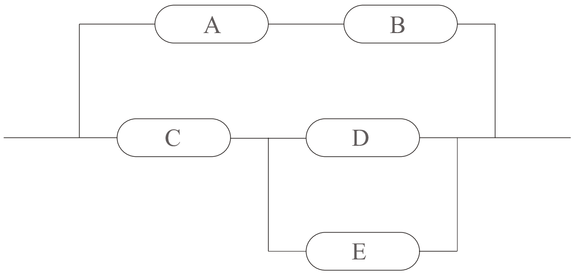


Figure 2.3 *Calculate reliability of this system (Example 2.17).*

Example 2.17 (Techniques for solving reliability problems). Calculate reliability of the system in Figure 2.3 if each component is operable with probability 0.92 independently of the other components.

Solution. This problem can be simplified and solved "step by step."

1. The upper link A-B works if both A and B work, which has probability

$$\boldsymbol{P}\{A\cap B\} = (0.92)^2 = 0.8464.$$

   We can represent this link as one component F that operates with probability 0.8464.

2. By the same token, components D and E, connected in parallel, can be

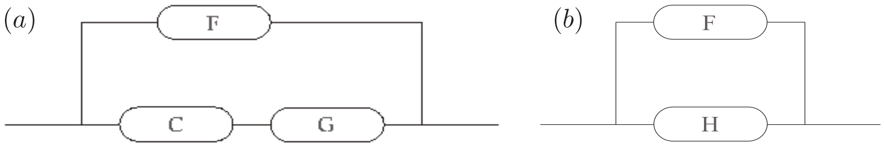


Figure 2.4 *Step by step solution of a system reliability problem.*

replaced by component G, operable with probability

$$\boldsymbol{P}\{D \cup E\} = 1 - (1 - 0.92)^2 = 0.9936,$$

as shown in Figure 2.4a.

3. Components C and G, connected sequentially, can be replaced by component H, operable with probability $\boldsymbol{P}\{C \cap G\} = 0.92 \cdot 0.9936 = 0.9141$, as shown in Figure 2.4b.
4. Last step. The system operates with probability

$$\boldsymbol{P}\{F \cup H\} = 1 - (1 - 0.8464)(1 - 0.9141) = \underline{0.9868},$$

which is the final answer.

In fact, the event "the system is operable" can be represented as $(A \cap B) \cup \{C \cap (D \cup E)\}$, whose probability we found step by step. ◇

2.3 Equally likely outcomes. Combinatorics

The case of equally likely outcomes

A simple situation for computing probabilities is the case of *equally likely outcomes*. That is, when the sample space Ω consists of n possible outcomes, $\omega_1, \ldots, \omega_n$, each having the same probability. Since

$$\sum_1^n \boldsymbol{P}\{\omega_k\} = \boldsymbol{P}\{\Omega\} = 1,$$

we have in this case $\boldsymbol{P}\{\omega_k\} = 1/n$ for all k. Further, a probability of any event E consisting of t outcomes, equals

$$\boldsymbol{P}\{E\} = \sum_{\omega_k \in E} \left(\frac{1}{n}\right) = t\left(\frac{1}{n}\right) = \frac{\text{number of outcomes in } E}{\text{number of outcomes in } \Omega}.$$

The outcomes forming event E are often called "favorable." Thus we have a formula

Equally likely outcomes

$$\boxed{\boldsymbol{P}\{E\} = \frac{\text{number of favorable outcomes}}{\text{total number of outcomes}} = \frac{\mathcal{N}_F}{\mathcal{N}_T}} \qquad (2.5)$$

where index "F" means "favorable" and "T" means "total."

Example 2.18. Tossing a die results in 6 equally likely possible outcomes, identified by the number of dots from 1 to 6. Applying (2.5), we obtain,

$$\boldsymbol{P}\{1\} = 1/6, \quad \boldsymbol{P}\{\text{ odd number of dots }\} = 3/6, \quad \boldsymbol{P}\{\text{ less than } 5\} = 4/6.$$

$\diamond$

The solution and even the answer to such problems may depend on our choice of outcomes and a sample space. Outcomes should be defined in such a way that they appear equally likely, otherwise formula (2.5) does not apply.

Example 2.19. A card is drawn from a bridge 52-card deck at random. Compute the probability that the selected card is a spade.

First solution. The sample space consists of 52 equally likely outcomes—cards. Among them, there are 13 favorable outcomes—spades. Hence, $\boldsymbol{P}\{\text{spade}\} = 13/52 = 1/4$.

Second solution. The sample space consists of 4 equally likely outcomes—suits: clubs, diamonds, hearts, and spades. Among them, one outcome is favorable—spades. Hence, $\boldsymbol{P}\{\text{spade}\} = 1/4$. $\diamond$

These two solutions relied on different sample spaces. However, in both cases, the defined outcomes were equally likely, therefore (2.5) was applicable, and we obtained the same result.

However, the situation may be different.

Example 2.20. A young family plans to have two children. What is the probability of two girls?

Solution 1 (wrong). There are 3 possible families with 2 children: two girls, two boys, and one of each gender. Therefore, the probability of two girls is 1/3.

Solution 2 (right). Each child is (supposedly) equally likely to be a boy or a girl. Genders of the two children are (supposedly) independent. Therefore,

$$\boldsymbol{P}\{\text{two girls}\} = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = 1/4.$$

◇

The second solution implies that the sample space consists of four, not three, equally likely outcomes: two boys, two girls, a boy and a girl, a girl and a boy. Each outcome in this sample has probability 1/4. Notice that the last two outcomes are counted separately, with the meaning, say, "the first child is a boy, the second one is a girl" and "the first child is a girl, the second one is a boy."

It is all right to define a sample space as in Solution 1. However, one must know in this case that the defined outcomes are *not equally likely*. Indeed, from Solution 2, we see that having one child of each gender is the most likely outcome, with probability of $1/4 + 1/4 = 1/2$. It was a mistake to apply (2.5) for such a sample space in Solution 1.

Example 2.21 (Paradox). There is a simple but controversial "situation" about a family with two children. Even some graduate students often fail to resolve it.

A family has two children. You met one of them, Jimmy, and he is a boy. What is the probability that the other child is a girl?

On one hand, why would the other child's gender be affected by Jimmy? Jimmy should have a sister or a brother with probabilities 1/2 and 1/2.

On the other hand, see Example 2.20. The sample space consists of 4 equally likely outcomes, $\{GG, BB, BG, GB\}$. You have already met one boy, thus the first outcome is automatically eliminated: $\{BB, BG, GB\}$. Among the remaining three outcomes, Jimmy has a sister in two outcomes and a brother in one. Thus, isn't the probability of a girl equal 2/3?

Where is the catch? Apparently, the sample space Ω has not been clearly defined in this example. The experiment is more complex than in Example 2.20 because we are now concerned not only about the gender of children but also about you meeting one of them. What is the mechanism, what are the probabilities for you to meet one or the other child? Once you met Jimmy, do the outcomes $\{BB, BG, GB\}$ remain equally likely?

A complete solution to this paradox is broken into steps in Exercise 2.30. ◇

In reality, business-related, sports-related, and political events are typically not equally likely. One outcome is usually more likely that another. For example, one team is always stronger than the other. Equally likely outcomes are usually associated with conditions of "a fair game" and "selected at random." In fair gambling, all cards, all dots on a die, all numbers in a roulette are equally likely. Also, when a survey is conducted, or a sample is selected "at random," the outcomes are "as close as possible" to being equally likely. This means, all the subjects have the same chance to be selected into a sample (otherwise, it is not a fair sample, and it can produce "biased" results).

2.3.1 Combinatorics

Formula (2.5) is simple, as long as its numerator and denominator can be easily evaluated. This is rarely the case; often the sample space consists of a multitude of outcomes. *Combinatorics* provides special techniques for the computation of $\mathcal{N}_T$ and $\mathcal{N}_F$, the total number and the number of favorable outcomes.

We shall consider a generic situation when objects is selected *at random* from a set of n. This general model has a number of useful applications.

The objects may be selected with replacement or without replacement. They may also be *distinguishable* or *indistinguishable*.

DEFINITION 2.11

Sampling **with replacement** means that every sampled item is replaced into the initial set, so that any of the objects can be selected with probability $1/N$ at any time. In particular, the same object may be sampled more than once.

DEFINITION 2.12

Sampling **without replacement** means that every sampled item is removed from further sampling, so the set of possibilities reduces by 1 after each selection.

DEFINITION 2.13

Objects are **distinguishable** if sampling of exactly the same objects *in a different order* yields a different outcome, that is, a different element of the sample space. For **indistinguishable** objects, the order is not important, it only matters which objects are sampled and which ones aren't.

Example 2.22 (Computer-generated passwords). When random passwords are generated, the order of characters is important because a different order yields a different password. Characters are distinguishable in this case. Further, if a password has to consist of different characters, they are sampled from the alphabet without replacement. ◇

Example 2.23 (Polls). When a sample of people is selected to conduct a poll, the same participants produce the same responses regardless of their order. They can be considered indistinguishable. ◇

Permutations with replacement

Possible selections of k *distinguishable* objects from a set of n are called *permutations*. When we sample with replacement, each time there are n possible selections, and the total of permutations is

Permutations with replacement

$$\boxed{P_r(n,k) = \overbrace{n \cdot n \cdot \ldots \cdot n}^{k \text{ terms}} = n^k}$$

Example 2.24 (Breaking passwords). From an alphabet consisting of 10 digits, 26 lower-case and 26 capital letters, one can create $P_r(62,8) =$ 218,340,105,584,896 (over 218 trillion) different 8-character passwords. At a speed of 1 million passwords per second, it will take a spy program almost 7 years to try all of them. Thus, on the average, it will guess your password in about 3.5 years.

At this speed, the spy program can test 604,800,000,000 passwords within 1 week. The probability that it guesses your password in 1 week is

$$\frac{\mathcal{N}_F}{\mathcal{N}_T} = \frac{\text{number of favorable outcomes}}{\text{total number of outcomes}} = \frac{604{,}800{,}000{,}000}{218{,}340{,}105{,}584{,}896} = 0.00277.$$

However, if capital letters are not used, the number of possible passwords is reduced to $P_r(36,8) =$ 2,821,109,907,456. This password will be guessed in about 16 days! The probability that it will happen in 1 week is 0.214. It is clear now why it is recommended to use three types of characters in your passwords and to change them every year. ◇

Permutations without replacement

During sampling without replacement, the number of possible selections reduces by 1 each time an object is sampled. Therefore, the number of permutations is

Permutations without replacement

$$\boxed{P(n,k) = \overbrace{n(n-1)(n-2)\cdot\ldots\cdot(n-k+1)}^{k \text{ terms}} = \frac{n!}{(n-k)!}}$$

where $n! = 1\cdot 2\cdot\ldots n$ (*n-factorial*) denotes the product of all integers from 1 to n.

The number of permutations without replacement also equals the number of possible allocations of k distinguishable objects among n available slots.

Example 2.25. In how many ways can 10 students be seated in a classroom with 15 chairs?

Solution. Students are distinguishable, and each student has a separate seat. Thus, the number of possible allocations is the number of permutations without replacement, $P(15,10) = 15\cdot 14\cdot\ldots\cdot 6 = 1.09\cdot 10^{10}$. Notice that if students enter the classroom one by one, the first student has 15 choices of seats, then one seat is occupied, and the second student has only 14 choices, etc., and the last student takes one of 6 chairs available at that time. ◇

Combinations without replacement

Possible selections of k *indistinguishable* objects from a set of n are called *combinations*. The number of combinations without replacement is also called "n choose k" and is denoted by $C(n,k)$ or $\binom{n}{k}$.

The only difference from $P(n,k)$ is disregarding the order. Now the same objects sampled in a different order produce the same outcome. Thus, $P(k,k) = k!$ different permutations (rearrangements) of the same objects yield only 1 combination. The total number of combinations is then

Combinations without replacement

$$\boxed{C(n,k) = \binom{n}{k} = \frac{P(n,k)}{P(k,k)} = \frac{n!}{k!(n-k)!}} \tag{2.6}$$

Example 2.26. An antivirus software reports that 3 folders out of 10 are infected. How many possibilities are there?

Solution. Folders A, B, C and folders C, B, A represent the same outcome, thus, the order is not important. A software clearly detected 3 different folders, thus it is sampling without replacement. The number of possibilities is

$$\binom{10}{3} = \frac{10!}{3!\,7!} = \frac{10 \cdot 9 \cdot \ldots \cdot 1}{(3 \cdot 2 \cdot 1)(7 \cdot \ldots \cdot 1)} = \frac{10 \cdot 9 \cdot 8}{3 \cdot 2 \cdot 1} = 120.$$

◇

Computational shortcuts

Instead of computing $C(n,k)$ directly by the formula, we can simplify the fraction. At least, the numerator and denominator can both be divided by either $k!$ or $(n-k)!$ (choose the larger of these for greater reduction). As a result,

$$C(n,k) = \binom{n}{k} = \frac{n \cdot (n-1) \cdot \ldots \cdot (n-k+1)}{k \cdot (k-1) \cdot \ldots \cdot 1},$$

the top and the bottom of this fraction being products of k terms. It is also handy to notice that

$$\begin{aligned} C(n,k) &= C(n,n-k) \text{ for any } k \text{ and } n \\ C(n,0) &= 1 \\ C(n,1) &= n \end{aligned}$$

Example 2.27. There are 20 computers in a store. Among them, 15 are brand new and 5 are refurbished. Six computers are purchased for a student lab. From the first look, they are indistinguishable, so the six computers are selected at random. Compute the probability that among the chosen computers, two are refurbished.

Solution. Compute the total number and the number of favorable outcomes. The *total* number of ways in which 6 computers are selected from 20 is

$$\mathcal{N}_T = \binom{20}{6} = \frac{20 \cdot 19 \cdot 18 \cdot 17 \cdot 16 \cdot 15}{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}.$$

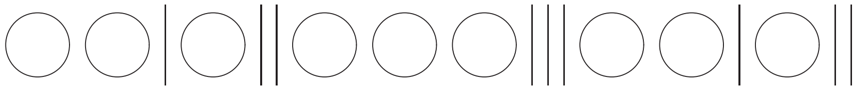


Figure 2.5 *Counting combinations with replacement. Vertical bars separate different classes of items.*

We applied the mentioned computational shortcut. Next, for the number of *favorable* outcomes, 2 refurbished computers are selected from a total of 5, and the remaining 4 new ones are selected from a total of 15. There are

$$\mathcal{N}_F = \binom{5}{2}\binom{15}{4} = \left(\frac{5\cdot 4}{2\cdot 1}\right)\left(\frac{15\cdot 14\cdot 13\cdot 12}{4\cdot 3\cdot 2\cdot 1}\right)$$

favorable outcomes. With further reduction of fractions, the probability equals

$$\boldsymbol{P}\left\{\text{ two refurbished computers }\right\} = \frac{\mathcal{N}_F}{\mathcal{N}_T} = \frac{7\cdot 13\cdot 5}{19\cdot 17\cdot 4} = 0.3522.$$

◇

Combinations with replacement

For combinations with replacement, the order is not important, and each object may be sampled more than once. Then each outcome consists of counts, how many times each of n objects appears in the sample. In Figure 2.5, we draw a circle for each time object #1 is sampled, then draw a separating bar, then a circle for each time object #2 is sampled, etc. Two bars next to each other mean that the corresponding object has never been sampled.

The resulting picture has to have k circles for a sample of size k and $(n-1)$ bars separating n objects. Each picture with these conditions represents an outcome. How many outcomes are there? It is the number of allocations of k circles and $(n-1)$ bars among $(k+n-1)$ slots available for them. Hence,

Combinations with replacement

$$\boxed{C_r(n,k) = \binom{k+n-1}{k} = \frac{(k+n-1)!}{k!(n-1)!}}$$

Notation

$P_r(n,k)$	$=$	number of permutations with replacement
$P(n,k)$	$=$	number of permutations without replacement
$C_r(n,k)$	$=$	number of combinations with replacement
$C(n,k)$ $\binom{n}{k}$	$=$	number of combinations without replacement

2.4 Conditional probability. Independence

Conditional probability

Suppose you are meeting someone at an airport. The flight is likely to arrive on time, the probability of that is 0.8. Suddenly it is announced that the flight departed one hour behind the schedule. Now it has the probability of only 0.05 to arrive on time. New information affected the probability of meeting this flight on time. The new probability is called *conditional probability*, where the new information, that the flight departed late, is a *condition*.

DEFINITION 2.14

Conditional probability of event A given event B is the probability that A occurs when B is *known to occur*.

Notation $\boldsymbol{P}\{A \mid B\} \quad = \quad$ conditional probability of A given B

How does one compute the conditional probability? First, consider the case of equally likely outcomes. In view of the new information, occurrence of the condition B, only the outcomes contained in B still have a non-zero chance to occur. Counting only such outcomes, the *unconditional probability* of A,

$$\boldsymbol{P}\{A\} = \frac{\text{number of outcomes in } A}{\text{number of outcomes in } \Omega}$$

is replaced by the *conditional probability* of A given B,

$$\boldsymbol{P}\{A \mid B\} = \frac{\text{number of outcomes in } A \cap B}{\text{number of outcomes in } B} = \frac{\boldsymbol{P}\{A \cap B\}}{\boldsymbol{P}\{B\}}.$$

This appears to be the general formula.

Conditional probability

$$\boxed{\boldsymbol{P}\{A \mid B\} = \frac{\boldsymbol{P}\{A \cap B\}}{\boldsymbol{P}\{B\}}} \tag{2.7}$$

From this, we also obtain the general formula for the probability of intersection.

Intersection, general case

$$\boxed{\boldsymbol{P}\{A \cap B\} = \boldsymbol{P}\{B\}\,\boldsymbol{P}\{A \mid B\}} \tag{2.8}$$

Independence

Now we can give an intuitively very clear definition of *independence*.

DEFINITION 2.15

Events A and B are **independent** if occurrence of B does not affect the probability of A, i.e.,

$$\boldsymbol{P}\{A \mid B\} = \boldsymbol{P}\{A\}.$$

According to this definition, *conditional* probability equals *unconditional* probability in case of independent events. Substituting this into (2.8) yields

$$\boldsymbol{P}\{A \cap B\} = \boldsymbol{P}\{A\}\,\boldsymbol{P}\{B\}.$$

This is our old formula for independent events.

Example 2.28. Ninety percent of flights depart on time. Eighty percent of flights arrive on time. Seventy-five percent of flights depart on time and arrive on time.

(a) You are meeting a flight that departed on time. What is the probability that it will arrive on time?

(b) You have met a flight, and it arrived on time. What is the probability that it departed on time?

(c) Are the events, departing on time and arriving on time, independent?

Solution. Denote the events,

$$\begin{array}{rcl} A & = & \{\text{arriving on time}\}\,, \\ D & = & \{\text{departing on time}\}\,. \end{array}$$

We have:

$$\boldsymbol{P}\{A\} = 0.8, \quad \boldsymbol{P}\{D\} = 0.9, \quad \boldsymbol{P}\{A \cap D\} = 0.75.$$

(a) $\boldsymbol{P}\{A \mid D\} = \dfrac{\boldsymbol{P}\{A \cap D\}}{\boldsymbol{P}\{D\}} = \dfrac{0.75}{0.9} = \underline{0.8333}$

(b) $\boldsymbol{P}\{D \mid A\} = \dfrac{\boldsymbol{P}\{A \cap D\}}{\boldsymbol{P}\{A\}} = \dfrac{0.75}{0.8} = \underline{0.9375}$

(c) Events are not independent because

$$\boldsymbol{P}\{A \mid D\} \neq \boldsymbol{P}\{A\}\,, \ \boldsymbol{P}\{D \mid A\} \neq \boldsymbol{P}\{D\}\,, \ \boldsymbol{P}\{A \cap D\} \neq \boldsymbol{P}\{A\}\,\boldsymbol{P}\{D\}\,.$$

Actually, any one of these inequalities is sufficient to prove that A and D are dependent. Further, we see that $\boldsymbol{P}\{A \mid D\} > \boldsymbol{P}\{A\}$ and $\boldsymbol{P}\{D \mid A\} > \boldsymbol{P}\{D\}$. In other words, departing on time increases the probability of arriving on time, and vise versa. This perfectly agrees with our intuition. ◇

Bayes Rule

The last example shows that two conditional probabilities, $\boldsymbol{P}\{A \mid B\}$ and $\boldsymbol{P}\{B \mid A\}$, are not the same, in general. Consider another example.

Example 2.29 (Reliability of a test). There exists a test for a certain virus disease (including a virus attack of a computer network). It is 95% reliable for infected patients and 99% reliable for others. That is, if a patient has the virus (event V), the test shows that (event S) with probability $\boldsymbol{P}\{S \mid V\} = 0.95$, and if the patient does not have the virus, the test shows that with probability $\boldsymbol{P}\left\{\overline{S} \mid \overline{V}\right\} = 0.99$.

Consider a patient whose test result is positive (i.e., the test shows that the patient has the virus). Knowing that sometimes the test is wrong, naturally, the patient is eager to know the probability that he or she indeed has the virus. However, this conditional probability, $\boldsymbol{P}\{V \mid S\}$, is not stated among the given characteristics of this test. ◇

This example is applicable to any testing procedure including software and hardware tests, pregnancy tests, paternity tests, alcohol tests, etc. The problem is to connect the given $\boldsymbol{P}\{S \mid V\}$ and the quantity in question, $\boldsymbol{P}\{V \mid S\}$. This was done in the eighteenth century by English minister *Thomas Bayes* (1702–1761) in the following way.

Notice that $A \cap B = B \cap A$. Therefore, using (2.8),

$$\boldsymbol{P}\{B\}\,\boldsymbol{P}\{A \mid B\} = \boldsymbol{P}\{A\}\,\boldsymbol{P}\{B \mid A\}.$$

Solve for $\boldsymbol{P}\{B \mid A\}$ to obtain

Bayes Rule

$$\boxed{\boldsymbol{P}\{B \mid A\} = \frac{\boldsymbol{P}\{A \mid B\}\,\boldsymbol{P}\{B\}}{\boldsymbol{P}\{A\}}} \qquad (2.9)$$

Example 2.30 (Situation on a midterm exam). On a midterm exam, students X, Y, and Z forgot to sign their papers. Professor knows that they can write a good exam with probabilities 0.8, 0.7, and 0.5, respectively. After the grading, he notices that two unsigned exams are good and one is bad. Given this information, and assuming that students worked independently of each other, what is the probability that the bad exam belongs to student Z?

Solution. Denote good and bad exams by G and B. Also, let GGB denote two good and one bad exams, XG denote the event "student X wrote a good exam," etc. We need to find $\boldsymbol{P}\{ZB \mid GGB\}$ given that $\boldsymbol{P}\{G \mid X\} = 0.8$, $\boldsymbol{P}\{G \mid Y\} = 0.7$, and $\boldsymbol{P}\{G \mid Z\} = 0.5$.

By the *Bayes Rule*,

$$\boldsymbol{P}\{ZB \mid GGB\} = \frac{\boldsymbol{P}\{GGB \mid ZB\}\,\boldsymbol{P}\{ZB\}}{\boldsymbol{P}\{GGB\}}.$$

Given ZB, event GGB occurs only when both X and Y write good exams. Thus, $\boldsymbol{P}\{GGB \mid ZB\} = (0.8)(0.7)$.

Event GGB consists of three outcomes depending on the student who wrote the bad exam. Adding their probabilities, we get

$$\begin{aligned}
&\boldsymbol{P}\{GGB\} \\
&= \boldsymbol{P}\{XG \cap YG \cap ZB\} + \boldsymbol{P}\{XG \cap YB \cap ZG\} + \boldsymbol{P}\{XB \cap YG \cap ZG\} \\
&= (0.8)(0.7)(0.5) + (0.8)(0.3)(0.5) + (0.2)(0.7)(0.5) = 0.47.
\end{aligned}$$

Then

$$\boldsymbol{P}\{ZB \mid GGB\} = \frac{(0.8)(0.7)(0.5)}{0.47} = \underline{0.5957}.$$

◇

In the Bayes Rule (2.9), the denominator is often computed by the Law of Total Probability.

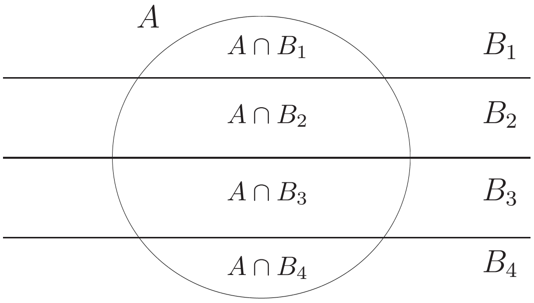


Figure 2.6 *Partition of the sample space* Ω *and the event* A.

Law of Total Probability

This law relates the unconditional probability of an event A with its conditional probabilities. It is used every time when it is easier to compute conditional probabilities of A given additional information.

Consider some partition of the sample space Ω with mutually exclusive and exhaustive events $B_1, \ldots, B_k$. It means that

$$B_i \cap B_j = \varnothing \text{ for any } i \neq j \text{ and } B_1 \cup \ldots \cup B_k = \Omega.$$

These events also partition the event A,

$$A = (A \cap B_1) \cup \ldots \cup (A \cap B_k),$$

and this is also a union of mutually exclusive events (Figure 2.6). Hence,

$$\boldsymbol{P}\{A\} = \sum_{j=1}^{k} \boldsymbol{P}\{A \cap B_j\},$$

and

Law of Total Probability

$$\boldsymbol{P}\{A\} = \sum_{j=1}^{k} \boldsymbol{P}\{A \mid B_j\} \boldsymbol{P}\{B_j\}$$

In case of two events ($k = 2$),

$$\boldsymbol{P}\{A\} = \boldsymbol{P}\{A \mid B\} \boldsymbol{P}\{B\} + \boldsymbol{P}\left\{A \mid \overline{B}\right\} \boldsymbol{P}\left\{\overline{B}\right\} \tag{2.10}$$

Together with the Bayes Rule, it makes the following popular formula

Bayes Rule for two events

$$\boxed{\boldsymbol{P}\left\{B \mid A\right\} = \frac{\boldsymbol{P}\left\{A \mid B\right\}\boldsymbol{P}\left\{B\right\}}{\boldsymbol{P}\left\{A \mid B\right\}\boldsymbol{P}\left\{B\right\} + \boldsymbol{P}\left\{A \mid \overline{B}\right\}\boldsymbol{P}\left\{\overline{B}\right\}}}$$

Example 2.31 (Reliability of a test, continued). Continue Example 2.29. Suppose that 4% of all the patients are infected with the virus, $\boldsymbol{P}\left\{V\right\} = 0.04$. Recall that $\boldsymbol{P}\left\{S \mid V\right\} = 0.95$ and $\boldsymbol{P}\left\{\overline{S} \mid \overline{V}\right\} = 0.99$. If the test shows positive results, the (conditional) probability that a patient has the virus equals

$$\begin{aligned}\boldsymbol{P}\left\{V \mid S\right\} &= \frac{\boldsymbol{P}\left\{S \mid V\right\}\boldsymbol{P}\left\{V\right\}}{\boldsymbol{P}\left\{S \mid V\right\}\boldsymbol{P}\left\{V\right\} + \boldsymbol{P}\left\{S \mid \overline{V}\right\}\boldsymbol{P}\left\{\overline{V}\right\}} \\ &= \frac{(0.95)(0.04)}{(0.95)(0.04) + (1-0.99)(1-0.04)} = \underline{0.7983}.\end{aligned}$$

◇

Example 2.32 (Diagnostics of computer codes). A new computer program consists of two modules. The first module contains an error with probability 0.2. The second module is more complex, it has a probability of 0.4 to contain an error, independently of the first module. An error in the first module alone causes the program to crash with probability 0.5. For the second module, this probability is 0.8. If there are errors in both modules, the program crashes with probability 0.9. Suppose the program crashed. What is the probability of errors in both modules?

Solution. Denote the events,

$$A = \left\{\text{errors in module I}\right\}, \quad B = \left\{\text{errors in module II}\right\}, \quad C = \left\{\text{crash}\right\}.$$

Further,

$$\begin{aligned}\left\{\text{errors in module I alone}\right\} &= A\backslash B = A\backslash(A \cap B) \\ \left\{\text{errors in module II alone}\right\} &= B\backslash A = B\backslash(A \cap B).\end{aligned}$$

It is given that $\boldsymbol{P}\left\{A\right\} = 0.2$, $\boldsymbol{P}\left\{B\right\} = 0.4$, $\boldsymbol{P}\left\{A \cap B\right\} = (0.2)(0.4) = 0.08$, by independence, $\boldsymbol{P}\left\{C \mid A\backslash B\right\} = 0.5$, $\boldsymbol{P}\left\{C \mid B\backslash A\right\} = 0.8$, and $\boldsymbol{P}\left\{C \mid A \cap B\right\} = 0.9$.

We need to compute $\boldsymbol{P}\left\{A \cap B \mid C\right\}$. Since A is a union of disjoint events $A\backslash B$ and $A \cap B$, we compute

$$\boldsymbol{P}\left\{A\backslash B\right\} = \boldsymbol{P}\left\{A\right\} - \boldsymbol{P}\left\{A \cap B\right\} = 0.2 - 0.08 = 0.12.$$

Similarly,

$$\boldsymbol{P}\{B\backslash A\} = 0.4 - 0.08 = 0.32.$$

Events $(A\backslash B)$, $(B\backslash A)$, $A \cap B$, and $\overline{(A \cup B)}$ form a partition of Ω, because they are mutually exclusive and exhaustive. The last of them is the event of no errors in the entire program. Given this event, the probability of a crash is 0. Notice that A, B, and $(A \cap B)$ are neither mutually exclusive nor exhaustive, so they cannot be used for the Bayes Rule. Now organize the data.

Location of errors	Probability of a crash
$\boldsymbol{P}\{A\backslash B\} = 0.12$	$\boldsymbol{P}\{C \mid A\backslash B\} = 0.5$
$\boldsymbol{P}\{B\backslash A\} = 0.32$	$\boldsymbol{P}\{C \mid B\backslash A\} = 0.8$
$\boldsymbol{P}\{A \cap B\} = 0.08$	$\boldsymbol{P}\{C \mid A \cap B\} = 0.9$
$\boldsymbol{P}\{\overline{A \cup B}\} = 0.48$	$\boldsymbol{P}\{C \mid \overline{A \cup B}\} = 0$

Combining the Bayes Rule and the Law of Total Probability,

$$\boldsymbol{P}\{A \cap B \mid C\} = \frac{\boldsymbol{P}\{C \mid A \cap B\}\,\boldsymbol{P}\{A \cap B\}}{\boldsymbol{P}\{C\}},$$

where

$$\begin{aligned} \boldsymbol{P}\{C\} \quad = \quad & \boldsymbol{P}\{C \mid A\backslash B\}\,\boldsymbol{P}\{A\backslash B\} + \boldsymbol{P}\{C \mid B\backslash A\}\,\boldsymbol{P}\{B\backslash A\} \\ & + \boldsymbol{P}\{C \mid A \cap B\}\,\boldsymbol{P}\{A \cap B\} + \boldsymbol{P}\left\{C \mid \overline{A \cup B}\right\}\boldsymbol{P}\left\{\overline{A \cup B}\right\}. \end{aligned}$$

Then

$$\boldsymbol{P}\{A \cap B \mid C\} = \frac{(0.9)(0.08)}{(0.5)(0.12) + (0.8)(0.32) + (0.9)(0.08) + 0} = \underline{0.1856}.$$

$\diamond$

Summary and conclusions

Probability of any event is a number between 0 and 1. The empty event has probability 0, and the sample space has probability 1. There are rules for computing probabilities of unions, intersections, and complements. For a union of disjoint events, probabilities are added. For an intersection of independent events, probabilities are multiplied. Combining these rules, one evaluates reliability of a system given reliabilities of its components.

In the case of equally likely outcomes, probability is a ratio of the number of favorable outcomes to the total number of outcomes. Combinatorics provides tools for computing these numbers in frequent situations involving permutations and combinations, with or without replacement.

Given occurrence of event B, one can compute conditional probability of event

A. Unconditional probability of A can be computed from its conditional probabilities by the Law of Total Probability. The Bayes Rule, often used in testing and diagnostics, relates conditional probabilities of A given B and of B given A.

Questions and exercises

2.1. Out of six computer chips, two are defective. If two chips are randomly chosen for testing (without replacement), compute the probability that both of them are defective. List all the outcomes in the sample space.

2.2. Suppose that after 10 years of service, 40% of computers have problems with motherboards (MB), 30% have problems with hard drives (HD), and 15% have problems with both MB and HD. What is the probability that a 10-year old computer still has fully functioning MB and HD?

2.3. A new computer virus can enter the system through the e-mail or through the internet. There is a 30% chance of receiving this virus through the e-mail. There is a 40% chance of receiving it through the internet. Also, the virus enters the system simultaneously through the e-mail and the internet with probability 0.15. What is the probability that the virus does not enter the system at all?

2.4. Among employees of a certain firm, 70% know C/C++, 60% know Fortran, and 50% know both languages. What portion of programmers

(a) does not know Fortran?

(b) does not know Fortran and C/C++?

(c) knows C/C++ but not Fortran?

(d) knows Fortran but not C/C++?

(e) If someone knows Fortran, what is the probability that he/she knows C/C++ too?

(f) If someone knows C/C++, what is the probability that he/she knows Fortran too?

2.5. A computer program is tested by 3 *independent* tests. When there is an error, these tests will discover it with probabilities 0.2, 0.3, and 0.5, respectively. Suppose that the program contains an error. What is the probability that it will be found by at least one test?

2.6. Under good weather conditions, 80% of flights arrive on time. During bad weather, only 30% of flights arrive on time. Tomorrow, the chance of good weather is 60%. What is the probability that your flight will arrive on time?

2.7. A system may become infected by some spyware through the internet or e-mail. Seventy percent of the time the spyware arrives via the internet, thirty percent of the time via e-mail. If it enters via the internet, the system detects it immediately with probability 0.6. If via e-mail, it is detected with probability 0.8. What percentage of times is this spyware detected?

2.8. A shuttle's launch depends on three key devices that may fail independently of each other with probabilities 0.01, 0.02, and 0.02, respectively. If any of the key devices fails, the launch will be postponed. Compute the probability for the shuttle to be launched on time, according to its schedule.

2.9. Successful implementation of a new system is based on three independent modules. Module 1 works properly with probability 0.96. For modules 2 and 3, these probabilities equal 0.95 and 0.90. Compute the probability that at least one of these three modules fails to work properly.

2.10. Three computer viruses arrived as an e-mail attachment. Virus A damages the system with probability 0.4. Independently of it, virus B damages the system with probability 0.5. Independently of A and B, virus C damages the system with probability 0.2. What is the probability that the system gets damaged?

2.11. A computer program is tested by 5 independent tests. If there is an error, these tests will discover it with probabilities 0.1, 0.2, 0.3, 0.4, and 0.5, respectively. Suppose that the program contains an error. What is the probability that it will be found

(a) by at least one test?

(b) by at least two tests?

(c) by all five tests?

2.12. A building is examined by policemen with four dogs that are trained to detect the scent of explosives. If there are explosives in a certain building, and each dog detects them with probability 0.6, independently of other dogs, what is the probability that the explosives will be detected by at least one dog?

2.13. An important module is tested by three independent teams of inspectors. Each team detects a problem in a defective module with probability 0.8. What is the probability that at least one team of inspectors detects a problem in a defective module?

2.14. A spyware is trying to break into a system by guessing its password. It does not give up until it tries 1 million different passwords. What is the probability that it will guess the password and break in if by rules, the password must consist of

(a) 6 different lower-case letters

(b) 6 different letters, some may be upper-case, and it is case-sensitive

(c) any 6 letters, upper- or lower-case, and it is case-sensitive

(d) any 6 characters including letters and digits

2.15. A computer program consists of two blocks written independently by two different programmers. The first block has an error with probability 0.2. The second block has an error with probability 0.3. If the program returns an error, what is the probability that there is an error in both blocks?

2.16. A computer maker receives parts from three suppliers, S1, S2, and S3. Fifty percent come from S1, twenty percent from S2, and thirty percent from S3. Among all the parts supplied by S1, 5% are defective. For S2 and S3, the portion of defective parts is 3% and 6%, respectively.

(a) What portion of all the parts is defective?

(b) A customer complains that a certain part in her recently purchased computer is defective. What is the probability that it was supplied by S1?

2.17. A computer assembling company receives 24% of parts from supplier X, 36% of parts from supplier Y, and the remaining 40% of parts from supplier Z. Five percent of parts supplied by X, ten percent of parts supplied by Y, and six percent of parts supplied by Z are defective. If an assembled computer has a defective part in it, what is the probability that this part was received from supplier Z?

2.18. A problem on a multiple-choice quiz is answered correctly with probability 0.9 if a student is prepared. An unprepared student guesses between 4 possible answers, so the probability of choosing the right answer is 1/4. Seventy-five percent of students prepare for the quiz. If Mr. X gives a correct answer to this problem, what is the chance that he did not prepare for the quiz?

2.19. At a plant, 20% of all the produced parts are subject to a special electronic inspection. It is known that any produced part which was inspected electronically has no defects with probability 0.95. For a part that was not inspected electronically this probability is only 0.7. A customer receives a part and find defects in it. What is the probability that this part went through an electronic inspection?

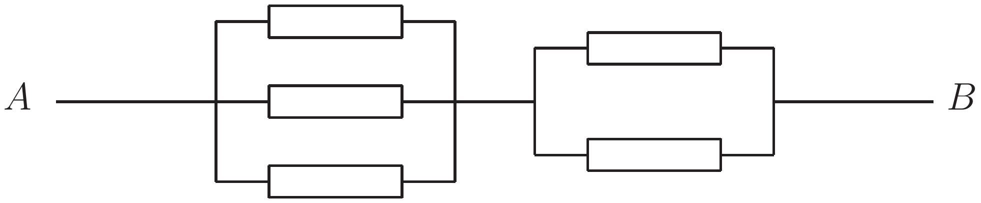


Figure 2.7 *Calculate reliability of this system (Exercise 2.21).*

2.20. All athletes at the Olympic games are tested for performance-enhancing steroid drug use. The imperfect test gives positive results (indicating drug use) for 90% of all steroid-users but also (and incorrectly) for 2% of those who do not use steroids. Suppose that 5% of all registered athletes use steroids. If an athlete is tested negative, what is the probability that he/she uses steroids?

2.21. In the system in Figure 2.7, each component fails with probability 0.3 independently of other components. Compute the system's reliability.

2.22. Three highways connect city A with city B. Two highways connect city B with city C. During a rush hour, each highway is blocked by a traffic accident with probability 0.2, independently of other highways.

(a) Compute the probability that there is at least one open route from A to C.

(b) How will a new highway, also blocked with probability 0.2 independently of other highways, change the probability in (a) if it is built

(α) between A and B?

(β) between B and C?

(γ) between A and C?

2.23. Calculate the reliability of each system shown in Figure 2.8, if components A, B, C, D, and E function properly with probabilities 0.9, 0.8, 0.7, 0.6, and 0.5, respectively.

2.24. Among 10 laptop computers, five are good and five have defects. Unaware of this, a customer buys 6 laptops.

(a) What is the probability of exactly 2 defective laptops among them?

(b) Given that *at least* 2 purchased laptops are defective, what is the probability that *exactly* 2 are defective?

2.25. This is known as *the Birthday Problem.*

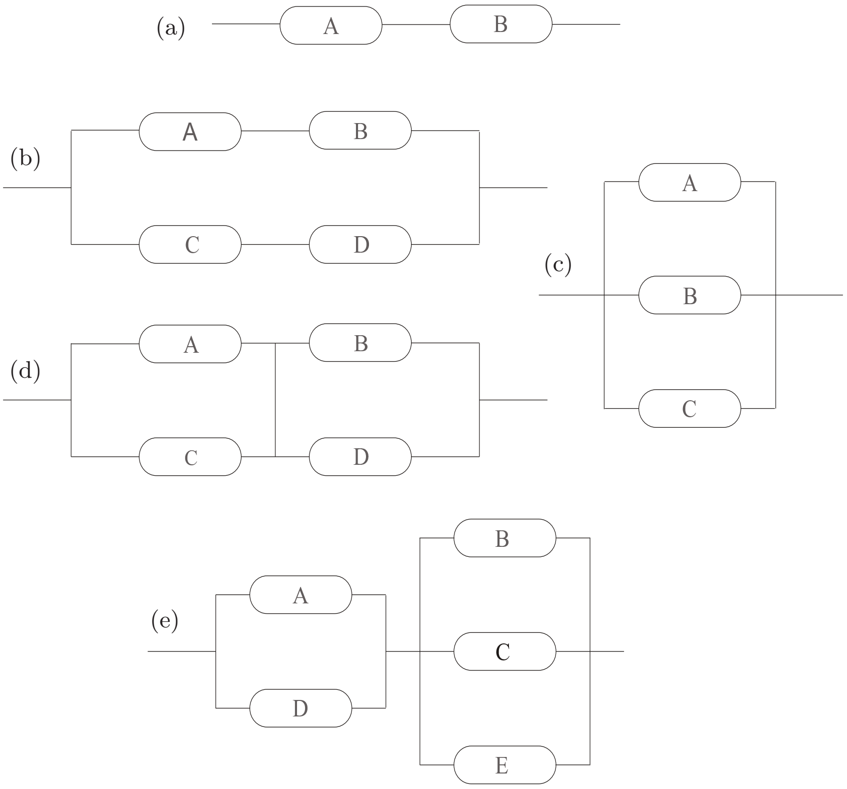


Figure 2.8 *Calculate reliability of each system (Exercise 2.23).*

(a) Consider a class with 30 students. Compute the probability that at least two of them have their birthdays on the same day. (For simplicity, ignore the leap year).

(b) How many students should be in class in order to have this probability above 0.5?

2.26. Two out of six computers in a lab have problems with hard drives. If three computers are selected at random for inspection, what is the probability that none of them has hard drive problems?

2.27. Among eighteen computers in some store, six have defects. Five randomly selected computers are bought for the university lab. Compute the probability that all five computers have no defects.

2.28. A quiz consists of 6 multiple-choice questions. Each question has 4 possible answers. A student is unprepared, and he has no choice but guessing answers completely at random. He passes the quiz if he gets at least 3 questions correctly. What is the probability that he will pass?

2.29. An internet search engine looks for a keyword in 9 databases, searching them in a random order. Only 5 of these databases contain the given keyword. Find the probability that it will be found in at least 2 of the first 4 searched databases.

2.30. Consider the situation described in Example 2.21 on p. 22, but this time let us define the sample space clearly. Suppose that one child is older, and the other is younger, their gender is independent of their age, and the child you meet is one or the other with probabilities 1/2 and 1/2.

(a) List all the outcomes in this sample space. Each outcome should tell the children's gender, which child is older, and which child you have met.

(b) Show that *unconditional* probabilities of outcomes BB, BG, and GB are equal.

(c) Show that *conditional* probabilities of BB, BG, and GB, after you met Jimmy, are not equal.

(d) Show that the *conditional* probability that Jim has a sister is 1/2.

2.31. Show that events $A, B, C, \ldots$ are disjoint if and only if $\overline{A}, \overline{B}, \overline{C}, \ldots$ are exhaustive.

2.32. Events A and B are independent. Show, intuitively and mathematically, that:

(a) Their complements are also independent.

(b) If they are disjoint, then $\boldsymbol{P}\{A\} = 0$ or $\boldsymbol{P}\{B\} = 0$.

(c) If they are exhaustive, then $\boldsymbol{P}\{A\} = 1$ or $\boldsymbol{P}\{B\} = 1$.

2.33. Derive a computational formula for the probability of a union of N arbitrary events. Assume that probabilities of all individual events and their intersections are given.

2.34. Prove that

$$\overline{E_1 \cap \ldots \cap E_n} = \overline{E_1} \cup \ldots \cup \overline{E_n}$$

for arbitrary events $E_1, \ldots, E_n$.

CHAPTER 3

Discrete Random Variables and their Distributions

This chapter introduces the concept of a random variable and studies discrete distributions in detail. Continuous distributions are discussed in Chapter 4.

3.1 Distribution of a random variable

3.1.1 Main concepts

DEFINITION 3.1

A **random variable** is a function of an outcome,

$$X = f(\omega).$$

In other words, it is a quantity that depends on chance.

The domain of a random variable is the sample space Ω. Its range can be the set of all real numbers $\boldsymbol{R}$, or only the positive numbers $(0, +\infty)$, or the integers $\boldsymbol{Z}$, or the interval $(0, 1)$, etc., depending on what possible values the random variable can potentially take.

Once an experiment is completed, and the outcome ω is known, the value of random variable $X(\omega)$ becomes determined.

Example 3.1. Consider an experiment of tossing 3 fair coins and counting the number of heads. Certainly, the same model suits the number of girls in a family with 3 children, the number of 1's in a random binary code consisting of 3 characters, etc.

Let X be the number of heads (girls, 1's). Prior to an experiment, its value is not known. All we can say is that X has to be an integer between 0 and 3. Since assuming each value is an event, we can compute probabilities,

$$\begin{aligned}
\boldsymbol{P}\{X=0\} &= \boldsymbol{P}\{\text{three tails}\} = \boldsymbol{P}\{TTT\} = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \frac{1}{8} \\
\boldsymbol{P}\{X=1\} &= \boldsymbol{P}\{HTT\} + \boldsymbol{P}\{THT\} + \boldsymbol{P}\{TTH\} = \frac{3}{8} \\
\boldsymbol{P}\{X=2\} &= \boldsymbol{P}\{HHT\} + \boldsymbol{P}\{HTH\} + \boldsymbol{P}\{THH\} = \frac{3}{8} \\
\boldsymbol{P}\{X=3\} &= \boldsymbol{P}\{HHH\} = \frac{1}{8}
\end{aligned}$$

Summarizing,

x	$\boldsymbol{P}\{X=x\}$
0	1/8
1	3/8
2	3/8
3	1/8
Total	1

◇

This table contains everything that is known about random variable X prior to the experiment. Before we know the outcome ω, we cannot tell what X equals to. However, we can list all the possible values of X and determine the corresponding probabilities.

DEFINITION 3.2

Collection of all the probabilities related to X is the **distribution** of X. The function

$$P(x) = \boldsymbol{P}\{X=x\}$$

is the **probability mass function**, or **pmf**. The **cumulative distribution function**, or **cdf** is defined as

$$F(x) = \boldsymbol{P}\{X \le x\} = \sum_{y \le x} \boldsymbol{P}(y). \tag{3.1}$$

For every outcome ω, the variable X takes one and only one value x. This makes events $\{X=x\}$ disjoint and exhaustive, and therefore,

$$\sum_x P(x) = \sum_x \boldsymbol{P}\{X=x\} = 1.$$

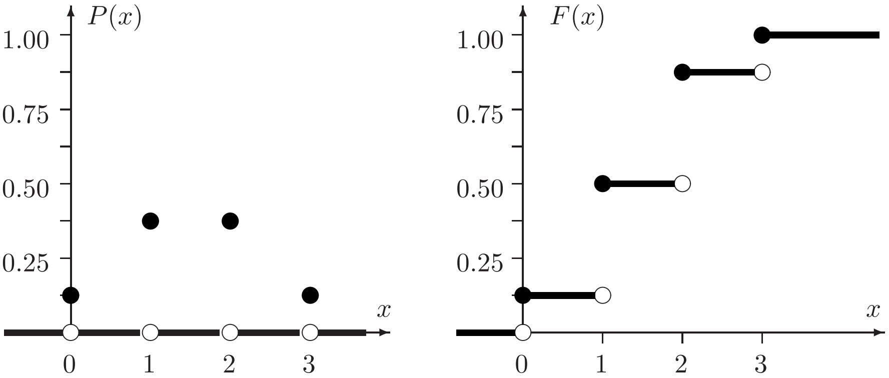


Figure 3.1 *The probability mass function $P(x)$ and the cumulative distribution function $F(x)$. White circles denote excluded points.*

Looking at (3.1), we can conclude that the cdf $F(x)$ is a non-decreasing function of x, always between 0 and 1, with

$$\lim_{x \downarrow -\infty} F(x) = 0 \text{ and } \lim_{x \uparrow +\infty} F(x) = 1.$$

Between any two subsequent values of X, $F(x)$ is constant. It jumps by $P(x)$ at each possible value x of X.

Recall that one way to compute the probability of an event is to add probabilities of all the outcomes in it. Hence, for any set A,

$$\boldsymbol{P}\{X \in A\} = \sum_{x \in A} P(x).$$

When A is an interval, its probability can be computed directly from the cdf $F(x)$,

$$\boldsymbol{P}\{a < X \leq b\} = F(b) - F(a).$$

Example 3.2. The pmf and cdf of X in Example 3.1 are shown in Fig. 3.1. $\Diamond$

MATLAB DEMO. The following MATLAB code simulates 3 coin tosses and computes a value of X 10,000 times.

```
N = 10000;        % Number of simulations
U = rand(3,N);    % a 3-by-N matrix of random numbers from [0,1]
Y = (U < 0.5);    % Y=1 (heads) if U < 0.5, otherwise Y=0 (tails)
X = sum(Y);       % Sums across columns. X = number of heads
hist(X);          % Histogram of X
```

On the obtained histogram, the two middle columns for $X = 1$ and $X = 2$ are about 3 times higher than the columns on each side, for $X = 0$ and $X = 3$. That is, in a run of 10,000 simulations, values 1 and 2 are attained three times more often than 0 and 3. This agrees with the pmf $P(0) = P(3) = 1/8$, $P(1) = P(2) = 3/8$.

Example 3.3 (Errors in independent modules). A program consists of two modules. The number of errors X_1 in the first module has the pmf $P_1(x)$, and the number of errors X_2 in the second module has the pmf $P_2(x)$, independently of X_1, where

x	$P_1(x)$	$P_2(x)$
0	0.5	0.7
1	0.3	0.2
2	0.1	0.1
3	0.1	0

Find the pmf and cdf of $Y = X_1 + X_2$, the total number of errors.

Solution. We break the problem into steps. First, determine all possible values of Y, then compute the probability of each value. Clearly, the number of errors Y is integer that can be as low as $0 + 0 = 0$ and as high as $3 + 2 = 5$. Since $P_2(3) = 0$, the second module has at most 2 errors. Next,

$$\begin{array}{lclcl}
P_Y(0) & = & P\{Y=0\} & = & \boldsymbol{P}\{X_1 = X_2 = 0\} = P_1(0)P_2(0) \\
 & & & = & (0.5)(0.7) = 0.35 \\
P_Y(1) & = & P\{Y=1\} & = & P_1(0)P_2(1) + P_1(1)P_2(0) \\
 & & & = & (0.5)(0.2) + (0.3)(0.7) = 0.31 \\
P_Y(2) & = & P\{Y=2\} & = & P_1(0)P_2(2) + P_1(1)P_2(1) + P_1(2)P_2(0) \\
 & & & = & (0.5)(0.1) + (0.3)(0.2) + (0.1)(0.7) = 0.18 \\
P_Y(3) & = & P\{Y=3\} & = & P_1(1)P_2(2) + P_1(2)P_2(1) + P_1(3)P_2(0) \\
 & & & = & (0.3)(0.1) + (0.1)(0.2) + (0.1)(0.7) = 0.12 \\
P_Y(4) & = & P\{Y=4\} & = & P_1(2)P_2(2) + P_1(3)P_2(1) \\
 & & & = & (0.1)(0.1) + (0.1)(0.2) = 0.03 \\
P_Y(5) & = & P\{Y=5\} & = & P_1(3)P_2(2) = (0.1)(0.1) = 0.01
\end{array}$$

Now check:

$$\sum_{y=0}^{5} P_Y(y) = 0.35 + 0.31 + 0.18 + 0.12 + 0.03 + 0.01 = 1,$$

thus we *probably* counted all the possibilities and did not miss any (we just wanted to emphasize that simply getting $\sum P(x) = 1$ does not guarantee that we made no mistake in our solution. However, if this equality is not satisfied, we have a mistake for sure).

The cumulative function can be computed as

$$
\begin{aligned}
F_Y(0) &= P_Y(0) = 0.35 \\
F_Y(1) &= F_Y(0) + P_Y(1) = 0.35 + 0.31 = 0.66 \\
F_Y(2) &= F_Y(1) + P_Y(2) = 0.66 + 0.18 = 0.84 \\
F_Y(3) &= F_Y(2) + P_Y(3) = 0.84 + 0.12 = 0.96 \\
F_Y(4) &= F_Y(3) + P_Y(4) = 0.96 + 0.03 = 0.99 \\
F_Y(5) &= F_Y(4) + P_Y(5) = 0.99 + 0.01 = 1.00
\end{aligned}
$$

Between the values of Y, $F(x)$ is constant. ◇

3.1.2 Types of random variables

So far, we are dealing with *discrete random variables*. These are variables whose range is finite or countable. In particular, it means that their values can be listed, or arranged in a sequence. Examples include the number of jobs submitted to a printer, number of errors, number of error-free modules, number of failed components. Discrete variables don't have to be integer. For example, the *proportion* of defective components in a lot of 100 can be 0, 1/100, 2/100, ..., 99/100, or 1. This variable assumes a finite number, 101 different values, so it is discrete, although not integer.

On the contrary, *continuous random variables* assume the whole interval of values. This could be a bounded interval (a, b), or an unbounded interval $(a, +\infty)$, $(-\infty, b)$, or $(-\infty, +\infty)$. Sometimes, it may be a union of several such intervals. Intervals are uncountable, therefore, all values of a random variable cannot be listed in this case. Examples of continuous variables include various times (software installation time, code execution time, connection time, waiting time, lifetime), also physical variables like weight, height, voltage, temperature, distance, the number of miles per gallon, etc. We shall discuss continuous random variables in detail in Chapter 4.

Example 3.4. For comparison, observe that the long jump is formally a continuous random variable because an athlete can jump any distance within some range. Results of a high jump, however, are discrete because the bar can only be placed on a finite number of heights. ◇

Notice that rounding a continuous random variable, say, to the nearest integer makes it discrete.

Sometimes we can see *mixed random variables* that are discrete on some range of values and continuous elsewhere.

Example 3.5. A job is sent to a printer. Let X be the waiting time before the job starts printing. With some probability, this job appears first in line and starts printing immediately, $X = 0$. It is also possible that the job is first in line but it takes 20 seconds for the printer to warm up, in which case $X = 20$. So far, the variable has a discrete behavior with a positive pmf $P(x)$ at $x = 0$ and $x = 20$. However, if there are other jobs in a queue, then X depends on the time it takes to print them, which is a continuous random variable. Using a popular jargon, besides *"point masses"* at $x = 0$ and $x = 20$, the variable is continuous, taking values in $(0, +\infty)$. Thus, X is neither discrete nor continuous. It is mixed. ◇

3.2 Distribution of a random vector

Often we deal with several random variables simultaneously. We may look at the size of a RAM and the speed of a CPU, the price of a computer and its capacity, temperature and humidity, technical and artistic performance, etc.

3.2.1 Joint distribution and marginal distributions

DEFINITION 3.3

If X and Y are random variables, then the pair (X, Y) is a **random vector**. Its distribution is called the **joint distribution** of X and Y. Individual distributions of X and Y are then called the **marginal distributions**.

Although we talk about two random variables in this section, all the concepts extend to a vector $(X_1, X_2, \ldots, X_n)$ of n components and its joint distribution.

Similarly to a single variable, the *joint distribution* of a vector is a collection of probabilities for a vector (X, Y) to take a value (x, y). Recall that two vectors are equal,

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

if $X = x$ and $Y = y$. This "and" means the intersection, therefore, the *joint probability mass function* of X and Y is

$$P(x, y) = \boldsymbol{P}\left\{(X, Y) = (x, y)\right\} = \boldsymbol{P}\left\{X = x \cap Y = y\right\}.$$

Again, $\{(X, Y) = (x, y)\}$ are exhaustive and mutually exclusive events for different pairs (x, y), therefore,

$$\sum_x \sum_y P(x, y) = 1.$$

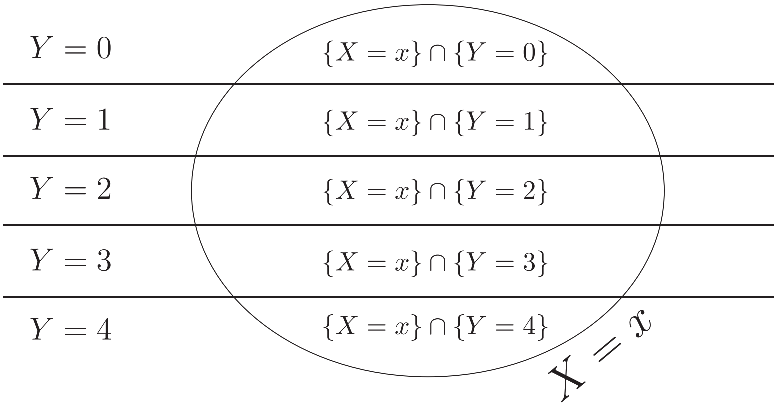

$Y = 0$
$\{X = x\} \cap \{Y = 0\}$
$Y = 1$
$\{X = x\} \cap \{Y = 1\}$
$Y = 2$
$\{X = x\} \cap \{Y = 2\}$
$Y = 3$
$\{X = x\} \cap \{Y = 3\}$
$Y = 4$
$\{X = x\} \cap \{Y = 4\}$
$X = x$

Figure 3.2 *Addition Rule: computing marginal probabilities from the joint distribution.*

The joint distribution of (X, Y) carries the full information about the behavior of this random vector. In particular, the marginal probability mass functions of X and Y can be obtained from the joint pmf by the Addition Rule.

Addition Rule

$$\boxed{\begin{array}{rcccl} P_X(x) & = & \boldsymbol{P}\{X = x\} & = & \sum_y P_{(X,Y)}(x, y) \\ P_Y(y) & = & \boldsymbol{P}\{Y = y\} & = & \sum_x P_{(X,Y)}(x, y) \end{array}} \tag{3.2}$$

That is, to get the marginal pmf of one variable, we add the joint probabilities over all values of the other variable.

The Addition Rule is illustrated in Figure 3.2. Events $\{Y = y\}$ for different values of y partition the sample space Ω. Hence, their intersections with $\{X = x\}$ partition the event $\{X = x\}$ into mutually exclusive parts. By the rule for the union of mutually exclusive events, formula (2.4) on p. 16, their probabilities should be added. These probabilities are precisely $P_{(X,Y)}(x, y)$.

In general, the joint distribution cannot be computed from marginal distributions because they carry no information about interrelations between random variables. For example, marginal distributions cannot tell whether variables X and Y are independent or dependent.

3.2.2 Independence of random variables

DEFINITION 3.4

Random variables X and Y are **independent** if

$$P_{(X,Y)}(x,y) = P_X(x)P_Y(y)$$

for *all* values of x and y. This means, events $\{X = x\}$ and $\{Y = y\}$ are independent for all x and y; in other words, variables X and Y take their values independently of each other.

In problems, to show independence of X and Y, we have to check whether the joint pmf factors into the product of marginal pmfs for *all* pairs x and y. To prove dependence, we only need to present one counterexample, a pair (x, y) with $P(x, y) \neq P_X(x)P_Y(y)$.

Example 3.6. A program consists of two modules. The number of errors, X, in the first module and the number of errors, Y, in the second module have the joint distribution, $P(0, 0) = P(0, 1) = P(1, 0) = 0.2$, $P(1, 1) = P(1, 2) = P(1, 3) = 0.1$, $P(0, 2) = P(0, 3) = 0.05$. Find (a) the marginal distributions of X and Y, (b) the probability of no errors in the first module, and (c) the distribution of the total number of errors in the program. Also, (d) find out if errors in the two modules occur independently.

Solution. It is convenient to organize the joint pmf of X and Y in a table. Adding rowwise and columnwise, we get the marginal pmfs,

$P_{(X,Y)}(x,y)$		y				$P_X(x)$
		0	1	2	3	
x	0	0.20	0.20	0.05	0.05	0.50
	1	0.20	0.10	0.10	0.10	0.50
$P_Y(y)$		0.40	0.30	0.15	0.15	1.00

This solves (a).

(b) $P_X(0) = 0.50$.

(c) Let $Z = X + Y$ be the total number of errors. To find the distribution of Z, we first identify its possible values, then find the probability of each value. We see that Z can be as small as 0 and as large as 4. Then,

$$\begin{aligned}
P_Z(0) &= \boldsymbol{P}\{X + Y = 0\} = \boldsymbol{P}\{X = 0 \cap Y = 0\} = P(0, 0) = 0.20,\\
P_Z(1) &= \boldsymbol{P}\{X = 0 \cap Y = 1\} + \boldsymbol{P}\{X = 1 \cap Y = 0\}\\
&= P(0, 1) + P(1, 0) = 0.20 + 0.20 = 0.40,
\end{aligned}$$

$$\begin{aligned}
P_Z(2) &= P(0,2)+P(1,1)=0.05+0.10=0.15,\\
P_Z(3) &= P(0,3)+P(1,2)=0.05+0.10=0.15,\\
P_Z(4) &= P(1,3)=0.10.
\end{aligned}$$

It is a good check to verify that $\sum_z P_Z(z)=1$.

(d) To verify independence of X and Y, check if their joint pmf factors into a product of marginal pmfs. We see that $P_{(X,Y)}(0,0)=0.2$ indeed equals $P_X(0)P_Y(0)=(0.5)(0.4)$. Keep checking... Next, $P_{(X,Y)}(0,1)=0.2$ whereas $P_X(0)P_Y(1)=(0.5)(0.3)=0.15$. There is no need to check further. We found a pair of x and y that violates the formula for independent random variables. Therefore, the numbers of errors in two modules are dependent. ◇

3.3 Expectation and variance

The distribution of a random variable or a random vector, the full collection of related probabilities, contains the entire information about its behavior. This detailed information can be summarized in a few vital characteristics describing the average value, the most likely value of a random variable, its spread, variability, etc. The most commonly used are the *expectation*, *variance*, *standard deviation*, *covariance*, and *correlation*, introduced in this section. Also rather popular and useful are the *mode*, *moments*, *quantiles*, and *interquartile range* that we discuss in Sections 8.1 and 9.1.1.

3.3.1 Expectation

DEFINITION 3.5

Expectation of a random variable X is its mean, the average value.

Example 3.7. Consider a variable that takes values 0 and 1 with probabilities $P(0)=P(1)=0.5$. That is,

$$X=\begin{cases} 0 & \text{with probability } 1/2\\ 1 & \text{with probability } 1/2\end{cases}$$

Observing this variable many times, we shall see $X=0$ about 50% of times and $X=1$ about 50% of times. The average value of X will then be close to 0.5, so it is reasonable to have $\mathbf{E}(X)=0.5$. ◇

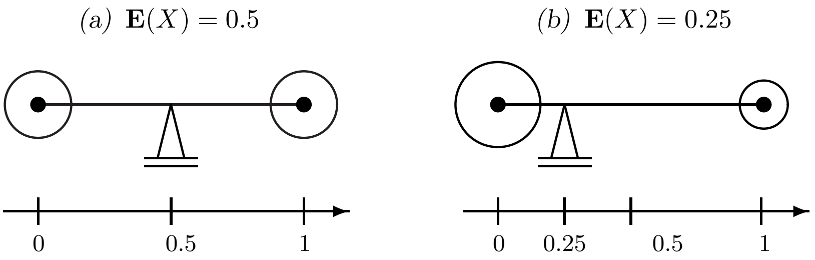


Figure 3.3 *Expectation as a center of gravity.*

It is wrong to think that $\mathbf{E}(X)$ is always the average of all possible values of X. Probabilities (frequencies) of these values should also be taken into account.

Example 3.8. Suppose now that $P(0) = 0.75$ and $P(1) = 0.25$. Then, in a long run, X is equal 1 only 1/4 of times, otherwise it equals 0. Suppose we earn \$1 every time we see $X = 1$. On the average, we earn \$1 every four times, or \$0.25 per each observation. Therefore, in this case $\mathbf{E}(X) = 0.25$. ◇

A physical model for these two examples is shown in Figure 3.3. In Figure 3.3a, we put two equal masses, 0.5 units each, at points 0 and 1 and connect them with a firm but weightless rod. The masses represent probabilities $P(0)$ and $P(1)$. Now we look for a point at which the system will be balanced. It is symmetric, hence its balance point, *the center of gravity*, is in the middle, 0.5.

Figure 3.3b represents a model for the second example. Here the masses at 0 and 1 equal 0.75 and 0.25 units, respectively, according to $P(0)$ and $P(1)$. This system is balanced at 0.25, which is also its center of gravity.

Similar arguments can be used to derive the general formula for the expectation.

Expectation, discrete case

$$\boxed{\mathbf{E}(X) = \sum_x xP(x)} \tag{3.3}$$

This formula returns the center of gravity for a system with masses $P(x)$ allocated at points x.

In a certain sense, expectation is the best forecast of X. The variable itself is random. It takes different values with different probabilities $P(x)$. At the same time, it has just one expectation $\mathbf{E}(X)$ which is non-random.

3.3.2 Expectation of a function

Often we are interested in another variable, Y, that is a function of X. For example, downloading time depends on the connection speed, profit of a computer store depends on the number of computers sold, and bonus of its manager depends on this profit. Expectation of $Y = g(X)$ is computed by a similar formula,

$$\mathbf{E}\left\{g(X)\right\} = \sum_x g(x)P(x). \tag{3.4}$$

Remark: Indeed, if g is a one-to-one function, then Y takes each value $y = g(x)$ with probability $P(x)$, and the formula for $\mathbf{E}(Y)$ can be applied directly. If g is not one-to-one, then some values of $g(x)$ will be repeated in (3.4). However, they are multiplied by the corresponding probabilities. When we add in (3.4), these probabilities are also added, thus each value of $g(x)$ is still multiplied by the probability $P_Y(g(x))$.

3.3.3 Properties

The following *linear* properties of expectations follow directly from (3.3) and (3.4). For *any* random variables X and Y and any non-random numbers a, b, and c, we have

Properties of expectations

$$\boxed{\begin{array}{lcl} \mathbf{E}(aX+bY+c) & = & a\,\mathbf{E}(X)+b\,\mathbf{E}(Y)+c \\ \text{In particular,} & & \\ \quad\mathbf{E}(X+Y) & = & \mathbf{E}(X)+\mathbf{E}(Y) \\ \quad\mathbf{E}(aX) & = & a\,\mathbf{E}(X) \\ \quad\mathbf{E}(c) & = & c \\ \text{For } \mathbf{independent}\ X \text{ and } Y, & & \\ \quad\mathbf{E}(XY) & = & \mathbf{E}(X)\,\mathbf{E}(Y) \end{array}} \tag{3.5}$$

Proof: The first property follows from the Addition Rule (3.2). For any X and Y,

$$\begin{aligned} \mathbf{E}(aX+bY+c) &= \sum_x\sum_y (ax+by+c)P_{(X,Y)}(x,y) \\ &= \sum_x ax\sum_y P_{(X,Y)}(x,y) + \sum_y by\sum_x P_{(X,Y)}(x,y) + c\sum_x\sum_y P_{(X,Y)}(x,y) \\ &= a\sum_x xP_X(x) + b\sum_y yP_Y(y) + c. \end{aligned}$$

The next 3 equalities are special cases. To prove the last property, we recall that $P_{(X,Y)}(x,y) = P_X(x)P_Y(y)$, therefore,

$$\mathbf{E}(XY) = \sum_x\sum_y (xy)P_X(x)P_Y(y) = \sum_x xP_X(x)\sum_y yP_Y(y) = \mathbf{E}(X)\,\mathbf{E}(Y). \qquad \square$$

Remark: The last property in (3.5) holds for some dependent variables too, hence it cannot be used to verify independence of X and Y.

Example 3.9. In Example 3.6 on p. 48,

$$\begin{aligned} \mathbf{E}(X) &= (0)(0.5) + (1)(0.5) = 0.5 \text{ and} \\ \mathbf{E}(Y) &= (0)(0.4) + (1)(0.3) + (2)(0.15) + (3)(0.15) = 1.05, \end{aligned}$$

therefore, the expected total number of errors is

$$\mathbf{E}(X + Y) = 0.5 + 1.05 = 1.65.$$

◇

Remark: Clearly, the program will never have 1.65 errors, because the number of errors is always integer. Then, should we round 1.65 to 2 errors? Absolutely not, it would be a mistake. Although both X and Y are integers, their expectations, or average values, do not have to be integers at all.

3.3.4 Variance and standard deviation

Expectation shows where the average value of a random variable is located, or where the variable is *expected* to be, plus or minus some error. How large could this "error" be, and how much can a variable *vary* around its expectation? Here we introduce measures of variability.

Example 3.10. Here is a rather artificial but illustrative scenario. Consider two users. One receives either 48 or 52 e-mail messages per day, with a 50-50% chance of each. The other receives either 0 or 100 e-mails, also with a 50-50% chance. What is a common feature of these two distributions, and how are they different?

We see that both users receive the same average number of e-mails:

$$\mathbf{E}(X) = \mathbf{E}(Y) = 50.$$

However, in the first case, the actual number of e-mails is always close to 50, whereas it always differs from it by 50 in the second case. The first random variable, X, is more stable, it has *low variability*. The second variable, Y, has *high variability*. ◇

This example shows that variability of a random variable is measured by its distance from the mean $\mu = \mathbf{E}(X)$. In its turn, this distance is random too, and therefore, cannot serve as a characteristic of a distribution. It remains to square it and take the expectation of the result.

DEFINITION 3.6

Variance of a random variable is defined as the expected squared deviation from the mean. For discrete random variables,

$$\mathrm{Var}(X) = \mathbf{E}\,(X - \mathbf{E}X)^2 = \sum_x (x-\mu)^2 P(x)$$

Remark: Notice that if the distance to the mean is not squared, then the result is always $\mu - \mu = 0$ bearing no information about the distribution of X.

According to this definition, variance is always non-negative. Further, it equals 0 only if $x = \mu$ for all values of x, i.e., when X is constantly equal μ. Certainly, a constant (non-random) variable has zero variability.

Variance can also be computed as

$$\mathrm{Var}(X) = \mathbf{E}(X^2) - \mu^2, \tag{3.6}$$

a proof of this is left as an exercise.

DEFINITION 3.7

Standard deviation is a square root of variance,

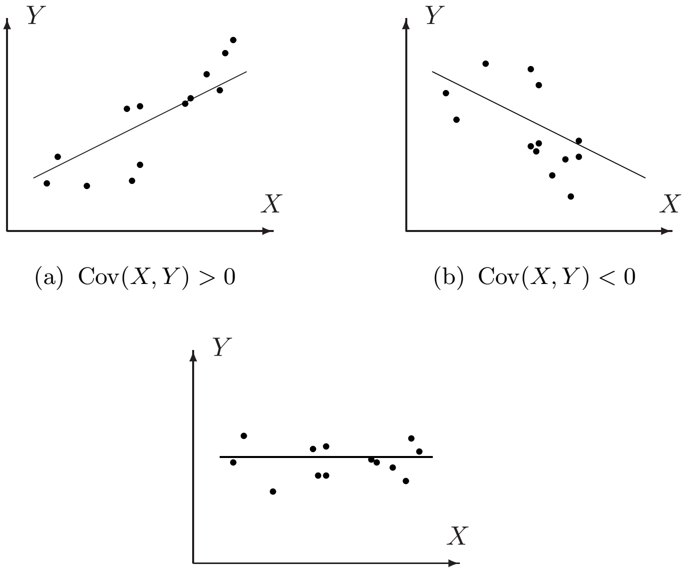
$$\mathrm{Std}(X) = \sqrt{\mathrm{Var}(X)}$$

Continuing the Greek-letter tradition, variance is often denoted by σ^2. Then, standard deviation is σ.

If X is measured in some units, then its mean μ has the same measurement unit as X. Variance σ^2 is measured in *squared units*, and therefore, it cannot be compared with X or μ. No matter how funny it sounds, it is rather normal to measure variance of profit in *squared dollars*, variance of class enrollment in *squared students*, and variance of available disk space in *squared* megabytes. When a squared root is taken, the resulting standard deviation σ is again measured in the same units as X. This is the main reason of introducing yet another measure of variability, σ.

3.3.5 Covariance and correlation

Expectation, variance, and standard deviation characterize the distribution of a single random variable. Now we introduce measures of *association* of two random variables.



(a) Cov$(X, Y) > 0$

(b) Cov$(X, Y) < 0$

(c) Cov$(X, Y) = 0$

Figure 3.4 *Positive, negative, and zero covariance.*

DEFINITION 3.8

Covariance $\sigma_{XY} = \text{Cov}(X, Y)$ is defined as

$$\begin{aligned}\text{Cov}(X, Y) &= \mathbf{E}\{(X - \mathbf{E}X)(Y - \mathbf{E}Y)\} \\ &= \mathbf{E}(XY) - \mathbf{E}(X)\,\mathbf{E}(Y)\end{aligned}$$

It summarizes interrelation of two random variables.

Covariance is the expected product of deviations of X and Y from their respective expectations. If Cov$(X, Y) > 0$, then positive deviations $(X - \mathbf{E}X)$ are more likely to be multiplied by positive $(Y - \mathbf{E}Y)$, and negative $(X - \mathbf{E}X)$ are more likely to be multiplied by negative $(Y - \mathbf{E}Y)$. In short, large X imply large Y, and small X imply small Y. These variables are *positively correlated*, Figure 3.4a.

Conversely, Cov$(X, Y) < 0$ means that large X generally correspond to small Y and small X correspond to large Y. These variables are *negatively correlated*, Figure 3.4b.

If Cov$(X, Y) = 0$, we say that X and Y are *uncorrelated*, Figure 3.4c.

DEFINITION 3.9

Correlation coefficient between variables X and Y is defined as

$$\rho = \frac{\mathrm{Cov}(X,Y)}{(\,\mathrm{Std}X)(\,\mathrm{Std}Y)}$$

Correlation coefficient is a rescaled, normalized covariance. Notice that covariance $\mathrm{Cov}(X,Y)$ has a measurement unit. It is measured in units of X multiplied by units of Y. As a result, it is not clear from its value whether X and Y are strongly or weakly correlated. Really, one has to compare $\mathrm{Cov}(X,Y)$ with the magnitude of X and Y. Correlation coefficient performs such a comparison.

How do we interpret the value of ρ? What possible values can it take?

As a special case of famous *Cauchy-Schwarz inequality*,

$$-1 \le \rho \le 1,$$

where $|\rho| = 1$ is possible only when all values of X and Y lie on a straight line (see Figure 3.5). Further, values of ρ near 1 indicate strong positive correlation, values near (-1) show strong negative correlation, and values near 0 show weak correlation or no correlation.

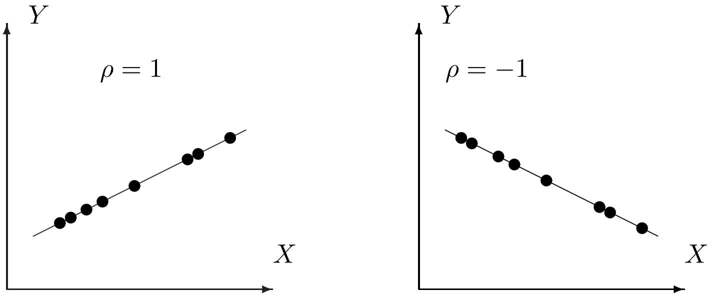


Figure 3.5 *Perfect correlation:* $\rho = \pm 1$.

3.3.6 Properties

For any random variables X, Y, Z, W, and any non-random numbers a, b, c, d,

Properties of variances and covariances

$$\boxed{\begin{array}{lcl}
\multicolumn{3}{l}{\mathrm{Var}(aX+bY+c) = a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y) + 2ab\,\mathrm{Cov}(X,Y)} \\
\multicolumn{3}{l}{\mathrm{Cov}(aX+bY, cZ+dW)} \\
\multicolumn{3}{l}{\quad = ac\,\mathrm{Cov}(X,Z) + ad\,\mathrm{Cov}(X,W) + bc\,\mathrm{Cov}(Y,Z) + bd\,\mathrm{Cov}(Y,W)} \\
\mathrm{Cov}(X,Y) & = & \mathrm{Cov}(Y,X) \\
\rho(X,Y) & = & \rho(Y,X) \\
\multicolumn{3}{l}{\text{In particular,}} \\
\mathrm{Var}(aX+b) & = & a^2\,\mathrm{Var}(X) \\
\mathrm{Cov}(aX+b, cY+d) & = & ac\,\mathrm{Cov}(X,Y) \\
\rho(aX+b, cY+d) & = & \rho(X,Y) \\
\multicolumn{3}{l}{\text{For independent } X \text{ and } Y,} \\
\mathrm{Cov}(X,Y) & = & 0 \\
\mathrm{Var}(X+Y) & = & \mathrm{Var}(X) + \mathrm{Var}(Y)
\end{array}} \tag{3.7}$$

Proof: To prove the first two formulas, we only need to multiply parentheses in the definitions of variance and covariance and apply (3.5):

$$\begin{aligned}
&\mathrm{Var}(aX+bY+c) = \mathbf{E}\left\{aX+bY+c - \mathbf{E}(aX+bY+c)\right\}^2 \\
&= \mathbf{E}\left\{(aX - a\,\mathbf{E}X) + (bY - b\,\mathbf{E}Y) + (c-c)\right\}^2 \\
&= \mathbf{E}\left\{a(X-\mathbf{E}X)\right\}^2 + \mathbf{E}\left\{b(Y-\mathbf{E}Y)\right\}^2 + \mathbf{E}\left\{a(X-\mathbf{E}X)b(Y-\mathbf{E}Y)\right\} \\
&\quad + \mathbf{E}\left\{b(Y-\mathbf{E}Y)a(X-\mathbf{E}X)\right\} \\
&= a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y) + 2ab\,\mathrm{Cov}(X,Y).
\end{aligned}$$

A formula for $\mathrm{Cov}(aX+bY, cZ+dW)$ is proved similarly and is left as an exercise.

For independent X and Y, we have $\mathbf{E}(XY) = \mathbf{E}(X)\,\mathbf{E}(Y)$ from (3.5). Then, according to the definition of covariance, $\mathrm{Cov}(X,Y) = 0$.

The other formulas follow directly from general cases, and their proofs are omitted.

□

We see that *independent* variables are always *uncorrelated*. The reverse is not always true. There exist some variables that are uncorrelated but not independent.

Notice that adding a constant does not affect the variables' variance or covariance. It *shifts* the whole distribution of X without changing its variability or degree of dependence of another variable. The correlation coefficient does not change even when multiplied by a constant because it is recomputed to the unit scale, due to $\mathrm{Std}(X)$ and $\mathrm{Std}(Y)$ in the denominator in Definition 3.9.

Example 3.11. Continuing Example 3.6, we compute

x	$P_X(x)$	$xP_X(x)$	$x - \mathbf{E}X$	$(x - \mathbf{E}X)^2 P_X(x)$
0	0.5	0	-0.5	0.125
1	0.5	0.5	0.5	0.125
$\mu_X = 0.5$			$\sigma_X^2 = 0.25$	

and (using the second method of computing variances)

y	$P_Y(y)$	$yP_Y(y)$	y^2	$y^2 P_Y(y)$
0	0.4	0	0	0
1	0.3	0.3	1	0.3
2	0.15	0.3	4	0.6
3	0.15	0.45	9	1.35
$\mu_Y = 1.05$			$\mathbf{E}(Y^2) = 2.25$	

Result: $\text{Var}(X) = 0.25$, $\text{Var}(Y) = 2.25 - 1.05^2 = 1.1475$, $\text{Std}(X) = \sqrt{0.25} = 0.5$, and $\text{Std}(Y) = \sqrt{1.1475} = 1.0712$.

Also,

$$\mathbf{E}(XY) = \sum_x \sum_y xyP(x,y) = (1)(1)(0.1) + (1)(2)(0.1) + (1)(3)(0.1) = 0.6$$

(the other five terms in this sum are 0). Therefore,

$$\text{Cov}(X,Y) = \mathbf{E}(XY) - \mathbf{E}(X)\,\mathbf{E}(Y) = 0.6 - (0.5)(1.05) = 0.075$$

and

$$\rho = \frac{\text{Cov}(X,Y)}{(\,\text{Std}X)(\,\text{Std}Y)} = \frac{0.075}{(0.5)(1.0712)} = 0.1400.$$

Thus, the numbers of errors in two modules are *positively and not very strongly correlated.* ◇

Notation

μ or $\mathbf{E}(X)$	=	expectation
σ_X^2 or $\text{Var}(X)$	=	variance
σ_X or $\text{Std}(X)$	=	standard deviation
σ_{XY} or $\text{Cov}(X,Y)$	=	covariance
ρ_{XY}	=	correlation coefficient

3.3.7 Chebyshev's inequality

Knowing just the expectation and variance, one can find the range of values most likely taken by this variable. Russian mathematician *Pafnuty Chebyshev* (1821–1894) showed any random variable X with expectation $\mu = \mathbf{E}(X)$ and variance $\sigma^2 = \text{Var}(X)$ belongs to the interval $\mu \pm \varepsilon = [\mu - \varepsilon, \mu + \varepsilon]$ with probability of at least $1 - (\sigma/\varepsilon)^2$. That is,

Chebyshev's inequality

$$\boxed{\begin{array}{c} \boldsymbol{P}\left\{|X-\mu|>\varepsilon\right\} \le \left(\frac{\sigma}{\varepsilon}\right)^2 \\ \\ \text{for any distribution with expectation } \mu \\ \text{and variance } \sigma^2 \text{ and any positive } \varepsilon. \end{array}} \qquad (3.8)$$

Proof: Here we consider discrete random variables. For other types, the proof is similar. According to Definition 3.6,

$$\begin{aligned} \sigma^2 &= \sum_{\text{all } x}(x-\mu)^2 P(x) \ge \sum_{\text{only } x:|x-\mu|>\varepsilon} (x-\mu)^2 P(x) \\ &\ge \sum_{x:|x-\mu|>\varepsilon} \varepsilon^2 P(x) = \varepsilon^2 \sum_{x:|x-\mu|>\varepsilon} P(x) = \varepsilon^2 \boldsymbol{P}\left\{|x-\mu|>\varepsilon\right\}. \end{aligned}$$

Hence, $\boldsymbol{P}\left\{|x-\mu|>\varepsilon\right\} \le \varepsilon^2/\sigma^2$. □

Chebyshev's inequality shows that only a large variance may allow a variable X to differ significantly from its expectation μ. In this case, the *risk* of seeing an extremely low or extremely high value of X increases. For this reason, risk is often measured in terms of a variance or standard deviation.

Example 3.12. Suppose the number of errors in a new software has expectation $\mu = 20$ and a standard deviation of 2. According to (3.8), there are more than 30 errors with probability

$$\boldsymbol{P}\left\{X>30\right\} \le \boldsymbol{P}\left\{|X-20|>10\right\} \le \left(\frac{2}{10}\right)^2 = 0.04.$$

However, if the standard deviation is 5 instead of 2, then the probability of more than 30 errors can only be bounded by $\left(\frac{5}{10}\right)^2 = 0.25$. ◇

Chebyshev's inequality is universal because it works for *any* distribution. Often it gives a rather loose bound for the probability of $|X-\mu|>\varepsilon$. With more information about the distribution, this bound may be improved.

3.3.8 Application to finance

Chebyshev's inequality shows that in general, higher variance implies higher probabilities of *large deviations* increasing the *risk* for a random variable to take values far from its expectation.

This finds a number of immediate applications. Here we focus on evaluating risks of financial deals, allocating funds, and constructing optimal portfolios.

This application is intuitively simple. The same methods can be used for the optimal allocation of computer memory, CPU time, customer support, or other resources.

Example 3.13 (Construction of an optimal portfolio). We would like to invest \$10,000 into shares of companies XX and YY. Shares of XX cost \$20 per share. The market analysis shows that their expected return is \$1 per share with a standard deviation of \$0.5. Shares of YY cost \$50 per share, with an expected return of \$2.50 and a standard deviation of \$1 per share, and returns from the two companies are independent. In order to maximize the expected return and minimize the risk (standard deviation or variance), is it better to invest (A) all \$10,000 into XX, (B) all \$10,000 into YY, or (C) \$5,000 in each company?

Solution. Let X be the actual (random) return from each share of XX, and Y be the actual return from each share of YY. Compute the expectation and variance of the return for each of the proposed portfolios (A, B, and C).

(a) At \$20 a piece, we can use \$10,000 to buy 500 shares of XX, thus $A = 500X$. Using (3.5) and (3.7),

$$\begin{aligned} \mathbf{E}(A) &= 500\,\mathbf{E}(X) = (500)(1) = 500; \\ \mathrm{Var}(A) &= 500^2\,\mathrm{Var}(X) = 500^2(0.5)^2 = 62,500. \end{aligned}$$

(b) Investing all \$10,000 into YY, we buy 10,000/50=200 shares of it, so that $B = 200Y$,

$$\begin{aligned} \mathbf{E}(B) &= 200\,\mathbf{E}(Y) = (200)(2.50) = 500; \\ \mathrm{Var}(A) &= 200^2\,\mathrm{Var}(Y) = 200^2(1)^2 = 40,000. \end{aligned}$$

(c) Investing \$5,000 into each company makes a portfolio consisting of 250 shares of XX and 100 shares of YY, so that $C = 250X + 100Y$. Since independence yields uncorrelation,

$$\begin{aligned} \mathbf{E}(C) &= 250\,\mathbf{E}(X) + 100\,\mathbf{E}(Y) = 250 + 250 = 500; \\ \mathrm{Var}(C) &= 250^2\,\mathrm{Var}(X) + 100^2\,\mathrm{Var}(Y) = 250^2(0.5)^2 + 100^2(1)^2 = 25,625. \end{aligned}$$

Result: The expected return is the same for each of the proposed three portfolios because each share of each company is expected to return 1/20 or 2.50/50, which is 5%. In terms of the expected return, all three portfolios are *equivalent*. Portfolio C, where investment is split between two companies, has the lowest variance, therefore, it is the least risky. This supports one of the basic principles in finance: *to minimize the risk, diversify the portfolio.* ◇

Example 3.14 (Optimal portfolio, correlated returns). Suppose now that the individual stock returns X and Y are no longer independent.

If the correlation coefficient is $\rho = 0.4$, how will it change the results of the previous example? What if they are negatively correlated with $\rho = -0.2$?

Solution. Only the volatility of the diversified portfolio C changes due to the correlation coefficient. Now $\operatorname{Cov}(X,Y) = \rho \operatorname{Std}(X) \operatorname{Std}(Y) = (0.4)(0.5)(1) = 0.2$, thus the variance of C increases by $2(250)(100)(0.2) = 10,000$,

$$\begin{aligned}\operatorname{Var}(C) &= \operatorname{Var}(250X + 100Y)\\ &= 250^2 \operatorname{Var}(X) + 100^2 \operatorname{Var}(Y) + 2(250)(100) \operatorname{Cov}(X,Y)\\ &= 25,625 + 10,000 = 35,625.\end{aligned}$$

Nevertheless, the diversified portfolio C is still optimal.

Why did the risk of portfolio C increase due to positive correlation of the two stocks? When X and Y are positively correlated, low values of X are likely to accompany low values of Y, therefore, the probability of the overall low return is higher, increasing the risk of the portfolio.

Conversely, negative correlation means that low values of X are likely to be compensated by high values of Y, and vice versa. Thus, the risk is reduced. Say, with the given $\rho = -0.2$, we compute $\operatorname{Cov}(X,Y) = \rho \operatorname{Std}(X) \operatorname{Std}(Y) = (-0.2)(0.5)(1) = -0.1$, and

$$\operatorname{Var}(C) = 25,625 + 2(250)(100) \operatorname{Cov}(X,Y) = 25,625 - 5,000 = 20,625.$$

Diversified portfolios consisting of negatively correlated components are the least risky. ◇

Example 3.15 (Optimizing even further). So, after all, with \$10,000 to invest, what is the most optimal portfolio consisting of shares of XX and YY, given their correlation coefficient of $\rho = -0.2$?

This is an *optimization* problem. Suppose t dollars are invested into XX and $(10,000 - t)$ dollars into YY, with the resulting profit is C_t. This amounts for $t/20$ shares of X and $(10,000 - t)/50 = 200 - t/50$ shares of YY. Plans A and B correspond to $t = 10,000$ and $t = 0$.

The expected return remains constant at \$500. If $\operatorname{Cov}(X,Y) = -0.1$, as in Example 3.14 above, then

$$\begin{aligned}\operatorname{Var}(C_t) =\ & \operatorname{Var}\{tX + (200 - t/50)Y\}\\ =\ & (t/20)^2 \operatorname{Var}(X) + (200 - t/50)^2 \operatorname{Var}(Y) + 2(t/20)(200 - t/50) \operatorname{Cov}(X,Y)\\ =\ & (t/20)^2(0.5)^2 + (200 - t/50)^2(1)^2 + 2(t/20)(200 - t/50)(-0.1)\\ =\ & \frac{49t^2}{40,000} - 10t + 40,000.\end{aligned}$$

See the graph of this function in Figure 3.6. Minimum of this variance is found at $t^* = 10/(\frac{49t^2}{40,000}) = 4081.63$. Thus, for the most optimal portfolio, we

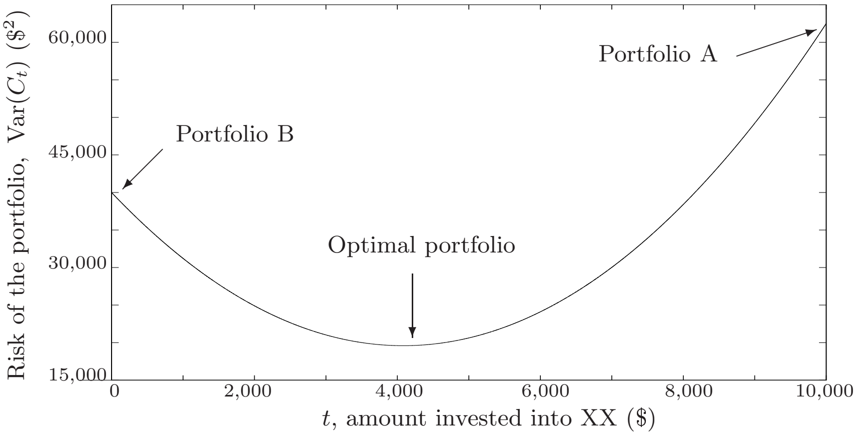


Figure 3.6 *Variance of a diversified portfolio.*

should invest $4081.63 into XX and the remaining $5919.37 into YY. Then we achieve the smallest possible risk (variance) of ($\$^2$)19, 592, measured, as we know, in squared dollars. ◇

3.4 Families of discrete distributions

Next, we introduce the most commonly used families of discrete distributions. Amazingly, absolutely different phenomena can be adequately described by the same mathematical model, or a family of distributions. Say, as we shall see below, the number of virus attacks, received e-mails, error messages, network blackouts, telephone calls, traffic accidents, earthquakes, and so on can all be modeled by the Poisson family of distributions.

3.4.1 Bernoulli distribution

The simplest random variable (excluding non-random ones!) takes just two possible values. Call them 0 and 1.

DEFINITION 3.10

> A random variable with two possible values, 0 and 1, is called a **Bernoulli variable**, its distribution is **Bernoulli distribution**, and any experiment with a *binary outcome* is called a **Bernoulli trial**.

This distribution is named after a Swiss mathematician *Jacob Bernoulli* (1654-1705) who discovered not only Bernoulli but also Binomial distribution.

Good or defective components, parts that pass or fail tests, transmitted or lost signals, working or malfunctioning hardware, sites that contain or do not contain a keyword, girls and boys, heads and tails, and so on, are examples of Bernoulli trials. All these experiments fit the same Bernoulli model, where we shall use generic names for the two outcomes: *"successes"* and *"failures."* These are nothing but commonly used generic names; in fact, successes do not have to be good, and failures do not have to be bad.

If $P(1) = p$ is the probability of a *success*, then $P(0) = q = 1 - p$ is the probability of a *failure*. We can then compute the expectation and variance as

$$\begin{aligned}
\mathbf{E}(X) &= \sum_x P(x) = (0)(1-p) + (1)(p) = p, \\
\mathrm{Var}(X) &= \sum_x (x-p)^2 P(x) = (0-p)^2(1-p) + (1-p)^2 p \\
&= p(1-p)(p+1-p) = p(1-p).
\end{aligned}$$

Bernoulli distribution

$$\boxed{\begin{array}{lcl}
p & = & \text{probability of success} \\
P(x) & = & \left\{\begin{array}{ccc} q = 1-p & \text{if} & x = 0 \\ p & \text{if} & x = 1 \end{array}\right. \\
\mathbf{E}(X) & = & p \\
\mathrm{Var}(X) & = & pq
\end{array}}$$

In fact, we see that there is a whole *family of Bernoulli distributions*, indexed by a *parameter* p. Every p between 0 and 1 defines another Bernoulli distribution. The distribution with $p = 0.5$ carries the highest level of uncertainty because $\mathrm{Var}(X) = pq$ is maximized by $p = q = 0.5$. Distributions with lower or higher p have lower variances. Extreme parameters $p = 0$ and $p = 1$ define non-random variables 0 and 1, respectively, their variance is 0.

3.4.2 Binomial distribution

Now consider a sequence of independent Bernoulli trials and count the number of successes in it. This may be the number of defective computers in a shipment, the number of updated files in a folder, the number of girls in a family, the number of e-mails with attachments, etc.

DEFINITION 3.11

A variable described as the number of successes in a sequence of independent Bernoulli trials has **Binomial distribution**. Its parameters are n, the number of trials, and p, the probability of success.

Remark: "Binomial" can be translated as "two numbers," *bi* meaning "two" and *nom* meaning "a number," thus reflecting the concept of binary outcomes.

Binomial probability mass function is

$$P(x) = \boldsymbol{P}\left\{X = x\right\} = \begin{pmatrix} n \\ x \end{pmatrix} p^x q^{n-x}, \qquad x = 0, 1, \ldots, n, \tag{3.9}$$

which is the probability of exactly x successes in n trials. In this formula, p^x is the probability of x successes, probabilities being multiplied due to independence of trials. Also, q^{n-x} is the probability of the remaining $(n-x)$ trials being failures. Finally, $\begin{pmatrix} n \\ x \end{pmatrix} = \frac{n!}{x!(n-x)!}$ is the number of elements of the sample space Ω that form the event $\{X = x\}$. This is the number of possible orderings of x successes and $(n-x)$ failures among n trials, and it is computed as $C(n, x)$ in (2.6).

Due to a somewhat complicated form of (3.9), practitioners use *a table of Binomial distribution*, Table A2. Its entries are values of the Binomial cdf $F(x)$. The pmf can then be obtained as

$$P(x) = F(x) - F(x-1).$$

Example 3.16. As part of a business strategy, randomly selected 20% of new internet service subscribers receive a special promotion from the provider. A group of 10 neighbors signs for the service. What is the probability that at least 4 of them get a special promotion?

<u>Solution</u>. We need to find the probability $\boldsymbol{P}\left\{X \geq 4\right\}$, where X is the number of people, out of 10, who receive a special promotion. This is the number of successes in 10 Bernoulli trials, therefore, X has Binomial distribution with parameters $n = 10$ and $p = 0.2$. From Table A2,

$$\boldsymbol{P}\left\{X \geq 4\right\} = 1 - F(3) = 1 - 0.8791 = \underline{0.1209}.$$

$\diamond$

The table does not go beyond $n = 20$. For large n, we shall learn to use reasonable approximations.

Computing the expectation directly by (3.3) results in a complicated formula,

$$\mathbf{E}(X) = \sum_{x=0}^{n} x \left(\begin{array}{c} n \\ x \end{array} \right) p^x q^{n-x} = \ldots \, ?$$

A shortcut can be obtained from the following important property.

Each Bernoulli trial is associated with a Bernoulli variable that equals 1 is the trial resulted in a success. Then, a sum of these variables is the overall number of successes. Thus, *any Binomial variable X can be represented as a sum of independent Bernoulli variables*,

$$X = X_1 + \ldots + X_n.$$

We can then compute (referring to (3.5) and (3.7))

$$\mathbf{E}(X) = \mathbf{E}(X_1 + \ldots + X_n) = \mathbf{E}(X_1) + \ldots + \mathbf{E}(X_n) = p + \ldots + p = np$$

and

$$\mathrm{Var}(X) = \mathrm{Var}(X_1 + \ldots + X_n) = \mathrm{Var}(X_1) + \ldots + \mathrm{Var}(X_n) = npq.$$

Binomial distribution

$$\boxed{\begin{array}{lcl} n & = & \text{number of trials} \\ p & = & \text{probability of success} \\ P(x) & = & \left(\begin{array}{c} n \\ x \end{array} \right) p^x q^{n-x} \\ \mathbf{E}(X) & = & np \\ \mathrm{Var}(X) & = & npq \end{array}}$$

Example 3.17. An exciting computer game is released. Sixty percent of players complete all the levels. Thirty percent of them will then buy an advanced version of the game. Among 15 users, what is the expected number of people who will buy the advanced version? What is the probability that at least two people will buy it?

Solution. Let X be the number of people (successes), among the mentioned 15 users (trials), who will buy the advanced version of the game. It has Binomial distribution with $n = 15$ trials and the probability of success

$$\begin{array}{rcl} p & = & \boldsymbol{P}\{\text{buy advanced}\} \\ & = & \boldsymbol{P}\{\text{buy advanced} \mid \text{complete all levels}\}\, \boldsymbol{P}\{\text{complete all levels}\} \\ & = & (0.30)(0.60) = 0.18. \end{array}$$

Then we have

$$\mathbf{E}(X) = np = (15)(0.18) = \underline{2.7}$$

and

$$\boldsymbol{P}\{X \geq 2\} = 1 - P(0) - P(1) = 1 - (1-p)^n - np(1-p)^{n-1} = \underline{0.7813}.$$

The last probability was computed directly by formula (3.9) because the probability of success, 0.18, is not in Table A2. ◇

3.4.3 Geometric distribution

Again, consider a sequence of independent Bernoulli trials. Each trial results in a "success" or a "failure."

DEFINITION 3.12

The number of Bernoulli trials needed to get the first success has **Geometric distribution**.

Example 3.18. A search engine goes through a list of sites looking for a given key phrase. Suppose the search terminates as soon as the key phrase is found. The number of sites visited is Geometric. ◇

Example 3.19. A hiring manager interviews candidates, one by one, to fill a vacancy. The number of candidates interviewed until one candidate receives an offer has Geometric distribution. ◇

Geometric random variables can take any integer value from 1 to infinity, because one needs at least 1 trial to have the first success, and the number of trials needed is not limited by any specific number. (For example, there is no guarantee that among the first 10 coin tosses there will be at least one head.) The only parameter is p, the probability of a "success."

Geometric probability mass function has the form

$$P(x) = \boldsymbol{P}\{\text{ the 1st success is the } x\text{-th trial }\} = (1-p)^{x-1}p, \quad x = 1, 2, \ldots,$$

which is the probability of $(x-1)$ failures on the first $(x-1)$ trials and a success on the last trial. Comparing with (3.9), there is no number of combinations in this formula because only one outcome has the first success coming on the x-th trial.

This is the first time we see an *unbounded* random variable, that is, with no upper bound. It is insightful to check whether $\sum_x P(x) = 1$, as it should hold for all pmf. Indeed,

$$\sum_x P(x) = \sum_{x=1}^{\infty}(1-p)^{x-1}p = \frac{(1-p)^0}{1-(1-p)}p = 1,$$

where we noticed that the sum on the left is a *geometric series* that gave its name to Geometric distribution.

Finally, Geometric distribution has expectation $\mu = 1/p$ and variance $\sigma^2 = (1-p)/p^2$.

Proof: The geometric series $s(q) = \sum_0^\infty q^x$ equals $(1-q)^{-1}$. Taking derivatives with respect to q,

$$\frac{1}{(1-q)^2} = \left(\frac{1}{1-q}\right)' = s'(q) = \left(\sum_0^\infty q^x\right)' = \sum_0^\infty xq^{x-1} = \sum_1^\infty xq^{x-1}.$$

It follows that for a Geometric variable X with parameter $p = 1-q$,

$$\mathbf{E}(X) = \sum_{x=1}^{\infty} xq^{x-1}p = \frac{1}{(1-q)^2}p = \frac{1}{p}.$$

Derivation of the variance is similar. One has to take the second derivative of $s(q)$ and get, after a few manipulations, an expression for $\sum x^2q^{x-1}$. □

Geometric distribution

$$\boxed{\begin{array}{lcl} p & = & \text{probability of success} \\ P(x) & = & (1-p)^{x-1}p, \quad x = 1, 2, \ldots \\ \mathbf{E}(X) & = & \dfrac{1}{p} \\ \text{Var}(X) & = & \dfrac{1-p}{p^2} \end{array}} \tag{3.10}$$

Example 3.20 (St. Petersburg Paradox). This paradox was noticed by a Swiss mathematician *Daniel Bernoulli* (1700–1782), a nephew of Jacob. It describes *a gambling strategy that enables one to win any desired amount of money with probability one.*

Consider a game that can be played any number of times. Rounds are independent, and each time your winning probability is p. The game does not have to be favorable to you or even fair. This p can be any positive probability. For each round, you bet some amount x. In case of a success, you win x. If you lose the round, you lose x.

The strategy is simple. Your initial bet is the amount that you desire to win eventually. Then, if you win a round, stop. If you lose a round, double your bet and continue.

Say, the desired profit is \$100. The game will progress as follows.

Round	Bet	Balance...	
		... if lose	... if win
1	100	-100	$+100$ and stop
2	200	-300	$+100$ and stop
3	400	-700	$+100$ and stop
...	...	...	...

Sooner or later, the game will stop, and at this moment, your balance will be \$100. Guaranteed! But this is not what D. Bernoulli called a paradox.

How many rounds should be played? Since each round is a Bernoulli trial, the number of them, X, until the first win is a Geometric random variable with parameter p.

Is the game endless? No, on the average, it will last $\mathbf{E}(X) = 1/p$ rounds. In a fair game with $p = 1/2$, one will need 2 rounds, on the average, to win the desired amount. In an "unfair" game, with $p < 1/2$, it will take longer to win, but still a finite number of rounds. For example, if $p = 0.2$, i.e., one win in five rounds, then on the average, one stops after $1/p = 5$ rounds. This is not a paradox yet.

Finally, how much money does one need to have in order to be able to follow this strategy? Let Y be the amount of the last bet. According to the strategy, $Y = 100 \cdot 2^{X-1}$. It is a discrete random variable whose expectation equals

$$\begin{aligned}
\mathbf{E}(Y) &= \sum_x \left(100 \cdot 2^{x-1}\right) P_X(x) = 100 \sum_{x=1}^{\infty} 2^{x-1}(1-p)^{x-1}p \\
&= 100p \sum_{x=1}^{\infty} (2(1-p))^{x-1} = \begin{cases} \dfrac{100p}{2(1-p)} & \text{if} \quad p > 1/2 \\ +\infty & \text{if} \quad p \le 1/2. \end{cases}
\end{aligned}$$

This is the St. Petersburg Paradox! A random variable that is always finite has an infinite expectation! Even when the game is fair offering a 50-50 chance to win, one has to be (on the average!) infinitely rich to follow this strategy.

To the best of our knowledge, every casino has a limit on the maximum bet, making sure gamblers cannot fully apply the described strategy. When such a limit is enforced, it can be proved that a winning strategy does not exist. ◇

3.4.4 Negative Binomial distribution

When we studied Geometric distribution and St. Petersburg paradox in Section 3.4.3, we played a game until the first win. Now keep playing until we reach a certain number of wins. The number of played games is then *Negative Binomial*.

DEFINITION 3.13

In a sequence of independent Bernoulli trials, the number of trials needed to obtain k successes has **Negative Binomial distribution**.

In some sense, Negative Binomial distribution is opposite to Binomial distribution. Binomial variables count the number of successes in a fixed number of trials whereas Negative Binomial variables count the number of trials needed to see a fixed number of successes. Other than this, there is nothing "negative" about this distribution.

Negative Binomial probability mass function is

$$\begin{aligned} P(x) = \boldsymbol{P}\left\{ \text{ the } x\text{-th trial is the } k\text{-th success } \right\} \\ = \ \boldsymbol{P}\left\{ \begin{array}{c} (k-1) \text{ successes in the first } (x-1) \text{ trials,} \\ \text{and the last trial is a success} \end{array} \right\} \\ = \ \left(\begin{array}{c} x-1 \\ k-1 \end{array} \right) (1-p)^{x-k} p^k. \end{aligned}$$

This formula accounts for the probability of k successes, the remaining $(x-k)$ failures, and the number of outcomes–sequences with the k-th success coming on the x-th trial.

Negative Binomial distribution has two parameters, k and p. With $k = 1$, it becomes Geometric. Also, each Negative Binomial variable can be represented as a sum of independent Geometric variables,

$$X = X_1 + \ldots + X_k, \tag{3.11}$$

with the same probability of success p. Indeed, the number of trials until the k-th success consists of a Geometric number of trials X_1 until the first success, an additional Geometric number of trials X_2 until the second success, etc.

Because of (3.11), we have

$$\begin{aligned} \mathbf{E}(X) &= \mathbf{E}(X_1 + \ldots + X_k) = \frac{k}{p}; \\ \mathrm{Var}(X) &= \mathrm{Var}(X_1 + \ldots + X_k) = \frac{k(1-p)}{p^2}. \end{aligned}$$

Negative Binomial distribution

$$\boxed{\begin{array}{lcl} k & = & \text{number of successes} \\ p & = & \text{probability of success} \\ P(x) & = & \begin{pmatrix} x-1 \\ k-1 \end{pmatrix} (1-p)^{x-k} p^k, \quad x = k, k+1, \ldots \\ \mathbf{E}(X) & = & \dfrac{k}{p} \\ \mathrm{Var}(X) & = & \dfrac{k(1-p)}{p^2} \end{array}} \tag{3.12}$$

Example 3.21 (Sequential testing). In a recent production, 5% of certain electronic components are defective. We need to find 12 non-defective components for our 12 new computers. Components are tested until 12 non-defective ones are found. What is the probability that more than 15 components will have to be tested?

Solution. Let X be the number of components tested until 12 non-defective ones are found. It is a number of trials needed to see 12 successes, hence X has Negative Binomial distribution with $k = 12$ and $p = 0.05$.

We need $\boldsymbol{P}\{X > 15\} = \sum_{16}^{\infty} P(x)$ or $1 - F(15)$, however, there is no table of Negative Binomial distribution in the Appendix, and applying the formula for $P(x)$ directly is rather cumbersome. What would be a quick solution?

Virtually any Negative Binomial problem can be solved by a Binomial distribution. Certainly, X is not Binomial, however, the probability $\boldsymbol{P}\{X > 15\}$ can be related to some Binomial variable. In our example,

$$\begin{aligned} \boldsymbol{P}\{X > 15\} &= \boldsymbol{P}\{\text{ more than 15 trials needed to get 12 successes }\} \\ &= \boldsymbol{P}\{\text{ 15 trials are not sufficient }\} \\ &= \boldsymbol{P}\{\text{ there are fewer than 12 successes in 15 trials }\} \\ &= \boldsymbol{P}\{Y < 12\}, \end{aligned}$$

where Y is the number of successes (non-defective components) in 15 trials, which is a Binomial variable with parameters $n = 15$ and $p = 0.95$. From Table A2 on p. 378,

$$\boldsymbol{P}\{X > 15\} = \boldsymbol{P}\{Y < 12\} = \boldsymbol{P}\{Y \leq 11\} = F(11) = \underline{0.0055}.$$

This technique, expressing a probability about one random variable in terms of another random variable, is rather useful. It will also help us relate Gamma and Poisson distributions and simplify computations significantly. ◇

3.4.5 Poisson distribution

The next distribution is related to a concept of *rare events*, or Poissonian events. Essentially it means that two such events are extremely unlikely to occur within a very short period of time or simultaneously. Arrivals of jobs, telephone calls, e-mail messages, traffic accidents, network blackouts, virus attacks, errors in software, floods, earthquakes are examples of rare events. The rigorous definition of rare events is given in Section 6.3.2.

DEFINITION 3.14

> The number of rare events occurring within a fixed period of time has **Poisson distribution**.

This distribution bears the name of a famous French mathematician *Siméon-Denis Poisson* (1781–1840).

Poisson distribution

$$
\begin{array}{lcl}
\lambda & = & \text{frequency, average number of events} \\
P(x) & = & e^{-\lambda}\dfrac{\lambda^x}{x!}, \; x = 0, 1, 2, \ldots \\
\mathbf{E}(X) & = & \lambda \\
\mathrm{Var}(X) & = & \lambda
\end{array}
$$

A Poisson variable can take any nonnegative integer value because there may be no rare events within the chosen period, on one end, and the possible number of events is not limited, on the other end. Poisson distribution has one parameter, $\lambda > 0$, which is the average number of the considered rare events. Values of its cdf are given in Table A3 on p. 384.

Example 3.22 (New accounts). Customers of an internet service provider initiate new accounts at the average rate of 10 accounts per day.

(a) What is the probability that more than 8 new accounts will be initiated today? (b) What is the probability that more than 16 accounts will be initiated within 2 days?

Solution. (a) New account initiations qualify as rare events because no two customers open accounts simultaneously. Then the number X of today's new accounts has Poisson distribution with parameter $\lambda = 10$. From Table A3,

$$\boldsymbol{P}\{X > 8\} = 1 - F_X(8) = 1 - 0.333 = \underline{0.667}.$$

(b) The number of accounts, Y, opened within 2 days does *not* equal $2X$. Rather, Y is another Poisson random variable whose parameter equals 20. Indeed, the parameter is the average number of rare events, which, over the period of two days, doubles the one-day average. Using Table A3 with $\lambda = 20$,

$$\boldsymbol{P}\{Y > 16\} = 1 - F_Y(16) = 1 - 0.221 = \underline{0.779}.$$

◇

3.4.6 Poisson approximation of Binomial distribution

Poisson distribution can be effectively used to approximate Binomial probabilities when the number of trials n is large, and the probability of success p is small. Such an approximation is adequate, say, for $n \geq 30$ and $p < 0.05$, and it becomes more accurate for larger n.

Example 3.23 (New accounts, continued). Indeed, the situation in Example 3.22 can be viewed as a sequence of Bernoulli trials. Suppose there are $n = 400,000$ potential internet users in the area, and on any specific day, each of them opens a new account with probability $p = 0.000025$. We see that the number of new accounts is the number of successes, hence a Binomial model with expectation $\mathbf{E}(X) = np = 10$ is possible. However, a distribution with such extreme n and p is unlikely to be found in any table, and computing its pmf by hand is tedious. Instead, one can use Poisson distribution with the same expectation $\lambda = 10$. ◇

Poisson approximation to Binomial

$$\boxed{\begin{array}{c}\text{Binomial}(n,p) \approx \text{ Poisson}(\lambda) \\ \text{where } n \geq 30,\ p \leq 0.05,\ np = \lambda\end{array}} \qquad (3.13)$$

Remark: Mathematically, it means closeness of Binomial and Poisson pmf,

$$\lim_{\substack{n \to \infty \\ p \to 0 \\ np \to \lambda}} \binom{n}{x} p^x (1-p)^{n-x} = e^{-\lambda}\frac{\lambda^x}{x!}$$

and this is what S. D. Poisson has shown.

When p is large ($p \geq 0.95$), the Poisson approximation is applicable too. The probability of a failure $q = 1-p$ is small in this case. Then, we can approximate the number of failures, which is also Binomial.

Example 3.24. Ninety-seven percent of electronic messages are transmitted with no error. What is the probability that out of 200 messages, at least 195 will be transmitted correctly?

Solution. Let X be the number of correctly transmitted messages. It is the number of successes in 200 Bernoulli trials, thus X is Binomial with $n = 200$ and $p = 0.97$. Poisson approximation cannot be applied to X because p is too large. However, the number of failures Y is also Binomial, with parameters $n = 200$ and $q = 0.03$, and it is approximately Poisson with $\lambda = nq = 6$. From Table A3,

$$\boldsymbol{P}\{X \geq 195\} = \boldsymbol{P}\{Y \leq 5\} = F_Y(5) \approx \underline{0.446}.$$

◇

There is a great variety of applications involving a large number of trials with a small probability of success. If the trials are not independent, the number of successes is in not Binomial. However, if dependence is weak, the use of Poisson approximation in such problems can still produce amazingly accurate results.

Example 3.25 (Birthday problem). This continues Exercise 2.25 on p. 38. Consider a class with $N \geq 10$ students. Compute the probability that at least two of them have their birthdays on the same day. How many students should be in class in order to have this probability above 0.5?

Solution. Poisson approximation will be used for the number of shared birthdays among all

$$n = \binom{N}{2} = \frac{N(N-1)}{2}$$

pairs of students in this class. In each pair, both students are born on the same day with probability $p = 1/365$. Each pair is a Bernoulli trial because the two birthdays either match or don't match. Besides, matches in two different pairs are "nearly" independent. Therefore, X, the number of pairs sharing birthdays, is "almost" Binomial. For $N \geq 10$, $n \geq 45$ is large, and p is small, thus, we shall use Poisson approximation with $\lambda = np = N(N-1)/730$,

$$\boldsymbol{P}\{\text{there are two students sharing birthday}\} = 1 - \boldsymbol{P}\{\text{no matches}\}$$

$$= 1 - \boldsymbol{P}\{X = 0\} \approx 1 - e^{-\lambda} \approx 1 - e^{-N^2/730}.$$

Solving the inequality $1 - e^{-N^2/730} > 0.5$, we obtain $N > \sqrt{730 \ln 2} = 22.5$. That is, in a class of at least $N = 23$ students, there is a more than 50% chance that at least two students were born on the same day of the year! ◇

The introduced method can only be applied to very small and very large values of p. For moderate p ($0.05 \leq p \leq 0.95$), the Poisson approximation may not be accurate. These cases are covered by the Central Limit Theorem on p. 101.

Summary and conclusions

Discrete random variables can take a finite or countable number of isolated values with different probabilities. Collection of all such probabilities is a distribution, which describes the behavior of a random variable. Random vectors are sets of random variables; their behavior is described by the joint distribution. Marginal probabilities can be computed from the joint distribution by the Addition Rule.

The average value of a random variable is its expectation. Variability around the expectation is measured by the variance and the standard deviation. Covariance and correlation measure association of two random variables. For any distribution, probabilities of large deviations can be bounded by Chebyshev's inequality, using only the expectation and variance.

Different phenomena can often be described by the same probabilistic model, or a family of distributions. The most commonly used discrete families are Binomial, including Bernoulli, Negative Binomial, including Geometric, and Poisson. Each family of distributions is used for a certain general type of situations, it has its parameters and a clear formula and/or a table for computing probabilities. These families are summarized in Section 11.1.1.

Questions and exercises

3.1. A computer virus is trying to corrupt two files. The first file will be corrupted with probability 0.4. Independently of it, the second file will be corrupted with probability 0.3.

(a) Compute the probability mass function (pmf) of X, the number of corrupted files.

(b) Draw a graph of its cumulative distribution function (cdf).

3.2. Every day, the number of network blackouts has a distribution (probability mass function)

x	0	1	2
$P(x)$	0.7	0.2	0.1

A small internet trading company estimates that each network blackout results in a \$500 loss. Compute expectation and variance of this company's daily loss due to blackouts.

3.3. There is one error in one of five blocks of a program. To find the error, we test three randomly selected blocks. Let X be the number of errors in these three blocks. Compute $\mathbf{E}(X)$ and $\mathrm{Var}(X)$.

3.4. Tossing a fair die is an experiment that can result in any integer number from 1 to 6 with equal probabilities. Let X be the number of dots on the top face of a die. Compute $\mathbf{E}(X)$ and $\mathrm{Var}(X)$.

3.5. A software package consists of 12 programs, five of which must be upgraded. If 4 programs are randomly chosen for testing,

(a) What is the probability that at least two of them must be upgraded?

(b) What is the expected number of programs, out of the chosen four, that must be upgraded?

3.6. A computer program contains one error. In order to find the error, we split the program into 6 blocks and test two of them, selected at random. Let X be the number of errors in these blocks. Compute $\mathbf{E}(X)$.

3.7. The number of home runs scored by a certain team in one game is a random variable with the distribution

x	0	1	2
$P(x)$	0.4	0.4	0.2

The team plays 2 games. The number of home runs scored in one game is independent of the number of home runs in the other game. Let Y be the *total* number of home runs. Find $\mathbf{E}(Y)$ and $\mathrm{Var}(Y)$.

3.8. A computer user tries to recall her password. She knows it can be one of 4 possible passwords. She tries her passwords until she finds the right one. Let X be the number of wrong passwords she uses before she finds the right one. Find $\mathbf{E}(X)$ and $\mathrm{Var}(X)$.

3.9. It takes an average of 40 seconds to download a certain file, with a standard deviation of 5 seconds. The actual distribution of the download time is unknown. Using Chebyshev's inequality, what can be said about the probability of spending more than 1 minute for this download?

3.10. Every day, the number of traffic accidents has a distribution (probability mass function)

x	0	1	2	more than 2
$P(x)$	0.6	0.2	0.2	0

independently of other days. What is the probability that there are more accidents on Friday than on Thursday?

3.11. Two dice are tossed. Let X be *the smaller* number of points. Let Y be *the* *larger* number of points. If both dice show the same number, say, z points, then $X = Y = z$.

(a) Find the joint probability mass function of (X, Y).

(b) Are X and Y independent? Explain.

(c) Find the distribution (probability mass function) of X.

(d) If $X = 2$, what is the probability that $Y = 5$?

3.12. Two random variables, X and Y, have the joint distribution $P(x, y)$,

		x	
$P(x,y)$		0	1
y	0	0.5	0.2
	1	0.2	0.1

(a) Are X and Y independent? Explain.

(b) Are $(X + Y)$ and $(X - Y)$ independent? Explain.

3.13. Two random variables X and Y have the joint distribution, $P(0, 0) = 0.2$, $P(0, 2) = 0.3$, $P(1, 1) = 0.1$, $P(2, 0) = 0.3$, $P(2, 2) = 0.1$, and $P(x, y) = 0$ for all other pairs (x, y).

(a) Find the distribution of $Z = X + Y$.

(b) Find the distribution of $U = X - Y$.

(c) Find the distribution of $V = XY$.

3.14. An internet service provider charges its customers for the time of the internet use rounding it up to the nearest hour. The joint distribution of the used time (X, hours) and the charge per hour (Y, cents) is given in the table below.

		x			
$P(x,y)$		1	2	3	4
y	1	0	0.06	0.06	0.10
	2	0.10	0.10	0.04	0.04
	3	0.40	0.10	0	0

Each customer is charged $Z = X \cdot Y$ cents, which is the number of hours multiplied by the price of each hour. Find the distribution of Z.

3.15. Let X and Y be the number of hardware failures in two computer labs in a given month. The joint distribution of X and Y is given in the table,

		x		
$P(x,y)$		0	1	2
y	0	0.52	0.20	0.04
	1	0.14	0.02	0.01
	2	0.06	0.01	0

(a) Compute the probability of at least one hardware failure.

(b) From the given distribution, are X and Y independent? Why or why not?

3.16. The number of hardware failures, X, and the number of software failures, Y, on any day in a small computer lab have the joint distribution $P(x,y)$, where $P(0,0) = 0.6$, $P(0,1) = 0.1$, $P(1,0) = 0.1$, $P(1,1) = 0.2$. Based on this information,

(a) Are X and Y (hardware and software failures) independent?

(b) Compute $\mathbf{E}(X+Y)$, i.e., the expected total number of failures during 1 day.

3.17. Shares of company A are sold at \$10 per share. Shares of company B are sold at \$50 per share. According to a market analyst, 1 share of each company can either gain \$1, with probability 0.5, or lose \$1, with probability 0.5, independently of the other company. Which of the following portfolios has the lowest risk:

(a) 100 shares of A

(b) 50 shares of A + 10 shares of B

(c) 40 shares of A + 12 shares of B

3.18. Shares of company A cost \$10 per share and give a profit of X%. Independently of A, shares of company B cost \$50 per share and give a profit of Y%. Deciding how to invest \$1,000, Mr. X chooses between 3 portfolios:

(a) 100 shares of A,

(b) 50 shares of A and 10 shares of B,

(c) 20 shares of B.

The distribution of X is given by probabilities:

$$P\{X=-3\} = 0.3, P\{X=0\} = 0.2, P\{X=3\} = 0.5.$$

The distribution of Y is given by probabilities:

$$P\{Y=-3\} = 0.4, P\{Y=3\} = 0.6.$$

Compute expectations and variances of the total dollar profit generated by portfolios (a), (b), and (c). What is the least risky portfolio? What is the most risky portfolio?

3.19. A and B are two competing companies. An investor decides whether to buy

(a) 100 shares of A, or

(b) 100 shares of B, or

(c) 50 shares of A and 50 shares of B.

A profit made on 1 share of A is a random variable X with the distribution $P(X = 2) = P(X = -2) = 0.5$.

A profit made on 1 share of B is a random variable Y with the distribution $P(Y = 4) = 0.2, P(Y = -1) = 0.8$.

If X and Y are independent, compute the expected value and variance of the total profit for strategies (a), (b), and (c).

3.20. A quality control engineer tests the quality of produced computers. Suppose that 5% of computers have defects, and defects occur independently of each other.

(a) Find the probability of exactly 3 defective computers in a shipment of twenty.

(b) Find the probability that the engineer has to test at least 5 computers in order to find 2 defective ones.

3.21. A lab network consisting of 20 computers was attacked by a computer virus. This virus enters each computer with probability 0.4, independently of other computers. Find the probability that it entered at least 10 computers.

3.22. Five percent of computer parts produced by a certain supplier are defective. What is the probability that a sample of 16 parts contains more than 3 defective ones?

3.23. Every day, a lecture may be canceled due to inclement weather with probability 0.05. Class cancelations on different days are independent.

(a) There are 15 classes left this semester. Compute the probability that at least 4 of them get canceled.

(b) Compute the probability that the tenth class this semester is the third class that gets canceled.

3.24. An internet search engine looks for a certain keyword in a sequence of independent web sites. It is believed that 20% of the sites contain this keyword.

(a) Compute the probability that at least 5 of the first 10 sites contain the given keyword.

(b) Compute the probability that the search engine had to visit at least 5 sites in order to find the first occurrence of a keyword.

3.25. About ten percent of users do not close Windows properly. Suppose that Windows is installed in a public library that is used by random people in a random order.

(a) On the average, how many users of this computer *do not* close Windows properly before someone *does* close it properly?

(b) What is the probability that exactly 8 of the next 10 users will close Windows properly?

3.26. After a computer virus entered the system, a computer manager checks the condition of all important files. She knows that each file has probability 0.2 to be damaged by the virus, independently of other files.

(a) Compute the probability that at least 5 of the first 20 files are damaged.

(b) Compute the probability that the manager has to check at least 6 files in order to find 3 undamaged files.

3.27. Messages arrive at an electronic message center at random times, with an average of 9 messages per hour.

(a) What is the probability of receiving *at least* five messages during the next hour?

(b) What is the probability of receiving *exactly* five messages during the next hour?

3.28. The number of received electronic messages has Poisson distribution with some parameter λ. Using Chebyshev inequality, show that the probability of receiving more than 4λ messages does not exceed $1/(9\lambda)$.

3.29. An insurance company divides its customers into 2 groups. Twenty percent of customers are in the high-risk group, and eighty percent are in the low-risk group. The high-risk customers make an average of 1 accident per year while the low-risk customers make an average of 0.1 accidents per year. Mr. X had no accidents last year. What is the probability that he is a high-risk driver?

3.30. Before the computer is assembled, its vital component (motherboard) goes through a special inspection. Only 80% of components pass this inspection.

(a) What is the probability that at least 18 of the next 20 components pass inspection?

(b) On the average, how many components should be inspected until a component that passes inspection is found?

3.31. On the average, 1 computer in 800 crashes during a severe thunderstorm. A certain company had 4,000 working computers when the area was hit by a severe thunderstorm.

(a) Compute the probability that less than 10 computers crashed.

(b) Compute the probability that exactly 10 computers crashed.

You may want to use a suitable approximation.

3.32. The number of computer shutdowns during any month has a Poisson distribution, averaging 0.25 shutdowns per month.

(a) What is the probability of at least 3 computer shutdowns during the next year?

(b) During the next year, what is the probability of at least 3 months (out of 12) with exactly 1 computer shutdown in each?

3.33. A dangerous computer virus attacks a folder consisting of 250 files. Files are affected by the virus independently of one another. Each file is affected with the probability 0.032. What is the probability that more than 7 files are affected by this virus?

3.34. In some city, the probability of a thunderstorm on any day is 0.6. During a thunderstorm, the number of traffic accidents has Poisson distribution with parameter 10. Otherwise, the number of traffic accidents has Poisson distribution with parameter 4. If there were 7 accidents yesterday, what is the probability that there was a thunderstorm?

3.35. An interactive system consists of ten terminals that are connected to the central computer. At any time, each terminal is ready to transmit a message with probability 0.7, independently of other terminals. Find the probability that exactly 6 terminals are ready to transmit at 8 o'clock.

3.36. Network breakdowns are unexpected rare events that occur every 3 weeks, on the average. Compute the probability of more than 4 breakdowns during a 21-week period.

3.37. Simplifying expressions, derive from the definitions of variance and covariance that

(a) $\mathrm{Var}(X) = \mathbf{E}(X^2) - \mathbf{E}^2(X)$;

(b) $\mathrm{Cov}(X, Y) = \mathbf{E}(XY) - \mathbf{E}(X)\,\mathbf{E}(Y)$.

3.38. Show that

$$\begin{aligned} &\mathrm{Cov}(aX + bY + c, dZ + eW + f) \\ &\quad = \quad ad\,\mathrm{Cov}(X, Z) + ae\,\mathrm{Cov}(X, W) + bd\,\mathrm{Cov}(Y, Z) + be\,\mathrm{Cov}(Y, W) \end{aligned}$$

for any random variables X, Y, Z, W, and any non-random numbers a, b, c, d, e, f.

CHAPTER 4

Continuous Distributions

Recall that any discrete distribution is concentrated on a finite or countable number of isolated values. Conversely, *continuous variables can take any value of an interval*, (a, b), $(a, +\infty)$, $(-\infty, +\infty)$, etc. Various times like service time, installation time, download time, failure time, and also physical measurements like weight, height, distance, velocity, temperature, and connection speed are examples of continuous random variables.

4.1 Probability density

For all continuous variables, the probability mass function (pmf) is always equal zero,*

$$P(x) = 0 \quad \text{ for all } x.$$

As a result, the pmf does not carry any information about a random variable. Rather, we can use the *cumulative distribution function* (cdf) $F(x)$. In the continuous case, it equals

$$F(x) = \boldsymbol{P}\left\{X \leq x\right\} = \boldsymbol{P}\left\{X < x\right\}.$$

These two expressions for $F(x)$ differ by $P\left\{X = x\right\} = P(x) = 0$.

In both continuous and discrete cases, the cdf $F(x)$ is a non-decreasing function that ranges from 0 to 1. Recall from Chapter 3 that in the discrete case, the graph of $F(x)$ has *jumps* of magnitude $P(x)$. For continuous distributions, $P(x) = 0$, which means no jumps. The cdf in this case is a continuous function.

* Remark: In fact, any probability mass function $P(x)$ can give positive probabilities to a finite or countable set only. Indeed, since $\sum_x P(x) = 1$, there can be at most 2 values of x with $P(x) \geq 1/2$, at most 4 values with $P(x) \geq 1/4$, etc. Continuing this way, we can list all x where $P(x) > 0$. Hence, the set of all such x is at most countable. It cannot be an interval because any interval is uncountable. This agrees with $P(x) = 0$ for continuous random variables that take entire intervals of values.

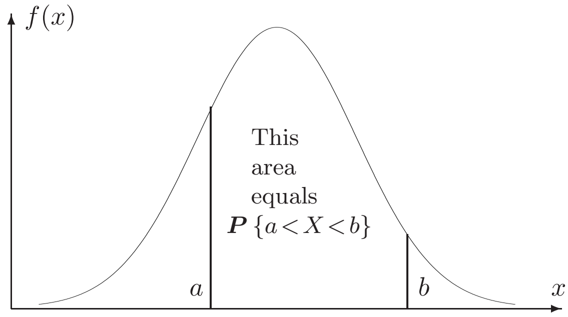


Figure 4.1 *Probabilities are areas under the density curve.*

DEFINITION 4.1

The **probability density function** (pdf) is the derivative of the cdf, $f(x) = F'(x)$.

Then, $F(x)$ is *antiderivative* of a density. By the Fundamental Theorem of Calculus, integral of a density from a to b equals the difference of antiderivatives, i.e.,

$$\int_a^b f(x)dx = F(b) - F(a) = \boldsymbol{P}\{a < X < b\},$$

where we notice again that the probability in the right-hand side also equals $\boldsymbol{P}\{a \le X < b\}$, $\boldsymbol{P}\{a < X \le b\}$, and $\boldsymbol{P}\{a \le X \le b\}$.

Probability density function

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

$$\boldsymbol{P}\{a < X < b\} = \int_a^b f(x)dx$$

Thus, probabilities can be calculated by integrating a density over the given sets. Furthermore, the integral $\int_a^b f(x)dx$ equals the area below the density curve between the points a and b. Therefore, geometrically, probabilities are represented by *areas* (Figure 4.1). Substituting $a = -\infty$ and $b = +\infty$, we obtain

$$\int_{-\infty}^b f(x)dx = \boldsymbol{P}\{-\infty < X < b\} = F(b) \tag{4.1}$$

and

$$\int_{-\infty}^{+\infty} f(x)dx = \boldsymbol{P}\{-\infty < X < +\infty\} = 1.$$

That is, the total area below the density curve equals 1.

Looking at Figure 4.1, we can see why $P(x) = 0$ for all continuous random variables. That is because

$$P(x) = \boldsymbol{P}\{x \le X \le x\} = \int_x^x f = 0.$$

Geometrically, it is the area below the density curve, where two sides of the region collapse into one.

Example 4.1. The lifetime, in years, of some electronic component is a continuous random variable with the density

$$f(x) = \begin{cases} \dfrac{k}{x^3} & \text{for} \quad x \ge 1 \\ 0 & \text{for} \quad x < 1. \end{cases}$$

Find k, draw a graph of the cdf $F(x)$, and compute the probability for the lifetime to exceed 5 years.

Solution. Find k from the condition $\int f(x)dx = 1$:

$$\int_{-\infty}^{+\infty} f(x)dx = \int_1^{+\infty} \frac{k}{x^3}dx = -\left.\frac{k}{2x^2}\right|_{x=1}^{+\infty} = \frac{k}{2} = 1.$$

Hence, $k = 2$. Integrating the density, we get the cdf,

$$F(x) = \int_{-\infty}^{x} f(y)dy = \int_1^x \frac{2}{y^3}dy = -\left.\frac{1}{y^2}\right|_{y=1}^{x} = 1 - \frac{1}{x^2}$$

for $x > 1$. Its graph is shown in Figure 4.2.

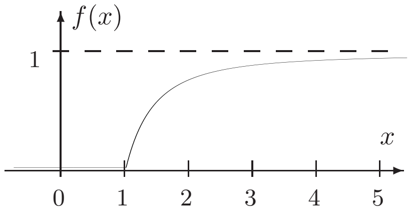


Figure 4.2 *Cumulative distribution function* $F(x)$.

Next, compute the probability for the lifetime to exceed 5 years,

$$\boldsymbol{P}\{X > 5\} = 1 - F(5) = 1 - \left(1 - \frac{1}{5^2}\right) = 0.04.$$

We can also obtain this probability by integrating the density,

$$\boldsymbol{P}\{X > 5\} = \int_5^{+\infty} f(x)dx = \int_5^{+\infty} \frac{2}{x^3}dx = -\left.\frac{1}{x^2}\right|_{x=5}^{+\infty} = \frac{1}{25} = 0.04.$$

◇

Analogy: pmf versus pdf

The role of a density for continuous distributions is very similar to the role of the probability mass function for discrete distributions. Most vital concepts can be translated from the discrete case to the continuous case by replacing pmf $P(x)$ with pdf $f(x)$ and integrating instead of summing.

	Discrete	**Continuous**
Definitions	$P(x) = P\{X = x\}$ (pmf)	$f(x) = F'(x)$ (pdf)
Computing probabilities	$\boldsymbol{P}\{X \in A\} = \sum_{x\in A} P(x)$	$\boldsymbol{P}\{X \in A\} = \int_A f(x)dx$
Cumulative distribution function	$F(x) = \boldsymbol{P}\{X \le x\} = \sum_{y\le x} P(y)$	$F(x) = \boldsymbol{P}\{X \le x\} = \int_{-\infty}^{x} f(y)dy$
Total probability	$\sum_x P(x) = 1$	$\int_{-\infty}^{\infty} f(x)dx = 1$

Joint and marginal densities

DEFINITION 4.2

For a vector of random variables, the **joint cumulative distribution function** is defined as

$$F_{(X,Y)}(x,y) = \boldsymbol{P}\{X \le x \ \cap Y \le y\}.$$

The **joint density** is the *mixed derivative* of the joint cdf,

$$f_{(X,Y)}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{(X,Y)}(x,y).$$

Similarly to the discrete case, a marginal density of X or Y can be obtained by integrating out the other variable. Variables X and Y are *independent* if their joint density factors into the product of marginal densities. Probabilities

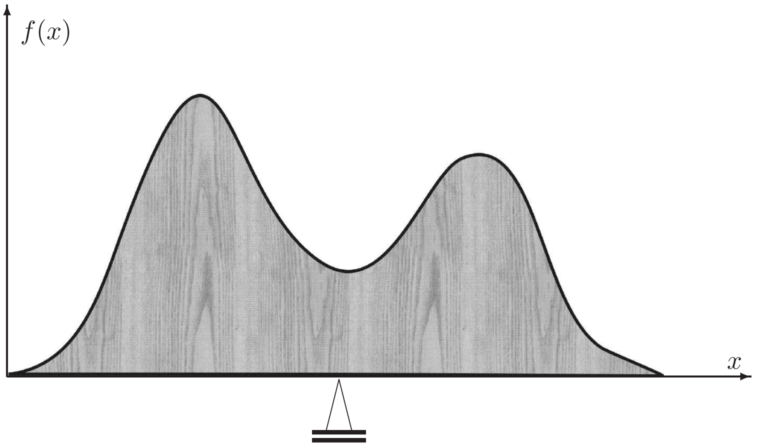


Figure 4.3 *Expectation of a continuous variable as a center of gravity.*

about X and Y can be computed by integrating the joint density over the corresponding set.

	Discrete	**Continuous**
Marginal distributions	$P(x) = \sum_y P(x,y)$ $P(y) = \sum_x P(x,y)$	$f(x) = \int f(x,y)dy$ $f(y) = \int f(x,y)dx$
Independence	$P(x,y) = P(x)P(y)$	$f(x,y) = f(x)f(y)$
Computing probabilities	$\boldsymbol{P}\{(X,Y) \in A\}$ $= \sum\sum_{(x,y)\in A} P(x,y)$	$\boldsymbol{P}\{(X,Y) \in A\}$ $= \iint_{(x,y)\in A} f(x,y)\,dx\,dy$

These concepts are directly extended to three or more variables.

Expectation and variance

Continuing our analogy with the discrete case, *expectation* of a continuous variable is also defined as a center of gravity (also, see p. 50, Figure 3.3). This time, if the entire region below the density curve is cut from a piece of wood, then it will be balanced at a point with coordinate $\mathbf{E}(X)$, as shown in Figure 4.3.

Variance, standard deviation, covariance, and correlation of continuous variables are defined similarly to the discrete case, see Definitions 3.6–3.9 and formula (3.6) on pp. 52–54. All the properties in (3.5), (3.7), and (3.8) extend to the continuous distributions. In calculations, don't forget to replace a pmf with a pdf, and a summation with an integral.

Discrete	**Continuous**
$\mathbf{E}(X) = \sum_x xP(x)$	$\mathbf{E}(X) = \int xf(x)dx$
$\mathrm{Var}(X) = \mathbf{E}(X-\mu)^2$ $= \sum_x (x-\mu)^2 P(x)$ $= \sum_x x^2 P(x) - \mu^2$	$\mathrm{Var}(X) = \mathbf{E}(X-\mu)^2$ $= \int (x-\mu)^2 f(x)dx$ $= \int x^2 f(x)dx - \mu^2$
$\mathrm{Cov}(X,Y) = \mathbf{E}(X-\mu_X)(Y-\mu_Y)$ $= \displaystyle\sum_x\sum_y (x-\mu_X)(y-\mu_Y)P(x,y)$ $= \displaystyle\sum_x\sum_y (xy)P(x,y) - \mu_x\mu_y$	$\mathrm{Cov}(X,Y) = \mathbf{E}(X-\mu_X)(Y-\mu_Y)$ $= \displaystyle\iint (x-\mu_X)(y-\mu_Y)f(x,y)\,dx\,dy$ $= \displaystyle\iint (xy)f(x,y)\,dx\,dy - \mu_x\mu_y$

Example 4.2. A random variable X in Example 4.1 has density

$$f(x) = 2x^{-3} \text{ for } x \geq 1.$$

Its expectation equals

$$\mu = \mathbf{E}(X) = \int x\,f(x)dx = \int_1^\infty 2x^{-2}dx = -2x^{-1}\big|_1^\infty = 2.$$

Computing its variance, we run into a "surprise,"

$$\sigma^2 = \mathrm{Var}(X) = \int x^2 f(x)dx - \mu^2 = \int_1^\infty 2x^{-1}dx - 4 = 2\ln x|_1^\infty - 4 = +\infty.$$

This variable does not have a finite variance! (Also, see the St. Petersburg paradox discussion on p. 66.) ◇

4.2 Families of continuous distributions

As in the discrete case, varieties of phenomena can be described by relatively few families of continuous distributions. Here, we shall discuss Uniform, Exponential, Gamma, and Normal families, adding Student's t and Fisher's F distributions in later chapters.

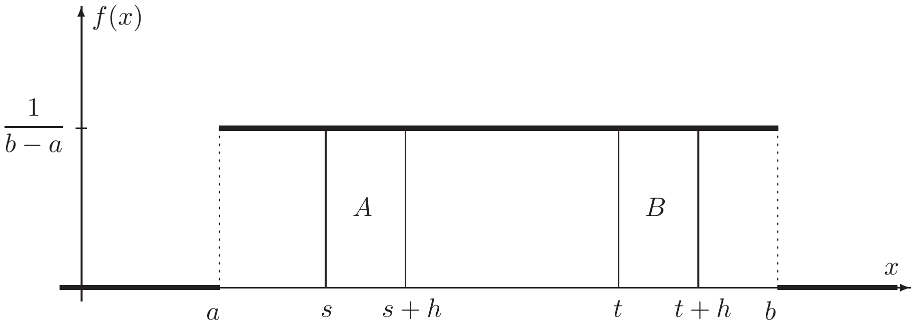


Figure 4.4 *The Uniform density and the Uniform property.*

4.2.1 Uniform distribution

Uniform distribution plays a unique role in stochastic modeling. As we shall see in Chapter 5, a random variable with any thinkable distribution can be generated from a Uniform random variable. Many computer languages and software are equipped with a random number generator that produces Uniform random variables. Users can convert them into variables with desired distributions and use for computer simulation of various events and processes.

Also, Uniform distribution is used in any situation when a value is picked "at random" from a given interval. That is, without any preference given to lower, higher, or medium values. For example, locations of errors in a program, birthdays throughout a year, and many continuous random variables modulo 1, modulo 0.1, 0.01, etc., are uniformly distributed over their corresponding intervals.

To give equal preference to all values, the Uniform distribution has a *constant* density (Figure 4.4). On the interval (a,b), its density equals

$$f(x) = \frac{1}{b-a}, \quad a < x < b,$$

because the rectangular area below the density graph must equal 1.

For the same reason, $|b-a|$ has to be a finite number. There does not exist a Uniform distribution on the entire real line. In other words, if you are asked to choose a random number from $(-\infty, +\infty)$, you cannot do it uniformly.

The Uniform property

For any $h > 0$ and $t \in [a, b-h]$, the probability

$$\boldsymbol{P}\{\, t < X < t+h \,\} = \int_t^{t+h} \frac{1}{b-a}\, dx = \frac{h}{b-a}$$

is *independent* of t. This is the *Uniform property*: the probability is only determined by the length of the interval, but not by its location.

Example 4.3. In Figure 4.4, rectangles A and B have the same area, showing that $\boldsymbol{P}\{s < X < s+h\} = \boldsymbol{P}\{t < X < t+h\}$. ◇

Example 4.4. If a flight scheduled to arrive at 5 pm actually arrives at a Uniformly distributed time between 4:50 and 5:10, then it is equally likely to arrive before 5 pm and after 5 pm, equally likely before 4:55 and after 5:05, etc. ◇

Standard Uniform distribution

The Uniform distribution with $a = 0$ and $b = 1$ is called *Standard Uniform distribution*. The Standard Uniform density is $f(x) = 1$ for $0 < x < 1$. Most random number generators return a Standard Uniform random variable.

All the Uniform distributions are related in the following way. If X is a Uniform(a, b) random variable, then

$$Y = \frac{x-a}{b-a}$$

is Standard Uniform. Likewise, if Y is Standard Uniform, then

$$X = a + (b-a)Y$$

is Uniform(a, b).

A number of other families of distributions have a "standard" member. Typically, a simple transformation converts a standard random variable into a non-standard one, and vice versa.

Expectation and variance

For a Standard Uniform variable Y,

$$\mathbf{E}(Y) = \int y f(y)dy = \int_0^1 y(1)dy = \frac{1}{2}$$

and

$$\mathrm{Var}(Y) = \mathbf{E}(Y^2) - \mathbf{E}^2(Y) = \int_0^1 y^2(1)dy - \left(\frac{1}{2}\right)^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}.$$

Now, consider the general case. Let $X = a+(b-a)Y$ which has a Uniform(a, b) distribution. By the properties of expectations and variances in (3.5) and (3.7),

$$\mathbf{E}(X) = \mathbf{E}\{a + (b-a)Y\} = a + (b-a)\,\mathbf{E}(Y) = a + \frac{b-a}{2} = \frac{a+b}{2}$$

and

$$\operatorname{Var}(X) = \operatorname{Var}\{a + (b-a)Y\} = (b-a)^2 \operatorname{Var}(Y) = \frac{(b-a)^2}{12}.$$

The expectation is precisely the middle of the interval $[a, b]$. Giving no preference to left or right sides, this agrees with the Uniform property and with the physical meaning of $\mathbf{E}(X)$ as a center of gravity.

Uniform distribution

$$\boxed{\begin{array}{lcl} (a,b) & = & \text{range of values} \\ f(x) & = & \dfrac{1}{b-a}, \quad a < x < b \\ \mathbf{E}(X) & = & \dfrac{a+b}{2} \\ \operatorname{Var}(X) & = & \dfrac{(b-a)^2}{12} \end{array}}$$

4.2.2 Exponential distribution

Exponential distribution is often used to model *time*: waiting time, interarrival time, hardware lifetime, failure time, time between telephone calls, etc. As we shall see below, in a sequence of rare events, when the number of events is Poisson, the time between events has Exponential distribution.

Exponential distribution has density

$$f(x) = \lambda e^{-\lambda x} \text{ for } x > 0. \tag{4.2}$$

With this density, we compute the Exponential cdf, mean, and variance as

$$\begin{aligned} F(x) &= \int_0^x f(t)dt = \int_0^x \lambda e^{-\lambda t} dt = 1 - e^{-\lambda x} \quad (x > 0), && (4.3) \\ \mathbf{E}(X) &= \int t f(t)dt = \int_0^\infty t\lambda e^{-\lambda t} dt = \frac{1}{\lambda} \quad \left(\begin{array}{c} \textit{integrating} \\ \textit{by parts} \end{array} \right), && (4.4) \\ \operatorname{Var}(X) &= \int t^2 f(t)dt - \mathbf{E}^2(X) \\ &= \int_0^\infty t^2 \lambda e^{-\lambda t} dt - \left(\frac{1}{\lambda}\right)^2 \quad (\textit{by parts twice}) \\ &= \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}. && (4.5) \end{aligned}$$

The quantity λ is a parameter of Exponential distribution, and its meaning is clear from $\mathbf{E}(X) = 1/\lambda$. If X is time, measured in minutes, then λ is a frequency, measured in min^{-1}. For example, if arrivals occur every half a minute, on the average, then $\mathbf{E}(X) = 0.5$ and $\lambda = 2$, saying that they occur with a frequency (arrival rate) of 2 arrivals per minute. This λ has the same meaning as the parameter of Poisson distribution in Section 3.4.5.

Times between rare events are Exponential

What makes Exponential distribution a good model for interarrival times? Apparently, this is not only experimental, but also a mathematical fact.

As in Section 3.4.5, consider the sequence of rare events, where the number of occurrences during time t has Poisson distribution with a parameter proportional to t. This process is rigorously defined in Section 6.3.2 where we call it *Poisson process*.

Event "the time T until the next event is greater than t" can be rephrased as "zero events occur by the time t," and further, as "$X = 0$," where X is the number of events during the time interval $[0, t]$. This X has Poisson distribution with parameter λt. It equals 0 with probability

$$P_X(0) = e^{-\lambda t}\frac{(\lambda t)^0}{0!} = e^{-\lambda t}.$$

Then we can compute the cdf of T as

$$F_T(t) = 1 - \boldsymbol{P}\{T > t\} = 1 - \boldsymbol{P}\{X = 0\} = 1 - e^{-\lambda t}, \tag{4.6}$$

and here we recognize the Exponential cdf. Therefore, the time until the next arrival has Exponential distribution.

Example 4.5. Jobs are sent to a printer at an average rate of 3 jobs per hour.
(a) What is the expected time between jobs?
(b) What is the probability that the next job is sent within 5 minutes?

Solution. Job arrivals represent rare events, thus the time T between them is Exponential with the given parameter $\lambda = 3\ \text{hrs}^{-1}$ (jobs per hour).
(a) $\mathbf{E}(T) = 1/\lambda = 1/3$ hours or 20 minutes between jobs;
(b) Convert to the same measurement unit: 5 min = (1/12) hrs. Then,

$$\begin{aligned} \boldsymbol{P}\{T < 1/12\,\text{hrs}\} &= F(1/12) = 1 - e^{-\lambda(1/12)} \\ &= 1 - e^{-1/4} = \underline{0.2212}. \end{aligned}$$

◇

Memoryless property

It is said that "Exponential variables lose memory." What does it mean?

Suppose that an Exponential variable T represents waiting time. Memoryless property means that the fact of having waited for t minutes gets "forgotten," and it does not affect the future waiting time. Regardless of the event $T > t$, when the total waiting time exceeds t, the remaining waiting time still has Exponential distribution with the same parameter. Mathematically,

$$\boldsymbol{P}\{T > t + x \mid T > t\} = \boldsymbol{P}\{T > x\} \qquad \text{for } t, x > 0. \tag{4.7}$$

In this formula, t is the already elapsed portion of waiting time, and x is the additional, remaining time.

Proof: From (4.3), $\boldsymbol{P}\{T > x\} = e^{-\lambda x}$. Also, by the formula (2.7) for conditional probability,

$$\begin{aligned} \boldsymbol{P}\{T > t + x \mid T > t\} &= \frac{\boldsymbol{P}\{T > t + x \cap T > t\}}{\boldsymbol{P}\{T > t\}} = \frac{\boldsymbol{P}\{T > t + x\}}{\boldsymbol{P}\{T > t\}} \\ &= \frac{e^{-\lambda(t+x)}}{e^{-\lambda t}} = e^{-\lambda x}. \end{aligned}$$

□

This property is unique for Exponential distribution. No other continuous variable $X \in (0, \infty)$ is memoryless. Among discrete variables, such a property belongs to *Geometric distribution* (Exercise 4.27).

In a sense, Geometric distribution is a discrete analogue of Exponential. Connection between these two families of distributions will become clear in Section 5.2.3.

Exponential distribution

$$\begin{aligned} \lambda &= \text{frequency, the number of events per time unit} \\ f(x) &= \lambda e^{-\lambda x}, \quad x > 0 \\ \mathbf{E}(X) &= \frac{1}{\lambda} \\ \mathrm{Var}(X) &= \frac{1}{\lambda^2} \end{aligned}$$

4.2.3 Gamma distribution

When a certain procedure consists of α independent steps, and each step takes Exponential(λ) amount of time, then the total time has *Gamma distribution* with parameters α and λ.

Thus, Gamma distribution can be widely used for the total time of a multistage scheme, for example, related to downloading or installing a number of files. In a process of rare events, with Exponential times between any two consecutive events, the time of the α-th event has Gamma distribution because it consists of α independent Exponential times.

Example 4.6 (Internet promotions). Users visit a certain internet site at the average rate of 12 hits per minute. Every sixth visitor receives some promotion that comes in a form of a flashing banner. Then the time between consecutive promotions has Gamma distribution with parameters $\alpha = 6$ and $\lambda = 12$. ◇

Having two parameters, Gamma distribution family offers a variety of models for positive random variables. Besides the case when a Gamma variable represents a sum of independent Exponential variables, Gamma distribution is often used for the amount of money being paid, amount of a commodity being used (gas, electricity, etc.), a loss incurred by some accident, etc.

Gamma distribution has a density

$$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)}\, x^{\alpha-1} e^{-\lambda x}, \qquad x > 0. \tag{4.8}$$

The denominator contains a *Gamma function*, Section 11.3.5. With certain techniques, this density can be mathematically derived for integer α by representing a Gamma variable X as a sum of Exponential variables each having a density (4.2).

In fact, α can take any positive value, not necessarily integer. With different α, the Gamma density takes different shapes (Figure 4.5). For this reason, α is called a *shape parameter*.

A special case of $\alpha = 1$ yields Exponential distribution,

$$\boxed{\text{Gamma}(1, \lambda) = \text{Exponential}(\lambda)}$$

This can be seen comparing (4.8) and (4.2) for $\alpha = 1$.

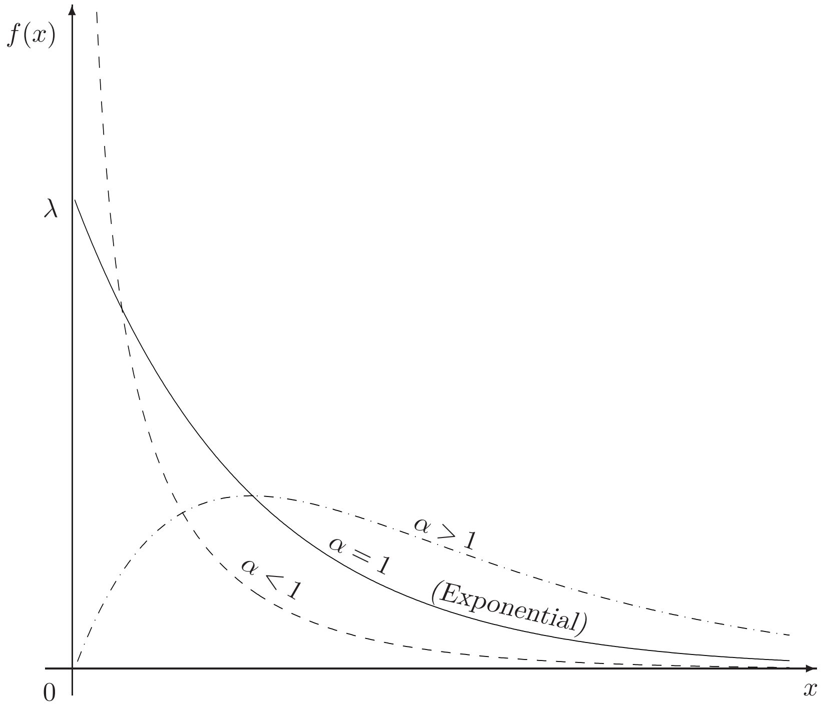


Figure 4.5 *Gamma densities with different shape parameters α.*

Expectation, variance, and some useful integration remarks

Gamma cdf has the form

$$F(t) = \int_0^t f(x)dx = \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^t x^{\alpha-1} e^{-\lambda x} dx. \tag{4.9}$$

This expression, related to a so-called *incomplete Gamma function*, does not simplify, and thus, computing probabilities is not always trivial. Let us offer several computational shortcuts.

First, let us notice that $\int_0^\infty f(x)dx = 1$ for Gamma and all the other densities. Then, integrating (4.8) from 0 to ∞, we obtain that

$$\boxed{\int_0^\infty x^{\alpha-1} e^{-\lambda x} dx = \frac{\Gamma(\alpha)}{\lambda^\alpha} \qquad \text{for any } \alpha > 0 \text{ and } \lambda > 0} \tag{4.10}$$

Substituting $\alpha + 1$ and $\alpha + 2$ in place of α, we get for a Gamma variable X,

$$\mathbf{E}(X) = \int x f(x) dx = \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty x^\alpha e^{-\lambda x} dx = \frac{\lambda^\alpha}{\Gamma(\alpha)} \cdot \frac{\Gamma(\alpha+1)}{\lambda^{\alpha+1}} = \frac{\alpha}{\lambda} \tag{4.11}$$

(using the equality $\Gamma(t+1) = t\Gamma(t)$ that holds for all $t > 0$),

$$\mathbf{E}(X^2) = \int x^2 f(x)dx = \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha+1} e^{-\lambda x} dx = \frac{\lambda^\alpha}{\Gamma(\alpha)} \cdot \frac{\Gamma(\alpha+2)}{\lambda^{\alpha+2}} = \frac{(\alpha+1)\alpha}{\lambda^2},$$

and therefore,

$$\operatorname{Var}(X) = \mathbf{E}(X^2) - \mathbf{E}^2(X) = \frac{(\alpha+1)\alpha - \alpha^2}{\lambda^2} = \frac{\alpha}{\lambda^2}. \tag{4.12}$$

For $\alpha = 1$, this agrees with (4.4) and (4.5). Moreover, for any integer α, (4.11) and (4.12) can be obtained directly from (4.4) and (4.5) by representing a Gamma variable X as a sum of independent Exponential(λ) variables $X_1, \ldots, X_\alpha$,

$$\begin{array}{rcl} \mathbf{E}(X) & = & \mathbf{E}(X_1 + \ldots + X_\alpha) = \mathbf{E}(X_1) + \ldots + \mathbf{E}(X_\alpha) = \alpha\left(\dfrac{1}{\lambda}\right), \\ \operatorname{Var}(X) & = & \operatorname{Var}(X_1 + \ldots + X_\alpha) = \operatorname{Var}(X_1) + \ldots + \operatorname{Var}(X_\alpha) = \alpha\left(\dfrac{1}{\lambda^2}\right). \end{array}$$

Gamma distribution

$$\boxed{\begin{array}{lcl} \alpha & = & \text{shape parameter} \\ \lambda & = & \text{frequency} \\ f(x) & = & \dfrac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, \quad x > 0 \\ \mathbf{E}(X) & = & \dfrac{\alpha}{\lambda} \\ \operatorname{Var}(X) & = & \dfrac{\alpha}{\lambda^2} \end{array}} \tag{4.13}$$

Example 4.7 (Total compilation time). Compilation of a computer program consists of 3 blocks that are processed sequentially, one after another. Each block takes Exponential time with the mean of 5 minutes, independently of other blocks.

(a) Compute the expectation and variance of the total compilation time.

(b) Compute the probability for the entire program to be compiled in less than 12 minutes.

Solution. The total time T is a sum of three independent Exponential times, therefore, it has Gamma distribution with $\alpha = 3$. The frequency parameter λ equals $(1/5)$ min^{-1} because the Exponential compilation time of each block has expectation $1/\lambda = 5$ min.

(a) For a Gamma random variable T with $\alpha = 3$ and $\lambda = 1/5$,

$$\mathbf{E}(T) = \frac{3}{1/5} = 15 \text{ (min)} \quad \text{and} \quad \mathrm{Var}(T) = \frac{3}{(1/5)^2} = 75 \text{ (min}^2\text{)}.$$

(b) A direct solution involves repeated integration by parts,

$$\begin{aligned}
\boldsymbol{P}\{T < 12\} &= \int_0^{12} f(t)dt = \frac{(1/5)^3}{\Gamma(3)} \int_0^{12} t^2 e^{-t/5} dt \\
&= \frac{(1/5)^3}{2!} \left(-5t^2 e^{-t/5} \Big|_{t=0}^{t=12} + \int_0^{12} 10te^{-t/5} dt \right) \\
&= \frac{1/125}{2} \left(-5t^2 e^{-t/5} - 50te^{-t/5} \Big|_{t=0}^{t=12} + \int_0^{12} 50e^{-t/5} dt \right) \\
&= \frac{1}{250} \left(-5t^2 e^{-t/5} - 50te^{-t/5} - 250e^{-t/5} \right) \Big|_{t=0}^{t=12} \\
&= 1 - e^{-2.4} - 2.4e^{-2.4} - 2.88e^{-2.4} = \underline{0.4303}. \qquad (4.14)
\end{aligned}$$

A much shorter way is to apply the Gamma-Poisson formula below (example 4.8). ◇

Gamma-Poisson formula

Computation of Gamma probabilities can be significantly simplified if a Gamma variable is interpreted as the time between rare events. In particular, one can avoid lengthy integration by parts, as in Example 4.7, and use Poisson distribution instead.

Indeed, let T be a Gamma variable with an integer parameter α and some positive λ. This is a distribution of the time of the α-th rare event. Then, the event $\{T > t\}$ means that the α-th rare event occurs after the moment t, and therefore, *fewer than α rare events occur before the time t*. We see that

$$\{T > t\} = \{X < \alpha\},$$

where X is the number of events that occur before the time t. This number of rare events X has Poisson distribution with parameter (λt), therefore, the probability

$$\boldsymbol{P}\{T > t\} = \boldsymbol{P}\{X < \alpha\}$$

and the probability of a complement

$$\boldsymbol{P}\{T \le t\} = \boldsymbol{P}\{X \ge \alpha\}$$

can both be computed using the Poisson distribution of X.

Gamma-Poisson formula

For a Gamma(α, λ) variable T and a Poisson(λt) variable X,

$$\boldsymbol{P}\{T > t\} = \boldsymbol{P}\{X < \alpha\}$$

$$\boldsymbol{P}\{T \leq t\} = \boldsymbol{P}\{X \geq \alpha\} \qquad (4.15)$$

Remark: Recall that $\boldsymbol{P}\{T > t\} = \boldsymbol{P}\{T \geq t\}$ and $\boldsymbol{P}\{T < t\} = \boldsymbol{P}\{T \leq t\}$ for a Gamma variable T, because it is continuous. Hence, (4.15) can also be used for the computation of $\boldsymbol{P}\{T \geq t\}$ and $\boldsymbol{P}\{T < t\}$. Conversely, the probability of $\{X = \alpha\}$ cannot be neglected for the Poisson (discrete!) variable X, thus the signs in the right-hand sides of (4.15) cannot be altered.

Example 4.8 (Total compilation time, continued). Here is an alternative solution to Example 4.7(b). According to the Gamma-Poisson formula with $\alpha = 3$, $\lambda = 1/5$, and $t = 12$,

$$\boldsymbol{P}\{T < 12\} = \boldsymbol{P}\{X \geq 3\} = 1 - F(2) = 1 - 0.5697 = \underline{0.430}$$

from Table A3 for the Poisson distribution of X with parameter $\lambda t = 2.4$.

Furthermore, we notice that the four-term expression we obtained in (4.14) after integrating by parts represents precisely

$$\boldsymbol{P}\{X \geq 3\} = 1 - P(0) - P(1) - P(2).$$

◇

Example 4.9. Lifetimes of computer memory chips have Gamma distribution with expectation $\mu = 12$ years and standard deviation $\sigma = 4$ years. What is the probability that such a chip has a lifetime between 8 and 10 years?

Solution.

Step 1, parameters. From the given data, compute parameters of this Gamma distribution. Using (4.13), obtain a system of two equations and solve them for α and λ,

$$\left\{\begin{array}{lcl} \mu & = & \alpha/\lambda \\ \sigma^2 & = & \alpha/\lambda^2 \end{array}\right. \Rightarrow \left\{\begin{array}{lclclcl} \alpha & = & \mu^2/\sigma^2 & = & (12/4)^2 & = & 9, \\ \lambda & = & \mu/\sigma^2 & = & 12/4^2 & = & 0.75. \end{array}\right.$$

Step 2, probability. We can now compute the probability,

$$\boldsymbol{P}\{8 < T < 10\} = F_T(10) - F_T(8). \qquad (4.16)$$

For each term in (4.16), we use the Gamma-Poisson formula with $\alpha = 9$,

$\lambda = 0.75$, and $t = 8, 10$,

$$F_T(10) = \boldsymbol{P}\{T \le 10\} = \boldsymbol{P}\{X \ge 9\} = 1 - F_X(8) = 1 - 0.662 = 0.338$$

from Table A3 for a Poisson variable X with parameter $\lambda t = (0.75)(10) = 7.5$;

$$F_T(8) = \boldsymbol{P}\{T \le 8\} = \boldsymbol{P}\{X \ge 9\} = 1 - F_X(8) = 1 - 0.847 = 0.153$$

from the same table, this time with parameter $\lambda t = (0.75)(8) = 6$. Then,

$$\boldsymbol{P}\{8 < T < 10\} = 0.338 - 0.153 = \underline{0.185}.$$

◇

4.2.4 Normal distribution

Normal distribution plays a vital role in Probability and Statistics, mostly because of the Central Limit Theorem, according to which sums and averages often have approximately Normal distribution. Due to this fact, various fluctuations and measurement errors that consist of accumulated number of small terms are normally distributed.

Remark: As said by a French mathematician *Jules Henri Poincaré*, "Everyone believes in the Normal law of errors, the experimenters because they think it is a mathematical theorem, the mathematicians because they think it is an experimental fact."

Besides sums, averages, and errors, Normal distribution is often found to be a good model for physical variables like weight, height, temperature, voltage, pollution level, and for instance, household incomes or student grades.

Normal distribution has a density

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\}, \qquad -\infty < x < +\infty,$$

where parameters μ and σ have a simple meaning of the expectation $\mathbf{E}(X)$ and the standard deviation $\mathrm{Std}(X)$. This density is known as the bell-shape curve, symmetric and centered at μ, its spread being controlled by σ. As seen in Figure 4.6, changing μ shifts the curve to the left or to the right without changing its shape, while changing σ makes it more concentrated or more flat. Often μ is called a *location parameter* and σ is called a *scale parameter*.

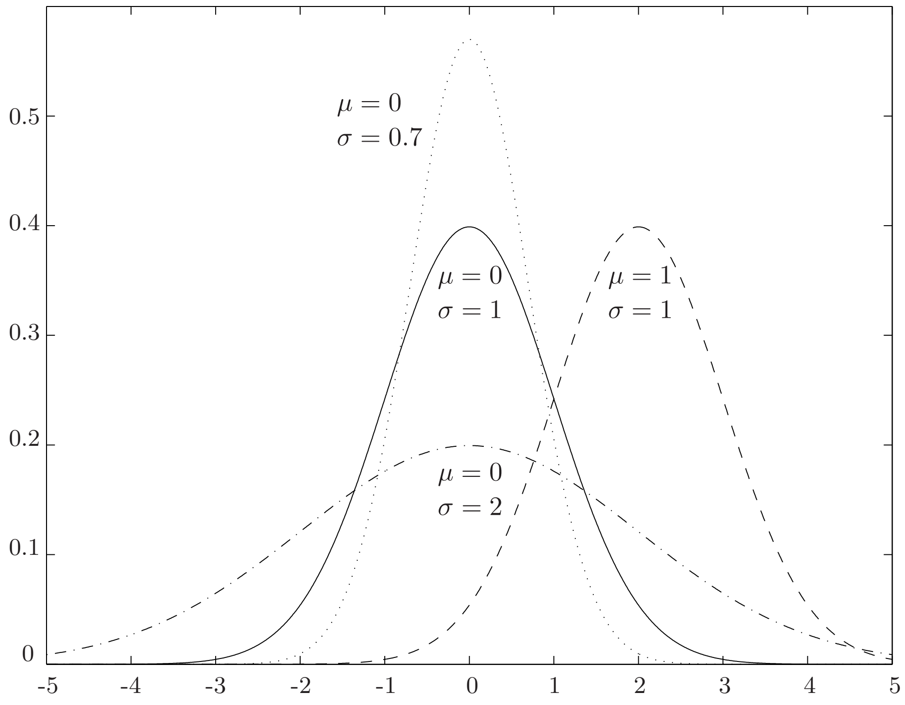


Figure 4.6 *Normal densities with different location and scale parameters.*

Normal distribution

$$
\begin{aligned}
\mu &= \text{expectation, location parameter} \\
\sigma &= \text{standard deviation, scale parameter} \\
f(x) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\}, \; -\infty < x < \infty \\
\mathbf{E}(X) &= \mu \\
\mathrm{Var}(X) &= \sigma^2
\end{aligned}
$$

Standard Normal distribution

DEFINITION 4.3

Normal distribution with "standard parameters" $\mu = 0$ and $\sigma = 1$ is called **Standard Normal distribution**.

Notation

$$\begin{array}{rcl} Z & = & \text{Standard Normal random variable} \\ \phi(x) & = & \dfrac{1}{\sqrt{2\pi}}\, e^{-x^2/2}, \text{ Standard Normal pdf} \\ \Phi(x) & = & \displaystyle\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}}\, e^{-z^2/2} dz, \text{ Standard Normal cdf} \end{array}$$

A Standard Normal variable, usually denoted by Z, can be obtained from a non-standard Normal(μ, σ) random variable X by *standardizing*, that is, subtracting the mean and dividing by the standard deviation,

$$Z = \frac{X - \mu}{\sigma}. \tag{4.17}$$

Unstandardizing Z, we can reconstruct the initial variable X,

$$X = \mu + \sigma\, Z. \tag{4.18}$$

Using these transformations, any Normal random variable can be obtained from a Standard Normal variable Z, therefore, we need a table of Standard Normal Distribution only (Table A4).

To find $\Phi(z)$ from Table A4, we locate a row with the first two digits of z and a column with the third digit of z and read the probability $\Phi(z)$ at their intersection. Notice that $\Phi(z) \approx 0$ (is "practically" zero) for all $z < -3.9$, and $\Phi(z) \approx 1$ (is "practically" one) for all $z > 3.9$.

Example 4.10 (Computing Standard Normal probabilities). For a Standard Normal random variable Z,

$$\begin{array}{rcl} \boldsymbol{P}\{Z < 1.35\} & = & \Phi(1.35) = 0.9115 \\ \boldsymbol{P}\{Z > 1.35\} & = & 1 - \Phi(1.35) = 0.0885 \\ \boldsymbol{P}\{-0.77 < Z < 1.35\} & = & \Phi(1.35) - \Phi(-0.77) = 0.9115 - 0.2206 = 0.6909. \end{array}$$

according to Table A4. Notice that $\boldsymbol{P}\{Z < -1.35\} = 0.0885 = \boldsymbol{P}\{Z > 1.35\}$, which is explained by the symmetry of the Standard Normal density in Figure 4.6. Due to this symmetry, "the left tail," or the area to the left of (-1.35) equals "the right tail," or the area to the right of 1.35. ◇

In fact, the symmetry of the Normal density, mentioned in this example, allows to obtain the first part of Table A4 on p. 386 directly from the second part,

$$\Phi(-z) = 1 - \Phi(z) \qquad \text{for } -\infty < z < +\infty.$$

To compute probabilities about an arbitrary Normal random variable X, we have to standardize it first, as in (4.17), then use Table A4.

Example 4.11 (Computing non-standard Normal probabilities). Suppose that the average household income in some country is 900 coins, and the standard deviation is 200 coins. Assuming the Normal distribution of incomes, compute the proportion of "middle class," whose income is between 600 and 1200 coins.

Solution. For a Normal($\mu = 900$, $\sigma = 200$) random variable X,

$$\begin{aligned} \boldsymbol{P}\{600 < X < 1200\} &= \boldsymbol{P}\left\{\frac{600-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{1200-\mu}{\sigma}\right\} \\ &= \boldsymbol{P}\left\{\frac{600-900}{200} < Z < \frac{1200-900}{200}\right\} = \boldsymbol{P}\{-1.5 < Z < 1.5\} \\ &= \Phi(1.5) - \Phi(-1.5) = 0.9332 - 0.0668 = \underline{0.8664}. \end{aligned}$$

◇

So far, we were computing probabilities of clearly defined events. These are *direct* problems. A number of applications require solution of an *inverse problem*, that is, finding a value of x given the corresponding probability.

Example 4.12 (Inverse problem). The government of the country in Example 4.11 decides to issue food stamps to the poorest 3% of households. Below what income will families receive food stamps?

Solution. We need to find such income x that $\boldsymbol{P}\{X < x\} = 3\% = 0.03$. This is an equation that can be solved in terms of x. Again, we standardize first, then use the table:

$$\boldsymbol{P}\{X < x\} = \boldsymbol{P}\left\{Z < \frac{x-\mu}{\sigma}\right\} = \Phi\left(\frac{x-\mu}{\sigma}\right) = 0.03,$$

from where

$$x = \mu + \sigma\Phi^{-1}(0.03).$$

In Table A4, we have to find the probability, the *table entry* of 0.03. We see that $\Phi(-1.88) \approx 0.03$. Therefore, $\Phi^{-1}(0.03) = -1.88$, and

$$x = \mu + \sigma(-1.88) = 900 + (200)(-1.88) = \underline{524} \text{ (coins)}$$

is the answer. ◇

As seen in this example, in order to solve an inverse problem, we use the table first, then *unstandardize*, as in (4.18), and find the required value of x.

4.3 Central Limit Theorem

We now turn our attention to *sums* of random variables,

$$S_n = X_1 + \ldots + X_n,$$

that appear in many applications. Let $\mu = \mathbf{E}(X_i)$ and $\sigma = \mathrm{Std}(X_i)$ for all $i = 1, \ldots, n$. How does S_n behave for large n?

The following MATLAB code is a good illustration to the behavior of partial sums S_n.

```
S(1)=0;
for n=2:1000; S(n)=S(n-1)+randn; end;   % $n^{th}$ partial sum
n=1:1000; comet(n,S); pause(3);         % Behavior of $S(n)$
comet(n,S./n); pause(3);                % Behavior of $S(n)/n$
comet(n,S./sqrt(n)); pause(3);          % Behavior of $S(n)/\sqrt{n}$
```

Apparently (users of MATLAB can see it from the obtained graphs),

- The *pure sum S_n diverges.* In fact, this should be anticipated because
$$\mathrm{Var}(S_n) = n\sigma^2 \to \infty,$$
so that variability of S_n grows unboundedly as n goes to infinity.
- The *average S_n/n converges.* Indeed, in this case, we have
$$\mathrm{Var}(S_n/n) = \mathrm{Var}(S_n)/n^2 = n\sigma^2/n^2 = \sigma^2/n \to 0,$$
so that variability of $\mathrm{Var}(S_n/n)$ vanishes as $n \to \infty$.
- An interesting normalization factor is $1/\sqrt{n}$. For $\mu = 0$, we can see from MATLAB simulations that $S_n/\sqrt{n}$ *neither diverges nor converges!* It does not tend to leave 0, but does not converge to 0 either. Rather, it behaves like some random variable. The following theorem states that this variable has approximately Normal distribution for large n.

Theorem 1 (Central Limit Theorem) *Let $X_1, X_2, \ldots$ be independent random variables with the same expectation $\mu = \mathbf{E}(X_i)$ and the same standard deviation $\sigma = \mathrm{Std}(X_i)$, and let*

$$S_n = \sum_{i=1}^{n} X_i = X_1 + \ldots + X_n.$$

As $n \to \infty$, the standardized sum

$$Z_n = \frac{S_n - \mathbf{E}(S_n)}{\mathrm{Std}(S_n)} = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

converges in distribution to a Standard Normal random variable, that is,

$$F_{Z_n}(z) = \boldsymbol{P}\left\{\frac{S_n - n\mu}{\sigma\sqrt{n}} \le z\right\} \to \Phi(z) \tag{4.19}$$

for all z.

This theorem is very powerful because it can be applied to random variables $X_1, X_2, \ldots$ having virtually any thinkable distribution with finite expectation and variance. As long as n is large (the rule of thumb is $n > 30$), one can use Normal distribution to compute probabilities about S_n.

Theorem 1 is only one basic version of the Central Limit Theorem. Over the last two centuries, it has been extended to large classes of dependent variables and vectors, stochastic processes, and so on.

Example 4.13 (Allocation of disk space). A disk has free space of 330 megabytes. Is it likely to be sufficient for 300 independent images, if each image has expected size of 1 megabyte with a standard deviation of 0.5 megabytes?

Solution. We have $n = 300$, $\mu = 1$, $\sigma = 0.5$. The number of images n is large, so the Central Limit Theorem applies. Then,

$$\begin{aligned} \boldsymbol{P}\left\{\text{sufficient space}\right\} &= \boldsymbol{P}\left\{S_n \le 330\right\} = \boldsymbol{P}\left\{\frac{S_n - n\mu}{\sigma\sqrt{n}} \le \frac{330 - (300)(1)}{0.5\sqrt{300}}\right\} \\ &\approx \Phi(3.46) = 0.9997. \end{aligned}$$

This probability probability is very high, hence, the available disk space is very likely to be sufficient. ◇

In the special case of *Normal* variables $X_1, X_2, \ldots$, the distribution of S_n is always Normal, and (4.19) becomes exact equality for arbitrary, even small n.

Example 4.14 (Elevator). You wait for an elevator, whose capacity is 2000 pounds. The elevator comes with ten adult passengers. Suppose your own weight is 150 lbs, and you heard that human weights are normally distributed with the mean of 165 lbs and the standard deviation of 20 lbs. Would you board this elevator or wait for the next one?

Solution. In other words, is overload likely? The probability of an overload equals

$$\begin{aligned} \boldsymbol{P}\left\{S_{10} + 150 > 2000\right\} &= \boldsymbol{P}\left\{\frac{S_{10} - (10)(165)}{20\sqrt{10}} > \frac{2000 - 150 - (10)(165)}{20\sqrt{10}}\right\} \\ &= 1 - \Phi(3.16) = 0.0008. \end{aligned}$$

In other words, it is safe with probability 0.9992. It is now for you to decide. ◇

Among the random variables discussed in Chapters 3 and 4, at least three have a form of S_n:

Binomial variable	=	sum of independent Bernoulli variables
Negative Binomial variable	=	sum of independent Geometric variables
Gamma variable	=	sum of independent Exponential variables

Hence, the Central Limit Theorem applies to all these cases, and with sufficiently large n in the case of Binomial, k for Negative Binomial, and α for Gamma variables.

In fact, *Abraham de Moivre* (1667–1754) obtained the first version of the Central Limit Theorem as the approximation of Binomial distribution.

Normal approximation to Binomial distribution

Binomial variables represent a special case of $S_n = X_1 + \ldots + X_n$, where all X_i have Bernoulli distribution with some parameter p. We know from Section 3.4.5 that small p allows to approximate Binomial distribution with Poisson, and large p allows such an approximation for the number of failures. For all other p (say, $0.05 \le p \le 0.95$), and large n, we can use Theorem 1:

$$\text{Binomial}(n,p) \approx Normal\left(\mu = np, \sigma = \sqrt{np(1-p)}\right) \tag{4.20}$$

To visualize how Binomial distribution gradually takes the shape of Normal distribution when $n \to \infty$, you may execute the following MATLAB code that graphs Binomial(n,p) pmf for increasing values of n.

```
 for n=1:2:100; x=0:n;
p=gamma(n+1)./gamma(x+1)./gamma(n-x+1).*(0.3).^x.*(0.7).^(n-x);
plot(x,p);
title('Binomial probabilities for increasing values of n');
pause(1); end;
```

Continuity correction

This correction is often needed when we approximate a discrete distribution (Binomial in this case) by a continuous distribution (Normal). Recall that the probability $\boldsymbol{P}\{X = x\}$ may be positive if X is discrete, whereas it is always 0 for continuous X. Thus, a direct use of (4.20) will always approximate this probability by 0. This is obviously a poor approximation.

This is resolved by introducing a *continuity correction*. Expand the interval by 0.5 units in each direction, then use the Normal approximation. Notice that

$$P_X(x) = \boldsymbol{P}\{X = x\} = \boldsymbol{P}\{x - 0.5 < X < x + 0.5\}$$

for a Binomial variable X, therefore, the continuity correction does not change the event and preserves its probability. It makes a difference for the Normal distribution. Now it is the probability of an interval instead of one number, and it is not zero.

Example 4.15. A new computer virus attacks a folder consisting of 200 files. Each file gets damaged with probability 0.2 independently of other files. What is the probability that fewer than 50 files get damaged?

Solution. The number X of damaged files has Binomial distribution with $n = 200$, $p = 0.2$, $\mu = np = 40$, and $\sigma = \sqrt{np(1-p)} = 5.657$. Applying the Central Limit Theorem with the continuity correction,

$$\begin{aligned} \boldsymbol{P}\{X < 50\} &= \boldsymbol{P}\{X < 49.5\} = \boldsymbol{P}\left\{\frac{X-40}{5.657} < \frac{49.5-40}{5.657}\right\} \\ &= \Phi(1.68) = \underline{0.9535}. \end{aligned}$$

Notice that the properly applied continuity correction replaces 50 with 49.5, not 50.5. Indeed, we are interested in the event that X is *strictly* less than 50. This includes all values up to 49 and corresponds to the interval $[0, 49]$ that we *expand* to $[0, 49.5]$. In other words, events $\{X < 50\}$ and $\{X < 49.5\}$ are the same, they include the same possible values of X. Events $\{X < 50\}$ and $\{X < 50.5\}$ are different because the former includes $X = 50$, and the latter does not. Replacing $\{X < 50\}$ with $\{X < 50.5\}$ would have changed its probability and would have given a wrong answer. ◇

When a continuous distribution (say, Gamma) is approximated by another continuous distribution (Normal), the continuity correction is not needed. In fact, it would be an error to use it in this case because it would no longer preserve the probability.

Summary and conclusions

Continuous distributions are used to model various times, sizes, measurements, and all other random variables that assume an entire interval of possible values.

Continuous distributions are described by their densities that play a role analogous to probability mass functions of discrete variables. Computing probabilities essentially reduces to integrating a density over the given set. Expectations and variances are defined similarly to the discrete case, replacing a probability mass function by a density and summation by integration.

In different situations, one uses Uniform, Exponential, Gamma, Normal distributions. A few other families are studied in later chapters.

The Central Limit Theorem states that a standardized sum of a large number of independent random variables is approximately Normal, thus Table A4 can be used to compute related probabilities. A continuity correction should be used when a discrete distribution is approximated by a continuous distribution.

Characteristics of continuous families are summarized in Section 11.1.2.

Questions and exercises

4.1. The lifetime, in years, of some electronic component is a continuous random variable with the density

$$f(x) = \begin{cases} \dfrac{k}{x^4} & \text{for} \quad x \geq 1 \\ 0 & \text{for} \quad x < 1. \end{cases}$$

Find k, the cumulative distribution function, and the probability for the lifetime to exceed 2 years.

4.2. The time, in minutes, it takes to reboot a certain system is a continuous variable with the density

$$f(x) = \begin{cases} C(10-x)^2, & \text{if } 0 < x < 10 \\ 0, & \text{otherwise} \end{cases}$$

(a) Compute C.

(b) Compute the probability that it takes between 1 and 2 minutes to reboot.

4.3. The installation time, in hours, for a certain software module has a probability density function $f(x) = k(1 - x^3)$ for $0 < x < 1$. Find k and compute the probability that it takes less than 1/2 hour to install this module.

4.4. Two continuous random variables X and Y have the joint density

$$f(x,y) = C(x^2 + y), \quad -1 \leq x \leq 1, \quad 0 \leq y \leq 1.$$

(a) Compute the constant C.

(b) Compute probabilities $P\{Y < 0.6\}$ and $P\{Y < 0.6 \mid X = 0.5\}$.

4.5. Lifetime of a certain hardware is a continuous random variable with density

$$f(x) = \begin{cases} K - x/50 & \text{for } 0 < x < 10 \text{ years} \\ 0 & \text{for all other } x \end{cases}$$

(a) Find K.

(b) What is the probability of a failure within the first 5 years?

(c) What is the expectation of the lifetime?

4.6. A program is divided into 3 blocks that are being compiled on 3 parallel computers. Each block takes an Exponential amount of time, 5 minutes on the average, independently of other blocks. The program is completed when *all* the blocks are compiled. Compute the expected time it takes the program to be compiled.

4.7. The time it takes a printer to print a job is an Exponential random variable with the expectation of 12 seconds. You send a job to the printer at 10:00 am, and it appears to be third in line. What is the probability that your job will be ready before 10:01?

4.8. For some electronic component, the time until failure has Gamma distribution with parameters $\alpha = 2$ and $\lambda = 2$ (years^{-1}). Compute the probability that the component fails within the first 6 months.

4.9. On the average, a computer experiences breakdowns every 5 months. The time until the first breakdown and the times between any two consecutive breakdowns are independent Exponential random variables. After the third breakdown, a computer requires a special maintenance.

(a) Compute the probability that a special maintenance is required within the next 9 months.

(b) Given that a special maintenance was not required during the first 12 months, what is the probability that it will not be required within the next 4 months?

4.10. Two computer specialists are completing work orders. The first specialist receives 60% of all orders. Each order takes her Exponential amount of time with parameter $\lambda_1 = 3$ hrs^{-1}. The second specialist receives the remaining 40% of orders. Each order takes him Exponential amount of time with parameter $\lambda_2 = 2$ hrs^{-1}.

A certain order was submitted 30 minutes ago, and it is still not ready. What is the probability that the first specialist is working on it?

4.11. Consider a satellite whose work is based on a certain block A. This block has an independent backup B. The satellite performs its task until both A and B fail. The lifetimes of A and B are exponentially distributed with the mean lifetime of 10 years.

(a) What is the probability that the satellite will work for more than 10 years?

(b) Compute the expected lifetime of the satellite.

4.12. A computer processes tasks in the order they are received. Each task takes an Exponential amount of time with the average of 2 minutes. Compute the probability that a package of 5 tasks is processed in less than 8 minutes.

4.13. On the average, it takes 25 seconds to download a file from the internet. If it takes an Exponential amount of time to download one file, then what is the probability that it will take more than 70 seconds to download 3 independent files?

4.14. The time X, in minutes, it takes to reboot a certain system has Gamma distribution with $\mathbf{E}(X) = 20$ and $\mathrm{Std}(X) = 10$.

(a) Compute parameters of this distribution.

(b) What is the probability that it takes less than 15 minutes to reboot this system?

4.15. A certain system is based on two independent modules, A and B. A failure of any module causes a failure of the whole system. The lifetime of each module has a Gamma distribution, with parameters α and λ given in the table,

Component	α	λ (years^{-1})
A	3	1
B	2	2

(a) What is the probability that the system works at least 2 years without a failure?

(b) Given that the system failed during the first 2 years, what is the probability that it failed due to the failure of component B (but not component A)?

4.16. The lifetime of a certain electronic component is a random variable with the expectation of 5000 hours and a standard deviation of 100 hours. What is the probability that the average lifetime of 400 components is less than 5012 hours?

4.17. Upgrading a certain software package requires installation of 82 new files. Files are installed consecutively. The installation time is random, but on the average, it takes 15 sec to install one file, with a variance of 16 sec^2. What is the probability that the whole package is upgraded in less than 20 minutes?

4.18. Among all the computer chips produced by a certain factory, 6 percent are defective. A sample of 400 chips is selected for inspection.

(a) What is the probability that this sample contains between 20 and 25 defective chips (including 20 and 25)?

(b) Suppose that each of 40 inspectors collects a sample of 400 chips. What is the probability that at least 8 inspectors will find between 20 and 25 defective chips in their samples?

4.19. An average scanned image occupies 0.6 megabytes of memory with a standard deviation of 0.4 megabytes. If you plan to install 80 images on your web site, what is the probability that their total size is between 47 megabytes and 50 megabytes?

4.20. A certain computer virus can damage any file with probability 35%, independently of other files. Suppose this virus enters a folder containing 2400 files. Compute the probability that between 800 and 850 files get damaged.

4.21. Seventy independent messages are sent from an electronic transmission center. Messages are processed sequentially, one after another. Transmission time of each message is Exponential with parameter $\lambda = 5$ min^{-1}. Find the probability that all 70 messages are transmitted in less than 12 minutes. Use the Central Limit Theorem.

4.22. A computer lab has two printers. Printer I handles 40% of all the jobs. Its printing time is Exponential with the mean of 2 minutes. Printer II handles the remaining 60% of jobs. Its printing time is Uniform between 0 minutes and 5 minutes. A job is printed in less than 1 minute. What is the probability that it was printed by Printer I?

4.23. An internet service provider has two connection lines for its customers. Eighty percent of customers are connected through Line I, and twenty percent are connected through Line II. Line I has a Gamma connection time with parameters $\alpha = 3$ and $\lambda = 2$ min^{-1}. Line II has a Uniform(a, b) connection time with parameters $a = 20$ sec and $b = 50$ sec. Compute the probability that it takes a randomly selected customer more than 30 seconds to connect to the internet.

4.24. Upgrading a certain software package requires installation of 68 new files. Files are installed consecutively. The installation time is random, but on the average, it takes 15 sec to install one file, with a variance of 11 sec^2.

(a) What is the probability that the whole package is upgraded in less than 12 minutes?

(b) A new version of the package is released. It requires only N new files to be installed, and it is promised that 95% of the time upgrading takes less than 10 minutes. Given this information, compute N.

4.25. Two independent customers are scheduled to arrive in the afternoon. Their arrival times are uniformly distributed between 2 pm and 8 pm. Compute

(a) the expected time of the first (earlier) arrival;

(b) the expected time of the last (later) arrival.

4.26. Let X and Y be independent Standard Uniform random variables.

(a) Find the probability of an event $\{0.5 < (X + Y) < 1.5\}$.

(b) Find the conditional probability $\boldsymbol{P}\{0.3 < X < 0.7 \mid Y > 0.5\}$.

This problem can be solved analytically as well as geometrically.

4.27. Prove the memoryless property of Geometric distribution. That is, if X has Geometric distribution with parameter p, show that

$$\boldsymbol{P}\{X > x + y \mid X > y\} = \boldsymbol{P}\{X > x\}$$

for any integer $x, y \geq 0$.

CHAPTER 5

Computer Simulations and Monte Carlo Methods

5.1 Introduction

Computer simulations refer to a regeneration of a process by writing a suitable computer program and observing its results. *Monte Carlo methods* are those based on computer simulations involving random numbers.

The main purpose of simulations is computing such quantities of interest whose direct computation is complicated, risky, consuming, expensive, or impossible. For example, suppose a complex device or machine is to be built and launched. Before it happens, its performance is simulated allowing experts to evaluate its adequacy and associated risks carefully and safely. For example, one surely prefers to evaluate safety and reliability of a new module of a space station by means of computer simulations rather than during the actual mission.

Monte Carlo methods are mostly used for the computation of probabilities, expected values, and other distribution characteristics. Recall that probability can be defined as *long-run proportion*. With the help of random number generators, computers can actually simulate a *long run*. Then, probability can be estimated by a mere computation of the associated frequency. The longer run is simulated, the more accurate result is obtained. Similarly, one can estimate expectations, variances, and other distribution characteristics from a long run of simulated random variables.

Figure 5.1 *Casino Monte Carlo in Principality of Monaco.*

Monte Carlo methods inherit their name from Europe's most famous *Monte Carlo casino* (Figure 5.1) located in Principality of Monaco since the 1850s. Probability distributions involved in gambling are often complicated, but they can be assessed via simulations. In early times, probabilists generated extra income by estimating vital probabilities and devising optimal gambling strategies.

5.1.1 Applications and examples

Let us briefly look at a few examples, where distributions are rather complicated, and thus, Monte carlo simulation appears simpler than any direct computation. Notice that in these examples, instead of generating the *actual* devices, computer systems, networks, viruses, and so on, we only simulate the associated *random variables*. For the study of probabilities and distributions, this is entirely sufficient.

Example 5.1 (Forecasting). Given just a basic distribution model, it is often very difficult to make reasonably *remote predictions*. Often a one-day development depends on the results obtained during all the previous days. Then prediction for tomorrow may be straightforward whereas computation of a one-month forecast is already problematic.

On the other hand, *simulation* of such a process can be easily performed day by day (or even minute by minute). Based on present results, we simulate the next day. Now we "know it," and thus, we can simulate the day after that, etc. For every time n, we simulate X_{n+1} based on already known X_1, X_2, ..., X_n. Controlling the length of this do-loop, we obtain forecasts for the next days, weeks, or months. Such simulations result, for example, in beautiful animated weather maps that we often see on TV news. They help predict future paths of storms and hurricanes.

Simulation of future failures reflects reliability of devices and systems. Simulation of future stock and commodity prices plays a crucial role in finance, as it allows valuations of options and other financial deals. ◇

Example 5.2 (Percolation). Consider a network of nodes. Some nodes are connected, say, with transmission lines, others are not (mathematicians would call such a network *a graph*). A signal is sent from a certain node. Once a node k receives a signal, it sends it along each of its output lines with some probability p_k. After a certain period of time, one desires to estimate the proportion of nodes that received a signal, the probability for a certain node to receive it, etc.

This general *percolation* model describes the way many phenomena may *spread*. The role of a signal may be played by a computer virus spreading from one computer to another, or by rumors spreading among people, or by fire spreading through a forest, or by a disease spreading between residents.

Technically, simulation of such a network reduces to generating Bernoulli random variables with parameters p_i. Line i transmits if the corresponding generated variable $X_i = 1$. In the end, we simply count the number of nodes that got the signal, or verify whether the given node received it. ◇

Example 5.3 (QUEUING). A queuing system is described by a number of random variables. It involves spontaneous arrivals of jobs, their random waiting time, assignment to servers, and finally, their random service time. In addition, some jobs may exit prematurely, others may not enter the system if it appears full, and also, intensity of the incoming traffic and the number of servers on duty may change during the day.

Designing a queuing system, or a server facility, it is important to evaluate its potential and vital characteristics. This will include the job's average waiting time, the proportion of "unsatisfied customers" (that exit prematurely or cannot enter), the proportion of jobs spending, say, more than an hour in the system, the expected usage of each server, the average number of available (idle) servers at the time when a job arrives. Queuing systems are discussed in detail in Chapter 7; methods of their simulation are in Section 7.6. ◇

Example 5.4 (MARKOV CHAIN MONTE CARLO). There is a modern technique of generating random variables from rather complex, often intractable distributions, as long as *conditional distributions* have a reasonably simple form. In semiconductor industry, for example, the joint distribution of good and defective chips on a produced wafer has a rather complicated correlation structure. As a result, it can only be written explicitly for rather simplified artificial models. On the other hand, the quality of each chip is predictable based on the quality of the surrounding, neighboring chips. Given its neighborhood, conditional probability for a chip to fail can be written, and thus, its quality can be simulated by generating a corresponding Bernoulli random variable with $X_i = 1$ indicating a failure.

According to the *Markov chain Monte Carlo* (MCMC) methodology, a long sequence of random variables is generated from conditional distributions. A wisely designed MCMC will then produce random variables that have the desired *unconditional* distribution, no matter how complex it is. ◇

In all the examples, we saw how different types of phenomena can be computer-simulated. However, one simulation is not enough for estimating probabilities and expectations. After we understand how to program the given phenomenon once, we can embed it in a do-loop and repeat similar simulations a large number of times, generating a *long run*. Since the simulated variables are random, we will generally obtain a number of different *realizations*, from which we calculate probabilities and expectations as long-run frequencies and averages.

5.2 Simulation of random variables

As we see, implementation of Monte Carlo methods reduces to generation of random variables from given distributions. Hence, it remains to design algorithms for generating random variables and vectors with given desired distributions.

Statistics software packages like SAS, Splus, SPSS, Minitab, and others have built-in procedures for the generation of random variables from the most common discrete and continuous distributions. Similar tools are found in recent versions of MATLAB, Microsoft Excel, and some libraries.

Majority of computer languages have a *random number generator* that returns only *Uniformly distributed* independent random variables. This section discusses general methods of transforming Uniform random variables, obtained by a standard random number generator, into variables with desired distributions.

5.2.1 Random number generators

Obtaining a good random variable is not a simple task. How do we know that it is “truly random” and does not have any undesired patterns? For example, quality random number generation is so important in coding and password creation that people design special tests to verify the “randomness” of generated numbers.

In the fields sensitive to good random numbers, their generation is typically related to an accurate measurement of some physical variable, for example, the computer time or noise. Certain transformations of this variable will generally give the desired result.

More often than not, a *pseudo-random number generator* is utilized. This is nothing but a very long list of numbers. A user specifies a *random number seed* that points to the location from which this list will be read. It should be noted that if the same computer code is executed with the same seed, then the same random numbers get generated, leading to identical results. Often each seed is generated within the system, which of course improves the quality of random numbers.

Instead of a computer, a **table of random numbers** is often used for small-size studies. For example, we can use Table A1 in Appendix.

Can a table adequately replace a computer random number generator? There are at least two issues that one should be aware of.

1. Generating a random number seed. A rule of thumb suggests to close your eyes and put your finger "somewhere" on the table. Start reading random numbers from this point, either horizontally, or vertically, or even diagonally, etc. Either way, this method is not pattern-free, and it does not guarantee "perfect" randomness.
2. Using the same table more than once. Most likely, we cannot find a new table of random numbers for each Monte Carlo study. As soon as we use the same table again, we may no longer consider our results independent of each other. Should we be concerned about it? It depends on the situation. For totally unrelated projects, this should not be a problem.

One way or another, a random number generator or a table of random numbers delivers to us Uniform random variables $U_1, U_2, \ldots \in (0,1)$. The next three subsections show how to transform them into a random variable (vector) X with the given desired distribution $F(x)$.

NOTATION | $U;\, U_1, U_2, \ldots$ = generated Uniform(0,1) random variables |

5.2.2 Discrete methods

At this point, we have obtained one or several independent Uniform(0,1) random variables by means of a random number generator or a table of random numbers. Variables from certain simple distributions can be immediately generated from this.

Example 5.5 (BERNOULLI). First, simulate a Bernoulli trial with probability of success p. For a Standard Uniform variable U, define

$$X = \begin{cases} 1 & \text{if} \quad U < p \\ 0 & \text{if} \quad U \geq p \end{cases}$$

We call it “a success” if $X = 1$ and “a failure” if $X = 0$. Using the Uniform distribution of U, we find that

$$\boldsymbol{P}\{\text{ success }\} = \boldsymbol{P}\{U < p\} = p.$$

Thus, we have generated a Bernoulli trial, and X has Bernoulli distribution with the desired probability p.

A MATLAB code for this scheme is

```
U  =  rand;
X  =  (U<p)
```

The value of p should be specified prior to this program. ◇

Example 5.6 (Binomial). Once we know how to generate Bernoulli variables, we can obtain a Binomial variable as a sum of n independent Bernoulli. For this purpose, we start with n Uniform random numbers, for example:

```
n  =  20; p = 0.68;
U  =  rand(n,1);
X  =  sum(U<p)
```

◇

Example 5.7 (Geometric). A while-loop of Bernoulli trials will generate a Geometric random variable, literarily according to its definition on p. 65. We run the loop of trials until the first success occurs. Variable X counts the number of failures:

```
X = 1;            % Need at least one trial
while rand > p;   % Continue while there are failures
   X = X+1;
end;              % Stop at the first success
X
```

The percentage sign (%) is used to separate MATLAB statements from comments. ◇

Example 5.8 (Negative Binomial). Just as in Example 5.6, once we know how to generate a Geometric variable, we can generate a number of them and obtain a Negative Binomial(k, p) variable as a sum of k independent Geometric(p) variables. ◇

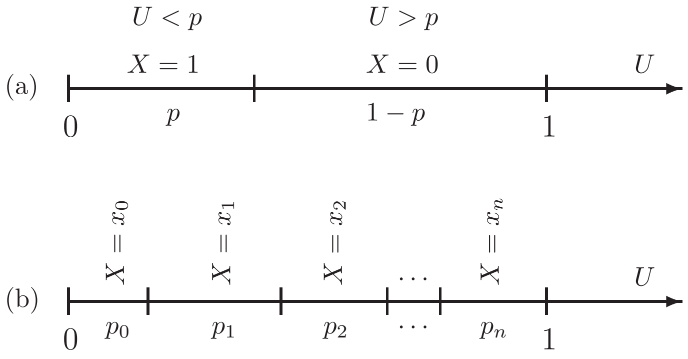


Figure 5.2 *Generating discrete random variables. The value of X is determined by the region where the generated value of U belongs.*

Arbitrary discrete distribution

Example 5.5 can be extended to any arbitrary discrete distribution. In this Example, knowing that the random number generator returns a number between 0 and 1, we divided the interval $[0, 1]$ into two parts, p and $(1 - p)$ in length. Then we determined the value of X according to the part where the generated value of U fell, as in Figure 5.2a.

Now consider an arbitrary discrete random variable X that takes values x_0, x_1, ... with probabilities p_0, p_1, ...,

$$p_i = \boldsymbol{P}\left\{X = x_i\right\}, \qquad \sum_i p_i = 1.$$

The scheme similar to Example 5.5 can be applied as follows.

Algorithm 5.1 *(Generating discrete variables)*

1. Divide the interval $[0, 1]$ into subintervals as shown in Figure 5.2b,

$$\begin{array}{rcl} A_0 & = & [0,\ p_0) \\ A_1 & = & [p_0,\ p_0 + p_1) \\ A_2 & = & [p_0 + p_1,\ p_0 + p_1 + p_2) \\ & & \text{etc.} \end{array}$$

   Subinterval A_i will have length p_i, there may be a finite or infinite number of them, according to possible values of X.

2. Obtain a Standard Uniform random variable from a random number generator or a table of random numbers.

3. If U belongs to A_i, let $X = x_i$.

From the Uniform distribution, it follows again that

$$\boldsymbol{P}\{X = x_i\} = \boldsymbol{P}\{U \in A_i\} = p_i.$$

Hence, the generated variable X has the desired distribution.

Notice that contrary to Examples 5.6 and 5.8, this algorithm is economic as it requires only one Uniform random number for each generated variable.

Values x_i can be written in any order, but they have to correspond to their probabilities p_i.

Example 5.9 (Poisson). Let us use Algorithm 5.1 to generate a Poisson variable with parameter $\lambda = 5$.

Recall from Section 3.4.5 that a Poisson variable takes values $x_0 = 0$, $x_1 = 1$, $x_2 = 2$, ... with probabilities

$$p_i = \boldsymbol{P}\{X = x_i\} = e^{-\lambda}\frac{\lambda^i}{i!} \text{ for } i = 0, 1, 2, \ldots$$

Following the algorithm, we generate a Uniform random number U and find the set A_i containing U, so that

$$p_0 + \ldots + p_{i-1} \leq U < p_0 + \ldots + p_{i-1} + p_i,$$

or, in terms of the cumulative distribution function,

$$F(i-1) \leq U < F(i).$$

This can be done in MATLAB by means of a while-loop:

```
lambda  =  5;                    % Parameter
U       =  rand;                 % Generated Uniform variable
i       =  0;                    % Initial value
F       =  exp(-lambda);         % Initial value, F(0)
while (U >= F);                  % The loop ends when U < F(i)
    F = F + exp(-lambda) * lambda^i/gamma(i+1);
    i = i + 1;
end;
X=i
```

◇

Remark: Statistical toolbox of MATLAB has built-in tools for generating random variables from certain given distributions. Generating Binomial, Geometric, Poisson, and some other variables can be done by a command `random(name,parameters)`, where "name" points to the desired family of distributions. In this chapter, we intentionally do not use these tools. The codes given here can be used as flow-charts and be directly translated into other languages.

5.2.3 Inverse transform method

As Algorithm 5.1 applies to all discrete distributions, we now turn attention to the *generation of continuous random variables*. The method will be based on the following simple yet surprising fact.

Theorem 2 *Let X be a continuous random variable with cdf $F_X(x)$. Define a random variable $U = F_X(X)$. The distribution of U is Uniform(0,1).*

Proof: First, we notice that $0 \leq F(x) \leq 1$ for all x, therefore, values of U lie in $[0, 1]$. Second, for any $u \in [0, 1]$, find the cdf of U,

$$\begin{array}{rcll} F_U(u) & = & \boldsymbol{P}\left\{U \leq u\right\} & \\ & = & \boldsymbol{P}\left\{F_X(X) \leq u\right\} & \\ & = & \boldsymbol{P}\left\{X \leq F_X^{-1}(u)\right\} & \text{(solve the inequality for } X) \\ & = & F_X(F_X^{-1}(u)) & \text{(by definition of cdf)} \\ & = & u & (F_X \text{ and } F_X^{-1} \text{ cancel}) \end{array}$$

We see that U has cdf $F_U(u) = u$ and density $f_U(u) = F_U'(u) = 1$ for $0 \leq u \leq 1$. This is the Uniform(0,1) density, hence, U has Uniform(0,1) distribution. □

Regardless of the initial distribution of X, it becomes Uniform(0,1), once X is substituted into its own cumulative distribution function!

Arbitrary continuous distribution

In order to generate variable X with the given continuous cdf F, let us revert the formula $U = F(X)$. Then X can be obtained from a generated Standard Uniform variable U as $X = F^{-1}(U)$.

Algorithm 5.2 *(Generating continuous variables)*

1. Obtain a Standard Uniform random variable from a random number generator.
2. Compute $X = F^{-1}(U)$. In other words, solve the equation $F(X) = U$ for X.

Does it follow directly from Theorem 2 that X has the desired distribution? Details are left in Exercise 5.11.

Example 5.10 (Exponential). How shall we generate an Exponential variable with parameter λ? According to Algorithm 5.2, we start by generating a Uniform random variable U. Then we recall that the Exponential cdf is $F(x) = 1 - e^{-\lambda x}$ and solve the equation

$$1 - e^{-\lambda X} = U.$$

The result is

$$X = -\frac{1}{\lambda} \ln(1-U). \tag{5.1}$$

Can this formula be simplified? By any rule of algebra, the answer is "no." On the other hand, $(1-U)$ has the same distribution as U, Standard Uniform. Therefore, we can replace U by $(1-U)$, and variable

$$X_1 = -\frac{1}{\lambda} \ln(U), \tag{5.2}$$

although different from X, will also have the desired Exponential(λ) distribution.

Let us check if our generated variables, X and X_1, have a suitable range of values. We know that $0 < U < 1$ with probability 1, hence $\ln(U)$ and $\ln(1-U)$ are negative numbers, so that both X and X_1 are positive, as Exponential variables should be. Just a check. ◇

Example 5.11 (Gamma). Gamma cdf has a complicated integral form (4.9), not mentioning F^{-1}, needed for Algorithm 5.2. There are numerical methods of solving the equation $F(X) = U$, but alternatively, for any integer α, a Gamma variable can be generated as a sum of α independent Exponential variables:

```
  X = sum( -1/lambda * log(rand(alpha,1)) )
```

◇

Discrete distributions revisited

Algorithm 5.2 is not directly applicable to discrete distributions because the inverse function F^{-1} does not exist in the discrete case. In other words, the key equation $F(X) = U$ may have either infinitely many roots or no roots at all (see Figure 3.1 on p. 43).

Moreover, a discrete variable X has a finite or countable range of possible values, and so does $F(X)$. The probability that U coincidentally equals one of these values is 0. We conclude that the equation $F(X) = U$ has no roots with probability 1.

Let us modify the scheme in order to accommodate discrete variables. Instead of solving $F(x) = U$ exactly, which is impossible, we solve it approximately by finding x, the smallest possible value of X such that $F(x) > U$.

The resulting algorithm is described below.

Algorithm 5.3 *(Generating discrete variables, revisited)*

1. Obtain a Standard Uniform random variable from a random number generator or a table of random numbers.
2. Compute $X = \min\{x \in S \text{ such that } F(x) > U\}$, where S is a set of possible values of X.

Algorithms 5.1 and 5.3 are equivalent if values x_i are arranged in their increasing order in Algorithm 5.1. Details are left in Exercise 5.12.

Example 5.12 (Geometric revisited). Applying Algorithm 5.3 to Geometric cdf

$$F(x) = 1 - (1-p)^x,$$

we need to find the smallest integer x satisfying the inequality (solve it for x)

$$1-(1-p)^x > U, \qquad (1-p)^x < 1-U, \qquad x\ln(1-p) < \ln(1-U),$$

$$x > \frac{\ln(1-U)}{\ln(1-p)} \quad \text{(changing the sign because } \ln(1-p)<0\text{)}$$

The smallest such integer is the ceiling* of $\ln(1-U)/\ln(1-p)$, so that

$$X = \left\lceil \frac{\ln(1-U)}{\ln(1-p)} \right\rceil. \tag{5.3}$$

◇

Exponential-Geometric relation

The formula (5.3) appears very similar to (5.1). With $\lambda = -\ln(1-p)$, our generated Geometric variable is just the ceiling of an Exponential variable! In other words, the ceiling of an Exponential variable has Geometric distribution.

This is not a coincidence. In a sense, Exponential distribution is a continuous analogue of a Geometric distribution. Recall that an Exponential variable describes the time until the next "rare event" whereas a Geometric variable is the time (the number of Bernoulli trials) until the next success. Also, both distributions have a memoryless property distinguishing them from other distributions (see (4.7) and Exercise 4.27).

5.2.4 Rejection method

Besides a good computer, one thing is needed to generate continuous variables using the inverse transform method of Section 5.2.3. It is a reasonably *simple form of the cdf* $F(x)$ that allows direct computation of $X = F^{-1}(U)$.

* Remark: Ceiling function $\lceil x \rceil$ is defined as the smallest integer that is no less than x.

When $F(x)$ has a complicated form but a density $f(x)$ is available, random variables with this density can be generated by *rejection method*.

Consider a point (X, Y) chosen at random from under the graph of density $f(x)$, as shown in Figure 5.3. What is the distribution of X, its first coordinate?

Theorem 3 *Let a pair (X, Y) have Uniform distribution over the region*

$$A = \{(x, y) \mid 0 \le y \le f(x)\}$$

for some density function f. Then f is the density of X.

PROOF: Uniform distribution has a constant density. In case of a pair (X, Y), this density equals 1 because

$$\iint_A 1\,dy\,dx = \int_x \left(\int_{y=0}^{f(x)} 1\,dy \right) dx = \int_x f(x)\,dx = 1$$

because $f(x)$ is a density.

The marginal density of X obtains from the joint density by integration,

$$f_X(x) = \int f_{(X,Y)}(x, y)\,dy = \int_{y=0}^{f(x)} 1\,dy = f(x).$$

□

It remains to generate a Uniform point in the region A. For this purpose, we select a *bounding box* around the graph of $f(x)$ as shown in Figure 5.3, generate a random point in this rectangle and reject all the points not belonging to A. The remaining points are Uniformly distributed in A.

Algorithm 5.4 *(Rejection method)*

1. Find such numbers a, b, and c that $0 \le f(x) \le c$ for $a \le x \le b$. The bounding box stretches along the x-axis from a to b and along the y-axis from 0 to c.
2. Obtain Standard Uniform random variables U and V from a random number generator or a table of random numbers.
3. Define $X = a + (b - a)U$ and $Y = cV$. Then X has Uniform(a, b) distribution, Y is Uniform$(0, c)$, and the point (X, Y) is Uniformly distributed in the bounding box.
4. If $Y > f(X)$, reject the point and return to step 2. If $Y \le f(X)$, then X is the desired random variable having the density $f(x)$.

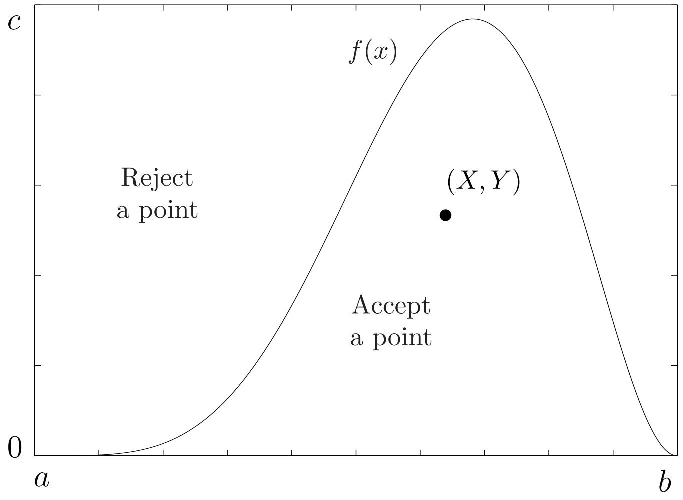


Figure 5.3 *Rejection method. A pair (X, Y) should have a Uniform distribution in the region under the graph of $f(x)$.*

Example 5.13 (Rejection method for Beta distribution). Beta distribution has density

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \text{ for } 0 \le x \le 1.$$

For $\alpha = 5.5$ and $\beta = 3.1$, this density is graphed in Figure 5.3. It never exceeds 2.5, therefore we choose the bounding box with $a = 0$, $b = 1$, and $c = 2.5$. MATLAB commands for generating a Beta(α, β) random variable are:

```
alpha=5.5; beta=3.1; a=0; b=1; c=2.5;
X=0; Y=c;             % Initial values
while Y > gamma(alpha+beta)/gamma(alpha)/gamma(beta)...
          * X.^(alpha-1) .* (1-X).^(beta-1);
   U=rand; V=rand; X=a+(b-a)*U; Y=c*V;
end; X
```

In this code, dot operations (preceded by a dot ".") are applied separately to each element of a matrix (instead of operations with whole matrices).

A histogram of 10,000 random variables generated (in a loop) by this algorithm is shown in Figure 5.4. Compare its shape with the graph of the density $f(x)$ in Figure 5.3. It is not as smooth as $f(x)$ because of randomness and a finite sample size, but if the simulation algorithm is designed properly, the shapes should be similar. ◇

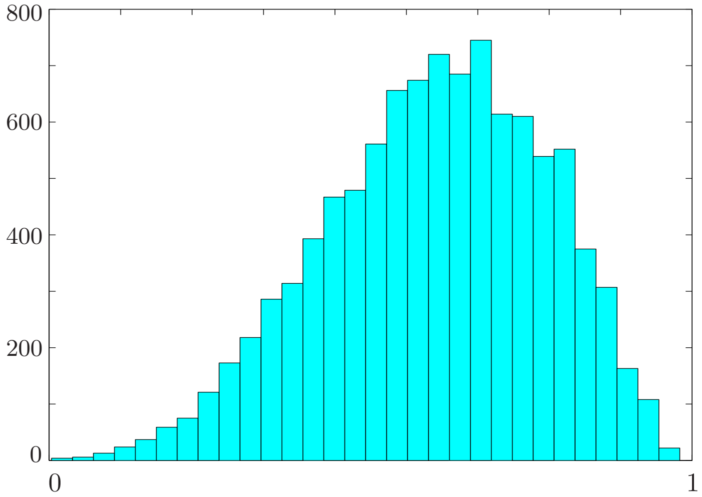


Figure 5.4 *A histogram of Beta random variables generated by rejection method. Compare with Figure 5.3.*

Generation of random vectors

Along the same lines, we can use rejection method to generate *random vectors* having desired joint densities. A bounding box now becomes a multi-dimensional cube, where we generate a Uniformly distributed random point $(X_1, X_2, \ldots, X_n, Y)$, accepted only if $Y \leq f(X_1, \ldots, X_n)$. Then, the generated vector $(X_1, \ldots, X_n)$ has the desired joint density

$$\begin{aligned} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) &= \int f_{X_1,\ldots,X_n,Y}(x_1, \ldots, x_n, y)dy \\ &= \int_0^{f(x_1,\ldots,x_n)} 1\, dy = f(x_1, \ldots, x_n). \end{aligned}$$

The inverse transform method can also be used to generate random vectors. Let us generate X_1 with the marginal cdf F_{X_1}. Observe its value, x_1, and generate X_2 from the *conditional cdf*

$$F_{X_2|X_1}(x_2|x_1) = \begin{cases} \dfrac{\displaystyle\sum_{x \leq x_2} P_{X_1,X_2}(x_1, x)}{P_{X_1}(x_1)} & \text{in the discrete case} \\ \dfrac{\displaystyle\int_{-\infty}^{x_2} f_{X_1,X_2}(x_1, x)dx}{f_{X_1}(x_1)} & \text{in the continuous case} \end{cases}$$

Then generate X_3 from the cdf $F_{X_3|X_1,X_2}(x_3|x_1, x_2)$, etc.

The inverse transform method requires tractable expressions for the cdf F and conditional cdf's. Rejection method requires only the joint density, however, it needs more random number generations due to rejections. This is not a significant disadvantage for relatively small Monte Carlo studies or relatively tight bounding boxes.

5.2.5 Special methods

Because of a complex form of their cdf $F(x)$, some random variables are generated by methods other than inverse transforms.

Poisson distribution

An alternative way of generating a Poisson variable is to count the number of "rare events" occurring during one unit of time.

Recall from Sections 3.4.5 and 4.2.2 that the number of "rare events" has Poisson distribution whereas the time between any two events is Exponential. We generate Exponential times between events according to (5.2), before their sum exceeds 1. The number of generated times equals the number of events, and this is our generated Poisson variable. In short,

1. Obtain Uniform variables $U_1, U_2, \ldots$ from a random number generator.
2. Compute Exponential variables $T_i = -\frac{1}{\lambda}\ln(U_i)$.
3. Let $X = \max\{k : T_1 + \ldots + T_k \leq 1\}$.

This algorithm can be simplified if we notice that

$$T_1 + \ldots + T_k = -\frac{1}{\lambda}\left(\ln(U_1) + \ldots + \ln(U_k)\right) = -\frac{1}{\lambda}\ln(U_1 \cdot \ldots \cdot U_k)$$

and therefore, X can be computed as

$$\begin{aligned} X &= \max\left\{k : -\frac{1}{\lambda}\ln(U_1 \cdot \ldots \cdot U_k) \leq 1\right\} \\ &= \max\left\{k : U_1 \cdot \ldots \cdot U_k \geq e^{-\lambda}\right\}. \end{aligned} \tag{5.4}$$

This formula for generating Poisson variables is rather popular.

Normal distribution

Box-Muller transformation

$$\begin{cases} Z_1 &= \sqrt{-2\ln(U_1)}\cos(2\pi U_2) \\ Z_2 &= \sqrt{-2\ln(U_2)}\sin(2\pi U_2) \end{cases}$$

converts a pair of generated Standard Uniform variables (U_1, U_2) into a pair of independent Standard Normal variables (Z_1, Z_2). This is a rather economic algorithm. To see why it works, solve Exercise 5.13.

5.3 Solving problems by Monte Carlo methods

We have learned how to generate random variables from any given distribution. Once we know how to generate one variable, we can put the algorithm in a loop and generate many variables, a "long run." Then, we shall estimate probabilities by the long-run proportions, expectations by the long-run averages, etc.

5.3.1 Estimating probabilities

This section discusses the most basic and most typical application of Monte Carlo methodology. Keeping in mind that probabilities are long-run proportions, we generate a long run of experiments and compute the proportion of times when our event occurred.

For a random variable X, the probability $p = \boldsymbol{P}\{X \in A\}$ is estimated by

$$\hat{p} = \widehat{\boldsymbol{P}}\{X \in A\} = \frac{\text{number of } X_1, \ldots, X_N \in A}{N},$$

where N is the size of Monte Carlo experiment, $X_1, \ldots, X_N$ are generated random variables with the same distribution as X, and a "hat" means the estimator. The latter is a very common and standard notation:

$$\underline{\text{Notation}} \quad \left| \; \hat{\theta} \quad = \quad \text{estimator of an unknown quantity } \theta \; \right|$$

How accurate is this method? To answer this question, compute $\mathbf{E}(\hat{p})$ and $\mathrm{Std}(\hat{p})$. Since the number of $X_1, \ldots, X_N \in A$ has Binomial(N, p) distribution with expectation (Np) and variance $Np(1-p)$, we obtain

$$\begin{aligned} \mathbf{E}(\hat{p}) &= \frac{1}{N}(Np) = p, \text{ and} \\ \mathrm{Std}(\hat{p}) &= \frac{1}{N}\sqrt{Np(1-p)} = \sqrt{\frac{p(1-p)}{N}}. \end{aligned}$$

The first result, $\mathbf{E}(\hat{p}) = p$ shows that our Monte Carlo estimator of p is *unbiased*, so that over a long run, it will on the average return the desired quantity p.

The second result, $\mathrm{Std}(\hat{p}) = \sqrt{p(1-p)/N}$, indicates that the standard deviation of our estimator $\hat{p}$ decreases with N at the rate of $1/\sqrt{N}$. Larger Monte Carlo experiments produce more accurate results. A 100-fold increase in the number of generated variables reduces the standard deviation (therefore, increasing accuracy) by a factor of 10.

Accuracy of a Monte Carlo study

In practice, how does it help to know the standard deviation of $\hat{p}$?

First, we can assess the accuracy of our results. For large N, we use Normal approximation of the Binomial distribution of $N\hat{p}$, as in (4.20) on p. 103. According to it,

$$\frac{N\hat{p} - Np}{\sqrt{Np(1-p)}} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{N}}} \approx \text{ Normal}(0,1),$$

therefore,

$$\boldsymbol{P}\left\{|\hat{p} - p| > \varepsilon\right\} = \boldsymbol{P}\left\{\frac{|\hat{p} - p|}{\sqrt{\frac{p(1-p)}{N}}} > \frac{\varepsilon}{\sqrt{\frac{p(1-p)}{N}}}\right\} \approx 2\Phi\left(-\frac{\varepsilon}{\sqrt{\frac{p(1-p)}{N}}}\right). \quad (5.5)$$

We have computed probabilities of this type in Section 4.2.4.

Second, we can design a Monte Carlo study that attains desired accuracy. That is, we can choose some small ε and α and conduct a Monte Carlo study of such a size N that will guarantee an error not exceeding ε with high probability $(1-\alpha)$. In other words, we can find such N that

$$\boldsymbol{P}\left\{|\hat{p} - p| > \varepsilon\right\} \leq \alpha. \quad (5.6)$$

If we knew the value of p, we could have equated the right-hand side of (5.5) to α and could have solved the resulting equation for N. This would have shown how many Monte Carlo simulations are needed in order to achieve the desired accuracy with the desired probability. However, p is unknown (if p is known, why do we need a Monte Carlo study to estimate it?). Then, we have two possibilities:

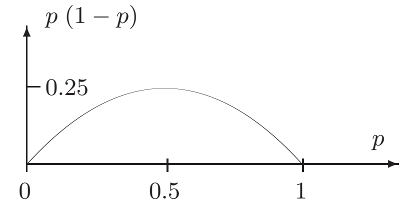


Figure 5.5 *Function* $p\,(1-p)$ attains its maximum at $p = 0.5$.

1. Use an "intelligent guess" (preliminary estimate) of p, if it is available.
2. Bound $p\,(1-p)$ by its largest possible value (see Figure 5.5),

$$p\,(1-p) \leq 0.25 \quad \text{for} \quad 0 \leq p \leq 1.$$

In the first case, if p^* is an "intelligent guess" of p, then we solve the inequality

$$2\Phi\left(-\frac{\varepsilon}{\sqrt{\frac{p^*(1-p^*)}{N}}}\right) \leq \alpha$$

in terms of N. In the second case, we solve

$$2\Phi\left(-2\varepsilon\sqrt{N}\right) \le \alpha.$$

Solutions of these inequalities give us the following rule.

Size of a Monte Carlo study

In order to guarantee that $\boldsymbol{P}\left\{|\hat{p}-p| > \varepsilon\right\} \le \alpha$, one needs to simulate

$$N \ge p^*(1-p^*)\left(\frac{z_{\alpha/2}}{\varepsilon}\right)^2$$

random variables, where p^* is a preliminary estimator of p, or

$$N \ge 0.25\left(\frac{z_{\alpha/2}}{\varepsilon}\right)^2$$

random variables, if such estimator is not available.

Recall (from Example 4.12) that $z_\alpha = \Phi^{-1}(1-\alpha)$ is such a value of a Standard Normal variable Z that can be exceeded with probability α. It is obtained from Table A4. The area under the Normal curve to the right of z_α equals α (Figure 5.6).

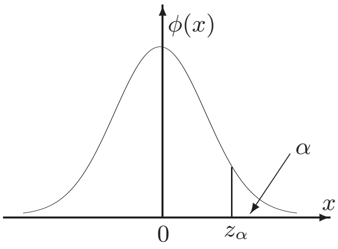


Figure 5.6 *Critical value* z_α.

If this formula returns N that is too small for the Normal approximation, we can use Chebyshev's inequality (3.8) on p. 58. We then obtain that

$$N \ge \frac{p^*(1-p^*)}{\alpha\varepsilon^2}, \tag{5.7}$$

if p^* is available, and

$$N \ge \frac{1}{4\alpha\varepsilon^2}, \tag{5.8}$$

otherwise, satisfy the desired condition (5.6) (Exercise 5.14).

Example 5.14 (Shared computer). The following problem does not have a simple analytic solution (by hand), therefore we use the Monte Carlo method.

A supercomputer is shared by 250 independent subscribers. Each day, each subscriber uses the facility with probability 0.3. The number of tasks sent by each active user has Geometric distribution with parameter 0.15, and each task takes a Gamma$(10, 3)$ distributed computer time (in minutes). Tasks are processed consecutively. What is the probability that all the tasks will be processed, that is, the total requested computer time is less than 24 hours? Estimate this probability, attaining the margin of error ±0.01 with probability 0.99.

Solution. The total requested time $T = T_1 + \ldots + T_X$ consists of times T_i requested by X active users. The number of active users X is Binomial(C, p), and each of them sends a Geometric(q) number of tasks Y_i. Thus, each $T_i = T_{i,1} + \ldots + T_{i,Y_i}$ is the sum of Y_i Gamma(β, λ) random variables. Overall, the distribution of T is rather complicated, although a Monte Carlo solution is simple.

It is hard to come up with an "intelligent guess" of the probability of interest $\boldsymbol{P}\{T < 24 \text{ hrs}\}$. To attain the required accuracy ($\alpha = 0.01$, $\varepsilon = 0.01$), we use

$$N \geq 0.25 \left(\frac{z_{\alpha/2}}{\varepsilon}\right)^2 = 0.25 \left(\frac{2.575}{0.01}\right)^2 = 16,577$$

simulations, where $z_{\alpha/2} = z_{0.005} = 2.575$ is found from Table A4. The obtained number N is large enough to justify the Normal approximation.

Next, we generate the number of active users, the number of tasks, and the time required by each task, repeat this procedure N times, and compute the proportion of times when the total time appears less than 24 hrs = 1440 min. The following MATLAB program can be used:

```
 N=16577;                    % number of simulations
 C=250; p=0.3; q=0.15;       % parameters
 alph=10; lambda=3;          % parameters of Gamma distribution
 TotalTime=zeros(N,1);       % save total time for each run
 for k=1:N;                  % do-loop of N runs
     X=sum( rand(C,1)<p );   % the number of active users
          % is generated as a sum of Bernoulli(p) variables
     Y=ceil( log(1-rand(X,1))/log(1-q) );
          % the number of tasks for each of X active users
          % is generated according to formula (5.3)
     TotalTasks=sum(Y);      % total daily number of tasks
     T=sum( -1/lambda * log(rand(alph,TotalTasks)) );
          % requested times are generated as in Example 5.11
     TotalTime(k)=sum(T);    % total time from the k-th run
 end;                        % end of simulations
P_est=mean(TotalTime<1440)  % proportion of runs with the total
        % time less than 24 hours; this is our estimator of p.
```

The resulting estimated probability should be close to 0.17. It is a rather low probability that all the tasks will be processed. ◇

5.3.2 Estimating means and standard deviations

Estimation of means, standard deviations, and other distribution characteristics is based on the same principle. We generate a Monte Carlo sequence of random variables $X_1, \ldots, X_N$ and compute the necessary long-run averages. The mean $\mathbf{E}(X)$ is estimated by the average (denoted by $\bar{X}$ and pronounced "X-bar")

$$\bar{X} = \frac{1}{N}\left(X_1 + \ldots + X_N\right).$$

If the distribution of $X_1, \ldots, X_N$ has mean μ and standard deviation σ,

$$\begin{aligned} \mathbf{E}(\bar{X}) &= \frac{1}{N}\left(\mathbf{E}X_1 + \ldots + \mathbf{E}X_N\right) = \frac{1}{N}(N\mu) = \mu, \text{ and} && (5.9) \\ \mathrm{Var}(\bar{X}) &= \frac{1}{N^2}\left(\mathrm{Var}X_1 + \ldots + \mathrm{Var}X_N\right) = \frac{1}{N^2}(N\sigma^2) = \frac{\sigma^2}{N}. && (5.10) \end{aligned}$$

From (5.9), we conclude that the estimator $\bar{X}$ is *unbiased* for estimating μ. Its standard deviation $\mathrm{Std}(\bar{X}) = \sigma/\sqrt{N}$ decreases like $1/\sqrt{N}$. For large N, we can use the Central Limit Theorem again to assess accuracy of our results and design a Monte Carlo study that attains the desired accuracy (the latter is possible if we have a "guess" about σ).

Variance of a random variable is defined as the expectation of $(X - \mu)^2$. Similarly, we estimate it by a long-run average, replacing the unknown value of μ by its Monte Carlo estimator. The resulting estimator is usually denoted by s^2,

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2.$$

Remark: Does the coefficient $\frac{1}{N-1}$ seem surprising? Proper averaging should divide the sum of N numbers by N, however, only $(N-1)$ in the denominator guarantees that $\mathbf{E}(s^2) = \sigma^2$, that is, s^2 is unbiased for σ^2. This fact will be proved in Section **??**. (Sometimes, coefficients $\frac{1}{N}$ and even $\frac{1}{N+1}$ are also used when estimating σ^2. The resulting estimates are not unbiased, but they have other attractive properties. For large N, the differences between all three estimators are negligible.)

Example 5.15 (Shared computer, continued). In Example 5.14, we generated the total requested computer time for each of N days and computed the proportion of days when this time was less than 24 hours. Further, we can estimate *the expectation of requested time* by computing the average,

```
ExpectedTime = mean(TotalTime);
```

The standard deviation of daily requested time will then be estimated by

```
 StandardDeviation = std(TotalTime);
```

The mean requested time appears to be 1667 minutes, exceeding the 24-hour period by 227 minutes, and the standard deviation is around 243 minutes. It is clear now why all the tasks get processed with such a low probability as estimated in Example 5.14. ◇

5.3.3 Forecasting

Forecasting the future is often an attractive but uneasy task. In order to predict what happens in t days, one usually has to study events occurring every day between now and t days from now. It is often found that tomorrow events depend on today, yesterday, etc. In such common situations, exact computation of probabilities may be long and difficult (although we'll learn some methods in Chapters 6 and 7), however, Monte Carlo forecasts are feasible.

The currently known information can be used to generate random variables for tomorrow. Then, they can be considered "known" information, and we can generate random variables for the day after tomorrow, and so on, until day t.

The rest is similar to the previous sections. We generate N Monte Carlo runs for the next t days and estimate probabilities, means, and standard deviations of day t variables by long-run averaging.

In addition to forecasts for a specific day t, we can predict *how long* a certain process will last, *when* a certain event will occur, and *how many* events will occur.

Example 5.16 (New software release). Here is a stochastic model for the number of errors found in a new software release. Every day, software developers find a random number of errors and correct them. The number of errors X_t found on day t has Poisson(λ_t) distribution whose parameter is the lowest number of errors found during the previous 3 days,

$$\lambda_t = \min\{X_{t-1}, X_{t-2}, X_{t-3}\}.$$

Suppose that during the first three days, software developers found 28, 22, and 18 errors.

(a) Predict the time it will take to find all the errors.
(b) Estimate the probability that some errors will remain undetected after 21 days.
(c) Predict the total number of errors in this new release.

Solution. Let us generate $N = 1000$ Monte Carlo runs. In each run, we generate the number of errors found on each day until one day this number equals 0. According to our model, no more errors can be found after that, and we conclude that all errors have been detected.

The following MATLAB code uses method (5.4) as follows,

```
N=1000;                  % number of Monte Carlo runs
Time=zeros(N,1);         % the day the last error is found
Nerrors=zeros(N,1);      % total number of errors
for k=1:N;               % do-loop of N runs
 Last3=[28,22,18];       % errors found during last 3 days
 DE=sum(Last3);          % detected errors so far
 T=0; X=min(Last3);      % T = # days, X = # errors on day T
 while X>0;              % while-loop until no errors are found
   lambda=min(Last3);    % parameter λ for day T
   U=rand; X=0;          % initial values
        while U>=exp(-lambda);
            U=U*rand; X=X+1;
        end;             % according to (5.4), X is Poisson(λ)
   T=T+1; DE=DE+X;       % update after day T
   Last3=[Last3(2:3), X];
 end;                    % the loop ends when X=0 on day T
 Time(k)=T-1;
 Nerrors(k)=DE;
end;
```

Now we estimate the expected time it takes to detect all the errors by `mean(Time)`, the probability of errors remaining after 21 days by `mean(Time>21)`, and the expected total number of errors by `mean(Nerrors)`. This Monte Carlo study should predict the expected time of about 16.75 days to detect all the errors, the probability 0.22 that errors remain after 21 days, and about 222 errors overall. ◇

5.3.4 Estimating lengths, areas, and volumes. Monte Carlo integration

Lengths

A Standard Uniform variable U has density $f_U(u) = 1$ for $0 \le u \le 1$. Hence, U belongs to a set $A \subset [0,1]$ with probability

$$\boldsymbol{P}\{U \in A\} = \int_A 1\, du = \text{ length of } A. \tag{5.11}$$

Monte Carlo methods can be used to estimate the probability in the left-hand side. At the same time, we estimate the right-hand side of (5.11), the length of A. Generate a long run of Standard Uniform random variables $U_1,\ U_2, \ldots, U_n$ and estimate the length of A by the proportion of U_i that fall into A.

What if set A does not lie within a unit interval? Well, we can always choose a suitable system of coordinates, with a suitable origin and scale, to make the interval $[0,1]$ cover the given bounded set as long as the latter is bounded. Alternatively, we can cover A with some interval $[a,b]$ and generate *non-standard* Uniform variables on $[a,b]$, in which case the estimated probability $\boldsymbol{P}\left\{U \in A\right\}$ should be multiplied by $(b-a)$.

Areas and volumes

Computing lengths rarely represents a serious problem, therefore, one would rarely use Monte Carlo methods for this purpose. However, the situation is different with estimating *areas* and *volumes*.

The method described for estimating lengths is directly translated into higher dimensions. Two independent Standard Uniform variables U and V have a *joint* density $f_{U,V}(u,v) = 1$ for $0 \le u, v \le 1$, hence,

$$\boldsymbol{P}\left\{(U,V) \in B\right\} = \iint_B 1\,du\,dv = \text{ area of } B$$

for any two-dimensional set B that lies within a unit square $[0,1]\times[0,1]$. Thus, the area of B can be estimated as a long-run frequency of vectors (U_i, V_i) that belong to set B.

Algorithm 5.5 *(Estimating areas)*

1. Obtain a large *even* number of independent Standard Uniform variables from a random number generator, call them $U_1, \ldots, U_n;\ V_1, \ldots, V_n$.
2. Count the number of pairs (U_i, V_i) such that the point with coordinates (U_i, V_i) belongs to set B. Call this number N_B.
3. Estimate the area of B by N_B/n.

Similarly, a long run of Standard Uniform triples (U_i, V_i, W_i) allows to estimate the volume of any three-dimensional set.

Areas of arbitrary regions with unknown boundaries

Notice that in order to estimate lengths, areas, and volumes by Monte Carlo methods, *knowing exact boundaries is not necessary*. To apply Algorithm 5.5, it is sufficient to determine which points belong to the given set.

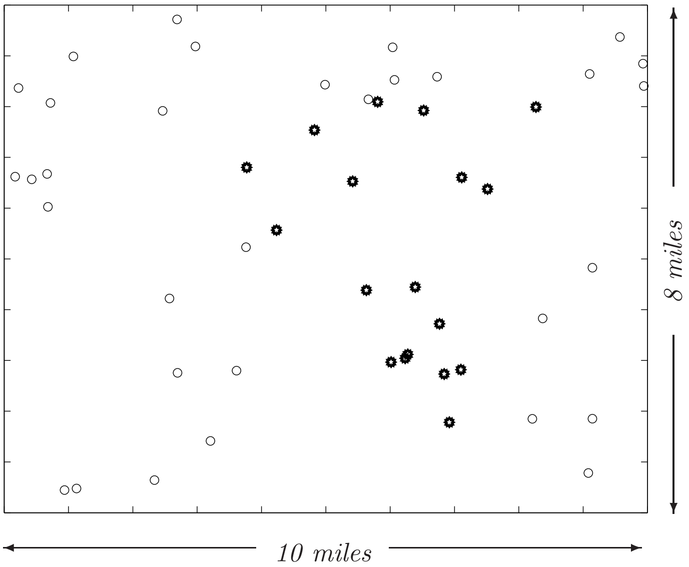


Figure 5.7 *Monte Carlo area estimation. Fifty sites are randomly selected; the marked sites belong to the exposed region.*

Also, the sampling region does not have to be a square. With different scales along the axes, random points may be generated on a rectangle or even a more complicated figure. One way to generate a random point in a region of arbitrary shape is to draw a larger square or rectangle around it and generate uniformly distributed coordinates until the corresponding point belongs to the region. In fact, by estimating the probability for a random point to fall into the area of interest, we estimate the proportion this area makes of the entire sampling region.

Example 5.17 (Size of the exposed region). Consider the following situation. An emergency is reported at a nuclear power plant. It is necessary to assess the size of the region exposed to radioactivity. Boundaries of the region cannot be determined, however, the level of radioactivity can be measured at any given location.

Algorithm 5.5 can be applied as follows. A rectangle of 10 by 8 miles is chosen that is likely to cover the exposed area. Pairs of Uniform random numbers (U_i, V_i) are generated, and the level of radioactivity is measured at all the obtained random locations. The area is then estimated as the proportion of measurements above the normal level, multiplied by the area of the sampling rectangle. In Figure 5.7, radioactivity is measured at 50 random sites, and it is found above the normal level at 18 locations. The exposed area is then estimated as

$$\frac{18}{50}(80 \text{ sq. miles}) = \underline{28.8 \text{ sq. miles.}}$$

◇

Notice that different scales on different axes in Figure 5.7 allowed to represent a rectangle as a unit square. Alternatively, we could have generated points with Uniform(0,10) x-coordinate and Uniform(0,8) y-coordinate.

Monte Carlo integration

We have seen how Monte Carlo methods can be used to estimate lengths, areas, and volumes. We can extend the method to *definite integrals* estimating areas below or above the graphs of corresponding functions. A MATLAB code for estimating an integral

$$\mathcal{I} = \int_0^1 g(x)dx$$

is

```
N  =  1000;                         % Number of simulations
U  =  rand(N,1); V = rand(N,1);     % Points in the bounding box
I  =  mean( V < g(U) )              % Estimator of integral I
```

Expression `V < g(U)` returns an $N \times 1$ vector. Each component equals 1 if the inequality holds for the given pair (U_i, V_i), and 0 otherwise. The average of these 0s and 1s, calculated by `mean`, is the proportion of 1s, and this is the Monte Carlo estimator of integral I.

Remark: This code assumes that the function g is already defined by the user in a file named "g.m." Otherwise, its full expression should be written in place of `g(U)`.

Remark: We also assumed that $0 \le x \le 1$ and $0 \le g(x) \le 1$. If not, we transform U and V into non-standard Uniform variables $X = a + (b - a)U$ and $Y = cV$, as in the rejection method, then the obtained integral should be multiplied by the box area $c(b - a)$.

Accuracy of results

So far, we estimated lengths, areas, volumes, and integrals by long-run proportions, the method described in Section 5.3.1. As we noted there, our estimates are unbiased, and their standard deviation is

$$\text{Std}\left(\hat{\mathcal{I}}\right) = \sqrt{\frac{\mathcal{I}(1-\mathcal{I})}{N}}, \tag{5.12}$$

where $\mathcal{I}$ is the actual quantity of interest.

Turns out, there are Monte Carlo integration methods that can beat this rate. Next, we derive an unbiased area estimator with a lower standard deviation. Also, it will not be restricted to an interval $[0, 1]$ or even $[a, b]$.

Improved Monte Carlo integration method

First, we notice that a definite integral

$$\mathcal{I} = \int_a^b g(x)dx = \frac{1}{b-a}\int_a^b (b-a)g(x)dx = \mathbf{E}\left\{(b-a)g(X)\right\}$$

equals the expectation of $(b-a)g(X)$ for a Uniform(a, b) variable X. Hence, instead of using proportions, we can estimate $\mathcal{I}$ by averaging $(b-a)g(X_i)$ for some large number of Uniform(a, b) variables $X_1, \ldots, X_N$.

Furthermore, with a proper adjustment, we can use *any continuous distribution* in place of Uniform(a, b). To do this, we choose some density $f(x)$ and write $\mathcal{I}$ as

$$\mathcal{I} = \int_a^b g(x)dx = \int_a^b \frac{g(x)}{f(x)} f(x)\,dx = \mathbf{E}\left(\frac{g(X)}{f(X)}\right),$$

where X has density $f(x)$. It remains to generate a long run of variables $X_1, \ldots, X_N$ with this density and compute the average of $g(X_i)/f(X_i)$.

In particular, we are no longer limited to a finite interval $[a, b]$. For example, by choosing $f(x)$ to be a Standard Normal density, we can perform Monte Carlo integration from $a = -\infty$ to $b = +\infty$:

```
N = 1000;                             % Number of simulations
Z = randn(N,1);                       % Standard Normal variables
f = 1/sqrt(2*Pi) * exp(-Z.^2/2);      % Standard Normal density
Iest = mean( g(Z)./f(Z) )             % Estimator of ∫_{-∞}^{∞} g(x) dx
```

Remark: recall that "dot" operations .^ and ./ stand for *pointwise* power and division. Without a dot, matrix operations would be done instead.

Accuracy of the improved method

As we already know from (5.9), using long-run averaging returns an *unbiased* estimator $\hat{\mathcal{I}}$, hence $\mathbf{E}(\hat{\mathcal{I}}) = \mathcal{I}$. We also know from (5.10) that

$$\mathrm{Std}(\hat{\mathcal{I}}) = \frac{\sigma}{\sqrt{N}},$$

where σ is the standard deviation of a random variable $R = g(X)/f(X)$. Hence, the estimator $\hat{\mathcal{I}}$ is more reliable if σ is small and N is large. Small σ

can be obtained by choosing a density $f(x)$ that is approximately proportional to $g(x)$ making R nearly a constant with $\operatorname{Std}(R) \approx 0$. However, generating variables from such a density will usually be just as difficult as computing the integral $\mathcal{I}$.

In fact, using a simple Standard Uniform distribution of X, we already obtain a lower standard deviation than in (5.12). Indeed, suppose that $0 \leq g(x) \leq 1$ for $0 \leq x \leq 1$. For a Uniform$(0,1)$ variable X, we have $f(X) = 1$ so that

$$\sigma^2 = \operatorname{Var} R = \operatorname{Var} g(X) = \mathbf{E}g^2(X) - \mathbf{E}^2 g(X) = \int_0^1 g^2(x)dx - \mathcal{I}^2 \leq \mathcal{I} - \mathcal{I}^2,$$

because $g^2 \leq g$ for $0 \leq g \leq 1$. We conclude that for this method,

$$\operatorname{Std}(\hat{\mathcal{I}}) \leq \sqrt{\frac{\mathcal{I} - \mathcal{I}^2}{N}} = \sqrt{\frac{\mathcal{I}(1-\mathcal{I})}{N}}.$$

Comparing with (5.12), we see that with the same number of simulations N, the latter method gives more accurate results. We can also say that to attain the same desired accuracy, the second method requires fewer simulations.

This can be extended to an arbitrary interval $[a,b]$ and any function $g \in [0,c]$ (Exercise 5.16).

Summary and conclusions

Monte Carlo methods are effectively used for estimating probabilities, expectations and other distribution characteristics in complex situations when computing these quantities by hand is difficult. According to Monte Carlo methodology, we generate a long sequence of random variables $X_1, \ldots, X_N$ from the distribution of interest and estimate probabilities by long-run proportions, expectation by long-run averages, etc. We extend these methods to the estimation of areas, volumes, and integrals. Similar techniques are used for forecasting.

All the discussed methods produce unbiased results. Standard deviations of the proposed estimators decrease at the rate of $1/\sqrt{N}$. Knowing the standard deviation enables us to assess the accuracy of obtained estimates, and also, to design a Monte Carlo study that attains the desired accuracy with the desired high probability.

In this chapter, we learned the inverse transform method of generating random variables, rejection method, discrete method, and some special methods. Monte Carlo simulation of more advanced models will be considered in chapters 6 and 7, where we simulate stochastic processes, Markov chains, and queuing systems.

Questions and exercises

5.1. Derive a formula and explain how to generate a random variable with the density $f(x) = (1.5)\sqrt{x}$ for $0 < x < 1$ if your random number generator produces a standard Uniform random variable U. Use the inverse transform method. Compute this variable if $U = 0.001$.

5.2. Let U be a Standard Uniform random variable. Show all the steps required to generate

(a) an Exponential random variable with the parameter $\lambda = 2.5$;

(b) a Bernoulli random variable with the probability of success 0.77;

(c) a Binomial random variable with parameters $n = 15$ and $p = 0.4$;

(d) a discrete random variable with the distribution $P(x)$, where $P(0) = 0.2$, $P(2) = 0.4$, $P(7) = 0.3$, $P(11) = 0.1$;

(e) a continuous random variable with the density $f(x) = 3x^2$, $0 < x < 1$;

(f) a continuous random variable with the density $f(x) = 1.5x^2$, $-1 < x < 1$;

(g) a continuous random variable with the density $f(x) = \frac{1}{12}\sqrt[3]{x}$, $0 \leq x \leq 8$.

If a computer generates U and the result is $U = 0.3972$, compute the variables generated in (a)–(g).

5.3. Explain how one can generate a random variable X that has a pdf

$$f(x) = \begin{cases} \frac{1}{2}(1+x) & \text{if } -1 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases},$$

given a computer-generated Standard Uniform variable U. Generate X using Table A1.

5.4. To evaluate the system parameters, one uses Monte Carlo methodology and simulates one of its vital characteristics X, a continuous random variable with the density

$$f(x) = \begin{cases} \frac{2}{9}(1+x) & \text{if } -1 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

Explain how one can generate X, given a computer-generated Standard Uniform variable U. If $U = 0.2396$, compute X.

5.5. Give an expression that transforms a standard Uniform variable U into a variable X with the following density,

$$f(x) = \frac{1}{3}x^2, \quad -1 < x < 2.$$

Compute X if a computer returns a value of $U = 0.8$.

5.6. Two mechanics are changing oil filters for the arrived customers. The service time has an Exponential distribution with the parameter $\lambda = 5$ for the first mechanic, and $\lambda = 20$ for the second mechanic. Since the second mechanic works faster, he is serving 4 times more customers than his partner. Therefore, when you arrive to have your oil filter changed, your probability of being served by the faster mechanic is $4/5$. Let X be your service time. Explain how to generate the random variable X.

5.7. Explain how to estimate the following probabilities.

(a) $\boldsymbol{P}\{X > Y\}$, where X and Y are independent Poisson random variables with parameters 3 and 5, respectively.

(b) The probability of a royal-flush (a ten, a jack, a queen, a king, and an ace of the same suit) in poker, if 5 cards are selected at random from a deck of 52 cards.

(c) The probability that it will take more than 35 minutes to have your oil filter changed in Exercise 5.6.

(d) With probability 0.95, we need to estimate each of the listed probabilities with a margin of error not exceeding 0.005. What should be the size of our Monte Carlo study in each case?

(e) (Computer mini-project) Conduct a Monte Carlo study and estimate probabilities (a)-(c) with an error not exceeding 0.005 with probability 0.95.

5.8. (Computer mini-project) Area of a unit circle equals π. Cover a circle with a 2 by 2 square and follow Algorithm 5.5 to estimate number π based on 100, 1,000, and 10,000 random numbers. Compare results with the exact value $\pi = 3.14159265358...$ and comment on precision.

5.9. (Computer project) Forty computers are connected in a network. One computer becomes infected with a virus. Every day, this virus spreads from any infected computer to any uninfected computer with probability 0.2. Also, every day, a computer technician takes 5 infected computers (or all infected computers, if their number is less than 5) and removes the virus from them. Estimate:

(a) the expected time it takes to remove the virus from the system;

(b) the probability that each computer gets infected at least once;

(c) the expected number of computers that get infected.

5.10. (Computer project) A forest consists of 1,000 trees forming a perfect 50×20 rectangle, Figure 5.8. The northwestern (top-left) corner tree catches fire. Wind blows from the west, therefore, the probability that any tree catches fire from its burning left neighbor is 0.8. The probability to catch fire from a tree immediately to the right, above, or below is only 0.3.

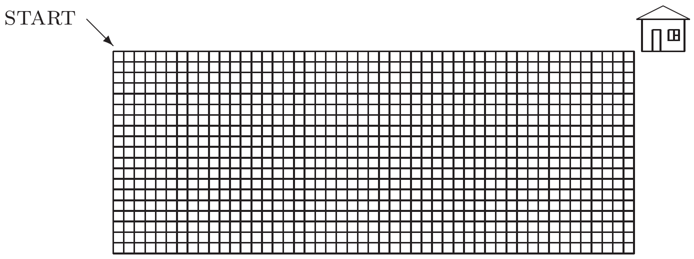


Figure 5.8 *The northwestern corner of the forest catches fire (Exercise 5.9).*

(a) Conduct a Monte Carlo study to estimate the probability that more than 30% of the forest will eventually be burning. With probability 0.95, your answer should differ from the true value by no more than 0.005.

(b) Based on the same study, predict the total number of affected trees X.

(c) Estimate $\text{Std}(X)$ and comment on the accuracy of your estimator of X.

(d) What is the probability that the actual number of affected trees differs from your estimator by more than 25 trees?

(e) A wooden house is located in the northeastern corner of the forest. Would you advise the owner that her house is in real danger?

5.11. Let F be a continuous cdf, and U be a Standard Uniform random variable. Show that random variable X obtained from U via a formula $X = F^{-1}(U)$ has cdf F.

5.12. Show that Algorithms 5.1 and 5.3 produce the same discrete variable X if they are based on the same value of a Uniform variable U and values x_i in Algorithm 5.1 are arranged in the increasing order, $x_0 < x_1 < x_2 < \ldots$.

5.13. Prove that the Box-Muller transformation

$$\begin{array}{rcl} Z_1 & = & \sqrt{-2\ln(U_1)}\cos(2\pi U_2) \\ Z_2 & = & \sqrt{-2\ln(U_1)}\sin(2\pi U_2) \end{array}$$

returns a pair of independent Standard Normal random variables by showing equality $\boldsymbol{P}\{Z_1 \leq a \ \cap \ Z_2 \leq b\} = \Phi(a)\Phi(b)$.

5.14. A Monte Carlo study is being designed to estimate some probability p by a long-run proportion $\hat{p}$. Suppose that condition (5.6) can be satisfied by some small N that does not allow the Normal approximation of the distribution of $\hat{p}$. Use Chebyshev's inequality instead. Show that the size N computed according to (5.7) or (5.8) satisfies the required condition (5.6).

5.15. A random variable is generated from a density $f(x)$ by rejection method. For this purpose, a bounding box with $a \le x \le b$ and $0 \le y \le c$ is selected. Show that this method requires a Geometric(p) number of generated pairs of Standard Uniform random variables and find p.

5.16. We estimate an integral

$$\mathcal{I} = \int_a^b g(x)dx$$

for arbitrary a and b and a function $0 \le g(x) \le c$ using both Monte Carlo integration methods introduced in this chapter. First, we generate N pairs of Uniform variables (U_i, V_i), $a \le U_i \le b$, $0 \le V_i \le c$, and estimate $\mathcal{I}$ by the properly rescaled proportion of pairs with $V_i \le g(U_i)$. Second, we generate only U_i and estimate $\mathcal{I}$ by the average value of $(b-a)g(U_i)$. Show that the second method produces more accurate results.

CHAPTER 6

Stochastic Processes

Let us summarize what we have already accomplished. Our ultimate goal was to learn to make decisions under uncertainty. We introduced a language of *probability* in Chapter 2 and learned how to measure uncertainty. Then, through Chapters 3–5, we studied *random variables*, *random vectors*, and their *distributions*. Have we learned enough to describe a situation involving uncertainty and be able to make good decisions?

Let us look around. If you say "Freeze!" and everything freezes for a moment, the situation will be completely described by random variables. However, the real world is dynamic. Many variables develop and change in real time: air temperatures, stock prices, interest rates, football scores, popularity of politicians, and also, the CPU usage, the speed of internet connection, the number of concurrent users, the number of running processes, available memory, and so on.

We now start the discussion of *stochastic processes* which are random variables that also evolve and change in time.

6.1 Definitions and Classifications

DEFINITION 6.1

A **stochastic process** is a random variable that also depends on time. It is therefore a function of two arguments, $X(t,\omega)$, where:

- $t \in \mathcal{T}$ is time, with $\mathcal{T}$ being a set of possible times, usually $[0,\infty)$, $(-\infty,\infty)$, $\{0,1,2,\ldots\}$, or $\{\ldots,-2,-1,0,1,2,\ldots\}$;
- $\omega \in \Omega$, as before, is an outcome of an experiment, with Ω being the whole sample space.

Values of $X(t,\omega)$ are called *states*.

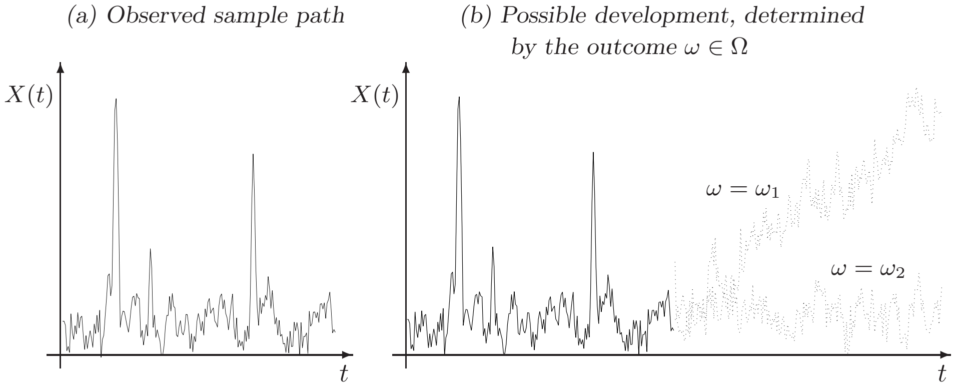


Figure 6.1 *Sample paths of CPU usage.*

At any fixed time t, we see a random variable $X_t(\omega)$, a function of a random outcome. On the other hand, if we fix ω, we obtain a function of time $X_\omega(t)$. This function is called a *realization*, a *sample path*, or a *trajectory* of a process $X(t, \omega)$.

Example 6.1 (CPU USAGE). Looking at the past usage of the central processing unit (CPU), we see a realization of this process until the current time (Figure 6.1a). However, the future behavior of the process is unclear. Depending on which outcome ω will actually take place, the process can develop differently (Figure 6.1b).

Remark: You can observe a similar stochastic process on your personal computer. In the latest versions of Windows, Ctrl-Alt-Del, pressed simultaneously and followed by "Windows Task Manager," will show the real-time sample path of CPU usage under the tab "Performance." ◇

Depending on possible values of T and X, stochastic processes are classified as follows.

DEFINITION 6.2

Stochastic process $X(t, \omega)$ is **discrete-state** if variable $X_t(\omega)$ is discrete for each time t, and it is a **continuous-state** if $X_t(\omega)$ is continuous.

DEFINITION 6.3

> Stochastic process $X(t, \omega)$ is a **discrete-time process** if the set of times $\mathcal{T}$ is discrete, that is, it consists of separate, isolated points. It is a **continuous-time process** if $\mathcal{T}$ is a connected, possibly unbounded interval.

Example 6.2. The CPU usage process, in percents, is continuous-state and continuous-time, as we can see in Figure 6.1a. ◇

Example 6.3. The *actual* air temperature $X(t, \omega)$ at time t is a continuous-time, continuous-state stochastic process. Indeed, it changes continuously and never jumps from one value to another. However, the temperature $Y(t, \omega)$ reported on a radio every 10 minutes is a discrete-time process. Moreover, since it is usually rounded to the nearest degree, it is also a discrete-state process. ◇

Example 6.4. In a printer shop, let $X(n, \omega)$ be the amount of time required to print the n-th job. This is a discrete-time, continuous-state stochastic process, because $n = 1, 2, 3, \ldots$, and $X \in (0, \infty)$.

Let $Y(n, \omega)$ be the number of pages of the n-th printing job. Now, $Y = 1, 2, 3, \ldots$ is discrete, therefore, this process is discrete-time and discrete-state. ◇

From now on, we shall not write ω as an argument of $X(t, \omega)$. Just keep in mind that behavior of a stochastic process depends on chance, just as we did with random variables and random vectors.

Another important class of stochastic processes is defined in the next section.

6.2 Markov processes and Markov chains

DEFINITION 6.4

> Stochastic process $X(t)$ is **Markov** if for any $t_1 < \ldots < t_n < t$ and any sets $A; A_1, \ldots, A_n$
>
> $$\begin{aligned} \boldsymbol{P}\left\{X(t) \in A \mid X(t_1) \in A_1, \ldots, X(t_n) \in A_n\right\} \\ = \boldsymbol{P}\left\{X(t) \in A \mid X(t_n) \in A_n\right\}. \end{aligned} \qquad (6.1)$$

Let us look at the equation (6.1). It means that the conditional distribution of $X(t)$ is the same under two different conditions,

(1) given observations of the process X at several moments in the past;

(2) given only *the latest* observation of X.

If a process is Markov, then its future behavior is the same under conditions (1) and (2). In other words, knowing the present, we get no information from the past that can be used to predict the future,

$$\boldsymbol{P}\{\text{ future } \mid \text{ past, present }\} = \boldsymbol{P}\{\text{ future } \mid \text{ present }\}$$

Then, for the future development of a Markov process, only its present state is important, and it does not matter *how* the process arrived to this state.

Some processes satisfy the Markov property, and some don't.

Example 6.5 (Internet connections). Let $X(t)$ be the total number of internet connections registered by some internet service provider by the time t. Typically, people connect to the internet at random times, regardless of how many connections have already been made. Therefore, the number of connections in a minute will only depend on the current number. For example, if 999 connections have been registered by 10 o'clock, then their total number will exceed 1000 during the next minute regardless of *when* and *how* these 999 connections were made in the past. This process is *Markov*. ◇

Example 6.6 (Stock prices). Let $Y(t)$ be the value of some stock or some market index at time t. If we know $Y(t)$, do we also want to know $Y(t-1)$ in order to predict $Y(t+1)$? One may argue that if $Y(t-1) < Y(t)$, then the market is rising, therefore, $Y(t+1)$ is likely (but not certain) to exceed $Y(t)$. On the other hand, if $Y(t-1) > Y(t)$, we may conclude that the market is falling and may expect $Y(t+1) < Y(t)$. It looks like knowing the past *in addition* to the present did help us to predict the future. Then, this process is *not Markov*. ◇

Due to a well-developed theory and a number of simple techniques available for Markov processes, it is important to know whether the process is Markov or not. The idea of Markov dependence was proposed and developed by *Andrei Markov* (1856–1922) who was a student of P. Chebyshev (p. 58) at St. Petersburg University in Russia.

6.2.1 Markov chains

DEFINITION 6.5

> A **Markov chain** is a discrete-time, discrete-state Markov stochastic process.

Introduce a few convenient simplifications. The time is discrete, so let us define the time set $\mathcal{T} = \{0, 1, 2, \ldots\}$. We can then look at a Markov chain as a random sequence

$$\{X(0), X(1), X(2), \ldots\}.$$

The state set is also discrete, so let us enumerate the states as $1, 2, \ldots, n$. Sometimes we'll start enumeration from state 0, and sometimes we'll deal with a Markov chain with infinitely many (discrete) states, then we'll have $n = \infty$.

The *Markov* property means that only the value of $X(t)$ matters for predicting $X(t+1)$, so the conditional probability

$$\begin{aligned} p_{ij}(t) &= \boldsymbol{P}\left\{X(t+1) = j \mid X(t) = i\right\} \qquad (6.2) \\ &= \boldsymbol{P}\left\{X(t+1) = j \mid X(t) = i,\ X(t-1) = h,\ X(t-2) = g, \ldots\right\} \end{aligned}$$

depends on i, j, and t only and equals the probability for the Markov chain X to make a *transition* from state i to state j at time t.

DEFINITION 6.6

> Probability $p_{ij}(t)$ in (6.2) is called a **transition probability**. Probability
>
> $$p_{ij}^{(h)}(t) = \boldsymbol{P}\left\{X(t+h) = j \mid X(t) = i\right\}$$
>
> of moving from state i to state j after h transitions is an **h-step transition probability**.

DEFINITION 6.7

> A Markov chain is **homogeneous** if all its transition probabilities are independent of t. Being homogeneous means that transition from i to j has the same probability at any time. Then $p_{ij}(t) = p_{ij}$ and $p_{ij}^{(h)}(t) = p_{ij}^{(h)}$.

Characteristics of a Markov chain

What do we need to know to describe a Markov chain?

By the Markov property, each next state should be predicted from the previous state only. Therefore, it is sufficient to know the distribution of its initial state $X(0)$ and the mechanism of transitions from one state to another.

The distribution of a Markov chain is completely determined by the initial distribution P_0 and one-step transition probabilities p_{ij}. Here P_0 is the probability mass function of X_0,

$$P_0(x) = \boldsymbol{P}\{X(0) = x\} \text{ for } x \in \{1, 2, \ldots, n\}$$

Based on this data, we would like to compute:

- *h-step transition probabilities $p_{ij}^{(h)}$;*
- *P_h, the distribution of states at time h, which is our forecast for $X(h)$;*
- *the limit of $p_{ij}^{(h)}$ and P_h as $h \to \infty$.*

Indeed, when making forecasts for many transitions ahead, computations will become rather lengthy, and thus, it will be more efficient to take the limit.

NOTATION	p_{ij}	$=$	$P\{X(t+1) = j \mid X(t) = i\}$, transition probability
	$p_{ij}^{(h)}$	$=$	$P\{X(t+h) = j \mid X(t) = i\}$, h-step transition probability
	$P_t(x)$	$=$	$P\{X(t) = x\}$, distribution of $X(t)$, distribution of states at time t
	$P_0(x)$	$=$	$P\{X(0) = x\}$, initial distribution

Example 6.7 (WEATHER FORECASTS). In some town, each day is either sunny or rainy. A sunny day is followed by another sunny day with probability 0.7, whereas a rainy day is followed by a sunny day with probability 0.4.

It rains on Monday. Make forecasts for Tuesday, Wednesday, and Thursday.

Solution. Weather conditions in this problem represent a homogeneous Markov chain with 2 states: state 1 = "sunny" and state 2 = "rainy." Transition probabilities are:

$$p_{11} = 0.7,\ p_{12} = 0.3,\ p_{21} = 0.4,\ p_{22} = 0.6,$$

where p_{12} and p_{22} were computed by the complement rule.

If it rains on Monday, then Tuesday is sunny with probability $p_{21} = 0.4$ (making a transition from a rainy to a sunny day), and Tuesday is rainy with probability $p_{22} = 0.6$. We can predict a 60% chance of rain.

Wednesday forecast requires 2-step transition probabilities, making one transition from Monday ($X(0)$) to Tuesday ($X(1)$) and another one from Tuesday to Wednesday ($X(2)$). We'll have to condition on weather conditions on Tuesday and use the Law of Total Probability from p. 32,

$$\begin{aligned} p_{21}^{(2)} &= \boldsymbol{P}\left\{ \text{ Wednesday is sunny } \mid \text{ Monday is rainy } \right\} \\ &= \sum_{i=1}^{2} \boldsymbol{P}\left\{X(1)=i \mid X(0)=2\right\} \boldsymbol{P}\left\{X(2)=1 \mid X(1)=i\right\} \\ &= \boldsymbol{P}\left\{X(1)=1 \mid X(0)=2\right\} \boldsymbol{P}\left\{X(2)=1 \mid X(1)=1\right\} \\ &\quad + \boldsymbol{P}\left\{X(1)=2 \mid X(0)=2\right\} \boldsymbol{P}\left\{X(2)=1 \mid X(1)=2\right\} \\ &= p_{21}p_{11} + p_{22}p_{21} = (0.4)(0.7) + (0.6)(0.4) = 0.52. \end{aligned}$$

By the Complement Rule, $p_{22}^{(2)} = 0.48$, and thus, we predict a 52% chance of sun and a 48% chance of rain on Wednesday.

For the Thursday forecast, we need to compute 3-step transition probabilities $p_{ij}^{(3)}$ because it takes 3 transitions to move from Monday to Thursday. We have to use the Law of Total Probability conditioning on *both* Tuesday and Wednesday. For example, going from rainy Monday to sunny Thursday means going from rainy Monday to either rainy or sunny Tuesday, then to either rainy or sunny Wednesday, and finally, to sunny Thursday,

$$p_{21}^{(3)} = \sum_{i=1}^{2}\sum_{j=1}^{2} p_{2i}p_{ij}p_{j1}.$$

This corresponds to a sequence of states $2 \to i \to j \to 1$. However, we have already computed 2-step transition probabilities $p_{21}^{(2)}$ and $p_{22}^{(2)}$, describing transition from Monday to Wednesday. It remains to add one transition to Thursday, hence,

$$p_{21}^{(3)} = p_{21}^{(2)}p_{11} + p_{22}^{(2)}p_{21} = (0.52)(0.7) + (0.48)(0.4) = 0.556.$$

So, we predict a 55.6% chance of sun on Thursday and a 44.4% chance of rain. ◇

The following *transition diagram* (Figure 6.2) reflects the behavior of this Markov chain. Arrows represent all possible one-step transitions, along with the corresponding probabilities. Check this diagram against the transition probabilities stated in Example 6.7. To obtain, say, a 3-step transition probability $p_{21}^{(3)}$, find all 3-arrow paths from state 2 "rainy" to state 1 "sunny." Multiply probabilities along each path and add over all 3-step paths.

Example 6.8 (Weather, continued). Suppose now that it does not rain yet, but meteorologists predict an 80% chance of rain on Monday. How does this affect our forecasts?

In Example 6.7, we have computed forecasts under the condition of rain on Monday. Now, a sunny Monday (state 1) is also possible. Therefore, in addition

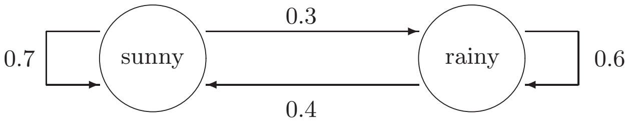


Figure 6.2 *Transition diagram for the Markov chain in Example 6.7.*

to probabilities $p_{2j}^{(h)}$ we also need to compute $p_{1j}^{(h)}$ (say, using the transition diagram, see Figure 6.2),

$$\begin{aligned} p_{11}^{(2)} &= (0.7)(0.7) + (0.3)(0.4) = 0.61, \\ p_{11}^{(3)} &= (0.7)^3 + (0.7)(0.3)(0.4) + (0.3)(0.4)(0.7) + (0.3)(0.6)(0.4) = 0.583. \end{aligned}$$

The initial distribution $P_0(x)$ is given as

$$P_0(1) = \boldsymbol{P}\{\text{ sunny Monday }\} = 0.2, \quad P_0(2) = \boldsymbol{P}\{\text{ rainy Monday }\} = 0.8.$$

Then, for each forecast, we use the Law of Total Probability, conditioning on the weather on Monday,

$$\begin{array}{ll} P_1(1) = P\{X(1) = 1\} = P_0(1)p_{11} + P_0(2)p_{21} = 0.46 & \text{for Tuesday} \\ P_2(1) = P\{X(2) = 1\} = P_0(1)p_{11}^{(2)} + P_0(2)p_{21}^{(2)} = 0.538 & \text{for Wednesday} \\ P_3(1) = P\{X(3) = 1\} = P_0(1)p_{11}^{(3)} + P_0(2)p_{21}^{(3)} = 0.5614 & \text{for Thursday} \end{array}$$

These are probabilities of a sunny day (state 1), respectively, on Tuesday, Wednesday, and Thursday. Then, the chance of rain (state 2) on these days is $P_1(2) = 0.54$, $P_2(2) = 0.462$, and $P_3(2) = 0.4386$. ◇

Noticeably, more remote forecasts require more lengthy computations. For a t-day ahead forecast, we have to account for all t-step paths on diagram Figure 6.2. Or, we use the Law of Total Probability, conditioning on *all* the intermediate states $X(1), X(2), \ldots, X(t-1)$.

To simplify the task, we shall employ *matrices.* If you are not closely familiar with basic matrix operations, refer to Section 11.4 in the Appendix.

6.2.2 Matrix approach

All one-step transition probabilities p_{ij} can be conveniently written in an $n \times n$ *transition probability matrix*

$$\begin{array}{cc|c}
 & & \begin{matrix}\text{To}\\ \text{state:}\end{matrix} \\
P \quad = & \left(\begin{matrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{matrix}\right) & \begin{matrix} 1 \\ 2 \\ \vdots \\ n \end{matrix} \\ \hline
\begin{matrix}\text{From}\\ \text{state:}\end{matrix} & \begin{matrix} 1 & 2 & \cdots & n \end{matrix} &
\end{array}$$

The entry on the intersection of the i-th row and the j-th column is p_{ij}, the transition probability from state i to state j.

From each state, a Markov chain makes a transition to one and only one state. State-destinations are disjoint and exhaustive events, therefore, *each row total equals 1*,

$$p_{i1} + p_{i2} + \ldots + p_{in} = 1. \tag{6.3}$$

We can also say that probabilities $p_{i1}, p_{i2}, \ldots, p_{in}$ form the *conditional distribution* of $X(1)$, given $X(0)$, so they have to add to 1.

In general, this does not hold for column totals. Some states may be "more favorable" than others, then they are visited more often the others, thus their column total will be larger. Matrices with property (6.3) are called *stochastic*.

Similarly, h-step transition probabilities can be written in an *h-step transition probability matrix*

$$P^{(h)} = \left(\begin{matrix} p_{11}^{(h)} & p_{12}^{(h)} & \cdots & p_{1n}^{(h)} \\ p_{21}^{(h)} & p_{22}^{(h)} & \cdots & p_{2n}^{(h)} \\ \vdots & \vdots & \vdots & \vdots \\ p_{n1}^{(h)} & p_{n2}^{(h)} & \cdots & p_{nn}^{(h)} \end{matrix}\right)$$

This matrix is also stochastic because each row represents the conditional distribution of $X(h)$, given $X(0)$ (which is a good way to check our results if we compute $P^{(h)}$ by hand).

Computing k-step transition probabilities

A simple matrix formula connects matrices $P^{(h)}$ and P. Let's start with the 2-step transition probabilities.

By the Law of Total Probability, conditioning and adding over all values of $k = X(1)$,

$$\begin{aligned} p_{ij}^{(2)} &= \boldsymbol{P}\left\{X(2)=j \mid X(0)=i\right\} \\ &= \sum_{k=1}^{n} \boldsymbol{P}\left\{X(1)=k \mid X(0)=i\right\} \boldsymbol{P}\left\{X(2)=j \mid X(1)=k\right\} \\ &= \sum_{k=1}^{n} p_{ik}p_{kj} = (p_{i1},\ldots,p_{in}) \begin{pmatrix} p_{1j} \\ \vdots \\ p_{nj} \end{pmatrix}. \end{aligned}$$

Each probability $p_{ij}^{(2)}$ is computed as a sum of $\boldsymbol{P}\left\{i \to k \to j\right\}$, over all 2-step paths leading from state i to state j (also, see Example 6.7). As a result, each $p_{ij}^{(2)}$ is a product of a the i-th row and the j-th column of matrix P! Hence, the entire 2-step transition probability matrix is

2-step transition probability matrix

$$\boxed{P^{(2)} = P \cdot P = P^2}$$

Further, h-step transition probabilities can be obtained from $(h-1)$-step transition probabilities by conditioning on $X(h-1) = k$,

$$\begin{aligned} p_{ij}^{(h)} &= \boldsymbol{P}\left\{X(h)=j \mid X(0)=i\right\} \\ &= \sum_{k=1}^{n} \boldsymbol{P}\left\{X(h-1)=k \mid X(0)=i\right\} \boldsymbol{P}\left\{X(h)=j \mid X(h-1)=k\right\} \\ &= \sum_{k=1}^{n} p_{ik}^{(h-1)}p_{kj} = \left(p_{i1}^{(h-1)},\ldots,p_{in}^{(h-1)}\right) \begin{pmatrix} p_{1j} \\ \vdots \\ p_{nj} \end{pmatrix}. \end{aligned}$$

This time, we consider all h-step routes from i to j. They consist of all $(h-1)$-step routes from i to some state k, followed by one step from k to j. Now, the result is the i-th row of matrix $P^{(h-1)}$ multiplied by the j-th column of P. Hence, $P^{(h)} = P^{(h-1)} \cdot P$, and we have a general formula

***h*-step transition probability matrix**

$$\boxed{P^{(h)} = \underbrace{P \cdot P \cdot \ldots \cdot P}_{h \text{ times}} = P^h}$$

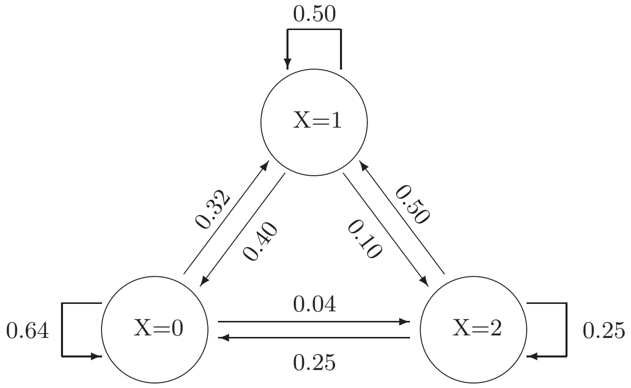


Figure 6.3 *Transition diagram for the Markov chain in Example 6.9.*

Example 6.9 (Shared device). A computer is shared by 2 users who send tasks to a computer remotely and work independently. At any minute, any connected user may disconnect with probability 0.5, and any disconnected user may connect with a new task with probability 0.2. Let $X(t)$ be the number of concurrent users at time t (minutes). This is a Markov chain with 3 states: 0, 1, and 2.

Compute transition probabilities. Suppose $X(0) = 0$, i.e., there are no users at time $t = 0$. Then $X(1)$ is the number of new connections within the next minute. It has Binomial(2,0.2) distribution, therefore,

$$p_{00} = (.8)^2 = .64,\ p_{01} = 2(.2)(.8) = .32,\ p_{02} = (.2)^2 = .04.$$

Next, suppose $X(0) = 1$, i.e., one user is connected, and the other is not. The number of new connections is Binomial(1,0.2), and the number of disconnections is Binomial(1,0.5). Considering all the possibilities, we obtain (verify)

$$p_{10} = (.8)(.5) = .40,\ p_{11} = (.2)(.5) + (.8)(.5) = .50,\ p_{12} = (.2)(.5) = .10.$$

Finally, when $X(0) = 2$, no new users can connect, and the number of disconnections is Binomial(2,0.5), so that

$$p_{20} = .25,\ p_{21} = .50,\ p_{22} = .25.$$

We obtain the following transition probability matrix,

$$P = \begin{pmatrix} .64 & .32 & .04 \\ .40 & .50 & .10 \\ .25 & .50 & .25 \end{pmatrix}.$$

The transition diagram corresponding to this matrix is shown in Figure 6.3. The 2-step transition probability matrix is then computed as

$$P^2 = \begin{pmatrix} .64 & .32 & .04 \\ .40 & .50 & .10 \\ .25 & .50 & .25 \end{pmatrix} \begin{pmatrix} .64 & .32 & .04 \\ .40 & .50 & .10 \\ .25 & .50 & .25 \end{pmatrix} = \begin{pmatrix} .5476 & .3848 & .0676 \\ .4810 & .4280 & .0910 \\ .4225 & .4550 & .1225 \end{pmatrix}.$$

For example, if both users are connected at 10:00, then at 10:02 there will be no users with probability 0.4225, one user with probability 0.4550, and two users with probability 0.1225.

Check that both matrices P and P^2 are stochastic. ◇

Computing the distribution of $X(h)$

The distribution of states after h transitions, or the probability mass function of $X(h)$, can be written in a matrix with $1 \times n$ matrix,

$$P_h = (P_h(1), \cdots, P_h(n)).$$

By the Law of Total Probability, this time conditioning on $X(0) = k$, we compute

$$\begin{aligned} P_h(j) &= \boldsymbol{P}\{X(h) = j\} = \sum_k \boldsymbol{P}\{X(0) = k\}\, \boldsymbol{P}\{X(h) = j \mid X(0) = k\} \\ &= \sum_k P_0(k) p_{kj}^{(h)} = \begin{pmatrix} P_0(1) & \cdots & P_0(n) \end{pmatrix} \begin{pmatrix} p_{1j}^{(h)} \\ \vdots \\ p_{1n}^{(h)} \end{pmatrix}. \end{aligned}$$

Each probability $P_h(j)$ is obtained when the entire row P_0 (initial distribution of X) is multiplied the j-th column of matrix P. Hence,

Distribution of $X(h)$

$$\boxed{P_h = P_0 P^h} \qquad (6.4)$$

Example 6.10 (Shared device, continued). If we know that there are 2 users connected at 10:00, we can write the initial distribution as

$$P_0 = (0, 0, 1).$$

Then the distribution of the number of users at 10:02, after $h = 2$ transitions, is computed as

$$P_2 = P_0 P^2 = (0, 0, 1) \begin{pmatrix} .5476 & .3848 & .0676 \\ .4810 & .4280 & .0910 \\ .4225 & .4550 & .1225 \end{pmatrix} = (.4225, .4550, .1225),$$

just as we concluded in the end of Example 6.9.

Suppose that all states are equally likely at 10:00. How can we compute the probability of 1 connected user at 10:02? Here, the initial distribution is

$$P_0 = (1/3, 1/3, 1/3)\,.$$

The distribution of $X(2)$ is

$$P_2 = P_0 P^2 = (1/3, 1/3, 1/3) \begin{pmatrix} .5476 & .3848 & .0676 \\ .4810 & .4280 & .0910 \\ .4225 & .4550 & .1225 \end{pmatrix} = (\ldots, .4226, \ldots)\,.$$

Thus, $\boldsymbol{P}\{X(2) = 1\} = 0.4226$. We only computed $P_2(1)$. The question did not require computation of the entire distribution P_2. ◇

In practice, knowing the distribution of the number of users at 10:00, one would rarely be interested in the distribution at 10:02. How should one figure the distribution at 11:00, the next day, or the next month?

In other words, how does one compute the distribution of $X(h)$ *for large* h? A direct solution is $P_h = P_0 \cdot P^h$, which can be computed for moderate n and h. For example, the distribution of the number of users at 11:00, $h = 60$ transitions after 10:00, can be computed in MATLAB as

```
P   =   [   .64   .32   .04
            .40   .50   .10
            .25   .50   .25   ];

P0  =   [   1/3   1/3   1/3   ];
h   =  60;
P0*P^h
```

For large h, one would really like to take a limit of P_h as $h \to \infty$ instead of a tedious computation of $P_0 \cdot P^h$. We learn how to do it in the next section.

6.2.3 Steady-state distribution

This section discusses the distribution of states of a Markov chain after a large number of transitions.

DEFINITION 6.8

A collection of limiting probabilities

$$\pi_x = \lim_{h \to \infty} P_h(x)$$

is called a **steady-state distribution** of a Markov chain $X(t)$.

When this limit exists, it can be used as a forecast of the distribution of X after *many* transitions. For a fast system (say, a processor of several gigahertz), it will not take long until a large number of transitions is reached. Then, its distribution of states at virtually any time can be thought of as a steady-state distribution.

Computing the steady-state distribution

When a steady-state distribution π *exists*, it can be computed as follows. We notice that π is a limit of not only P_h but also P_{h+1}. The latter two are related by the formula

$$P_h P = P_0 P^h P = P_0 P^{h+1} = P_{h+1}.$$

Taking the limit of P_h and P_{h+1}, as $h \to \infty$, we obtain

$$\pi P = \pi. \tag{6.5}$$

Then, solving (6.5) for π, we get the steady-state distribution of a Markov chain with a transition probability matrix P.

The system of steady-state equations (6.5) consists of n equations with n unknowns which are probabilities of n states. However, this system is *singular*, it has infinitely many solutions. That is because both sides of the system (6.5) can be multiplied by any constant C, and thus, any multiple of π is also a solution of (6.5).

There is yet one more condition that we have not used. Being a distribution, all the probabilities π_x must add to 1,

$$\pi_1 + \ldots + \pi_n = 1.$$

Only one solution of (6.5) can satisfy this condition, and this is the steady-state distribution.

Steady-state distribution

$$\pi = \lim_{h\to\infty} P_h$$

is computed as a solution of

$$\left\{ \begin{array}{rcl} \pi P & = & \pi \\ \sum\limits_x \pi_x & = & 1 \end{array} \right.$$

Example 6.11 (Weather, continued). In Example 6.7 on p. 148, the transition probability matrix of sunny and rainy days is

$$P = \begin{pmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix}.$$

The steady-state equation for this Markov chain is

$$\pi P = \pi,$$

or

$$(\pi_1,\ \pi_2) = (\pi_1,\ \pi_2) \begin{pmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix} = (0.7\pi_1 + 0.4\pi_2,\ 0.3\pi_1 + 0.6\pi_2).$$

We obtain a system of equations,

$$\left\{ \begin{array}{rcl} 0.7\pi_1 + 0.4\pi_2 & = & \pi_1 \\ 0.3\pi_1 + 0.6\pi_2 & = & \pi_2 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{rcl} 0.4\pi_2 & = & 0.3\pi_1 \\ 0.3\pi_1 & = & 0.4\pi_2 \end{array} \right. \Leftrightarrow \pi_2 = \frac{3}{4}\pi_1.$$

We see that two equations in our system reduced to one. This will always happen: *one equation will follow from the others*, and this is because the system $\pi P = \pi$ is singular. It remains to use the *normalizing equation* $\sum \pi_x = 1$,

$$\pi_1 + \pi_2 = \pi_1 + \frac{3}{4}\pi_1 = \frac{7}{4}\pi_1 = 1,$$

from where

$$\underline{\pi_1 = 4/7} \quad \text{and} \quad \underline{\pi_2 = 3/7}.$$

Hence, in this city, $4/7 \approx 57\%$ of days are sunny, and $3/7 \approx 43\%$ of days are rainy. ◇

Example 6.12 (Shared device, continued). In Example 6.9 on p. 153, the transition probability matrix for the number of concurrent users is

$$P = \begin{pmatrix} .64 & .32 & .04 \\ .40 & .50 & .10 \\ .25 & .50 & .25 \end{pmatrix}.$$

Let us find the steady-state distribution. It will automatically serve as our forecast for the number of users next day or next month. Thus, it will answer the question posed in the end of Section 6.2.2.

The steady-state equations are:

$$\left\{ \begin{array}{rcl} .64\pi_0 + .40\pi_1 + .25\pi_2 & = & \pi_0 \\ .32\pi_0 + .50\pi_1 + .50\pi_2 & = & \pi_1 \\ .04\pi_0 + .10\pi_1 + .25\pi_2 & = & \pi_2 \end{array} \right. ; \left\{ \begin{array}{rcl} -.36\pi_0 + .40\pi_1 + .25\pi_2 & = & 0 \\ .32\pi_0 - .50\pi_1 + .50\pi_2 & = & 0 \\ .04\pi_0 + .10\pi_1 - .75\pi_2 & = & 0 \end{array} \right.$$

Solving this system by method of elimination, we express π_2 from the first equation and substitute into the other equations,

$$\left\{\begin{array}{rcl} \multicolumn{3}{c}{\pi_2 = 1.44\pi_0 - 1.6\pi_1} \\ .32\pi_0 - .50\pi_1 + .50(1.44\pi_0 - 1.6\pi_1) & = & 0 \\ .04\pi_0 + .10\pi_1 - .75(1.44\pi_0 - 1.6\pi_1) & = & 0 \end{array}\right. ; \left\{\begin{array}{rcl} \multicolumn{3}{c}{\pi_2 = 1.44\pi_0 - 1.6\pi_1} \\ 1.04\pi_0 - 1.3\pi_1 & = & 0 \\ -1.04\pi_0 + 1.3\pi_1 & = & 0 \end{array}\right.$$

The last two equations are equivalent. This was anticipated; one equation should always follow from the others, so we are "probably" on a right track.

Express π_1 from the second equation and substitute into the first one,

$$\left\{\begin{array}{rcl} \pi_2 & = & 1.44\pi_0 - 1.6(.8\pi_0) \\ \pi_1 & = & .8\pi_0 \end{array}\right. ; \left\{\begin{array}{rcl} \pi_2 & = & .16\pi_0 \\ \pi_1 & = & .8\pi_0 \end{array}\right.$$

Finally, use the normalizing equation,

$$\pi_0 + \pi_1 + \pi_2 \;=\; \pi_0 + .8\pi_0 + .16\pi_0 \;=\; 1.96\pi_0 \;=\; 1,$$

from where we compute the answer,

$$\left\{\begin{array}{rcccl} \pi_0 & = & 1/1.96 & = & .5102 \\ \pi_1 & = & .8(.5102) & = & .4082 \\ \pi_2 & = & .16(.5102) & = & .0816 \end{array}\right.$$

This is a direct but a little long way of finding the steady-state distribution π. For this particular problem, a shorter solution is offered in Exercise 6.24. ◇

The limit of P^h

Then, what will be h-step transition probabilities for large h? It turns out that matrix $P^{(h)}$ has a limit, as $h \to \infty$, and the limiting matrix has the form

$$\Pi = \lim_{h\to\infty} P^{(h)} = \begin{pmatrix} \pi_1 & \pi_2 & \cdots & \pi_n \\ \pi_1 & \pi_2 & \cdots & \pi_n \\ \vdots & \vdots & \ldots & \vdots \\ \pi_1 & \pi_2 & \cdots & \pi_n \end{pmatrix}.$$

All the rows of the limiting matrix Π are equal, and they consist of the steady-state probabilities π_x!

How can this phenomenon be explained? First, the forecast for some very remote future should not depend on the current state $X(0)$. Say, our weather forecast for the next century should not depend on the weather we have today. Therefore, $p_{ik} = p_{jk}$ for all i, j, k, and this is why all rows of Π coincide.

Second, our forecast, independent of $X(0)$, will just be given in terms of long-term proportions, that is, π_x. Indeed, if your next-year vacation is in August, then the best weather forecast you can get will likely be given as the historical average for this time of the year.

Steady state

What is the steady state of a Markov chain? Suppose the system has reached its steady state, so that the current distribution of states is $P_t = \pi$. A system makes one more transition, and the distribution becomes $P_{t+1} = \pi P$. But $\pi P = \pi$, and thus, $P_t = P_{t+1}$. We see that *in a steady state, transitions do not affect the distribution*. In this sense, it is *steady*.

Existence of a steady-state distribution. Regular Markov chains

As we know from Calculus, there are situations when a limit simply does not exist. Similarly, there are Markov chains with no steady-state distribution.

Example 6.13 (PERIODIC MARKOV CHAIN, NO STEADY-STATE DISTRIBUTION). In chess, a knight can only move to a field of different color, from white to black, and from black to white. Then, the transition probability matrix of the color of its field is

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Computing P^2, P^3, etc., we find that

$$P^{(h)} = P^h = \begin{cases} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} & \text{for all odd } h \\ & \\ \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} & \text{for all even } h \end{cases}$$

Indeed, after any odd number of moves, the knight's field will change its color, and after any even number of moves, it will return to the initial color. Thus, there is no limit of P_h and $P^{(h)}$. ◇

The Markov chain in Example 6.13 is *periodic* with period 2 because $X(t) = X(t+2)$ for all t with probability 1. Periodic Markov chains cannot be regular; from any state, some h-step transitions are possible and some are not, depending on whether or not h is divisible by the period.

There are other situations when steady-state probabilities cannot be found. What we need is a criterion for the existence of a steady-state distribution.

DEFINITION 6.9

A Markov chain is **regular** if

$$p_{ij}^{(h)} > 0$$

for some h and all i, j. That is, for some h, matrix $P^{(h)}$ has only non-zero entries, and h-step transitions from any state to any state are possible.

Any regular Markov chain has a steady-state distribution.

Example 6.14. Markov chains in Examples 6.7 and 6.9 are *regular* because all transitions are possible for $h = 1$ already, and matrix P does not contain any zeros. ◇

Example 6.15. A Markov chain with transition probability matrix

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0.99 & 0 & 0 & 0.01 \end{pmatrix}.$$

is also regular. Matrix P contains zeros, and so do P^2, P^3, P^4, and P^5. The 6-step transition probability matrix

$$P^{(6)} = \begin{pmatrix} .009 & .090 & .900 & .001 \\ .001 & .009 & .090 & .900 \\ .810 & .001 & .009 & .180 \\ .162 & .810 & .001 & .027 \end{pmatrix}$$

contains no zeros and proves regularity of this Markov chain.

In fact, computation of all P^h up to $h = 6$ is not required in this problem. Regularity can also be seen from the transition diagram in Figure 6.4. Looking at this diagram, any state i can be reached in 6 steps from any state j. Moving counterclockwise, we can always reach state 4, and then state i in 6 states or less. The record route goes from state 1 to state 3, where it takes 3 steps from 1 to 4, and then 3 more steps from 4 to 3. If we can reach state i from state j in fewer than 6 steps, we just use the remaining steps circling around state 4. For example, state 2 is reached from state 1 in 6 steps as follows:

$$1 \to 2 \to 3 \to 4 \to 4 \to 4 \to 1.$$

Notice that we don't have to compute $p_{ij}^{(h)}$. We only need to verify that they are all positive for some h. ◇

Example 6.16 (IRREGULAR MARKOV CHAIN, ABSORBING STATES). When there is a state i with $p_{ii} = 1$, a Markov chain cannot be regular. There is no

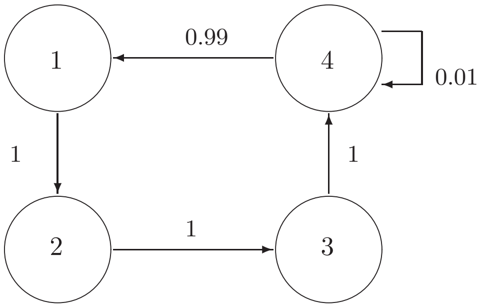


Figure 6.4 *Transition diagram for a regular Markov chain in Example 6.15.*

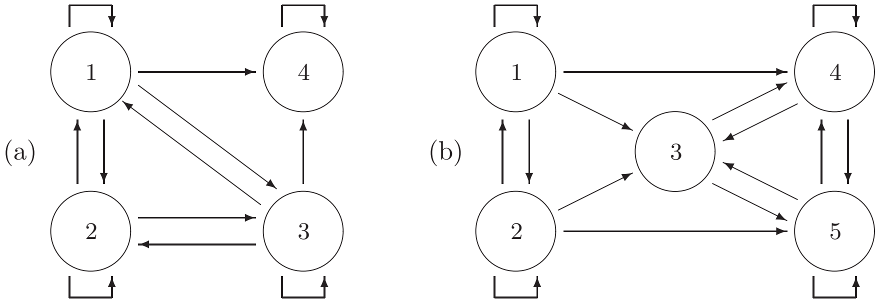


Figure 6.5 *Absorbing states and absorbing zones (Example 6.16).*

exit from state i, therefore, $p_{ij}^{(h)} = 0$ for all h and all $j \neq i$. Such a state is called *absorbing*. For example, state 4 in Figure 6.5a is absorbing, therefore, the Markov chain is irregular.

There may be several absorbing states or an entire absorbing zone, from which the remaining states can never be reached. For example, states 3, 4, and 5 in Figure 6.5b form an *absorbing zone*, some kind of a Bermuda triangle. When this process finds itself in the set $\{3, 4, 5\}$, there is no route from there to the set $\{1, 2\}$. As a result, probabilities $p_{31}^{(h)}$, $p_{52}^{(h)}$ and some others equal 0 for all h. Although all p_{ij} may be less than 1, such Markov chains are still irregular.

Notice that both Markov chains do have steady-state distributions. The first process will eventually reach state 4 and will stay there for good. Therefore, the limiting distribution of $X(h)$ is $\pi = \lim P_h = (0, 0, 0, 1)$. The second Markov chain will eventually leave states 1 and 2 for good, thus its limiting (steady-state) distribution has the form $\pi = (0, 0, \pi_3, \pi_4, \pi_5)$. ◇

Conclusion

This section gives us an important method of analyzing rather complicated stochastic systems. Once it is proved that the process is Markov, it remains to find one-step transition probabilities. Then, the steady-state distribution can be computed, and thus, we obtain the distribution of the process at any time, after a sufficient number of transitions.

This methodology will be our main working tool in Chapter 7, when we study queuing systems and evaluate their performance.

6.3 Counting processes

A large number of situations can be described by *counting processes*. As becomes clear from their name, they *count*. What they count differs greatly from one process to another. These may be counts of arrived jobs, completed tasks, transmitted messages, detected errors, scored goals, and so on.

DEFINITION 6.10

A stochastic process X is **counting** if $X(t)$ is the number of items counted by the time t.

As time passes, one can count additional items, therefore, sample paths of a counting process are always *non-decreasing*. Also, counts are nonnegative integers, $X(t) \in \{0, 1, 2, 3, ...\}$. Hence, counting processes are *discrete state*.

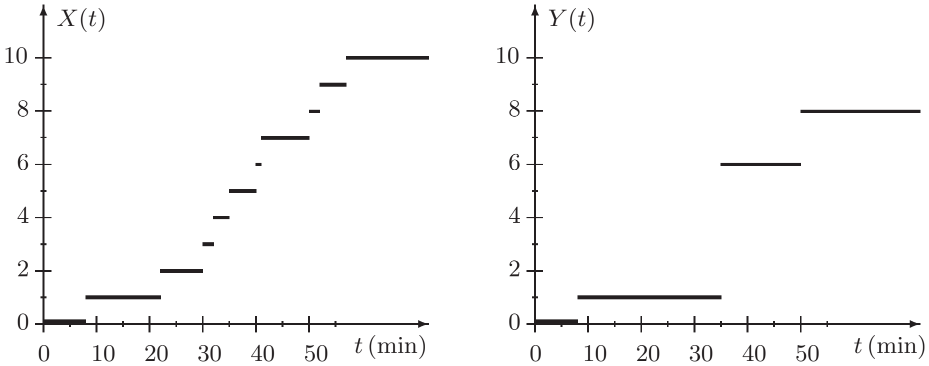


Figure 6.6 *Counting processes in Example 6.17.*

Example 6.17 (E-MAILS AND ATTACHMENTS). Figure 6.6 shows sample paths of two counting process, $X(t)$ being the number of transmitted e-mails by the time t and $Y(t)$ being the number of transmitted attachments. According to the graphs, e-mails were transmitted at $t = 8, 22, 30, 32, 35, 40, 41, 50, 52,$ and 57 min. The e-mail counting process $X(t)$ increments by 1 at each of these times. Only 3 of these e-mails contained attachments. One attachment was sent at $t = 8$, five more at $t = 35$, making the total of $Y(35) = 6$, and two more attachments at $t = 50$, making the total of $Y(50) = 8$. ◇

Two classes of counting processes will be discussed in detail, a discrete-time *Binomial process* and a continuous-time *Poisson process*.

6.3.1 Binomial process

We again consider a sequence of independent Bernoulli trials with probability of success p (Sections 3.4.1–3.4.4) and count "successes."

DEFINITION 6.11

Binomial process $X(n)$ is the number of successes in the first n independent Bernoulli trials, where $n = 0, 1, 2, \ldots$.

It is a discrete-time discrete-space counting stochastic process. Moreover, it is *Markov*, and therefore, a Markov chain.

As we know from Sections 3.4.2–3.4.3, the distribution of $X(n)$ at any time n is Binomial(n, p), and the number of trials Y between two consecutive successes is Geometric(p) (Figure 6.7).

NOTATION			
	$X(n)$	=	number of successes in n trials
	Y	=	number of trials between consecutive successes

Relation to real time: frames

The "time" variable n actually measures the number of trials. It is not expressed in minutes or seconds. However, it can be related to real time.

Suppose that Bernoulli trials occur at equal time intervals, every Δ seconds. Then n trials occur during time $t = n\Delta$, and thus, the value of the process at time t has Binomial distribution with parameters $n = t/\Delta$ and p. The expected number of successes during t seconds is therefore

$$\mathbf{E}\left\{X\left(\frac{t}{\Delta}\right)\right\} = \frac{t}{\Delta}\,p,$$

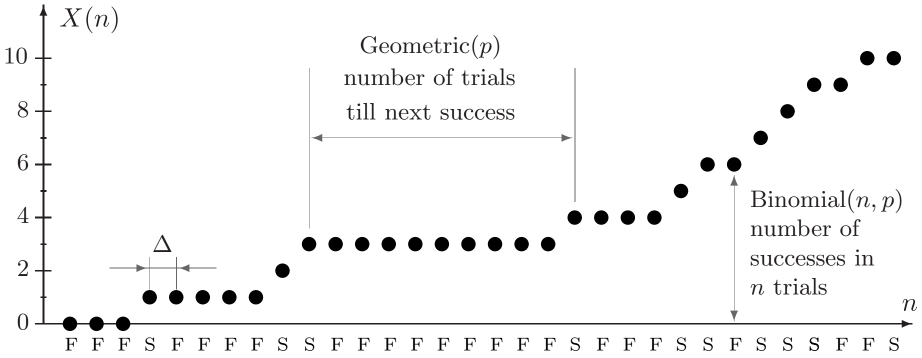


Figure 6.7 *Binomial process (sample path). Legend: S = success, F=failure.*

which amounts to

$$\lambda = \frac{p}{\Delta}$$

successes per second.

DEFINITION 6.12

> **Arrival rate** $\lambda = p/\Delta$ is the average number of successes per one unit of time. The time interval Δ of each Bernoulli trial is called a **frame**. The **interarrival time** is the time between successes.

These concepts, arrival rate and interarrival time, deal with modeling arrivals of jobs, messages, customers, and so on with a Binomial counting process, which is a common method in discrete-time queuing systems (Section 7.3). The key assumption in such models is that no more than 1 arrival is allowed during each Δ-second frame. If this assumption appears unreasonable, and two or more arrivals can occur during the same frame, one has to model the process with a smaller Δ.

NOTATION

λ	=	arrival rate
Δ	=	frame size
p	=	probability of arrival (success) during one frame (trial)
$X(t/\Delta)$	=	number of arrivals by the time t
T	=	interarrival time

The interarrival period consists of a Geometric number of frames Y, each frame taking Δ seconds. Hence, the interarrival time can be computed as

$$T = Y\Delta.$$

It is a *rescaled Geometric* random variable, its possible values are Δ, 2Δ, 3Δ, etc.; its expectation and variance are

$$\begin{aligned}\mathbf{E}(T) &= \mathbf{E}(Y)\Delta = \frac{1}{p}\Delta = \frac{1}{\lambda};\\ \mathrm{Var}(T) &= \mathrm{Var}(Y)\Delta^2 = (1-p)\left(\frac{\Delta}{p}\right)^2 \text{ or } \frac{1-p}{\lambda^2}.\end{aligned}$$

Binomial counting process

$$\boxed{\begin{aligned}\lambda &= p/\Delta\\ n &= t/\Delta\\ X(n) &= \mathit{Binomial}(n,p)\\ Y &= \mathit{Geometric}(p)\\ T &= Y\Delta\end{aligned}}$$

Example 6.18 (Mainframe computer). Jobs are sent to a mainframe computer at a rate of 2 jobs per minute. Arrivals are modeled by a Binomial counting process.

(a) Choose such a frame size that makes the probability of a new job during each frame equal 0.1.

(b) Using the chosen frames, compute the probability of more than 3 jobs received during one minute.

(c) Compute the probability of more than 30 jobs during 10 minutes.

(d) What is the average interarrival time, and what is the variance?

(e) Compute the probability that the next job does not arrive during the next 30 seconds.

Solution.

(a) We have $\lambda = 2$ min^{-1} and $p = 0.1$. Then

$$\Delta = \frac{p}{\lambda} = 0.05 \text{ min or } 3 \text{ sec.}$$

(b) During $t = 1$ min, we have $n = t/\Delta = 20$ frames. The number of jobs during this time is Binomial$(n = 20, p = 0.1)$. From Table A2 (in the Appendix),

$$\boldsymbol{P}\{X(n) > 3\} = 1 - \boldsymbol{P}\{X(n) \le 3\} = 1 - 0.8670 = 0.1330.$$

(c) Here $n = 10/0.05 = 200$ frames, and we use Normal approximation to

Binomial(n, p) distribution (recall p. 103). With a proper continuity correction, and using Table A4,

$$\begin{aligned} \boldsymbol{P}\{X(n) > 30\} &= \boldsymbol{P}\{X(n) > 30.5\} \\ &= \boldsymbol{P}\left\{ \frac{X(n) - np}{\sqrt{np(1-p)}} > \frac{30.5 - (200)(0.1)}{\sqrt{(200)(0.1)(1-0.1)}} \right\} \\ &= \boldsymbol{P}\{Z > 2.48\} = 1 - 0.9934 = 0.0066. \end{aligned}$$

Comparing questions (b) and (c), notice that 3 jobs during 1 minute is not the same as 30 jobs during 10 minutes!

(d) $\mathbf{E}(T) = 1/\lambda = 1/2$ min or 30 sec. Intuitively, this is rather clear because the jobs arrive at a rate of two per minute. The interarrival time has variance

$$\operatorname{Var}(T) = \frac{1-p}{\lambda^2} = \frac{0.9}{2^2} = 0.225.$$

(e) For the interarrival time $T = Y\Delta = Y(0.05)$ and a Geometric variable Y,

$$\begin{aligned} \boldsymbol{P}\{T > 0.5 \text{ min}\} &= \boldsymbol{P}\{Y(0.05) > 0.5\} = \boldsymbol{P}\{Y > 10\} \\ &= \sum_{k=11}^{\infty} (1-p)^{k-1} p = (1-p)^{10} = (0.9)^{10} = 0.3138. \end{aligned}$$

Alternatively, this is also the probability of 0 arrivals during $n = t/\Delta = 0.5/0.05 = 10$ frames, which is also $(1-p)^{10}$. ◇

Markov property

Binomial counting process is *Markov*, with transition probabilities

$$p_{ij} = \begin{cases} p & \text{if} \quad j = i+1 \\ 1-p & \text{if} \quad j = i \\ 0 & \text{otherwise} \end{cases}$$

That is, during each frame, the count increments by 1 in case of a success or remains the same in case of a failure (see Figure 6.8). Transition probabilities are constant over time and independent of the past values of $X(n)$, therefore, it is a *stationary Markov chain*.

This Markov chain is *irregular* because $X(n)$ is non-decreasing, thus $p_{10}^{(h)} = 0$ for all h. Once we see one success, the number of successes will never return to zero. Thus, this process has no steady-state distribution.

The h-step transition probabilities simply form a Binomial distribution. Indeed, $p_{ij}^{(h)}$ is the probability of going from i to j successes in h transitions,

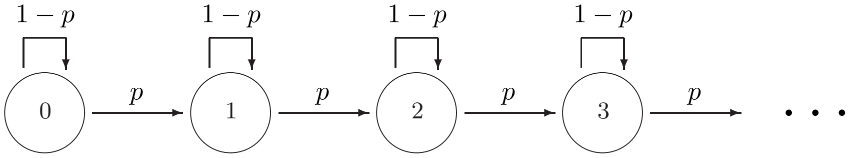


Figure 6.8 *Transition diagram for a Binomial counting process.*

i.e.,

$$\begin{aligned} p_{ij}^{(h)} &= \boldsymbol{P}\left\{(j-i) \text{ successes in } h \text{ trials}\right\} \\ &= \begin{cases} \binom{h}{j-i} p^{j-i}(1-p)^{h-j+i} & \text{if} \quad 0 \le j-i \le h \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Notice that the transition probability matrix has ∞ rows and ∞ columns because $X(n)$ can reach any large value for sufficiently large n:

$$P = \begin{pmatrix} 1-p & p & 0 & 0 & \cdots \\ 0 & 1-p & p & 0 & \cdots \\ 0 & 0 & 1-p & p & \cdots \\ 0 & 0 & 0 & 1-p & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}$$

6.3.2 Poisson process

Continuous time

We now turn attention to *continuous-time* stochastic processes. The time variable t will run continuously through the whole interval, and thus, even during one minute there will be infinitely many moments when the process $X(t)$ may change. Then, how can one study such models?

Often a continuous-time process can be viewed as a limit of some discrete-time process whose frame size gradually decreases to zero, therefore allowing more frames during any period of time,

$$\Delta \downarrow 0 \text{ and } n \uparrow \infty.$$

Example 6.19. For illustration, let us recall how movies are made. Although all motions on the screen seem continuous, we realize that an infinite amount of information could not be stored in a video cassette. Instead, we see a discrete sequence of exposures that run so fast that each motion seems continuous and smooth.

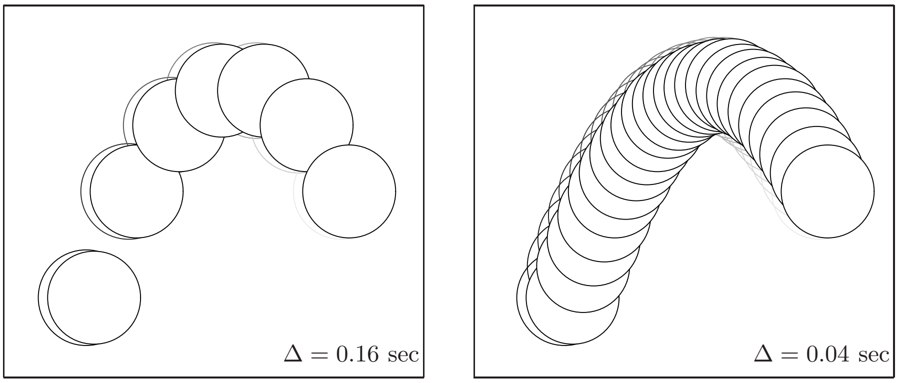


Figure 6.9 *From discrete motion to continuous motion: reducing the frame size Δ.*

Early-age video cameras shot exposures rather slowly, the interval Δ between successive shots was pretty long (~ 0.2–0.5 sec). As a result, the quality of recorded video was rather low. Movies were "too discrete."

Modern camcorders can shoot up to 200 exposures per second attaining $\Delta = 0.005$. With such a small Δ, the resulting movie seems perfectly continuous. A shorter frame Δ results in a "more continuous" process (Figure 6.9). ◇

Poisson process as the limiting case

Going from discrete time to continuous time, Poisson process is the limiting case of a Binomial counting process as $\Delta \downarrow 0$.

DEFINITION 6.13

> **Poisson process** is a continuous-time counting process obtained from a Binomial counting process when its frame size Δ decreases to 0 while the arrival rate λ remains constant.

We can now study a Poisson process by taking a Binomial counting process and letting its frame size decrease to zero.

Consider a Binomial counting process that counts arrivals or other events occurring at a rate λ. Let $X(t)$ denote the number of arrivals occurring until time t. What will happen with this process if we let its frame size Δ converge to 0?

The arrival rate λ remains constant. Arrivals occur at the same rate (say, messages arrive at a server) regardless of your choice of frame Δ.

The number of frames during time t increases to infinity,

$$n = \frac{t}{\Delta} \uparrow \infty \text{ as } \Delta \downarrow 0.$$

The probability of an arrival during each frame is proportional to Δ, so it also decreases to 0,

$$p = \lambda\Delta \downarrow 0 \text{ as } \Delta \downarrow 0.$$

Then, *the number of arrivals* during time t is a Binomial(n, p) variable with expectation

$$\mathbf{E}\,X(t) = np = \frac{tp}{\Delta} = \lambda t.$$

In the limiting case, as $\Delta \downarrow 0$, $n \uparrow \infty$, and $p \downarrow 0$, it becomes a Poisson variable with parameter $np = \lambda t$,

$$X(t) = \text{ Binomial}(n, p) \rightarrow \text{ Poisson}(\lambda)$$

(if in doubt, see p. 71).

The *interarrival time* T becomes a random variable with the c.d.f.

$$\begin{array}{rcll}
F_T(t) & = & \boldsymbol{P}\{T \le t\} = \boldsymbol{P}\{Y \le n\} & \text{because } T = Y\Delta \text{ and } t = n\Delta \\
 & = & 1 - (1-p)^n & \text{Geometric distribution of } Y \\
 & = & 1 - \left(1 - \dfrac{\lambda t}{n}\right)^n & \text{because } p = \lambda\Delta = \lambda t/n \\
 & \rightarrow & 1 - e^{\lambda t}. & \begin{array}{l}\text{This is the “Euler limit”:} \\ (1 + x/n)^n \rightarrow e^x \text{ as } n \rightarrow \infty\end{array}
\end{array}$$

What we got is the c.d.f. of Exponential distribution! Hence, the interarrival time is Exponential with parameter λ.

Further, the time T_k of the k-th arrival is the sum of k Exponential interrarival times that has *Gamma*(k, λ) distribution. From this, we immediately obtain our familiar *Gamma-Poisson formula* (4.15),

$$\boldsymbol{P}\{T_k \le t\} = \boldsymbol{P}\{\text{ k-th arrival before time } t\} = \boldsymbol{P}\{X(t) \ge k\}$$

where T_k is Gamma(k, λ) and $X(t)$ is Poisson(λt).

Poisson process

$$\boxed{\begin{array}{ccl}
X(t) & = & \textit{Poisson}(\lambda t) \\
T & = & \textit{Exponential}(\lambda) \\
T_k & = & \textit{Gamma}(k, \lambda) \\
\boldsymbol{P}\{T_k \le t\} & = & \boldsymbol{P}\{X(t) \ge k\} \\
\boldsymbol{P}\{T_k > t\} & = & \boldsymbol{P}\{X(t) < k\}
\end{array}}$$

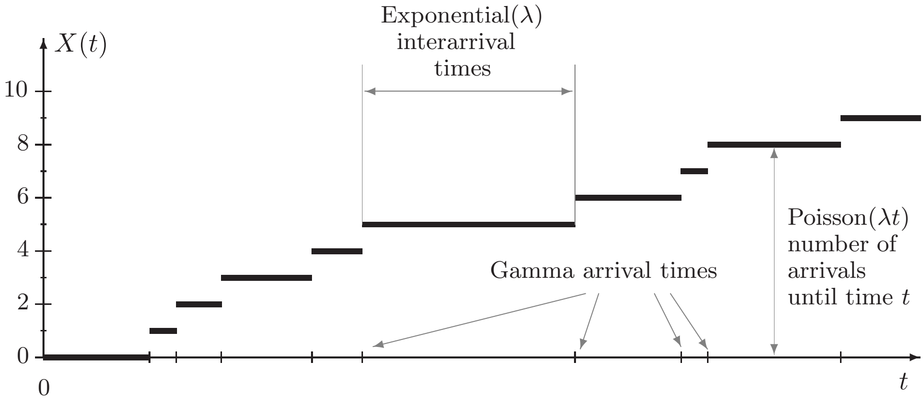


Figure 6.10 *Poisson process (sample path).*

A sample path of some Poisson process is shown in Figure 6.10.

Example 6.20 (Web site hits). The number of hits to a certain web site follows a Poisson process with the intensity parameter $\lambda = 7$ hits per minute.

On the average, how much time is needed to get 10,000 hits? What is the probability that this will happen within 24 hours?

Solution. The time of the $10,000$-th hit T_k has Gamma distribution with parameters $k = 10,000$ and $\lambda = 7$ min^{-1}. Then, the expected time of the k-th hit is

$$\mu = \mathbf{E}(T_k) = \frac{k}{\lambda} = \underline{\text{1,429 min or 23.81 hrs}}.$$

Also,

$$\sigma = \text{Std}(T_k) = \frac{\sqrt{k}}{\lambda} = 14.3 \text{ min}.$$

By the Cental Limit Theorem of Section 4.3, we can use the Normal approximation to the Gamma distribution of T_k. The probability that the 10,000-th hit occurs within 24 hours (1440 min) is

$$\boldsymbol{P}\left\{T_k < 1440\right\} = \boldsymbol{P}\left\{\frac{T_k - \mu}{\sigma} < \frac{1440 - 1429}{14.3}\right\} = \boldsymbol{P}\left\{Z < 0.77\right\} = \underline{0.7794}.$$

◇

Rare events and modeling

A more conventional definition of a Poisson process sounds like this.

DEFINITION 6.14

Poisson process $X(t)$ is a continuous-time counting process with independent increments, such that
(a) $\boldsymbol{P}\{X(t+\Delta)-X(t)=1\}=\lambda\Delta+o(\Delta)$ as $\Delta\to 0$;
(b) $\boldsymbol{P}\{X(t+\Delta)-X(t)>1\}=o(\Delta)$ as $\Delta\to 0$.

In this definition, differences $X(t+\Delta)-X(t)$ are called *increments*. For a Poisson process, an increment is the number of arrivals during time interval $(t,t+\Delta]$.

Properties (a) and (b) imply the known fact that Binomial process probabilities differ from their limits (as $\Delta\to 0$) by a small quantity converging to 0 faster than Δ. This quantity is in general denoted by $o(\Delta)$;

$$\frac{o(\Delta)}{\Delta}\to 0 \text{ as } \Delta\to 0.$$

For a Binomial counting process, $\boldsymbol{P}\{X(t+\Delta)-X(t)=1\}$ is the probability of 1 arrival during 1 frame, and it equals $p=\lambda\Delta$, whereas the probability of more than 1 arrival is zero. For the Poisson process, these probabilities may be different, but only by a little.

The last definition describes formally what we called **rare events**. These events occur at random times; probability of a new event during a short interval of time is proportional to the length of this interval. Probability of more than 1 event during that time is much smaller comparing with the length of such interval. For such sequences of events a Poisson process is a suitable stochastic model. Examples of rare events include telephone calls, message arrivals, virus attacks, errors in codes, traffic accidents, natural disasters, network blackouts, and so on.

Example 6.21 (MAINFRAME COMPUTER, REVISITED). Let us now model the job arrivals in Example 6.18 on p. 165 with a *Poisson process*. Keeping the same arrival rate $\lambda=2$ min^{-1}, this will remove a perhaps unnecessary assumption that no more than 1 job can arrive during any given frame.

Revisiting questions (b)–(e) in Example 6.18, we now obtain,

(b) The probability of more than 3 jobs during 1 minute is

$$\boldsymbol{P}\{X(1)>3\}=1-\boldsymbol{P}\{X(1)\le 3\}=1-0.8571=0.1429,$$

from Table A3 with parameter $\lambda t=(2)(1)=2$.

(c) The probability of more than 30 jobs during 10 minutes is

$$\boldsymbol{P}\{X(10)>30\}=1-\boldsymbol{P}\{X(10)\le 30\}=1-0.9865=0.0135,$$

from Table A3 with parameter $\lambda t=(2)(10)=20$.

(d) For the Exponential($\lambda = 2$) distribution of interarrival times, we get again that

$$\mathbf{E}(T) = \frac{1}{\lambda} = 0.5 \text{ min or } 30 \text{ sec.}$$

This is because the jobs arrive to the computer at the same rate regardless of whether we model their arrivals with a Binomial or Poisson process. Also,

$$\text{Var}(T) = \frac{1}{\lambda^2} = 0.25.$$

(e) The probability that the next job does not arrive during the next 30 seconds is

$$\boldsymbol{P}\{T > 0.5 \text{ min}\} = e^{-\lambda(0.5)} = e^{-1} = 0.3679.$$

Alternatively, this is also the probability of 0 arrivals during 0.5 min, i.e.,

$$\boldsymbol{P}\{X(0.5) = 0\} = 0.3679,$$

from Table A3 with parameter $\lambda t = (2)(0.5) = 1$.

We see that most of the results are somewhat different from Example 6.18, when the same events were modeled by a Binomial counting process. Noticeably, the variance of interarrival times increased. Binomial process introduces a restriction on the number of arrivals during each frame, therefore reducing variability. ◇

6.4 Simulation of stochastic processes

A number of important characteristics of stochastic processes require lengthy complex computations unless they are estimated be means of Monte Carlo methods. One may be interested to explore the time it takes a process to attain a certain level, the time the process spends above some level or above another process, the probability that one process reaches a certain value ahead of another process, etc. Also, it is often important to predict future behavior of a stochastic process.

Discrete-time processes

Sample paths of discrete-time stochastic processes are usually simulated *sequentially*. We start with $X(0)$, then simulate $X(1)$ from the conditional distribution of $X(1)$ given $X(0)$,

$$P_{X_1 \mid X_0}(x_1 \mid x_0) = \frac{P_{X_0,X_1}(x_0,x_1)}{P_{X_0}(x_0)}, \tag{6.6}$$

then $X(2)$ from the conditional distribution of $X(2)$ given $X(0)$ and X_1,

$$P_{X_2 \mid X_0,X_1}(x_2 \mid x_0, x_1) = \frac{P_{X_0,X_1,X_2}(x_0, x_1, x_2)}{P_{X_0,X_1}(x_0, x_1)}, \tag{6.7}$$

etc. In the case of *continuous-state* (but still, discrete-time) processes, the probability mass functions $P_{X_j}(x_j)$ in (6.6) and (6.7) will be replaced with densities $f_{X_j}(x_j)$.

Markov chains

For *Markov chains*, all conditional distributions in (6.6) and (6.7) have a simple form given by by *one-step transition probabilities*. Then, the simulation algorithm reduces to the following.

Algorithm 6.1 *(Simulation of Markov chains)*

1. Initialize: generate $X(0)$ from the initial distribution $P_0(x)$.
2. Transition: having generated $X(t) = i$, generate $X(t + 1)$ as a discrete random variables that takes value j with probability p_{ij}. See Algorithm 5.1 on p. 117.
3. Return to step 2 until a sufficiently long sample path is generated.

Example 6.22 (Weather forecasts and water resources). In Example 6.8, the Markov chain of sunny and rainy days has initial distribution

$$\left\{ \begin{array}{lllll} P_0(\text{sunny}) & = & P_0(1) & = & 0.2, \\ P_0(\text{rainy}) & = & P_0(2) & = & 0.8, \end{array} \right.$$

and transition probability matrix

$$P = \begin{pmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix}.$$

For simplicity, sunny days are denoted by $X = 1$, and rainy days by $X = 2$. The following MATLAB code will generate the forecast for the next 100 days.

```
N   =  100;           % length of sample path
X   =  zeros(N,1);
p   =  [0.2 0.8];     % initial distribution
P   =  [0.7 0.3       % transition probability matrix
         0.4 0.6];
U   =  rand(N,1);     % N Uniform variables to start with
for t=1:N;            % simulate X(1),...,X(N) sequentially
   X(t) = 1*( U(t) <= p(1) ) + 2*( U(t) > p(1) );
                      % X(t)=1 with probability p(1) and
                      % X(t)=2 with probability 1-p(1)=p(2)
   p = P(X(t),:);     % prepare the distribution of X(t+1);
                      % its pmf is {X(t)}^th row of matrix P
end; X
```

This program returns a sequence of states that looks like this:

2211122212221121111111122222222112111222211111111112221211221...

Notice fairly long segments of sunny days ($X = 1$) and rainy days ($X = 2$), showing dependence among the generated variables. This is because a sunny day is more likely to be followed by another sunny day, and a rainy day is more likely to be followed by a rainy day.

Generating a large number of such sequences, we can estimate, say, the probability of 20 consecutive days without rain at least once during the next year, the average time between such droughts, the expected number of times this may happen during the same year, etc. Based on this, water resources can be allocated accordingly. ◇

Binomial process

Simulation of a Binomial process is rather straightforward, as it is based on a sequence of independent Bernoulli trials. After simulating such a sequence, compute partial sums as follows.

```
N      =  100;  p   = 0.4;
X      =  zeros(N,1);       % initialization
Y      =  (rand(N,1) < p);  % sequence of Bernoulli(p) variables
X(1)   =  Y(1);
for t=2:N;                  % X(t) is the number of successes
      X(t)=X(t-1)+Y(t);     % in the first t trials
end; X
```

It is a good illustration to look at the simulated stochastic process in real time.

Animation of the generated discrete-time process can be created as follows.

```
plot(1,X(1),'o');          % Start the plot and allocate the box
axis([0 N 0 max(X)]);      % for the entire simulated segment
hold on;                   % Keep all the plotted points
for t=2:N;
   plot(t,X(t),'o');       % Plot each point with a circle
   pause(0.5);             A half-second pause after each point
end; hold off
```

If you have access to MATLAB, try this code. You will see a real discrete-time process with half-second frames Δ.

Continuous-time processes

Simulation of continuous-time processes has a clear problem. The time t runs continuously through the time interval, taking infinitely many values in this range, however, we cannot store an infinite number of random variables in memory of our computer!

For most practical purposes, it suffices to generate a discrete-time process with a rather short frame Δ (discretization). For example a three-dimensional *Brownian motion process*, a continuous-time process with continuous sample paths and independent Normal increments, can be simulated by the following code.

```
N=5000; X=zeros(N,3);      % Initialize X(t)
Z=randn(N,3);              % N × 3 matrix of Normal increments
X(1,:)=Z(1,:);
for t=2:N;  X(t,:)=X(t-1,:)+Z(t,:);  end;
comet3(X(:,1),X(:,2),X(:,3));
```

The last command, `comet3`, creates a three-dimensional animation of the generated process.

Poisson process

Poisson processes can be generated without discretization. Indeed, although they are continuous-time, the value of $X(t)$ can change only a finite number of times during each interval. The process changes every time when a new "rare event" or arrival occurs, which happens a Poisson(λt) number of times during an interval of length t.

Then, it suffices to generate these moments of arrival. As we know from Section 6.3.2, the first arrival time is Exponential(λ), and each interarrival time is

Exponential(λ) too. Then, a segment of a Poisson process during a time interval $[0, M]$ can be generated and "animated" with the following MATLAB code.

```
M=1000; lambda=0.04;
S=0; T=0;             % T is a growing vector of arrival times
while S<=M;           % Loop ends when arrival time S exceeds M
  Y=-1/lambda * log(rand);     % Exponential interarrival time
  S=S+Y;              % new arrival time
  T=[T S];            % vector of arrival times extends
end;                  % by one element
N=length(T);          % generated number of arrivals
X=zeros(M,1);         % initialize the Poisson process
for t=1:M;
  X(t)=sum(T<=t);     % X(t) is the number of arrivals
end;                  % by the time t
comet(X);             % Animation of the generated process!
```

Summary and conclusions

Stochastic processes are random variables that change, evolve, and develop in time. There are discrete- and continuous-time, discrete- and continuous-state processes, depending on their possible values and possible times.

Markov processes form an important class, where only the most recent value of the process is needed to predict its future probabilities. Then, a Markov chain is fully described by its initial distribution and transition probabilities. Its limiting behavior, after a large number of transitions, is determined by a steady-state distribution which we compute by solving a system of steady-state equations. Binomial and Poisson counting processes are both Markov, with discrete and continuous time, respectively.

Thus we developed an important tool for studying rather complex stochastic (involving uncertainty) systems. As long as the process is Markov, we compute its forecast, steady-state distribution, and other probabilities, expectations, and quantities of interest. Continuous-time processes can be viewed as limits of some discrete-time processes, when the frame size reduces to zero.

In the next chapter, we use this approach to evaluate performance of queuing systems.

Questions and exercises

6.1. A small computer lab has 2 terminals. The number of students working in this lab is recorded at the end of every hour. A computer assistant notices the following pattern:

- If there are 0 or 1 students in a lab, then the number of students in 1 hour has a 50-50% chance to increase by 1 or remain unchanged.
- If there are 2 students in a lab, then the number of students in 1 hour has a 50-50% chance to decrease by 1 or remain unchanged.

(a) Write the transition probability matrix for this Markov chain.

(b) Is this a regular Markov chain? Justify your answer.

(c) Suppose there is nobody in the lab at 7 am. What is the probability of nobody working in the lab at 10 am?

6.2. A computer system can operate in two different modes. Every hour, it remains in the same mode or switches to a different mode according to the transition probability matrix

$$\begin{pmatrix} 0.4 & 0.6 \\ 0.6 & 0.4 \end{pmatrix}.$$

(a) Compute the 2-step transition probability matrix.

(b) If the system is in Mode I at 5:30 pm, what is the probability that it will be in Mode I at 8:30 pm on the same day?

6.3. Markov chains find direct applications in genetics. Here is an example.

An offspring of a black dog is black with probability 0.6 and brown with probability 0.4. An offspring of a brown dog is black with probability 0.2 and brown with probability 0.8.

(a) Write the transition probability matrix of this Markov chain.

(b) Rex is a brown dog. Compute the probability that his grandchild is black.

6.4. Every day, Bill takes the same street from his home to the university. There are 4 street lights along his way, and Bill has noticed the following Markov dependence. If he sees a green light at an intersection, then 60% of time the next light is also green, and 40% of time the next light is red. However, if he sees a red light, then 70% of time the next light is also red, and 30% of time the next light is green.

(a) Construct the transition probability matrix for the street lights.

(b) If the first light is green, what is the probability that the third light is red?

(c) Bill's classmate Jack has *many* street lights between his home and the university. If the *first* street light is green, what is the probability that the *last* street light is red? (Use the steady-state distribution.)

6.5. The pattern of sunny and rainy days on planet Rainbow is a homogeneous Markov chain with two states. Every sunny day is followed by another sunny day with probability 0.8. Every rainy day is followed by another rainy day with probability 0.6. Compute the probability that April 1 next year is rainy on Rainbow.

6.6. A computer device can be either in a busy mode (state 1) processing a task, or in an idle mode (state 2), when there are no tasks to process. Being in a busy mode, it can finish a task and enter an idle mode any minute with the probability 0.2. Thus, with the probability 0.8 it stays another minute in a busy mode. Being in an idle mode, it receives a new task any minute with the probability 0.1 and enters a busy mode. Thus, it stays another minute in an idle mode with the probability 0.9. The initial state is idle. Let X_n be the state of the device after n minutes.

(a) Find the distribution of X_2.

(b) Find the steady-state distribution of X_n.

6.7. A Markov chain has the transition probability matrix

$$P = \begin{pmatrix} 0.3 & ... & 0 \\ 0 & 0 & ... \\ 1 & ... & ... \end{pmatrix}$$

(a) Fill in the blanks.

(b) Show that this is a regular Markov chain.

(c) Compute the steady-state probabilities.

6.8. A Markov chain has 3 possible states: A, B, and C. Every hour, it makes a transition to a *different* state. From state A, transitions to states B and C are equally likely. From state B, transitions to states A and C are equally likely. From state C, it always makes a transition to state A. Find the steady-state distribution of states.

6.9. Tasks are sent to a supercomputer at an average rate of 6 tasks per minute. Their arrivals are modeled by a Binomial counting process with 2-second frames.

(a) Compute the probability of more than 2 tasks sent during 10 seconds.

(b) Compute the probability of more than 20 tasks sent during 100 seconds. You may use a suitable approximation.

6.10. The number of electronic messages received by an interactive message center is modeled by Binomial counting process with 15-second frames. The average arrival rate is 1 message per minute. Compute the probability of receiving more than 3 messages during a 2-minute interval.

6.11. Jobs are sent to a printer at the average rate of 2 jobs per minute. Binomial counting process is used to model these jobs.

(a) What frame length Δ gives the probability 0.1 of an arrival during any given frame?

(b) With this value of Δ, compute the expectation and standard deviation for the number of jobs sent to the printer during a 1-hour period.

6.12. On the average, 2 airplanes per minute land at a certain international airport. We would like to model the number of landings by a Binomial counting process.

(a) What frame length should one use to guarantee that the probability of a landing during any frame does not exceed 0.1?

(b) Using the chosen frames, compute the probability of no landings during the next 5 minutes.

(c) Using the chosen frames, compute the probability of more than 100 landed airplanes during the next hour.

6.13. On the average, every 12 seconds a customer makes a call using a certain phone card. Calls are modeled by a Binomial counting process with 2-second frames. Find the mean and the variance for the time, in seconds, between two consecutive calls.

6.14. Customers of a certain internet service provider connect to the internet at the average rate of 3 customers per minute. Assuming Binomial counting process with 5-second frames, compute the probability of more than 10 new connections during the next 3 minutes. Compute the mean and the standard deviation of the number of seconds between connections.

6.15. Customers of an internet service provider connect to the internet at the average rate of 12 new connections per minute. Connections are modeled by a Binomial counting process.

(a) What frame length Δ gives the probability 0.15 of an arrival during any given frame?

(b) With this value of Δ, compute the expectation and standard deviation for the number of seconds between two consecutive connections.

6.16. Messages arrive at a transmission center according to a Binomial counting process with 30 frames per minute. The average arrival rate is 40 messages per hour. Compute the mean and standard deviation of the number of messages arrived between 10 am and 10:30 am.

6.17. Messages arrive at an electronic message center at random times, with an average of 9 messages per hour.

(a) What is the probability of receiving at least five messages during the next hour?

(b) What is the probability of receiving exactly five messages during the next hour?

6.18. Messages arrive at an interactive message center according to a Binomial counting process with the average interarrival time of 15 seconds. Choosing a frame size of 5 seconds, compute the probability that during 200 minutes of operation, no more than 750 messages arrive.

6.19. Power outages are unexpected rare events occurring according to a Poisson process with the average rate of 3 outages per month. Compute the probability of more than 5 power outages during three summer months.

6.20. Telephone calls to a customer service center occur according to a Poisson process with the rate of 1 call every 3 minutes. Compute the probability of receiving more than 5 calls during the next 12 minutes.

6.21. Network blackouts occur at an average rate of 5 blackouts per month.

(a) Compute the probability of more than 3 blackouts during a given month.

(b) Each blackout costs \$1500 for computer assistance and repair. Find the expectation and standard deviation of the monthly total cost due to blackouts.

6.22. An internet service provider offers special discounts to every third connecting customer. Its customers connect to the internet according to a Poisson process with the rate of 5 customers per minute. Compute:

(a) the probability that no offer is made during the first 2 minutes

(b) expectation and variance of the time of the first offer

6.23. On the average, Mr. Z drinks and drives once in 4 years. He knows that

- Every time when he drinks and drives, he is caught by police.
- According to the laws of his state, the third time when he is caught drinking and driving results in the loss of his driver's license.

- Poisson process is the correct model for such "rare events" as drinking and driving.

What is the probability that Mr. Z will keep his driver's license for at least 10 years?

6.24. Refer to Example 6.9. Find the steady-state distribution for the number of users by writing $X(t)$ as a sum of two independent Markov chains,

$$X(t) = Y_1(t) + Y_2(t),$$

where $Y_i(t) = 1$ if user i is connected at time t and $Y_i(t) = 0$ if user i is not connected, for $i = 1, 2$. Find the steady-state distribution for each Y_i, then use it to find the steady-state distribution for X. Compare your result with Example 6.12.

6.25. (Computer mini-project) Generate 10,000 transitions of the Markov chain in Example 6.9 on p. 153. How many times do you find your generated Markov chain in each of its three states? Does it match the distribution found in Example 6.12?

6.26. (Computer mini-project) In this project, we explore the effect of Markov dependence.

Start with a sunny day and generate weather forecasts for the next 100 days according to the Markov chain in Example 6.7 on p. 148. Observe periods of sunny days and periods of rainy days.

Then consider another city where each day, sunny or rainy, is followed by a sunny day with probability (4/7) and by a rainy day with probability (3/7). As we know from Example 6.11 on p. 156, this is the steady-state distribution of sunny and rainy days, so the overall proportion of sunny days should be the same in both cities. Generate weather forecasts for 100 days and compare with the first city. What happened with periods of consecutive sunny and rainy days?

Each day in the first city, the weather depends on the previous day. In the second city, weather forecasts for different days are independent.

6.27. (Computer mini-project) Generate a 24-hour segment of a Poisson process of arrivals with the arrival rate $\lambda = 5$ hours^{-1}. Graph its trajectory. Do you observe several arrivals in a short period of time followed by relatively long periods with no arrivals at all? Why are the arrivals so unevenly distributed over the 24-hour period?

CHAPTER 7

Queuing Systems

We are now ready to analyze a broad range of *queuing systems* that play a crucial role in Computer Science and other fields.

DEFINITION 7.1

A **queuing system** is a server facility consisting of one or several servers designed to perform certain tasks or process certain jobs and a queue of jobs waiting to be processed.

Jobs arrive at the queuing system, wait for an available server, get processed by this server, and leave.

Examples of queuing systems are:

- a personal or shared computer executing tasks sent by its users
- an internet service provider whose customers connect to the internet, browse, and disconnect
- a printer processing jobs sent to it from different computers
- a customer service with one or several representatives on duty answering calls from their customers
- a TV channel viewed by many people at various times
- a toll area on a highway, or an automated teller machine (ATM) in a bank, where cars arrive, get the required service and depart
- a medical office serving patients

7.1 Main components of a queuing system

How does a queuing system work? What happens with a job when it goes through a queuing system? The main stages are depicted in Figure 7.1.

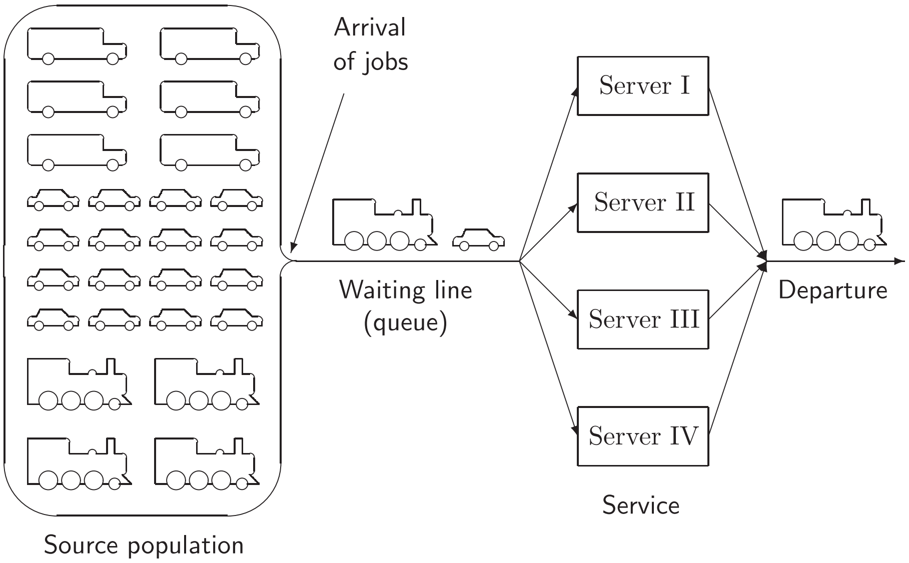


Figure 7.1 *Main components of a queuing system.*

Arrivals

Typically, jobs arrive to a queuing system at random times. A *counting process* $A(t)$ tells the number of arrivals that occurred by the time t. In stationary queuing systems (whose distribution characteristics do not change over time), arrivals occur at *arrival rate*

$$\lambda_A = \frac{\mathbf{E}A(t)}{t}$$

for any $t > 0$, which is the expected number of arrivals per 1 unit of time. Then, the expected time between arrivals is

$$\mu_A = \frac{1}{\lambda_A}.$$

Queuing and routing to servers

Arrived jobs are typically processed according to the order of their arrivals, on a "first come–first serve" basis.

When a new job arrives, it may find the system in different states. If one server is available at that time, it will certainly take the new job. If several servers are available, the job may be randomized to one of them, or the server may be chosen according to some rules. For example, the fastest server or the least loaded server may be assigned to process the new job. Finally, if all servers are busy working on other jobs, the new job will join the queue, wait until all the previously arrived jobs are completed, and get routed to the next available server.

Various additional constraints may take place. For example, a queue may have a *buffer* that limits the number of waiting jobs. Such a queuing system will have *limited capacity*; the total number of jobs in it at any time is bounded by some constant C. If the capacity is full (for example, in a parking garage), a new job cannot enter the system until another job departs.

Also, jobs may leave the queue prematurely, say, after an excessively long waiting time. Servers may also open and close during the day as people need rest and servers need maintenance. Complex queuing systems with many extra conditions may be difficult to study analytically, however, we shall learn Monte Carlo methods for queuing systems in Section 7.6.

Service

Once a server becomes available, it immediately starts processing the next assigned job. In practice, service times are random because they depend on the amount of work required by each task. The average service time is μ_S. It may vary from one server to another as some computers or customer service representatives work faster than others. The *service rate* is defined as the average number of jobs processed by a continuously working server during one unit of time. It equals

$$\lambda_S = \frac{1}{\mu_S}.$$

Departure

When the service is completed, the job leaves the system.

The following parameters and random variables describe performance of a queuing system.

NOTATION

Parameters

λ_A = arrival rate
λ_S = service rate

μ_A = $1/\lambda_A$ = mean interarrival time
μ_S = $1/\lambda_S$ = mean service time

r = $\lambda_A/\lambda_S = \mu_S/\mu_A$ = utilization, or arrival-to-service ratio

<u>Notation</u>

<u>Random variables</u>

$X_s(t)$ = number of jobs receiving service at time t

$X_w(t)$ = number of jobs waiting in a queue at time t

$X(t)$ = $X_s(t) + X_w(t)$,
the total number of jobs in the system at time t

S_k = service time of the k-th job

W_k = waiting time of the k-th job

R_k = $S_k + W_k$, response time, the total time a job spends in the system from its arrival until the departure

Utilization r is an important parameter. As we see in later sections, it shows whether or not a system can function under the current or even higher rate of arrivals, and how much the system is over- or underloaded.

A queuing system is *stationary* if the distributions of S_k, W_k, and R_k are independent of k. In this case, index k will often be omitted.

Most of the time, our goal will be finding the distribution of $X(t)$, the total number of jobs in the system. The other characteristics of a queuing system will be assessed from that. As a result, we shall obtain a comprehensive performance evaluation of a queuing system.

7.2 The Little's Law

The Little's Law gives a simple relationship between the expected number of jobs, the expected response time, and the arrival rate. It is valid for any stationary queuing system.

Little's Law $$\boxed{\lambda_A \, \mathbf{E}(R) = \mathbf{E}(X)}$$

Proof: For an elegant derivation of the Little's Law, calculate the shaded area in Figure 7.2. In this figure, rectangles represent the jobs, stretching between their arrival and departure times. Thus, the length of each rectangle equals

$$\text{Departure time} - \text{Arrival time} = R,$$

and so does its area.

By the time T, there are $A(T)$ arrivals. Among them, $X(T)$ jobs remain in the

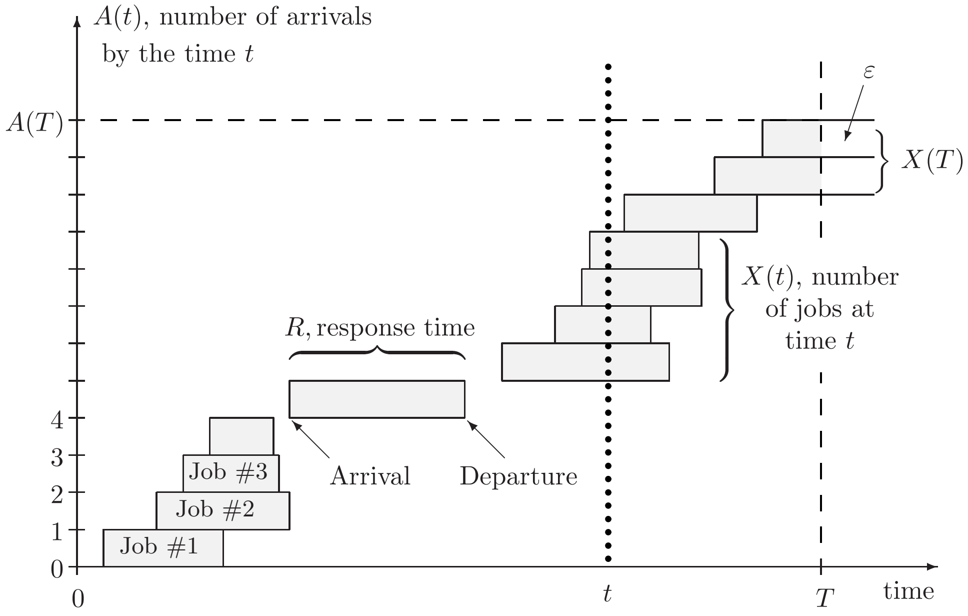


Figure 7.2 *Queuing system and the illustration to the Little's Law.*

system at time T. Only a portion of these jobs is completed by time T, the other portion (call it ε) will take place after time T. Then, the total shaded area equals

$$\text{Shaded area} = \sum_{k=1}^{A(T)} R_k - \varepsilon. \tag{7.1}$$

Alternatively, we can recall from Calculus that every area can be computed by integration. We let t run from 0 to T and integrate the cross-section of the shaded region at t. As seen on the picture, the length of this cross-section is $X(t)$, the number of jobs in the system at time t. Hence,

$$\text{Shaded area} = \int_0^T X(t)dt. \tag{7.2}$$

Combining (7.1) and (7.2), we get

$$\sum_{k=1}^{A(T)} R_k - \varepsilon = \int_0^T X(t)\,dt. \tag{7.3}$$

It remains to take expectations, divide by T, and let $T \to \infty$. Then $\varepsilon/T \to 0$. In the left-hand side of (7.3), we get

$$\lim_{T\to\infty} \frac{1}{T}\,\mathbf{E}\left(\sum_{k=1}^{A(T)} R_k - \varepsilon\right) = \lim_{T\to\infty} \frac{\mathbf{E}(A(T))\,\mathbf{E}(R)}{T} - 0 = \lambda_A\,\mathbf{E}(R).$$

Recall that the arrival rate is $\lambda_A = \mathbf{E}(A(T))/T$.

In the right-hand side of (7.3), we get the average value of $X(t)$,

$$\lim_{T\to\infty} \frac{1}{T}\,\mathbf{E}\int_0^T X(t)\,dt = \mathbf{E}(X).$$

Therefore, $\lambda_A\,\mathbf{E}(R) = \mathbf{E}(X)$. □

Example 7.1 (Queue in a bank). You walk into a bank at 10:00. Being there, you count a total of 10 customers and assume that this is the typical, average number. You also notice that on the average, customers walk in every 2 minutes. When should you expect to finish services and leave the bank?

Solution. We have $\mathbf{E}(X) = 10$ and $\mu_A = 2$ min. By the Little's Law,

$$\mathbf{E}(R) = \frac{\mathbf{E}(X)}{\lambda_A} = \mathbf{E}(X)\mu_A = (10)(2) = \underline{20\text{ min}}.$$

That is, your expected response time is 20 minutes, and you should expect to leave at 10:20. ◇

The Little's Law is universal, it applies to any stationary queuing system and even the system's components — the queue and the servers. Thus, we can immediately deduce that

$$\mathbf{E}(X_w) = \lambda_A\,\mathbf{E}(W),$$

and for the system with only one server,

$$\mathbf{E}(X_s) = \lambda_A\,\mathbf{E}(S).$$

This law is a fairly recent result. It was obtained by *John D. C. Little* who is currently an Institute Professor at Massachusetts Institute of Technology in the United States.

The Little's Law only relates *expectations* of the number of jobs and their response time. In the rest of this chapter, we evaluate the entire distribution of $X(t)$ that directs us to various probabilities and expectations of interest. These quantities will describe and predict performance of a queuing system.

DEFINITION 7.2

The number of jobs in a queuing system, $X(t)$, is called a **queuing process**. In general, it is not a counting process because jobs arrive and depart, therefore, their number may increase and decrease whereas any counting process is non-decreasing.

7.3 Bernoulli single-server queuing process

DEFINITION 7.3

> **Bernoulli single-server queuing process** is a discrete-time queuing process with the following characteristics:
>
> – one server
>
> – unlimited capacity
>
> – arrivals occur according to a Binomial process the probability of a new arrival during each frame is p_A
>
> – the probability of a service completion (and a departure) during each frame is p_S provided that there is at least one job in the system at the beginning of the frame
>
> – service times and interarrival times are independent

Everything learned in Section 6.3.1 about Binomial *counting* processes applies to arrivals of jobs. It also applies to service completions all the time when there is at least one job in the system. We can then deduce that

– there is a Geometric(p_A) number of frames between successive arrivals
– each service takes a Geometric(p_S) number of frames
– service of any job takes at least one frame
– $p_A = \lambda_A \Delta$
– $p_S = \lambda_S \Delta$

Markov property

Moreover, Bernoulli single-server queuing process is a *homogeneous Markov chain* because probabilities p_A and p_S never change. The number of jobs in the system increments by 1 with each arrival and decrements by 1 with each departure (Figure 7.3). Conditions of a Binomial process guarantee that at most one arrival and at most one departure may occur during each frame. Then, we can compute all transition probabilities,

$$\begin{array}{lllll} p_{00} & = & \boldsymbol{P}\{\text{ no arrivals }\} & = & 1 - p_A \\ p_{01} & = & \boldsymbol{P}\{\text{ new arrival }\} & = & p_A \end{array}$$

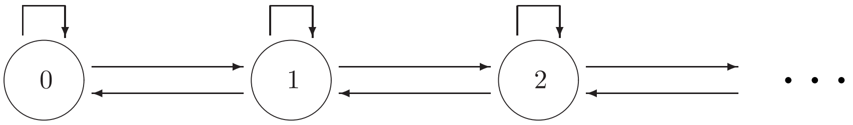


Figure 7.3 *Transition diagram for a Bernoulli single-server queuing process.*

and for all $i \geq 1$,

$$\begin{array}{lcll}
p_{i,i-1} & = & \boldsymbol{P}\{\text{ no arrivals } \bigcap \text{ one departure }\} & = (1-p_A)p_S \\
p_{i,i} & = & \boldsymbol{P}\{\text{ no arrivals } \bigcap \text{ no departures }\} & \\
 & & +\boldsymbol{P}\{\text{ one arrival } \bigcap \text{ one departure }\} & = (1-p_A)(1-p_S)+p_Ap_S \\
p_{i,i+1} & = & \boldsymbol{P}\{\text{ one arrival } \bigcap \text{ no departures }\} & = p_A(1-p_S)
\end{array}$$

The transition probability matrix (of an interesting size $\infty \times \infty$) is three-diagonal,

$$P = \begin{pmatrix}
1-p_A & p_A & 0 & \cdots \\
(1-p_A)p_S & \begin{array}{c}(1-p_A)(1-p_S)\\+p_Ap_S\end{array} & p_A(1-p_S) & \cdots \\
0 & (1-p_A)p_S & \begin{array}{c}(1-p_A)(1-p_S)\\+p_Ap_S\end{array} & \cdots \\
0 & 0 & (1-p_A)p_S & \ddots \\
\vdots & \vdots & \ddots & \ddots
\end{pmatrix} \tag{7.4}$$

All the other transition probabilities equal 0 because the number of jobs cannot change by more than one during any single frame.

This transition probability matrix may be used, for example, to simulate this queuing system and study its performance, as we did with general Markov chains in Section 6.4. One can also compute k-step transition probabilities and predict the load of a server or the length of a queue at any time in future.

Example 7.2 (Printer). Any printer represents a single-server queuing system because it can process only one job at a time while other jobs are stored in a queue. Suppose the jobs are sent to a printer at the rate of 20 per hour, and that it takes an average of 40 seconds to print each job. Currently a printer is printing a job, and there is another job stored in a queue. Assuming Bernoulli single-server queuing process with 20-second frames,

(a) what is the probability that the printer will be idle in 2 minutes?
(b) what is the expected length of a queue in 2 minutes? Find the expected number of waiting jobs and the expected total number of jobs in the system.

Solution. We are given:

$$\begin{aligned}&\lambda_A = 20 \text{ hrs}^{-1} = 1/3 \text{ min}^{-1},\\ &\lambda_S = 1/\mu_S = 1/40 \text{ sec}^{-1} = 1.5 \text{ min}^{-1},\\ &\Delta = 1/3 \text{ min}.\end{aligned}$$

Compute

$$\begin{array}{rcccl} p_A &=& \lambda_A\Delta &=& 1/9\\ p_S &=& \lambda_S\Delta &=& 1/2,\end{array}$$

and all the transition probabilities for the number of jobs X,

$$\begin{array}{rcccl} p_{00} &=& 1-p_A &=& 8/9 = 0.889\\ p_{01} &=& p_A &=& 1/9 = 0.111\end{array}$$

and for $i \geq 1$,

$$\begin{array}{rcl} p_{i,i-1} &=& (1-p_A)p_S = 4/9 = 0.444\\ p_{i,i+1} &=& p_A(1-p_S) = 1/18 = 0.056\\ p_{i,i} &=& 1 - 0.444 - 0.056 = 0.5\end{array} \qquad (7.5)$$

Notice the "shortcut" used in computing $p_{i,i}$. Indeed, the row sum in a transition probability matrix is always 1, therefore, $p_{i,i} = 1 - p_{i,i-1} - p_{i,i+1}$.

We got the following transition probability matrix,

$$P = \begin{pmatrix} .889 & .111 & 0 & 0 & \cdots\\ .444 & .5 & .056 & 0 & \cdots\\ 0 & .444 & .5 & .056 & \cdots\\ 0 & 0 & .444 & .5 & \ddots\\ \vdots & \vdots & \vdots & \ddots & \ddots \end{pmatrix}$$

Since there are 2 min/Δ = 6 frames in 2 minutes, we need a distribution of X after 6 frames, which is

$$P_6 = P_0 P^6,$$

as we know from Section 6.2, formula (6.4). Here there is an interesting problem. How do we deal with matrix P that has infinitely many rows and columns?

Fortunately, we only need a small portion of this matrix. There are 2 jobs currently in the system, the initial distribution is then

$$P_0 = (0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0).$$

In a course of 6 frames, their number can change by 6 at most (look at Figure 7.3), and thus, it is sufficient to consider the the first 9 rows and 9 columns of P only, corresponding to states 0, 1, ..., 8. Then, we compute the distribution after 6 frames,

$$P_6 = P_0 P^6 = (0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0) \begin{pmatrix} .889 & .111 & 0 & \cdots & 0 & 0 \\ .444 & .5 & .056 & \cdots & 0 & 0 \\ 0 & .444 & .5 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & .5 & .056 \\ 0 & 0 & 0 & \cdots & .444 & .5 \end{pmatrix}^6$$

$$= (.644\ .250\ .080\ .022\ .004\ .000\ .000\ .000\ .000).$$

Based on this distribution,

(a) The probability that the printer is idle after 2 minutes is the probability of no jobs in the system at that time,

$$P_6(0) = \underline{0.644}.$$

(b) The expected number of jobs in the system is

$$\mathbf{E}(X) = \sum_{x=0}^{8} x P_6(x) = \underline{0.494 \text{ jobs}}.$$

Out of these jobs, X_w are waiting in a queue and X_s are getting service. However, the server processes at most 1 job at a time, therefore, X_s equals 0 and 1. It has *Bernoulli* distribution, and its expectation is

$$\mathbf{E}(X_s) = \boldsymbol{P}\{\text{ printer is busy }\} = 1 - \boldsymbol{P}\{\text{printer is idle}\} = 1 - 0.644 = 0.356.$$

Therefore, the expected length of a queue equals

$$\mathbf{E}(X_w) = \mathbf{E}(X) - \mathbf{E}(X_s) = 0.494 - 0.356 = \underline{0.138 \text{ waiting jobs}}.$$

◇

Matrix computations in Example 7.2 are rather lengthy if you do them by hand; however, they present no problem for MATLAB, matrix calculators, or any other tool that can handle matrices. A MATLAB code for computing $P_0 P^6$ and $\mathbf{E}(X)$ is

```
P0 = zeros(1,9);   % Initial distribution with P_0(2) = 1.
P0(3) = 1;         % The 1st element of P0 is P_0(0),
                   the 2nd element is P_0(1), etc.
P = zeros(9);      % A 9x9 transition probability matrix
P(1,1) = 8/9; P(1,2) = 1/9; P(9,8) = 4/9; P(9,9) = .5;
for k=2:8; P(k,k-1)=4/9; P(k,k)=.5; P(k,k+1)=1/18; end;
P6 = P0*P^6        % Distribution after 6 frames
EX = (0:8)*P6'     % This computes E(X) = ∑ xP_6(x) as inner
                   % product of two vectors. Transposition (')
                   % converts a row into a column
```

Steady-state distribution

Bernoulli single-server queuing process is an *irregular Markov chain*. Indeed, any k-step transition probability matrix contains zeros because a k-step transition from 0 to $(k+1)$ is impossible as it requires at least $(k+1)$ arrivals, and this is impossible by the conditions of the Binomial process of arrivals.

Nevertheless, any system whose service rate exceeds the arrival rate (that is, jobs can be served faster than they arrive, so there is no overload),

$$\lambda_S > \lambda_A,$$

does have a steady-state distribution. Its computation is possible, despite the infinite dimension of P, but a little cumbersome. Instead, we shall compute the steady-state distribution for a continuous-time queuing process, taking, as usual, the limit of P as $\Delta \to 0$. Computations will simplify significantly.

7.3.1 Systems with limited capacity

As we see, the number of jobs in a Bernoulli single-server queuing system may potentially reach any number. However, many systems have limited resources for storing jobs. Then, there is a maximum number of jobs C that can possibly be in the system simultaneously. This number is called *capacity*.

How does limited capacity change the behavior of a queuing system? Until the capacity C is reached, the system operates without any limitation, as if $C = \infty$. All transition probabilities are the same as in (7.4).

The situation changes only when $X = C$. At this time, the system is full; it can accept new jobs into its queue only if some job departs. As before, the number of jobs decrements by 1 if there is a departure and no new arrival,

$$p_{C,C-1} = (1 - p_A)p_S.$$

In *all* other cases, the number of jobs remains at $X = C$. If there is no

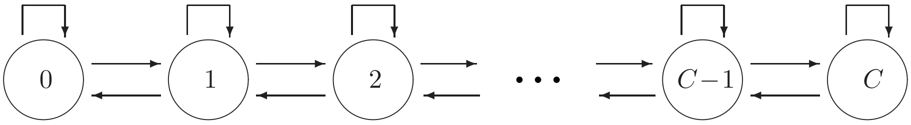


Figure 7.4 *Transition diagram for a Bernoulli single-server queuing process with limited capacity.*

departure during some frame, and a new job arrives, this job cannot enter the system. Hence,

$$p_{C,C} = (1 - p_A)(1 - p_S) + p_A p_S + (1 - p_A)p_S.$$

This Markov chain has states 0, 1, ..., C (Figure 7.4), its transition probability matrix is finite, any state can be reached in C steps, hence, the Markov chain is regular, and its steady-state distribution is readily available.

Example 7.3 (Telephone with two lines). Having a telephone with 2 lines, a customer service representative can talk to a customer and have another one "on hold." This is a system with limited capacity $C = 2$. When the capacity is reached and someone tries to call, (s)he will get a busy signal or voice mail.

Suppose the representative gets an average of 10 calls per hour, and the average phone conversation lasts 4 minutes. Modeling this by a Bernoulli single-server queuing process with limited capacity and 1-minute frames, compute the steady-state distribution and interpret it.

Solution. We have $\lambda_A = 10$ hrs^{-1}= 1/6 min^{-1}, $\lambda_S = 1/4$ min^{-1}, and $\Delta =$ 1 min. Then

$$\begin{array}{rcrcl} p_A &=& \lambda_A\Delta &=& 1/6, \\ p_S &=& \lambda_S\Delta &=& 1/4. \end{array}$$

The Markov chain $X(t)$ has 3 states, $X = 0$, $X = 1$, and $X = 2$. The transition probability matrix is

$$\begin{array}{rcl} P &=& \left(\begin{array}{ccc} 1-p_A & p_A & 0 \\ (1-p_A)p_S & (1-p_A)(1-p_S) + p_A p_S & p_A(1-p_S) \\ 0 & (1-p_A)p_S & 1-(1-p_A)p_S \end{array} \right) \\ &=& \left(\begin{array}{ccc} 5/6 & 1/6 & 0 \\ 5/24 & 2/3 & 1/8 \\ 0 & 5/24 & 19/24 \end{array} \right). \end{array} \qquad (7.6)$$

Next, we solve the steady-state equations

$$\pi P = \pi \;\Rightarrow\; \left\{ \begin{array}{l} \frac{5}{6}\,\pi_0 + \frac{5}{24}\,\pi_1 = \pi_0 \\ \frac{1}{6}\,\pi_0 + \frac{2}{3}\,\pi_1 + \frac{5}{24}\,\pi_2 = \pi_1 \\ \frac{1}{8}\,\pi_1 + \frac{19}{24}\,\pi_2 = \pi_2 \end{array} \right. \;\Rightarrow\; \left\{ \begin{array}{l} \frac{5}{24}\,\pi_1 = \frac{1}{6}\,\pi_0 \\ \frac{1}{6}\,\pi_0 + \frac{2}{3}\,\pi_1 + \frac{5}{24}\,\pi_2 = \pi_1 \\ \frac{1}{8}\,\pi_1 = \frac{5}{24}\,\pi_2 \end{array} \right.$$

As we expect, the second equation follows from the others. After substitution, it becomes an identity

$$\frac{5}{24}\,\pi_1 + \frac{2}{3}\,\pi_1 + \frac{1}{8}\,\pi_1 = \pi_1.$$

We use this as a double-check and turn to the normalizing equation

$$\pi_0 + \pi_1 + \pi_2 = \frac{5}{4}\,\pi_1 + \pi_1 + \frac{3}{5}\,\pi_1 = \frac{57}{20}\,\pi_1 = 1,$$

from where

$$\pi_0 = 25/57 = \underline{0.439}, \quad \pi_1 = 20/57 = \underline{0.351}, \quad \pi_2 = 12/57 = \underline{0.210}.$$

Interpreting this result, 43.9% of time the representative is not talking on the phone, 35.1% of time (s)he talks but has the second line open, and 21.0% of time both lines are busy and new calls don't get through. ◇

7.4 M/M/1 system

We now turn our attention to *continuous-time* queuing processes. Our usual approach is to move from discrete time to continuous time gradually by reducing the frame size Δ to zero.

First, let us explain what the notation "M/M/1" actually means.

NOTATION

A queuing system can be denoted as A/S/n/C, where

A	denotes the distribution of interarrival times
S	denotes the distribution of service times
n	is the number of servers
C	is the capacity

Default capacity is C $= \infty$ (unlimited capacity)

Letter M denotes *Exponential distribution* because it is *memoryless*, and the resulting process is *Markov*.

DEFINITION 7.4

An **M/M/1 queuing process** is a continuous-time queuing process with the following characteristics,

- one server;
- unlimited capacity;
- Exponential interarrival times with the arrival rate λ_A;
- Exponential service times with the service rate λ_S;
- service times and interarrival times are independent.

From Section 6.3.2, we know that Exponential interarrival times imply a *Poisson process* of arrivals with parameter λ_A. This is a very popular model for telephone calls and many other types of arriving jobs.

M/M/1 as a limiting case of a Bernoulli queuing process

We study M/M/1 systems by considering a Bernoulli single-server queuing process and letting its frame Δ go to zero. Our goal is to derive the steady-state distribution and other quantities of interest that evaluate the system's performance.

When the frame Δ gets small, its square, Δ^2 becomes practically negligible, and transition probabilities for a Bernoulli single-server queuing process can be written as

$$\begin{array}{rcccl} p_{00} & = & 1 - p_A & = & 1 - \lambda_A\Delta \\ p_{10} & = & p_A & = & \lambda_A\Delta \end{array}$$

and for all $i \geq 1$,

$$\begin{array}{rcccccl} p_{i,i-1} & = & (1-p_A)p_S & = & (1-\lambda_A\Delta)\lambda_S\Delta & \approx & \lambda_S\Delta \\ p_{i,i+1} & = & p_A(1-p_S) & = & \lambda_A\Delta(1-\lambda_S\Delta) & \approx & \lambda_A\Delta \\ p_{i,i} & = & (1-p_A)(1-p_S)+p_Ap_S & \approx & 1-\lambda_A\Delta-\lambda_S\Delta & & \end{array}$$

Remark: To be rigorous, these probabilities are written up to a small term of order $O(\Delta^2)$, as $\Delta \to 0$. These terms with Δ^2 will eventually cancel out in our derivation.

We have obtained the following transition probability matrix,

$$P \approx \begin{pmatrix} 1-\lambda_A\Delta & \lambda_A\Delta & 0 & 0 & \cdots \\ \lambda_S\Delta & 1-\lambda_A\Delta-\lambda_S\Delta & \lambda_A\Delta & 0 & \cdots \\ 0 & \lambda_S\Delta & 1-\lambda_A\Delta-\lambda_S\Delta & \lambda_A\Delta & \cdots \\ 0 & 0 & \lambda_S\Delta & 1-\lambda_A\Delta-\lambda_S\Delta & \ddots \\ \vdots & \vdots & \vdots & \ddots & \ddots \end{pmatrix} \tag{7.7}$$

Steady-state distribution for an M/M/1 system

As we solve the standard steady-state system of equations (infinitely many equations with infinitely many unknowns, actually),

$$\begin{cases} \pi P = \pi \\ \sum \pi_i = 1 \end{cases}$$

a rather simple, almost familiar final answer may come as a surprise!

Indeed, multiplying $\pi = (\pi_0, \pi_1, \pi_2, \ldots)$ by the first column of P, we get

$$\pi_0(1 - \lambda_A \Delta) + \pi_1 \lambda_S \Delta = \pi_0 \quad \Rightarrow \quad \lambda_A \Delta \pi_0 = \lambda_S \Delta \pi_1 \quad \Rightarrow \quad \boxed{\lambda_A \pi_0 = \lambda_S \pi_1}.$$

The framed expression is called *the first balance equation*.

Remark: We divided both sides of the equation by Δ. At this point, if we kept the Δ^2 terms, we would have gotten rid of them anyway by taking the limit as $\Delta \to 0$.

Next, multiplying π by the second column of P, we get

$$\pi_0 \lambda_A \Delta + \pi_1(1 - \lambda_A \Delta - \lambda_S \Delta) + \pi_2 \lambda_S \Delta = \pi_1 \quad \Rightarrow \quad (\lambda_A + \lambda_S)\pi_1 = \lambda_A \pi_0 + \lambda_S \pi_2.$$

Thanks to the first balance equation, $\lambda_A \pi_0$ and $\lambda_S \pi_1$ cancel each other, and we obtain the *second balance equation*,

$$\boxed{\lambda_A \pi_1 = \lambda_S \pi_2}.$$

This trend of balance equations will certainly continue because every next column of matrix P is just the same as the previous column, only shifted down by 1 position. Thus, the *general balance equation* looks like

$$\boxed{\lambda_A \pi_{i-1} = \lambda_S \pi_i} \quad \text{or} \quad \boxed{\pi_i = r\,\pi_{i-1}} \tag{7.8}$$

where $r = \lambda_A / \lambda_S$ is called *utilization*, or *arrival-to-service ratio*.

Repeatedly applying (7.8) for i, $i-1$, $i-2$, etc., we express each π_i via π_0,

$$\pi_i = r\,\pi_{i-1} = r^2 \pi_{i-2} = r^3 \pi_{i-3} = \ldots = r^i \pi_0.$$

Finally, we recognize the geometric series and apply the normalizing condition,

$$\sum_{i=0}^{\infty} \pi_i = \sum_{i=0}^{\infty} r^i \pi_0 = \frac{\pi_0}{1-r} = 1 \quad \Rightarrow \quad \begin{cases} \pi_0 &= 1 - r \\ \pi_1 &= r\pi_0 = r(1-r) \\ \pi_2 &= r^2 \pi_0 = r^2(1-r) \\ & \quad \text{etc.} \end{cases}$$

This distribution of $X(t)$ is *Shifted Geometric*, because $Y = X + 1$ has the standard Geometric distribution with parameter $p = 1 - r$,

$$\boldsymbol{P}\{Y = y\} = \boldsymbol{P}\{X = y - 1\} = \pi_{y-1} = r^{y-1}(1-r) = (1-p)^{y-1}p \text{ for } y \geq 1,$$

as in Section 3.4.3. This helps us compute the expected number of jobs in the system at any time and its variance,

$$\mathbf{E}(X) = \mathbf{E}(Y-1) = \mathbf{E}(Y) - 1 = \frac{1}{1-r} - 1 = \frac{r}{1-r}$$

and

$$\mathrm{Var}(X) = \mathrm{Var}(Y-1) = \mathrm{Var}(Y) = \frac{r}{(1-r)^2}$$

(for Geometric expectation and variance, see (3.10) on p. 66).

M/M/1 system: steady-state distribution of the number of jobs

$$\boxed{\begin{aligned} &\pi_x = \boldsymbol{P}\{X = x\} = r^x(1-r) \\ &\qquad \text{for } x = 0, 1, 2, \ldots \\ &\mathbf{E}(X) = \frac{r}{1-r} \\ &\mathrm{Var}(X) = \frac{r}{(1-r)^2} \\ &\text{where } r = \lambda_A/\lambda_S = \mu_S/\mu_A \end{aligned}} \tag{7.9}$$

7.4.1 Evaluating the system's performance

Many important system characteristics can be obtained directly from the distribution (7.9).

Utilization

We now see the actual meaning of the *arrival-to-service ratio*, or *utilization*. $r = \lambda_A/\lambda_S$. According to (7.9), it equals

$$r = 1 - \pi_0 = \boldsymbol{P}\{X > 0\},$$

which is the probability that there are jobs in the system, and therefore, the server is busy processing a job. Thus,

$$\begin{aligned} \boldsymbol{P}\{\text{ server is busy }\} &= r \\ \boldsymbol{P}\{\text{ server is idle }\} &= 1-r \end{aligned}$$

We can also say that r is the proportion of time when the server is put to work. In other words, r (utilization!) shows how much the server is utilized.

The system is functional if $r < 1$. In fact, our derivation of the distribution of X is only possible when $r < 1$, otherwise the geometric series used there diverges.

If $r \geq 1$, the system gets *overloaded*. Arrivals are too frequent comparing with the service rate, and the system cannot manage the incoming flow of jobs. The number of jobs in the system will accumulate in this case (unless, of course, it has a limited capacity).

Waiting time

When a new job arrives, it finds the system with X jobs in it. While these X jobs are being served, the new job awaits for its turn in a queue. Thus, its waiting time consists of service times of X earlier jobs,

$$W = S_1 + S_2 + S_3 + \ldots + S_X.$$

Perhaps, the first job in this queue has already started its service. This possibility, however, does not affect the distribution of its service time S_1. Recall that service times in M/M/1 systems are *Exponential*, and this distribution has a *memoryless property*. At any moment, the remaining service time for this job still has Exponential(λ_S) distribution regardless of how long it is already being served!

We can then conclude that the new job has *expected waiting time*

$$\mathbf{E}(W) = \mathbf{E}(S_1 + \ldots + S_X) = \mathbf{E}(S)\,\mathbf{E}(X) = \frac{\mu_S\, r}{1-r} \quad \text{or} \quad \frac{r}{\lambda_S(1-r)}.$$

(By writing the product of expectations, we actually used the fact that service times are independent of the number of jobs in the system at that time).

Response time

Response time is the time a job spends in the system, from its arrival until its departure. It consists of waiting time (if any) and service time. The *expected response time* can then be computed as

$$\mathbf{E}(R) = \mathbf{E}(W) + \mathbf{E}(S) = \frac{\mu_S\, r}{1-r} + \mu_S = \frac{\mu_S}{1-r} \quad \text{or} \quad \frac{1}{\lambda_S(1-r)}.$$

Remark: Notice that W is a rare example of a variable whose distribution is neither discrete nor continuous. On one hand, it has a probability mass function at 0 because $\boldsymbol{P}\{W=0\} = 1-r$ is the probability that the server is idle and available and there is no waiting time for a new job. On the other hand, W has a density $f(x)$ for all $x > 0$. Given any positive number of jobs $X = n$, the waiting time is the sum of n Exponential times which is Gamma(n, λ_S). Such a distribution of W is *mixed*.

Queue

The length of a queue is the number of waiting jobs,

$$X_w = X - X_s.$$

The number of jobs X_s getting service at any time is 0 or 1. It is therefore a Bernoulli variable with parameter

$$\boldsymbol{P}\{\text{ server is busy }\} = r.$$

Hence, the *expected queue length* is

$$\mathbf{E}(X_w) = \mathbf{E}(X) - \mathbf{E}(X_s) = \frac{r}{1-r} - r = \frac{r^2}{1-r}.$$

Little's Law revisited

The Little's Law certainly applies to the M/M/1 queuing system and its components, the queue and the server. Assuming the system is functional ($r < 1$), all the jobs go through the entire system, and thus, each component is subject to the same arrival rate λ_A. The Little's Law then guarantees that

$$\begin{array}{rcl} \lambda_A\,\mathbf{E}(R) & = & \mathbf{E}(X), \\ \lambda_A\,\mathbf{E}(S) & = & \mathbf{E}(X_s), \\ \lambda_A\,\mathbf{E}(W) & = & \mathbf{E}(X_w). \end{array}$$

Using our results in (7.10), it wouldn't be a problem for you to verify all three equations, would it?

M/M/1: main performance characteristics

$$\boxed{\begin{array}{lcccc} \mathbf{E}(R) & = & \dfrac{\mu_S}{1-r} & = & \dfrac{1}{\lambda_S(1-r)} \\[2mm] \mathbf{E}(W) & = & \dfrac{\mu_S\, r}{1-r} & = & \dfrac{r}{\lambda_S(1-r)} \\[2mm] \mathbf{E}(X) & = & \dfrac{r}{1-r} & & \\[2mm] \mathbf{E}(X_w) & = & \dfrac{r^2}{1-r} & & \\[2mm] \multicolumn{2}{l}{\boldsymbol{P}\{\text{server is busy}\}} & = & r & \\ \multicolumn{2}{l}{\boldsymbol{P}\{\text{server is idle}\}} & = & 1-r & \end{array}} \tag{7.10}$$

Example 7.4 (Message transmission with a single channel). Messages arrive to a communication center at random times with an average of 5 messages per minute. They are transmitted through a single channel in the order they were received. On average, it takes 10 seconds to transmit a message. Conditions of an M/M/1 queue are satisfied. Compute the main performance characteristics for this center.

Solution. The arrival rate $\lambda_A = 5$ min^{-1}and the expected service time $\mu_S = 10$ sec or $(1/6)$ min are given. Then, the utilization is

$$r = \lambda_A/\lambda_S = \lambda_A\mu_S = \underline{5/6}.$$

This also represents the proportion of time when the channel is busy and the probability of a non-zero waiting time.

The average number of messages stored in the system at any time is

$$\mathbf{E}(X) = \frac{r}{1-r} = \underline{5}.$$

Out of these, an average of

$$\mathbf{E}(X_w) = \frac{r^2}{1-r} = \underline{4.17}$$

messages are waiting, and

$$\mathbf{E}(X_s) = r = \underline{0.83}$$

are being transmitted.

When a message arrives to the center, its waiting time until its transmission begins averages

$$\mathbf{E}(W) = \frac{\mu_S r}{1-r} = \underline{50 \text{ seconds}},$$

whereas the total amount of time since its arrival until the end of its transmission has an average of

$$\mathbf{E}(R) = \frac{\mu_S}{1-r} = \underline{1 \text{ minute}}.$$

◇

Example 7.5 (Forecast). Let's continue Example 7.4. Suppose that next year the customer base of our transmission center is projected to increase by 10%, and thus, the intensity of incoming traffic λ_A increases by 10% too. How will this affect the center's performance?

Solution. Recompute the main performance characteristics under the new arrival rate

$$\lambda_A^{\text{NEW}} = (1.1)\lambda_A^{\text{OLD}} = 5.5 \text{ min}^{-1}.$$

Now the utilization equals $r = 11/12$, getting dangerously close to 1 where the system gets overloaded. For high values of r, various parameters of the system increase rapidly. A 10% increase in the arrival rate will result in rather significant changes in other variables. Using (7.10), we now get

$$\begin{array}{lcl}
\mathbf{E}(X) & = & 11 \text{ jobs}, \\
\mathbf{E}(X_w) & = & 10.08 \text{ jobs}, \\
\mathbf{E}(W) & = & 110 \text{ seconds, and} \\
\mathbf{E}(R) & = & 2 \text{ minutes}.
\end{array}$$

We see that the response time, the waiting time, the average number of stored messages, and therefore, the average required amount of memory more than doubled when the number of customers increased by mere 10%. ◇

When a system gets nearly overloaded

As we observed in Example 7.5, the system slowed down significantly as a result of a 10% increase in the intensity of incoming traffic, projected for the next year. One may try to forecast the two-year future of the system, assuming a 10% increase of a customer base each year. It will appear that during the second year the utilization will exceed 1, and the system will be unable to function.

What is a practical solution to this problem? Another channel or two may be added to the center to help the existing channel handle all the arriving messages! The new system will then have more than one channel-server. Such systems are analyzed in the next section.

7.5 Multiserver queuing systems

We now turn our attention to queuing systems with several servers. We assume that each server can perform the same range of services, however, in general, some servers may be faster than others. Thus, the service times for different servers may potentially have different distributions.

When a job arrives, it either finds all servers busy serving jobs, or it finds one or several available servers. In the first case, the job will wait in a queue for its turn whereas in the second case, it will be routed to one of the idle servers. A mechanism assigning jobs to available servers may be random, or it may be based on some rule. For example, some companies will make sure that each call to their customer service is handled by the least loaded customer service representative.

The number of servers may be finite or infinite. A system with infinitely many servers can afford an unlimited number of concurrent users. For example, any number of people can watch a TV channel simultaneously. A job served by a system with $k = \infty$ servers will never wait; it will always find available idle servers.

As in previous sections, here is our plan for analyzing multiserver systems:

– first, verify if the number of jobs in the system at time t is a Markov process and write its transition probability matrix P. For continuous-time processes, we select "very short" frames Δ;

– next, compute the steady-state distribution π, and

– finally, use π to compute the system's long-term performance characteristics.

We treat a few common and analytically simple cases in detail. Advanced theory goes further; in this book we shall analyze more complex and non-Markov queuing systems by Monte Carlo methods in Section 7.6.

Notice that utilization r no longer has to be less than 1. A system with k servers can handle k times the traffic of a single-server system, therefore, it will function with any $r < k$.

7.5.1 Bernoulli k-server queuing process

DEFINITION 7.5

Bernoulli k-server queuing process is a discrete-time queuing process with the following characteristics,

- k servers
- unlimited capacity
- arrivals occur according to a Binomial counting process; the probability of a new arrival during each frame is p_A
- during each frame, each *busy* server completes its job with probability p_S independently of the other servers and independently of the process of arrivals

Markov property

As a result, all service times are *Geometric*(p_S), and all interarrival times are *Geometric*(p_A), multiplied by the frame length Δ. This is similar to a single-server process in Section 7.3. Geometric variables have a *memoryless property*, they forget the past, and therefore, again our process is Markov.

The new feature is that now several jobs may finish during the same frame.

Suppose that $X_s = n$ jobs are currently getting service. During the next frame, each of them may finish and depart, independently of the other jobs. Then the number of departures is the number of successes in n independent Bernoulli trials, and therefore, it has *Binomial* distribution with parameters n and p_S.

This will help us compute the transition probability matrix.

Transition probabilities

Let's compute transition probabilities

$$p_{ij} = \boldsymbol{P}\left\{X(t+\Delta) = j \mid X(t) = i\right\}.$$

To do this, suppose there are i jobs in a k-server system. Then,

- for $i \leq k$, the number of servers is sufficient for the current jobs, all jobs are getting service, and the number of departures X_d during the next frame is Binomial(i, p_S);
- for $i > k$, there are more jobs than servers. Then all k servers are busy, and the number of departures X_d during the next frame is Binomial(k, p_S).

Every time, the number n of busy servers is the smaller of the number of jobs i and the total number of servers k,

$$n = \min\left\{i, k\right\}.$$

In addition, a new job arrives during the next frame with probability p_A.

Accounting for all these possibilities, we get:

$$\begin{array}{lll}
p_{i,i+1} & = & \boldsymbol{P}\left\{ \text{ 1 arrival, 0 departures } \right\} = p_A \cdot (1-p_S)^n; \\
p_{i,i} & = & \boldsymbol{P}\left\{ \text{ 1 arrival, 1 departure } \right\} + \boldsymbol{P}\left\{ \text{ 0 arrivals, 0 departures } \right\} \\
 & = & p_A \cdot np_S(1-p_S)^{n-1} + (1-p_A)\cdot(1-p_S)^n; \\
p_{i,i-1} & = & \boldsymbol{P}\left\{ \text{ 1 arrival, 2 departures } \right\} + \boldsymbol{P}\left\{ \text{ 0 arrivals, 1 departure } \right\} \\
 & = & p_A \cdot \left(\begin{array}{c} n \\ 2 \end{array}\right) p_S^2(1-p_S)^{n-2} + (1-p_A)\cdot np_S(1-p_S)^{n-1}; \\
\cdots & \cdots & \cdots\;\cdots\;\cdots \\
p_{i,i-n} & = & \boldsymbol{P}\left\{ \text{ 0 arrivals, } n \text{ departures } \right\} \\
 & = & (1-p_A)\cdot p_S^n.
\end{array} \tag{7.11}$$

Seemingly long, these formulas only need the computation of Binomial probabilities for the number of departing jobs X_d.

A transition diagram for a 2-server system is shown in Figure 7.5. The number of concurrent jobs can make transitions from i to $i-2$, $i-1$, i, and $i+1$.

Example 7.6 (Customer service as a two-server system with limited capacity). There are two customer service representatives on duty answering customers' calls. When both of them are busy, two more customers may be "on hold," but other callers will receive a "busy signal." Customers

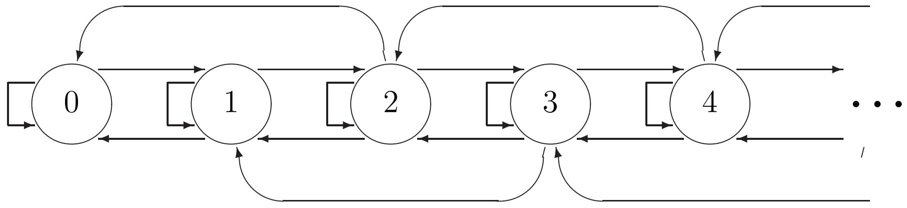


Figure 7.5 *Transition diagram for a Bernoulli queuing system with 2 servers.*

call at the rate of 1 call every 5 minutes, and the average service takes 8 minutes. Assuming a two-server Bernoulli queuing system with limited capacity and 1-minute frames, compute

(a) the steady-state distribution of the number of concurrent jobs;

(b) the proportion of callers who get a "busy signal";

(c) the percentage of time each representative is busy, if each of them takes 50% of all calls.

Solution. This system has $k = 2$ servers, capacity $C = 4$, $\lambda_A = 1/5 \text{ min}^{-1}$, $\lambda_S = 1/8 \text{ min}^{-1}$, and $\Delta = 1$ min. Then

$$p_A = \lambda_A \Delta = 0.2 \text{ and } p_S = \lambda_S \Delta = 0.125.$$

Using (7.11), we compute transition probabilities:

$$\begin{array}{cc|c}
 & & \text{To} \\
 & & \text{state:} \\
P = & \begin{pmatrix} 0.8000 & 0.2000 & 0 & 0 & 0 \\ 0.1000 & 0.7250 & 0.1750 & 0 & 0 \\ 0.0125 & 0.1781 & 0.6562 & 0.1531 & 0 \\ 0 & 0.0125 & 0.1781 & 0.6562 & 0.1531 \\ 0 & 0 & 0.0125 & 0.1781 & 0.8094 \end{pmatrix} & \begin{array}{c} 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \\
\hline
\begin{array}{c}\text{From} \\ \text{state:}\end{array} & \begin{array}{ccccc} 0 \quad & 1 \quad & 2 \quad & 3 \quad & 4 \end{array} &
\end{array}$$

Frankly, we used the following MATLAB code to compute this matrix,

```
k=2; C=4; pa=0.2; ps=0.125;
P = zeros(C+1,C+1); for i=0:C;
  n = min(i,k);          % number of busy servers
  for j=0:n;
    P(i+1,i-j+1) = ...   % transition i -> i-j
    pa * nchoosek(n,j+1)*(1-ps)^(n-j-1)*ps^(j+1)...
    + (1-pa) * nchoosek(n,j)*(1-ps)^(n-j)*ps^j;
  end;
  if i<C;                % capacity is not reached
    P(i+1,i+2) = pa * (1-ps)^n;
      else               % capacity is reached
    P(C+1,C+1) = P(C+1,C+1) + pa * (1-ps)^n;
                         % new calls are not accepted
  end;
end; P
```

(a) The steady-state distribution for this system is

$$\left\{\begin{array}{lcl} \pi_0 & = & 0.1527 \\ \pi_1 & = & 0.2753 \\ \pi_2 & = & 0.2407 \\ \pi_3 & = & 0.1837 \\ \pi_4 & = & 0.1476 \end{array}\right.$$

(b) The proportion of callers who hear a "busy signal" is the probability that the system is full when a job arrives, which is

$$\boldsymbol{P}\left\{X = C\right\} = \pi_4 = \underline{0.1476}.$$

(c) Each customer service rep is busy when there are two, three, or four jobs in the system, plus a half of the time when there is one job. This totals

$$\pi_2 + \pi_3 + \pi_4 + 0.5\pi_1 = \underline{0.7090} \text{ or } \underline{70.90\%}.$$

◇

7.5.2 M/M/k systems

An M/M/k system is a multiserver extention of M/M/1. According to the general "A/S/n/C" notation in Section 7.4, it means

DEFINITION 7.6

An **M/M/k queuing process** is a continuous-time queuing process with

- k servers
- unlimited capacity
- Exponential interarrival times with the arrival rate λ_A
- Exponential service time for each server with the service rate λ_S, independent of all the arrival times and the other servers

Once again, we move from the discrete-time Bernoulli multiserver process to the continuous-time M/M/k by letting the frame Δ go to 0. For very small Δ, transition probabilities (7.11) simplify,

$$\begin{aligned}
p_{i,i+1} &= \lambda_A\Delta \cdot (1-\lambda_S\Delta)^n \approx \lambda_A\Delta = p_A \\
p_{i,i} &= \lambda_A\Delta \cdot n\lambda_S\Delta(1-\lambda_S\Delta)^{n-1} + (1-\lambda_A\Delta)\cdot(1-\lambda_S\Delta)^n \\
&\approx 1-\lambda_A\Delta - n\lambda_S\Delta = 1-p_A-np_S \\
p_{i,i-1} &\approx n\lambda_S\Delta = np_S \\
p_{i,j} &= 0 \;\text{ for all other } j.
\end{aligned} \tag{7.12}$$

Again, $n = \min\{i, k\}$ is the number of jobs receiving service among the total of i jobs in the system.

For example, for $k = 3$ servers, the transition probability matrix is

$$P = \begin{pmatrix}
1-p_A & p_A & 0 & 0 & 0 & \ddots \\
p_S & 1-p_A-p_S & p_A & 0 & 0 & \ddots \\
0 & \mathbf{2}p_S & 1-p_A-2p_S & p_A & 0 & \ddots \\
0 & 0 & \mathbf{3}p_S & 1-p_A-3p_S & p_A & \ddots \\
0 & 0 & 0 & \mathbf{3}p_S & 1-p_A-3p_S & \ddots \\
0 & 0 & 0 & 0 & \mathbf{3}p_S & \ddots \\
\ddots & \ddots & \ddots & \ddots & \ddots & \ddots
\end{pmatrix} \tag{7.13}$$

Let's try to understand these probabilities intuitively. Recall that Δ is very small, so we ignored terms proportional to Δ^2, Δ^3, etc. Then, no more than one event, arrival or departure, may occur during each frame. Probability of more than one event is of the order $O(\Delta^2)$. Changing the number of jobs by 2 requires at least 2 events, and thus, such changes cannot occur during one frame.

At the same time, transition from i to $i-1$ may be caused by a departure of any of n currently served jobs. This is why we see the departure probability p_S multiplied by n.

Steady-state distribution

From matrix P, it would not be hard for us to find π, the steady-state distribution for the number of concurrent jobs X, and further performance characteristics. A similar derivation for a special case $k = 1$ is in Section 7.4.

Again, we solve the system

$$\begin{cases} \pi P = \pi \\ \sum\limits_i \pi_i = 1 \end{cases}$$

Multiplying π by the first column of P gives our familiar *balance equation*

$$\pi_0(1-p_A)+\pi_1 p_S = \pi_0 \quad \Rightarrow \quad \pi_0 p_A = \pi_1 p_S \quad \Rightarrow \quad \pi_0 p_A = \pi_1 p_S \quad \Rightarrow \quad \boxed{\pi_1 = r\pi_0},$$

where

$$r = \frac{p_A}{p_S} = \frac{\lambda_A}{\lambda_S}$$

is *utilization*. Things become different from the second balance equation, affected by multiple servers,

$$\pi_0 p_A + \pi_1(1-p_A-p_S) + 2\pi_2 p_S = \pi_1 \quad \Rightarrow \quad \pi_1 p_A = 2\pi_2 p_S \quad \Rightarrow \quad \boxed{\pi_2 = 2r\pi_1}.$$

Continuing, we get all the balance equations,

$$\begin{cases} \pi_1 &=& r\pi_0 \\ \pi_2 &=& r\pi_1/2 &=& r^2\pi_0/2! \\ \pi_3 &=& r\pi_2/3 &=& r^3\pi_0/3! \\ \ldots & \ldots & \ldots & \ldots & \ldots \quad \ldots \\ \pi_k &=& r\pi_{k-1}/k &=& r^k\pi_0/k! \end{cases}$$

until the number of jobs k occupies all the servers, after which

$$\begin{cases} \pi_{k+1} &=& (r/k)\pi_k &=& (r/k)\, r^k\pi_0/k! \\ \pi_{k+2} &=& (r/k)\pi_{k+1} &=& (r/k)^2\, r^k\pi_0/k! \\ \quad \text{etc.} \end{cases}$$

The normalizing equation $\sum \pi_i = 1$ returns

$$\begin{aligned} 1 &= \pi_0 + \pi_1 + \ldots \\ &= \pi_0 \left(1 + r + \frac{r^2}{2!} + \frac{r^3}{3!} + \ldots + \frac{r^k}{k!} + \frac{r^k}{k!}(r/k) + \frac{r^k}{k!}(r/k)^2 + \ldots \right) \\ &= \pi_0 \left(\sum_{i=0}^{k-1} \frac{r^i}{i!} + \frac{r^k}{(1-r/k)k!} \right), \end{aligned}$$

using in the last line the formula for a geometric series.

We have obtained the following result.

M/M/k system: steady-state distribution of the number of jobs

$$\pi_x = \boldsymbol{P}\{X = x\} = \begin{cases} \dfrac{r^x}{x!}\pi_0 & \text{for } x \le k \\ \dfrac{r^x}{k!}\pi_0 \left(\dfrac{r}{k}\right)^{x-k} & \text{for } x > k \end{cases}$$

where

$$\pi_0 = \boldsymbol{P}\{X = 0\} = \frac{1}{\displaystyle\sum_{i=0}^{k-1} \frac{r^i}{i!} + \frac{r^k}{(1-r/k)k!}}$$

and $r = \lambda_A / \lambda_S$

(7.14)

Example 7.7 (A MULTICHANNEL MESSAGE TRANSMISSION CENTER). In Example 7.5 on p. 201, we worried about the system's incapability to handle the increasing stream of messages. That was a single-server queuing system.

Suppose now that in view of our forecast 2 additional channels were built with the same parameters as the first channel. As a result, we have an M/M/3 system. Now it can easily handle even twice as much traffic!

Indeed, suppose now that the arrival rate λ_A has doubled since Example 7.4. Now it equals 10 min^{-1}. The service rate is still 6 min^{-1} for each server. The system's utilization is

$$r = \lambda_A / \lambda_S = 10/6 = 1.67 > 1,$$

which is fine. Having 3 servers, the system will function with any $r < 3$.

What percentage of messages will be sent immediately, with no waiting time?

Solution. A message does not wait at all times when there is an idle server (channel) to transmit it. The latter happens when the number of jobs in the system is less than the number of servers. Hence,

$$\boldsymbol{P}\{W=0\} = \boldsymbol{P}\{X<3\} = \pi_0+\pi_1+\pi_2 = \underline{0.70},$$

where the steady-state probabilities are computed by formula (7.14) as

$$\pi_0 = \frac{1}{1+1.67+\dfrac{(1.67)^2}{2!}+\dfrac{(1.67)^3}{(1-1.67/3)3!}} = \frac{1}{5.79} = 0.17,$$

$$\pi_1 = (1.67)\pi_0 = 0.29, \quad \text{and} \quad \pi_2 = \frac{(1.67)^2}{2!}\pi_0 = 0.24.$$

◇

7.5.3 Unlimited number of servers and M/M/∞

An unlimited number of servers completely eliminates the waiting time. Whenever a job arrives, there will always be servers available to handle it, and thus, the response time T consists of the service time only. In other words,

$$X = X_s, \quad R = S, \quad X_w = 0, \quad \text{and} \quad W = 0.$$

Have we seen such queuing systems?

Clearly, nobody can physically build an *infinite* number of devices. Having an unlimited number of servers simply means that the number of concurrent users is unlimited. For example, most internet service providers and most long-distance telephone companies allow virtually any number of concurrent connections; an unlimited number of people can watch a TV channel, listen to a radio station, or take a sun bath.

Sometimes a model with infinitely many servers is a reasonable approximation for a system where jobs typically don't wait and get their service immediately. This may be appropriate for a computer server, a grocery store, or a street with practically no traffic jams.

M/M/∞ queueing system

The definition and all the theory of M/M/k systems applies to M/M/∞ when we substitute $k=\infty$ or take a limit as the number of servers k goes to infinity. The number of jobs will always be less than the number of servers ($i<k$), and therefore, we always have $n=i$. That is, with i jobs in the system, exactly i servers are busy.

The number of jobs X in an M/M/∞ system has a transition probability matrix

$$P = \begin{pmatrix} 1-p_A & p_A & 0 & 0 & \ddots \\ p_S & 1{-}p_A{-}p_S & p_A & 0 & \ddots \\ 0 & \mathbf{2}p_S & 1{-}p_A{-}2p_S & p_A & \ddots \\ 0 & 0 & \mathbf{3}p_S & 1{-}p_A{-}3p_S & \ddots \\ 0 & 0 & 0 & \mathbf{4}p_S & \ddots \\ \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix} \qquad (7.15)$$

and its steady-state distribution is ... *doesn't it look familiar*?

Taking a limit of (7.14) as $k \to \infty$, we get

$$\pi_0 = \frac{1}{\displaystyle\sum_{i=0}^{\infty} \frac{r^i}{i!} + \lim_{k\to\infty} \frac{r^k}{(1-r/k)k!}} = \frac{1}{e^r + 0} = e^{-r},$$

and for all $x \geq 0$,

$$\pi_x = \frac{r^x}{x!}\,\pi_0 = e^{-r}\,\frac{r^x}{x!}.$$

We certainly recognize *Poisson distribution*! Yes, the number of concurrent jobs in an M/M/∞ system is Poisson with parameter $r = \lambda_A/\lambda_S$.

M/M/∞ system: steady-state distribution of the number of jobs

The number of jobs is **Poisson**(r)

$$\pi_x = \boldsymbol{P}\{X = x\} = e^{-r}\,\frac{r^x}{x!} \qquad (7.16)$$

$$\mathbf{E}(X) = \operatorname{Var}(X) = r$$

Example 7.8 (A POWERFUL SERVER). A certain powerful server can afford practically any number of concurrent users. The users connect to a server at random times, every 3 minutes, on the average, according to a Poisson counting process. Each user spends an Exponential amount of time on the server with an average of 1 hour and disconnects from it, independently of other users.

Such a description fits an M/M/∞ queuing system with

$$r = \mu_S/\mu_A = 60\,\text{min}\,/3\,\text{min} = 20.$$

The number of concurrent users is Poisson(20). We can use Table A3 in the Appendix to find out that

$$\boldsymbol{P}\{X = 0\} = 0.0000,$$

that is, the server practically never becomes idle.

Also, if an urgent message is sent to all the users, then $\mathbf{E}(X) = 20$ users, on the average, will see it immediately. Fifteen or more users will receive this message immediately with probability

$$\boldsymbol{P}\{X \geq 15\} = 1 - F(14) = 1 - 0.1049 = \underline{0.8951}.$$

◇

7.6 Simulation of queuing systems

Let us summarize what we have achieved in this chapter. We developed a theory and understood how to analyze and evaluate rather basic queuing systems: Bernoulli and M/M/k. We have covered the cases of one server and several servers, limited and unlimited capacity, but never considered such complications as customers' premature dropout, service interruption, non-Markov processes of arrivals and/or services, and distinction between different servers.

Most of our results were obtained from the Markov property of the considered queuing processes. For these systems, we derived a steady-state distribution of the number of concurrent jobs and computed the vital performance characteristics from it.

The only general result is the Little's Law of Section 7.2, which can be applied to any queuing system.

In practice, however, many queuing systems have a much more complex structure. Jobs may arrive according to a non-Poisson process; often the rate of arrivals changes during the day (there is a rush hour on highways and on the internet). Service times may have different distributions, and they are not always memoryless, thus the Markov property may not be satisfied. The number of servers may also change during the day (additional servers may turn on during rush hours). Some customers may get dissatisfied with a long waiting time and quit in the middle of their queue. And so on.

Queuing theory does not cover all the possible situations. On the other hand, we can simulate the behavior of almost any queuing system and study its properties by *Monte Carlo methods*.

Markov case

A queuing system is *Markov* only when its interarrival and service times are memoryless. Then the future can be predicted from the present without relying on the past (Section 6.2).

If this is the case, we compute the transition probability matrix as we did in (7.4), (7.6), (7.7), (7.11), (7.12), (7.13), or (7.15), and simulate the Markov chain according to Algorithm 6.1 on p. 173. To study long-term characteristics of a queuing system, the initial distribution of X_0 typically does not matter, so we may start this algorithm with 0 jobs in the system and immediately "switch on" the servers.

Even when the system is Markov, some interesting characteristics do not follow from its steady-state distribution directly. They can now be estimated from a Monte Carlo study by simulated long-run proportions and averages. For example, we may be interested in estimating the percentage of dissatisfied customers, the expected number of users during a period of time from 3 pm till 5 pm, etc. We may also compute various forecasts (Section 5.3.3).

General case

Monte Carlo methods of Chapter 5 let us simulate and evaluate rather complex queuing systems far beyond Bernoulli and M/M/k. As long as we know the distributions of interarrival and service times, we can generate the processes of arrivals and services. To assign jobs to servers, we keep track of servers that are available each time when a new job arrives. When all the servers are busy, the new job will enter a queue.

As we simulate the work of a queuing system, we keep records of events and variables that are of interest to us. After a large number of Monte Carlo runs, we average our records in order to estimate probabilities and means by long-run proportions and averages.

Example: simulation of a multiserver queuing system

Let us simulate one day, from 8 am till 10 pm, of a queuing system that has

- four servers;
- Gamma distributed service times with parameters given in the table,

Server	α	λ
I	6	0.3 min^{-1}
II	10	0.2 min^{-1}
III	7	0.7 min^{-1}
IV	5	1.0 min^{-1}

– a Poisson process of arrivals with the rate of 1 arrival every 4 min, independent of service times;

– random assignment of servers, when more than 1 server is available.

In addition, suppose that after 15 minutes of waiting, jobs withdraw from a queue if their service has not started.

For this system, we are interested to estimate the daily averages of:

– the total time each server was busy with jobs;

– the total number of jobs served by each server;

– the average waiting time;

– the maximum waiting time;

– the number of withdrawn jobs;

– the number of times a server was available immediately (this is also the number of jobs with no waiting time);

– the number of jobs remaining in the system at 10 pm.

We start by entering parameters of the system.

```
k  = 4;                      % number of servers
mu = 4;                      % mean interarrival time
alpha  = [6 10 7 5];         % parameters of service times
lambda = [0.3 0.2 0.7 1.0];
```

Then we initialize variables and start keeping track of arriving jobs. These are empty arrays so far. They get filled with each arriving job.

```
arrival = [ ];   % arrival time
start   = [ ];   % service starts
finish  = [ ];   % service finishes; departure time
server  = [ ];   % assigned server
j = 0;           % The job number is initialized
T = 0;           % arrival time of a new job
A = zeros(1,k);  % array of times when each server
                 % becomes available
```

The queuing system is ready to work! We start a “while”-loop over the number of arriving jobs. It will run until the end of the day, when arrival time T reaches 14 hours, or 840 minutes. The length of this loop, the total number of arrived jobs, is random.

```
while T < 840;             % until the end of the day
  j=j+1;                   % next job
  T = T-mu*log(rand);      % arrival time of job j
  arrival = [arrival T];
```

Generated in accordance with Example 5.10 on p. 119, the arrival time T is obtained by incrementing the previous arrival time by an Exponential interarrival time (Example 5.10). Generated within this while-loop, the last job actually arrives after 10 pm. You may either accept it or delete it.

Next, we need to assign the new job j to a server, following the rule of random assignment. There are two cases here: either all servers are busy at the arrival time T, or some servers are available.

```
Nfree = sum( A < T );   % number of free servers at time T
u = 1;                  % u = server that will take job j

if Nfree == 0;          % If all servers are busy at time T ...
  for v=2:k;
    if A(v)<A(u);       % Find the server that gets
      u = v;            % available first and assign
    end;                % job j to it
  end;

  if A(u)-T > 15;       % If job j waits more than 15 min,
    start=[start 0];    % then it withdraws at time T+15.
    finish=[finish T+15];
    u = 0;              % No server is assigned.
  else                  % If job j doesn't withdraw ...
    start = [start A(u)];
  end;

else                    % If there are servers available ...
  u = ceil(rand*k);     % Generate a random server, from 1 to k
  while A(u) > T;       % Random search for an available
    u = ceil(rand*k);   % server
  end;
start = [start T];      % Service starts immediately at T
end;
server = [server u];
```

The server (u) for the new job (j) is determined. We can now generate its service time S from the suitable Gamma distribution and update our records.

```
  if u>0;                  % if job j doesn't withdraw ...
    S     = sum( -1/lambda(u) * log(rand(alpha(u),1) ) );
    finish = [finish start(j)+S];
    A(u)   = start(j) + S; % This is the time when server u
  end;                     % will now become available
end;                       % End of the while-loop
disp([(1:j)' arrival' start' finish' server'])
```

The day (our "while"-loop) has ended. The last command displayed the following records,

Job	Arrival time	Service starts	Service ends	Server
...	.....	.....	.....	.....

From this table, all variables of our interest can be computed.

```
for u=1:k;
  Twork(u) = sum((server==u).*(finish-start));
                     % The total working time for server u
  Njobs(u) = sum(server==u);  % number of jobs served by u
end;
Wmean = mean(start-arrival);  % average waiting time
Wmax = max(start-arrival);    % maximum waiting time
Nwithdr = sum(server==0);     % number of withdrawn jobs
Nav = sum(start-arrival < 0.00001);
                        % number of jobs that did not wait
Nat10 = sum(finish > 840);    % number of jobs at 10 pm
```

We have computed all the required quantities for one day. Put this code into a "do"-loop and simulate a large number of days. The required expected values are then estimated by averaging variables `Wmean`, `Nwithdr`, etc., over all days.

Summary and conclusions

Queuing systems are service facilities designed to serve randomly arriving jobs.

Several classes of queuing systems were studied in this chapter. We discussed discrete-time and continuous-time, Markov and non-Markov, one-server and multiserver queuing processes with limited and unlimited capacity.

Detailed theory was developed for Markov systems, where we derived the distribution of the number of concurrent jobs and obtained analytic expressions for a number of vital characteristics. These results are used to evaluate performance of a queuing system, forecast its future efficiency when parameters change, and see if it is still functional under the new conditions.

Performance of more complicated and advanced queuing systems can be evaluated by Monte Carlo methods. One needs to simulate arrivals of jobs, assignment of servers, and service times and to keep track of all variables of interest. A sample computer code for such a simulation is given.

Questions and exercises

7.1. Customers arrive at the ATM at the rate of 10 customers per hour and spend 2 minutes, on average, on all the transactions. This system is modeled by the single-server Bernoulli queuing process with 10-second frames. Write the transition probability matrix for the number of customers at the ATM at the end of each frame.

7.2. Consider Bernoulli single-server queuing process with an arrival rate of 2 jobs per minute, a service rate of 4 jobs per minute, frames of 0.2 minutes, and a capacity limited by 2 jobs. Compute the steady-state distribution of the number of jobs in the system.

7.3. Performance of a car wash center is modeled by the single-server Bernoulli queuing process with 2-minute frames. The cars arrive every 10 minutes, on the average. The average service time is 6 minutes. Capacity is unlimited. If there are no cars at the center at 10 am, compute the probability that one car will be washed and another car will be waiting at 10:04 am.

7.4. Mary has a telephone with two lines that allows her to talk with a friend and have at most one other friend on hold. On the average, she gets 10 calls every hour, and an average conversation takes 2 minutes. Assuming a single-server limited-capacity Bernoulli queuing process with 1-minute frames, compute the fraction of time Mary spends using her telephone.

7.5. A customer service representative can work with one person at a time and have at most one other customer waiting. Compute the steady-state distribution of the number of customers in this queuing system at any time, assuming that customers arrive according to a Bernoulli counting process with 3-minute frames and the average interarrival time of 10 minutes, and the average service takes 15 minutes.

7.6. Performance of a telephone with 2 lines is modeled by the Bernoulli single-server queuing process with limited capacity ($C = 2$). If both lines of a telephone are busy, the new callers receive a busy signal and cannot enter the queue. On the average, there are 5 calls per hour, and the average call takes 20 minutes. Compute the steady-state probabilities using four-minute frames.

7.7. Messages arrive at an electronic mail server at the average rate of 4 messages every 5 minutes. Their number is modeled by a Binomial counting process.

(a) What frame length makes the probability of a new message arrival during a given frame equal 0.05?

(b) Suppose that 50 messages arrived during some 1-hour period. Does this indicate that the arrival rate is on the increase? Use frames computed in (a).

7.8. Jobs arrive at the server at the rate of 8 jobs per hour. The service takes 3 minutes, on the average. This system is modeled by the single-server Bernoulli queuing process with 5-second frames and capacity limited by 3 jobs. Write the transition probability matrix for the number of jobs in the system at the end of each frame.

7.9. For an M/M/1 queuing process with the arrival rate of 5 min^{-1} and the average service time of 4 seconds, compute

(a) the proportion of time when there are exactly 2 customers in the system

(b) the expected response time (the expected time from the arrival till the departure)

7.10. Jobs are sent to a printer at random times, according to a Poisson process of arrivals, with a rate of 12 jobs per hour. The time it takes to print a job is an Exponential random variable, independent of the arrival time, with the average of 2 minutes per job.

(a) A job is sent to a printer at noon. When is it expected to be printed?

(b) How often does the total number of jobs in a queue and currently being printed exceeds 2?

7.11. Performance of an ATM is modeled by an M/M/1 queue with the arrival rate of 20 customers per hour and the average service time of 2 minutes.

(a) A customer arrives at 8 pm. What is the expected waiting time?

(b) What is the probability that nobody is using the ATM at 3 pm?

7.12. Customers come to a teller's window according to a Poisson process with a rate of 10 customers every 25 minutes. Service times are Exponential. The average service takes 2 minutes. Compute

(a) the average number of customers in the system and the average number of customers waiting in a line

(b) the fraction of time when the teller is busy with a customer

(c) the fraction of time when the teller is busy and at least five other customers are waiting in a line

7.13. For an M/M/1 queuing system with the average interarrival time of 5 minutes and the average service time of 3 minutes, compute

(a) the expected response time

(b) the fraction of time when there are fewer than 2 jobs in the system

(c) the fraction of customers who have to wait before their service starts

7.14. Jobs arrive at the service facility according to a Poisson process with the average interarrival time of 4.5 minutes. A typical job spends a Gamma distributed time with parameters $\alpha = 12$, $\lambda = 5$ min^{-1} in the system and leaves.

(a) Compute the average number of jobs in the system at any time.

(b) Suppose that only 20 jobs arrived during the last three hours. Is this evidence that the expected interarrival time has increased?

7.15. Cars arrive at an ATM according to a Poisson process with the average rate of 1 car every 10 minutes. The time each customer spends on bank operations is Exponential with the average time of 3 minutes. When a customer uses ATM, the other arrived customers stay in a line waiting for their turn. Compute

(a) the expected number of cars in the line at any time

(b) the proportion of time when nobody is using this ATM

(c) the expected time each customer spends at the ATM, from arrival till departure

7.16. Messages arrive at an electronic mail server according to a Poisson process with the average frequency of 5 messages per minute. The server can process only one message at a time, and messages are processed on a "first come – first serve" basis. It takes an Exponential amount of time X to process any text message, *plus* an Exponential amount of time Y, independent of X, to process attachments (if there are any). Forty percent of messages contain attachments. Compute the expected response time of this server, if $\mathbf{E}(X) = 2$ seconds, and $\mathbf{E}(Y) = 7$ seconds.

7.17. Consider a hot-line telephone that has no second line. When the telephone is busy, the new callers get a busy signal. People call at the average rate of 2 calls per minute. The average duration of a telephone conversation is 1 minute. The system behaves like a Bernoulli single-server queuing process with a frame size of 1 second.

(a) Compute the steady-state distribution for the number of concurrent jobs.

(b) What is the probability that more than 150 people attempted to call this number between 2 pm and 3 pm?

7.18. On the average, every 6 minutes a customer arrives at an M/M/k queuing system, spends an average of 20 minutes there, and departs. What is the mean number of customers in the system at any given time?

7.19. Verify the Little's Law for the M/M/1 queuing system and its components.

7.20. (Computer project) A multiserver system (computer lab, customer service, telephone company) consists of $n = 4$ servers (computers, customer service representatives, telephone cables). Every server is able to process any job, but some of them work faster than the others. The service times are distributed according to the table.

Server	Distribution	Parameters
I	Gamma	$\alpha = 7$, $\lambda = 3$ min^{-1}
II	Gamma	$\alpha = 5$, $\lambda = 2$ min^{-1}
III	Exponential	$\lambda = 0.3$ min^{-1}
IV	Uniform	$a = 4$ min, $b = 9$ min

The jobs (customers, telephone calls) arrive to the system at random times, independently of each other, according to a Poisson process. The average interarrival time is 2 minutes. If a job arrives, and there are free servers available, then the job is *equally likely* to be processed by any of the available servers. If no servers are available at the time of arrival, the job enters a queue. After waiting for 10 minutes, if the service has not started, the job leaves the system. The system works 10 hours a day, from 8 am till 6 pm.

Run at least 1000 Monte Carlo simulations and estimate the following quantities:

(a) the expected waiting time for a randomly selected job
(b) the expected response time
(c) the expected length of a queue, when a new job arrives
(d) the expected *maximum* waiting time during a 10-hour day
(e) the expected *maximum* length of a queue during a 10-hour day
(f) the probability that at least one server is available, when a job arrives
(g) the probability that at least two servers are available, when a job arrives
(h) the expected number of jobs processed by each server
(i) the expected time each server is idle during the day
(j) the probability that no jobs are waiting or being processed at 6pm
(k) the expected percentage of jobs that left the queue prematurely

CHAPTER 8

Introduction to Statistics

The first seven chapters of this book taught us to analyze problems and systems involving uncertainty, to find probabilities, expectations, and other characteristics for a variety of situations, and to produce forecasts that may lead to important decisions.

What was given to us in all these problems? Ultimately, *we needed to know the distribution and its parameters*, in order to compute probabilities or at least to estimate them by means of Monte Carlo. Often the distribution may not be given, and we learned how to fit the right model, say, Binomial, Exponential, or Poisson, given the type of variables we deal with. In any case, parameters of the fitted distribution had to be reported to us explicitly, or they had to follow directly from the problem.

This, however, is rarely the case in practice. Only sometimes the situation may be under our control, where, for example, produced items have predetermined specifications, and therefore, one knows parameters of their distribution.

Much more often *parameters are not known.* Then, how can one apply the knowledge of Chapters 1–7 and compute probabilities? The answer is simple: *we need to collect data.* A properly collected sample of data can provide rather sufficient information about parameters of the observed system. In the next sections and chapters, we learn how to use this sample

- to visualize data, understand the patterns, and make quick statements about the system's behavior;
- to characterize this behavior in simple terms and quantities;
- to estimate the distribution parameters;
- to assess reliability of our estimates;
- to test statements about parameters and the entire system;
- to understand relations among variables;
- to fit suitable models and use them to make forecasts.

8.1 Population and sample, parameters and statistics

Data collection is a crucially important step in Statistics. We use the collected and observed *sample* to make statements about a much larger set — the *population*.

DEFINITION 8.1

> A **population** consists of all units of interest. Most of them are not observed. A **sample** consists of observed units collected from the population. It is used to make statements about the population.

In real problems, we would like to make statements about the population. To compute probabilities, expectations, and make optimal decisions under uncertainty, we need to know the population *parameters*. However, the only way to know these parameters is to measure the entire population, i.e., to conduct a *census*.

Instead of a census, we may collect data in a form of a random sample from a population (Figure 8.1). This is our data. We can measure them, perform calculations, and *estimate* the unknown parameters of the population up to a certain *measurable* degree of accuracy.

<u>Notation</u>

θ	$=$	population parameter
$\hat{\theta}$	$=$	its estimator, obtained from a sample

Example 8.1 (Customer satisfaction). For example, even if 80% of users are satisfied with their internet connection, it does not mean that exactly 8 out of 10 customers in your observed sample are satisfied. As we can see from Table A2 in the Appendix, with probability 0.0328, only a half of ten sampled customers are satisfied. In other words, there is a 3% chance for a random sample to suggest that contrary to the population parameter, no more than 50% of users are satisfied. ◇

This example shows that although with low probability, a sample may give a rather misleading information about the population. *Sampling errors are inevitable.*

Sampling and non-sampling errors

Sampling and non-sampling errors refer to any discrepancy between a sample and a population.

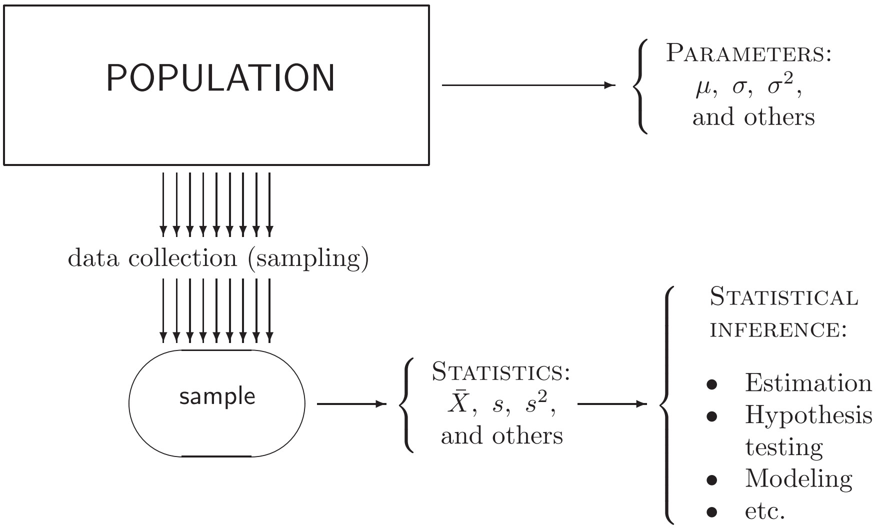


Figure 8.1 *Population parameters and sample statistics.*

Sampling errors are caused by the mere fact that only a sample, a portion of a population, is observed. For most of reasonable statistical procedures, sampling errors decrease (and converge to zero) as the sample size increases.

Non-sampling errors are caused by inappropriate sampling schemes or wrong statistical techniques. Often no wise statistical techniques can rescue a poorly collected sample of data.

Look at some examples of wrong sampling practices.

Example 8.2 (SAMPLING FROM A WRONG POPULATION). To evaluate the work of a Windows help desk, a survey of social science students of some university is conducted. This sample poorly represents the whole population of Windows users. For example, computer science students and especially computer professionals may have a totally different opinion about the Windows help desk. ◇

Example 8.3 (DEPENDENT OBSERVATIONS). Comparing two brands of notebooks, a senior manager asks all employees of her group to state which notebook they like better. Again, these employees are not randomly selected from the population of all users of these notebooks. Also, their opinions are likely to be *dependent*. Working together, these people often communicate, and their points of view affect each other. Dependent observations do not necessarily cause non-sampling errors, if they are handled properly. The fact is, in such cases, we cannot assume independence. ◇

Example 8.4 (NOT EQUALLY LIKELY). A survey among passengers of some airline is conducted in the following way. A sample of random flights is selected from a list, and ten passengers on each of these flights are also randomly chosen. Each sampled passenger is asked to fill a questionnaire. Is this a representative sample?

Suppose Mr. X flies only once a year whereas Ms. Y has business trips twice a month. Obviously, Ms. Y has a much higher chance to be sampled than Mr. X. Unequal probabilities have to be taken into account, otherwise a non-sampling error will inevitably occur. ◇

In this book, we focus on *simple random sampling*.

DEFINITION 8.2

Simple random sampling is a sampling design where units are collected from the entire population independently of each other, all being equally likely to be sampled.

Observations collected by means of a simple random sampling design are **iid** (independent, identically distributed) random variables.

Example 8.5. To evaluate its customers' satisfaction, a bank makes a list of all the accounts. A Monte Carlo method is used to choose a random number between 1 and N, where N is the total number of bank accounts. Say, we generate a Uniform(0,N) variable X and sample an account number $\lceil X \rceil$ from the list. Similarly, we choose the second account, uniformly distributed among the remaining $N-1$ accounts, etc., until we get a sample of the desired size n. This is a simple random sample. ◇

Obtaining a good, representative random sample is rather important in Statistics. Although we have only a portion of the population in our hands, a rigorous sampling design followed by a suitable statistical inference allows to estimate parameters and make statements with a certain measurable degree of confidence.

8.2 Simple descriptive statistics

Suppose a good random sample

$$S = (X_1, X_2, \ldots, X_n)$$

has been collected. For example, to evaluate effectiveness of a processor for a certain type of tasks, we recorded the CPU time for $n = 30$ randomly chosen jobs (in seconds),

70	36	43	69	82	48	34	62	35	15	
59	139	46	37	42	30	55	56	36	82	(8.1)
38	89	54	25	35	24	22	9	56	19	

What information do we get from this collection of numbers?

We know that X, the CPU time of a random job, is a random variable, and it does not have to be one of the observed thirty. We'll use the collected data to describe the distribution of X.

Simple **descriptive statistics** measuring the location, spread, variability, and other characteristics can be computed immediately. In this section, we discuss the following statistics,

- **mean**, measuring the average value of a sample;
- **median**, measuring the central value;
- **quantiles** and **quartiles**, showing where certain portions of a sample are located;
- **variance**, **standard deviation**, and **interquartile range**, measuring variability and spread of data.

Each statistic is a random variable because it is computed from random data. It has a so-called *sampling distribution*.

Each statistic estimates the corresponding population parameter and adds certain information about the distribution of X, the variable of interest.

We used similar methods in Section 5.3.2, where we estimated parameters from Monte Carlo samples obtained by computer simulations. Here we estimate parameters and make conclusions based on real, not simulated, data.

8.2.1 Mean

Sample mean $\bar{X}$ estimates the population mean $\mu = \mathbf{E}(X)$.

DEFINITION 8.3

Sample mean $\bar{X}$ is the arithmetic average,

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n}$$

Naturally, being the average of sampled observations, $\bar{X}$ estimates the average value of the whole distribution of X. Computed from random data, $\bar{X}$ does not necessarily equal μ, however, we would expect it to converge to μ when a large sample is collected.

Sample means possess a number of good properties. They are unbiased, consistent, and asymptotically normal.

Remark: This is true if the population has finite mean and variance, which is the case for almost all the distributions in this book (see, however, Example 3.20 on p. 66).

Unbiasedness

DEFINITION 8.4

An estimator $\hat{\theta}$ is **unbiased** for a parameter θ if its expectation equals the parameter,

$$\mathbf{E}(\hat{\theta}) = \theta$$

for all possible values of θ.

Unbiasedness means that in a long run, collecting a large number of samples and computing $\hat{\theta}$ from each of them, on the average we hit the unknown parameter θ exactly. In other words, unbiased estimators neither underestimate nor overestimate the parameter.

Sample mean estimates μ unbiasedly because its expectation is

$$\mathbf{E}(\bar{X}) = \mathbf{E}\left(\frac{X_1 + \ldots + X_n}{n}\right) = \frac{\mathbf{E}X_1 + \ldots + \mathbf{E}X_n}{n} = \frac{n\mu}{n} = \mu.$$

Consistency

DEFINITION 8.5

An estimator $\hat{\theta}$ is **consistent** for a parameter θ if the probability of its sampling error of any size converges to 0 as the sample size increases to infinity. Stating it rigorously,

$$\boldsymbol{P}\left\{|\hat{\theta} - \theta| > \varepsilon\right\} \to 0 \text{ as } n \to \infty$$

for any $\varepsilon > 0$. That is, when we estimate θ from a large sample, the estimation error $|\hat{\theta} - \theta|$ is unlikely to exceed ε, and it does it with smaller and smaller probabilities as we increase the sample size.

Consistency of $\bar{X}$ follows directly from Chebyshev's inequality on p. 58.

To use this inequality, we find the variance of $\bar{X}$,

$$\text{Var}(\bar{X}) = \text{Var}\left(\frac{X_1 + \ldots + X_n}{n}\right) = \frac{\text{Var}X_1 + \ldots + \text{Var}X_n}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}. \tag{8.2}$$

Then, using Chebyshev's inequality for the random variable $\bar{X}$, we get

$$\boldsymbol{P}\left\{|\bar{X} - \mu| > \varepsilon\right\} \leq \frac{\text{Var}(\bar{X})}{\varepsilon^2} = \frac{\sigma^2/n}{\varepsilon^2} \to 0,$$

as $n \to \infty$.

Thus, a sample mean is *consistent*. Its sampling error will be small with a higher and higher probability, as we collect larger and larger samples.

Asymptotic Normality

By the Central Limit Theorem, the sum of observations, and therefore, the sample mean have approximately Normal distribution if they are computed from a large sample. That is, the distribution of

$$Z = \frac{\bar{X} - \mathbf{E}\bar{X}}{\text{Std}\bar{X}} = \frac{\bar{X} - \mu}{\sigma\sqrt{n}}$$

converges to Standard Normal as $n \to \infty$. This property is called **Asymptotic Normality**.

Example 8.6 (CPU TIMES). Looking at the CPU data on p. 225, we estimate the average (expected) CPU time μ by

$$\bar{X} = \frac{70 + 36 + \ldots + 56 + 19}{30} = \frac{1447}{30} = 48.2333.$$

We may conclude that the actual mean CPU time of all the jobs is "near" 48.2333 seconds. ◇

NOTATION	μ	=	population mean
	$\bar{X}$	=	sample mean, estimator of μ
	σ	=	population standard deviation
	s	=	sample standard deviation, estimator of σ
	σ^2	=	population variance
	s^2	=	sample variance, estimator of σ

8.2.2 Median

One disadvantage of a sample mean is its *sensitivity to extreme observations*. For example, if the first job in our sample is unusually heavy, and it takes 30 minutes to get processed instead of 70 seconds, this one extremely large observation shifts the sample mean from 48.2333 sec to 105.9 sec. Can we call such an estimator "reliable"?

Another simple measure of location is a *sample median* which estimates the *population median*. It is much less sensitive than the sample mean.

DEFINITION 8.6

Median means a "central" value.

Sample median $\hat{M}$ is a number that is exceeded by at most a half of observations and is preceded by at most a half of observations.

Population median M is a number that is exceeded with probability no greater than 0.5 and is preceded with probability no greater than 0.5. That is, M is such that

$$\left\{ \begin{array}{lcl} \boldsymbol{P}\{X > M\} & \leq & 0.5 \\ \boldsymbol{P}\{X < M\} & \leq & 0.5 \end{array} \right.$$

Understanding the shape of a distribution

Comparing the mean and the median, one can tell whether the distribution of X is right-skewed, left-skewed, or symmetric (Figure 8.2):

$$\begin{array}{lcl} \text{Symmetric distribution} & \Rightarrow & M = \mu \\ \text{Right-skewed distribution} & \Rightarrow & M < \mu \\ \text{Left-skewed distribution} & \Rightarrow & M > \mu \end{array}$$

Computation of a population median

For continuous distributions, computing a population median reduces to solving one equation:

$$\left\{ \begin{array}{lclcl} \boldsymbol{P}\{X > M\} & = & 1 - F(M) & \leq & 0.5 \\ \boldsymbol{P}\{X < M\} & = & F(M) & \leq & 0.5 \end{array} \right. \quad \Rightarrow \quad F(M) = 0.5.$$

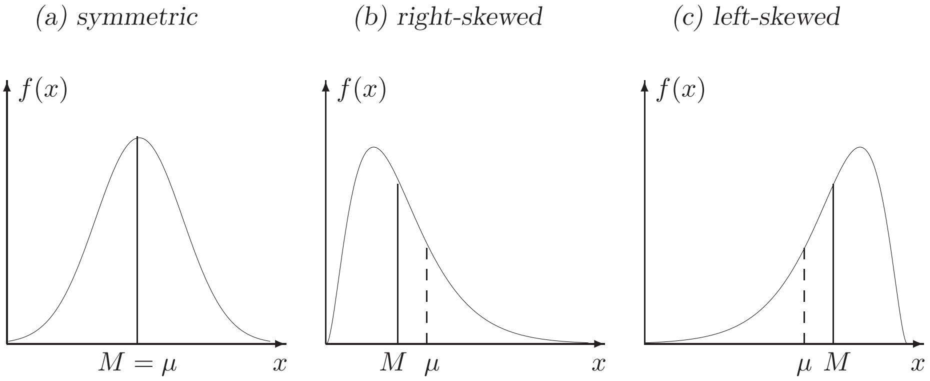


Figure 8.2 *A mean μ and a median M for distributions of different shapes.*

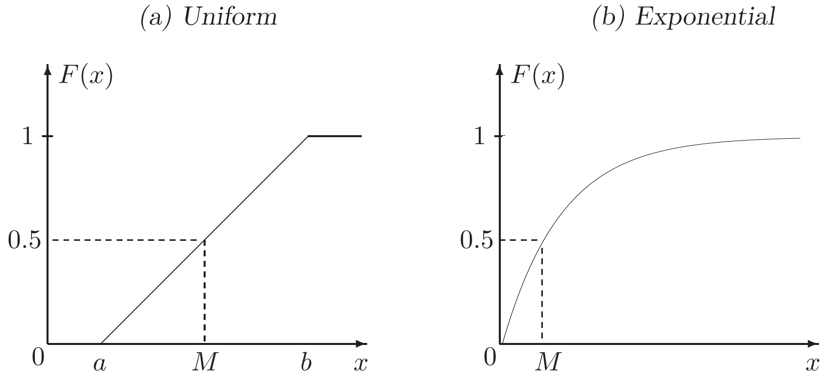


Figure 8.3 *Computing medians of continuous distributions.*

Example 8.7 (Uniform, Figure 8.3a). Uniform(a,b) distribution has a cdf

$$F(x) = \frac{x-a}{b-a} \text{ for } 0 < x < 1.$$

Solving the equation $F(M) = (M-a)/(b-a) = 0.5$, we get

$$M = \frac{a+b}{2}.$$

It coincides with the mean because the Uniform distribution is symmetric. ◇

Example 8.8 (Exponential, Figure 8.3b). Exponential(λ) distribution has a cdf

$$F(x) = 1 - e^{-\lambda x} \text{ for } x > 0.$$

Solving the equation $F(M) = 1 - e^{-\lambda x} = 0.5$, we get

$$M = \frac{\ln 2}{\lambda} = \frac{0.6931}{\lambda}.$$

We know that $\mu = 1/\lambda$ for Exponential distribution. Here the median is smaller than the mean because Exponential distribution is right-skewed. $\Diamond$

For discrete distributions, equation $F(x) = 0.5$ has either the whole interval of roots or no roots at all (see Figure 8.4).

In the first case, any number in this interval, excluding the ends, is a median. Notice that the median in this case is not unique (Figure 8.4a). Often the middle of this interval is reported as the median.

In the second case, the smallest x with $F(x) \geq 0.5$ is the median. It is the value of x where the cdf jumps over 0.5 (Figure 8.4b).

Example 8.9 (Symmetric Binomial, Figure 8.4a). Consider Binomial distribution with $n = 5$ and $p = 0.5$. From Table A2, we see that for all $2 < x < 3$,

$$\begin{cases} \boldsymbol{P}\{X < x\} & = & F(2) & = & 0.5 \\ \boldsymbol{P}\{X > x\} & = & 1 - F(2) & = & 0.5 \end{cases}$$

By Definition 8.6, any number of the interval (2,3) is a median.

This result agrees with our intuition. With $p = 0.5$, successes and failures are equally likely. Pick, for example, $x = 2.4$ in the interval (2,3). Having fewer than 2.4 successes (i.e., at most two) has the same chance as having fewer than 2.4 failures (i.e., at least 3 successes). Therefore, $X < 2.4$ with the same probability as $X > 2.4$, which makes $x = 2.4$ a central value, a median. We can say that $x = 2.4$ (and any other x between 2 and 3) splits the distribution into two equal parts. Then, it is a median. $\Diamond$

Example 8.10 (Asymmetric Binomial, Figure 8.4b). For the Binomial distribution with $n = 5$ and $p = 0.4$,

$$\begin{array}{lll} F(x) < 0.5 & \text{for} & x < 2 \\ F(x) > 0.5 & \text{for} & x \geq 2 \end{array}$$

but there is not value of x where $F(x) = 0.5$. Then, $M = 2$ is the median.

Seeing a value on either side of $x = 2$ has probability less than 0.5, which makes $x = 2$ a center. $\Diamond$

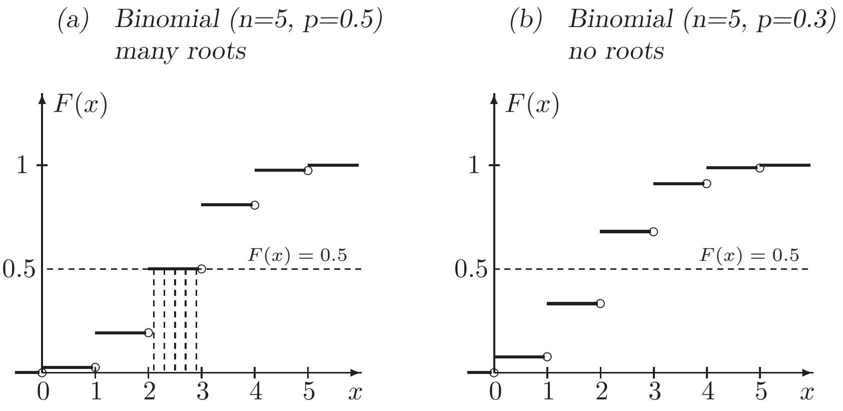


Figure 8.4 *Computing medians of discrete distributions.*

Computing sample medians

A sample is always discrete, it consists of a finite number of observations. Then, computing a sample median is similar to the case of discrete distributions.

In simple random sampling, all observations are equally likely, and thus, equal probabilities on each side of a median translate into an equal number of observations.

Again, there are two cases, depending on the sample size n.

Sample median

If n is odd, the $\left(\frac{n+1}{2}\right)$-th smallest observation is a median.

If n is even, any number between the $\left(\frac{n}{2}\right)$-th smallest and the $\left(\frac{n+2}{2}\right)$-th smallest observations is a median.

Example 8.11 (MEDIAN CPU TIME). Let's compute the median of $n = 30$ CPU times from the data on p. 225.

First, order the data,

$$\begin{array}{rrrrrrrrrr}
9 & 15 & 19 & 22 & 24 & 25 & 30 & 34 & 35 & 35 \\
36 & 36 & 37 & 38 & \mathbf{42} & \mathbf{43} & 46 & 48 & 54 & 55 \\
56 & 56 & 59 & 62 & 69 & 70 & 82 & 82 & 89 & 139
\end{array} \tag{8.3}$$

Next, since $n = 30$ is even, find $n/2 = 15$-th smallest and $(n+2)/2 = 16$-th smallest observations. These are 42 and 43. Any number between them is a sample median (typically reported as 42.5). ◇

We see why medians are not sensitive to extreme observations. If in the previous example, the first CPU time happens to be 30 minutes instead of 70 seconds, it does not affect the sample median at all!

Sample medians are easy to compute. In fact, no computations are needed, only the ordering. If you are driving (and only if you find it safe!), here is a simple experiment that you can conduct yourself.

Example 8.12 (MEDIAN SPEED ON A HIGHWAY). How can you measure the median speed of cars on a multilane road without a radar? It's very simple. Adjust your speed so that a half of cars overtake you, and you overtake the other half. Then you are driving with the median speed. ◇

8.2.3 Quantiles, percentiles, and quartiles

Generalizing the notion of a median, we replace 0.5 in its definition on p. 228 by some $0 < p < 1$.

DEFINITION 8.7

A p-**quantile** of a population is such a number x that solves inequalities

$$\left\{ \begin{array}{lcc} \boldsymbol{P}\{X < x\} & \leq & p \\ \boldsymbol{P}\{X > x\} & \leq & 1-p \end{array} \right.$$

A **sample** p**-quantile** is any number that exceeds at most $100p\%$ of the sample, and is exceeded by at most $100(1-p)\%$ of the sample.

A γ-**percentile** is (0.01γ)-quantile.

First, second, and third **quartiles** are the 25th, 50th, and 75th percentiles. They split a population or a sample into four equal parts.

A **median** is at the same time a 0.5-quantile, 50th percentile, and 2nd quartile.

Notation

q_p	$=$	population p-quantile
$\hat{q}_p$	$=$	sample p-quantile, estimator of q_p
π_γ	$=$	population γ-percentile
$\hat{\pi}_\gamma$	$=$	sample γ-percentile, estimator of π_γ
Q_1, Q_2, Q_3	$=$	population quantiles
$\hat{Q}_1, \hat{Q}_2, \hat{Q}_3$	$=$	sample quartiles, estimators of $Q_1, Q_2,$ and Q_3
M	$=$	population median
$\hat{M}$	$=$	sample median, estimator of M

Quantiles, quartiles, and percentiles are related as follows.

Quantiles, quartiles, percentiles

$$\boxed{\begin{array}{c} q_p = \pi_{100p} \\ Q_1 = \pi_{25} = q_{1/4} \qquad Q_3 = \pi_{75} = q_{3/4} \\ M = Q_2 = \pi_{50} = q_{1/2} \end{array}}$$

Sample statistics are of course in a similar relation.

Computing quantiles is very similar to computing medians.

Example 8.13 (Sample quartiles). Let us compute the 1st and 3rd quartiles of CPU times. Again, we look at the ordered sample

9	15	19	22	24	25	30	**34**	35	35
36	36	37	38	42	43	46	48	54	55
56	56	**59**	62	69	70	82	82	89	139

First quartile $\hat{Q}_1$. For $p = 0.25$, we find that 25% of the sample equals $np = 7.5$, and 75% of the sample is $n(1-p) = 22.5$ observations. From the ordered sample, we see that only the 8th element, 34, has no more than 7.5 observations to the left and no more than 22.5 observations to the right of it. Hence, $\hat{Q}_1 = 34$.

Third quartile $\hat{Q}_3$. Similarly, the third sample quartile is the 23rd smallest element, $\hat{Q}_3 = 59$. ◇

Example 8.14 (Calculating factory warranties from population percentiles). A computer maker sells extended warranty on the produced computers. It agrees to issue a warranty for x years if it knows that only 10% of computers will fail before the warranty expires. It is known from past experience that lifetimes of these computers have Gamma distribution with $\alpha = 60$ and $\lambda = 5$ years^{-1}. Compute x and advise the company on the important decision under uncertainty about possible warranties.

Solution. We just need to find the tenth percentile of the specified Gamma distribution and let

$$x = \pi_{10}.$$

As we know from Section 4.3, being a sum of Exponential variables, a Gamma variable is approximately Normal for large $\alpha = 60$. Using (4.13), compute

$$\begin{aligned} \mu &= \alpha/\lambda = 12, \\ \sigma &= \sqrt{\alpha/\lambda^2} = 1.55. \end{aligned}$$

From Table A4, the 10th percentile of a standardized variable

$$Z = \frac{X - \mu}{\sigma}$$

equals (-1.28) (find the probability closest to 0.10 in the table and read the corresponding value of z). Unstandardizing it, we get

$$x = \mu + (-1.28)\sigma = 12 - (1.28)(1.55) = \underline{10.02}.$$

Thus, the company can issue a 10-year warranty rather safely.

Remark: Of course, one does not have to use Normal approximation in the last example. A number of computer packages have built-in commands for the exact evaluation of probabilities and quantiles. For example, in recent versions of MATLAB, the 10th percentile of Gamma$(\alpha = 60, \lambda = 5)$ distribution can be obtained by the command

```
gaminv(0.10, 60, 1/5)
```

◇

8.2.4 Variance and standard deviation

Statistics introduced in the previous sections showed where the average value and certain percentages of a population are located. Now we are going to measure *variability* of our variable, how unstable the variable can be, and how much the actual value can differ from its expectation. As a result, we'll be able to assess reliability of our estimates and accuracy of our forecasts.

DEFINITION 8.8

For a sample $(X_1, X_2, \ldots, X_n)$, a **sample variance** is defined as

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2. \tag{8.4}$$

It measures variability among observations and estimates the population variance $\sigma^2 = \mathrm{Var}(X)$.

Sample standard deviation is a square root of a sample variance,

$$s = \sqrt{s^2}.$$

It measures variability in the same units as X and estimates the population standard deviation $\sigma = \mathrm{Std}(X)$.

Both population and sample variances are measured in squared units (in^2, sec^2, \$^{2}, etc.). Therefore, it is convenient to have standard deviations that are comparable with our variable of interest, X.

The formula for s^2 follows the same idea as that for σ^2. It is also the average squared deviation from the mean, this time computed for a sample. Like σ^2, sample variance measures how far the actual values of X are from their average.

Computation

Often it is easier to compute the sample variance using another formula,

$$\textbf{Sample variance} \quad \boxed{s^2 = \frac{\sum\limits_{i=1}^{n} X_i^2 - n\bar{X}^2}{n-1}.} \tag{8.5}$$

Remark: Expressions (8.4) and (8.5) are equivalent because

$$\begin{aligned} \sum \left(X_i - \bar{X}\right)^2 &= \sum X_i^2 - 2\bar{X} \sum X_i + \sum \bar{X}^2 \\ &= \sum X_i^2 - 2\bar{X}\left(n\bar{X}\right) + n\bar{X}^2 = \sum X_i^2 - n\bar{X}^2. \end{aligned}$$

When $X_1, \ldots, X_n$ are integers, but $(X_1 - \bar{X}), \ldots, (X_n - \bar{X})$ are fractions, it may be easier to use (8.5). However, $(X_n - \bar{X})$ are generally smaller in magnitude, and thus, we'd rather use (8.4) if $X_1, \ldots, X_n$ are rather large numbers.

Example 8.15 (CPU TIME, CONTINUED). For the data in (8.1) on p. 225, we have computed $\bar{X} = 48.2333$. Following Definition 8.8, we can compute the sample variance as

$$s^2 = \frac{(70 - 48.2333)^2 + \ldots + (19 - 48.2333)^2}{29} = \frac{20,391}{29} = 703.1506 \text{ (sec}^2\text{)}.$$

Alternatively, using (8.5),

$$s^2 = \frac{70^2 + \ldots + 19^2 - (30)(48.2333)^2}{29} = \frac{90,185 - 69,794}{29} = 703.1506 \text{ (sec}^2\text{)}.$$

The sample standard deviation is

$$s = \sqrt{703.1506} = 26.1506 \text{ (sec}^2\text{)}.$$

We can use these results, for example, as follows. Since $\bar{X}$ and s estimate the population mean and standard deviation, we can make a claim that at least 8/9 of all tasks require less than

$$\bar{X} + 3s = 127.78 \text{ seconds} \tag{8.6}$$

of CPU time. We used Chebyshev's inequality (3.8) to derive this (also see Exercise 8.3). ◇

A seemingly strange coefficient $\left(\frac{1}{n-1}\right)$ ensures that s^2 is an **unbiased** estimator of σ^2.

Proof: Let us prove the unbiasedness of s^2.

Case 1. Suppose for a moment that the population mean $\mu = \mathbf{E}(X) = 0$. Then

$$\mathbf{E}X_i^2 = \text{Var}X_i = \sigma^2,$$

and by (8.2),

$$\mathbf{E}\bar{X}^2 = \text{Var}\bar{X} = \sigma^2/n.$$

Then,

$$\mathbf{E}s^2 = \frac{\mathbf{E}\sum X_i^2 - n\,\mathbf{E}\bar{X}^2}{n-1} = \frac{n\sigma^2 - \sigma^2}{n-1} = \sigma^2.$$

Case 2. If $\mu \neq 0$, consider auxiliary variables $Y_i = X_i - \mu$. Variances don't depend on constant shifts (see (3.7), p. 56), therefore, Y_i have the same variance as X_i. Their sample variances are equal too,

$$s_Y^2 = \frac{\sum\left(Y_i - \bar{Y}\right)^2}{n-1} = \frac{\sum\left(X_i + \mu - (\bar{X} - \mu)\right)^2}{n-1} = \frac{\sum\left(X_i - \bar{X}\right)^2}{n-1} = s_X^2.$$

Since $\mathbf{E}(Y_i) = 0$, Case 1 applies to these variables. Thus,

$$\mathbf{E}(s_X^2) = \mathbf{E}(s_Y^2) = \sigma_Y^2 = \sigma_X^2.$$

□

Similarly to $\bar{X}$, it can be shown that under rather mild assumptions, sample variance and sample standard deviation are **consistent** and **asymptotically Normal**.

8.2.5 Interquartile range

Sample mean, variance, and standard deviation are *sensitive to outliers*. If an extreme observation (an **outlier**) erroneously appears in our data set, it can rather significantly affect the values of $\bar{X}$ and s^2.

In practice, outliers may be a real problem that is hard to avoid. To detect and identify outliers, we need measures of variability that are not very sensitive to them.

One such measure is an interquartile range.

DEFINITION 8.9

An **interquartile range** is defined as the difference between the first and the third quartiles,

$$IQR = Q_3 - Q_1.$$

It measures variability of data. Not much affected by outliers, it is often used to detect them.

Detection of outliers

A "rule of thumb" for identifying outliers is the rule of **1.5(IQR)**. Measure $1.5(Q_3 - Q_1)$ down from the first quartile and up from the third quartile. All the data points observed outside of this interval are assumed suspiciously far. They are the first candidates to be handled as outliers.

Remark: The rule of $1.5(IQR)$ comes from the assumption that the data are nearly normally distributed. If this is a valid assumption, then 99.3% of the population should appear within 1.5 interquartile ranges from quartiles (Exercise 8.4). It is so unlikely to see a value of X outside of this range that such an observation may be treated as an outlier.

Example 8.16 (Any outlying CPU times?). Can we suspect that sample (8.1) has outliers? Compute

$$IQR = Q_3 - Q_1 = 59 - 34 = 25$$

and measure 1.5 interquartile ranges from each quartile:

$$\begin{array}{rcrcr} Q_1 - 1.5(IQR) & = & 34 - 37.5 & = & -3.5; \\ Q_3 + 1.5(IQR) & = & 59 + 37.5 & = & 96.5. \end{array}$$

In our data, one task took 139 seconds, which is well outside of the interval $[-3.5, 96.5]$. This may be an outlier. ◇

Handling of outliers

What should we do if the 1.5(IQR) rule suggests possible outliers in the sample?

Many people simply delete suspicious observations, keeping in mind that one outlier can significantly affect sample mean and standard deviation and therefore spoil our statistical analysis. However, deleting them immediately may not be the best idea.

It is rather important to track the history of outliers and understand the reason they appeared in the data set. There may be a pattern that a practitioner would want to be aware of. It may be a new trend that was not known before. Or, it may be an observation from a very special part of the population. Sometimes important phenomena are discovered by looking at outliers.

If it is confirmed that a suspected observation entered the data set by a mere mistake, it can be deleted.

8.3 Graphical statistics

Despite highly developed theory and methodology of Statistics, when it comes to analysis of real data, experienced statisticians will often follow a very simple advice:

Before you do anything with a data set,
look at it!

A quick look at a sample may clearly suggest

- a probability model, i.e., a family of distributions to be used;
- statistical methods suitable for the given data;
- presence or absence of outliers;
- presence or absence of heterogeneity;
- existence of time trends and other patterns;
- relation between two or several variables.

There is a number of simple and advanced ways to *visualize* data. This section introduces

- histograms,
- stem-and-leaf plots,
- boxplots,
- time plots, and
- scatter plots.

Each graphical method serves a certain purpose and discovers certain information about data.

8.3.1 Histogram

A **histogram** shows the shape of a pmf or a pdf of data, checks for homogeneity, and suggests possible outliers. To construct a histogram, we split the range of data into equal intervals, "bins," and count how many observations fall into each bin.

A **frequency histogram** consists of columns, one for each bin, whose height is determined by the *number* of observations in the bin.

A **relative frequency histogram** has the same shape but a different vertical scale. Its column heights represent the *proportion* of all data that appeared in each bin.

The sample of CPU times on p. 225 stretches from 9 to 139 seconds. Choosing intervals [0,14], [14,28], [28,42], . . . as bins, we count

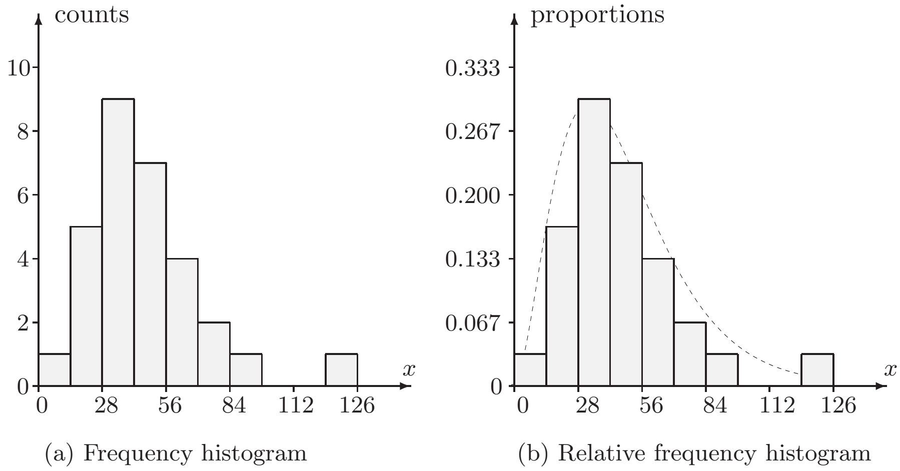
1	observation	between	0	and	14
5	observations	"	14	"	28
9	"	"	28	"	42
7	"	"	42	"	56
4	"	"	56	"	70

. . . .

Using this for column heights, a (frequency) histogram of CPU times is then constructed (Figure 8.5a). A relative frequency histogram (Figure 8.5b) is only different in the vertical scale. Each count is now divided by the sample size $n = 30$.

What information can we draw from these histograms?

Histograms have a shape similar to the pmf or pdf of data, especially in large samples.



(a) Frequency histogram

(b) Relative frequency histogram

Figure 8.5 *Histograms of CPU data.*

Remark: To understand the last statement, let's imagine for a moment that the data are integers and all columns in a relative frequency histogram have a unit width. Then the height of a column above a number x equals the proportion of x's in a sample, and in large samples it approximates the probability $P(x)$, the pmf (probability is a long-run proportion, Chapter 2).

For continuous distributions, the height of a unit-width column equals its area. Probabilities are areas under the density curve (Chapter 4). Thus, we get approximately the same result either computing sample proportions or integrating the curve that connects the column tops on a relative frequency histogram.

Now, if columns have a non-unit (but still, equal) width, it will only change the horizontal scale but will not alter the shape of a histogram. In each case, this shape will be similar to the graph of the population pmf or pdf.

The following information can be drawn from the histograms shown in Figure 8.5:

- Continuous distribution of CPU times is not symmetric; it is right-skewed as we see 5 columns to the right of the highest column and only 2 columns to the left.
- Among continuous distributions in Chapter 4, only Gamma distribution has a similar shape; a Gamma family seems appropriate for CPU times. We sketched a suitable Gamma pdf with a dashed curve in Figure 8.5b. It is rescaled because our columns don't have a unit width.
- The time of 139 seconds stands alone suggesting that it is in fact an outlier.

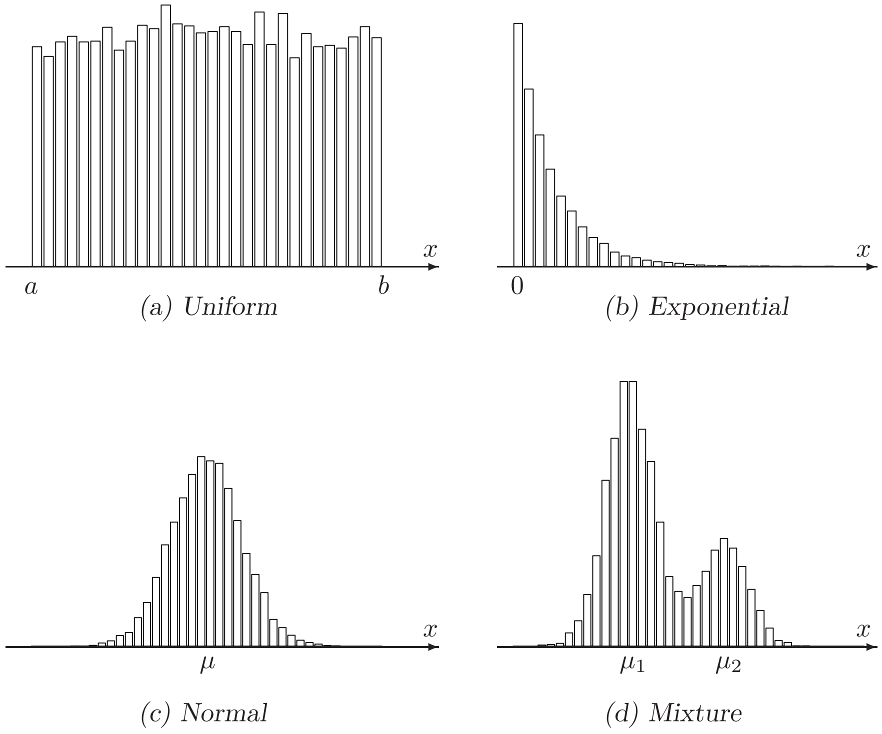


Figure 8.6 *Histograms of various samples.*

- There is no indication of heterogeneity; all data points except $x = 139$ form a rather homogeneous group that fits the sketched Gamma curve.

How else may histograms look like?

We saw a rather nice fit of a Gamma distribution in Figure 8.5b, except for one outlier. What other shapes of histograms can we see, and what other conclusions about the population can we make?

Certainly, histograms come in all shapes and sizes. Four examples are shown in Figure 8.6.

In Figure 8.6a, the distribution is almost symmetric, and columns have almost the same height. Slight differences can be attributed to the randomness of our sample, i.e., the *sampling error*. The histogram suggest a Uniform or Discrete Uniform distribution between a and b.

In Figure 8.6b, the distribution is heavily right-skewed, column heights decrease exponentially fast. This sample should come from an Exponential distribution, if variables are continuous, or from Geometric, if they are discrete.

In Figure 8.6c, the distribution is symmetric, with very quickly vanishing "tails." Its bell shape reminds a Normal density that, as we know from Section 4.2.4, decays at a rate of $\sim e^{-cx^2}$. We can locate the center μ of a histogram and conclude that this sample is likely to come from a Normal distribution with a mean close to μ.

Figure 8.6d presents a rather interesting case that deserves special attention.

Mixtures

Let us look at Figure 8.6d . We have not seen a distribution with two "humps" in the previous chapters. Most likely, here we deal with a **mixture of distributions**. Each observation comes from distribution F_1 with some probability p_1 and comes from distribution F_2 with probability $p_2 = 1 - p_1$.

Mixtures typically appear in heterogeneous populations that consist of several groups: females and males, graduate and undergraduate students, daytime and nighttime internet traffic, Windows, Unix, or Macintosh users, etc. In such cases, we can either study each group separately, or use the Law of Total Probability on p. 32 and write the (unconditional) cdf as

$$F(x) = p_1 F_1(x) + p_2 F_2(x) + \ldots$$

and study the whole population at once.

Bell shapes of both humps in Figure 8.6d suggests that the sample came from a mixture of two Normal distributions (with means around μ_1 and μ_2), with a higher probability of having mean μ_1, since the left hump is bigger.

The choice of bins

Experimenting with histograms, you can notice that their shape depends on the choice of bins. One can hear various rules of thumb about a good choice of bins, but in general,

- there should not be too few or too many bins;
- their number may increase with a sample size;
- they should be chosen to make the histogram informative, so that we can see shapes, outliers, etc.

In Figure 8.5, we simply divided the range of CPU data into 10 equal intervals, 14 sec each, and apparently this was sufficient for drawing important conclusions.

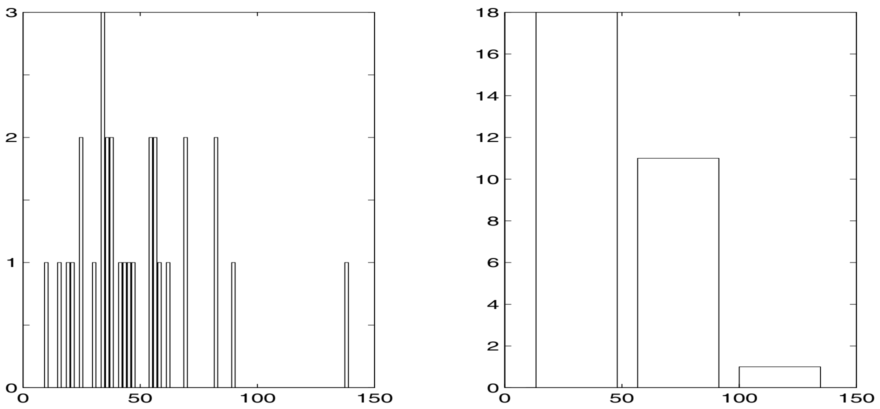


Figure 8.7 *Wrong choice of bins for CPU data: too many bins, too few bins.*

As two extremes, consider histograms in Figure 8.7 constructed from the same CPU data.

The first histogram has too many columns, therefore, each column is short. Most bins have only 1 observation. This tells little about the actual shape of the distribution; however, we can still notice an outlier $X = 139$.

The second histogram has only 3 columns. It is hard to guess the family of distributions here, although a flat Uniform distribution is already ruled out. The outlier is not seen, it merged with the rightmost bin.

Both histograms in Figure 8.7 can be made more informative by a better choice of bins.

A MATLAB command for constructing histograms is `hist(x)` or `hist(x,n)` where x is the data, and n is the desired number of bins.

8.3.2 Stem-and-leaf plot

Stem-and-leaf plots are similar to histograms although they carry more information. Namely, they also show how the data are distributed *within* columns.

To construct a stem-and-leaf plot, we need to draw a stem and a leaf. The first one or several digits form a stem, and the next digit forms a leaf. Other digits are dropped; in other words, the numbers get rounded. For example, a number 239 can be written as

$$23 \;\big|\; 9$$

with 23 going to the stem and 9 to the leaf, or as

$$2 \,\big|\, 3$$

with 2 joining the stem, 3 joining the leaf, and digit 9 being dropped. In the first case, the *leaf unit* equals 1 while in the second case, the leaf unit is 10 showing that the (rounded) number is not 23 but 230.

For the CPU data on p. 225, let the last digits form a leaf. The remaining digits go to the stem. Each CPU time is then written as

$$10\text{ “stem”} + \text{“leaf”},$$

making the following stem-and-leaf plot,

Leaf unit = 1

0	9							
1	5	9						
2	2	4	5					
3	0	4	5	5	6	6	7	8
4	2	3	6	8				
5	4	5	6	6	9			
6	2	9						
7	0							
8	2	2	9					
9								
10								
11								
12								
13	9							

Turning this plot by 90 degrees counterclockwise, we get a *histogram* with 10-unit bins (because each stem unit equals 10). Thus, all the information seen on a histogram can be obtained here too. In addition, now we can see individual values within each column. We have the entire sample sorted and written in the form of a stem-and-leaf plot. If needed, we can even compute sample mean, median, quartiles, and other statistics from it.

Example 8.17 (Comparison). Sometimes stem-and-leaf plots are used to compare two samples. For this purpose, one can put two leaves on the same stem. Consider, for example, samples of round-trip transit times (known as pings) received from two locations.

Location I: 0.0156, 0.0396, 0.0355, 0.0480, 0.0419, 0.0335, 0.0543, 0.0350, 0.0280, 0.0210, 0.0308, 0.0327, 0.0215, 0.0437, 0.0483 seconds

Location II: 0.0298, 0.0674, 0.0387, 0.0787, 0.0467, 0.0712, 0.0045, 0.0167, 0.0661, 0.0109, 0.0198, 0.0039 seconds

Choosing a leaf unit of 0.001, a stem unit of 0.01, and dropping the last digit, we construct the following two stem-and-leaf plots.

Leaf unit = 0.001

						0	3	4	
					5	1	0	6	9
			1	1	8	2			
0	2	3	5	5	9	3	8		
		1	3	8	8	4	6		
					4	5			
						6	1	6	7
						7	8		

Looking at these two plots, one will see about the same average ping from the two locations. One will also realize that the first location has a more stable connection because its pings have lower variability and lower variance. For the second location, the fastest ping will be understood as

$$\{10(\text{leaf } 0) + \text{stem } 3\}\,(\text{leaf unit } 0.001) = 0.003,$$

and the slowest ping as

$$\{10(\text{leaf } 7) + \text{stem } 8\}\,(\text{leaf unit } 0.001) = 0.078.$$

◇

8.3.3 Boxplot

The main descriptive statistics of a sample can be represented graphically by a **boxplot**. To construct a boxplot, we draw a box between the first and the third quartiles, a line inside a box for a median, and extend whiskers to the smallest and the largest observations, thus representing a so-called *five-point summary*

$$\text{five-point summary} = \left(\min X_i,\ \hat{Q}_1,\ \hat{M},\ \hat{Q}_3,\ \max X_i\right).$$

Often a sample mean $\bar{X}$ is also depicted with a dot or a cross. Observations further than 1.5 interquartile ranges are usually drawn separately from whiskers, indicating the possibility of outliers. This is in accordance with the 1.5(IQR) rule (see Section 8.2.5).

The mean and five-point summary of CPU times were found in Examples 8.6, 8.11, and 8.13,

$$\bar{X} = 48.2333;\ \min X_i = 9,\ \hat{Q}_1 = 34,\ \hat{M} = 42.5,\ \hat{Q}_3 = 59,\ \max X_i = 139.$$

We also know that $X = 139$ is more than 1.5(IQR) away from the third quartile, and we suspect that it may be an outlier.

A boxplot is drawn in Figure 8.8. The mean is depicted with a "+," and

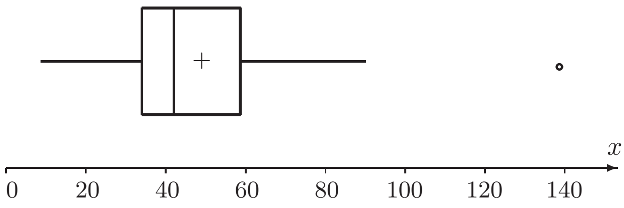


Figure 8.8 *Boxplot of CPU time data.*

the right whisker extends till the second largest observation $X = 89$ because $X = 139$ is a suspected outlier (depicted with a little circle).

From this boxplot, one can conclude:

- The distribution of CPU times is right skewed because (1) the mean exceeds the median, and (2) the right half of the box is larger than the left half.
- Each half of a box and each whisker represents approximately 25% of the population. For example, we expect about 25% of all CPU times to fall between 42.5 and 59 seconds.

Parallel boxplots

Boxplots are often used to compare different populations or parts of the same population. For such a comparison, samples of data are collected from each part, and their boxplots are drawn on the next scale next to each other.

For example, seven parallel boxplots in Figure 8.9 represent the amount of internet traffic handled by a certain center during a week. We can see the following general patterns:

- The heaviest internet traffic occurs on Fridays.
- Fridays also have the highest variability.
- The lightest traffic is seen on weekends, with an increasing trend from Saturday to Monday.
- Each day, the distribution is right-skewed, with a few outliers on each day except Saturday. Outliers indicate occurrences of unusually heavy internet traffic.

Trends can also be seen on scatter plots and time plots.

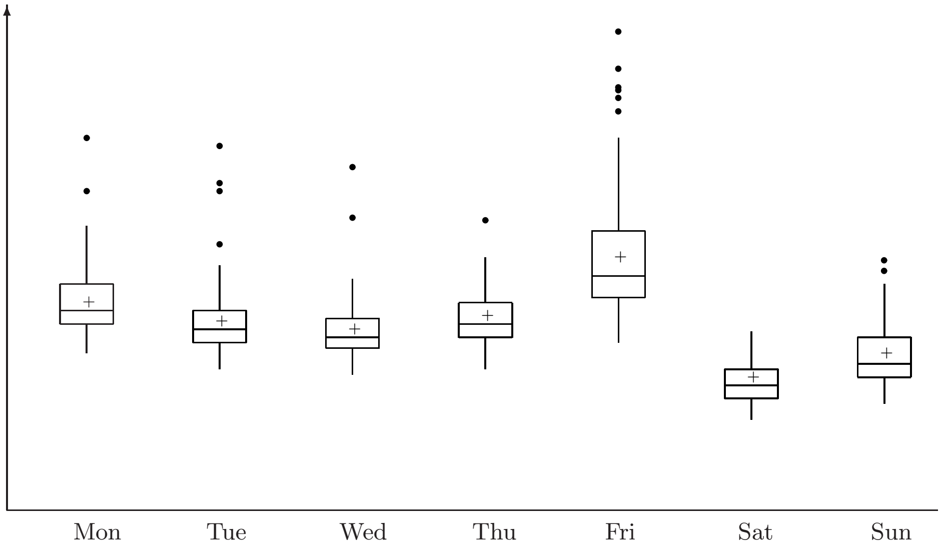


Figure 8.9 *Parallel boxplots of internet traffic.*

8.3.4 Scatter plots and time plots

Scatter plots are used to see and understand a relationship between two variables. These can be temperature and humidity, experience and salary, age of a network and its speed, number of servers and the expected response time, and so on.

To study the relationship, both variables are measured on each sampled item. For example, temperature and humidity during n days, age and speed of n networks, or experience and salary of n randomly chosen computer scientists are recorded. Then, a **scatter plot** consists of n points on an (x, y)-plane, with x- and y-coordinates representing the two recorded variables.

Example 8.18 (Antivirus maintenance). Protection of a personal computer largely depends on the frequency of running antivirus software on it. One can set to run it every day, once a week, once a month, etc.

During a scheduled maintenance of computer facilities, a computer manager records the number of times the antivirus software was launched on each computer during 1 month (variable X) and the number of detected worms (variable Y). The data for 30 computers are in the table.

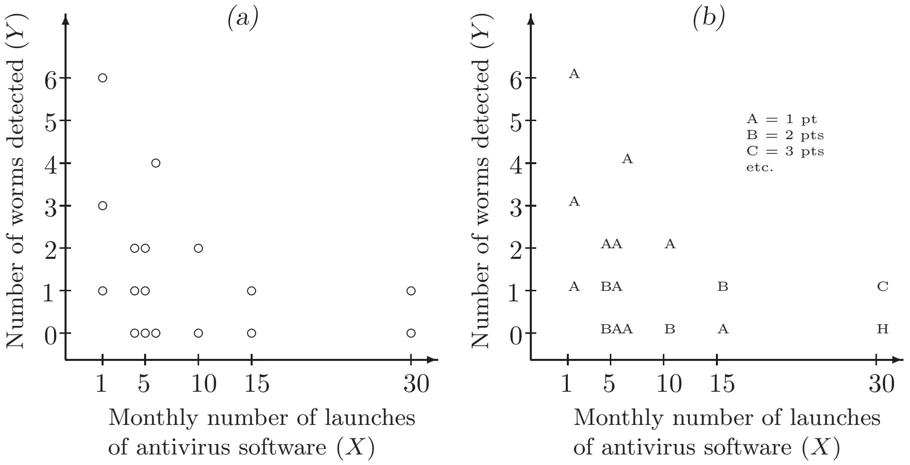


Figure 8.10 *Scatter plots for Examples 8.18 and 8.19.*

X	30	30	30	30	30	30	30	30	30	30	30	15	15	15	10
Y	0	0	1	0	0	0	1	1	0	0	0	0	1	1	0

X	10	10	6	6	5	5	5	4	4	4	4	4	1	1	1
Y	0	2	0	4	1	2	0	2	1	0	1	0	6	3	1

Is there a connection between the frequency of running antivirus software and the number of worms in the system? A scatter plot of these data is given in Figure 8.10a. It clearly shows that the number of worms reduces, in general, when the antivirus is employed more frequently. This relationship, however, is not certain because no worm was detected on some "lucky" computers although the antivirus software was launched only once a week on them. ◇

Example 8.19 (Plotting identical points). Looking at the scatter plot in Figure 8.10a, the manager in Example 8.18 realized that a portion of data is hidden there because there are identical observations. For example, no worms were detected on 8 computers where the antivirus software is used daily (30 times a month). Then, Figure 8.10a may be misleading.

When the data contain identical pairs of observations, the points on a scatter plot are often depicted with either numbers or letters ("A" for 1 point, "B" for two identical points, "C" for three, etc.). You can see the result in Figure 8.10b. ◇

When we study time trends and development of variables over time, we use **time plots**. These are scatter plots with x-variable representing time.

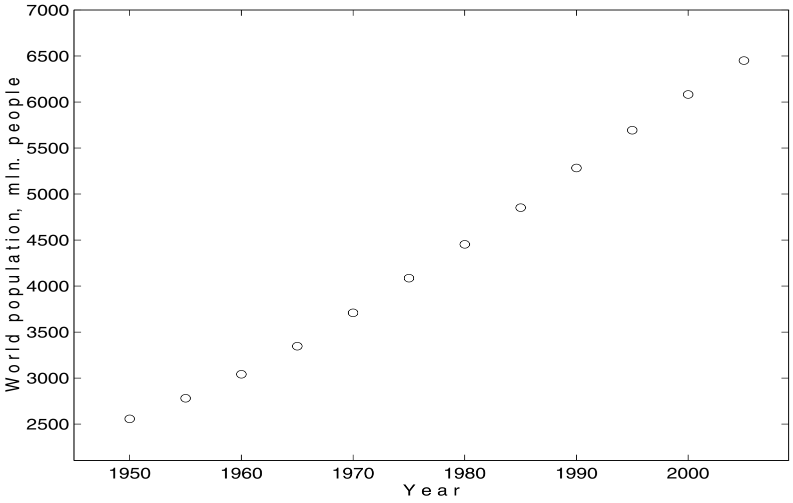


Figure 8.11 *Time plot of the world population in 1950–2005.*

Example 8.20 (World population). For example, here is how the world population increased between 1950 and 2005 (Figure 8.11). We can clearly see that the population increases at an almost steady rate. The actual data are given in Table 10.1 on p. 328. We estimate the trends seen on time plots and scatter plots and even make forecasts for the future in Chapter 10. ◇

Summary and conclusions

Methods discussed in this chapter provide a quick description of a data set, general conclusions about the population, and help to formulate conjectures about its features.

Sample expectation, variance, moments, and quantiles estimate and give us an idea about the corresponding population parameters.

Before we start working with a data set, we look at it! There is a number of methods to display data: histograms, stem-and-leaf plots, boxplots, scatter plots, time plots. Good data displays show the shape of the distribution, the type of skewness or symmetry, interrelations, general trends, and possible outliers.

More advanced and accurate statistical methods are introduced in the next chapter.

Questions and exercises

8.1. The numbers of blocked intrusion attempts on each day during the first two weeks of the month were

$$56, 47, 49, 37, 38, 60, 50, 43, 43, 59, 50, 56, 54, 58$$

After the change of firewall settings, the numbers of blocked intrusions during the next 20 days were

$$53, 21, 32, 49, 45, 38, 44, 33, 32, 43, 53, 46, 36, 48, 39, 35, 37, 36, 39, 45.$$

Comparing the number of blocked intrusions before and after the change,

(a) construct side-by-side stem-and-leaf plots

(b) compute the five-number summaries and construct parallel boxplots

(c) comment on your findings

8.2. A network provider investigates the load of its network. The number of concurrent users is recorded at fifty locations (thousands of people),

17.2	22.1	18.5	17.2	18.6	14.8	21.7	15.8	16.3	22.8
24.1	13.3	16.2	17.5	19.0	23.9	14.8	22.2	21.7	20.7
13.5	15.8	13.1	16.1	21.9	23.9	19.3	12.0	19.9	19.4
15.4	16.7	19.5	16.2	16.9	17.1	20.2	13.4	19.8	17.7
19.7	18.7	17.6	15.9	15.2	17.1	15.0	18.8	21.6	11.9

(a) Compute the sample mean, variance, and standard deviation of the number of concurrent users.

(b) Compute the five-point summary and construct a boxplot.

(c) Compute the interquartile range. Are there any outliers?

(d) It is reported that the number of concurrent users follows approximately Normal distribution. Does the histogram support this claim?

8.3. Verify the use of Chebyshev's inequality in (8.6) of Example 8.15. Show that if the population mean is indeed 48.2333 and the population standard deviation is indeed 26.5170, then at least 8/9 of all tasks require less than 127.78 seconds of CPU time.

8.4. Use Table A4 to compute the probability for *any* Normal random variable to take a value within 1.5 interquartile ranges from population quartiles.

8.5. The following data shows population of the United States (in million) since 1790,

Year	1790	1800	1810	1820	1830	1840	1850	1860
Population	3.9	5.3	7.2	9.6	12.9	17.1	23.2	31.4

Year	1870	1880	1890	1900	1910	1920	1930	1940
Population	38.6	50.2	63.0	76.2	92.2	106.0	123.2	132.2

Year	1950	1960	1970	1980	1990	2000
Population	151.3	179.3	203.3	226.5	248.7	281.4

Construct a time plot for the U.S. population. What kind of trend do you see? What information can be extracted from this plot?

8.6. Refer to Exercise 8.5. Compute 10-year *increments* of the population growth $x_1 = 5.3 - 3.9$, $x_2 = 7.2 - 5.3$, etc.

(a) Compute sample mean, median, and variance of 10-year increments. Discuss how the U.S. population changes during a decade.

(b) Construct a time plot of 10-year increments and discuss the observed pattern.

8.7. Refer to Exercise 8.5. Compute 10-year *relative population change* $y_1 = (5.3 - 3.9)/3.9$, $y_2 = (7.2 - 5.3)/5.3$, etc.

(a) Compute sample mean, median, and variance of the relative population change.

(b) Construct a time plot of the relative population change. What trend do you see now?

(c) Comparing the time plots in Exercises 8.6 and 8.7, what kind of correlation between x_i and y_i would you expect? Verify by computing the sample correlation coefficient

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})/(n-1)}{s_x s_y}.$$

What can you conclude? How would you explain this phenomenon?

8.8. Consider three data sets.

(1) 19, 24, 12, 19, 18, 24, 8, 5, 9, 20, 13, 11, 1, 12, 11, 10, 22, 21, 7, 16, 15, 15, 26, 16, 1, 13, 21, 21, 20, 19

(2) 17, 24, 21, 22, 26, 22, 19, 21, 23, 11, 19, 14, 23, 25, 26, 15, 17, 26, 21, 18, 19, 21, 24, 18, 16, 20, 21, 20, 23, 33

(3) 56, 52, 13, 34, 33, 18, 44, 41, 48, 75, 24, 19, 35, 27, 46, 62, 71, 24, 66, 94, 40, 18, 15, 39, 53, 23, 41, 78, 15, 35

(a) For each data set, draw a histogram and determine whether the distribution is right-skewed, left-skewed, or symmetric.

(b) Compute sample means and sample medians. Do they support your findings about skewness and symmetry? How?

8.9. The following data set represents the number of new computer accounts registered during ten consecutive days.

$$43, 37, 50, 51, 58, 105, 52, 45, 45, 10.$$

(a) Compute the mean, median, quartiles, and standard deviation.

(b) Check for outliers using the 1.5(IQR) rule.

(c) Delete the detected outliers and compute the mean, median, quartiles, and standard deviation again.

(d) Make a conclusion about the effect of outliers on basic descriptive statistics.

CHAPTER 9

Statistical Inference

After taking a general look at the data, we are ready for more advanced and more informative statistical analysis.

In this chapter, we learn how

- *to estimate parameters* of the distribution. Methods of Chapter 8 mostly concern measure of location (mean, median, quantiles) and variability (variance, standard deviation, interquartile range). As we know, this does not cover all possible parameters, and thus, we still lack a general methodology of estimation.
- *to construct confidence intervals*. Any estimator, computed from a collected random sample instead of the whole population, is understood as only an approximation of the corresponding parameter. Instead of one estimator that is subject to a *sampling error*, it is often more reasonable to produce an interval that will contain the true population parameter with a certain known high probability.
- *to test hypotheses*. That is, we shall use the collected sample to verify statements and claims about the population. As a result of each test, a statement is either rejected on basis of the observed data or accepted (not rejected). Sampling error in this analysis results in a possibility of wrongfully accepting or rejecting the hypothesis, however, we can design tests to control the probability of such errors.

Results of such statistical analysis are used for making decisions under uncertainty, developing optimal strategies, forecasting, evaluating and controlling performance, and so on.

9.1 Parameter estimation

By now, we have learned a few elementary ways to determine the *family of distributions*. We take into account the nature of our data, basic description, and range; propose a suitable family of distributions; and support our conjecture by looking at a histogram.

In this section, we learn how to estimate parameters of distributions. As a result, a large family will be reduced to just one distribution that we can use for performance evaluation, forecasting, etc.

Example 9.1 (Poisson). For example, consider a sample of computer chips with a certain type of rare defects. The number of defects on each chip is recorded. This is the number of rare events, and thus, it should follow a Poisson distribution with *some* parameter λ.

We know that $\lambda = \mathbf{E}(X)$ is the expectation of a Poisson variable (Section 3.4.5). Then, should we estimate it with a sample mean $\bar{X}$, as in the previous section? Or, should we use a sample variance s^2 because λ also equals $\mathrm{Var}(X)$? ◇

Example 9.2 (Gamma). Suppose now that we deal with a Gamma(α, λ) family of distributions. Its parameters α and λ do not represent the mean, variance, standard deviation, or any other measures discussed in Chapter 8. What would the estimation algorithm be this time? ◇

Questions raised in these examples do not have unique answers. Statisticians developed a number of estimation techniques, each having certain optimal properties.

Two rather popular methods are discussed in this section:

– method of moments, and

– method of maximum likelihood.

Two other methods are introduced later: Bayesian parameter estimation in Section 9.5 and least squares estimation in Chapter 10.

9.1.1 Method of moments

Moments

First, let us define the moments.

DEFINITION 9.1

The k-th **population moment** is defined as

$$\mu_k = \mathbf{E}(X^k).$$

The k-th **sample moment**

$$m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

estimates μ_k from a sample $(X_1, \ldots, X_n)$.

The first sample moment is the sample mean $\bar{X}$.

Central moments are computed similarly, after centralizing the data, that is, subtracting the mean.

DEFINITION 9.2

The k-th **population central moment** is defined as

$$\mu_k' = \mathbf{E}(X - \mu_1)^k.$$

The k-th **sample central moment**

$$m_k' = \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^k$$

estimates μ_k from a sample $(X_1, \ldots, X_n)$.

Remark: The second population central moment is variance $\mathrm{Var}(X)$. The second sample central moment is sample variance, although $(n-1)$ in its denominator is now replaced by n. We mentioned that estimation methods are not unique. For unbiased estimation of $\sigma^2 = \mathrm{Var}(X)$, we use

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

however, method of moments and method of maximum likelihood produce a different version,

$$S^2 = m_2' = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

And this is not all! We'll see other estimates of σ^2 as well.

Estimation

Method of moments is based on a simple idea. Since our sample comes from a family of distributions $\{F(\theta)\}$, we choose such a member of this family whose properties are close to properties of our data. Namely, we shall match the *moments*.

To estimate k parameters, equate the first k population and sample moments,

$$\left\{ \begin{array}{ccc} \mu_1 & = & m_1 \\ \ldots & \ldots & \ldots \\ \mu_k & = & m_k \end{array} \right.$$

The left-hand sides of these equations depend on the distribution parameters. The right-hand sides can be computed from data. The **method of moments estimator** is the solution of this system of equations.

Example 9.3 (Poisson). To estimate parameter λ of Poisson(λ) distribution, we recall that

$$\mu_1 = \mathbf{E}(X) = \lambda.$$

There is only one unknown parameter, hence we write one equation,

$$\mu_1 = \lambda = m_1 = \bar{X}.$$

"Solving" it for λ, we obtain

$$\hat{\lambda} = \bar{X},$$

the method of moments estimator of λ. ◇

If it is easier, one may opt to equate central moments.

Example 9.4 (Gamma distribution of CPU times). The histogram in Figure 8.5 suggested that CPU times have Gamma distribution with some parameters α and λ. To estimate them, we need two equations. From data on p. 225, we compute

$$m_1 = \bar{X} = 48.2333 \quad \text{and} \quad m_2' = S^2 = 679.7122.$$

and write two equations,

$$\left\{ \begin{array}{ccccccc} \mu_1 & = & \mathbf{E}(X) & = & \alpha/\lambda & = & m_1 \\ \mu_2' & = & \mathrm{Var}(X) & = & \alpha/\lambda^2 & = & m_2'. \end{array} \right.$$

It is convenient to use the second *central* moment here because we already know the expression for the variance $m_2' = \mathrm{Var}(X)$ of a Gamma variable.

Solving this system in terms of α and λ, we get the method of moment estimates

$$\left\{ \begin{array}{ccccc} \hat{\alpha} & = & m_1^2/m_2' & = & 3.4227 \\ \hat{\lambda} & = & m_1/m_2' & = & 0.0710. \end{array} \right.$$

◇

Of course, we solved these two examples so quickly because we already knew the moments of Poisson and Gamma distributions from Sections 3.4.5 and 4.2.3. When we see a new distribution for us, we'll have to compute its moments.

Consider, for example, *Pareto distribution* that plays an increasingly vital role in modern internet modeling due to very heavy internet traffic nowadays.

Example 9.5 (Pareto). A two-parameter *Pareto distribution* has a cdf

$$F(x) = 1 - \left(\frac{x}{\sigma}\right)^{-\theta} \quad \text{for } x > \sigma.$$

How should we compute a method of moments estimator of σ and θ?

We have not seen Pareto distribution in this book so far, so we'll have to compute its first two moments.

We start with the density

$$f(x) = F'(x) = \frac{\theta}{\sigma}\left(\frac{x}{\sigma}\right)^{-\theta-1} = \theta\sigma^{\theta}x^{-\theta-1}$$

and use it to find the expectation

$$\begin{aligned} \mu_1 &= \mathbf{E}(X) = \int_{\sigma}^{\infty} x\, f(x)\, dx = \theta\sigma^{\theta}\int_{\sigma}^{\infty} x^{-\theta} dx \\ &= \theta\sigma^{\theta}\, \frac{x^{-\theta+1}}{-\theta+1}\Bigg|_{x=\sigma}^{x=\infty} = \frac{\theta\sigma}{\theta-1}, \quad \text{for } \theta > 1, \end{aligned}$$

and the second moment

$$\mu_2 = \mathbf{E}(X^2) = \int_{\sigma}^{\infty} x^2\, f(x)\, dx = \theta\sigma^{\theta}\int_{\sigma}^{\infty} x^{-\theta+1}dx = \frac{\theta\sigma^2}{\theta-2}, \quad \text{for } \theta > 2.$$

For $\theta \leq 1$, a Pareto variable has an infinite expectation, and for $\theta \leq 2$, it has an infinite second moment.

Then we solve the method of moments equations

$$\left\{ \begin{array}{lllll} \mu_1 & = & \dfrac{\theta\sigma}{\theta-1} & = & m_1 \\ \mu_2 & = & \dfrac{\theta\sigma^2}{\theta-2} & = & m_2 \end{array} \right.$$

and find that

$$\hat{\theta} = \sqrt{\frac{m_2}{m_2 - m_1^2}} + 1 \quad \text{and} \quad \hat{\sigma} = \frac{m_1(\hat{\theta}-1)}{\hat{\theta}}. \tag{9.1}$$

When we collect a sample from Pareto distribution, we can compute sample moments m_1 and m_2 and estimate parameters by (9.1). ◇

On rare occasions, when k equations are not enough to estimate k parameters, we'll consider higher moments.

Example 9.6 (NORMAL). Suppose we already know the mean μ of a Normal distribution and would like to estimate the variance σ^2. Only one parameter σ^2 is unknown, however, the first method of moments equation

$$\mu_1 = m_1$$

does not contain σ^2 and therefore does not produce its estimate. We then consider the second equation, say,

$$\mu_2' = \sigma^2 = m_2' = S^2,$$

which gives us the method of moments estimate immediately, $\hat{\sigma}^2 = S^2$. ◇

Method of moments estimates are typically easy to compute. They can serve as a quick tool for estimating parameters of interest.

9.1.2 Method of maximum likelihood

Another interesting idea is behind the method of *maximum likelihood estimation.*

Since the sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ has already been observed, we find such parameters that maximize the probability (likelihood) for this to happen. In other words, we make the event that has happened to be as likely as possible. This is yet another way to make the chosen distribution consistent with the observed data.

DEFINITION 9.3

> **Maximum likelihood estimator** is the parameter value that maximizes the likelihood of the observed sample. For a discrete distribution, we maximize the joint pmf of data $P(X_1, \ldots, X_n)$. For a continuous distribution, we maximize the joint density $f(X_1, \ldots, X_n)$.

Both cases, discrete and continuous, are explained below.

Discrete case

For a discrete distribution, the probability of a given sample is the joint pmf of data,

$$\boldsymbol{P}\left\{\boldsymbol{X} = (X_1, \ldots, X_n)\right\} = P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i),$$

because in a simple random sample, all observed X_i are independent.

To maximize this likelihood, we consider the critical points by taking derivatives with respect to all unknown parameters and equating them to 0 (see Calculus review in Appendix 11.3). A nice shortcut is to take logarithms first. Differentiating the sum

$$\ln \prod_{i=1}^{n} P(X_i) = \sum_{i=1}^{n} \ln P(X_i)$$

is easier than differentiating the product $\prod P(X_i)$. Besides, since logarithm is an increasing function, the log-likelihood $\ln P(\boldsymbol{X})$ is maximized at exactly the same point as the likelihood $P(\boldsymbol{X})$.

Example 9.7 (Poisson). The pmf of Poisson distribution is

$$P(x) = e^{-\lambda} \frac{\lambda^x}{x!},$$

and its logarithm is

$$\ln P(x) = -\lambda + x \ln \lambda - \ln(x!).$$

Thus, we need to maximize

$$\ln P(\boldsymbol{X}) = \sum_{i=1}^{n} (-\lambda + X_i \ln \lambda) + C = -n\lambda + \ln \lambda \sum_{i=1}^{n} X_i,$$

where $C = -\sum \ln(x!)$ is a constant that does not contain the unknown parameter λ.

Find the critical point(s) of this log-likelihood. Differentiating it and equating its derivative to 0, we get

$$\frac{\partial}{\partial \lambda} \ln P(\boldsymbol{X}) = -n + \frac{1}{\lambda} \sum_{i=1}^{n} X_i = 0.$$

This equation has only one solution

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}.$$

Since this is the only critical point, and since the likelihood vanishes (converges to 0) as $\lambda \downarrow 0$ or $\lambda \uparrow \infty$, we conclude that $\hat{\lambda}$ is the maximizer. Therefore, it is the maximum likelihood estimator of λ.

For the Poisson distribution, the method of moments and the method of maximum likelihood returned the same estimator, $\hat{\lambda} = \bar{X}$. ◇

Continuous case

In the continuous case, the probability to observe exactly the given number $X = x$ is 0, as we know from Chapter 4. Instead, the method of maximum

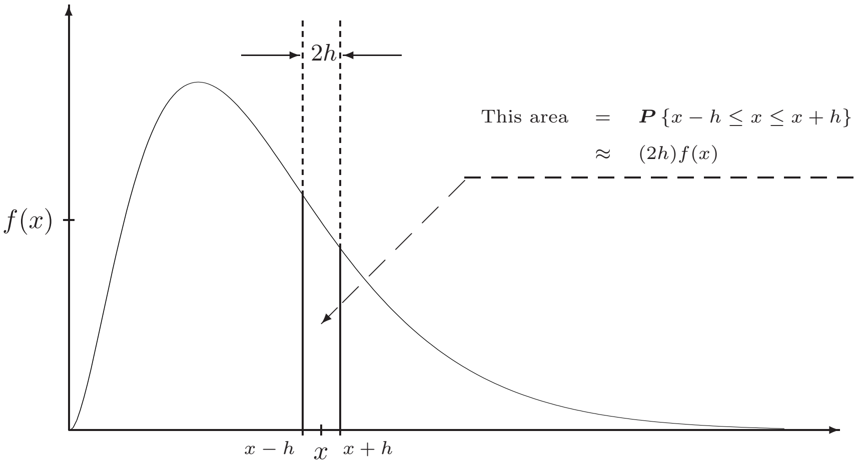


Figure 9.1 *Probability of observing "almost"* $X = x$.

likelihood will maximize the probability of observing "almost" the same number.

For a very small h,

$$\boldsymbol{P}\{x-h < X < x+h\} = \int_{x-h}^{x+h} f(y)dy \approx (2h)f(x).$$

That is, the probability of observing a value close to x is proportional to the density $f(x)$ (see Figure 9.1). Then, for a sample $\boldsymbol{X} = (X_1, \ldots, X_n)$, the maximum likelihood method will maximize the joint density $f(X_1, \ldots, X_n)$.

Example 9.8 (Exponential). The Exponential density is

$$f(x) = \lambda e^{-\lambda x},$$

so the log-likelihood of a sample can be written as

$$\ln f(\boldsymbol{X}) = \sum_{i=1}^{n} \ln\left(\lambda e^{-\lambda X_i}\right) = \sum_{i=1}^{n} (\ln\lambda - \lambda X_i) = n\ln\lambda - \lambda\sum_{i=1}^{n} X_i.$$

Taking its derivative with respect to the unknown parameter λ, equating to 0, and solving for λ, we get

$$\frac{\partial}{\partial\lambda}\ln f(\boldsymbol{X}) = \frac{n}{\lambda} - \sum_{i=1}^{n} X_i = 0,$$

resulting in

$$\hat{\lambda} = \frac{n}{\sum X_i} = \frac{1}{\bar{X}}.$$

Again, this is the only critical point, and the likelihood $f(\boldsymbol{X})$ vanishes as $\lambda \downarrow 0$ or $\lambda \uparrow \infty$. Thus, $\hat{\lambda} = \bar{X}$ is the maximum likelihood estimator of λ. This time, it also coincides with the method of moments estimator (Exercise 9.3b). $\diamond$

Sometimes the likelihood has no critical points, then it is maximized at the boundary.

Example 9.9 (Uniform). Based on a sample from Uniform$(0, b)$ distribution, how can we estimate the parameter b?

The Uniform$(0, b)$ density is

$$f(x) = \frac{1}{b} \quad \text{for } \; 0 \leq x \leq b.$$

It is decreasing in b, and therefore, it is maximized at the the smallest possible value of b, which is x.

For a sample $(X_1, \ldots, X_n)$, the joint density

$$f(X_1, \ldots, X_n) = \left(\frac{1}{b}\right)^n \quad \text{for } \; 0 \leq X_1, \ldots, X_n \leq b$$

also attains its maximum at the smallest possible value of b which is now the largest observation. Indeed, $b \geq X_i$ for all i only if $b \geq \max(X_i)$.

Therefore, the maximum likelihood estimator is $\hat{b} = \max(X_i)$. $\diamond$

When we estimate more than 1 parameter, all the partial derivatives should be equal 0 at the critical point. If no critical points exist, the likelihood is again maximized on the boundary.

Example 9.10 (Pareto). For the Pareto distribution in Example 9.5, the log-likelihood is

$$\ln f(\boldsymbol{X}) = \sum_{i=1}^{n} \ln \left(\theta \sigma^{\theta} X_i^{-\theta-1}\right) = n \ln \theta + n\theta \ln \sigma - (\theta + 1) \sum_{i=1}^{n} \ln X_i$$
$$\text{for} \quad X_1, \ldots, X_n \geq \sigma.$$

Maximizing this function over both σ and θ, we notice that it always increases in σ. Thus, we estimate σ by its largest possible value, which is the smallest observation,

$$\hat{\sigma} = \min(X_i).$$

We can substitute this value of σ into the log-likelihood and maximize with respect to θ,

$$\frac{\partial}{\partial \theta} \ln f(\boldsymbol{X}) = \frac{n}{\theta} + n \ln \hat{\sigma} - \sum_{i=1}^{n} \ln X_i = 0;$$

$$\hat{\theta} = \frac{n}{\sum \ln X_i - n \ln \hat{\sigma}} = \frac{n}{\sum \ln \left(X_i/\hat{\sigma}\right)}.$$

The maximum likelihood estimates of σ and θ are

$$\hat{\sigma} = \min(X_i) \quad \text{and} \quad \hat{\theta} = \frac{n}{\sum \ln \left(X_i/\hat{\sigma}\right)}.$$

◇

Maximum likelihood estimators are rather popular because of their nice properties. Under mild conditions, these estimators are consistent, and for large samples, they have an approximately Normal distribution. Often in complicated problems, finding a good estimation scheme may be challenging whereas the maximum likelihood method always gives a reasonable solution.

9.2 Confidence intervals

When we report an estimator $\hat{\theta}$ of a population parameter θ, we know that most likely

$$\hat{\theta} \neq \theta$$

due to a sampling error. We realize that we have estimated θ *up to some error*. Likewise, nobody understands the internet connection of 11 megabytes per second as exactly 11 megabytes going through the network every second, and nobody takes a meteorological forecast as the promise of exactly the predicted temperature.

Then how much can we trust the reported estimator? How far can it be from the actual parameter of interest, and what is the probability that it will be reasonably close?

To answer these questions, to assess the accuracy of reported estimates and their *margin of error*, statisticians use *confidence intervals*.

DEFINITION 9.4

An interval $[a, b]$ is a $(1-\alpha)100\%$ **confidence interval** for the parameter θ if it contains the parameter with probability $(1-\alpha)$,

$$\boldsymbol{P}\{a \leq \theta \leq b\} = 1 - \alpha.$$

The **coverage probability** $(1-\alpha)$ is also called a **confidence level**.

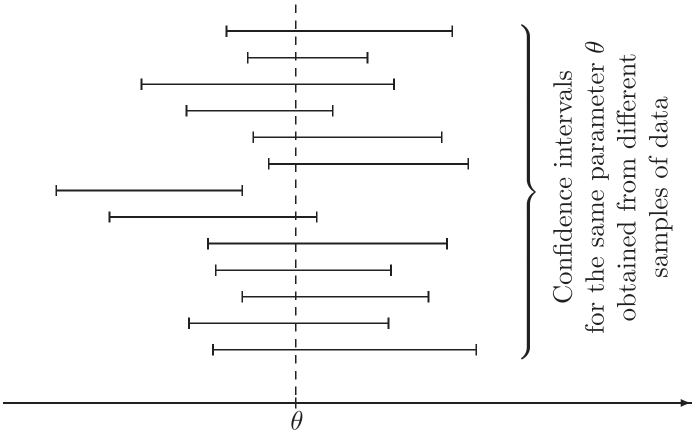


Figure 9.2 *Confidence intervals and coverage of parameter θ.*

Let us take a moment to think about this definition. The probability of a random event $\{a \leq \theta \leq b\}$ has to be $(1-\alpha)$. What randomness is involved in this event?

The population parameter θ is not random. It characterizes the population, independently of any random sampling procedure, and therefore, remains constant. On the other hand, the interval is computed from random data, and therefore, it is random. The *coverage probability* refers to the chance that our interval covers a constant parameter θ.

This is illustrated in Figure 9.2. Suppose that we collect many random samples and produce a confidence interval from each of them. If these are $(1-\alpha)100\%$ confidence intervals, then we expect $(1-\alpha)100\%$ of them to cover θ and $100\alpha\%$ of them to miss it. In Figure 9.2, we see one interval that does not cover θ. No mistake was made in data collection and construction of this interval. It missed the parameter only due to a *sampling error*.

It is therefore *wrong* to say, *"I computed a 90% confidence interval, it is* $[3, 6]$*. Parameter belongs to this interval with probability 90%."* The parameter is constant, it either belongs to the interval $[3, 6]$ (with probability 1) or does not.

9.2.1 Construction of confidence intervals: a general method

Given a sample of data and a desired confidence level $(1-\alpha)$, how can we construct a confidence interval $[a, b]$ that will satisfy the coverage condition

$$\boldsymbol{P}\{a \leq \theta \leq b\} = 1-\alpha$$

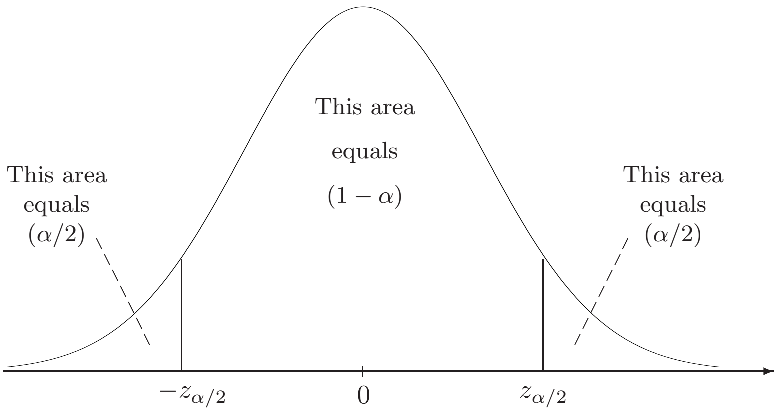


Figure 9.3 *Standard Normal quantiles* $\pm z_{\alpha/2}$ *and partition of the area under the density curve.*

in Definition 9.4?

We start by estimating parameter θ. *Assume there is an unbiased estimator* $\hat{\theta}$ *that has a Normal distribution.* When we standardize it, we get a Standard Normal variable

$$Z = \frac{\hat{\theta} - \mathbf{E}(\hat{\theta})}{\mathrm{Std}(\hat{\theta})} = \frac{\hat{\theta} - \theta}{\mathrm{Std}(\hat{\theta})}. \tag{9.2}$$

This variable falls between the quantiles $q_{\alpha/2}$ and $q_{1-\alpha/2}$, denoted by

$$\begin{aligned} -z_{\alpha/2} &= q_{\alpha/2} \\ z_{\alpha/2} &= q_{1-\alpha/2} \end{aligned}$$

with probability $(1-\alpha)$, as you can see in Figure 9.3.

Then,

$$\boldsymbol{P}\left\{ -z_{\alpha/2} \le \frac{\hat{\theta} - \theta}{\mathrm{Std}(\hat{\theta})} \le z_{\alpha/2} \right\} = 1 - \alpha.$$

Solving the inequality for θ, we get

$$\boldsymbol{P}\left\{ \hat{\theta} - z_{\alpha/2}\,\mathrm{Std}(\hat{\theta}) \le \theta \le \hat{\theta} - z_{\alpha/2}\,\mathrm{Std}(\hat{\theta}) \right\} = 1 - \alpha.$$

The problem is solved! We have obtained two numbers

$$\begin{aligned} a &= \hat{\theta} - z_{\alpha/2}\,\mathrm{Std}(\hat{\theta}) \\ b &= \hat{\theta} + z_{\alpha/2}\,\mathrm{Std}(\hat{\theta}) \end{aligned}$$

such that

$$\boldsymbol{P}\left\{a \le \theta \le b\right\} = 1 - \alpha.$$

Confidence interval, Normal distribution

If parameter θ has an unbiased, Normally distributed estimator $\hat{\theta}$, then

$$\hat{\theta} \pm z_{\alpha/2}\,\mathrm{Std}(\hat{\theta}) \;=\; \left[\hat{\theta} - z_{\alpha/2}\,\mathrm{Std}(\hat{\theta}),\; \hat{\theta} + z_{\alpha/2}\,\mathrm{Std}(\hat{\theta})\right]$$

is a $(1-\alpha)100\%$ confidence interval for θ.

If the distribution of $\hat{\theta}$ is *approximately* Normal, we get an *approximately* $(1-\alpha)100\%$ confidence interval.

(9.3)

In this formula, $\hat{\theta}$ is the **center of the interval**, and $z_{\alpha/2}\,\mathrm{Std}(\hat{\theta})$ is the **margin**. The margin of error is often reported along with poll and survey results. It is usually computed for a 95% confidence interval.

We have seen quantiles $\pm z_{\alpha/2}$ in inverse problems (Example 4.12 on p. 100). Now, in confidence estimation, and also, in the next section on hypothesis testing, they will play a crucial role as we'll need to attain the desired confidence level α. The most commonly used values are

$$\begin{aligned} z_{0.10} &= 1.282 \\ z_{0.05} &= 1.645 \\ z_{0.025} &= 1.960 \\ z_{0.01} &= 2.326 \\ z_{0.005} &= 2.576 \end{aligned} \tag{9.4}$$

Several important applications of this general method are discussed below. In each problem, we

(a) find an unbiased estimator of θ,

(b) check if it has a Normal distribution,

(c) find its standard deviation $\mathrm{Std}(\hat{\theta})$,

(d) obtain quantiles $\pm z_{\alpha/2}$ from the table of Normal distribution (Table A4 in the Appendix), and finally,

(e) apply the rule (9.3).

9.2.2 Confidence interval for the population mean

Let us construct a confidence interval for the population mean

$$\theta = \mu = \mathbf{E}(X).$$

Start with an estimator,

$$\hat{\theta} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

The rule (9.3) is applicable in two cases.

1. If a sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ comes from Normal distribution, then $\bar{X}$ is also Normal, and rule (9.3) can be applied.
2. If a sample comes from *any* distribution, but the sample size n is large, then $\bar{X}$ has an approximately Normal distribution according to the Central Limit Theorem on p. 101. Then rule (9.3) gives an approximately $(1-\alpha)100\%$ confidence interval.

In Section 8.2.1, we derived

$$\begin{array}{rcll} \mathbf{E}(\bar{X}) & = & \mu & \text{(thus, it is an unbiased estimator);} \\ \mathrm{Std}(\bar{X}) & = & \sigma/\sqrt{n}. & \end{array}$$

Then, (9.3) reduces to the following $(1-\alpha)100\%$ confidence interval for μ.

Confidence interval for the mean; σ is known

$$\boxed{\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}} \qquad (9.5)$$

Example 9.11. Construct a 95% confidence interval for the population mean based on a sample of measurements

2.5, 7.4, 8.0, 4.5, 7.4, 9.2

if measurement errors have Normal distribution, and the measurement device guarantees a standard deviation of $\sigma = 2.2$.

Solution. This sample has size $n = 6$ and sample mean $\bar{X} = 6.50$. To attain a confidence level of

$$1 - \alpha = 0.95,$$

we need $\alpha = 0.05$ and $\alpha/2 = 0.025$. Hence, we are looking for quantiles

$$q_{0.025} = -z_{0.025} \quad \text{and} \quad q_{0.975} = z_{0.025}.$$

From (9.4) or Table A4, we find that $q_{0.975} = 1.960$. Substituting these values into (9.5), we obtain a 95% confidence interval for μ,

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 6.50 \pm (1.960)\frac{2.2}{\sqrt{6}} = \underline{6.50 \pm 1.76} \text{ or } \underline{[4.74,\ 8.26]}.$$

◇

The only situation when method (9.3) cannot be applied is when the sample size is small and the distribution of data is not Normal. Special methods for the given distribution of $\boldsymbol{X}$ are required in this case.

9.2.3 Confidence interval for the difference between two means

Under the same conditions as in the previous section,

– Normal distribution of data or
– sufficiently large sample size,

we can construct a confidence interval for the *difference* between two means.

This problem arises when we compare two populations. It may be a comparison of two materials, two suppliers, two service providers, two communication channels, two labs, etc. From each population, a sample is collected,

$$\begin{array}{ll} \boldsymbol{X} = (X_1, \ldots, X_n) & \text{from one population,} \\ \boldsymbol{Y} = (Y_1, \ldots, Y_m) & \text{from the other population.} \end{array}$$

Suppose that the two samples are collected independently of each other.

To construct a confidence interval for the difference between population means

$$\theta = \mu_X - \mu_Y,$$

we complete the usual steps (a)–(e) below.

(a) Propose an estimator of θ,

$$\hat{\theta} = \bar{X} - \bar{Y}.$$

It is natural to come up with this estimator because $\bar{X}$ estimates μ_X and $\bar{Y}$ estimates μ_Y.

(b) Check that $\hat{\theta}$ is unbiased. Indeed,

$$\mathbf{E}(\hat{\theta}) = \mathbf{E}\left(\bar{X} - \bar{Y}\right) = \mathbf{E}\left(\bar{X}\right) - \mathbf{E}\left(\bar{Y}\right) = \mu_X - \mu_Y = \theta.$$

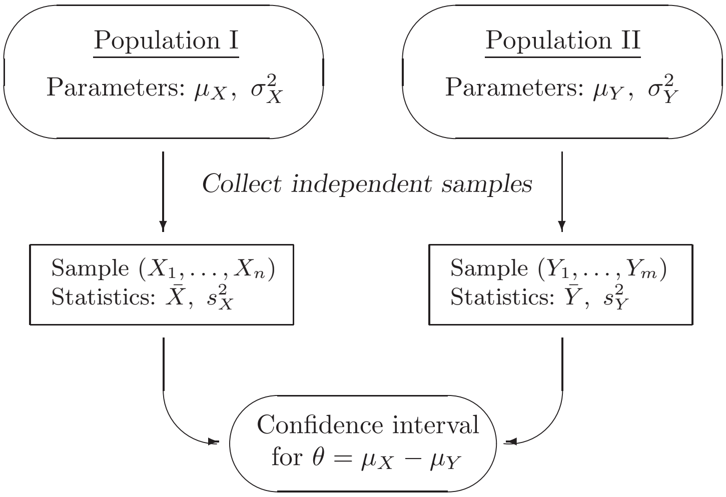

Population I
Parameters: $\mu_X,\ \sigma_X^2$
Population II
Parameters: $\mu_Y,\ \sigma_Y^2$
Collect independent samples
Sample $(X_1,\ldots,X_n)$
Statistics: $\bar{X},\ s_X^2$
Sample $(Y_1,\ldots,Y_m)$
Statistics: $\bar{Y},\ s_Y^2$
Confidence interval
for $\theta = \mu_X - \mu_Y$

Figure 9.4 *Comparison of two populations.*

(c) Check that $\hat{\theta}$ has a Normal or approximately Normal distribution. This is true if the observations are Normal or *both* sample sizes m and n are large.

(d) Find the standard deviation (using independence of $\boldsymbol{X}$ and $\boldsymbol{Y}$),

$$\text{Std}(\hat{\theta}) = \sqrt{\text{Var}\left(\bar{X}-\bar{Y}\right)} = \sqrt{\text{Var}\left(\bar{X}\right) + \text{Var}\left(\bar{Y}\right)} = \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}.$$

(e) Find quantiles $\pm z_{\alpha/2}$ and compute the confidence interval according to (9.3). This results in the following formula.

Confidence interval for the difference of means; known standard deviations

$$\boxed{\bar{X} - \bar{Y} \pm z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \tag{9.6}$$

Example 9.12 (Effect of an upgrade). A manager evaluates effectiveness of a major hardware upgrade by running a certain process 50 times before the upgrade and 50 times after it. Based on these data, the average running time is 8.5 minutes before the upgrade, 6.2 minutes after it. Historically, the standard deviation has been 1.8 minutes, and presumably it has not changed. Construct a 90% confidence interval showing how much the mean running time reduced due to the hardware upgrade.

Solution. We have $n = m = 50$, $\sigma_X = \sigma_Y = 1.8$, $\bar{X} = 8.5$, and $\bar{Y} = 6.2$. Also, the confidence level $(1 - \alpha)$ equals 0.9, hence $\alpha/2 = 0.05$, and $z_{\alpha/2} = 1.645$.

The distribution of times may not be Normal, however, due to large sample sizes, the estimator

$$\hat{\theta} = \bar{X} - \bar{Y}$$

is approximately Normal by the Central Limit Theorem. Thus, formula (9.6) is applicable, and a 90% confidence interval for the difference of means $(\mu_X - \mu_Y)$ is

$$8.5 - 6.2 \pm (1.645)\sqrt{1.8^2\left(\frac{1}{50} + \frac{1}{50}\right)} = \underline{2.3 \pm 0.6 \text{ or } [1.7, 2.9]}.$$

We can say that the hardware upgrade resulted in a 2.3-minute reduction of the mean running time, with a 90% confidence margin of 0.6 minutes. ◇

9.2.4 Selection of a sample size

Formula (9.3) describes a confidence interval as

$$\text{center} \pm \text{margin}$$

where

$$\begin{array}{rcl} \text{center} & = & \hat{\theta}, \\ \text{margin} & = & z_{\alpha/2}\,\text{Std}(\hat{\theta}). \end{array}$$

We can revert the problem and ask a very practical question: *How large a sample should be collected to provide a certain desired precision of our estimator?*

In other words, what sample size n guarantees that the margin of a $(1-\alpha)100\%$ confidence interval does not exceed a specified limit Δ?

To answer this question, we only need to solve the inequality

$$\text{margin} \le \Delta \tag{9.7}$$

for n. Typically, parameters are estimated more accurately based on larger samples, so that $\text{Std}(\hat{\theta})$ is a decreasing function of sample size n. Then, (9.7) is satisfied for sufficiently large n.

9.2.5 Estimating means with the given precision

When we estimate a population mean, the margin of error is

$$\text{margin} = z_{\alpha/2}\sigma/\sqrt{n}.$$

Solving inequality (9.7) for n results in the following rule.

Sample size for a given precision

In order to attain a margin of error Δ for estimating a population mean with a confidence level $(1-\alpha)$,

a sample of size $\quad n \geq \left(\frac{z_{\alpha/2}\sigma}{\Delta}\right)^2 \quad$ is required.

(9.8)

When we compute the expression in (9.8), it will most likely be a fraction. Notice that we can only *round it up* to the nearest integer sample size. If we round it down, our margin will exceed Δ.

Looking at (9.8), we see that a large sample is needed

- to attain a narrow margin,
- to attain a high confidence level, and
- to provide a reasonably narrow confidence interval under high variability of data (large σ).

In particular, we need to quadruple the sample size in order to half the margin of the interval.

Example 9.13. In Example 9.11, we constructed a 95% confidence with the center 6.50 and margin 1.76 based on a sample of size 6. Now, that was too wide, right? How large a sample do we need to estimate the population mean with a margin of at most 0.4 units with 95% confidence?

Solution. We have $\Delta = 0.4$, $\alpha = 0.05$, and from Example 9.11, $\sigma = 2.2$. By (9.8), we need a sample of

$$n \geq \left(\frac{z_{0.05/2}\sigma}{\Delta}\right)^2 = \left(\frac{(1.960)(2.2)}{0.4}\right)^2 = 116.2.$$

Keeping in mind that we can only round up, we need a sample of at least 117 observations. ◇

9.3 Unknown standard deviation

A rather heavy condition was assumed when we constructed all the confidence intervals. We assumed a *known standard deviation* σ and used it in all the derived formulas.

Sometimes this assumption is perfectly valid. We may know the variance from a large archive of historical data, or it may be given as precision of a measuring device.

Much more often, however, the population variance is unknown. We'll then estimate it from data and see if we can still apply methods of the previous section.

Two broad situations will be considered:

– large samples from any distribution,

– any sample from a Normal distribution.

In the only remaining case, a small non-Normal sample, usually a confidence interval can still be constructed by special methods.

9.3.1 Large samples

A large sample should produce a rather accurate estimator of a variance. We can then replace the true variance $\mathrm{Std}(\hat{\theta})$ in (9.3) by its estimator $\widehat{\mathrm{Std}(\hat{\theta})}$, and obtain an approximate confidence interval

$$\hat{\theta} \pm z_{\alpha/2}\,\widehat{\mathrm{Std}(\hat{\theta})}.$$

Example 9.14 (Delays at nodes). Internet connections are often slowed by delays at nodes. Let us determine if the delay time increases during heavy-volume times.

Five hundred packets are sent through the same network between 5 pm and 6 pm (sample $\boldsymbol{X}$), and three hundred packets are sent between 10 pm and 11 pm (sample $\boldsymbol{Y}$). The early sample has a mean delay time of 0.8 sec with a standard deviation of 0.1 sec whereas the second sample has a mean delay time of 0.5 sec with a standard deviation of 0.08 sec. Construct a 99.5% confidence interval for the difference between the mean delay times.

Solution. We have $n = 500$, $\bar{X} = 0.8$, $s_X = 0.1$; $m = 300$, $\bar{Y} = 0.5$, $s_Y = 0.08$. Large sample sizes allow us to replace unknown population standard deviations by their estimates and use an approximately Normal distribution of sample means.

For a confidence level of $1 - \alpha = 0.995$, we need

$$z_{\alpha/2} = z_{0.0025} = q_{0.9975}.$$

Look for the *probability* 0.9975 in the body of Table A4 and find the corresponding value of z,

$$z_{0.0025} = 2.81.$$

Then, a 99.5% confidence interval for the difference of mean execution times is

$$\begin{aligned}\bar{X}-\bar{Y} \quad &\pm \quad z_{0.0025}\sqrt{\frac{s_X^2}{n}+\frac{s_Y^2}{m}} = (0.8-0.5)\pm(2.81)\sqrt{\frac{(0.1)^2}{500}+\frac{(0.08)^2}{300}} \\ &= \quad \underline{0.3\pm 0.018} \text{ or } \underline{[0.282,\, 0.318]}.\end{aligned}$$

◇

9.3.2 Confidence intervals for proportions

In particular, we surely don't know the variance when we estimate a population proportion.

DEFINITION 9.5

We assume a subpopulation A of items that have a certain *attribute*. By the **population proportion** we mean the probability

$$p = \boldsymbol{P}\{i \in A\}$$

for a randomly selected item i to have this attribute.

A **sample proportion**

$$\hat{p} = \frac{\text{number of sampled items from } A}{n}$$

is used to estimate p.

It is convenient to use *indicator* variables

$$X_i = \begin{cases} 1 & \text{if} \quad i \in A \\ 0 & \text{if} \quad i \notin A \end{cases}$$

Each X_i has Bernoulli distribution with parameter p. In particular,

$$\mathbf{E}(X_i) = p \quad \text{and} \quad \mathrm{Var}(X_i) = p(1-p).$$

Also,

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is nothing but a sample mean of X_i.

Therefore,

$$\mathbf{E}(\hat{p}) = p \quad \text{and} \quad \mathrm{Var}(\hat{p}) = \frac{p(1-p)}{n},$$

as we know from properties of sample means on p. 227.

We conclude that

1. a sample proportion $\hat{p}$ is unbiased for the population proportion p;
2. it has approximately Normal distribution for large samples, because it has a form of a sample mean;
3. when we construct a confidence interval for p, we do not know the standard deviation $\mathrm{Std}\,(\hat{p})$.

Indeed, knowing the standard deviation is equivalent to knowing p, and if we know p, why would we need a confidence interval for it?

Thus, we estimate the unknown standard deviation

$$\mathrm{Std}\,(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

by

$$\widehat{\mathrm{Std}\,(\hat{p})} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

and use it in the general formula

$$\hat{p} \pm z_{\alpha/2}\,\widehat{\mathrm{Std}\,(\hat{p})}$$

to construct an approximate $(1-\alpha)100\%$ confidence interval.

Confidence interval
for a population proportion

$$\boxed{\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

Similarly, we can construct a confidence interval for the *difference between two proportions*. In two populations, we have proportions p_1 and p_2 of items with an attribute. Independent samples of size n_1 and n_2 are collected, and both parameters are estimated by sample proportions $\hat{p}_1$, $\hat{p}_2$.

Summarizing, we have

Parameter of interest:	$\theta = p_1 - p_2$
Estimated by:	$\hat{\theta} = \hat{p}_1 - \hat{p}_2$
Its variance:	$\mathrm{Var}(\hat{\theta}) = \dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}$
Estimated by:	$\widehat{\mathrm{Var}(\hat{\theta})} = \dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}$

Confidence interval for the difference of proportions

$$\boxed{\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

Example 9.15 (Pre-election poll). A candidate prepares for the local elections. During his campaign, 42 out of 70 randomly selected people in town A and 59 out of 100 randomly selected people in town B showed they would vote for this candidate. Estimate the difference in support this candidate is getting in towns A and B with 95% confidence. Can we state affirmatively that the candidate gets a stronger support in town A?

Solution. We have $n_1 = 70$, $n_2 = 100$, $\hat{p}_1 = 42/70 = 0.6$, and $\hat{p}_2 = 59/100 = 0.59$. For the confidence interval, we have

$$\text{center} = \hat{p}_1 - \hat{p}_2 = 0.01,$$

and

$$\begin{aligned} \text{margin} &= z_{0.05/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \\ &= (1.960)\sqrt{\frac{(0.6)(0.4)}{70} + \frac{(0.59)(0.41)}{100}} = 0.15. \end{aligned}$$

Then

$$0.01 \pm 0.15 = \underline{\text{[-0.14, 0.16]}}$$

is a 95% confidence interval for the difference in support $(p_1 - p_2)$ in the two towns.

So, is the support stronger in town A? On one hand, the estimator $\hat{p}_1 - \hat{p}_2 = 0.01$ suggests that the support is 1% higher in town A than in town B. On the other hand, the difference could appear positive just because of a sampling error. As we see, the 95% confidence interval includes a large range of negative values too. Therefore, the obtained data does *not* indicate affirmatively that the support in town A is stronger. (In fact, we attempted to test a hypothesis here and concluded that there was no evidence for it or against it. A formal procedure will be introduced in Section 9.4.) ◇

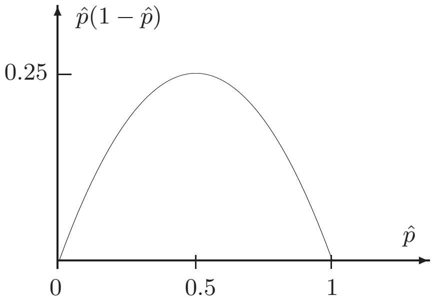


Figure 9.5 *Function* $\hat{p}(1-\hat{p})$ *attains its maximum at* $\hat{p} = 0.5$.

9.3.3 Estimating proportions with a given precision

Out confidence interval for a population proportion has a margin

$$\text{margin} = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

A standard way of finding the sample size that provides the desired margin Δ is to solve the inequality

$$\text{margin} \le \Delta \quad \text{or} \quad n \ge \hat{p}(1-\hat{p})\left(\frac{z_{\alpha/2}}{\Delta}\right)^2.$$

However, this inequality includes $\hat{p}$. To know $\hat{p}$, we first need to collect a sample, but to know the sample size, we first need to know $\hat{p}$!

A way out of this circle is shown in Figure 9.5. As we see, the function $\hat{p}(1-\hat{p})$ never exceeds 0.25. Therefore, we can replace the unknown value of $\hat{p}(1-\hat{p})$ by 0.25 and find a sample size n, perhaps larger than we actually need, that will ensure that we estimate $\hat{p}$ with a margin not exceeding Δ. That is, choose a sample size

$$\boxed{n \ge 0.25\left(\frac{z_{\alpha/2}}{\Delta}\right)^2,}$$

and it will automatically be at least as large as the required $\hat{p}(1-\hat{p})(z_{\alpha/2}/\Delta)^2$, regardless of the unknown value of $\hat{p}$.

Example 9.16. A sample of size

$$n \ge 0.25\left(\frac{1.960}{0.1}\right)^2 = 96.04$$

(that is, at least 97 observations) always guarantees that a population proportion is estimated with an error of at most 0.01 with a 95% confidence.

◇

9.3.4 Small samples: Student's t distribution

Having a small sample, we can no longer pretend that a sample standard deviation s is an accurate estimator of the population standard deviation σ. Then, how should we adjust the confidence interval when we replace σ by s, or more generally, when we replace $\text{Std}(\hat{\theta})$ by $\widehat{\text{Std}(\hat{\theta})}$?

A famous solution was proposed by *William Gosset* (1876–1937), known by his pseudonym Student. Working for the Irish brewery Guinness, he derived the T-distribution for the quality control problems in brewing.

Student followed the steps similar to our derivation of a confidence interval on p. 264, replaced the true but unknown standard deviation of $\hat{\theta}$ by its estimator $\widehat{\text{Std}(\hat{\theta})}$ and concluded that the **T-ratio**

$$t = \frac{\hat{\theta} - \theta}{\widehat{\text{Std}(\hat{\theta})}},$$

the *ratio* of two random variables, no longer has a Normal distribution!

Student figured the distribution of a T-ratio. For the problem of estimating the mean based on n Normal observations $X_1, \ldots, X_n$, this was **T-distribution** with $(n-1)$ *degrees of freedom*.

Table A6 gives critical values t_α of the T-distribution that we'll use for confidence intervals.

Confidence interval for the mean; σ is unknown

$$\bar{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

where $t_{\alpha/2}$ is a critical value from T-distribution with $n-1$ degrees of freedom

(9.9)

Example 9.17 (Unauthorized use of a computer account). If an unauthorized person accesses a computer account with the correct username and password (stolen or cracked), can this intrusion be detected? Recently, a number of methods have been proposed to detect such unauthorized use. The time between keystrokes, the time a key is depressed, the frequency of various keywords are measured and compared with the account owner. If there are noticeable differences, an intruder is detected.

The following times between keystrokes were recorded when a user typed the username and password:

.46, .38, .31, .24, .20, .31, .34, .42, .09, .18, .46, .21 seconds

As the first step in detecting if this is an intrusion, construct a 90% confidence interval for the mean time between keystrokes assuming Normal distribution of these times.

Solution. The sample size is $n = 12$, the sample mean time is $\bar{X} = 0.3$ sec, and the sample standard deviation is $s = 0.1183$. The critical value of t distribution with $n - 1 = 11$ degrees of freedom is $t_{\alpha/2} = t_{0.05} = 1.796$. Then, the 90% confidence interval for the mean time is

$$0.3 \pm (1.796)\frac{0.1183}{\sqrt{12}} = 0.3 \pm 0.0613 = \underline{[0.2387;\, 0.3613]}$$

Example 9.26 on p. 294 will show whether these data signal an intrusion. ◇

The density of Student's T-distribution is a bell-shaped symmetric curve that can be easily confused with Normal. Comparing with the Normal density, its peak is lower and its tails are thicker. Therefore, a larger number t_α is generally needed to cut area α from the right tail. That is,

$$t_\alpha > z_\alpha$$

for small α. As a consequence, the confidence interval (9.9) is wider than the interval (9.5) for the case of known σ. This wider margin is the price paid for not knowing the standard deviation σ. When we lack a certain piece of information, we cannot get a more accurate estimator.

However, we see in Table A6 that

$$t_\alpha \to z_\alpha,$$

as the number of degrees of freedom ν tends to infinity. Indeed, having a large sample (so large $\nu = n - 1$), we can count on a very accurate estimator of σ, and thus, the confidence interval is almost as narrow as for the known σ in this case.

Degrees of freedom ν is the parameter of T-distribution controlling the shape of the T-density curve. Its meaning is the *dimension* of a vector used to estimate the variance. Here we estimate σ^2 by a sample variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

and thus, we use a vector

$$\boldsymbol{X}' = \left(X_1 - \bar{X}, \ldots, X_n - \bar{X}\right).$$

The initial vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ has dimension n, therefore, it has n degrees of freedom. However, when the sample mean $\bar{X}$ is subtracted from each observation, there appears a linear relation among the elements,

$$\sum_{i=1}^{n}(X_i - \bar{X}) = 0.$$

We lose 1 degree of freedom due to this constraint; the vector $\boldsymbol{X}'$ belongs to a $(n-1)$-dimensional hyperplane, and this is why we have only $\nu = n - 1$ degrees of freedom.

In many similar problems, degrees of freedom can be computed as

$$\begin{array}{c}\text{number of}\\ \text{degrees of freedom}\end{array} = \text{sample size} - \begin{array}{c}\text{number of}\\ \text{estimated}\\ \text{location parameters}\end{array} \qquad (9.10)$$

9.3.5 Comparison of two populations

We now construct a confidence interval for the difference of two means $\mu_X - \mu_Y$, comparing the population of X's with the population of Y's.

Again, independent random samples are collected,

$$\boldsymbol{X} = (X_1, \ldots, X_n) \quad \text{and} \quad \boldsymbol{Y} = (Y_1, \ldots, Y_m),$$

one from each population, as in Figure 9.4 on p. 268. This time, however, population variances σ_X^2 and σ_Y^2 are unknown to us, and we use their estimates.

Two important cases need to be considered here. In one case, there exists an exact and simple solution based on T-distribution. The other case suddenly appears to be a famous *Behrens-Fisher problem*, where no exact solution exists, and only approximations are available.

Case 1. Equal variances

Suppose there are reasons to assume that the two populations have equal variances,

$$\sigma_X^2 = \sigma_Y^2 = \sigma^2.$$

For example, two sets of data are collected with the same measurement device, thus, measurements have different means but the same precision.

In this case, there is only one variance σ^2 to estimate instead of two. We should use both samples $\boldsymbol{X}$ and $\boldsymbol{Y}$ to estimate their common variance. This estimator of σ^2 is called a **pooled sample variance**, and it is computed as

$$s_p^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{m}(Y_i - \bar{Y})^2}{n+m-2} = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}. \tag{9.11}$$

Substituting this variance estimator in (9.6) for σ_X^2 and σ_Y^2, we get the following confidence interval.

Confidence interval for the difference of means; equal, unknown standard deviations

$$\bar{X} - \bar{Y} \pm t_{\alpha/2}\, s_p \sqrt{\frac{1}{n} + \frac{1}{m}}$$

where s_p is the *pooled standard deviation*, a root of the pooled variance in (9.11)

and $t_{\alpha/2}$ is a critical value from T-distribution with $n-1$ degrees of freedom

Example 9.18 (CD WRITER AND BATTERY LIFE). CD writing is energy consuming, therefore, it affects the battery lifetime on laptops. To estimate the effect of CD writing, 30 users are asked to work on their laptops until the "low battery" sign comes on.

Eighteen users without a CD writer worked an average of 5.3 hours with a standard deviation of 1.4 hours. The other twelve, who used their CD writer, worked an average of 4.8 hours with a standard deviation of 1.6 hours. Assuming Normal distributions with equal population variances ($\sigma_X^2 = \sigma_2^2$), construct a 95% confidence interval for the battery life reduction caused by CD writing.

Solution. Effect of the CD writer is measured by the reduction of the mean battery life. We have $n = 12$, $\bar{X} = 4.8$, $s_X = 1.6$ for users with a CD writer and $m = 18$, $\bar{Y} = 5.3$, $s_Y = 1.4$ for users without it. The pooled standard deviation is

$$s_p = \sqrt{\frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}} = \sqrt{\frac{(11)(1.6)^2 + (17)(1.4)^2}{28}} = 1.4818$$

(check: it has to be between s_X and s_Y). The critical value is $t_{0.025} = 2.048$ (use 28 d.f.). The 95% confidence interval for the difference between the mean battery lives is

$$(4.8 - 5.3) \pm (2.048)(1.4818)\sqrt{\frac{1}{18} + \frac{1}{12}} = -0.5 \pm 1.13 = \underline{[\text{-1.63; 0.63}]}.$$

◇

Remark: Let's discuss formula (9.11). First, notice that different sample means $\bar{X}$ and $\bar{Y}$ are used for X-terms and Y-terms. Indeed, our two populations may have different means. As we know, variance of any variable measures its deviation from its mean. Thus, from each observation we subtract its own mean-estimate.

Second, we lose 2 degrees of freedom due to the estimation of two means. Two constraints,

$$\sum_{i=1}^{n}(X_i - \bar{X}) = 0 \quad \text{and} \quad \sum_{i=1}^{m}(Y_i - \bar{Y}) = 0,$$

show that the number of **degrees of freedom** is only $(n+m-2)$ instead of $(n+m)$. We see this coefficient in the denominator, and it makes s_p^2 an unbiased estimator of σ^2 (see Exercise 9.6).

Case 2. Unequal variances

The most difficult case is when both variances are unknown and unequal. Confidence estimation of $\mu_X - \mu_Y$ is known as the *Behrens-Fisher problem*. Certainly, we can replace unknown variances σ_X^2, σ_Y^2 by their estimates s_X^2, s_Y^2 and form a T-ratio

$$t = \frac{\left(\bar{X} - \bar{Y}\right) - (\mu_X - \mu_Y)}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}}.$$

However, it won't have a T-distribution.

An approximate solution was proposed by Satterthwaite, who used the method of moments to estimate degrees of freedom ν of a T-distribution that is "closest" to this T-ratio. This number depends on unknown variances. Estimating them by sample variances, Satterthwaite obtained the formula

$$\nu = \frac{\left(\dfrac{s_X^2}{n} + \dfrac{s_Y^2}{m}\right)^2}{\dfrac{s_X^4}{n^2(n-1)} + \dfrac{s_Y^4}{m^2(m-1)}}. \qquad (9.12)$$

This number of degrees of freedom often appears non-integer. There are T-distributions with non-integer ν, see Section 11.1.1. To use Table A6, take the closest ν given in the table.

Formula (9.12) is widely used for t-intervals and t-tests.

Confidence interval for the difference of means; unequal, unknown standard deviations

$$\bar{X} - \bar{Y} \pm t_{\alpha/2}\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}$$

where $t_{\alpha/2}$ is a critical value from T-distribution with ν degrees of freedom given by formula (9.12)

Example 9.19 (Comparison of two servers). An account on server A is more expensive than an account on server B. However, server A is faster. To see if whether it's optimal to go with the faster but more expensive server, a manager needs to know how much faster it is. A certain computer algorithm is executed 20 times on server A and 30 times on server B with the following results,

	Server A	**Server B**
Sample mean	6.7 min	7.5 min
Sample standard deviation	0.6 min	1.2 min

Construct a 95% confidence interval for the difference $\mu_1 - \mu_2$ between the mean execution times on server A and server B.

Solution. We have $n = 30$, $m = 20$, $\bar{X} = 6.7$, $\bar{Y} = 7.5$, $s_X = 0.6$, and $s_Y = 1.2$. The second standard deviation is twice larger than the first one, therefore, equal variances population variances can hardly be assumed. We use the method for unknown, unequal variances.

Using Satterthwaite approximation (9.12), we find degrees of freedom:

$$\nu = \frac{\left(\dfrac{(0.6)^2}{30} + \dfrac{(1.2)^2}{20}\right)^2}{\dfrac{(0.6)^4}{30^2(29)} + \dfrac{(1.2)^4}{20^2(19)}} = 25.4.$$

To use Table A6, we round this ν to 25 and find $t_{0.025} = 2.060$. Then, the confidence interval is

$$\begin{aligned}
\bar{X} - \bar{Y} \pm t_{\alpha/2}\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}} &= 6.7 - 7.5 \pm (2.060)\sqrt{\frac{(0.6)^2}{30} + \frac{(1.2)^2}{20}} \\
&= \underline{-0.8 \pm 0.6} \quad \text{or} \quad \underline{[-1.4, -0.2]}.
\end{aligned}$$

◇

9.4 Hypothesis testing

A vital role of Statistics is in verifying statements, claims, conjectures, and in general - *testing hypotheses*. Based on a random sample, we can use Statistics to verify whether

- a system has not been infected,
- a hardware upgrade was efficient,
- the average number of concurrent users increased by 2000 this year,
- the average connection speed is 54 Mbps, as claimed by the internet service provider,
- the proportion of defective products is at most 3%, as promised by the manufacturer,
- service times have Gamma distribution,
- the number of errors in software is independent of the manager's experience,
- etc.

Testing statistical hypotheses has wide applications far beyond Computer Science. These methods are used to prove efficiency of a new medical treatment, safety of a new automobile brand, innocence of a defendant, and authorship of a document; to establish cause-and-effect relationships; to identify factors that can significantly improve the response; to fit stochastic models; to detect information leaks; and so forth.

9.4.1 Hypothesis and alternative

To begin, we need to state exactly what we are testing. These are *hypothesis* and *alternative*.

NOTATION	H_0	=	hypothesis (the null hypothesis)
	H_A	=	alternative (the alternative hypothesis)

H_0 and H_A are simply two mutually exclusive statements. Each test results either in acceptance of H_0 or its rejection in favor of H_A.

A null hypothesis is always an equality, absence of an effect or relation, some "normal," usual statement that people have believed in for years. In order to overturn the common belief and to reject the hypothesis, we need *significant evidence*. Such an evidence can only be provided by data. Only when such evidence is found, and when it strongly supports the alternative H_A, can the hypothesis H_0 be rejected in favor of H_A.

Based on a random sample, a statistician cannot tell whether the hypothesis is true or the alternative. We need to see the entire population to tell that.

The purpose of each test is to determine whether the data provides sufficient evidence against H_0 in favor of H_A.

This is similar to a criminal trial. Typically, the jury cannot tell whether the defendant is guilty or innocent. It is not their task. They are only required to determine if the presented evidence against the defendant is sufficient. By default, called *presumption of innocence*, insufficient evidence leads to acquittal.

Example 9.20. To verify that the the average connection speed is 54 Mbps, we test the hypothesis $H_0: \ \mu = 54$ against the alternative $H_A: \ \mu \neq 54$, where μ is the average speed of all connections.

However, if we worry about a *low* connection speed only, we can conduct a one-sided test of

$$H_0: \ \mu = 54 \ \text{ vs } \ H_A: \ \mu < 54.$$

In this case, we only measure the amount of evidence supporting the alternative $H_A: \ \mu < 54$. In the absence of such evidence, we gladly accept the null hypothesis. ◇

DEFINITION 9.6

Alternative of the type $H_A: \ \mu \neq \mu_0$ covering regions on both sides of the hypothesis ($H_0: \ \mu = \mu_0$) is a **two-sided alternative**.

Alternative $H_A: \ \mu < \mu_0$ covering the region to the left of H_0 is **one-sided, left-tail**.

Alternative $H_A: \ \mu > \mu_0$ covering the region to the right of H_0 is **one-sided, right-tail**.

Example 9.21. To verify whether the average number of concurrent users increased by 2000, we test

$$H_0: \ \mu_2 - \mu_1 = 2000 \ \text{ vs } \ H_A: \ \mu_2 - \mu_1 \neq 2000,$$

where μ_1 is the average number of concurrent users last year, and μ_2 is the average number of concurrent users this year. Depending on the situation, we may replace the *two-sided alternative* $H_A: \ \mu_2 - \mu_1 \neq 2000$ with a one-sided alternative $H_A: \ \mu_2 - \mu_1 < 2000$ or $H_A: \ \mu_2 - \mu_1 > 2000$. ◇

Example 9.22. To verify if the proportion of defective products is at most 3%, we test

$$H_0: \ p = 0.03 \ \text{ vs } \ H_A: \ p > 0.03,$$

where p is the proportion of defects in the whole shipment.

Why do we choose the *right-tail alternative* $H_A: \ p > 0.03$? That is because we reject the shipment only if significant evidence supporting this alternative is collected. We don't need to collect such evidence to accept the shipment.

◇

9.4.2 Type I and Type II errors: level of significance

When testing hypotheses, we realize that all we see is a random sample. Therefore, with all the best statistics skills, our decision to accept or to reject H_0 may still be wrong. This is a *sampling error* (Section 8.1).

Four situations are possible,

	Result of the test	
	Reject H_0	**Accept** H_0
H_0 **is true**	*Type I error*	correct
H_0 **is false**	correct	*Type II error*

In two of the four cases, the test results in a *correct decision*. Either we accepted a true hypothesis, or we rejected a false hypothesis. The other two situations are sampling errors.

DEFINITION 9.7

A **type I error** occurs when we reject the true hypothesis.

A **type II error** occurs when we accept the false hypothesis.

Each error occurs with a certain probability that we hope to keep small. A good test results in an erroneous decision only if the observed data are somewhat extreme.

A type I error is often considered more dangerous and undesired than a type II error. Making a type I error can be compared with convicting an innocent defendant or sending a patient to a surgery when (s)he does not need one.

For this reason, we shall design tests that bound the probability of type I error by a pre-assigned small number α. Under this conditions, we may want to minimize the probability of type II error.

DEFINITION 9.8

Probability of a type I error is the **significance level** of a test,

$$\alpha = \boldsymbol{P}\left\{\text{reject } H_0 \mid H_0 \text{ is true}\right\}.$$

Probability of rejecting a false hypothesis is the **power of the test**.

Typically, hypotheses are tested at significance levels as small as 0.01, 0.05, or 0.10, although there are exceptions. Testing at a low level of significance means that only a large amount of evidence can force rejection of H_0. Rejecting a hypothesis at a very low level of significance is done with a lot of confidence that this decision is right.

9.4.3 Level α tests: general approach

A standard algorithm for a level α test of a hypothesis H_0 against an alternative H_A consists of 3 steps.

Step 1. Test statistic

Testing hypothesis is based on a **test statistic** T, a quantity computed from the data that has some known, tabulated distribution F_0 if the hypothesis H_0 is true.

Test statistics are used to discriminate between the hypothesis and the alternative. When we verify a hypothesis about some parameter θ, the test statistic is usually obtained by a suitable transformation of its estimator $\hat{\theta}$.

We find a suitable test statistic and compute it from the data.

Step 2. Acceptance region and rejection region

Next, we consider the **null distribution** F_0. This is the distribution of test statistic T when the hypothesis H_0 is true. If it has a density f_0, then the whole area under the density curve is 1, and we can always find a portion of it whose area is $(1-\alpha)$, as shown in Figure 9.6. It is called **acceptance region**.

The remaining part, the complement of the acceptance region, is called **rejection region**. By the complement rule, its area is α.

As another example, look at Figure 9.3 on p. 264. If the null distribution of T is *Standard Normal*, then the area between $(-z_{\alpha/2})$ and $z_{\alpha/2}$ equals exactly $(1-\alpha)$. The interval

$$A = [-z_{\alpha/2}, z_{\alpha/2}]$$

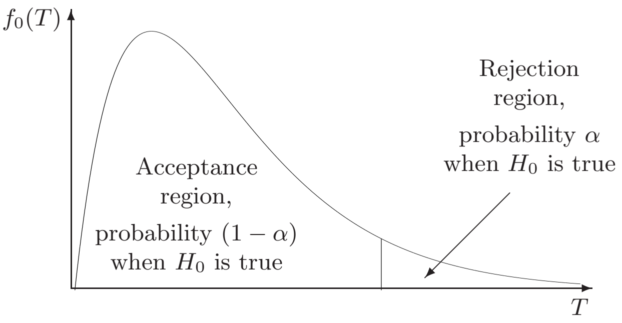


Figure 9.6 *Acceptance and rejection regions.*

can serve as a level α acceptance region, and the remaining part

$$R = \bar{A} = (-\infty, -z_{\alpha/2}) \cup (z_{\alpha/2}, +\infty)$$

is the rejection region.

Areas under the density curve are probabilities, and we conclude that

$$\boldsymbol{P}\{T \in \text{ acceptance region } \mid H_0\} = 1 - \alpha$$

and

$$\boldsymbol{P}\{T \in \text{ rejection region } \mid H_0\} = \alpha.$$

Step 3: Result and its interpretation

Accept the hypothesis H_0 if the test statistic T belongs to the acceptance region. Reject H_0 in favor of the alternative H_A if T belongs to the rejection region.

Our acceptance region guarantees that the significance level of our test is

$$\begin{aligned} \text{Significance level} &= \boldsymbol{P}\{\text{ Type I error }\} \\ &= \boldsymbol{P}\{\text{ Reject } \mid H_0\} \\ &= \boldsymbol{P}\{T \in \text{ rejection region } \mid H_0\} \\ &= \alpha. \end{aligned} \tag{9.13}$$

Therefore, indeed, we have a level α test!

The interesting part is to interpret our result correctly. Notice that conclusions like *"My level α test accepted the hypothesis. Therefore, the hypothesis is true with probability* $(1-\alpha)$" are *wrong*! Statements H_0 and H_A are about a non-random population, and thus, the hypothesis can either be true with probability 1 or false with probability 1.

If the test rejects the hypothesis, all we can state is that the data provides sufficient evidence against H_0. It may either happen because H_0 is not true, or because our sample is too extreme. The latter, however, can only happen with probability α.

If the test accepts the hypothesis, it only means that the evidence obtained from the data is not sufficient to reject it. In the absence of sufficient evidence, by default, we accept the null hypothesis.

Notation

α	=	level of significance, probability of type I error
T	=	test statistic
F_0, f_0	=	null distribution of T and its density
A	=	acceptance region
R	=	rejection region

9.4.4 Acceptance regions and power

Our construction of the acceptance region guaranteed the desired significance level α, as we proved in (9.13). However, many other regions will also have probability $(1-\alpha)$ (see Figure 9.7). Among them, how do we choose the best one?

To avoid *type II errors*, we choose such an acceptance region that will unlikely cover the test statistic T in case if the *alternative* H_A is true. This maximizes the *power* of our test because we'll rarely accept H_0 in this case.

Then, we look at our test statistic T under the alternative. Often

(a) a *right-tail alternative* forces T to be large,

(b) a *left-tail alternative* forces T to be small,

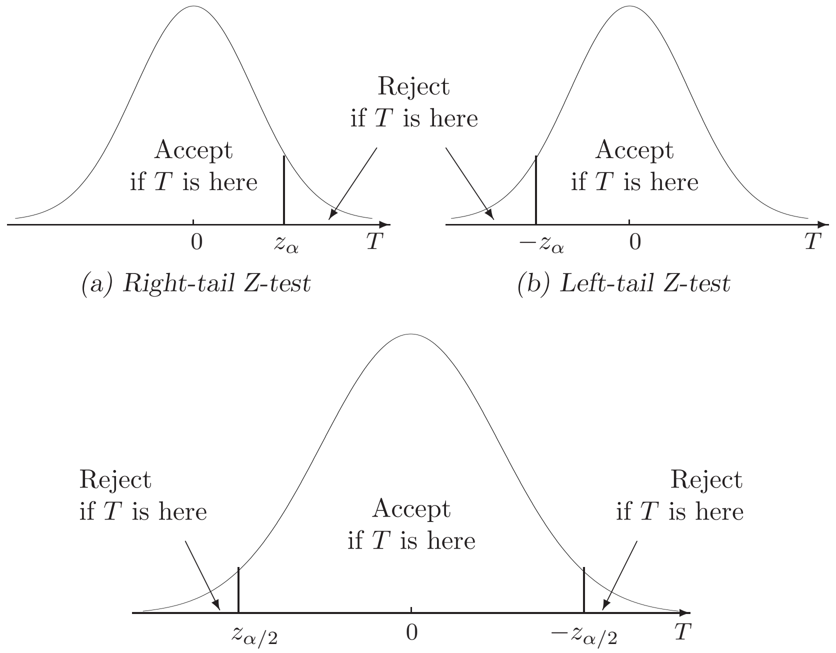
(c) a *two-sided alternative* forces T to be either large or small

(although it certainly depends on how we choose T). If this is the case, it tells us exactly when we should reject the null hypothesis:

(a) For a **right-tail alternative**, the rejection region R should consist of large values of T. Choose R on the right, A on the left (Figure 9.7a).

(b) For a **left-tail alternative**, the rejection region R should consist of small values of T. Choose R on the left, A on the right (Figure 9.7b).

(c) For a **two-sided alternative**, the rejection region R should consist of very small and very large values of T. Let R consist of two extreme regions, while A covers the middle (Figure 9.7c).



(a) Right-tail Z-test

(b) Left-tail Z-test

(c) Two-sided Z-test

Figure 9.7 *Acceptance and rejection regions for a Z-test with (a) a one-sided right-tail alternative; (b) a one-sided left-tail alternative; (c) a two-sided alternative.*

9.4.5 Standard Normal null distribution (Z-test)

An important case, in terms of a large number of applications, is when the null distribution of the test statistic is *Standard Normal.*

The test in this case is called a **Z-test**, and the test statistic is usually denoted by Z.

(a) A level α test with a **right-tail alternative** should

$$\left\{ \begin{array}{lll} \text{reject } H_0 & \text{if} & Z > z_\alpha \\ \text{accept } H_0 & \text{if} & Z \leq z_\alpha \end{array} \right. \tag{9.14}$$

The rejection region in this case consists of large values of Z only,

$$R = (z_\alpha, +\infty), \qquad A = (-\infty, z_\alpha].$$

See Figure 9.7a.

Under the null hypothesis, Z belongs to A and we accept the hypothesis with probability

$$\Phi(z_\alpha) = 1 - \alpha,$$

making the probability of false rejection (type I error) equal α .

For example, we use this acceptance region to test

$$H_0: \ \mu = \mu_0 \ \text{ vs } \ H_A: \ \mu > \mu_0.$$

(b) With a **left-tail alternative**, we should

$$\begin{cases} \text{reject } H_0 & \text{if} \quad Z < -z_\alpha \\ \text{accept } H_0 & \text{if} \quad Z \geq -z_\alpha \end{cases} \tag{9.15}$$

The rejection region consists of small values of Z only,

$$R = (-\infty, -z_\alpha), \qquad A = [-z_\alpha, +\infty).$$

Similarly, $\boldsymbol{P}\{Z \in A\} = 1 - \alpha$ under H_0, thus, the probability of type I error equals α.

For example, this is how we should test

$$H_0: \ \mu = \mu_0 \ \text{ vs } \ H_A: \ \mu < \mu_0.$$

(c) With a **two-sided alternative**, we

$$\begin{cases} \text{reject } H_0 & \text{if} \quad |Z| > z_{\alpha/2} \\ \text{accept } H_0 & \text{if} \quad |Z| \leq z_{\alpha/2} \end{cases} \tag{9.16}$$

The rejection region consists of very small and very large values of Z,

$$R = (-\infty, z_{\alpha/2}) \cup (z_{\alpha/2}, +\infty), \qquad A = [-z_{\alpha/2}, z_{\alpha/2}].$$

Again, the probability of type I error equals α in this case.

For example, we use this test for

$$H_0: \ \mu = \mu_0 \ \text{ vs } \ H_A: \ \mu \neq \mu_0.$$

This is easy to remember:

- for a two-sided test, divide α by two and use $z_{\alpha/2}$;
- for a one-sided test, use z_α keeping in mind that the rejection region consists of just one piece.

Now consider testing a hypothesis about a population parameter θ. Suppose that its estimator $\hat{\theta}$ has Normal distribution, at least approximately, and we know $\mathbf{E}(\hat{\theta})$ and $\operatorname{Var}(\hat{\theta})$ if the hypothesis is true.

Then the test statistic

$$Z = \frac{\hat{\theta} - \mathbf{E}(\hat{\theta})}{\sqrt{\operatorname{Var}(\hat{\theta})}} \tag{9.17}$$

has Standard Normal distribution, and we can use (9.14), (9.15), and (9.16) to construct acceptance and rejection regions for a level α test. We call Z a **Z-statistic**.

Examples of Z-tests are in the next section.

9.4.6 Z-tests for means and proportions

As we already know,

– sample means have Normal distribution when the distribution of data is Normal;

– sample means have approximately Normal distribution when they are computed from large samples (the distribution of data can be arbitrary);

– sample proportions have approximately Normal distribution when they are computed from large samples;

– this extends to differences between means and between proportions

(see Sections 8.2.1 and 9.2.2–9.3.2).

For all these cases, we can use a Z-statistic (9.17) and acceptance regions (9.14)–(9.16) to design a powerful level α test.

Z-tests are summarized in Table 9.1 (you can certainly derive them without our help; see Exercise 9.7).

Example 9.23 (Z-TEST ABOUT A POPULATION MEAN). The number of concurrent users for some internet service provider has always averaged 5000 with a standard deviation of 800. After an equipment upgrade, the average number of users at 100 randomly selected moments of time is 5200. Does it indicate, at a 5% level of significance, that the mean number of concurrent users has increased? Assume that the standard deviation of the number of concurrent users has not changed.

Solution. We test the null hypothesis $H_0: \ \mu = 5000$ against a *one-sided right-tail alternative* $H_A: \ \mu > 5000$, because we are only interested to know if the mean number of users μ has increased.

Step 1: Test statistic. We are given: $\sigma = 800$, $n = 100$, $\alpha = 0.05$, $\mu_0 = 5000$, and from the sample, $\bar{X} = 5200$. The test statistic is

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{5200 - 5000}{800/\sqrt{100}} = 2.5.$$

Step 2: Acceptance and rejection regions. The critical value is

$$z_\alpha = z_{0.05} = 1.645$$

(don't divide α by 2 because it is a one-sided test). With the right-tail alternative, we

$$\left\{ \begin{array}{lll} \text{reject } H_0 & \text{if} & Z > 1.645 \\ \text{accept } H_0 & \text{if} & Z \leq 1.645 \end{array} \right.$$

Step 3: Result. Our test statistic $Z = 2.5$ belongs to the *rejection region*,

Null hypothesis	Parameter, estimator	If H_0 is true:		Test statistic
H_0	θ, $\hat{\theta}$	$\mathbf{E}(\hat{\theta})$	$\mathrm{Var}(\hat{\theta})$	$Z = \dfrac{\hat{\theta} - \theta_0}{\sqrt{\mathrm{Var}(\hat{\theta})}}$
One-sample Z-tests for means and proportions, based on a sample of size n				
$\mu = \mu_0$	μ, $\bar{X}$	μ_0	$\dfrac{\sigma^2}{n}$	$\dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$
$p = p_0$	p, $\hat{p}$	p_0	$\dfrac{p(1-p)}{n}$	$\dfrac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$
Two-sample Z-tests comparing means and proportions of two populations, based on independent samples of size n and m				
$\mu_X - \mu_Y$ $= D$	$\mu_X - \mu_Y$, $\bar{X} - \bar{Y}$	D	$\dfrac{\sigma_X^2}{n} + \dfrac{\sigma_Y^2}{m}$	$\dfrac{\bar{X} - \bar{Y} - D}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}$
$p_1 - p_2$ $= D$	$p_1 - p_2$, $\hat{p}_1 - \hat{p}_2$	D	$\frac{p_1(1-p_1)}{n}$ $+\frac{p_2(1-p_2)}{m}$	$\dfrac{\hat{p}_1 - \hat{p}_2 - D}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{m}}}$

Table 9.1 *Summary of Z-tests.*

therefore, we *reject the null hypothesis*. The data (5200 users, on the average, at 100 times) provided sufficient evidence against the hypothesis that the mean number of users remains 5000. ◇

Example 9.24 (Two-sample Z-test of proportions). A quality inspector finds 10 defective parts in a sample of 500 parts received from manufacturer A. Out of 400 parts from manufacturer B, she finds 12 defective ones. A computer-making company uses these parts in their computers and claims that the quality of parts produced by A and B is the same. At the 5% level of significance, do we have enough evidence to disprove this claim?

Solution. We test $H_0 : p_A = p_B$, or $H_0 : p_A - p_B = 0$, against $H_A : p_A \neq p_B$. This is a two-sided test because no direction of the alternative has been indicated. We only need to verify whether or not the proportions of defective parts are equal for manufacturers A and B.

Step 1: Test statistic. We are given: $\hat{p}_A = 10/500 = 0.02$ from a sample of size $n = 500$; $\hat{p}_B = 12/400 = 0.03$ from a sample of size $m = 400$. The tested value is $D = 0$.

The test statistic equals

$$Z = \frac{\hat{p}_A - \hat{p}_B - D}{\sqrt{\frac{p_A(1-p_A)}{n} + \frac{p_B(1-p_B)}{m}}} = \frac{0.02 - 0.03}{\sqrt{\frac{(0.02)(0.98)}{500} + \frac{(0.03)(0.97)}{400}}} = -0.945.$$

Step 2: Acceptance and rejection regions. This is a two-sided test, thus we divide α by 2, find $z_{0.05/2} = z_{0.025} = 1.96$, and

$$\begin{cases} \text{reject } H_0 & \text{if} \quad |Z| > 1.96 \\ \text{accept } H_0 & \text{if} \quad |Z| \leq 1.96 \end{cases}$$

Step 3: Result. The evidence against H_0 is insufficient because $|Z| \leq 1.96$. Although *sample proportions* of defective parts are unequal, the difference between them appears too small to claim that *population proportions* are different. ◇

9.4.7 Pooled sample proportion

The test in Example 9.24 can be conducted differently and perhaps, more efficiently.

Indeed, we standardize the estimator $\hat{\theta} = \hat{p}_A - \hat{p}_B$ using its expectation $\mathbf{E}(\hat{\theta})$ and variance $\mathrm{Var}(\hat{\theta})$ under the null distribution, i.e., when H_0 is true. However, under the null hypothesis $p_A = p_B$, and therefore, $\hat{p}_A$ and $\hat{p}_B$ have the same variance.

What does it mean for us? It means that instead of two variances, we only need to estimate one common variance. First we estimate the common population proportion by the overall proportion of defective parts,

$$\hat{p}(\text{pooled}) = \frac{\text{number of defective parts}}{\text{total number of parts}} = \frac{n\hat{p}_A + m\hat{p}_B}{n+m}.$$

Then we estimate the common variance as

$$\widehat{\mathrm{Var}}(\hat{p}_A - \hat{p}_B) = \frac{\hat{p}(1-\hat{p})}{n} + \frac{\hat{p}(1-\hat{p})}{m} = \hat{p}(1-\hat{p})\left(\frac{1}{n} + \frac{1}{m}\right)$$

and use it for the Z-statistic,

$$Z = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n} + \frac{1}{m}\right)}}.$$

Example 9.25 (Example 9.24, continued). Here the pooled proportion equals

$$\hat{p} = \frac{10+12}{500+400} = 0.0244,$$

so that

$$Z = \frac{0.02 - 0.03}{\sqrt{(0.0244)(0.9756)\left(\frac{1}{500} + \frac{1}{400}\right)}} = -0.966.$$

This does not affect our result. We obtained a different value of Z-statistic, but it also belongs to the acceptance region. We still don't have a significant evidence against the equality of two population proportions. ◇

9.4.8 Unknown σ: T-tests

As we decided in Section 9.3, when we don't know the population standard deviation, we estimate it. The resulting *T-statistic* has the form

$$t = \frac{\hat{\theta} - \mathbf{E}(\hat{\theta})}{\sqrt{\widehat{\mathrm{Var}}(\hat{\theta})}}.$$

In the case *when the distribution of* $\hat{\theta}$ *is Normal*, the test is based on *Student's T-distribution* with acceptance and rejection regions according to the direction of H_A:

(a) For a **right-tail alternative**,

$$\begin{cases} \text{reject } H_0 & \text{if} \quad t > t_\alpha \\ \text{accept } H_0 & \text{if} \quad t \le t_\alpha \end{cases} \tag{9.18}$$

(b) For a **left-tail alternative**,

$$\begin{cases} \text{reject } H_0 & \text{if} \quad t < -t_\alpha \\ \text{accept } H_0 & \text{if} \quad t \ge -t_\alpha \end{cases} \tag{9.19}$$

(c) For a **two-sided alternative**,

$$\begin{cases} \text{reject } H_0 & \text{if} \quad |t| > t_{\alpha/2} \\ \text{accept } H_0 & \text{if} \quad |t| \le t_{\alpha/2} \end{cases} \tag{9.20}$$

Quantiles t_α and $t_{\alpha/2}$ are given in Table A6. As in Section 9.3.4, the number of degrees of freedom depends on the problem and the sample size, see Table 9.2 and formula (9.10).

As in Section 9.3.4, the **pooled sample variance**

$$s_p^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{m}(Y_i - \bar{Y})^2}{n+m-2} = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}$$

Hypothesis H_0	Conditions	Test statistic t	Degrees of freedom
$\mu = \mu_0$	Sample size n; unknown σ	$t = \dfrac{\bar{X} - \mu_0}{s/\sqrt{n}}$	$n - 1$
$\mu_X - \mu_Y = D$	Sample sizes n, m; unknown but equal $\sigma_X = \sigma_y$	$t = \dfrac{\bar{X} - \bar{Y} - D}{s_p\sqrt{\frac{1}{n} + \frac{1}{m}}}$	$n + m - 2$
$\mu_X - \mu_Y = D$	Sample sizes n, m; unknown, unequal $\sigma_X \neq \sigma_y$	$t = \dfrac{\bar{X} - \bar{Y} - D}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}}$	Formula (9.12)

Table 9.2 *Summary of T-tests.*

is computed for the case of equal unknown variances. When variances are not equal, degrees of freedom are computed by Satterthwaite approximation (9.12).

Example 9.26 (Unauthorized use of a computer account, continued). A long-time authorized user of the account makes 0.2 seconds between keystrokes. One day, the data in Example 9.17 on p. 276 are recorded as someone typed the correct username and password. At a 1% level of significance, is this an evidence of an unauthorized attempt?

Let us test

$$H_0:\ \mu = 0.2 \ \text{ vs } \ H_A:\ \mu \neq 0.2$$

at a significance level $\alpha = 0.01$. From Example 9.17, we have sample statistics $n = 12$, $\bar{X} = 0.3$ and $s = 0.1183$. Compute the T-statistic,

$$t = \frac{\bar{X} - 0.2}{s/\sqrt{n}} = \frac{0.3 - 0.2}{0.1183/\sqrt{12}} = 5.8565.$$

The acceptance region is $[-3.106, 3.106]$ (we used T-distribution with $12 - 1 =$ 11 degrees of freedom and $\alpha/2 = 0.005$ because of the two-sided alternative), and it does not include our test statistic.

Therefore, we reject the null hypothesis and conclude that *there is a significant evidence of an unauthorized use of that account.* ◇

Example 9.27 (CD writer and battery life). Does a CD writer consume extra energy, and therefore, does it reduce the battery life on a laptop?

Example 9.18 on p. 279 provides data on battery lives for laptops with a CD writer (sample $\boldsymbol{X}$) and without a CD writer (sample $\boldsymbol{Y}$):

$$n = 12,\ \bar{X} = 4.8,\ s_X = 1.6;\ m = 18,\ \bar{Y} = 5.3,\ s_Y = 1.4;\ s_p = 1.4818.$$

Testing

$$H_0: \ \mu_X = \mu_Y \quad \text{vs} \quad H_A: \ \mu_X < \mu_Y$$

at $\alpha = 0.05$, we obtain

$$t = \frac{\bar{X} - \bar{Y}}{s_p\sqrt{\frac{1}{n} + \frac{1}{m}}} = \frac{4.8 - 5.3}{(1.4818)\sqrt{\frac{1}{18} + \frac{1}{12}}} = -0.9054.$$

The acceptance region for this left-tail test is $[-z_\alpha, \infty) = [-1.645, \infty)$. We accept H_0 concluding that there is *no evidence that laptops with a CD writer have a shorter battery life*. ◇

Example 9.28 (Comparison of two servers, continued). Is server A faster in Example 9.19 on p. 281? Formulate and test the hypothesis at a level $\alpha = 0.05$.

Solution. To see if server A is faster, we need to test

$$H_0: \ \mu_X = \mu_Y \quad \text{vs} \quad H_A: \ \mu_X < \mu_Y.$$

This is the case of unknown, unequal standard deviations. In Example 9.19, we used Satterthwaite approximation for the number of degrees of freedom and obtained $\nu = 25.4$. We should accept the hypothesis if $t \geq -1.708$. Since

$$t = \frac{6.7 - 7.5}{\sqrt{\frac{(0.6)^2}{30} + \frac{(1.2)^2}{20}}} = -2.7603,$$

we reject H_0 and conclude that *server A is faster*. ◇

When the distribution of $\hat{\theta}$ is not Normal, the Student's *T-distribution* cannot be used. The distribution of a T-statistic and all its probabilities will be different from Student's T, and as a result, our test may not have the desired significance level.

9.4.9 Duality: two-sided tests and two-sided confidence intervals

An interesting fact can be discovered if we look into our derivation of tests and confidence intervals. It turns out that we can conduct two-sided tests using nothing but the confidence intervals!

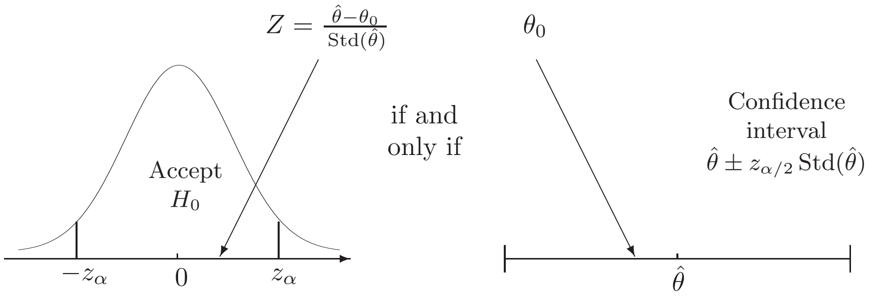

$Z = \frac{\hat\theta-\theta_0}{\text{Std}(\hat\theta)}$
Accept H_0
$-z_\alpha$
0
z_α
if and only if
θ_0
Confidence interval $\hat\theta \pm z_{\alpha/2}\,\text{Std}(\hat\theta)$
$\hat\theta$

Figure 9.8 *Duality of tests and confidence intervals.*

A level α Z-test of $H_0:\ \theta=\theta_0$ vs $H_A:\theta\neq\theta_0$
accepts the null hypothesis

if and only if

a symmetric $(1-\alpha)100\%$ confidence Z-interval for θ contains θ_0.

(9.21)

Proof: This rule is easy to understand.
The null hypothesis H_0 is accepted if and only if the Z-statistic belongs to the acceptance region, i.e.,

$$\left|\frac{\hat\theta-\theta_0}{\text{Std}(\hat\theta)}\right| \le z_{\alpha/2}.$$

This is equivalent to

$$\left|\hat\theta-\theta_0\right| \le z_{\alpha/2}\,\text{Std}(\hat\theta).$$

We see that the distance from θ_0 to the center of Z-interval $\hat\theta$ does not exceed its margin, $z_{\alpha/2}\,\text{Std}(\hat\theta)$ (see (9.3) and Figure 9.8). Therefore, θ_0 belongs to the Z-interval.

□

Rule (9.21) applies *only* when

- we are testing against a two-sided alternative (notice that our confidence intervals are two-sided too);
- significance level α of the test matches confidence level $(1-\alpha)$ of the confidence interval. For example, a two-sided 3% level test can be conducted using a 97% confidence interval.

Example 9.29. A sample of 6 measurements

$$2.5, 7.4, 8.0, 4.5, 7.4, 9.2$$

is collected from a Normal distribution with mean μ and standard deviation $\sigma = 2.2$. Test whether $\mu = 6$ against a two-sided alternative $H_A: \ \mu \neq 6$ at the 5% level of significance.

Solution. Solving Example 9.11 on p. 266, we have already constructed a 95% confidence interval for μ,

$$[4.74, \ 8.26].$$

The value of $\mu_0 = 6$ belongs to it, therefore, at the 5% level, the null hypothesis is accepted. $\diamond$

Example 9.30. Use data in Example 9.29 to test whether $\mu = 7$.

Solution. The interval $[4.74, \ 8.26]$ contains $\mu_0 = 7$ too, therefore, the hypothesis $H_0: \ \mu = 7$ is accepted as well. $\diamond$

In the last two examples, how could we possibly accept both hypotheses, $\mu = 6$ and $\mu = 7$? Obviously, μ cannot be equal 6 and 7 at the same time! This is true. By accepting both null hypotheses, we only acknowledge that the data does not present sufficient evidence against either of them.

Example 9.31 (Pre-election poll). In Example 9.15 on p. 274, we computed a 95% confidence interval for the difference of proportions supporting a candidate in towns A and B: $[-0.14, \ 0.16]$. This interval contains 0, therefore, the test of

$$H_0: \ p_1 = p_2 \ \text{ vs } \ H_A: \ p_1 \neq p_2$$

accepts the null hypothesis at the 5% level. Apparently, there is no evidence of unequal support of this candidate in the two towns. $\diamond$

Example 9.32 (Hardware upgrade). In Example 9.12, we studied effectiveness of the hardware upgrade. We constructed a 90% confidence interval for the difference $(\mu_X - \mu_Y)$ in mean running times of a certain process: $[1.7, 2.9]$.

So, can we conclude that the upgrade was successful? Ineffective upgrade corresponds to a null hypothesis $H_0: \ \mu_X = \mu_Y$, or $\mu_X - \mu_Y = 0$. Since the interval $[1.7, 2.9]$ does not contain 0, the no-effect hypothesis should be rejected at the 10% level of significance. $\diamond$

Example 9.33 (WAS THE UPGRADE SUCCESSFUL? EXAMPLE 9.32, CONTINUED). On the second thought, we can only use Rule (9.21) to test the **two-sided alternative** $H_A: \ \mu_X \neq \mu_Y$, right? At the same time, the hardware upgrade is successful only when the running time reduces, i.e., $\mu_X > \mu_Y$. Then, we can only judge effectiveness of the upgrade by a **one-sided, right-tail test** of

$$H_0: \ \mu_X = \mu_Y \ \text{ vs } \ H_A: \ \mu_X > \mu_Y. \tag{9.22}$$

Let us try to use the interval $[1.7, 2.9]$ for this test too. The null hypothesis in Example 9.32 is rejected at the 10% level in favor of a two-sided alternative, thus

$$|Z| > z_{\alpha/2} = z_{0.05}.$$

Then, either $Z < -z_{0.05}$ or $Z > z_{0.05}$. The first case is ruled out because the interval $[1.7, 2.9]$ consists of positive numbers, hence it cannot possibly support a left-tail alternative.

We conclude that $Z > z_{0.05}$, hence the test (9.22) results in rejection of H_0 at the 5% level of significance.

Conclusion. Our 90% confidence interval for $(\mu_X - \mu_Y)$ shows significant evidence, at the 5% level of significance, that the hardware upgrade was successful. ◇

Similarly, for the case of unknown variance(s),

A level α T-test of $H_0: \ \theta = \theta_0$ vs $H_A: \theta \neq \theta_0$
accepts the null hypothesis

if and only if

a symmetric $(1-\alpha)100\%$ confidence T-interval for θ contains θ_0.

Example 9.34 (UNAUTHORIZED USE OF A COMPUTER ACCOUNT, CONTINUED). A 90% confidence interval for the mean time between keystrokes is

$$[0.2387; \ 0.3613]$$

(Example 9.17 on p. 276). Example 9.26 on p. 294 tests whether such data could have been obtained if the account owner used the account who has a mean time of 0.2 seconds. The interval does not contain 0.2. Therefore, at a 10% level of significance, we have significant evidence that *the account was used by an unauthorized person.* ◇

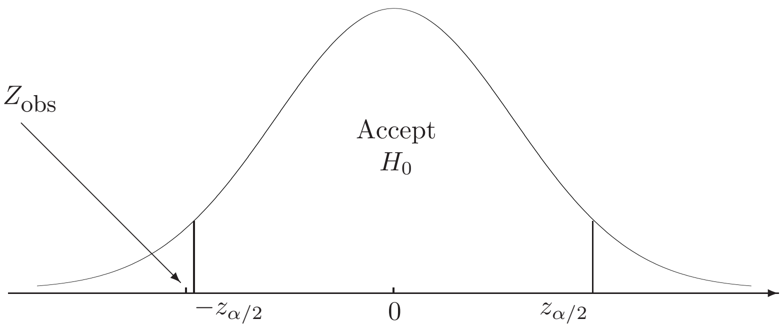


Figure 9.9 *This test is "too close to call": formally we reject the null hypothesis but Z-statistic is almost on the boundary.*

9.4.10 P-value

How do we choose α?

So far, we were testing hypotheses by means of acceptance and rejection regions. In the last section, we learned how to use confidence intervals for two-sided tests. Either way, we need to know the *significance level* α in order to conduct a test. Results of our test depend on it.

How do we choose α, the probability of making type I sampling error, rejecting the true hypothesis? Of course, when it seems too dangerous to reject true H_0, we choose a low significance level. How low? Should we choose $\alpha = 0.01$? Perhaps, 0.001? Or even 0.0001?

Also, if our *observed* test statistic $Z = Z_{\text{obs}}$ belongs to a rejection region but it is "too close to call" (see, for example, Figure 9.9), then how do we report the result? Formally, we should reject the null hypothesis, but practically, we realize that a slightly different significance level α could have expanded the acceptance region just enough to cover Z_{obs} and force us to accept H_0.

Suppose that the result of our test is crucially important. For example, the choice of a business strategy depends on it. In this case, can we rely so heavily on the choice of α? And if we rejected the true hypothesis just because we chose $\alpha = 0.05$ instead of $\alpha = 0.04$, then how do we explain that the situation was marginal? What is the statistical term for "too close to call"?

P-value

Using a P-value approach, we try not to rely on the level of significance. In fact, let us try to test a hypothesis using *all levels of significance*!

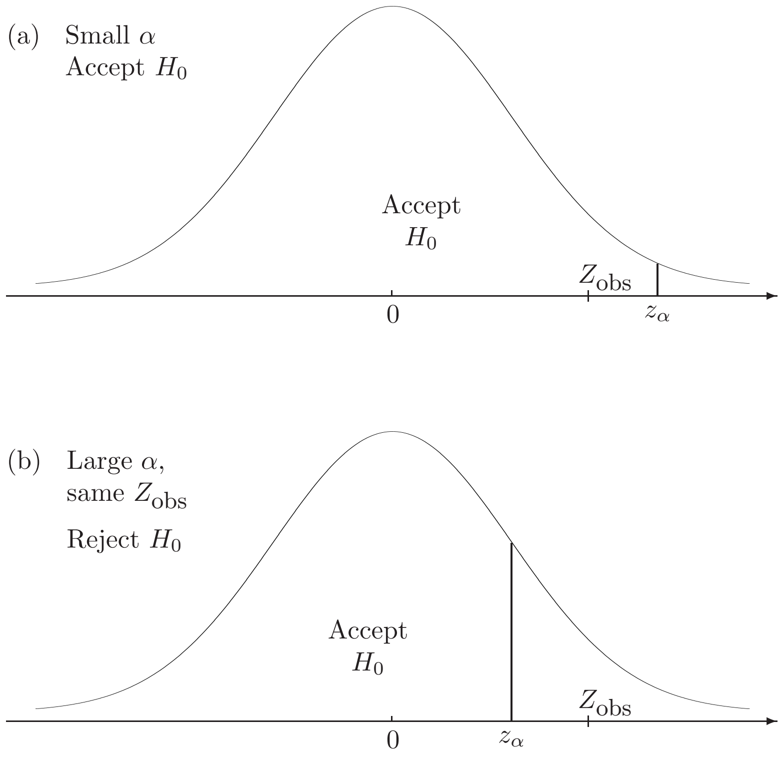


Figure 9.10 *(a) Under a low level of significance α, we accept the null hypothesis. (b) Under a high level of significance, we reject it.*

Considering all levels of significance (between 0 and 1 because α is a probability of Type I error), we notice:

Case 1. If a level of significance is *very low*, we *accept* the null hypothesis (see Figure 9.10a). A low value of

$$\alpha = \boldsymbol{P}\left\{ \text{ reject the null hypothesis when it is true } \right\}$$

makes it very unlikely to reject the hypothesis because it yields a very small rejection region. The right-tail area above the rejection region equals α.

Case 2. On the other extreme end, a *high significance level* α makes it likely to reject the null hypothesis and corresponds to a large rejection region. A sufficiently large α will produce such a large rejection region that will cover our test statistic, forcing us to *reject* H_0 (see Figure 9.10b).

Conclusion: there exists a boundary value between α-to-accept (case 1) and α-to-reject (case 2). This number is *a P-value* (Figure 9.11).

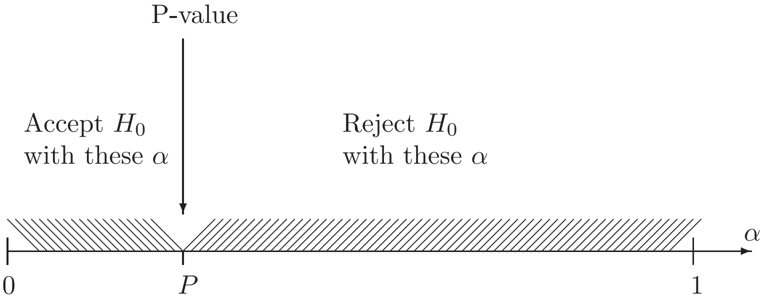

P-value
Accept H_0 with these α
Reject H_0 with these α
α
0
P
1

Figure 9.11 *P-value separates α-to-accept and α-to-reject.*

DEFINITION 9.9

> **P-value** is the lowest significance level α that forces rejection of the null hypothesis.
>
> **P-value** is also the highest significance level α that forces acceptance of the null hypothesis.

Testing hypotheses with a P-value

Once we know a P-value, we can indeed test hypotheses at *all* significance levels. Figure 9.11 clearly shows that for all $\alpha < P$ we accept the null hypothesis, and for all $\alpha > P$, we reject it.

Usual significance levels α lie in the interval [0.01, 0.1] (although there are exceptions, depending on the problem). Then, a P-value greater than 0.1 exceeds all natural significance levels, and the null hypothesis should be accepted. Conversely, if a P-value is less than 0.01, then it is smaller than all natural significance levels, and the null hypothesis should be rejected. Notice that we did not even have to specify the level α for these tests!

Only if the P-value happens to fall between 0.01 and 0.1, we really have to think about the level of significance. This is the "marginal case," "too close to call." When we report the conclusion, accepting or rejecting the hypothesis, we should always remember that with a slightly different α, the decision could have been reverted. When the matter is crucially important, a good decision is to collect more data until a more definitive answer can be obtained.

Testing H_0 with a P-value

For	$\alpha < P$,	accept H_0
For	$\alpha > P$,	reject H_0
Practically,		
If	$P < 0.01$,	reject H_0
If	$P > 0.1$,	accept H_0

Computing P-values

Here is how a P-value can be computed from data.

Let us look at Figure 9.10 again. Start from Figure 9.10a, gradually increase α, and keep your eye at the vertical bar separating the acceptance and rejection region. It will move to the left until it hits the observed test statistic Z_{obs}. At this point, our decision changes, and we switch from case 1 (Figure 9.10a) to case 2 (Figure 9.10b). Increasing α further, we pass the Z-statistic and start accepting the null hypothesis.

What happens at the border of α-to-accept and α-to-reject? Definition 9.9 says that this borderline α is the *P-value*,

$$P = \alpha.$$

Also, at this border our observed Z-statistic coincides with the critical value z_α,

$$Z_{\text{obs}} = z_\alpha,$$

and thus,

$$P = \alpha = \boldsymbol{P}\left\{Z > z_\alpha\right\} = \boldsymbol{P}\left\{Z > Z_{\text{obs}}\right\}.$$

In this formula, Z is any Standard Normal random variable, and Z_{obs} is our observed test statistic, which is a concrete number, computed from data. First, we compute Z_{obs}, then use Table A4 to calculate

$$\boldsymbol{P}\left\{Z > Z_{\text{obs}}\right\} = 1 - \Phi(Z_{\text{obs}}).$$

P-values for the left-tail one-sided and for the two-sided alternatives are computed similarly, as given in Table 9.3.

This table applies to all the Z-tests in this chapter. It can be directly extended to the case of unknown standard deviations and T-tests (Table 9.4).

Hypothesis H_0	Alternative H_A	P-value	Computation
$\theta = \theta_0$	right-tail $\theta > \theta_0$	$\boldsymbol{P}\{Z > Z_{\text{obs}}\}$	$1 - \Phi(Z_{\text{obs}})$
	left-tail $\theta < \theta_0$	$\boldsymbol{P}\{Z < Z_{\text{obs}}\}$	$\Phi(Z_{\text{obs}})$
	two-sided $\theta \neq \theta_0$	$\boldsymbol{P}\{\|Z\| > \|Z_{\text{obs}}\|\}$	$2(1 - \Phi(\|Z_{\text{obs}}\|))$

Table 9.3 *P-values for Z-tests.*

Hypothesis H_0	Alternative H_A	P-value	Computation
$\theta = \theta_0$	right-tail $\theta > \theta_0$	$\boldsymbol{P}\{t > t_{\text{obs}}\}$	$1 - F_\nu(t_{\text{obs}})$
	left-tail $\theta < \theta_0$	$\boldsymbol{P}\{t < t_{\text{obs}}\}$	$F_\nu(t_{\text{obs}})$
	two-sided $\theta \neq \theta_0$	$\boldsymbol{P}\{\|t\| > \|t_{\text{obs}}\|\}$	$2(1 - F_\nu(\|t_{\text{obs}}\|))$

Table 9.4 *P-values for T-tests (F_ν is the cdf of T-distribution with the suitable number ν of degrees of freedom).*

Understanding P-values

Looking at Tables 9.3 and 9.4, we see that *P-value* is the probability of observing a more extreme value of a test statistic than Z_{obs} or t_{obs}. Being "extreme" is determined by the alternative. For a right-tail alternative, large numbers are extreme; for a left-tail alternative, small numbers are extreme; and for a two-sided alternative, both large and small numbers are extreme. In general, the more extreme test statistic we observe, the stronger support of the alternative it provides.

This creates another interesting definition of a P-value.

DEFINITION 9.10

P-value is the probability of observing a more extreme test statistic than the test statistic computed from a given sample.

The following philosophy can be used when we test hypotheses by means of a P-value.

We are deciding between the null hypothesis H_0 and the alternative H_A. Observed is a test statistic Z_{obs}. If H_0 were true, how likely would it be to observe such a statistic?

A high P-value tells that this or even more extreme value of Z_{obs} is quite possible, and therefore, we see no contradiction with H_0. The null hypothesis is not rejected.

Conversely, a low P-value signals that such an extreme test statistic was unlikely if H_0 were true. However, we really observed it. Then, our data are in contradiction with the hypothesis, and we should reject it.

For example, if $P = 0.0001$, there is only 1 chance in 10,000 to observe what we really observed. The evidence supporting the alternative is highly significant in this case.

Example 9.35 (How significant was the upgrade?). Refer to Examples 9.12 and 9.32. At the 5% level of significance, we know that the hardware upgrade was successful. Was it marginally successful or very highly successful? Let us compute the P-value.

Start with computing a Z-statistic,

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} = \frac{8.5 - 6.2}{\sqrt{\frac{1.8^2}{50} + \frac{1.8^2}{50}}} = 6.39.$$

From Table A4, we find that the P-value for the right-tail alternative is

$$P = \boldsymbol{P}\left\{Z > Z_{\text{obs}}\right\} = \boldsymbol{P}\left\{Z > 6.39\right\} = 1 - \Phi(6.39) = 0.0000.$$

The P-value is extremely low, therefore, we can reject the null hypothesis not only at the 5%, but also at the 1% and even 0.01% level of significance! We see now that the hardware upgrade was extremely successful. ◇

Example 9.36 (Quality inspection). In Example 9.24, we compared the quality of parts produced by two manufacturers by a two-sided test. We obtained a test statistic

$$Z_{\text{obs}} = -0.94.$$

The P-value for this test equals

$$P = \boldsymbol{P}\left\{|Z| > |-0.94|\right\} = 2(1 - \Phi(0.94)) = 2(1 - 0.8264) = 0.3472.$$

This is a rather high P-value (greater than 0.1), and the null hypothesis is not rejected. If it is true, there is a 34% chance of observing what we really observed. No contradiction with H_0, and therefore, no evidence saying that the quality of parts is not the same. ◇

Table A6 is not as detailed as Table A4. Often we can only use it to bound the P-value from below and from above. Typically, it suffices for hypothesis testing.

Example 9.37 (UNAUTHORIZED USE OF A COMPUTER ACCOUNT, CONTINUED). How significant is the evidence in Example 9.26 on p. 294 that the account was used by an unauthorized person?

Under the null hypothesis, our T-statistic has T-distribution with 11 degrees of freedom. Comparing $t = 5.8565$ from Example 9.26 with the entire row 11 of Table A6, we find that it exceeds all the critical values including $t_{0.0001} = 5.453$. A two-sided test would reject the hypothesis even at a very low level $\alpha = 0.0002$. *The evidence of an unauthorized use is overwhelming!* ◇

9.5 Bayesian estimation and hypothesis testing

Interesting results and many new statistical methods can be obtained when we take a rather different look at statistical problems.

The difference is in our treatment of *uncertainty*.

So far, random samples were the only source of uncertainty in all the discussed statistical tools. The only distributions, expectations, and variances considered so far were distributions, expectations, and variances of data and various statistics computed from data. Population parameters were considered fixed. Statistical procedures were based on the distribution of data given these parameters,

$$f(\boldsymbol{x} \mid \theta) = f(X_1, \ldots, X_n \mid \theta).$$

This is the **frequentist approach**. According to it, all probabilities refer to random samples of data and possible long-run frequencies, and so do such concepts as unbiasedness, consistency, confidence level, and significance level:

- an estimator $\hat{\theta}$ is *unbiased* if in a long run of random samples, it averages to the parameter θ;
- a test has significance level α if in a long run of random samples, $100\alpha\%$ of times the true hypothesis is rejected;
- an interval has confidence level $(1 - \alpha)$ if in a long run of random samples, $(1 - \alpha)100\%$ of obtained confidence intervals contain the parameter, as shown in Figure 9.2, p. 263;

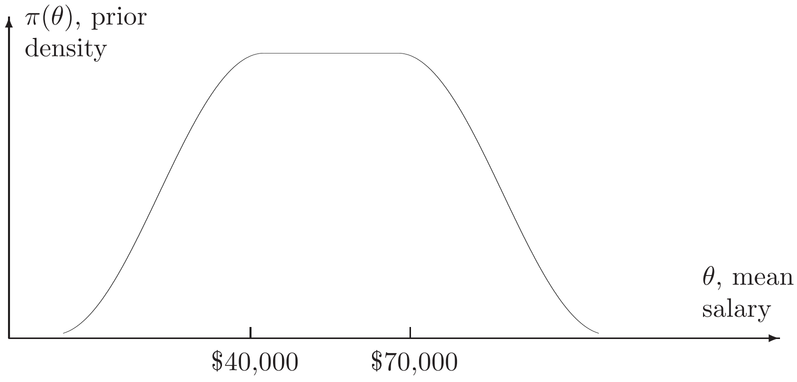


Figure 9.12 *Our prior distribution for the average starting salary.*

– and so on.

However, there is another approach: the **Bayesian approach**. According to it, uncertainty is also attributed to the unknown parameter θ. Some values of θ are more likely than others. Then, as long as we talk about the likelihood, we can define a whole distribution of values of θ. We call it a *prior distribution*, and it reflects our ideas, beliefs, and past experiences about the parameter before we collect and use the data.

Example 9.38 (Salaries). What do you think is the average starting annual salary of a Computer Science graduate? Is it \$20,000 per year? Unlikely, that's too low. Perhaps, \$200,000 per year? No, that's too high for a fresh graduate. Between \$40,000 and \$70,000 sounds like a reasonable range. We can certainly collect data on 100 recent graduates, compute their average salary and use it as an estimate, but before that, we already have our beliefs on what the mean salary may be. We can express it as some distribution with the most likely range between \$40,000 and \$70,000 (Figure 9.12). ◇

Collected data may change our idea about the parameter. Probabilities of different values of θ may change. Then we'll have a *posterior distribution* of θ.

One benefit of this approach is that we no longer have to explain our results in terms of a "long run." Often we collect just one sample for our analysis and don't experience any long runs of samples. Instead, with the Bayesian approach, we can state the result in terms of of the distribution of θ. For example, we clearly state the posterior probability for a parameter to belong to the obtained confidence interval, or the posterior probability that the hypothesis is true.

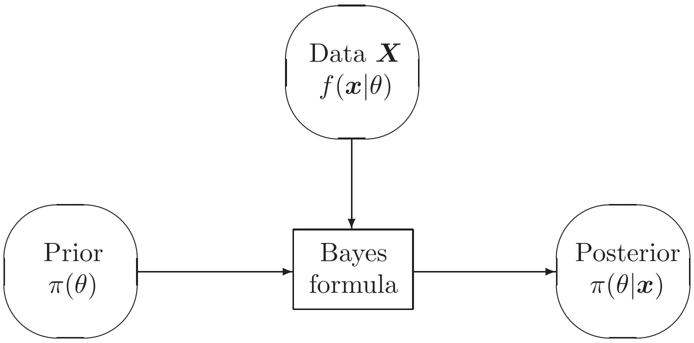


Figure 9.13 *Two sources of information about the parameter θ.*

9.5.1 Prior and posterior

Now we have two sources of information to use in our Bayesian inference:

1. collected and observed data;
2. prior distribution of the parameter.

These two pieces are combined via the **Bayes formula** (p. 31).

Prior to the experiment, our knowledge about the parameter θ is expressed in terms of the **prior distribution** (prior pmf or pdf)

$$\pi(\theta).$$

The observed sample of data $\boldsymbol{X} = (X_1, \ldots, X_n)$ has distribution (pmf or pdf)

$$f(\boldsymbol{x}|\theta) = f(x_1, \ldots, x_n|\theta).$$

This distribution is conditional on θ. That is, different values of the parameter θ generate different distributions of data, and thus, conditional probabilities about $\boldsymbol{X}$ generally depend on the condition, θ.

Observed data adds information about the parameter. The updated knowledge about θ can be expressed as the **posterior distribution**.

Posterior distribution

$$\boxed{\pi(\theta|\boldsymbol{x}) = \pi(\theta|\boldsymbol{X} = \boldsymbol{x}) = \frac{f(\boldsymbol{x}|\theta)\pi(\theta)}{m(\boldsymbol{x})}.} \tag{9.23}$$

Posterior distribution of the parameter θ is now conditioned on data $\boldsymbol{X} = \boldsymbol{x}$. Naturally, conditional distributions $f(\boldsymbol{x}|\theta)$ and $\pi(\theta|\boldsymbol{x})$ are related via the Bayes Rule (2.9).

According to the Bayes Rule, the denominator of (9.23), $m(\boldsymbol{x})$, represents the unconditional distribution of data $\boldsymbol{X}$. This is the **marginal distribution** (pmf or pdf) of the sample $\boldsymbol{X}$. Being unconditional means that it is constant for different values of the parameter θ. It can be computed by the *Law of Total Probability* (see p. 32) or its continuous-case version.

Marginal distribution of data

$$\boxed{\begin{array}{l} m(\boldsymbol{x}) = \displaystyle\sum_{\theta} f(x|\theta)\pi(\theta) \\ \quad \text{for discrete prior distributions } \pi \\[2ex] m(\boldsymbol{x}) = \displaystyle\int_{\theta} f(x|\theta)\pi(\theta)d\theta \\ \quad \text{for continuous prior distributions } \pi \end{array}} \tag{9.24}$$

Example 9.39 (Quality inspection). A manufacturer claims that the shipment contains only 5% of defective items, but the inspector feels that in fact it is 10%. We have to decide whether to accept or to reject the shipment based on θ, the proportion of defective parts.

Before we see the real data, let's assign a 50-50 chance to both suggested values of θ, i.e.,

$$\pi(0.05) = \pi(0.10) = 0.5.$$

A random sample of 20 parts has 3 defective ones. Calculate the posterior distribution of θ.

Solution. Apply the Bayes formula (9.23). Given θ, the distribution of the number of defective parts X is Binomial($n = 20, \theta$). For $x = 3$, from Table A2, we have

$$f(x \mid \theta = 0.05) = F(3 \mid \theta = 0.05) - F(2 \mid \theta = 0.05) = 0.9841 - 0.9245 = 0.0596$$

and

$$f(x \mid \theta = 0.10) = F(3 \mid \theta = 0.10) - F(2 \mid \theta = 0.10) = 0.8670 - 0.6769 = 0.1901.$$

The marginal distribution of X (for $x = 3$) is

$$\begin{aligned} m(3) &= f(x \mid 0.05)\pi(0.05) + f(x \mid 0.10)\pi(0.10) \\ &= (0.0596)(0.5) + (0.1901)(0.5) = 0.12485. \end{aligned}$$

Posterior probabilities of $\theta = 0.05$ and $\theta = 0.10$ are now computed as

$$\begin{aligned} \pi(0.05 \mid X = 3) &= \frac{f(x \mid 0.05)\pi(0.05)}{m(3)} = \frac{(0.0596)(0.5)}{0.1248} = 0.2387; \\ \pi(0.10 \mid X = 3) &= \frac{f(x \mid 0.10)\pi(0.10)}{m(3)} = \frac{(0.1901)(0.5)}{0.1248} = 0.7613. \end{aligned}$$

Conclusion. In the beginning, we had no preference between the two suggested values of θ. Then we observed a rather high proportion of defective parts, 3/20=15%. Taking this into account, $\theta = 0.10$ is now about three times as likely than $\theta = 0.05$. ◇

Notation

$\pi(\theta)$	$=$	prior distribution
$\pi(\theta \mid \boldsymbol{x})$	$=$	posterior distribution
$f(x\|\theta)$	$=$	distribution of data (model)
$m(\boldsymbol{x})$	$=$	marginal distribution of data
$\boldsymbol{X}$	$=$	$(X_1, \ldots, X_n)$, sample of data
$\boldsymbol{x}$	$=$	$(x_1, \ldots, x_n)$, observed values of $X_1, \ldots, X_n$.

9.5.2 Conjugate distribution families

A suitably chosen prior distribution of θ may lead to a very tractable form of the posterior.

DEFINITION 9.11

A family of prior distributions π is **conjugate** to the model $f(\boldsymbol{x}|\theta)$ if the posterior distribution belongs to the same family.

Three classical example of conjugate families are given below.

Gamma family is conjugate to the Poisson model

Let $(X_1, \ldots, X_n)$ be a sample from Poisson(θ) distribution with a Gamma(α, λ) prior distribution of θ.

Then

$$f(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta) = \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!} \sim e^{-n\theta}\theta^{\sum x_i}. \qquad (9.25)$$

Remark about dropping constant coefficients. In the end of (9.25), we dropped $(x_i!)$ and wrote that the result is "proportional" ($\sim$) to $e^{-n\theta}\theta^{\sum x_i}$. Dropping terms that don't contain θ often simplifies the computation. The form of the posterior distribution can be obtained without the constant term, and if needed, we can eventually evaluate the normalizing constant in the end, making $\pi(\theta|\boldsymbol{x})$ a fine distribution with the total probability 1, for example, as we did in Example 4.1 on p. 83. In particular, the marginal distribution

$m(\boldsymbol{x})$ can be dropped because it is θ-free. But keep in mind that in this case we obtain the posterior distribution "up to a constant coefficient."

The Gamma prior distribution of θ has density

$$\pi(\theta) \sim \theta^{\alpha-1} e^{-\lambda\theta}.$$

As a function of θ, this prior density has the same form as the model $f(\boldsymbol{x}|\theta)$. This is the general idea behind conjugate families.

Then, the posterior distribution of θ given $\boldsymbol{X} = \boldsymbol{x}$ is

$$\begin{array}{rcl} \pi(\theta|\boldsymbol{x}) & \sim & f(\boldsymbol{x}|\theta)\pi(\theta) \\ & \sim & \left(e^{-n\theta}\theta^{\sum x_i}\right)\left(\theta^{\alpha-1}e^{-\lambda\theta}\right) \\ & \sim & \theta^{\alpha+\sum x_i - 1} e^{-(\lambda+n)\theta}. \end{array}$$

Comparing with the general form of a Gamma density (say, (4.8) on p. 92), we see that $\pi(\theta|\boldsymbol{x})$ is the Gamma distribution with new parameters,

$$\alpha_x = \alpha + \sum_{i=1}^{n} x_i \text{ and } \lambda_x = \lambda + n.$$

We can conclude that:

1. Gamma family of prior distributions is conjugate to Poisson models.
2. Having observed a Poisson sample $\boldsymbol{X} = \boldsymbol{x}$, we update the Gamma$(\alpha, \lambda)$ prior distribution of θ to the Gamma$(\alpha + \sum x_i, \lambda + n)$ posterior.

Gamma distribution family is rather rich, it has two parameters. There is often a good chance to find a member of this large family that suitably reflects our knowledge about θ.

Example 9.40 (Network blackouts). The number of network blackouts each week has Poisson(θ) distribution. The weekly rate of blackouts θ is not known exactly, but according to the past experience with similar networks, it averages 4 blackouts with a standard deviation of 2.

There exists a Gamma distribution with the given mean $\mu = \alpha/\lambda = 4$ and standard deviation $\sigma = \sqrt{\alpha}/\lambda = 2$. Its parameters α and λ can be obtained by solving the system,

$$\left\{ \begin{array}{rcl} \alpha/\lambda & = & 4 \\ \sqrt{\alpha}/\lambda & = & 2 \end{array} \right. \Rightarrow \left\{ \begin{array}{rcccl} \alpha & = & (4/2)^2 & = & 4 \\ \lambda & = & 2^2/4 & = & 1 \end{array} \right.$$

We can assume the Gamma$(\alpha = 4,\ \lambda = 1)$ prior distribution θ. It is convenient to have a conjugate prior because the posterior will then belong to the Gamma family too.

Suppose there were $X_1 = 2$ blackouts this week. Given that, the posterior distribution of θ is Gamma with parameters

$$\alpha_x = \alpha + 2 = 6, \quad \lambda_x = \lambda + 1 = 2.$$

If no blackouts occur during the next week, the posterior parameters become

$$\alpha_x = \alpha + 2 + 0 = 6, \quad \lambda_x = \lambda + 2 = 3.$$

This posterior distribution has the average weekly rate of $6/3 = 2$ blackouts per week. Two weeks with very few blackouts reduced the average rate from 4 to 2. ◇

Beta family is conjugate to the Binomial model

A sample from Binomial(k, θ) distribution (assume k is known) has the probability mass function

$$f(\boldsymbol{x} \mid \theta) = \prod_{i=1}^{n} \left(\begin{array}{c} k \\ x_i \end{array} \right) \theta^{x_i} (1-\theta)^{k-x_i} \sim \theta^{\sum x_i} (1-\theta)^{nk - \sum x_i}.$$

Density of Beta(α, β) prior distribution has the same form, as a function of θ,

$$\pi(\theta) \sim \theta^{\alpha-1}(1-\theta)^{\beta-1} \quad \text{for } 0 < \theta < 1$$

(see Section 11.1.2 in the Appendix). Then, the posterior density of θ is

$$\pi(\theta \mid \boldsymbol{x}) \sim f(\boldsymbol{x} \mid \theta)\pi(\theta) \sim \theta^{\alpha + \sum_{i=1}^{n} x_i - 1} (1-\theta)^{\beta + nk - \sum_{i=1}^{n} x_i - 1},$$

and we recognize the Beta density with new parameters

$$\alpha_x = \alpha + \sum_{i=1}^{n} x_i \text{ and } \beta_x = \beta + nk - \sum_{i=1}^{n} x_i.$$

Hence,

1. Beta family of prior distributions is conjugate to the Binomial model.
2. Posterior parameters are $\alpha_x = \alpha + \sum x_i$ and $\beta_x = \beta + nk - \sum x_i$.

Normal family is conjugate to the Normal model

Consider now a sample from Normal distribution with an unknown mean θ and a known variance σ^2:

$$\begin{aligned} f(\boldsymbol{x} \mid \theta) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{(x_i - \theta)^2}{2\sigma^2} \right\} \sim \exp\left\{ -\sum_{i=1}^{n} \frac{(x_i-\theta)^2}{2\sigma^2} \right\} \\ &\sim \exp\left\{ \theta \frac{\sum x_i}{\sigma^2} - \theta^2 \frac{n}{2\sigma^2} \right\} = \exp\left\{ \left(\theta \bar{X} - \frac{\theta^2}{2} \right) \frac{n}{\sigma^2} \right\}. \end{aligned}$$

If the prior distribution of θ is also Normal, with prior mean μ and prior variance τ^2, then

$$\pi(\theta) \sim \exp\left\{-\frac{(\theta-\mu)^2}{2\tau^2}\right\} \sim \exp\left\{\left(\theta\mu - \frac{\theta^2}{2}\right)\frac{1}{\tau^2}\right\},$$

and again, it has a similar form as $f(\boldsymbol{x}|\theta)$.

The posterior density of θ equals

$$\begin{aligned}\pi(\theta \mid \boldsymbol{x}) \quad &\sim \quad f(\boldsymbol{x} \mid \theta)\pi(\theta) \sim \exp\left\{\theta\left(\frac{n\bar{X}}{\sigma^2} + \frac{\mu}{\tau^2}\right) - \frac{\theta^2}{2}\left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)\right\} \\ &\sim \quad \exp\left\{-\frac{(\theta-\mu_x)^2}{2\tau_x^2}\right\},\end{aligned}$$

where

$$\mu_x = \frac{n\bar{X}/\sigma^2 + \mu/\tau^2}{n/\sigma^2 + 1/\tau^2} \text{ and } \tau_x^2 = \frac{1}{n/\sigma^2 + 1/\tau^2}. \tag{9.26}$$

This posterior distribution is certainly Normal with parameters μ_x, τ_x.

Hence,

1. Normal family of prior distributions is conjugate to the Normal model with unknown mean;
2. Posterior parameters are given by (9.26).

We see that the posterior mean μ_x is a weighted average of the prior mean μ and the sample mean $\bar{X}$. This is how the prior information and the observed data are combined in case of Normal distributions.

Moreover, as the sample size n increases, we get more information from the sample, and as a result, the posterior mean converges to the frequentist estimator $\bar{X}$ as $n \to \infty$.

Posterior mean will also converge to $\bar{X}$ when $\tau \to \infty$. Large τ means a lot of uncertainty in the prior distribution of θ, thus, naturally, we should rather use observed data as a more reliable source of information in this case.

On the other hand, large σ indicates a lot of uncertainty in the observed sample. If that is the case, the prior distribution is more reliable, and as we see in (9.26), $\mu_x \approx \mu$ for large σ.

Results of this section are summarized in Table 9.5.

9.5.3 Bayesian estimation

We have already completed the most important step in Bayesian inference. We obtained the posterior distribution. All the knowledge about the unknown parameter is now included in the posterior, and that is what we'll use for further statistical analysis (Figure 9.14).

Model $f(\boldsymbol{x}\|\theta)$	Prior $\pi(\theta)$	Posterior $\pi(\theta\|\boldsymbol{x})$
Poisson(θ)	Gamma$(\alpha,\,\lambda)$	Gamma$(\alpha+n\bar{X},\,\lambda+n)$
Binomial(k,θ)	Beta$(\alpha,\,\beta)$	Beta$(\alpha+n\bar{X},\,\beta+n(k-\bar{X}))$
Normal(θ,σ)	Normal(μ,τ)	Normal$\left(\dfrac{n\bar{X}/\sigma^2+\mu/\tau^2}{n/\sigma^2+1/\tau^2},\,\dfrac{1}{\sqrt{n/\sigma^2+1/\tau^2}}\right)$

Table 9.5 *Three classical conjugate families.*

To estimate θ, we simply compute the **posterior mean**,

$$\hat{\theta}_{\mathrm{B}} = \mathbf{E}\left\{\theta|\boldsymbol{X}=\boldsymbol{x}\right\} = \begin{cases} \displaystyle\sum_{\theta}\theta\pi(\theta|\boldsymbol{x}) \quad = \quad \frac{\sum\theta f(\boldsymbol{x}|\theta)\pi(\theta)}{\sum f(\boldsymbol{x}|\theta)\pi(\theta)} \\ \qquad\qquad \text{if } \theta \text{ is discrete} \\ \\ \displaystyle\int_{\theta}\theta\pi(\theta|\boldsymbol{x})d\theta \quad = \quad \frac{\int\theta f(\boldsymbol{x}|\theta)\pi(\theta)d\theta}{\int f(\boldsymbol{x}|\theta)\pi(\theta)d\theta} \\ \qquad\qquad \text{if } \theta \text{ is continuous} \end{cases}$$

The result is a conditional expectation of θ given data $\boldsymbol{X}$. In abstract terms, the **Bayes estimator** $\hat{\theta}_{\mathrm{B}}$ is what we "expect" θ to be, after we observed a sample.

How accurate is this estimator? Among all estimates of θ, $\hat{\theta}_{\mathrm{B}} = \mathbf{E}\left\{\theta|\boldsymbol{x}\right\}$ has the lowest squared-error *posterior risk*

$$\mathbf{E}\left\{(\hat{\theta}-\theta)^2 \mid \boldsymbol{X}=\boldsymbol{x}\right\}$$

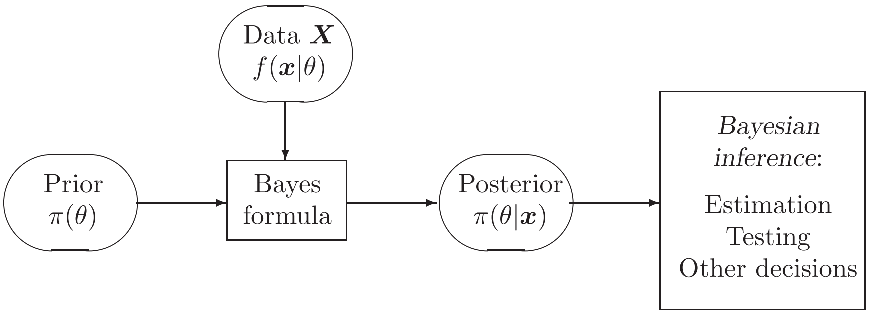


Figure 9.14 *Posterior distribution is the basis for Bayesian inference.*

and also, the lowest *Bayes risk*

$$\mathbf{E}\,\mathbf{E}(\hat{\theta}-\theta)^2,$$

where this double expectation is taken over the joint distribution of $\boldsymbol{X}$ and θ.

For the Bayes estimator $\hat{\theta}_{\mathrm{B}}$, posterior risk equals **posterior variance**, which shows variability of θ around $\hat{\theta}_{\mathrm{B}}$.

Example 9.41 (Normal case). The Bayes estimator of the mean θ of Normal(θ, σ) distribution with a Normal(μ, τ) prior is

$$\hat{\theta}_{\mathrm{B}} = \mu_x = \frac{n\bar{X}/\sigma^2 + \mu/\tau^2}{n/\sigma^2 + 1/\tau^2},$$

and its posterior risk is

$$\tau_x^2 = \frac{1}{n/\sigma^2 + 1/\tau^2},$$

(Table 9.5). As we expect, this risk decreases to 0 as the sample size grows to infinity. ◇

Example 9.42 (Network blackouts, continued). After two weeks of data, the weekly rate of network blackouts, according to Example 9.40 on p. 310, has Gamma posterior distribution with parameters $\alpha_x = 6$ and $\lambda_x = 3$.

The Bayes estimator of the weekly rate θ is

$$\hat{\theta}_{\mathrm{B}} = \mathbf{E}\,\{\theta|\boldsymbol{x}\} = \frac{\alpha_x}{\lambda_x} = 2 \text{ (blackouts per week)}$$

with a posterior risk of

$$\mathrm{Var}\,\{\theta|\boldsymbol{x}\} = \frac{\alpha_x}{\lambda_x^2} = 0.6667.$$

◇

Although conjugate priors simplify our statistics, Bayesian inference can certainly be done for other priors too.

Example 9.43 (Quality inspection, continued). In Example 9.39 on p. 308, we computed posterior distribution of the proportion of defective parts θ. This was a discrete distribution,

$$\pi(0.05 \mid \boldsymbol{x}) = 0.2387; \qquad \pi(0.10 \mid \boldsymbol{x}) = 0.7613.$$

Now, the Bayes estimator of θ is

$$\hat{\theta}_{\mathrm{B}} = \sum_{\theta} \theta\pi(\theta \mid \boldsymbol{x}) = (0.05)(0.2387) + (0.10)(0.7613) = 0.0881.$$

It does not agree with the manufacturer (who claims $\theta = 0.05$) or with the quality inspector (who feels that $\theta = 0.10$) but its value is much closer to the inspector's estimate.

The posterior risk of θ is

$$\begin{aligned} \text{Var}\{\theta|\boldsymbol{x}\} &= \mathbf{E}\left\{\theta^2|\boldsymbol{x}\right\} - \mathbf{E}^2\{\theta|\boldsymbol{x}\} \\ &= (0.05)^2(0.2387) + (0.10)^2(0.7613) - (0.0881)^2 = 0.0004, \end{aligned}$$

which means a rather low posterior standard deviation of 0.02. ◇

9.5.4 Bayesian credible sets

Confidence intervals have a totally different meaning in Bayesian analysis. Having a posterior distribution of θ, we no longer have to explain the confidence level $(1-\alpha)$ in terms of a long run of samples. Instead, we can give an interval $[a, b]$ or a set C that has a posterior probability $(1-\alpha)$ and state that *the parameter θ belongs to this set with probability* $(1-\alpha)$. Such a statement was impossible before we considered prior and posterior distributions. This set is called a $(1-\alpha)100\%$ *credible set*.

DEFINITION 9.12

Set C is a $(1-\alpha)100\%$ **credible set** for the parameter θ if the posterior probability for θ to belong to C equals $(1-\alpha)$. That is,

$$\boldsymbol{P}\{\theta \in C \mid \boldsymbol{X} = \boldsymbol{x}\} = \int_C \pi(\theta|\boldsymbol{x})d\theta = 1 - \alpha.$$

Such a set is not unique. Recall that for two-sided, left-tail, and right-tail hypothesis testing, we took different portions of the area under the Normal curve, all equal $(1-\alpha)$.

Minimizing the length of set C among all $(1-\alpha)100\%$ credible sets, we just have to include all the points θ with a high posterior density $\pi(\theta|\boldsymbol{x})$,

$$C = \{\theta : \ \pi(\theta|\boldsymbol{x}) \geq c\}$$

(see Figure 9.15). Such a set is called the **highest posterior density credible set**, or just the **HPD set**.

For the Normal(μ_x, τ_x) posterior distribution of θ, the $(1-\alpha)100\%$ HPD set is

$$\mu_x \pm z_{\alpha/2}\tau_x = [\mu_x - z_{\alpha/2}\tau_x, \mu_x + z_{\alpha/2}\tau_x].$$

Example 9.44 (Salaries, continued). In Example 9.38 on p. 306, we "decided" that the most likely range for the mean starting salary θ of Computer

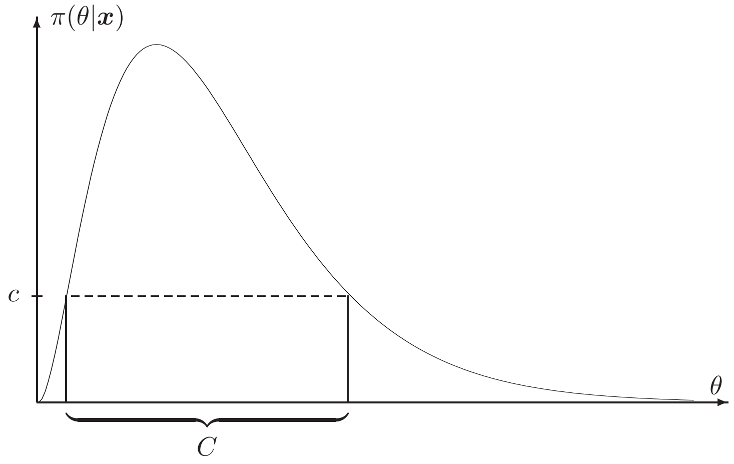


Figure 9.15 *The* $(1-\alpha)$*100% highest posterior density credible set.*

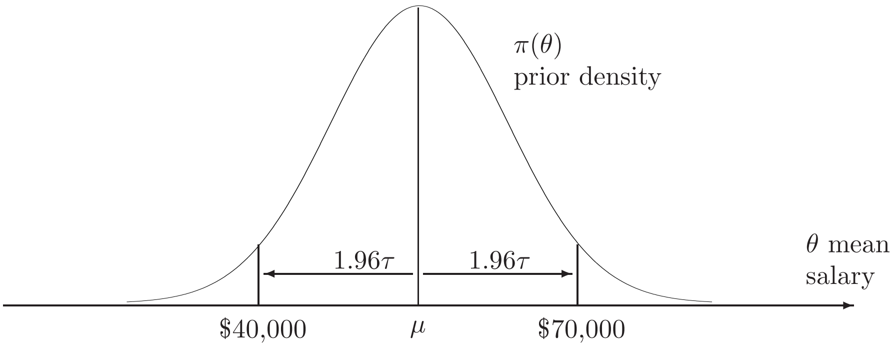


Figure 9.16 *Normal prior distribution and the 95% HPD credible set for the mean starting salary of Computer Science graduates (Example 9.44).*

Science graduates is between \$40,000 and \$70,000. Expressing this in a form of a prior distribution, we let the prior mean be $\mu = (40,000 + 70,000)/2 = 55,000$. Further, if we feel that the range $[40,000; 70,000]$ is 95% likely, and we accept a Normal prior distribution for θ, then this range should be equal

$$[40,000; 70,000] = \mu \pm z_{0.025/2}\tau = \mu \pm 1.96\tau,$$

where τ is the prior standard deviation (Figure 9.16). It can be computed from the information given,

$$\tau = \frac{70,000 - 40,000}{2(1.96)} = 7,653.$$

This is the advantage of using a rich (two-parameter) family of prior distributions: we are likely to find a member of this family that reflects our prior beliefs adequately.

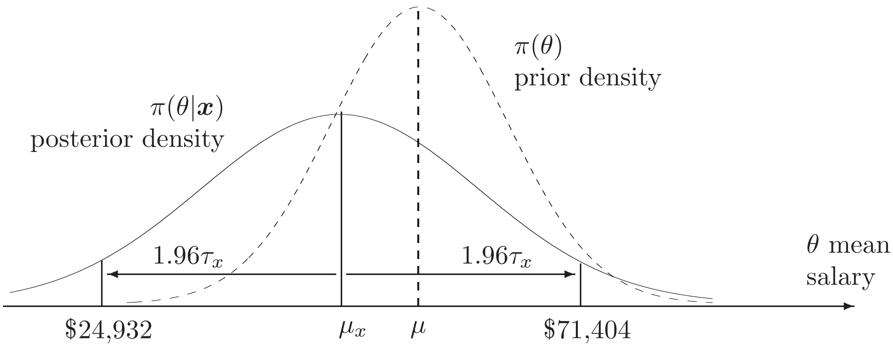


Figure 9.17 *Normal prior and posterior distributions for the mean starting salary (Example 9.44).*

Then, *prior to collecting any data*, the 95% HPD credible set of the mean starting salary θ is

$$\mu \pm z_{0.025}\tau = \underline{[40{,}000;\ 70{,}000]}.$$

Suppose a random sample of 100 graduates has the mean starting salary of $\bar{X} = 48,000$ with a sample standard deviation $s = 12,000$. From Table 9.5, we determine the posterior mean and standard deviation,

$$\begin{aligned} \mu_x &= \frac{n\bar{X}/\sigma^2 + \mu/\tau^2}{n/\sigma^2 + 1/\tau^2} = \frac{(100)(48,000)/(12,000)^2 + (55,000)/(7,653)^2}{100/(12,000)^2 + 1/(7653)^2} \\ &= 48,168; \\ \tau_x &= \frac{1}{\sqrt{n/\sigma^2 + 1/\tau^2}} = \frac{1}{\sqrt{100/(12,000)^2 + 1/(7653)^2}} = 11,855. \end{aligned}$$

We used the sample standard deviation s in place of the population standard deviation σ assuming that a sample of size 100 estimates the latter rather accurately. Alternatively, we could put a prior distribution on unknown σ too and estimate it by Bayesian methods. Since the observed sample mean is smaller than our prior mean, the resulting posterior distribution is shifted to the left of the prior (Figure 9.17).

Conclusion. After seeing the data, the Bayes estimator for the mean starting salary of CS graduates is

$$\hat{\theta}_{\mathrm{B}} = \mu_x = 48,168 \text{ dollars},$$

and the 95% HPD credible set for this mean salary is

$$\mu_x \pm z_{0.025}\tau_x = 48,168 \pm (1.96)(11,855) = 48,168 \pm 23,236 = \underline{[24{,}932;\ 71{,}404]}$$

Lower than a priori predicted starting salaries extended the lower end of our credible interval. ◇

Example 9.45 (TELEPHONE COMPANY). A new telephone company predicts to handle an average of 1000 calls per hour. During 10 randomly selected hours of operation, it handled 7265 calls.

How should it update the initial estimate of the frequency of telephone calls? Construct a 95% HPD credible set. Telephone calls are placed according to a Poisson process. Exponential prior distribution of the hourly rate of calls is applicable.

Solution. We need a Bayesian estimator of the frequency θ of telephone calls. The number of calls during 1 hour has Poisson(θ) distribution, where θ is unknown, with

$$\text{Exponential}(\lambda) = \text{Gamma}(1, \lambda)$$

prior distribution that has an expectation of

$$\mathbf{E}(\theta) = \frac{1}{\lambda} = 1000 \text{ calls.}$$

Hence, $\lambda = 0.001$. We observe a sample of size $n = 10$, totaling

$$\sum_{i=1}^{n} X_i = n\bar{X} = 7265 \text{ calls.}$$

As we know (see Table 9.5 on p. 313), the posterior distribution in this case is Gamma(α_x, λ_x) with

$$\begin{array}{rcl} \alpha_x & = & \alpha + n\bar{X} = 7266, \\ \lambda_x & = & \lambda + n = 10.001. \end{array}$$

This distribution has mean

$$\mu_x = \alpha_x / \lambda_x = 726.53$$

and standard deviation

$$\tau_x = \alpha_x / \lambda_x^2 = 72.65.$$

The Bayes estimator of θ is

$$\mathbf{E}(\theta|\boldsymbol{X}) = \mu_x = \underline{726.53 \text{ calls per hour}}.$$

It almost coincides with the sample mean $\bar{X}$ showing the sample was informative enough to dominate over the prior information.

For the credible set, we notice that α_x is sufficiently large to make the Gamma posterior distribution approximately equal the Normal distribution with parameters μ_x and τ_x. The 95% HPD credible set is then

$$\mu_x \pm z_{0.05/2}\tau_x = 726.53 \pm (1.96)(72.65) = 726.53 \pm 142.39 = \underline{[584.14,\ 868.92]}$$

◇

9.5.5 Bayesian hypothesis testing

Bayesian hypothesis testing is very easy to interpret. We can compute prior and posterior probabilities for the hypothesis H_0 and alternative H_A to be true and decide from there which one to accept or to reject.

Computing such probabilities was not possible without prior and posterior distributions of the parameter θ. In non-Bayesian statistics, θ was not random, thus H_0 and H_A were either true (with probability 1) or false (with probability 1).

For Bayesian tests, in order for H_0 and H_A to have meaningful, non-zero probabilities, they often represent disjoint sets of parameter values,

$$H_0: \ \theta \in \Theta_0 \ \text{ vs } \ H_A: \ \theta \in \Theta_1$$

(which makes sense because exact equality $\theta = \theta_0$ is unlikely to hold anyway, and in practice it is understood as $\theta \approx \theta_0$).

Comparing posterior probabilities of H_0 and H_A,

$$\boldsymbol{P}\{\Theta_0 \mid \boldsymbol{X} = \boldsymbol{x}\} \ \text{ and } \ \boldsymbol{P}\{\Theta_1 \mid \boldsymbol{X} = \boldsymbol{x}\},$$

we decide whether $\boldsymbol{P}\{\Theta_1 \mid \boldsymbol{X} = \boldsymbol{x}\}$ is large enough to present significant evidence and to reject the null hypothesis. One can again compare it with $(1-\alpha)$ (0.90, 0.95, 0.99, etc.) or state the result in terms of likelihood, “the null hypothesis is this much likely to be true.”

Example 9.46 (Quality inspection, continued). In Example 9.39 on p. 308, we are actually testing

$$H_0: \ \theta = 0.05 \ \text{ vs } \ H_A: \ \theta = 0.10$$

for the proportion θ of defective parts.

Example 9.43 gives posterior probabilities

$$\pi(\Theta_0 \mid \boldsymbol{X} = \boldsymbol{x}) = 0.2387 \ \text{ and } \ \pi(\Theta_1 \mid \boldsymbol{X} = \boldsymbol{x}) = 0.7613.$$

Thus, having observed the real data, we give only a 23.87% chance for the null hypothesis to be true. It may not be enough to reject the whole shipment, but it certainly questions the manufacturer claim and perhaps prompts for further data collection and more detailed inspection. ◇

Example 9.47 (Telephone company, continued). Let us test now whether the telephone company in Example 9.45 can actually face a call rate of 1000 calls *or more* per hour. We are testing

$$H_0: \ \theta \geq 1000 \ \text{ vs } \ H_A: \ \theta < 1000,$$

where θ is the hourly rate of telephone calls.

According to the Gamma(α_x, λ_x) posterior distribution of θ and its Normal $(\mu_x = 726.53, \tau_x = 72.65)$ approximation,

$$\boldsymbol{P}\{H_0 \mid \boldsymbol{X} = \boldsymbol{x}\} = \boldsymbol{P}\left\{\frac{\theta - \mu_x}{\tau_x} \geq \frac{1000 - \mu_x}{\tau_x}\right\} = 1 - \Phi(3.76) = 0.0001.$$

By the complement rule, $\boldsymbol{P}\{H_A \mid \boldsymbol{X} = \boldsymbol{x}\} = 0.9999$, and this presents a sufficient evidence against H_0.

We conclude that it's extremely unlikely for this company to face a frequency of 1000+ calls per hour. ◇

Summary and conclusions

After taking a general look at the data by methods of Chapter 8, we proceeded with the more advanced and informative statistical inference described in this chapter.

There is a number of methods for estimating the unknown population parameters. Each method provides estimates with certain good and desired properties. We learned three methods of **parameter estimation**.

Method of moments is based on matching the population and sample moments. It is relatively simple to implement, and it makes the estimated distribution "close" to the actual distribution of data in terms of their moments.

Maximum likelihood estimators maximize the probability of observing the sample that was really observed, thus making the actually occurring situation as likely as possible under the estimated distribution.

Bayesian parameter estimation combines the information contained in the prior distribution and the data. Bayesian inference is based on the posterior distribution.

In addition to parameter estimates, **confidence intervals** show the margin of error. A $(1-\alpha)100\%$ confidence interval contains the unknown parameter with probability $(1-\alpha)$. In a non-Bayesian setting, it means that $(1-\alpha)100\%$ of all confidence intervals constructed from a large number of samples should contain the parameter, and only $100\alpha\%$ may miss it.

Bayesian $(1-\alpha)100\%$ *credible sets* also contain the parameter θ with probability $(1-\alpha)$, but this time the probability refers to the distribution of θ. Explaining a $(1-\alpha)100\%$ credible set, we can say that given the observed data, θ belongs to the obtained set with probability $(1-\alpha)$.

We can use data to verify statements and **test hypotheses**. Essentially, we measure the evidence provided by the data against the *null hypothesis* H_0. Then we decide whether it is sufficient for rejecting H_0.

Given *significance level* α, we can construct acceptance and rejection regions, compute a suitable *test statistic*, and make a decision depending on which region it belongs to. Alternatively, we may compute a P-value of the test. It shows how significant the evidence against H_0 is. Low P-values suggest rejection of the null hypothesis. P-value is the boundary between levels α-to reject and α-to-accept. It also represents the probability of observing the same or more extreme sample than the one that was actually observed.

Depending on what we are testing against, the *alternative hypothesis* may be two-sided or one-sided. We account for the direction of the alternative when we construct acceptance and rejection regions and when we compute a P-value.

For *Bayesian hypothesis testing*, we compute posterior probabilities of H_0 and H_A and decide if the former is sufficiently smaller than the latter to force rejection of H_0.

Questions and exercises

9.1. Estimate the unknown parameter θ from a sample

$$3, 3, 3, 3, 3, 7, 7, 7$$

drawn from a discrete distribution with the probability mass function

$$\begin{array}{rcl} P(3) & = & \theta \\ P(7) & = & 1-\theta \end{array}$$

Obtain two estimators:

(a) the method of moments estimator

(b) the maximum likelihood estimator

9.2. The number of times a computer code is executed until it runs without errors has a Geometric distribution with unknown parameter p. For 5 independent computer projects, a student records the following numbers of runs:

$$3 \quad 7 \quad 5 \quad 3 \quad 2$$

Estimate p

(a) by the method of moments

(b) by the method of maximum likelihood

9.3. Use method of moments and method of maximum likelihood to estimate

(a) parameters a and b if a sample from Uniform(a, b) distribution is observed

(b) parameter λ if a sample from Exponential(λ) distribution is observed

(c) parameter μ if a sample from Normal(μ, σ) distribution is observed, and we already know σ

(d) parameter σ if a sample from Normal(μ, σ) distribution is observed, and we already know μ

(e) parameters μ and σ if a sample from Normal(a, b) distribution is observed, and both μ and σ are unknown

9.4. A sample of 3 observations $(X_1 = 0.4,\ X_2 = 0.7,\ X_3 = 0.9)$ is collected from a continuous distribution with density

$$f(x) = \begin{cases} \theta x^{\theta-1} & \text{if } \ 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

Estimate θ by *your favorite method.*

9.5. A sample $(X_1, ..., X_{10})$ is drawn from a distribution with a probability density function

$$\frac{1}{2}\left(\frac{1}{\theta}e^{-x/\theta} + \frac{1}{10}e^{-x/10}\right), \quad 0 < x < \infty$$

The sum of all 10 observations equals 150. Estimate θ by the method of moments.

9.6. Consider two populations (X's and Y's) with different means but the same variance. Two independent samples, sizes n and m, are collected. Show that the pooled variance estimator

$$s_p^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{m}(Y_i - \bar{Y})^2}{n + m - 2}$$

estimates their common variance unbiasedly.

9.7. Verify columns 3-5 in Table 9.1 on p. 291.

9.8. In order to ensure efficient usage of a server, it is necessary to estimate the mean number of concurrent users. According to records, the average number of concurrent users at 100 randomly selected times is 37.7, with a standard deviation $\sigma = 9.2$.

(a) Construct a 90% confidence interval for the expectation of the number of concurrent users.

(b) At the 1% significance level, do these data provide significant evidence that the mean number of concurrent users is greater than 35?

9.9. Installation of a certain hardware takes random time with a standard deviation of 5 minutes.

(a) A computer technician installs this hardware on 64 different computers, with the average installation time of 42 minutes. Give a 95% confidence interval for the mean installation time.

(b) Suppose that the population mean installation time is 40 minutes. A technician installs the hardware on your PC. What is the probability that the installation time will be within the interval computed in (a)?

9.10. Salaries of entry-level computer engineers have Normal distribution with unknown mean and variance. Three randomly selected computer engineers have salaries (in \$ 1000s):

$$30, 50, 70$$

(a) Construct a 90% confidence interval for the average salary of an entry-level computer engineer.

(b) Does this sample provide a significant evidence, at a 10% level of significance, that the average salary of all entry-level computer engineers is different from \$80,000? Explain.

9.11. We have to accept or reject a large shipment of items. For quality control purposes, we collect a sample of 200 items and find 24 defective items in it.

(a) Construct a 96% confidence interval for the proportion of defective items in the whole shipment.

(b) The manufacturer claims that at most one in 10 items in the shipment is defective. At the 4% level of significance, do we have sufficient evidence to disprove this claim? Do we have it at the 3% level?

9.12. Refer to Exercise 9.11. Having looked at the collected sample, we consider an alternative supplier. A sample of 150 items produced by the new supplier contains 13 defective items.

Is there a significant evidence that the quality of items produced by the new supplier is higher than the quality of items in Exercise 9.11? What is the P-value?

9.13. An electronic parts factory produces resistors. Statistical analysis of the output suggests that resistances follow an approximately Normal distribution with a standard deviation of 0.2 ohms. A sample of 52 resistors has the average resistance of 0.62 ohms.

(a) Based on these data, construct a 95% confidence interval for the population mean resistance.

(b) If the actual population mean resistance is exactly 0.6 ohms, what is the probability that an average of 52 resistances is 0.62 ohms or higher?

9.14. Is there a significant difference between two servers in Example 9.19 on p. 281?

(a) Use the confidence interval in Example 9.19 to conduct a two-sided test at the 5% level of significance.

(b) Compute a P-value of the two-sided test in (a).

(c) Is server A really faster? How strong is the evidence? Formulate the suitable hypothesis and alternative and compute the corresponding P-value.

State your conclusions in (a), (b), and (c).

9.15. According to Example 9.15 on p. 274, there is no significant difference, at the 5% level, between towns A and B in their support for the candidate. However, the level $\alpha = 0.05$ was chosen rather arbitrarily, and the candidate still does not know if he can trust the results when planning his campaign. Can we compare the two towns at *all* reasonable levels of significance? Compute the P-value of this test and state conclusions.

9.16. A sample of 250 items from lot A contains 10 defective items, and a sample of 300 items from lot B is found to contain 18 defective items.

(a) Construct a 98% confidence interval for the difference of proportions of defective items.

(b) At a significance level $\alpha = 0.02$, is there a significant difference between the quality of the two lots?

9.17. Compute a P-value for the right-tail test in Example 9.23 on p. 290 and state your conclusions about a significant increase in the number of concurrent users.

9.18. Consider the data about the number of blocked intrusions in Exercise 8.1, p. 250.

(a) Construct a 95% confidence interval for the difference between the average number of intrusion attempts per day before and after the change firewall setting (assume equal variances).

(b) Can we claim a significant reduction in the rate of intrusion attempts? The number of intrusion attempts each day has approximately Normal distribution. Compute P-values and state your conclusions under the assumption of equal variances and without it. Does this assumption make a difference?

9.19. *Inverse Gamma distribution* is given by the density

$$f(x) = \frac{e^{-1/x\beta}}{\Gamma(\alpha)\beta^{\alpha}x^{\alpha+1}} \text{ for } x > 0,$$

where α and β are positive parameters.

(a) Show that the Inverse Gamma family is conjugate to the Exponential(θ^{-1}) model. Compute posterior parameters.

(b) Show that it is also conjugate to a Gamma(r, θ^{-1}) model with known r and unknown θ. Compute posterior parameters.

(c) The expectation and variance of an Inverse Gamma random variable are

$$\mu = \frac{1}{\beta(\alpha-1)};\, \tau^2 = \frac{1}{\beta^2(\alpha-1)^2(\alpha-2)}.$$

Given a sample of Exponential(θ^{-1}) lifetimes

$$\boldsymbol{X} = (5,\ 3,\ 5,\ 8,\ 10)$$

and an Inverse Gamma(α, β) prior distribution with $\alpha = \beta = 3$, compute the Bayes estimator of θ.

(d) What is the posterior risk of your Bayes estimator?

9.20. An internet service provider studies the distribution of the number of concurrent users of the network. This number has Normal distribution with mean θ and standard deviation 4,000 people. The prior distribution of θ is Normal with mean 15,000 and standard deviation 2,000.

The data on the number of concurrent users are collected, see Exercise 8.2 on p. 250.

(a) Give the Bayes estimator for the mean number of concurrent users θ.

(b) Construct a 90% credible set for θ and interpret it.

(c) Is there a significant evidence that the mean number of concurrent users exceeds 16,000?

9.21. Another statistician conducts a non-Bayesian analysis of the data in Exercise 8.2 on p. 250 about concurrent users.

(a) Give the non-Bayesian estimator for the mean number of concurrent users θ.

(b) Construct a 90% confidence interval for θ and interpret it.

(c) Is there a significant evidence that the mean number of concurrent users exceeds 16,000?

(d) How do your results differ from the previous exercise?

9.22. In Example 9.11 on p. 266, we constructed a confidence interval for the population mean μ based on the observed Normally distributed measurements. Suppose that prior to the experiment we thought this mean should be between 5.0 and 6.0 with probability 0.95.

(a) Find a conjugate prior distribution that fully reflects your prior beliefs.

(b) Derive the posterior distribution and find the Bayes estimator of μ. Compute its Bayes risk.

(c) Compute a 95% HPD credible set for μ. Is it different from the 95% confidence interval? What causes the differences?

9.23. If ten coin tosses result in ten straight heads, can this coin be fair and unbiased?

By looking at a coin, you believe that it is fair (a 50-50 chance of each side) with probability 0.99. This is your prior probability. Then you toss the coin ten times, and each time it turns up heads. Compute the posterior probability that it is a fair coin.

9.24. Observed is a sample from Uniform$(0, \theta)$ distribution.

(a) Find a conjugate family of prior distributions.

(b) Assuming a prior distribution from this family, derive a form of the Bayes estimator, its posterior risk, and its Bayes risk.

CHAPTER 10

Regression

In Chapters 8 and 9, we were concerned about the distribution of *one random variable*, its parameters, expectation, variance, symmetry, skewness, etc. In this chapter, we study *relations* among variables.

Many variables observed in real life are related. The type of their relation can often be expressed in a mathematical form called *regression*. Establishing and testing such a relation enables us:

– to understand interactions, causes and effects among variables;

– to predict unobserved variables based on the observed ones;

– to determine which variables significantly affect the variable of interest.

10.1 Least squares estimation

Regression models relate a *response variable* to one or several predictors. Having observed predictors, we can forecast the response by computing its *conditional expectation*, given all the available predictors.

DEFINITION 10.1

Response or *dependent variable* Y is a variable of interest that we predict based on one or several predictors.

Predictors or *independent variables* $X^{(1)}, \ldots, X^{(k)}$ are used to predict the values and behavior of the response variable Y.

Regression of Y on $X^{(1)}, \ldots, X^{(k)}$ is the conditional expectation,

$$G(x^{(1)}, \ldots, x^{(k)}) = \mathbf{E}\left\{Y \mid X^{(1)} = x^{(1)}, \ldots, X^{(k)} = x^{(k)}\right\}.$$

It is a function of $x^{(1)}, \ldots, x^{(k)}$ whose form can be estimated from data.

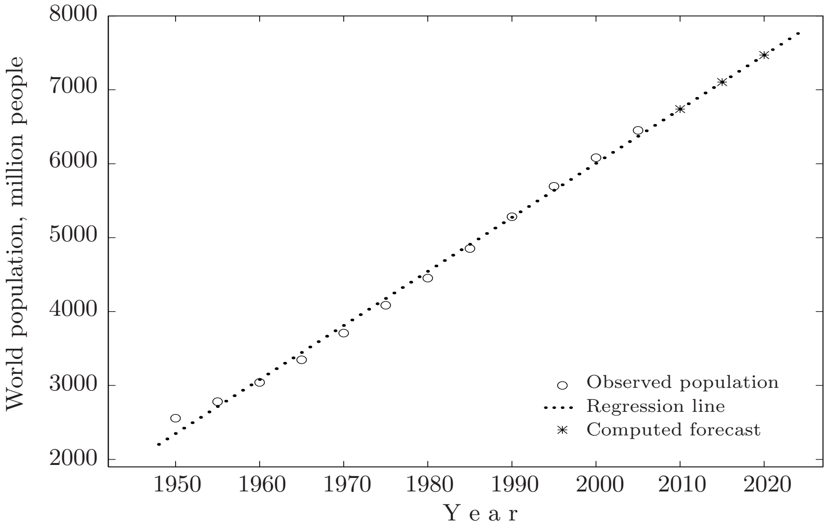


Figure 10.1 *World population in 1950–2005 and its regression forecast for 2010–2020.*

Year	Population mln. people	Year	Population mln. people	Year	Population mln. people
1950	2557	1975	4086	2000	6082
1955	2781	1980	4453	2005	6451
1960	3041	1985	4852	2010	?
1965	3347	1990	5283	2015	?
1970	3709	1995	5694	2020	?

Table 10.1 *Population of the world, 1950–2020.*

10.1.1 Examples

Consider several situations when we can predict a *dependent* variable of interest from *independent* predictors.

Example 10.1 (World population). According to the International Data Base of the *U.S. Census Bureau*, population of the world grows according to Table 10.1. How can we use these data to predict the world population in years 2010, 2015, and 2020?

Figure 10.1 shows that the population (response) is tightly related to the year (predictor),

$$\text{population} \approx G(\text{year}).$$

It increases every year, and its growth is almost linear. If we estimate the *regression function G* relating our response and our predictor (see the dotted

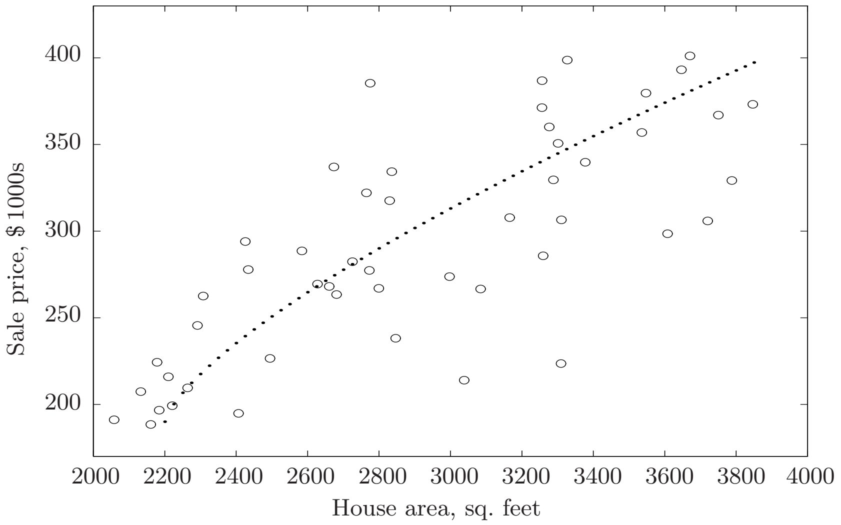


Figure 10.2 *House sale prices and their footage.*

line on Figure 10.1) and extend its graph to the year 2020, the forecast is ready. We can simply compute $G(2010)$, $G(2015)$, and $G(2020)$.

For example, a straight line that fits the observed data for years 1950–2005 predicts the population of 6.7 billion people in year 2010, 7.1 billion in 2015, and 7.5 billion in 2020. ◇

How accurate is the forecast obtained in this example? The observed population during 1950–2005 appears rather close to the estimated regression line in Figure 10.1. It is reasonable to hope that it will continue to do so through 2020.

The situation is different in the next example.

Example 10.2 (HOUSE PRICES). Seventy house sale prices in a certain county are depicted in Figure 10.2 along with the house area.

First, we see a clear relation between these two variables, and in general, bigger houses are more expensive. However, the trend no longer seems linear.

Second, there is a large amount of variability around this trend. Indeed, area is not the only factor determining the house price. Houses with the same area may still be priced differently.

Then, how can we estimate the price of a 3200-square-foot house? We can estimate the general trend (the dotted line in Figure 10.2) and plug 3200 into the resulting formula, but due to obviously high variability, our estimation will not be as accurate as in Example 10.1. ◇

To improve our estimation in the last example, we may take other factors into account: the number of bedrooms and bathrooms, the backyard area, the average income of the neighborhood, etc. If all the added variables are relevant for pricing a house, our model will have a closer fit and will provide more accurate predictions. Regression models with multiple predictors are studied in Section 10.3.

10.1.2 Method of least squares

Our immediate goal is to estimate the **regression function** G that connects response variable Y with predictors $X^{(1)}, \ldots, X^{(k)}$. First we focus on *univariate regression* predicting response Y based on one predictor X. The method will be extended to k predictors in Section 10.3.

In univariate regression, we observe *pairs* $(x_1, y_1), \ldots, (x_n, y_n)$, shown in Figure 10.3a.

For accurate forecasting, we are looking for the function $\widehat{G}(x)$ that passes as close as possible to the observed data points. This is achieved by minimizing distances between observed data points

$$y_1, \ldots, y_n$$

and the corresponding points on the fitted regression line,

$$\widehat{y}_1 = \widehat{G}(x_1), \ldots, \widehat{y}_n = \widehat{G}(x_n)$$

(see Figure 10.3b). Method of least squares minimizes the sum of squared distances.

DEFINITION 10.2

Residuals

$$e_i = y_i - \widehat{y}_i$$

are differences between observed responses y_i and their **fitted values** $\widehat{y}_i = \widehat{G}(x_i)$.

Method of least squares finds a regression function $\widehat{G}(x)$ that minimizes the sum of squared residuals

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 . \tag{10.1}$$

Function $\widehat{G}$ is usually sought in a suitable form, linear, quadratic, logarithmic, etc. The simplest form is linear.

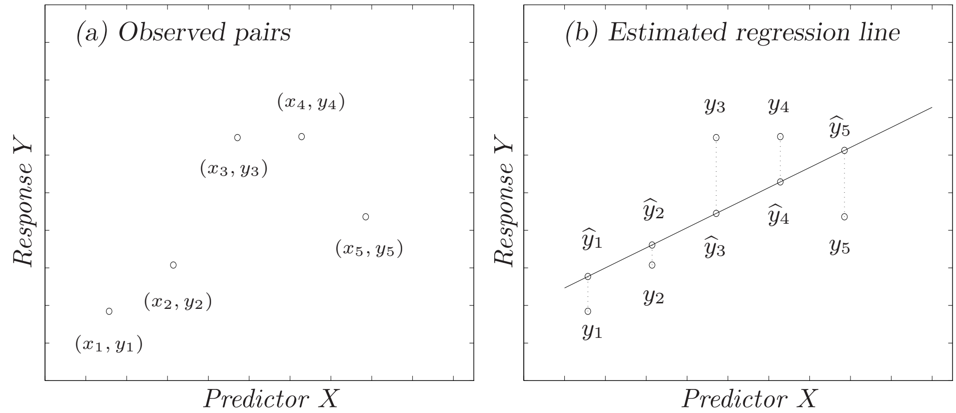


Figure 10.3 *Least squares estimation of the regression line.*

10.1.3 Linear regression

Linear regression model assumes that the conditional expectation

$$G(x) = \mathbf{E}\{Y \mid X = x\} = \beta_0 + \beta_1 x$$

is a *linear function* of x. As any linear function, it has an intercept β_0 and a slope β_1.

The **intercept**

$$\beta_0 = G(0)$$

equals the value of the regression function for $x = 0$. Sometimes it has no physical meaning. For example, nobody will try to predict the value of a computer with 0 random access memory, and nobody will consider the Federal reserve rate in year 0. In other cases, intercept is quite important. For example, according to the *Ohm's Law* ($V = R\,I$) the voltage across an *ideal* conductor is proportional to the current. A non-zero intercept ($V = V_0 + R\,I$) would show that the circuit is not ideal, and there is a external loss of voltage.

The **slope**

$$\beta_1 = G(x+1) - G(x)$$

is the predicted change in the response variable when predictor changes by 1. This is a very important parameter that shows how fast we can change the expected response by varying the predictor. For example, customer satisfaction will increase by $\beta_1(\Delta x)$ when the quality of produced computers increases by (Δx).

A zero slope means absence of a linear relationship between X and Y. In this case, Y is expected to stay constant when X changes.

Estimation in linear regression

Let us estimate the slope and intercept by **method of least squares**. Following (10.1), we minimize the sum of squared residuals

$$Q = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \widehat{G}(x_i) \right)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 .$$

We can do it by taking partial derivatives of Q, equating them to 0 and solving the resulting equations for β_0 and β_1.

The partial derivatives are

$$\begin{aligned} \frac{\partial Q}{\partial \beta_0} &= -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) ; \\ \frac{\partial Q}{\partial \beta_1} &= -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i . \end{aligned}$$

Equating them to 0, we obtain so-called *normal equations*,

$$\begin{cases} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) & = 0 \\ \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) & = 0 \end{cases}$$

From the first normal equation,

$$\beta_0 = \frac{\sum y_i - \beta_1 \sum x_i}{n} = \bar{y} - \beta_1 \bar{x}. \tag{10.2}$$

Substituting this into the second normal equation, we get

$$\sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = \sum_{i=1}^{n} x_i \left((y_i - \bar{y}) - \beta_1 (x_i - \bar{x}) \right) = S_{xy} - \beta_1 S_{xx} = 0, \tag{10.3}$$

where

$$S_{xx} = \sum_{i=1}^{n} x_i (x_i - \bar{x}) = \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{10.4}$$

and

$$S_{xy} = \sum_{i=1}^{n} x_i (y_i - \bar{y}) = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \tag{10.5}$$

are *sums of squares and cross-products*. Notice that it is all right to subtract $\bar{x}$ from x_i in the right-hand sides of (10.4) and (10.5) because $\sum (x_i - \bar{x}) = 0$ and $\sum (y_i - \bar{y}) = 0$.

Finally, we obtain the **least squares estimates** of intercept β_0 and slope β_1 from (10.2) and (10.3).

Regression estimates

$$\boxed{\begin{array}{rcl} b_0 &=& \widehat{\beta}_0 \;=\; \bar{y} - b_1\,\bar{x} \\ b_1 &=& \widehat{\beta}_1 \;=\; S_{xy}/S_{xx} \\ \text{where} && \\ S_{xx} &=& \displaystyle\sum_{i=1}^{n}(x_i-\bar{x})^2 \\ S_{xy} &=& \displaystyle\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \end{array}} \qquad (10.6)$$

Example 10.3 (World population). In Example 10.1, x_i is the year, and y_i is the world population during that year. To estimate the regression line in Figure 10.1, we compute

$$\bar{x} = 1977.5; \quad \bar{y} = 4361.3;$$

$$\begin{array}{rcl} S_{xx} &=& (1950-\bar{x})^2 + \ldots + (2005-\bar{x})^2 = 3757; \\ S_{xy} &=& (1950-\bar{x})(2557-\bar{y}) + \ldots + (2005-\bar{x})(6451-\bar{y}) = 261475. \end{array}$$

Then

$$\begin{array}{rcl} b_1 &=& S_{xy}/S_{xx} = 73.14 \\ b_0 &=& \bar{y} - b_1\bar{x} = -140273. \end{array}$$

The estimated regression line is

$$\widehat{G}(x) = b_0 + b_1\,x = \underline{\text{-140273 + 73.14x}}.$$

We conclude that the average world population grows at the rate of 73.14 million every year.

We can use the obtained equation to predict the future growth of the world population. Regression prediction for years 2010, 2015, and 2020 are

$$\begin{array}{rcl} \widehat{G}(2010) &=& b_0 + 2010\,b_1 = \underline{\text{6738 million people}} \\ \widehat{G}(2015) &=& b_0 + 2015\,b_1 = \underline{\text{7104 million people}} \\ \widehat{G}(2020) &=& b_0 + 2020\,b_1 = \underline{\text{7470 million people}} \end{array}$$

◇

10.1.4 Regression and correlation

Recall from Section 3.3.5 that **covariance**

$$\mathrm{Cov}(X,Y) = \mathbf{E}(X - \mathbf{E}(X))(Y - \mathbf{E}(Y))$$

and **correlation coefficient**

$$\rho = \frac{\mathrm{Cov}(X,Y)}{(\,\mathrm{Std}X)(\,\mathrm{Std}Y)}$$

measure the direction and strength of a linear relationship between variables X and Y. From observed data, we estimate $\mathrm{Cov}(X,Y)$ and ρ by the **sample covariance**

$$s_{xy} = \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

(it is unbiased for the population covariance) and the **sample correlation coefficient**

$$r = \frac{s_{xy}}{s_x s_y}, \tag{10.7}$$

where

$$s_x = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} \quad \text{and} \quad s_y = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n-1}}$$

are sample standard deviations of X and Y.

Comparing (10.3) and (10.7), we see that the estimated slope b_1 and the sample regression coefficient r are proportional to each other. Now we have two new formulas for the regression slope.

Estimated regression slope

$$\boxed{b_1 = \frac{S_{xy}}{S_{xx}} = \frac{s_{xy}}{s_x^2} = r\left(\frac{s_y}{s_x}\right)}$$

Like the correlation coefficient, regression slope is positive for positively correlated X and Y and negative for negatively correlated X and Y. The difference is the r is dimensionless whereas the slope is measured in units of Y per units of X. Thus, its value by itself does not indicate whether the dependence is weak or strong. It depends on the units, the scale of X and Y. We test significance of the regression slope in Section 10.2.

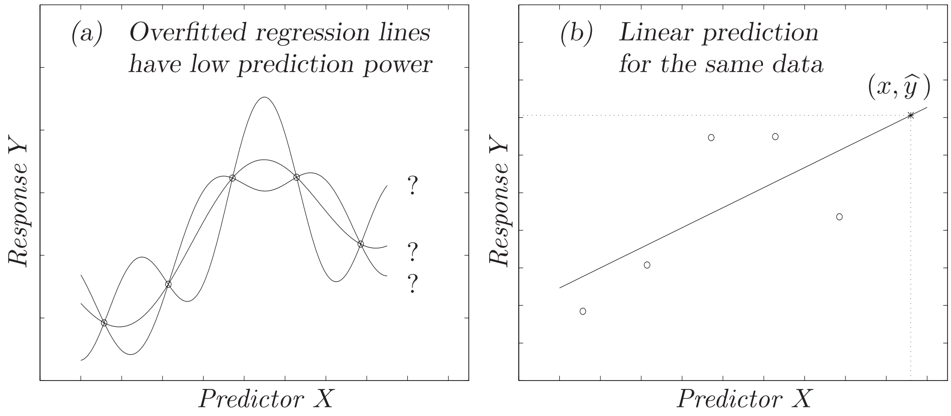


Figure 10.4 *Regression-based prediction.*

10.1.5 Overfitting a model

Among all possible straight lines, the method of least squares chooses one line that is closest to the observed data. Still, as we see in Figure 10.3b, we had some residuals $e_i = (y_i - \widehat{y}_i)$ and some positive sum of squared residuals. The straight line has not accounted for all 100% of variation among y_i.

Why, one might ask, have we considered only linear models? As long as all x_i are different, we can always find a regression function $\widehat{G}(x)$ that passes through all the observed points without any error. Then, the sum $\sum e_i^2 = 0$ will truly be minimized!

Trying to fit the data perfectly is a rather dangerous habit. Although we can achieve an excellent fit to the observed data, it never guarantees a good prediction. The model will be *overfitted*, too much taylored to the given data. Using it to predict unobserved responses is very questionable (see Figure 10.4a,b).

10.2 Analysis of variance, prediction, and further inference

In this section, we

- evaluate the *goodness of fit* of the chosen regression model to the observed data,

- estimate the response variance,
- test significance of regression parameters,
- construct confidence and prediction intervals.

10.2.1 ANOVA and R-square

Analysis of variance (**ANOVA**) explores variation among the observed responses. A portion of this variation can be explained by predictors. The rest is attributed to "error."

For example, there exists some variation among the house sale prices on Figure 10.2. Why are the houses priced differently? Well, the price depends on the house area, and bigger houses tend to be more expensive. So, to some extent, variation among prices is explained by variation among house areas. However, two houses with the same area may still have different prices. These differences cannot be explained by the area.

The total variation among observed responses is measured by the **total sum of squares**

$$SS_{\text{TOT}} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = (n-1)s_y^2.$$

This is the variation of y_i about their sample mean *regardless* of our regression model.

A portion of this total variation is attributed to predictor X and the regression model connecting predictor and response. This portion is measured by the **regression sum of squares**

$$SS_{\text{REG}} = \sum_{i=1}^{n} (\widehat{y}_i - \bar{y})^2.$$

This is the portion of total variation *explained by the model*. It is often computed as

$$\begin{aligned} SS_{\text{REG}} &= \sum_{i=1}^{n} (b_0 + b_1 x_i - \bar{y})^2 \\ &= \sum_{i=1}^{n} (\bar{y} - b_1\bar{x} + b_1 x_i - \bar{y})^2 \\ &= \sum_{i=1}^{n} b_1^2 (x_i - \bar{x})^2 \\ &= b_1^2 S_{xx} \ \text{ or } \ (n-1) b_1^2 s_x^2. \end{aligned}$$

The rest of total variation is attributed to "error." It is measured by the **error**

sum of squares

$$SS_{\mathrm{ERR}} = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} e_i^2.$$

This is the portion of total variation *not explained by the model*. It equals the sum of squared residuals that the method of least squares minimizes. Thus, applying this method, we minimize the *error sum of squares*.

Regression and error sums of squares partition SS_{TOT} into two parts (Exercise 10.6),

$$SS_{\mathrm{TOT}} = SS_{\mathrm{REG}} + SS_{\mathrm{ERR}}.$$

The *goodness of fit*, appropriateness of the predictor and the chosen regression model can be judged by the proportion of SS_{TOT} that the model can explain.

DEFINITION 10.3

R-square, or **coefficient of determination** is the proportion of the total variation explained by the model,

$$R^2 = \frac{SS_{\mathrm{REG}}}{SS_{\mathrm{TOT}}}.$$

It is always between 0 and 1, with high values generally suggesting a good fit.

In univariate regression, R-square also equals the squared sample correlation coefficient (Exercise 10.7),

$$R^2 = r^2.$$

Example 10.4 (World population, continued). Continuing Example 10.3, we find

$$\begin{aligned} SS_{\mathrm{TOT}} &= (n-1)s_y^2 = (11)(1.748 \cdot 10^6) = 1.923 \cdot 10^7, \\ SS_{\mathrm{REG}} &= b_1^2 S_{xx} = 1.912 \cdot 10^7, \\ SS_{\mathrm{ERR}} &= SS_{\mathrm{TOT}} - SS_{\mathrm{REG}} = 0.011 \cdot 10^5. \end{aligned}$$

A linear model for the growth of the world population has a very high R-square of

$$R^2 = \frac{SS_{\mathrm{REG}}}{SS_{\mathrm{TOT}}} = \underline{0.994} \text{ or } \underline{99.4\%}.$$

This is a very good fit although some portion of the remaining 0.6% of total variation can still be explained by adding non-linear terms into the model. ◇

10.2.2 Tests and confidence intervals

Methods of estimating a regression line and partitioning the total variation do not rely on any distribution, thus, we can apply them to virtually any data.

For further analysis, we introduce **standard regression assumptions**. We will assume that observed responses y_i are independent Normal random variables with mean

$$\mathbf{E}(Y_i) = \beta_0 + \beta_1 \, x_i$$

and constant variance σ^2. Predictors x_i are considered *non-random*.

As a consequence, regression estimates b_0 and b_1 have Normal distribution. After we estimate the variance σ^2, they can be studied by T-tests and T-intervals.

Degrees of freedom and variance estimation

According to the standard assumptions, responses $Y_1, \ldots, Y_n$ have different means but the same variance. This variance equals the mean squared deviation of responses from their respective expectations. Let us estimate it.

First, we estimate each expectation $\mathbf{E}(Y_i) = G(x_i)$ by

$$\widehat{G}(x_i) = b_0 + b_1 x_i = \widehat{y}_i.$$

Then, we consider deviations $e_i = y_i - \widehat{y}_i$, square them and add. We obtain the *error sum of squares*

$$SS_{\text{ERR}} = \sum_{i=1}^{n} e_i^2.$$

It remains to divide this sum by its number of degrees of freedom (this is how we estimated variances in Section **??**).

Let us compute degrees of freedom for all three SS in the regression ANOVA.

The total sum of squares $SS_{\text{TOT}} = (n-1)s_y^2$ has $\text{df}_{\text{TOT}} = n - 1$ degrees of freedom because it is computed directly from the sample variance s_y^2.

Out of them, the regression sum of squares SS_{REG} has $\text{df}_{\text{REG}} = 1$ degree of freedom. Recall (from p. 278) that the number of degrees of is the dimension of the corresponding space. Certainly, regression line has dimension 1.

This leaves $\text{df}_{\text{ERR}} = n - 2$ degrees of freedom for the error sum of squares, so that

$$\text{df}_{\text{TOT}} = \text{df}_{\text{REG}} + \text{df}_{\text{ERR}}.$$

The error degrees of freedom also follow from formula (9.10),

$$\text{df}_{\text{ERR}} = \text{ sample size } - \begin{array}{c}\text{number of estimated}\\ \text{location parameters}\end{array} = n - 2,$$

with 2 degrees of freedom deducted for 2 estimated parameters, β_0 and β_1. Equipped with this, we now estimate the variance.

Regression variance $\boxed{s^2 = \dfrac{SS_{\text{ERR}}}{n-2}}$

It estimates $\sigma^2 = \text{Var}(Y)$ unbiasedly.

Remark: Notice that the usual sample variance

$$s_y^2 = \frac{SS_{\text{TOT}}}{n-1} = \frac{\sum(y_i - \bar{y})^2}{n-1}$$

is biased because $\bar{y}$ no longer estimates the expectation of Y_i.

A standard way to present analysis of variance is the *ANOVA table*.

Univariate ANOVA

Source	Sum of squares	Degrees of freedom	Mean squares	F
Model	SS_{REG} $= \sum(\hat{y}_i - \bar{y})^2$	1	MS_{REG} $= SS_{\text{REG}}$	$\dfrac{MS_{\text{REG}}}{MS_{\text{ERR}}}$
Error	SS_{ERR} $= \sum(y_i - \hat{y}_i)^2$	$n-2$	MS_{ERR} $= \dfrac{SS_{\text{ERR}}}{n-2}$	
Total	SS_{TOT} $= \sum(y_i - \bar{y})^2$	$n-1$		

Mean squares MS_{REG} and MS_{ERR} are obtained from the corresponding sums of squares dividing them by their degrees of freedom. We see that the sample regression variance is the mean squared error,

$$s^2 = MS_{\text{ERR}}.$$

The estimated standard deviation s is usually called *root mean squared error* or *RMSE*.

The *F-ratio*

$$F = \frac{MS_{\text{REG}}}{MS_{\text{ERR}}}$$

is used to test significance of the entire regression model.

Inference about the regression slope

Having estimated the regression variance σ^2, we can proceed with tests and confidence intervals for the regression slope β_1. As usually, we start with the estimator of β_1 and its sampling distribution.

The slope is estimated by

$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{S_{xx}} = \frac{\sum(x_i - \bar{x})y_i}{S_{xx}}$$

(we can drop $\bar{y}$ because it is multiplied by $\sum(x_i - \bar{x}) = 0$).

According to *standard regression assumptions*, y_i are Normal random variables and x_i are non-random. Being a linear function of y_i, the estimated slope b_1 is also Normal with the expectation

$$\mathbf{E}(b_1) = \frac{\sum(x_i - \bar{x})\,\mathbf{E}(y_i)}{S_{xx}} = \frac{\sum(x_i - \bar{x})(\beta_0 + \beta_1 x_i)}{S_{xx}} = \frac{\sum(x_i - \bar{x})^2(\beta_1)}{\sum(x_i - \bar{x})^2} = \beta_1,$$

(which shows that b_1 is an *unbiased estimator* of β_1), and the variance

$$\operatorname{Var}(b_1) = \frac{\sum(x_i - \bar{x})^2 \operatorname{Var}(y_i)}{S_{xx}^2} = \frac{\sum(x_i - \bar{x})^2\sigma^2}{S_{xx}^2} = \frac{\sigma^2}{S_{xx}}.$$

Summarizing the results,

Sampling distribution of a regression slope

$$\boxed{\begin{array}{l} b_1 \text{ is Normal}(\mu_b, \sigma_b), \\ \text{where} \\ \mu_b = \mathbf{E}(b_1) = \beta_1 \\ \sigma_b = \operatorname{Std}(b_1) = \dfrac{\sigma}{\sqrt{S_{xx}}} \end{array}}$$

We estimate the standard deviation of b_1 by

$$s_b = \frac{s}{\sqrt{S_{xx}}},$$

and therefore, use T-intervals and T-tests.

Following the general principles, a $(1 - \alpha)100\%$ **confidence interval** for the slope is

$$\text{Estimator} \pm t_{\alpha/2} \left(\begin{array}{c} \text{estimated} \\ \text{st. deviation} \\ \text{of the estimator} \end{array} \right) = b_1 \pm t_{\alpha/2} \frac{s}{\sqrt{S_{xx}}}.$$

Testing hypotheses $H_0: \ \beta_1 = B$ about the regression slope, use the T-statistic

$$t = \frac{b_1 - B}{\widehat{\text{Std}b_1}} = \frac{b_1 - B}{s/\sqrt{S_{xx}}}.$$

P-values, acceptance and rejection regions are computed from Table A6 in the Appendix, T-distribution with $(n-2)$ degrees of freedom. These are degrees of freedom used in the estimation of σ^2.

As always, the form of the alternative determines whether it is a two-sided, right-tail, or left-tail test.

Significance of the model, relevance of predictor X in the inference about response Y, existence of a linear relation among them correspond to a non-zero slope. It means that a change in X causes changes in Y. In the absence of such relation, $\mathbf{E}(Y) = \beta_0$ remains constant.

To see if X is significant for the prediction of Y, test the null hypothesis

$$H_0: \ \beta_1 = 0 \ \text{ vs } \ H_a: \ \beta_1 \neq 0.$$

ANOVA F-test

A more universal, and therefore, more popular method of testing significance of a model is the **ANOVA F-test**. It compares the portion of variation explained by regression with the portion that remained unexplained. Significant models explain a relatively large portion.

Each portion of the total variation is measured by the corresponding *sum of squares*, SS_{REG} for the explained portion and SS_{ERR} for the unexplained portion (error). Dividing each SS by the number of degrees of freedom, we obtain **mean squares**,

$$MS_{\text{REG}} = \frac{SS_{\text{REG}}}{\text{df}_{\text{REG}}} = \frac{SS_{\text{REG}}}{1} = SS_{\text{REG}}$$

and

$$MS_{\text{ERR}} = \frac{SS_{\text{ERR}}}{\text{df}_{\text{ERR}}} = \frac{SS_{\text{ERR}}}{n-2} = s^2.$$

Under the null hypothesis

$$H_0: \ \beta_1 = 0,$$

both mean squares, MS_{REG} and MS_{ERR} are independent, and their ratio

$$F = \frac{MSR}{MSE} = \frac{SSR}{s^2}$$

has **F-distribution** with $\mathrm{df}_{\mathrm{REG}} = 1$ and $\mathrm{df}_{\mathrm{ERR}} = n - 2$ degrees of freedom.

This distribution has two parameters, numerator d.f. and denominator d.f., and it is very popular for testing ratios of variances and significance of models. Its critical values for significance levels $\alpha = 0.01$ and $\alpha = 0.05$ are tabulated in Table A5.

ANOVA F-test is always *one-sided* and *right-tail* because only large values of the F-statistic show a large portion of explained variation and the overall significance of the model.

F-test and T-test

We now have two tests for the model significance, a T-test for the regression slope and the ANOVA F-test. For the univariate regression, they are absolutely equivalent. In fact, the F-statistic equals the squared T-statistic for testing $H_0: \ \beta_1 = 0$ because

$$t^2 = \frac{b_1^2}{s^2/S_{xx}} = \frac{(S_{xy}/S_{xx})^2}{s^2/S_{xx}} = \frac{S_{xy}^2}{S_{xx}S_{yy}} \frac{S_{yy}}{s^2} = \frac{r^2 SS_{\mathrm{TOT}}}{s^2} = \frac{SS_{\mathrm{REG}}}{s^2} = F.$$

Hence, both tests give us the same result.

Example 10.5 (World population, continued). Can we apply the introduced methods to Examples 10.1–10.3? Estimating the correlation coefficient between y_i and y_{i-1}, we get a rather high value of 0.76. Hence, we cannot assume independence of y_i, and one of the standard assumptions is violated.

Our least squares regression line is still correct, however, in order to proceed with tests and confidence intervals, we need slightly more advanced methods accounting not only for the variance but also for covariances among the observed responses. ◇

Example 10.6 (Efficiency of computer programs). A computer manager needs to know how efficiency of her new computer program depends on the size of incoming data. Efficiency will be measured by the number of processed requests per hour. Applying the program to data sets of different sizes, she gets the following results,

Data size (gigabytes), x	6	7	7	8	10	10	15
Processed requests, y	40	55	50	41	17	26	16

In general, larger data sets require more computer time, and therefore, fewer requests are processed within 1 hour. The response variable here is the number of processed requests (y), and we attempt to predict it from the size of a data set (x).

(a) Estimation of the regression line. We can start by computing

$$n = 7,\ \bar{x} = 9,\ \bar{y} = 35,\ S_{xx} = 56,\ S_{xy} = -232,\ S_{yy} = 1452.$$

Estimate regression slope and intercept by

$$b_1 = \frac{S_{xy}}{S_{xx}} = -4.14 \quad \text{and} \quad b_0 = \bar{y} - b_1\bar{x} = 72.3.$$

Then, the estimated regression line has an equation

$$y = 72.3 - 4.14x.$$

Notice the negative slope. It means that *increasing* incoming data sets by 1 gigabyte we expect to process 4.14 *fewer* requests per hour.

(b) ANOVA table and variance estimation. Let us compute all components of the ANOVA. We have

$$SS_{\text{TOT}} = S_{yy} = 1452$$

partitioned into

$$SS_{\text{REG}} = b_1^2 S_{xx} = 961 \quad \text{and} \quad SS_{\text{ERR}} = SS_{\text{TOT}} - SS_{\text{REG}} = 491.$$

Simultaneously, $n-1 = 6$ degrees of freedom of SS_{TOT} are partitioned into $\text{df}_{\text{REG}} = 1$ and $\text{df}_{\text{ERR}} = 5$ degrees of freedom.

Fill the rest of the ANOVA table,

Source	Sum of squares	Degrees of freedom	Mean squares	F
Model	961	1	961	9.79
Error	491	5	98.2	
Total	1452	6		

Regression variance σ^2 is estimated by

$$s^2 = MS_{\text{ERR}} = 98.2.$$

R-square is

$$R^2 = \frac{SS_{\text{REG}}}{SS_{\text{TOT}}} = \frac{961}{1452} = 0.662.$$

That is, 66.2% of the total variation of the number of processed requests is explained by sizes of data sets only.

(c) Inference about the slope. Is the slope statistically significant? Does the number of process requests really depend on the size of data sets? To test the null hypothesis $H_0 : \ \beta_1 = 0$, compute the T-statistic

$$t = \frac{b_1}{\sqrt{s^2/S_{xx}}} = \frac{-4.14}{\sqrt{98.2/56}} = -3.12.$$

Checking the T-distribution table (Table A6) with 5 d.f., we find that the P-value for the *two-sided* test is between 0.01 and 0.02. We conclude that the slope is *moderately significant*. Precisely, it is significant at any level $\alpha \geq 0.02$ and not significant at any $\alpha \leq 0.01$.

(d) ANOVA F-TEST. A similar result is suggested by the F-test. From Table A5, the F-statistic of 9.79 is not significant at the 0.01 level, but significant at the 0.05 level.

$\diamond$

10.2.3 Prediction

One of the main applications of regression analysis is making forecasts, predictions of the response variable Y based on the known or controlled predictors X.

Let x_* be the value of the predictor X. The corresponding value of the response Y is computed by evaluating the estimated regression line at x_*,

$$\hat{y}_* = \widehat{G}(x_*) = b_0 + b_1 x_*.$$

This is how we predicted the population in years 2010–2020 in Example 10.3. As happens with any forecast, our predicted values are understood as the most intelligent guesses, and not as guaranteed exact sizes of the population during these years.

How reliable are regression predictions, and how close are they to the real true values? As a good answer, we can construct

– a $(1-\alpha)100\%$ **confidence interval** for the expectation

$$\mu_* = \mathbf{E}(Y \mid X = x_*)$$

and

– a $(1-\alpha)100\%$ **prediction interval** for the actual value of $Y = y_*$ when $X = x_*$.

Confidence interval for the mean of responses

The expectation

$$\mu_* = \mathbf{E}(Y \mid X = x_*) = G(x_*) = \beta_0 + \beta_1 x_*$$

is a population parameter. This is the mean response for the entire subpopulation of units where the independent variable X equals x_*. For example, it corresponds to the average price of all houses with the area $x_* = 2500$ square feet.

First, we estimate μ_* by

$$\begin{aligned} \hat{y}_* &= b_0 + b_1 x_* \\ &= \bar{y} - b_1\bar{x} + b_1 x_* \\ &= \bar{y} + b_1(x_* - \bar{x}) \\ &= \frac{1}{n}\sum y_i + \frac{\sum (x_i - \bar{x}) y_i}{S_{xx}}(x_* - \bar{x}) \\ &= \sum_{i=1}^{n}\left(\frac{1}{n} + \frac{\sum (x_i - \bar{x})}{S_{xx}}(x_* - \bar{x})\right) y_i. \end{aligned}$$

We see again that the estimator is a linear function of responses y_i. Then, under standard regression assumptions, $\hat{y}_*$ is Normal, with expectation

$$\mathbf{E}(\hat{y}_*) = \mathbf{E}b_0 + \mathbf{E}b_1 x_* = \beta_0 + \beta_1 x_* = \mu_*$$

(it is unbiased), and variance

$$\begin{aligned} \operatorname{Var}(\hat{y}_*) &= \sum\left(\frac{1}{n} + \frac{\sum (x_i - \bar{x})}{S_{xx}}(x_* - \bar{x})\right)^2 \operatorname{Var}(y_i) \\ &= \sigma^2\left(\sum_{i=1}^{n}\frac{1}{n^2} + 2\sum_{i=1}^{n}(x_i - \bar{x})\frac{x_* - \bar{x}}{S_{xx}} + \frac{S_{xx}(x_* - \bar{x})^2}{S_{xx}^2}\right) \\ &= \sigma^2\left(\frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{xx}}\right) \end{aligned}$$

(because $\sum (x_i - \bar{x}) = 0$).

Then, we estimate the regression variance σ^2 by s^2 and obtain the following confidence interval.

$(1-\alpha)$100% confidence interval for the mean $\mu_* = \mathbf{E}(Y \mid X = x_*)$ of all responses with $X = x_*$

$$\boxed{b_0 + b_1 x_* \pm t_{\alpha/2}\, s\sqrt{\frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{xx}}}}$$

Prediction interval for the individual response

Often we are more interested in predicting the actual response rather than the mean of all possible responses. For example, we may be interested in the price of one particular house that we are planning to buy, not in the average price of all similar houses.

Instead of estimating a *population parameter*, we are now predicting the *actual value* of a random variable.

DEFINITION 10.4

An interval $[a, b]$ is a $(1-\alpha)100\%$ **prediction interval** for the individual response Y corresponding to predictor $X = x_*$ if it contains the value of Y with probability $(1-\alpha)$,

$$\boldsymbol{P}\left\{a \leq Y \leq b \mid X = x_*\right\} = 1-\alpha.$$

This time, all three quantities, Y, a, and b, are random variables. Predicting Y by $\hat{y}_*$ and standardizing all three parts of the inequality

$$a \leq Y \leq b,$$

we get

$$\begin{aligned}\boldsymbol{P}\left\{\frac{a-\hat{y}_*}{\operatorname{Std}(Y-\hat{y}_*)} \leq \frac{Y-\hat{y}_*}{\operatorname{Std}(Y-\hat{y}_*)} \leq \frac{b-\hat{y}_*}{\operatorname{Std}(Y-\hat{y}_*)}\right\} &= 1-\alpha \\ = \boldsymbol{P}\left\{-t_{\alpha/2} \leq \frac{Y-\hat{y}_*}{\operatorname{Std}(Y-\hat{y}_*)} \leq t_{\alpha/2}\right\}&,\end{aligned}$$

where

$$\operatorname{Var}(Y-\hat{y}_*) = \operatorname{Var}(Y) + \operatorname{Var}(\hat{y}_*) = \sigma^2\left(1+\frac{1}{n}+\frac{(x_*-\bar{x})^2}{S_{xx}}\right).$$

We see that the response of interest, Y, made its contribution into the variance. This is the difference between a confidence interval for the mean of all responses and a prediction interval for the individual response. Predicting the individual value is a more difficult task.

A prediction interval is now computed by solving equations

$$\frac{a-\hat{y}_*}{\operatorname{Std}(Y-\hat{y}_*)} = -t_{\alpha/2} \quad \text{and} \quad \frac{b-\hat{y}_*}{\operatorname{Std}(Y-\hat{y}_*)} = t_{\alpha/2}$$

for a and b and estimating σ by s.

$(1-\alpha)$100% prediction interval for the individual response Y when $X = x_*$

$$\boxed{b_0 + b_1 x_* \pm t_{\alpha/2}\, s\sqrt{1+\frac{1}{n}+\frac{(x_*-\bar{x})^2}{S_{xx}}}} \tag{10.8}$$

Several conclusions are apparent from this formula.

First, it is harder to predict the individual value than to estimate its mean. More uncertainty is involved, and as a result, the margin of a prediction interval is larger than the margin of a confidence interval.

Second, we get more accurate estimates and more accurate predictions from large samples. When the sample size n (and therefore, typically, S_{xx}), tends to ∞, the margin of the confidence interval converges to 0.

On the other hand, the margin of a prediction interval converges to $(t_{\alpha/2}\sigma)$. As we collect more and more observations, our estimates of b_0 and b_1 become more accurate; however, uncertainty about the individual response Y will never vanish.

Third, we see that it regression estimation and prediction are most accurate when x_* is close to $\bar{x}$ so that

$$(x_* - \bar{x})^2 \approx 0.$$

The margin increases as the independent variable x_* drifts away from $\bar{x}$. We conclude that it is easiest to make forecasts under normal and "standard" conditions, and it is hardest to predict anomalies.

Example 10.7 (Predicting the program efficiency). Suppose we need to start processing requests that refer to $x^* = 16$ gigabytes of data. Based on our regression analysis of the program efficiency in Example 10.6, we predict

$$y^* = b_0 + b_1 x^* = 72.3 - 4.14(16) = 6$$

requests processed within 1 hour. A 95% prediction interval for the number of processed requests is

$$\begin{aligned} y_* \pm t_{0.025}\, s\sqrt{1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{xx}}} &= 6 \pm (2.571)\sqrt{98.2}\sqrt{1 + \frac{1}{7} + \frac{(16-9)^2}{56}} \\ &= 6 \pm 36.2 = [0; 42]. \end{aligned}$$

(using Table A6 with 5 d.f.). We rounded both ends of the prediction interval knowing that there cannot be a negative or fractional number of requests. ◇

Prediction bands

For all possible values of a predictor x^*, we can prepare a graph of $(1-\alpha)$ **prediction bands** given by (10.8). Then, for each value of x^*, one can draw a vertical line and obtain a $100(1-\alpha)\%$ prediction interval.

Figure 10.5 shows the 95% prediction bands for the number of processed requests in Example 10.7. These are two curves on each side of the fitted regression line. As we have already noticed, prediction is most accurate when x^* is near the sample mean $\bar{x}$. Prediction intervals get wider when we move away from $\bar{x}$.

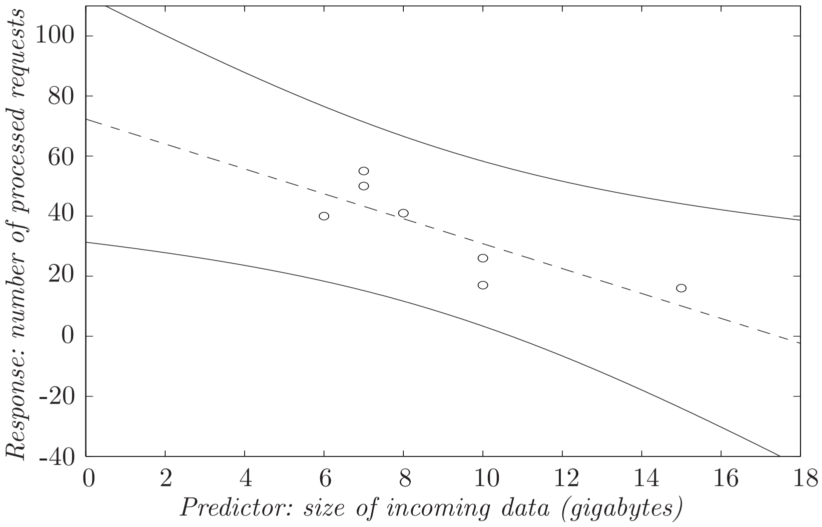


Figure 10.5 *Regression prediction of program efficiency.*

10.3 Multivariate regression

In the last two sections, we learned how to predict a response variable Y from a predictor variable X. We also hoped that including more information and using several predictors instead of one will enhance our prediction.

Now we introduce **multiple linear regression** that will connect a response Y with several predictors $X^{(1)}$, $X^{(2)}$, …, $X^{(k)}$.

10.3.1 Introduction and examples

Example 10.8 (Additional information). In Example 10.2, we discussed predicting price of a house based on its area. We decided that perhaps, this prediction is not very accurate due to a high variability among house prices.

What is the source of this variability? Why are houses of the same size priced differently?

Certainly, area is not the only parameter of a house. Prices are different due to different design, location, number of rooms and bathrooms, presence of a basement, a garage, a swimming pool, different size of a backyard, etc. When we take all this information into account, we'll have a rather accurate description of a house and hopefully, a rather accurate prediction of its price. ◇

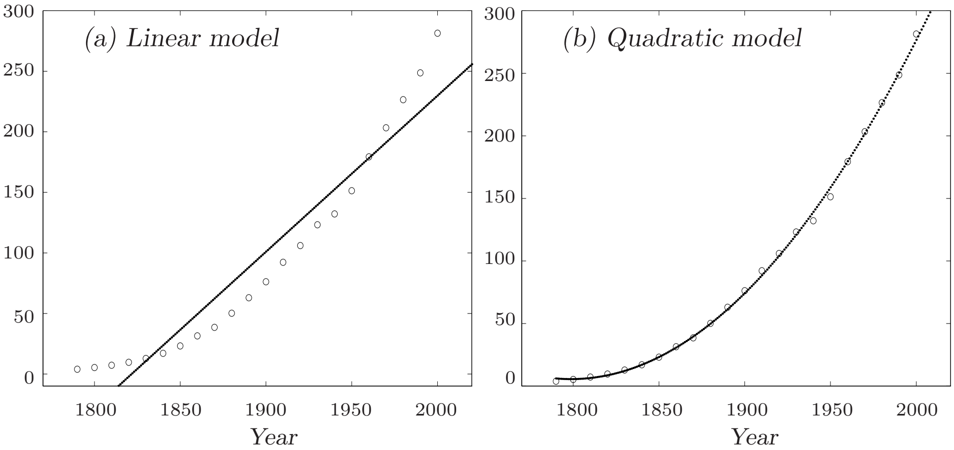


Figure 10.6 *U.S. population in 1790–2000 (million people).*

Example 10.9 (U.S. POPULATION AND NONLINEAR TERMS). One can often reduce variability and do more accurate analysis by adding nonlinear terms into the linear regression model. In Example 10.3, we predicted the world population for years 2010–2020 based on the *linear model*

$$\mathbf{E}(\text{population}) = \beta_0 + \beta_1(\text{year}).$$

We showed in Example 10.4 that this model has a pretty good fit.

However, a linear model does a poor prediction of the U.S. population between 1790 and 2000 (see Figure 10.6a). The population growth is clearly nonlinear.

On the other hand, a quadratic model in Figure 10.6b gives an amazingly excellent fit! It misses only a slight decrease in the rate of growth during the World War II.

For this model, we assume

$$\mathbf{E}(\text{population}) = \beta_0 + \beta_1(\text{year}) + \beta_2(\text{year})^2,$$

or even better,

$$\mathbf{E}(\text{population}) = \beta_0 + \beta_1(\text{year-1800}) + \beta_2(\text{year-1800})^2.$$

◇

A **multivariate regression model** assumes that the conditional expectation of a response

$$\mathbf{E}\left\{Y \mid X^{(1)} = x^{(1)}, \ldots, X^{(k)} = x^{(k)}\right\} = \beta_0 + \beta_1 x^{(1)} + \ldots + \beta_k x^{(k)} \quad (10.9)$$

is a linear function of predictors $x^{(1)}, \ldots, x^{(k)}$.

This regression model has one intercept and a total of k slopes, and therefore, it defines a k-dimensional *regression plane* in a $(k+1)$-dimensional space of $(X^{(1)}, \ldots, X^{(k)}, Y)$.

The **intercept** β_0 is the expected response when all predictors equal zero.

Each **regression slope** β_j is the expected change of the response Y when the corresponding predictor $X^{(j)}$ changes by 1 *while all the other predictors remain constant.*

In order to estimate all the parameters of model (10.9), we collect a sample of n *multivariate observations*

$$\left\{ \begin{array}{ccc} \boldsymbol{X}_1 & = & \left(X_1^{(1)}, X_1^{(2)}, \ldots, X_1^{(k)}\right) \\ \boldsymbol{X}_2 & = & \left(X_2^{(1)}, X_2^{(2)}, \ldots, X_2^{(k)}\right) \\ \vdots & \vdots & \vdots \\ \boldsymbol{X}_n & = & \left(X_n^{(1)}, X_n^{(2)}, \ldots, X_n^{(k)}\right) \end{array} \right. .$$

Essentially, we collect a sample of n units (say, houses) and measure all k predictors on each house (area, number of rooms, etc.). Also, we measure responses, $Y_1, \ldots, Y_n$. We then estimate $\beta_0, \beta_1, \ldots, \beta_k$ by the method of least squares.

10.3.2 Matrix approach and least squares estimation

According to the *method of least squares*, we find such slopes $\beta_1, \ldots, \beta_k$ and such an intercept β_0 that will minimize the sum of squared "errors"

$$Q = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i^{(1)} - \ldots - \beta_k x_i^{(k)}\right)^2 .$$

Minimizing Q, we can again take partial derivatives of Q with respect to all the unknown parameters and solve the resulting system of equations. It can be conveniently written in a *matrix form* (which requires basic knowledge of linear algebra; if needed, refer to Appendix, Section 11.4).

Matrix approach to multivariate linear regression

We start with the data. Observed are an $n \times 1$ response vector $\boldsymbol{Y}$ and an $n \times (k+1)$ predictor matrix $\boldsymbol{X}$,

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \quad \text{and} \quad \boldsymbol{X} = \begin{pmatrix} 1 & \boldsymbol{X}_1 \\ \vdots & \vdots \\ 1 & \boldsymbol{X}_n \end{pmatrix} = \begin{pmatrix} 1 & X_1^{(1)} & \cdots & X_1^{(k)} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_n^{(1)} & \cdots & X_n^{(k)} \end{pmatrix} .$$

It is convenient to augment the predictor matrix with a column of 1's because now the multivariate regression model (10.9) can be written as

$$\mathbf{E}\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_1^{(1)} & \cdots & X_1^{(k)} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_n^{(1)} & \cdots & X_n^{(k)} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix},$$

or simply

$$\mathbf{E}(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}.$$

Now the multidimensional parameter

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}.$$

includes the intercept and all the slopes. In fact, the intercept β_0 can also be treated as one of the slopes that corresponds to the added column of 1's.

Estimating $\boldsymbol{\beta}$, we get a vector of **sample regression slopes**

$$\boldsymbol{b} = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{pmatrix}.$$

Fitted values are then computed as

$$\widehat{\boldsymbol{y}} = \begin{pmatrix} \widehat{y}_1 \\ \vdots \\ \widehat{y}_n \end{pmatrix} = \boldsymbol{X}\boldsymbol{b},$$

and thus, the least squares problem reduces to minimizing

$$\begin{aligned} Q(\boldsymbol{b}) &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = (\boldsymbol{y} - \widehat{\boldsymbol{y}})^T(\boldsymbol{y} - \widehat{\boldsymbol{y}}) \\ &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}). \end{aligned} \tag{10.10}$$

with T denoting a transposed vector.

Least squares estimates

In the matrix form, the minimum of the sum of squares

$$Q(\boldsymbol{b}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) = \boldsymbol{b}^T(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{b} - 2\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{b} + \boldsymbol{y}^T\boldsymbol{y}$$

is attained by

Estimated slopes in multivariate regression $\boxed{\boldsymbol{b} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}}$

As we can see from this formula, all the estimated slopes are

– *linear* functions of observed responses $(y_1, \ldots, y_n)$,

– *unbiased* for the regression slopes because

$$\mathbf{E}(\boldsymbol{b}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\,\mathbf{E}(\boldsymbol{y}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta},$$

– *Normal* if the response variable Y is Normal.

This is a multivariate analogue of $b = S_{xy}/S_{xx}$ that we derived for the univariate case.

10.3.3 Analysis of variance, tests, and prediction

We can again partition the *total sum of squares* measuring the total variation of responses into the *regression sum of squares* and the *error sum of squares.*

The **total sum of squares** is still

$$SS_{\text{TOT}} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = (\boldsymbol{y} - \bar{\boldsymbol{y}})^T(\boldsymbol{y} - \bar{\boldsymbol{y}}),$$

where

$$\bar{\boldsymbol{y}} = \begin{pmatrix} \bar{y} \\ \vdots \\ \bar{y} \end{pmatrix} = \bar{y}\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

This is the **total sum of squares corrected for the mean**; it has $\text{df}_{\text{TOT}} = (n-1)$ degrees of freedom.

For regression models without an intercept, it is customary to consider

$$SS'_{\text{TOT}} = \sum_{i=1}^{n} y_i^2 = \boldsymbol{y}^T\boldsymbol{y},$$

which is the **uncorrected total sum of squares** instead of SS_{TOT}, with all $\text{df}'_{\text{TOT}} = n$ degrees of freedom.

Again,

$$SS_{\text{TOT}} = SS_{\text{REG}} + SS_{\text{ERR}},$$

where

$$SS_{\mathrm{REG}} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = (\widehat{\boldsymbol{y}} - \bar{\boldsymbol{y}})^T (\widehat{\boldsymbol{y}} - \bar{\boldsymbol{y}})$$

is the **regression sum of squares**, and

$$SS_{\mathrm{ERR}} = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = (\boldsymbol{y} - \widehat{\boldsymbol{y}})^T (\boldsymbol{y} - \widehat{\boldsymbol{y}}) = \boldsymbol{e}^T \boldsymbol{e}$$

is the **error sum of squares**, the quantity we minimized when we applied the method of least squares.

The multivariate regression model (10.9) defines a k-dimensional regression plane where the fitted values belong to. Therefore, the regression sum of squares has

$$\mathrm{df}_{\mathrm{REG}} = k$$

degrees of freedom, whereas by subtraction,

$$\mathrm{df}_{\mathrm{ERR}} = \mathrm{df}_{\mathrm{TOT}} - \mathrm{df}_{\mathrm{REG}} = n - k - 1$$

degrees of freedom are left for SS_{ERR}. This is again the sample size n minus k estimated slopes and 1 estimated intercept.

We can then write the ANOVA table,

Multivariate ANOVA

Source	Sum of squares	Degrees of freedom	Mean squares	F
Model	SS_{REG} $= (\widehat{\boldsymbol{y}} - \bar{\boldsymbol{y}})^T (\widehat{\boldsymbol{y}} - \bar{\boldsymbol{y}})$	k	MS_{REG} $= \dfrac{SS_{\mathrm{REG}}}{k}$	$\dfrac{MS_{\mathrm{REG}}}{MS_{\mathrm{ERR}}}$
Error	SS_{ERR} $= (\boldsymbol{y} - \widehat{\boldsymbol{y}})^T (\boldsymbol{y} - \widehat{\boldsymbol{y}})$	$n - k - 1$	MS_{ERR} $= \dfrac{SS_{\mathrm{ERR}}}{n-k-1}$	
Total	SS_{TOT} $= (\boldsymbol{y} - \bar{\boldsymbol{y}})^T (\boldsymbol{y} - \bar{\boldsymbol{y}})$	$n - 1$		

The **coefficient of determination**

$$R^2 = \frac{SS_{\mathrm{REG}}}{SS_{\mathrm{TOT}}}$$

again measures the proportion of the total variation explained by regression. It can be shown that R^2 can only increase when we increase k by adding new predictors to our model. Thus, we should expect to increase R^2 and generally, get a better fit by going from univariate to multivariate regression.

Testing significance of the entire model

Further inference requires **standard multivariate regression assumptions** of Y_i being independent Normal random variables with means

$$\mathbf{E}(Y_i) = \beta_0 + \beta_1 X_i^{(1)} + \ldots + \beta_k X_i^{(k)}$$

and constant variance σ^2 while all predictors $X_i^{(j)}$ are non-random.

ANOVA F-test in multivariate regression tests significance of the entire model. The model is significant as long as at least one slope is not zero. Thus, we are testing

$$H_0 : \ \beta_1 = \ldots = \beta_k = 0 \quad \text{vs} \quad H_A : \ \text{not } H_0; \ \text{at least one } \beta_j \neq 0.$$

We compute the F-statistic

$$F = \frac{MS_{\text{REG}}}{MS_{\text{ERR}}} = \frac{SS_{\text{REG}}/k}{SS_{\text{ERR}}/(n-k-1)}$$

and check it against the F-distribution with k and $(n-k-1)$ degrees of freedom in Table A5.

This is always a one-sided right-tail test. Only large values of F correspond to large SS_{REG} indicating that fitted values $\hat{y}_i$ are far from the overall mean $\bar{y}$, and therefore, the expected response really changes along the regression plane according to predictors.

Variance estimator

Regression variance $\sigma^2 = \text{Var}(Y)$ is then estimated by the mean squared error

$$s^2 = MS_{\text{ERR}} = \frac{SS_{\text{ERR}}}{n-k-1}.$$

It is an unbiased estimator of σ^2 that can be used in further inference.

Testing individual slopes

For the inference about **individual regression slopes** β_j, we compute all the variances $\text{Var}(\beta_j)$. Matrix

$$\text{VAR}(\boldsymbol{b}) = \begin{pmatrix} \text{Var}(b_1) & \text{Cov}(b_1, b_2) & \cdots & \text{Cov}(b_1, b_k) \\ \text{Cov}(b_2, b_1) & \text{Var}(b_2) & \cdots & \text{Cov}(b_2, b_k) \\ \vdots & \vdots & \vdots & \vdots \\ \text{Cov}(b_k, b_1) & \text{Cov}(b_k, b_2) & \cdots & \text{Var}(b_k) \end{pmatrix}$$

is called a **variance-covariance matrix** of a vector $\boldsymbol{b}$. It equals

$$\begin{aligned} \mathrm{VAR}(\boldsymbol{b}) &= \mathrm{VAR}\left((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}\right) \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\,\mathrm{VAR}(\boldsymbol{y})\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\ &= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\ &= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}. \end{aligned}$$

Diagonal elements of this $k \times k$ matrix are variances of individual regression slopes,

$$\sigma^2_{b_1} = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{11}, \ldots,\ \sigma^2_{b_k} = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{kk}.$$

We estimate them by sample variances,

$$s^2_{b_1} = s^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{11}, \ldots,\ s^2_{b_k} = s^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{kk}.$$

Now we are ready for the inference about individual slopes. Hypothesis

$$H_0:\ \beta_j = B$$

can be tested with a T-statistics

$$t = \frac{\beta_j - B}{s_{b_j}}.$$

Compare this T-statistic against the T-distribution with $\mathrm{df}_{\mathrm{ERR}} = n - k - 1$ degrees of freedom, Table A6. This test may be two-sided or one-sided, depending on the alternative.

A test of

$$H_0:\ \beta_j = 0 \quad \text{vs}\ \ H_A:\ \beta_j \neq 0$$

shows whether predictor $X^{(j)}$ is relevant for the prediction of Y. If the alternative is true, the expected response

$$\mathbf{E}(Y) = \beta_0 + \beta_1 X^{(1)} + \ldots + \beta_j X^{(j)} + \ldots + \beta_j X^{(k)}$$

changes depending on $X^{(j)}$ even if all the other predictors remain constant.

Prediction

For the given vector of predictors $\boldsymbol{X}_* = (X^{(1)}_* = x^{(1)}_*, \ldots, X^{(k)}_* = x^{(k)}_*)$, we estimate the expected response by

$$\hat{y}_* = \widehat{\mathbf{E}}\{Y \mid \boldsymbol{X}_* = \boldsymbol{x}_*\} = \boldsymbol{x}_*\boldsymbol{b},$$

and predict the individual response by the same statistic.

To produce confidence and prediction intervals, we compute the variance,

$$\mathrm{Var}(\hat{y}_*) = \mathrm{Var}(\boldsymbol{x}_*\boldsymbol{b}) = \boldsymbol{x}_*^T\,\mathrm{Var}(\boldsymbol{b})\boldsymbol{x}_* = \sigma^2\boldsymbol{x}_*^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_*,$$

where $\boldsymbol{X}$ is the matrix of predictors used to estimate the regression slope $\boldsymbol{\beta}$.

Estimating σ^2 by s^2, we obtain a $(1-\alpha)100\%$ **confidence interval** for $\mu_* = \mathbf{E}(Y)$.

$(1-\alpha)$**100% confidence**
interval for the mean
$\mu_* = \mathbf{E}(Y \mid \boldsymbol{X}_* = \boldsymbol{x}_*)$
of all responses with $X_* = x^*$

$$\boxed{\boldsymbol{x}_* \boldsymbol{b} \pm t_{\alpha/2}\, s \sqrt{\boldsymbol{x}_*^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_*}}$$

Accounting for the additional variation of the individual response y_*, we get a $(1-\alpha)100\%$ **prediction interval** for y_*.

$(1-\alpha)$**100% prediction**
interval for
the individual response Y
when $\boldsymbol{X}_* = \boldsymbol{x}_*$

$$\boxed{\boldsymbol{x}_* \boldsymbol{b} \pm t_{\alpha/2}\, s \sqrt{1 + \boldsymbol{x}_*^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_*}}$$

In both expressions, $t_{\alpha/2}$ refers to the T-distribution with $(n-k-1)$ degrees of freedom.

Example 10.10 (Database structure). The computer manager in Examples 10.6 and 10.7 tries to improve the model by adding another predictor. She decides that in addition to the size of data sets, efficiency of the program may depend on the database structure. In particular, it may be important to know how many tables were used to arrange each data set. Putting all this information together, we have

Data size (gigabytes), x_1	6	7	7	8	10	10	15
Number of tables, x_2	4	20	20	10	10	2	1
Processed requests, y	40	55	50	41	17	26	16

(a) Least squares estimation. The *predictor matrix* and the *response vec-*

tor are

$$\boldsymbol{X} = \begin{pmatrix} 1 & 6 & 4 \\ 1 & 7 & 20 \\ 1 & 7 & 20 \\ 1 & 8 & 10 \\ 1 & 10 & 10 \\ 1 & 10 & 2 \\ 1 & 15 & 1 \end{pmatrix}, \qquad \boldsymbol{Y} = \begin{pmatrix} 40 \\ 55 \\ 50 \\ 41 \\ 17 \\ 26 \\ 16 \end{pmatrix}.$$

We then compute

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{pmatrix} 7 & 63 & 67 \\ 63 & 623 & 519 \\ 67 & 519 & 1021 \end{pmatrix} \quad \text{and} \quad \boldsymbol{X}^T\boldsymbol{Y} = \begin{pmatrix} 245 \\ 1973 \\ 2908 \end{pmatrix},$$

to obtain the estimated *vector of slopes*

$$\boldsymbol{b} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{Y}) = \begin{pmatrix} 52.7 \\ -2.87 \\ 0.85 \end{pmatrix}.$$

Thus, the regression equation is

$$y = 52.7 - 2.87x_1 + 0.85x_2,$$

or

$$\begin{pmatrix} \text{number of} \\ \text{requests} \end{pmatrix} = 52.7 - 2.87 \begin{pmatrix} \text{size of} \\ \text{data} \end{pmatrix} + 0.85 \begin{pmatrix} \text{number of} \\ \text{tables} \end{pmatrix}.$$

(b) ANOVA AND F-TEST. The total sum of squares is still $SS_{\text{TOT}} = S_{yy} = 1452$. It is the same for all the models with this response.

Having figured a vector of *fitted values*

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{b} = \begin{pmatrix} 38.9 \\ 49.6 \\ 49.6 \\ 38.2 \\ 32.5 \\ 25.7 \\ 10.5 \end{pmatrix},$$

we can immediately compute

$$SS_{\text{REG}} = (\widehat{\boldsymbol{y}} - \bar{\boldsymbol{y}})^T(\widehat{\boldsymbol{y}} - \bar{\boldsymbol{y}}) = 1143.3 \quad \text{and} \quad SS_{\text{ERR}} = (\boldsymbol{y} - \widehat{\boldsymbol{y}})^T(\boldsymbol{y} - \widehat{\boldsymbol{y}}) = 308.7.$$

The ANOVA table is then completed as

Source	Sum of squares	Degrees of freedom	Mean squares	F
Model	1143.3	2	571.7	7.41
Error	308.7	4	77.2	
Total	1452	6		

Notice 2 degrees of freedom for the model because we now use two predictor variables.

R-SQUARE of $R^2 = SS_{\text{REG}}/SS_{\text{TOT}} = 0.787$ is 12.5% higher than in Example 10.6. These additional 12.5% of the total variation are explained by the new predictor x_2, in addition to x_1 that has already been used in the model. R-square can only increase when new variables are added.

ANOVA F-TEST statistic of 7.41 with 2 and 4 d.f. shows that the model is significant at the level of 0.05 but not 0.01.

REGRESSION VARIANCE σ^2 is estimated by $s^2 = 77.2$.

(c) INFERENCE ABOUT THE NEW SLOPE. Is the new predictor variable x_2 significant? It is, as long as the corresponding slope β_2 is proved to be non-zero. Let us test $H_0: \ \beta_2 = 0$.

The vector of slopes $\boldsymbol{b}$ has an estimated variance-covariance matrix

$$\widehat{\text{VAR}}(\boldsymbol{b}) = s^2(\boldsymbol{X'X})^{-1} = \begin{pmatrix} 284.7 & -22.9 & -7.02 \\ -22.9 & 2.06 & 0.46 \\ -7.02 & 0.46 & 0.30 \end{pmatrix}.$$

From this, $\text{Std}(b_2) = \sqrt{0.30} = 0.55$. The T-statistic is then

$$t = \frac{b_2}{\text{Std}(b_2)} = \frac{0.85}{0.55} = 1.54,$$

and for a two-sided test this is not significant at any level up to 0.10. This suggests that adding the data structure into the model does not bring a significant improvement. ◇

10.4 Model building

Multivariate regression opens an almost unlimited opportunity for us to improve prediction by adding more and more new predictors into our model. On the other hand, we saw in Section 10.1.5 that overfitting a model leads to a low prediction power. Moreover, it will often result in large variances $\sigma^2_{b_j}$ and therefore, unstable regression estimates.

Then, how can we build a model with the right, optimal set of predictors $X^{(j)}$ that will give us a good, accurate fit?

Two methods of variable selection are introduced here. One is based on the *adjusted R-square* criterion, the other is derived from the *extra sum of squares principle*.

10.4.1 Adjusted R-square

It is shown mathematically that R^2, the coefficient of determination, can only increase when we add predictors to the regression model. No matter how irrelevant it is for the response Y, any new predictor can only increase the proportion of explained variation.

Therefore, R^2 is not a fair criterion when we compare models with different numbers of predictors (k). Including irrelevant predictors should be penalized whereas R^2 can only reward for this.

A fair measure of goodness-of-fit is the *adjusted R-square*.

DEFINITION 10.5

Adjusted R-square

$$R^2_{\text{adj}} = 1 - \frac{SS_{\text{ERR}}/(n-k-1)}{SS_{\text{TOT}}/(n-1)} = 1 - \frac{SS_{\text{ERR}}/\text{df}_{\text{ERR}}}{SS_{\text{TOT}}/\text{df}_{\text{TOT}}}$$

is a criterion of variable selection. It rewards for adding a predictor only if it considerably reduces the error sum of squares.

Comparing with

$$R^2 = \frac{SS_{\text{REG}}}{SS_{\text{TOT}}} = \frac{SS_{\text{TOT}} - SS_{\text{ERR}}}{SS_{\text{TOT}}} = 1 - \frac{SS_{\text{ERR}}}{SS_{\text{TOT}}},$$

adjusted R-square includes degrees of freedom into this formula. As a result, R^2 always increases when a new variable is added whereas R^2_{adj} may decrease.

Imagine adding a non-significant predictor into the regression model. The number of estimated slopes k increments by 1. However, if this variable is not able to explain any variation of the response, the sums of squares, SS_{REG} and SS_{ERR} will remain the same. Then, $SS_{\text{ERR}}/(n-k-1)$ will increase and R^2_{adj} will decrease.

Adjusted R-square criterion: *choose a model with the highest adjusted R-square.*

10.4.2 Extra sum of squares, partial F-tests, and sequential variable selection

Suppose we have K predictors available for predicting a response. Technically, to select a subset that maximizes adjusted R-square, we need to fit all 2^K models and choose the one with the highest R^2_{adj}. This is possible for rather moderate K, this scheme is built in some statistical software.

Fitting all models is not feasible when the total number of predictors is large. Instead, we consider a sequential scheme that will follow a reasonable path and consider only a few models. At every step, it will compare some set of predictors

$$\boldsymbol{X}(Full) = \left(X^{(1)}, \ldots, X^{(k)}, X^{(k+1)}, \ldots, X^{(m)}\right)$$

and the corresponding *full model*

$$\mathbf{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}) = \beta_0 + \beta_1 x^{(1)} + \ldots + \beta_k x^{(k)} + \beta_{k+1} x^{(k+1)} + \ldots + \beta_m x^{(m)}$$

with a subset

$$\boldsymbol{X}(Reduced) = \left(X^{(1)}, \ldots, X^{(k)}\right)$$

and the corresponding *reduced model*

$$\mathbf{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}) = \beta_0 + \beta_1 x^{(1)} + \ldots + \beta_k x^{(k)}.$$

If the full model is significantly better, expanding the set of predictors is justified. If it is just as good as the reduced model, we should keep the smaller number of predictors.

DEFINITION 10.6

A model with a larger set of predictors is called a **full model**.

Including only a subset of predictors, we obtain a **reduced model**.

The difference in the variation explained by the two models is the **extra sum of squares**,

$$\begin{aligned} SS_{\text{EX}} &= SS_{\text{REG}}(Full) - SS_{\text{REG}}(Reduced) \\ &= SS_{\text{ERR}}(Reduced) - SS_{\text{ERR}}(Full). \end{aligned}$$

Extra sum of squares measures the *additional* amount of variation explained by additional predictors $X^{(k+1)}, \ldots, X^{(m)}$. By subtraction, it has

$$\text{df}_{\text{EX}} = \text{df}_{\text{REG}}(Full) - \text{df}_{\text{REG}}(Reduced) = m - k$$

degrees of freedom.

Significance of the additional explained variation (measured by SS_{EX}) is tested by a **partial F-statistic**

$$F = \frac{SS_{\text{EX}}/\text{df}_{\text{EX}}}{MS_{\text{ERR}}(Full)} = \frac{SS_{\text{ERR}}(Reduced) - SS_{\text{ERR}}(Full)}{SS_{\text{ERR}}(Full)} \left(\frac{n-m-1}{m-k}\right).$$

As a set, $X^{(k+1)}, \ldots, X^{(m)}$ affect the response Y if at least one of the slopes $\beta_{k+1}, \ldots, \beta_m$ is not zero in the full model. The partial F-test is a test of

$$H_0: \ \beta_{k+1} = \ldots = \beta_m = 0 \quad \text{vs} \quad H_A: \ \text{not } H_0.$$

If the null hypothesis is true, the partial F-statistic has the *F-distribution* with

$$\mathrm{df}_{\mathrm{EX}} = m - k \quad \text{and} \quad \mathrm{df}_{\mathrm{ERR}}(\mathit{Full}) = n - m - 1$$

degrees of freedom, Table A5.

The *partial F-test* is used for sequential selection of predictors in multivariate regression. We describe two such algorithms, *stepwise selection* and *backward elimination*.

Stepwise (forward) selection

The **stepwise selection** *algorithm* starts with the simplest model that excludes all the predictors,

$$G(\boldsymbol{x}) = \beta_0.$$

Then, predictors enter the model sequentially, one by one. Every new predictor should make the most significant contribution, among all the predictors that have not been included yet.

According to this rule, the first predictor $X^{(s)}$ to enter the model is the one that has the most significant univariate ANOVA F-statistic

$$F_1 = \frac{MS_{\mathrm{REG}}(X^{(s)})}{MS_{\mathrm{ERR}}(X^{(s)})}.$$

All F-tests considered at this step refer to the same F-distribution with 1 and $(n-2)$ d.f. Therefore, the largest F-statistic implies the lowest P-value and the most significant slope β_s

The model is now

$$G(\boldsymbol{x}) = \beta_0 + \beta_s x^{(s)}.$$

The next predictor $X^{(t)}$ to be selected is the one that makes the most significant contribution to $X^{(s)}$. Among all the remaining predictors, it should maximize the partial F-statistic

$$F_2 = \frac{SS_{\mathrm{ERR}}(\mathit{Reduced}) - SS_{\mathrm{ERR}}(\mathit{Full})}{MS_{\mathrm{ERR}}(\mathit{Full})}$$

designed to test significance of the slope β_t when the first predictor $X^{(s)}$ is already included. Such a partial F-statistic is also called **F-to-enter**.

All F-statistics at this step are compared against the same F-distribution with 1 and $(n-3)$ d.f., and again, the largest F-statistic points to the most significant slope β_t.

If the second predictor is included, the model becomes

$$G(\boldsymbol{x}) = \beta_0 + \beta_s x^{(s)} + \beta_t x^{(t)}.$$

The algorithm continues until the F-to-enter statistic is not significant for all the remaining predictors, according to a pre-selected significance level α. The final model will have all predictors significant at this level.

Backward elimination

The **backward elimination algorithm** works in the direction opposite to stepwise selection.

It starts with the full model that contains all possible predictors,

$$G(\boldsymbol{x}) = \beta_0 + \beta_1 x^{(1)} + \ldots + \beta_m x^{(m)}.$$

Predictors are *removed* from the model sequentially, one by one, until all the remaining predictors are statistically significant.

Significance is again determined by a partial F-test. In this scheme, it is called **F-to-remove**.

The first predictor to be removed is the one that *minimizes* the F-to-remove statistic

$$F_{-1} = \frac{SS_{\text{ERR}}(\textit{Reduced}) - SS_{\text{ERR}}(\textit{Full})}{MS_{\text{ERR}}(\textit{Full})}.$$

Again, the test with the lowest value of F_{-1} has the highest P-value indicating the least significance.

Suppose the slope β_u is found the least significant. Predictor $X^{(u)}$ is removed, and the model becomes

$$G(\boldsymbol{x}) = \beta_0 + \beta_1 x^{(1)} + \ldots + \beta_{u-1} x^{(u-1)} + \beta_{u+1} x^{(u+1)} + \ldots + \beta_m x^{(m)}.$$

Then we choose the next predictor to be removed by comparing all F_{-2} statistics, then go to F_{-3}, etc. The algorithm stops at the stage when all F-to-remove tests reject the corresponding null hypotheses. In the resulting model, all the remaining slopes are significant.

Both sequential model selection schemes, stepwise and backward elimination, involve fitting at most K models. This requires much less computing power than the adjusted R^2 method, where all 2^K models are considered.

Modern statistical computing packages (SAS, Splus, SPSS, JMP, and others) are equipped with all three considered model selection procedures.

Example 10.11 (Program efficiency: choice of a model). How should we predict the program efficiency in Examples 10.6–10.10 after all? Should we use the size of data sets x_1 alone, or the data structure x_2 alone, or both variables?

(a) Adjusted R-square criterion. For the *full model*,

$$R^2_{\text{adj}} = 1 - \frac{SS_{\text{ERR}}/\text{df}_{\text{ERR}}}{SS_{\text{TOT}}/\text{df}_{\text{TOT}}} = 1 - \frac{308.7/4}{1452/6} = 0.681.$$

Reduced model with only one predictor x_1 (Example 10.6) has

$$R^2_{\text{adj}} = 1 - \frac{491/5}{1452/6} = 0.594,$$

and another *reduced model* with only x_2 has $R^2_{\text{adj}} = 0.490$ (Exercise 10.9).

How do we interpret these R^2_{adj}? The price paid for including both predictors x_1 and x_2 is the division by 4 d.f. instead of 5 when we computed R^2_{adj} for the full model. Nevertheless, the full model explains such a large portion of the total variation that fully compensates for this penalty and makes the full model preferred to reduced ones. According to the *adjusted R-square criterion, the full model is best.*

(b) Partial F-test. How significant was addition of a new variable x_2 into our model? Comparing the *full model* in Example 10.10 with the *reduced model* in Example 10.6, we find the *extra sum of squares*

$$SS_{\text{EX}} = SS_{\text{REG}}(Full) - SS_{\text{REG}}(\textit{Reduced}) = 1143 - 961 = 182.$$

This is the additional amount of the total variation of response explained by x_2 when x_1 is already in the model. It has 1 d.f. because we added only 1 variable. The *partial F-test statistic* is

$$F = \frac{SS_{\text{EX}}/\text{df}_{\text{EX}}}{MS_{\text{ERR}}(\textit{Reduced})} = \frac{182/1}{98.2} = 1.85.$$

From Table A5 with 1 and 5 d.f., we see that this F-statistic is *not significant* at the 0.05 level. A relatively small additional amount of 182 of the total variation that the second predictor can explain does not justify its inclusion into the model.

(c) Sequential model selection. What models should be selected by stepwise and backward elimination routines?

Stepwise model selection starts by including the first predictor x_1. It is significant at the 5% level, as we know from Example 10.6, hence we keep it in the model. Next, we include x_2. As we have just seen, it fails to result in a significant gain, $F_2 = 1.85$, and thus, we do not keep it in the model. The resulting model predicts the program efficiency y based on the size of data sets x_1 only.

Backward elimination scheme starts with the full model and looks for ways to reduce it. Among the two reduced models, the model with x_1 has a higher regression sum of squares SS_{REG}, hence the other variable x_2 is the first one to be removed. The remaining variable x_1 is significant at the 5% level, therefore, we again arrive to the reduced model predicting y based on x_1.

Two different model selection criteria, adjusted R-square and partial F-tests, lead us to two different models. Each of them is best in a different sense. Not a surprise. ◇

10.4.3 Categorical predictors and dummy variables

Careful model selection is one of the most important steps in practical statistics. In regression, only a wisely chosen subset of predictors delivers accurate estimates and good prediction.

On the other hand, any useful information should be incorporated into our model. We conclude this chapter with a note on using *categorical* (that is, non-numerical) predictors in regression modeling.

Often a good portion of the variation of response Y can be explained by *attributes* rather than numbers. Examples are

– computer manufacturer (Dell, IBM, Hewlett Packard, etc.);

– operating system (Unix, Linux, Windows, etc.);

– major (Statistics, Computer Science, Electrical Engineering, etc.);

– gender (female, male);

– color (white, blue, green, etc.).

Unlike numerical predictors, attributes have no particular order. For example, it is totally *wrong* to code operating systems with numbers (1 = Unix, 2 = Linux, 3 = Windows), create a new predictor $X^{(k+1)}$ and include it into the regression model

$$G(\boldsymbol{x}) = \beta_0 + \beta_1 x^{(1)} + \ldots + \beta_k x^{(k)} + \beta_{k+1} x^{(k+1)}.$$

If we do so, it puts Linux right in the middle between Unix and Windows and tells that changing an operating system from Unix to Linux has exactly the same effect on the response Y as changing it from Linux to Windows!

However, performance of a computer depends on the operating system, manufacturer, type of the processor, and other categorical variables. How can we use them in our regression model?

We need to create so-called **dummy variables**. A dummy variable is binary, taking values 0 or 1,

$$Z_i^{(j)} = \begin{cases} 1 & \text{if unit } i \text{ in the sample has category } j \\ 0 & \text{otherwise} \end{cases}$$

For a categorical variable with C categories, we create $C-1$ dummy predictors, $\boldsymbol{Z}^{(1)}, \ldots, \boldsymbol{Z}^{(C-1)}$. They carry the entire information about the attribute. Sampled items from category C will be marked by all $C-1$ dummies equal to 0.

Example 10.12 (DUMMY VARIABLES FOR THE OPERATING SYSTEM). In addition to numerical variables, we would like to include the operating system into the regression model. Suppose that each sampled computer has one of three operating systems: Unix, Linux, or Windows.

$$Z_i^{(1)} = \begin{cases} 1 & \text{if computer } i \text{ has Unix} \\ 0 & \text{otherwise} \end{cases}$$

$$Z_i^{(2)} = \begin{cases} 1 & \text{if computer } i \text{ has Linux} \\ 0 & \text{otherwise} \end{cases}$$

Together with numerical predictors $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(k)}$, the regression model will be

$$G(\boldsymbol{x}, \boldsymbol{z}) = \beta_0 + \gamma_1 z^{(1)} + \gamma_2^{(2)} z^{(2)} + \beta_1 x^{(1)} + \ldots + \beta_k x^{(k)}.$$

◇

Fitting the model, all dummy variables are included into the *predictor matrix* $\boldsymbol{X}$ as columns.

Avoid singularity by creating only $(C-1)$ *dummies*

Notice that creating C dummies for an attribute with C categories causes a linear relation

$$\boldsymbol{Z}^{(1)} + \ldots + \boldsymbol{Z}^{(C)} = 1.$$

A column of 1's is already included into the predictor matrix $\boldsymbol{X}$, and therefore, such a linear relation will cause singularity of $(\boldsymbol{X}^T\boldsymbol{X})$ when we compute the least squares estimates $\boldsymbol{b} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$. Thus, it is necessary and sufficient to have only $(C-1)$ dummy variables.

Interpretation of slopes for dummy variables

Each slope γ_j for a dummy predictor $Z^{(j)}$ is the expected change in the response caused by incrementing $Z^{(j)}$ by 1 while keeping all other predictors constant. Such an increment occurs when we compare the last category C with category j.

Thus, the slope γ_j is the difference in the expected response comparing category C with category j. The difference of two slopes $(\gamma_j - \gamma_m)$ compares category j with category m.

Testing significance of a categorical variable is reasonable when we test all the slopes γ_j simultaneously. This is done by a partial F-test.

Summary and conclusions

This chapter provides methods of estimating mathematical relations between one or several predictor variables and a response variable. Results are used to explain behavior of the response and to predict its value for any new set of predictors.

Method of least squares is used to estimate regression parameters. Coefficient of determination R^2 shows the portion of the total variation that the included predictors can explain. The unexplained portion is considered as "error."

Analysis of variance (ANOVA) partitions the total variation into explained and unexplained parts and estimates regression variance by the mean squared error. This allows further statistical inference, testing slopes, constructing confidence intervals for mean responses and prediction intervals for individual responses. ANOVA F-test is used to test significance of the entire model.

For accurate estimation and efficient prediction, it is important to select the right subset of predictors. Sequential model selection algorithms are based on partial F-tests comparing full and reduced models at each step.

Categorical predictors are included into regression modeling by creating dummy variables.

Questions and exercises

10.1. The time it takes to transmit a file always depends on the file size. Suppose you transmitted 30 files, with the average size of 126 Kbytes and the standard deviation of 35 Kbytes. The average transmittance time was 0.04 seconds with the standard deviation of 0.01 seconds. The correlation coefficient between the time and the size was 0.86.

Based on this data, fit a linear regression model and predict the time it will take to transmit a 400 Kbyte file.

10.2. The following data were obtained from a sample of size $n = 75$:
– the predictor variable X has mean 32.2, variance 6.4;
– the response variable Y has mean 8.4, variance 2.8; and
– the sample covariance between X and Y is 3.6.

(a) Estimate the linear regression equation predicting Y based on X.

(b) Complete the ANOVA table. What portion of the total variation of Y is explained by X only?

(c) Assuming Normal distribution of the response, construct a 99% confidence interval for the regression slope. Is the slope significant?

10.3. At a gas station, 180 drivers were asked to record the mileage of their cars and the number of miles per gallon. The results are summarized in the table.

	Sample mean	Standard deviation
Mileage	24,598	14,634
Miles per gallon	23.8	3.4

The sample correlation coefficient is $r = -0.17$.

(a) Compute the least squares regression line which describes how the number of miles per gallon depends on the mileage.

(b) Use R^2 to evaluate its goodness of fit. Is this a good model?

(c) You purchase a used car with 35,000 miles on it. Predict the number of miles per gallon. Give a 95% prediction interval and a 95% confidence interval for the average number of miles per gallon for all cars with such a mileage.

10.4. The data below represent investments in development of new software by some computer company over an 11-year period,

Year, X	1995	1996	1997	1998	1999	2000
Investment, Y (in \$1000s)	17	23	31	29	33	39

Year, X	2001	2002	2003	2004	2005
Investment, Y (in \$1000s)	39	40	41	44	47

(a) In the regression model with Y as a dependent variable, estimate the variance of Y.

(b) Test whether the investment increases by *more* than \$ 1,800 every year, on the average.

(c) Give a 95% prediction interval for the investment in new-product development in the year 2009.

(d) Interpret this interval (explain the meaning of 95%) and state all assumptions used in this procedure.

10.5. In the previous problem, a market analyst notices that the investment amount may depend on whether the company shows profit in its financial reports. A categorical variable Z contains the additional information.

Year, X	1995	1996	1997	1998	1999	2000
Reporting profit, Z	no	no	yes	no	yes	yes

Year, X	2001	2002	2003	2004	2005
Reporting profit, Z	yes	yes	no	yes	yes

(a) Add a dummy variable to the model considered in Exercise 10.4 and estimate all parameters of the new multivariate regression model.

(b) If the company reports a profit during year 2007, predict the investment amount.

(c) How would your prediction change if the company reports a loss during year 2007?

(d) Complete a multivariate ANOVA table and test significance of the entire model.

(e) Does the new variable Z explain a significant portion of the total variation, in addition to the time trend?

10.6. For univariate linear regression, show that

$$SS_{\text{TOT}} = SS_{\text{REG}} + SS_{\text{ERR}}.$$

Hint: Write $SS_{\text{TOT}} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}((y_i - \widehat{y}_i) + (\widehat{y}_i - \bar{y}))^2$.

10.7. For univariate linear regression, show that R-square is the squared sample correlation coefficient,

$$R^2 = r^2.$$

Hint: Write the regression sum of squares as

$$SS_{\text{REG}} = \sum_{i=1}^{n}(b_0 + b_1\,x_i - \bar{y})^2$$

and substitute our derived formula for the regression slope b_1.

10.8. John wants to know if there is a relation between the number of hours he spends preparing for his weekly quiz and the grade he receives on it. He keeps records for 10 weeks.

It turns out that on the average, he spends 3.6 hours a week preparing for the quiz, with the standard deviation of 0.5 hours. His average grade is 82 (out of 100), with the standard deviation of 14. The correlation between the two variables is $r = 0.62$.

(a) Find the equation of the regression line predicting the quiz grade based on the time spent on preparation.

(b) This week, John studied for 4 hours. Predict his grade.

(c) Does this linear model explain most of the variation? Is it a good fit? Why?

10.9. For the data in Example 10.10, fit a linear regression model predicting the program efficiency (number of processed requests) based on the database structure x_2 only. Complete the ANOVA table, compute R-square and adjusted R-square. Is this reduced model significant?

10.10. Refer to Exercise 8.5 on p. 250.

(a) Fit a linear regression model estimating the time trend of the U.S. population. For simplicity, subtract 1800 from each year and let $x = \text{year} - 1800$ serve as a predictor.

(b) Complete the ANOVA table and compute R-square.

(c) According to the linear model, what population would you predict for years 2010, 2015, and 2020?

10.11. Here we improve the regression modeling of the previous exercise.

(a) Add the quadratic component and fit a multivariate regression model

$$\mathbf{E}(\text{population}) = \beta_0 + \beta_1(\text{year-1800}) + \beta_2(\text{year-1800})^2.$$

Estimate all the regression parameters.

(b) What population does this model predict for years 2010, 2015, and 2020?

(c) Complete the ANOVA table and compute R-square. Comparing with the previous exercise, how much of the total variation does the quadratic term explain?

(d) Now we have two competing models for the U.S. population. A linear (reduced) model is studied in Exercise 10.10, and a quadratic (full) model is considered here. Also, consider a model with a quadratic term only (without a linear term). Which model is preferred, according to the adjusted R-square criterion?

(e) Do your findings agree with Figure 10.6 on p. 349? Comment.

10.12. Refer to Exercise 8.6 on p. 251. Does a linear regression model provide an adequate fit? Estimate regression parameters, make a plot of 10-year increments in the U.S. population, along with your estimated regression line. What can you infer about the U.S. population growth?

10.13. Refer to Exercise 8.7 on p. 251.

(a) Fit a linear regression model to the 10-year relative change of the U.S. population. Estimate the regression parameters.

(b) Complete the ANOVA table and compute R-square.

(c) Under the standard regression assumptions, conduct ANOVA F-test and comment on the significance of the fitted model.

(d) Compute a 95% confidence interval for the regression slope.

(e) Compute a 95% prediction interval for the relative change of the population between years 2000 and 2010, then between years 2010 and 2020.

(f) Construct a histogram of regression residuals. Does it support our assumption of the normal distribution?

10.14. Consider the program efficiency study in Examples 10.6–10.11. The computer manager makes another attempt to improve the prediction power. This time she would like to consider the fact that the first four times the program worked under the operational system A and then switched to the operational system B.

Data size (gigabytes), x_1	6	7	7	8	10	10	15
Number of tables, x_2	4	20	20	10	10	2	1
Operational system, x_3	A	A	A	A	B	B	B
Processed requests, y	40	55	50	41	17	26	16

(a) Introduce a dummy variable responsible for the operational system and include it into the regression analysis.

(b) Does the new variable improve the goodness of fit? What is the new R-square?

(c) Is the new variable significant?

(d) What is the final regression equation that you would recommend to the computer manager any time when she needs to predict the number of processed requests given the size of data sets, the number of tables, and the operational system? Select the best regression equation using different model selection criteria.

CHAPTER 11

Appendix

11.1 Inventory of distributions

11.1.1 Discrete families

Bernoulli(p)

Generic description:	0 or 1, success or failure, result of one Bernoulli trial
Range of values:	$x = 0, 1$
Parameter:	$p \in (0, 1)$, probability of success
Probability mass function:	$P(x) = p^x (1-p)^{1-x}$
Expectation:	$\mu = p$
Variance:	$\sigma^2 = p(1-p)$
Relations:	Special case of *Binomial*(n, p) when $n = 1$ A sum of n independent Bernoulli(p) variables is *Binomial*(n, p)

Binomial(n, p)

Generic description:	Number of successes in n independent Bernoulli trials
Range of values:	$x = 0, 1, 2, \ldots, n$
Parameters:	$n = 1, 2, 3, \ldots$, number of Bernoulli trials $p \in (0, 1)$, probability of success
Probability mass function:	$P(x) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x}$
Cdf:	Table A2
Expectation:	$\mu = np$
Variance:	$\sigma^2 = np(1-p)$
Relations:	Binomial(n, p) is a sum of n independent *Bernoulli*(p) variables

$\text{Binomial}(1, p) = \textit{Bernoulli}(p)$

Table: Appendix, Table A2

Geometric(p)

Generic description: Number of Bernoulli trials until the first success

Range of values: $x = 1, 2, 3, \ldots$

Parameter: $p \in (0, 1)$, probability of success

Probability mass function: $P(x) = p(1-p)^{x-1}$

Cdf: $1 - (1-p)^x$

Expectation: $\mu = \dfrac{1}{p}$

Variance: $\sigma^2 = \dfrac{1-p}{p^2}$

Relations: Special case of *Negative Binomial*(k, p) when $k = 1$

A sum of n independent Geometric(p) variables is *Negative Binomial*(n, p)

Negative Binomial(k, p)

Generic description: Number of Bernoulli trials until the k-th success

Range of values: $x = k, k+1, k+2, \ldots$

Parameters: $k = 1, 2, 3, \ldots$, number of successes

$p \in (0, 1)$, probability of success

Probability mass function: $P(x) = \begin{pmatrix} x-1 \\ k-1 \end{pmatrix} (1-p)^{x-k} p^k$

Expectation: $\mu = \dfrac{k}{p}$

Variance: $\sigma^2 = \dfrac{k(1-p)}{p^2}$

Relations: Negative Binomial(k, p) is a sum of n independent *Geometric*(p) variables

Negative Binomial$(1, p)$ = *Geometric*(p)

Poisson(λ)

Generic description: Number of “rare events” during a fixed time interval

Range of values: $x = 0, 1, 2, \ldots$

Parameter: $\lambda \in (0, \infty)$, frequency of “rare events”

Probability mass function: $P(x) = e^{-\lambda}\dfrac{\lambda^x}{x!}$
Cdf: Table A3
Expectation: $\mu = \lambda$
Variance: $\sigma^2 = \lambda$
Relations: Limiting case of *Binomial*(n, p) when $n \to \infty$, $p \to 0$, $np \to \lambda$
Table: Appendix, Table A3

11.1.2 Continuous families

Beta(α, β)

Generic description: In a sample from Standard Uniform distribution, it is the distribution of the k^{th} smallest observation
Range of values: $0 < x < 1$
Parameter: $\alpha, \beta \in (0, \infty)$, frequency of events, inverse scale parameter
Density: $f(x) = \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, x^{\alpha-1}(1-x)^{\beta-1}$
Expectation: $\mu = \alpha/\alpha + \beta$
Variance: $\sigma^2 = \alpha\beta/(\alpha+\beta)^2(\alpha+\beta+1)$
Relations: As a prior distribution, Beta family is conjugate to the *Binomial* model.
Beta(1,1) distribution is *Uniform(0,1)*.

Exponential(λ)

Generic description: In a Poisson process, time between consecutive events
Range of values: $x > 0$
Parameter: $\lambda \in (0, \infty)$, frequency of events, inverse scale parameter
Density: $f(x) = \lambda e^{-\lambda x}$
Cdf: $F(x) = 1 - e^{-\lambda x}$
Expectation: $\mu = \dfrac{1}{\lambda}$
Variance: $\sigma^2 = 1/\lambda^2$
Relations: Special case of *Gamma*(α, λ) when $\alpha = 1$
A sum of α independent Exponential(λ) variables is *Gamma*(α, λ)

F(ν_1, ν_2)

Generic description: Ratio of independent mean squares
Range of values: $0 < x < +\infty$
Parameters: $\nu_1 > 0$, numerator degrees of freedom
$\nu_2 > 0$, denominator degrees of freedom

Density function: $f(x) = \dfrac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{x\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)}\sqrt{\dfrac{(\nu_1 x)^{\nu_1}\nu_2^{\nu_2}}{(\nu_1 x+\nu_2)^{\nu_1+\nu_2}}}$

Expectation: $\dfrac{\nu_2}{\nu_2 - 2}$ for $\nu_2 > 2$

Variance: $\dfrac{2\nu^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)}$ for $\nu_2 > 4$

Relations: A square of $t(\nu)$ variable is $F(1,\nu)$

Table: Appendix, Table A5

Gamma(α, λ)

Generic description: In a Poisson process, the time of the α-th event
Range of values: $x > 0$
Parameters: $\alpha \in (0,\infty)$, shape parameter
$\lambda \in (0,\infty)$, frequency of events, inverse scale parameter
Density function: $f(x) = \dfrac{\lambda^\alpha}{\Gamma(\alpha)}\, x^{\alpha-1}e^{-\lambda x}$
Expectation: $\mu = \alpha/\lambda$
Variance: $\sigma^2 = \alpha/\lambda^2$
Relations: For integer α, Gamma(α,λ) is a sum of α independent *Exponential*(λ) variables
(2) Gamma$(1,\lambda)$ = *Exponential*(λ)
(3) As a prior distribution, Gamma family is conjugate to the *Poisson*(θ) model.

Normal(μ, σ)

Generic description: Often used as distribution of errors, measurements, sums, averages, etc.
Range of values: $-\infty < x < +\infty$
Parameters: $\mu \in (-\infty,\infty)$, expectation, location parameter
$\sigma \in (0,\infty)$, standard deviation, scale parameter
Density function: $\dfrac{1}{\sigma\sqrt{2\pi}}\exp\left\{\dfrac{-(x-\mu)^2}{2\sigma^2}\right\}$

Cdf: $F(x) = \Phi\left(\dfrac{x-\mu}{\sigma}\right)$, use Table A4 for $\Phi(z)$

Expectation: μ

Variance: σ^2

Relations: Limiting case of standardized sums of random variables, including *Binomial*(n,p), *Negative Binomial*(k,p), and *Gamma*(α,λ) as $n,k,\alpha \to \infty$

Table: Appendix, Table A4

Pareto(θ, σ)

Generic description: A heavy-tail distribution often used to model amount of internet traffic, various financial and sociological data

Range of values: $x > \sigma$

Parameters: $\theta \in (0,\infty)$, shape parameter
$\sigma \in (0,\infty)$, scale parameter

Density function: $\theta\sigma^{\theta}x^{-\theta-1}$

Cdf: $F(x) = 1 - \left(\dfrac{x}{\sigma}\right)^{-\theta}$

Expectation: $\dfrac{\theta\sigma}{\theta-1}$ for $\theta > 1$; does not exist for $\theta \le 1$

Variance: $\dfrac{\theta\sigma^2}{(\theta-1)^2(\theta-2)}$ for $\theta > 2$; does not exist for $\theta \le 2$

Relations: If X is *Exponential*(θ) then $Y = \sigma e^X$ is *Pareto*(θ,σ)

Student's T(ν)

Generic description: Standardized statistic with an estimated standard deviation

Range of values: $-\infty < x < +\infty$

Parameter: $\nu > 0$, number of degrees of freedom

Density function: $f(x) = \dfrac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\dfrac{x^2}{\nu}\right)^{-\dfrac{\nu+1}{2}}$

Expectation: 0

Variance: $\dfrac{\nu}{\nu-2}$ for $\nu > 2$

Relations: Converges to Normal distribution as $\nu \to \infty$

Table: Appendix, Table A6

Uniform(a, b)

Generic description:	A number selected “at random” from a given interval
Range of values:	$a < x < b$
Parameters:	$-\infty < a < b < +\infty$, ends of the interval
Density function:	$f(x) = \dfrac{1}{b-a}$
Expectation:	$\mu = \dfrac{a+b}{2}$
Variance:	$\sigma^2 = \dfrac{(b-a)^2}{12}$
Relations:	Uniform(0,1) distribution is *Beta(1,1)*

11.2 Distribution tables

Table A1. Table of Uniform(0,1) random numbers

z	1	2	3	4	5	6	7	8	9	10
1	.9501	.8381	.7948	.4154	.6085	.4398	.2974	.7165	.7327	.8121
2	.2311	.0196	.9568	.3050	.0158	.3400	.0492	.5113	.4222	.6101
3	.6068	.6813	.5226	.8744	.0164	.3142	.6932	.7764	.9614	.7015
4	.4860	.3795	.8801	.0150	.1901	.3651	.6501	.4893	.0721	.0922
5	.8913	.8318	.1730	.7680	.5869	.3932	.9830	.1859	.5534	.4249
6	.7621	.5028	.9797	.9708	.0576	.5915	.5527	.7006	.2920	.3756
7	.4565	.7095	.2714	.9901	.3676	.1197	.4001	.9827	.8580	.1662
8	.0185	.4289	.2523	.7889	.6315	.0381	.1988	.8066	.3358	.8332
9	.8214	.3046	.8757	.4387	.7176	.4586	.6252	.7036	.6802	.8386
10	.4447	.1897	.7373	.4983	.6927	.8699	.7334	.4850	.0534	.4516
11	.6154	.1934	.1365	.2140	.0841	.9342	.3759	.1146	.3567	.9566
12	.7919	.6822	.0118	.6435	.4544	.2644	.0099	.6649	.4983	.1472
13	.9218	.3028	.8939	.3200	.4418	.1603	.4199	.3654	.4344	.8699
14	.7382	.5417	.1991	.9601	.3533	.8729	.7537	.1400	.5625	.7694
15	.1763	.1509	.2987	.7266	.1536	.2379	.7939	.5668	.6166	.4442
16	.4057	.6979	.6614	.4120	.6756	.6458	.9200	.8230	.1133	.6206
17	.9355	.3784	.2844	.7446	.6992	.9669	.8447	.6739	.8983	.9517
18	.9169	.8600	.4692	.2679	.7275	.6649	.3678	.9994	.7546	.6400
19	.4103	.8537	.0648	.4399	.4784	.8704	.6208	.9616	.7911	.2473
20	.8936	.5936	.9883	.9334	.5548	.0099	.7313	.0589	.8150	.3527
21	.0579	.4966	.5828	.6833	.1210	.1370	.1939	.3603	.6700	.1879
22	.3529	.8998	.4235	.2126	.4508	.8188	.9048	.5485	.2009	.4906
23	.8132	.8216	.5155	.8392	.7159	.4302	.5692	.2618	.2731	.4093
24	.0099	.6449	.3340	.6288	.8928	.8903	.6318	.5973	.6262	.4635
25	.1389	.8180	.4329	.1338	.2731	.7349	.2344	.0493	.5369	.6109
26	.2028	.6602	.2259	.2071	.2548	.6873	.5488	.5711	.0595	.0712
27	.1987	.3420	.5798	.6072	.8656	.3461	.9316	.7009	.0890	.3143
28	.6038	.2897	.7604	.6299	.2324	.1660	.3352	.9623	.2713	.6084
29	.2722	.3412	.5298	.3705	.8049	.1556	.6555	.7505	.4091	.1750
30	.1988	.5341	.6405	.5751	.9084	.1911	.3919	.7400	.4740	.6210
31	.0153	.7271	.2091	.4514	.2319	.4225	.6273	.4319	.9090	.2460
32	.7468	.3093	.3798	.0439	.2393	.8560	.6991	.6343	.5962	.5874
33	.4451	.8385	.7833	.0272	.0498	.4902	.3972	.8030	.3290	.5061
34	.9318	.5681	.6808	.3127	.0784	.8159	.4136	.0839	.4782	.4648
35	.4660	.3704	.4611	.0129	.6408	.4608	.6552	.9455	.5972	.5414
36	.4186	.7027	.5678	.3840	.1909	.4574	.8376	.9159	.1614	.9423
37	.8462	.5466	.7942	.6831	.8439	.4507	.3716	.6020	.8295	.3418
38	.5252	.4449	.0592	.0928	.1739	.4122	.4253	.2536	.9561	.4018
39	.2026	.6946	.6029	.0353	.1708	.9016	.5947	.8735	.5955	.3077
40	.6721	.6213	.0503	.6124	.9943	.0056	.5657	.5134	.0287	.4116

Table A2. Binomial distribution

$$F(x) = \boldsymbol{P}\{X \le x\} = \sum_{k=0}^{x} \binom{n}{k} p^k (1-p)^{n-k}$$

		p									
n	x	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
1	0	.9500	.9000	.8500	.8000	.7500	.7000	.6500	.6000	.5500	.5000
2	0	.9025	.8100	.7225	.6400	.5625	.4900	.4225	.3600	.3025	.2500
	1	.9975	.9900	.9775	.9600	.9375	.9100	.8775	.8400	.7975	.7500
3	0	.8574	.7290	.6141	.5120	.4219	.3430	.2746	.2160	.1664	.1250
	1	.9927	.9720	.9393	.8960	.8438	.7840	.7182	.6480	.5748	.5000
	2	.9999	.9990	.9966	.9920	.9844	.9730	.9571	.9360	.9089	.8750
4	0	.8145	.6561	.5220	.4096	.3164	.2401	.1785	.1296	.0915	.0625
	1	.9860	.9477	.8905	.8192	.7383	.6517	.5630	.4752	.3910	.3125
	2	.9995	.9963	.9880	.9728	.9492	.9163	.8735	.8208	.7585	.6875
	3	1.00	.9999	.9995	.9984	.9961	.9919	.9850	.9744	.9590	.9375
5	0	.7738	.5905	.4437	.3277	.2373	.1681	.1160	.0778	.0503	.0312
	1	.9774	.9185	.8352	.7373	.6328	.5282	.4284	.3370	.2562	.1875
	2	.9988	.9914	.9734	.9421	.8965	.8369	.7648	.6826	.5931	.5000
	3	1.00	.9995	.9978	.9933	.9844	.9692	.9460	.9130	.8688	.8125
	4	1.00	1.00	.9999	.9997	.9990	.9976	.9947	.9898	.9815	.9687
6	0	.7351	.5314	.3771	.2621	.1780	.1176	.0754	.0467	.0277	.0156
	1	.9672	.8857	.7765	.6554	.5339	.4202	.3191	.2333	.1636	.1094
	2	.9978	.9842	.9527	.9011	.8306	.7443	.6471	.5443	.4415	.3438
	3	.9999	.9987	.9941	.9830	.9624	.9295	.8826	.8208	.7447	.6563
	4	1.00	.9999	.9996	.9984	.9954	.9891	.9777	.9590	.9308	.8906
	5	1.00	1.00	1.00	.9999	.9998	.9993	.9982	.9959	.9917	.9844
7	0	.6983	.4783	.3206	.2097	.1335	.0824	.0490	.0280	.0152	.0078
	1	.9556	.8503	.7166	.5767	.4449	.3294	.2338	.1586	.1024	.0625
	2	.9962	.9743	.9262	.8520	.7564	.6471	.5323	.4199	.3164	.2266
	3	.9998	.9973	.9879	.9667	.9294	.8740	.8002	.7102	.6083	.5000
	4	1.00	.9998	.9988	.9953	.9871	.9712	.9444	.9037	.8471	.7734
	5	1.00	1.00	.9999	.9996	.9987	.9962	.9910	.9812	.9643	.9375
	6	1.00	1.00	1.00	1.00	.9999	.9998	.9994	.9984	.9963	.9922
8	0	.6634	.4305	.2725	.1678	.1001	.0576	.0319	.0168	.0084	.0039
	1	.9428	.8131	.6572	.5033	.3671	.2553	.1691	.1064	.0632	.0352
	2	.9942	.9619	.8948	.7969	.6785	.5518	.4278	.3154	.2201	.1445
	3	.9996	.9950	.9786	.9437	.8862	.8059	.7064	.5941	.4770	.3633
	4	1.00	.9996	.9971	.9896	.9727	.9420	.8939	.8263	.7396	.6367
	5	1.00	1.00	.9998	.9988	.9958	.9887	.9747	.9502	.9115	.8555
	6	1.00	1.00	1.00	.9999	.9996	.9987	.9964	.9915	.9819	.9648
	7	1.00	1.00	1.00	1.00	1.00	.9999	.9998	.9993	.9983	.9961
9	0	.6302	.3874	.2316	.1342	.0751	.0404	.0207	.0101	.0046	.0020
	1	.9288	.7748	.5995	.4362	.3003	.1960	.1211	.0705	.0385	.0195
	2	.9916	.9470	.8591	.7382	.6007	.4628	.3373	.2318	.1495	.0898
	3	.9994	.9917	.9661	.9144	.8343	.7297	.6089	.4826	.3614	.2539
	4	1.00	.9991	.9944	.9804	.9511	.9012	.8283	.7334	.6214	.5000
	5	1.00	.9999	.9994	.9969	.9900	.9747	.9464	.9006	.8342	.7461
	6	1.00	1.00	1.00	.9997	.9987	.9957	.9888	.9750	.9502	.9102
	7	1.00	1.00	1.00	1.00	.9999	.9996	.9986	.9962	.9909	.9805
	8	1.00	1.00	1.00	1.00	1.00	1.00	.9999	.9997	.9992	.9980
10	0	.5987	.3487	.1969	.1074	.0563	.0282	.0135	.0060	.0025	.0010
	1	.9139	.7361	.5443	.3758	.2440	.1493	.0860	.0464	.0233	.0107
	2	.9885	.9298	.8202	.6778	.5256	.3828	.2616	.1673	.0996	.0547
	3	.9990	.9872	.9500	.8791	.7759	.6496	.5138	.3823	.2660	.1719
	4	.9999	.9984	.9901	.9672	.9219	.8497	.7515	.6331	.5044	.3770
	5	1.00	.9999	.9986	.9936	.9803	.9527	.9051	.8338	.7384	.6230
	6	1.00	1.00	.9999	.9991	.9965	.9894	.9740	.9452	.8980	.8281
	7	1.00	1.00	1.00	.9999	.9996	.9984	.9952	.9877	.9726	.9453
	8	1.00	1.00	1.00	1.00	1.00	.9999	.9995	.9983	.9955	.9893
	9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.9999	.9997	.9990

Table A2, continued. Binomial distribution

		p								
n	*x*	.55	.60	.65	.70	.75	.80	.85	.90	.95
1	0	.4500	.4000	.3500	.3000	.2500	.2000	.1500	.1000	.0500
2	0	.2025	.1600	.1225	.0900	.0625	.0400	.0225	.0100	.0025
	1	.6975	.6400	.5775	.5100	.4375	.3600	.2775	.1900	.0975
3	0	.0911	.0640	.0429	.0270	.0156	.0080	.0034	.0010	.0001
	1	.4252	.3520	.2817	.2160	.1562	.1040	.0607	.0280	.0072
	2	.8336	.7840	.7254	.6570	.5781	.4880	.3859	.2710	.1426
4	0	.0410	.0256	.0150	.0081	.0039	.0016	.0005	.0001	.0000
	1	.2415	.1792	.1265	.0837	.0508	.0272	.0120	.0037	.0005
	2	.6090	.5248	.4370	.3483	.2617	.1808	.1095	.0523	.0140
	3	.9085	.8704	.8215	.7599	.6836	.5904	.4780	.3439	.1855
5	0	.0185	.0102	.0053	.0024	.0010	.0003	.0001	.0000	.0000
	1	.1312	.0870	.0540	.0308	.0156	.0067	.0022	.0005	.0000
	2	.4069	.3174	.2352	.1631	.1035	.0579	.0266	.0086	.0012
	3	.7438	.6630	.5716	.4718	.3672	.2627	.1648	.0815	.0226
	4	.9497	.9222	.8840	.8319	.7627	.6723	.5563	.4095	.2262
6	0	.0083	.0041	.0018	.0007	.0002	.0001	.0000	.0000	.0000
	1	.0692	.0410	.0223	.0109	.0046	.0016	.0004	.0001	.0000
	2	.2553	.1792	.1174	.0705	.0376	.0170	.0059	.0013	.0001
	3	.5585	.4557	.3529	.2557	.1694	.0989	.0473	.0158	.0022
	4	.8364	.7667	.6809	.5798	.4661	.3446	.2235	.1143	.0328
	5	.9723	.9533	.9246	.8824	.8220	.7379	.6229	.4686	.2649
7	0	.0037	.0016	.0006	.0002	.0001	.0000	.0000	.0000	.0000
	1	.0357	.0188	.0090	.0038	.0013	.0004	.0001	.0000	.0000
	2	.1529	.0963	.0556	.0288	.0129	.0047	.0012	.0002	.0000
	3	.3917	.2898	.1998	.1260	.0706	.0333	.0121	.0027	.0002
	4	.6836	.5801	.4677	.3529	.2436	.1480	.0738	.0257	.0038
	5	.8976	.8414	.7662	.6706	.5551	.4233	.2834	.1497	.0444
	6	.9848	.9720	.9510	.9176	.8665	.7903	.6794	.5217	.3017
8	0	.0017	.0007	.0002	.0001	.0000	.0000	.0000	.0000	.0000
	1	.0181	.0085	.0036	.0013	.0004	.0001	.0000	.0000	.0000
	2	.0885	.0498	.0253	.0113	.0042	.0012	.0002	.0000	.0000
	3	.2604	.1737	.1061	.0580	.0273	.0104	.0029	.0004	.0000
	4	.5230	.4059	.2936	.1941	.1138	.0563	.0214	.0050	.0004
	5	.7799	.6846	.5722	.4482	.3215	.2031	.1052	.0381	.0058
	6	.9368	.8936	.8309	.7447	.6329	.4967	.3428	.1869	.0572
	7	.9916	.9832	.9681	.9424	.8999	.8322	.7275	.5695	.3366
9	0	.0008	.0003	.0001	.0000	.0000	.0000	.0000	.0000	.0000
	1	.0091	.0038	.0014	.0004	.0001	.0000	.0000	.0000	.0000
	2	.0498	.0250	.0112	.0043	.0013	.0003	.0000	.0000	.0000
	3	.1658	.0994	.0536	.0253	.0100	.0031	.0006	.0001	.0000
	4	.3786	.2666	.1717	.0988	.0489	.0196	.0056	.0009	.0000
	5	.6386	.5174	.3911	.2703	.1657	.0856	.0339	.0083	.0006
	6	.8505	.7682	.6627	.5372	.3993	.2618	.1409	.0530	.0084
	7	.9615	.9295	.8789	.8040	.6997	.5638	.4005	.2252	.0712
	8	.9954	.9899	.9793	.9596	.9249	.8658	.7684	.6126	.3698
10	0	.0003	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	1	.0045	.0017	.0005	.0001	.0000	.0000	.0000	.0000	.0000
	2	.0274	.0123	.0048	.0016	.0004	.0001	.0000	.0000	.0000
	3	.1020	.0548	.0260	.0106	.0035	.0009	.0001	.0000	.0000
	4	.2616	.1662	.0949	.0473	.0197	.0064	.0014	.0001	.0000
	5	.4956	.3669	.2485	.1503	.0781	.0328	.0099	.0016	.0001
	6	.7340	.6177	.4862	.3504	.2241	.1209	.0500	.0128	.0010
	7	.9004	.8327	.7384	.6172	.4744	.3222	.1798	.0702	.0115
	8	.9767	.9536	.9140	.8507	.7560	.6242	.4557	.2639	.0861
	9	.9975	.9940	.9865	.9718	.9437	.8926	.8031	.6513	.4013

Table A2, continued. Binomial distribution

		p									
n	*x*	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
11	0	.5688	.3138	.1673	.0859	.0422	.0198	.0088	.0036	.0014	.0005
	1	.8981	.6974	.4922	.3221	.1971	.1130	.0606	.0302	.0139	.0059
	2	.9848	.9104	.7788	.6174	.4552	.3127	.2001	.1189	.0652	.0327
	3	.9984	.9815	.9306	.8389	.7133	.5696	.4256	.2963	.1911	.1133
	4	.9999	.9972	.9841	.9496	.8854	.7897	.6683	.5328	.3971	.2744
	5	1.00	.9997	.9973	.9883	.9657	.9218	.8513	.7535	.6331	.5000
	6	1.00	1.00	.9997	.9980	.9924	.9784	.9499	.9006	.8262	.7256
	7	1.00	1.00	1.00	.9998	.9988	.9957	.9878	.9707	.9390	.8867
	8	1.00	1.00	1.00	1.00	.9999	.9994	.9980	.9941	.9852	.9673
	9	1.00	1.00	1.00	1.00	1.00	1.00	.9998	.9993	.9978	.9941
	10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.9998	.9995
12	0	.5404	.2824	.1422	.0687	.0317	.0138	.0057	.0022	.0008	.0002
	1	.8816	.6590	.4435	.2749	.1584	.0850	.0424	.0196	.0083	.0032
	2	.9804	.8891	.7358	.5583	.3907	.2528	.1513	.0834	.0421	.0193
	3	.9978	.9744	.9078	.7946	.6488	.4925	.3467	.2253	.1345	.0730
	4	.9998	.9957	.9761	.9274	.8424	.7237	.5833	.4382	.3044	.1938
	5	1.00	.9995	.9954	.9806	.9456	.8822	.7873	.6652	.5269	.3872
	6	1.00	.9999	.9993	.9961	.9857	.9614	.9154	.8418	.7393	.6128
	7	1.00	1.00	.9999	.9994	.9972	.9905	.9745	.9427	.8883	.8062
	8	1.00	1.00	1.00	.9999	.9996	.9983	.9944	.9847	.9644	.9270
	9	1.00	1.00	1.00	1.00	1.00	.9998	.9992	.9972	.9921	.9807
	10	1.00	1.00	1.00	1.00	1.00	1.00	.9999	.9997	.9989	.9968
	11	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.9999	.9998
13	0	.5133	.2542	.1209	.0550	.0238	.0097	.0037	.0013	.0004	.0001
	1	.8646	.6213	.3983	.2336	.1267	.0637	.0296	.0126	.0049	.0017
	2	.9755	.8661	.6920	.5017	.3326	.2025	.1132	.0579	.0269	.0112
	3	.9969	.9658	.8820	.7473	.5843	.4206	.2783	.1686	.0929	.0461
	4	.9997	.9935	.9658	.9009	.7940	.6543	.5005	.3530	.2279	.1334
	5	1.00	.9991	.9925	.9700	.9198	.8346	.7159	.5744	.4268	.2905
	6	1.00	.9999	.9987	.9930	.9757	.9376	.8705	.7712	.6437	.5000
	7	1.00	1.00	.9998	.9988	.9944	.9818	.9538	.9023	.8212	.7095
	8	1.00	1.00	1.00	.9998	.9990	.9960	.9874	.9679	.9302	.8666
	9	1.00	1.00	1.00	1.00	.9999	.9993	.9975	.9922	.9797	.9539
	10	1.00	1.00	1.00	1.00	1.00	.9999	.9997	.9987	.9959	.9888
	11	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.9999	.9995	.9983
	12	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.9999
14	0	.4877	.2288	.1028	.0440	.0178	.0068	.0024	.0008	.0002	.0001
	1	.8470	.5846	.3567	.1979	.1010	.0475	.0205	.0081	.0029	.0009
	2	.9699	.8416	.6479	.4481	.2811	.1608	.0839	.0398	.0170	.0065
	3	.9958	.9559	.8535	.6982	.5213	.3552	.2205	.1243	.0632	.0287
	4	.9996	.9908	.9533	.8702	.7415	.5842	.4227	.2793	.1672	.0898
	5	1.00	.9985	.9885	.9561	.8883	.7805	.6405	.4859	.3373	.2120
	6	1.00	.9998	.9978	.9884	.9617	.9067	.8164	.6925	.5461	.3953
	7	1.00	1.00	.9997	.9976	.9897	.9685	.9247	.8499	.7414	.6047
	8	1.00	1.00	1.00	.9996	.9978	.9917	.9757	.9417	.8811	.7880
	9	1.00	1.00	1.00	1.00	.9997	.9983	.9940	.9825	.9574	.9102
	10	1.00	1.00	1.00	1.00	1.00	.9998	.9989	.9961	.9886	.9713
	11	1.00	1.00	1.00	1.00	1.00	1.00	.9999	.9994	.9978	.9935
	12	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.9999	.9997	.9991
	13	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.9999
15	0	.4633	.2059	.0874	.0352	.0134	.0047	.0016	.0005	.0001	.0000
	1	.8290	.5490	.3186	.1671	.0802	.0353	.0142	.0052	.0017	.0005
	2	.9638	.8159	.6042	.3980	.2361	.1268	.0617	.0271	.0107	.0037
	3	.9945	.9444	.8227	.6482	.4613	.2969	.1727	.0905	.0424	.0176
	4	.9994	.9873	.9383	.8358	.6865	.5155	.3519	.2173	.1204	.0592
	5	.9999	.9978	.9832	.9389	.8516	.7216	.5643	.4032	.2608	.1509
	6	1.00	.9997	.9964	.9819	.9434	.8689	.7548	.6098	.4522	.3036
	7	1.00	1.00	.9994	.9958	.9827	.9500	.8868	.7869	.6535	.5000
	8	1.00	1.00	.9999	.9992	.9958	.9848	.9578	.9050	.8182	.6964
	9	1.00	1.00	1.00	.9999	.9992	.9963	.9876	.9662	.9231	.8491
	10	1.00	1.00	1.00	1.00	.9999	.9993	.9972	.9907	.9745	.9408
	11	1.00	1.00	1.00	1.00	1.00	.9999	.9995	.9981	.9937	.9824
	12	1.00	1.00	1.00	1.00	1.00	1.00	.9999	.9997	.9989	.9963
	13	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.9999	.9995
	14	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table A2, continued. Binomial distribution

		p								
n	*x*	.55	.60	.65	.70	.75	.80	.85	.90	.95
11	0	.0002	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	1	.0022	.0007	.0002	.0000	.0000	.0000	.0000	.0000	.0000
	2	.0148	.0059	.0020	.0006	.0001	.0000	.0000	.0000	.0000
	3	.0610	.0293	.0122	.0043	.0012	.0002	.0000	.0000	.0000
	4	.1738	.0994	.0501	.0216	.0076	.0020	.0003	.0000	.0000
	5	.3669	.2465	.1487	.0782	.0343	.0117	.0027	.0003	.0000
	6	.6029	.4672	.3317	.2103	.1146	.0504	.0159	.0028	.0001
	7	.8089	.7037	.5744	.4304	.2867	.1611	.0694	.0185	.0016
	8	.9348	.8811	.7999	.6873	.5448	.3826	.2212	.0896	.0152
	9	.9861	.9698	.9394	.8870	.8029	.6779	.5078	.3026	.1019
	10	.9986	.9964	.9912	.9802	.9578	.9141	.8327	.6862	.4312
12	0	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	1	.0011	.0003	.0001	.0000	.0000	.0000	.0000	.0000	.0000
	2	.0079	.0028	.0008	.0002	.0000	.0000	.0000	.0000	.0000
	3	.0356	.0153	.0056	.0017	.0004	.0001	.0000	.0000	.0000
	4	.1117	.0573	.0255	.0095	.0028	.0006	.0001	.0000	.0000
	5	.2607	.1582	.0846	.0386	.0143	.0039	.0007	.0001	.0000
	6	.4731	.3348	.2127	.1178	.0544	.0194	.0046	.0005	.0000
	7	.6956	.5618	.4167	.2763	.1576	.0726	.0239	.0043	.0002
	8	.8655	.7747	.6533	.5075	.3512	.2054	.0922	.0256	.0022
	9	.9579	.9166	.8487	.7472	.6093	.4417	.2642	.1109	.0196
	10	.9917	.9804	.9576	.9150	.8416	.7251	.5565	.3410	.1184
	11	.9992	.9978	.9943	.9862	.9683	.9313	.8578	.7176	.4596
13	0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	1	.0005	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	2	.0041	.0013	.0003	.0001	.0000	.0000	.0000	.0000	.0000
	3	.0203	.0078	.0025	.0007	.0001	.0000	.0000	.0000	.0000
	4	.0698	.0321	.0126	.0040	.0010	.0002	.0000	.0000	.0000
	5	.1788	.0977	.0462	.0182	.0056	.0012	.0002	.0000	.0000
	6	.3563	.2288	.1295	.0624	.0243	.0070	.0013	.0001	.0000
	7	.5732	.4256	.2841	.1654	.0802	.0300	.0075	.0009	.0000
	8	.7721	.6470	.4995	.3457	.2060	.0991	.0342	.0065	.0003
	9	.9071	.8314	.7217	.5794	.4157	.2527	.1180	.0342	.0031
	10	.9731	.9421	.8868	.7975	.6674	.4983	.3080	.1339	.0245
	11	.9951	.9874	.9704	.9363	.8733	.7664	.6017	.3787	.1354
	12	.9996	.9987	.9963	.9903	.9762	.9450	.8791	.7458	.4867
14	0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	1	.0003	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	2	.0022	.0006	.0001	.0000	.0000	.0000	.0000	.0000	.0000
	3	.0114	.0039	.0011	.0002	.0000	.0000	.0000	.0000	.0000
	4	.0426	.0175	.0060	.0017	.0003	.0000	.0000	.0000	.0000
	5	.1189	.0583	.0243	.0083	.0022	.0004	.0000	.0000	.0000
	6	.2586	.1501	.0753	.0315	.0103	.0024	.0003	.0000	.0000
	7	.4539	.3075	.1836	.0933	.0383	.0116	.0022	.0002	.0000
	8	.6627	.5141	.3595	.2195	.1117	.0439	.0115	.0015	.0000
	9	.8328	.7207	.5773	.4158	.2585	.1298	.0467	.0092	.0004
	10	.9368	.8757	.7795	.6448	.4787	.3018	.1465	.0441	.0042
	11	.9830	.9602	.9161	.8392	.7189	.5519	.3521	.1584	.0301
	12	.9971	.9919	.9795	.9525	.8990	.8021	.6433	.4154	.1530
	13	.9998	.9992	.9976	.9932	.9822	.9560	.8972	.7712	.5123
15	0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	1	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	2	.0011	.0003	.0001	.0000	.0000	.0000	.0000	.0000	.0000
	3	.0063	.0019	.0005	.0001	.0000	.0000	.0000	.0000	.0000
	4	.0255	.0093	.0028	.0007	.0001	.0000	.0000	.0000	.0000
	5	.0769	.0338	.0124	.0037	.0008	.0001	.0000	.0000	.0000
	6	.1818	.0950	.0422	.0152	.0042	.0008	.0001	.0000	.0000
	7	.3465	.2131	.1132	.0500	.0173	.0042	.0006	.0000	.0000
	8	.5478	.3902	.2452	.1311	.0566	.0181	.0036	.0003	.0000
	9	.7392	.5968	.4357	.2784	.1484	.0611	.0168	.0022	.0001
	10	.8796	.7827	.6481	.4845	.3135	.1642	.0617	.0127	.0006
	11	.9576	.9095	.8273	.7031	.5387	.3518	.1773	.0556	.0055
	12	.9893	.9729	.9383	.8732	.7639	.6020	.3958	.1841	.0362
	13	.9983	.9948	.9858	.9647	.9198	.8329	.6814	.4510	.1710
	14	.9999	.9995	.9984	.9953	.9866	.9648	.9126	.7941	.5367

Table A2, continued. Binomial distribution

		p									
n	x	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
16	0	.4401	.1853	.0743	.0281	.0100	.0033	.0010	.0003	.0001	.0000
	1	.8108	.5147	.2839	.1407	.0635	.0261	.0098	.0033	.0010	.0003
	2	.9571	.7892	.5614	.3518	.1971	.0994	.0451	.0183	.0066	.0021
	3	.9930	.9316	.7899	.5981	.4050	.2459	.1339	.0651	.0281	.0106
	4	.9991	.9830	.9209	.7982	.6302	.4499	.2892	.1666	.0853	.0384
	5	.9999	.9967	.9765	.9183	.8103	.6598	.4900	.3288	.1976	.1051
	6	1.00	.9995	.9944	.9733	.9204	.8247	.6881	.5272	.3660	.2272
	7	1.00	.9999	.9989	.9930	.9729	.9256	.8406	.7161	.5629	.4018
	8	1.00	1.00	.9998	.9985	.9925	.9743	.9329	.8577	.7441	.5982
	9	1.00	1.00	1.00	.9998	.9984	.9929	.9771	.9417	.8759	.7728
	10	1.00	1.00	1.00	1.00	.9997	.9984	.9938	.9809	.9514	.8949
	11	1.00	1.00	1.00	1.00	1.00	.9997	.9987	.9951	.9851	.9616
	12	1.00	1.00	1.00	1.00	1.00	1.00	.9998	.9991	.9965	.9894
	13	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.9999	.9994	.9979
18	0	.3972	.1501	.0536	.0180	.0056	.0016	.0004	.0001	.0000	.0000
	1	.7735	.4503	.2241	.0991	.0395	.0142	.0046	.0013	.0003	.0001
	2	.9419	.7338	.4797	.2713	.1353	.0600	.0236	.0082	.0025	.0007
	3	.9891	.9018	.7202	.5010	.3057	.1646	.0783	.0328	.0120	.0038
	4	.9985	.9718	.8794	.7164	.5187	.3327	.1886	.0942	.0411	.0154
	5	.9998	.9936	.9581	.8671	.7175	.5344	.3550	.2088	.1077	.0481
	6	1.00	.9988	.9882	.9487	.8610	.7217	.5491	.3743	.2258	.1189
	7	1.00	.9998	.9973	.9837	.9431	.8593	.7283	.5634	.3915	.2403
	8	1.00	1.00	.9995	.9957	.9807	.9404	.8609	.7368	.5778	.4073
	9	1.00	1.00	.9999	.9991	.9946	.9790	.9403	.8653	.7473	.5927
	10	1.00	1.00	1.00	.9998	.9988	.9939	.9788	.9424	.8720	.7597
	11	1.00	1.00	1.00	1.00	.9998	.9986	.9938	.9797	.9463	.8811
	12	1.00	1.00	1.00	1.00	1.00	.9997	.9986	.9942	.9817	.9519
	13	1.00	1.00	1.00	1.00	1.00	1.00	.9997	.9987	.9951	.9846
	14	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.9998	.9990	.9962
20	0	.3585	.1216	.0388	.0115	.0032	.0008	.0002	.0000	.0000	.0000
	1	.7358	.3917	.1756	.0692	.0243	.0076	.0021	.0005	.0001	.0000
	2	.9245	.6769	.4049	.2061	.0913	.0355	.0121	.0036	.0009	.0002
	3	.9841	.8670	.6477	.4114	.2252	.1071	.0444	.0160	.0049	.0013
	4	.9974	.9568	.8298	.6296	.4148	.2375	.1182	.0510	.0189	.0059
	5	.9997	.9887	.9327	.8042	.6172	.4164	.2454	.1256	.0553	.0207
	6	1.00	.9976	.9781	.9133	.7858	.6080	.4166	.2500	.1299	.0577
	7	1.00	.9996	.9941	.9679	.8982	.7723	.6010	.4159	.2520	.1316
	8	1.00	.9999	.9987	.9900	.9591	.8867	.7624	.5956	.4143	.2517
	9	1.00	1.00	.9998	.9974	.9861	.9520	.8782	.7553	.5914	.4119
	10	1.00	1.00	1.00	.9994	.9961	.9829	.9468	.8725	.7507	.5881
	11	1.00	1.00	1.00	.9999	.9991	.9949	.9804	.9435	.8692	.7483
	12	1.00	1.00	1.00	1.00	.9998	.9987	.9940	.9790	.9420	.8684
	13	1.00	1.00	1.00	1.00	1.00	.9997	.9985	.9935	.9786	.9423
	14	1.00	1.00	1.00	1.00	1.00	1.00	.9997	.9984	.9936	.9793
	15	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.9997	.9985	.9941
25	0	.2774	.0718	.0172	.0038	.0008	.0001	.0000	.0000	.0000	.0000
	1	.6424	.2712	.0931	.0274	.0070	.0016	.0003	.0001	.0000	.0000
	2	.8729	.5371	.2537	.0982	.0321	.0090	.0021	.0004	.0001	.0000
	3	.9659	.7636	.4711	.2340	.0962	.0332	.0097	.0024	.0005	.0001
	4	.9928	.9020	.6821	.4207	.2137	.0905	.0320	.0095	.0023	.0005
	5	.9988	.9666	.8385	.6167	.3783	.1935	.0826	.0294	.0086	.0020
	6	.9998	.9905	.9305	.7800	.5611	.3407	.1734	.0736	.0258	.0073
	7	1.00	.9977	.9745	.8909	.7265	.5118	.3061	.1536	.0639	.0216
	8	1.00	.9995	.9920	.9532	.8506	.6769	.4668	.2735	.1340	.0539
	9	1.00	.9999	.9979	.9827	.9287	.8106	.6303	.4246	.2424	.1148
	10	1.00	1.00	.9995	.9944	.9703	.9022	.7712	.5858	.3843	.2122
	11	1.00	1.00	.9999	.9985	.9893	.9558	.8746	.7323	.5426	.3450
	12	1.00	1.00	1.00	.9996	.9966	.9825	.9396	.8462	.6937	.5000
	13	1.00	1.00	1.00	.9999	.9991	.9940	.9745	.9222	.8173	.6550
	14	1.00	1.00	1.00	1.00	.9998	.9982	.9907	.9656	.9040	.7878
	15	1.00	1.00	1.00	1.00	1.00	.9995	.9971	.9868	.9560	.8852
	16	1.00	1.00	1.00	1.00	1.00	.9999	.9992	.9957	.9826	.9461
	17	1.00	1.00	1.00	1.00	1.00	1.00	.9998	.9988	.9942	.9784
	18	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.9997	.9984	.9927

Table A2, continued. Binomial distribution

		p								
n	x	.55	.60	.65	.70	.75	.80	.85	.90	.95
16	2	.0006	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	3	.0035	.0009	.0002	.0000	.0000	.0000	.0000	.0000	.0000
	4	.0149	.0049	.0013	.0003	.0000	.0000	.0000	.0000	.0000
	5	.0486	.0191	.0062	.0016	.0003	.0000	.0000	.0000	.0000
	6	.1241	.0583	.0229	.0071	.0016	.0002	.0000	.0000	.0000
	7	.2559	.1423	.0671	.0257	.0075	.0015	.0002	.0000	.0000
	8	.4371	.2839	.1594	.0744	.0271	.0070	.0011	.0001	.0000
	9	.6340	.4728	.3119	.1753	.0796	.0267	.0056	.0005	.0000
	10	.8024	.6712	.5100	.3402	.1897	.0817	.0235	.0033	.0001
	11	.9147	.8334	.7108	.5501	.3698	.2018	.0791	.0170	.0009
	12	.9719	.9349	.8661	.7541	.5950	.4019	.2101	.0684	.0070
	13	.9934	.9817	.9549	.9006	.8029	.6482	.4386	.2108	.0429
	14	.9990	.9967	.9902	.9739	.9365	.8593	.7161	.4853	.1892
	15	.9999	.9997	.9990	.9967	.9900	.9719	.9257	.8147	.5599
18	3	.0010	.0002	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	4	.0049	.0013	.0003	.0000	.0000	.0000	.0000	.0000	.0000
	5	.0183	.0058	.0014	.0003	.0000	.0000	.0000	.0000	.0000
	6	.0537	.0203	.0062	.0014	.0002	.0000	.0000	.0000	.0000
	7	.1280	.0576	.0212	.0061	.0012	.0002	.0000	.0000	.0000
	8	.2527	.1347	.0597	.0210	.0054	.0009	.0001	.0000	.0000
	9	.4222	.2632	.1391	.0596	.0193	.0043	.0005	.0000	.0000
	10	.6085	.4366	.2717	.1407	.0569	.0163	.0027	.0002	.0000
	11	.7742	.6257	.4509	.2783	.1390	.0513	.0118	.0012	.0000
	12	.8923	.7912	.6450	.4656	.2825	.1329	.0419	.0064	.0002
	13	.9589	.9058	.8114	.6673	.4813	.2836	.1206	.0282	.0015
	14	.9880	.9672	.9217	.8354	.6943	.4990	.2798	.0982	.0109
	15	.9975	.9918	.9764	.9400	.8647	.7287	.5203	.2662	.0581
	16	.9997	.9987	.9954	.9858	.9605	.9009	.7759	.5497	.2265
	17	1.00	.9999	.9996	.9984	.9944	.9820	.9464	.8499	.6028
20	4	.0015	.0003	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	5	.0064	.0016	.0003	.0000	.0000	.0000	.0000	.0000	.0000
	6	.0214	.0065	.0015	.0003	.0000	.0000	.0000	.0000	.0000
	7	.0580	.0210	.0060	.0013	.0002	.0000	.0000	.0000	.0000
	8	.1308	.0565	.0196	.0051	.0009	.0001	.0000	.0000	.0000
	9	.2493	.1275	.0532	.0171	.0039	.0006	.0000	.0000	.0000
	10	.4086	.2447	.1218	.0480	.0139	.0026	.0002	.0000	.0000
	11	.5857	.4044	.2376	.1133	.0409	.0100	.0013	.0001	.0000
	12	.7480	.5841	.3990	.2277	.1018	.0321	.0059	.0004	.0000
	13	.8701	.7500	.5834	.3920	.2142	.0867	.0219	.0024	.0000
	14	.9447	.8744	.7546	.5836	.3828	.1958	.0673	.0113	.0003
	15	.9811	.9490	.8818	.7625	.5852	.3704	.1702	.0432	.0026
	16	.9951	.9840	.9556	.8929	.7748	.5886	.3523	.1330	.0159
	17	.9991	.9964	.9879	.9645	.9087	.7939	.5951	.3231	.0755
	18	.9999	.9995	.9979	.9924	.9757	.9308	.8244	.6083	.2642
	19	1.00	1.00	.9998	.9992	.9968	.9885	.9612	.8784	.6415
25	6	.0016	.0003	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	7	.0058	.0012	.0002	.0000	.0000	.0000	.0000	.0000	.0000
	8	.0174	.0043	.0008	.0001	.0000	.0000	.0000	.0000	.0000
	9	.0440	.0132	.0029	.0005	.0000	.0000	.0000	.0000	.0000
	10	.0960	.0344	.0093	.0018	.0002	.0000	.0000	.0000	.0000
	11	.1827	.0778	.0255	.0060	.0009	.0001	.0000	.0000	.0000
	12	.3063	.1538	.0604	.0175	.0034	.0004	.0000	.0000	.0000
	13	.4574	.2677	.1254	.0442	.0107	.0015	.0001	.0000	.0000
	14	.6157	.4142	.2288	.0978	.0297	.0056	.0005	.0000	.0000
	15	.7576	.5754	.3697	.1894	.0713	.0173	.0021	.0001	.0000
	16	.8660	.7265	.5332	.3231	.1494	.0468	.0080	.0005	.0000
	17	.9361	.8464	.6939	.4882	.2735	.1091	.0255	.0023	.0000
	18	.9742	.9264	.8266	.6593	.4389	.2200	.0695	.0095	.0002
	19	.9914	.9706	.9174	.8065	.6217	.3833	.1615	.0334	.0012
	20	.9977	.9905	.9680	.9095	.7863	.5793	.3179	.0980	.0072
	21	.9995	.9976	.9903	.9668	.9038	.7660	.5289	.2364	.0341
	22	.9999	.9996	.9979	.9910	.9679	.9018	.7463	.4629	.1271
	23	1.00	.9999	.9997	.9984	.9930	.9726	.9069	.7288	.3576
	24	1.00	1.00	1.00	.9999	.9992	.9962	.9828	.9282	.7226

Table A3. Poisson distribution

$$F(x) = \boldsymbol{P}\{X \le x\} = \sum_{k=0}^{x} \frac{e^{-\lambda}\lambda^k}{k!}$$

	λ														
x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.4	1.5
0	.905	.819	.741	.670	.607	.549	.497	.449	.407	.368	.333	.301	.273	.247	.223
1	.995	.982	.963	.938	.910	.878	.844	.809	.772	.736	.699	.663	.627	.592	.558
2	1.00	.999	.996	.992	.986	.977	.966	.953	.937	.920	.900	.879	.857	.833	.809
3	1.00	1.00	1.00	.999	.998	.997	.994	.991	.987	.981	.974	.966	.957	.946	.934
4	1.00	1.00	1.00	1.00	1.00	1.00	.999	.999	.998	.996	.995	.992	.989	.986	.981
5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.999	.999	.998	.998	.997	.996
6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.999	.999
7	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

	λ														
x	1.6	1.7	1.8	1.9	2.0	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	2.9	3.0
0	.202	.183	.165	.150	.135	.122	.111	.100	.091	.082	.074	.067	.061	.055	.050
1	.525	.493	.463	.434	.406	.380	.355	.331	.308	.287	.267	.249	.231	.215	.199
2	.783	.757	.731	.704	.677	.650	.623	.596	.570	.544	.518	.494	.469	.446	.423
3	.921	.907	.891	.875	.857	.839	.819	.799	.779	.758	.736	.714	.692	.670	.647
4	.976	.970	.964	.956	.947	.938	.928	.916	.904	.891	.877	.863	.848	.832	.815
5	.994	.992	.990	.987	.983	.980	.975	.970	.964	.958	.951	.943	.935	.926	.916
6	.999	.998	.997	.997	.995	.994	.993	.991	.988	.986	.983	.979	.976	.971	.966
7	1.00	1.00	.999	.999	.999	.999	.998	.997	.997	.996	.995	.993	.992	.990	.988
8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.999	.999	.999	.999	.998	.998	.997	.996
9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.999	.999	.999	.999
10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

	λ														
x	3.5	4.0	4.5	5.0	5.5	6.0	6.5	7.0	7.5	8.0	8.5	9.0	9.5	10.0	10.5
0	.030	.018	.011	.007	.004	.002	.002	.001	.001	.000	.000	.000	.000	.000	.000
1	.136	.092	.061	.040	.027	.017	.011	.007	.005	.003	.002	.001	.001	.000	.000
2	.321	.238	.174	.125	.088	.062	.043	.030	.020	.014	.009	.006	.004	.003	.002
3	.537	.433	.342	.265	.202	.151	.112	.082	.059	.042	.030	.021	.015	.010	.007
4	.725	.629	.532	.440	.358	.285	.224	.173	.132	.100	.074	.055	.040	.029	.021
5	.858	.785	.703	.616	.529	.446	.369	.301	.241	.191	.150	.116	.089	.067	.050
6	.935	.889	.831	.762	.686	.606	.527	.450	.378	.313	.256	.207	.165	.130	.102
7	.973	.949	.913	.867	.809	.744	.673	.599	.525	.453	.386	.324	.269	.220	.179
8	.990	.979	.960	.932	.894	.847	.792	.729	.662	.593	.523	.456	.392	.333	.279
9	.997	.992	.983	.968	.946	.916	.877	.830	.776	.717	.653	.587	.522	.458	.397
10	.999	.997	.993	.986	.975	.957	.933	.901	.862	.816	.763	.706	.645	.583	.521
11	1.00	.999	.998	.995	.989	.980	.966	.947	.921	.888	.849	.803	.752	.697	.639
12	1.00	1.00	.999	.998	.996	.991	.984	.973	.957	.936	.909	.876	.836	.792	.742
13	1.00	1.00	1.00	.999	.998	.996	.993	.987	.978	.966	.949	.926	.898	.864	.825
14	1.00	1.00	1.00	1.00	.999	.999	.997	.994	.990	.983	.973	.959	.940	.917	.888
15	1.00	1.00	1.00	1.00	1.00	.999	.999	.998	.995	.992	.986	.978	.967	.951	.932
16	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.999	.998	.996	.993	.989	.982	.973	.960
17	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.999	.998	.997	.995	.991	.986	.978
18	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.999	.999	.998	.996	.993	.988
19	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.999	.999	.998	.997	.994
20	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.999	.998	.997

Table A3, continued. Poisson distribution

	λ														
x	11	12	13	14	15	16	17	18	19	20	22	24	26	28	30
0	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
1	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
2	.001	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
3	.005	.002	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
4	.015	.008	.004	.002	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
5	.038	.020	.011	.006	.003	.001	.001	.000	.000	.000	.000	.000	.000	.000	.000
6	.079	.046	.026	.014	.008	.004	.002	.001	.001	.000	.000	.000	.000	.000	.000
7	.143	.090	.054	.032	.018	.010	.005	.003	.002	.001	.000	.000	.000	.000	.000
8	.232	.155	.100	.062	.037	.022	.013	.007	.004	.002	.001	.000	.000	.000	.000
9	.341	.242	.166	.109	.070	.043	.026	.015	.009	.005	.002	.000	.000	.000	.000
10	.460	.347	.252	.176	.118	.077	.049	.030	.018	.011	.004	.001	.000	.000	.000
11	.579	.462	.353	.260	.185	.127	.085	.055	.035	.021	.008	.003	.001	.000	.000
12	.689	.576	.463	.358	.268	.193	.135	.092	.061	.039	.015	.005	.002	.001	.000
13	.781	.682	.573	.464	.363	.275	.201	.143	.098	.066	.028	.011	.004	.001	.000
14	.854	.772	.675	.570	.466	.368	.281	.208	.150	.105	.048	.020	.008	.003	.001
15	.907	.844	.764	.669	.568	.467	.371	.287	.215	.157	.077	.034	.014	.005	.002
16	.944	.899	.835	.756	.664	.566	.468	.375	.292	.221	.117	.056	.025	.010	.004
17	.968	.937	.890	.827	.749	.659	.564	.469	.378	.297	.169	.087	.041	.018	.007
18	.982	.963	.930	.883	.819	.742	.655	.562	.469	.381	.232	.128	.065	.030	.013
19	.991	.979	.957	.923	.875	.812	.736	.651	.561	.470	.306	.180	.097	.048	.022
20	.995	.988	.975	.952	.917	.868	.805	.731	.647	.559	.387	.243	.139	.073	.035
21	.998	.994	.986	.971	.947	.911	.861	.799	.725	.644	.472	.314	.190	.106	.054
22	.999	.997	.992	.983	.967	.942	.905	.855	.793	.721	.556	.392	.252	.148	.081
23	1.00	.999	.996	.991	.981	.963	.937	.899	.849	.787	.637	.473	.321	.200	.115
24	1.00	.999	.998	.995	.989	.978	.959	.932	.893	.843	.712	.554	.396	.260	.157
25	1.00	1.00	.999	.997	.994	.987	.975	.955	.927	.888	.777	.632	.474	.327	.208
26	1.00	1.00	1.00	.999	.997	.993	.985	.972	.951	.922	.832	.704	.552	.400	.267
27	1.00	1.00	1.00	.999	.998	.996	.991	.983	.969	.948	.877	.768	.627	.475	.333
28	1.00	1.00	1.00	1.00	.999	.998	.995	.990	.980	.966	.913	.823	.697	.550	.403
29	1.00	1.00	1.00	1.00	1.00	.999	.997	.994	.988	.978	.940	.868	.759	.623	.476
30	1.00	1.00	1.00	1.00	1.00	.999	.999	.997	.993	.987	.959	.904	.813	.690	.548
31	1.00	1.00	1.00	1.00	1.00	1.00	.999	.998	.996	.992	.973	.932	.859	.752	.619
32	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.999	.998	.995	.983	.953	.896	.805	.685
33	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.999	.997	.989	.969	.925	.850	.744
34	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.999	.999	.994	.979	.947	.888	.797
35	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.999	.996	.987	.964	.918	.843
36	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.998	.992	.976	.941	.880
37	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.999	.995	.984	.959	.911
38	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.999	.997	.990	.972	.935
39	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.998	.994	.981	.954
40	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.999	.996	.988	.968
41	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.999	.998	.992	.978
42	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.999	.995	.985
43	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.999	.997	.990
44	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.998	.994
45	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.999	.996
46	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.999	.998
47	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.999
48	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.999
49	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.999
50	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table A4. Standard Normal distribution

$$\Phi(z) = \boldsymbol{P}\left\{Z \leq z\right\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2} dx$$

z	-0.09	-0.08	-0.07	-0.06	-0.05	-0.04	-0.03	-0.02	-0.01	-0.00
-(3.9+)	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.8	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
-3.7	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
-3.6	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0002	.0002
-3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
-3.4	.0002	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003
-3.3	.0003	.0004	.0004	.0004	.0004	.0004	.0004	.0005	.0005	.0005
-3.2	.0005	.0005	.0005	.0006	.0006	.0006	.0006	.0006	.0007	.0007
-3.1	.0007	.0007	.0008	.0008	.0008	.0008	.0009	.0009	.0009	.0010
-3.0	.0010	.0010	.0011	.0011	.0011	.0012	.0012	.0013	.0013	.0013
-2.9	.0014	.0014	.0015	.0015	.0016	.0016	.0017	.0018	.0018	.0019
-2.8	.0019	.0020	.0021	.0021	.0022	.0023	.0023	.0024	.0025	.0026
-2.7	.0026	.0027	.0028	.0029	.0030	.0031	.0032	.0033	.0034	.0035
-2.6	.0036	.0037	.0038	.0039	.0040	.0041	.0043	.0044	.0045	.0047
-2.5	.0048	.0049	.0051	.0052	.0054	.0055	.0057	.0059	.0060	.0062
-2.4	.0064	.0066	.0068	.0069	.0071	.0073	.0075	.0078	.0080	.0082
-2.3	.0084	.0087	.0089	.0091	.0094	.0096	.0099	.0102	.0104	.0107
-2.2	.0110	.0113	.0116	.0119	.0122	.0125	.0129	.0132	.0136	.0139
-2.1	.0143	.0146	.0150	.0154	.0158	.0162	.0166	.0170	.0174	.0179
-2.0	.0183	.0188	.0192	.0197	.0202	.0207	.0212	.0217	.0222	.0228
-1.9	.0233	.0239	.0244	.0250	.0256	.0262	.0268	.0274	.0281	.0287
-1.8	.0294	.0301	.0307	.0314	.0322	.0329	.0336	.0344	.0351	.0359
-1.7	.0367	.0375	.0384	.0392	.0401	.0409	.0418	.0427	.0436	.0446
-1.6	.0455	.0465	.0475	.0485	.0495	.0505	.0516	.0526	.0537	.0548
-1.5	.0559	.0571	.0582	.0594	.0606	.0618	.0630	.0643	.0655	.0668
-1.4	.0681	.0694	.0708	.0721	.0735	.0749	.0764	.0778	.0793	.0808
-1.3	.0823	.0838	.0853	.0869	.0885	.0901	.0918	.0934	.0951	.0968
-1.2	.0985	.1003	.1020	.1038	.1056	.1075	.1093	.1112	.1131	.1151
-1.1	.1170	.1190	.1210	.1230	.1251	.1271	.1292	.1314	.1335	.1357
-1.0	.1379	.1401	.1423	.1446	.1469	.1492	.1515	.1539	.1562	.1587
-0.9	.1611	.1635	.1660	.1685	.1711	.1736	.1762	.1788	.1814	.1841
-0.8	.1867	.1894	.1922	.1949	.1977	.2005	.2033	.2061	.2090	.2119
-0.7	.2148	.2177	.2206	.2236	.2266	.2296	.2327	.2358	.2389	.2420
-0.6	.2451	.2483	.2514	.2546	.2578	.2611	.2643	.2676	.2709	.2743
-0.5	.2776	.2810	.2843	.2877	.2912	.2946	.2981	.3015	.3050	.3085
-0.4	.3121	.3156	.3192	.3228	.3264	.3300	.3336	.3372	.3409	.3446
-0.3	.3483	.3520	.3557	.3594	.3632	.3669	.3707	.3745	.3783	.3821
-0.2	.3859	.3897	.3936	.3974	.4013	.4052	.4090	.4129	.4168	.4207
-0.1	.4247	.4286	.4325	.4364	.4404	.4443	.4483	.4522	.4562	.4602
-0.0	.4641	.4681	.4721	.4761	.4801	.4840	.4880	.4920	.4960	.5000

Table A4, continued. Standard Normal distribution

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998
3.5	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998
3.6	.9998	.9998	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
3.7	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
3.8	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
3.9+	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table A5. Table of F-distribution for $\alpha = 0.01$

$F_{0.01}$; critical values, such that $\boldsymbol{P}\{F > F_{0.01}\} = 0.01$

ν_2, denominator degrees of freedom	ν_1, numerator degrees of freedom												
	1	2	3	4	5	6	7	8	9	10	15	20	30
1	4052	4999	5404	5624	5764	5859	5928	5981	6022	6056	6157	6209	6260
2	98.5	99.0	99.2	99.3	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.5
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	26.9	26.7	26.5
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.2	14.0	13.8
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.72	9.55	9.38
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.56	7.40	7.23
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.31	6.16	5.99
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.52	5.36	5.20
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	4.96	4.81	4.65
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.56	4.41	4.25
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.25	4.10	3.94
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.01	3.86	3.70
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.82	3.66	3.51
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.66	3.51	3.35
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.52	3.37	3.21
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.41	3.26	3.10
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.31	3.16	3.00
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.23	3.08	2.92
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.15	3.00	2.84
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.09	2.94	2.78
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	2.98	2.83	2.67
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	2.89	2.74	2.58
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.81	2.66	2.50
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.75	2.60	2.44
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.70	2.55	2.39
35	7.42	5.27	4.40	3.91	3.59	3.37	3.20	3.07	2.96	2.88	2.60	2.44	2.28
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.52	2.37	2.20
45	7.23	5.11	4.25	3.77	3.45	3.23	3.07	2.94	2.83	2.74	2.46	2.31	2.14
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70	2.42	2.27	2.10
55	7.12	5.01	4.16	3.68	3.37	3.15	2.98	2.85	2.75	2.66	2.38	2.23	2.06
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.35	2.20	2.03
70	7.01	4.92	4.07	3.60	3.29	3.07	2.91	2.78	2.67	2.59	2.31	2.15	1.98
80	6.96	4.88	4.04	3.56	3.26	3.04	2.87	2.74	2.64	2.55	2.27	2.12	1.94
90	6.93	4.85	4.01	3.53	3.23	3.01	2.84	2.72	2.61	2.52	2.24	2.09	1.92
100	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50	2.22	2.07	1.89
150	6.81	4.75	3.91	3.45	3.14	2.92	2.76	2.63	2.53	2.44	2.16	2.00	1.83
200	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50	2.41	2.13	1.97	1.79
300	6.72	4.68	3.85	3.38	3.08	2.86	2.70	2.57	2.47	2.38	2.10	1.94	1.76
400	6.70	4.66	3.83	3.37	3.06	2.85	2.68	2.56	2.45	2.37	2.08	1.92	1.75
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.04	1.88	1.70

Table A5, continued. Table of F-distribution for $\alpha = 0.05$

$F_{0.05}$; critical values, such that $\boldsymbol{P}\{F > F_{0.05}\} = 0.05$

	ν_1, numerator degrees of freedom												
ν_2, denominator degrees of freedom	1	2	3	4	5	6	7	8	9	10	15	20	30
1	161	199	216	225	230	234	237	239	241	242	246	248	250
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.70	8.66	8.62
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.86	5.80	5.75
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.62	4.56	4.50
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.94	3.87	3.81
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.51	3.44	3.38
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.22	3.15	3.08
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.01	2.94	2.86
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.85	2.77	2.70
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.72	2.65	2.57
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.62	2.54	2.47
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.53	2.46	2.38
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.46	2.39	2.31
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.40	2.33	2.25
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.35	2.28	2.19
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.31	2.23	2.15
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.27	2.19	2.11
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.23	2.16	2.07
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.20	2.12	2.04
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.15	2.07	1.98
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.11	2.03	1.94
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.07	1.99	1.90
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.04	1.96	1.87
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.01	1.93	1.84
35	4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.16	2.11	1.96	1.88	1.79
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	1.92	1.84	1.74
45	4.06	3.20	2.81	2.58	2.42	2.31	2.22	2.15	2.10	2.05	1.89	1.81	1.71
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.87	1.78	1.69
55	4.02	3.16	2.77	2.54	2.38	2.27	2.18	2.11	2.06	2.01	1.85	1.76	1.67
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.84	1.75	1.65
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02	1.97	1.81	1.72	1.62
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95	1.79	1.70	1.60
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94	1.78	1.69	1.59
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.77	1.68	1.57
150	3.90	3.06	2.66	2.43	2.27	2.16	2.07	2.00	1.94	1.89	1.73	1.64	1.54
200	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88	1.72	1.62	1.52
300	3.87	3.03	2.63	2.40	2.24	2.13	2.04	1.97	1.91	1.86	1.70	1.61	1.50
400	3.86	3.02	2.63	2.39	2.24	2.12	2.03	1.96	1.90	1.85	1.69	1.60	1.49
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.67	1.57	1.46

Table A6. Table of Student's T-distribution

t_α; critical values, such that $\boldsymbol{P}\{t > t_\alpha\} = \alpha$

ν	α, the right-tail probability									
(d.f.)	.10	.05	.025	.02	.01	.005	.0025	.001	.0005	.0001
1	3.078	6.314	12.706	15.89	31.82	63.66	127.3	318.3	636.6	3185
2	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60	70.71
3	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92	22.20
4	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610	13.04
5	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.894	6.869	9.676
6	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959	8.023
7	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408	7.064
8	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041	6.442
9	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781	6.009
10	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587	5.694
11	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437	5.453
12	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318	5.263
13	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221	5.111
14	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140	4.985
15	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073	4.880
16	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015	4.790
17	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965	4.715
18	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.610	3.922	4.648
19	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883	4.590
20	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850	4.539
21	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819	4.492
22	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792	4.452
23	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768	4.416
24	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745	4.382
25	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725	4.352
26	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707	4.324
27	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.689	4.299
28	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674	4.276
29	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.660	4.254
30	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646	4.234
32	1.309	1.694	2.037	2.141	2.449	2.738	3.015	3.365	3.622	4.198
34	1.307	1.691	2.032	2.136	2.441	2.728	3.002	3.348	3.601	4.168
36	1.306	1.688	2.028	2.131	2.434	2.719	2.990	3.333	3.582	4.140
38	1.304	1.686	2.024	2.127	2.429	2.712	2.980	3.319	3.566	4.115
40	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551	4.094
45	1.301	1.679	2.014	2.115	2.412	2.690	2.952	3.281	3.520	4.049
50	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496	4.014
55	1.297	1.673	2.004	2.104	2.396	2.668	2.925	3.245	3.476	3.985
60	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460	3.962
70	1.294	1.667	1.994	2.093	2.381	2.648	2.899	3.211	3.435	3.926
80	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416	3.899
90	1.291	1.662	1.987	2.084	2.368	2.632	2.878	3.183	3.402	3.878
100	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390	3.861
200	1.286	1.653	1.972	2.067	2.345	2.601	2.838	3.131	3.340	3.789
∞	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.090	3.290	3.719

11.3 Calculus review

This section is a very brief summary of Calculus skills required for reading this book.

11.3.1 Inverse function

Function g is the **inverse function** of function f if

$$g(f(x)) = x \text{ and } f(g(y)) = y$$

for all x and y where $f(x)$ and $g(y)$ exist.

NOTATION | Inverse function $g = f^{-1}$ |

(Don't confuse the inverse $f^{-1}(x)$ with $1/f(x)$. These are different functions!)

To find the inverse function, solve the equation

$$f(x) = y.$$

The solution $g(y)$ is the inverse of f.

For example, to find the inverse of $f(x) = 3 + 1/x$, we solve the equation

$$3 + 1/x = y \quad \Rightarrow \quad 1/x = y - 3 \quad \Rightarrow \quad x = \frac{1}{y-3}.$$

The inverse function of f is $g(y) = 1/(y - 3)$.

11.3.2 Limits and continuity

A function $f(x)$ has a **limit** L at a point x_0 if $f(x)$ approaches L when x approaches x_0. To say more rigorously, for any ε there exists such δ that $f(x)$ is ε-close to L when x is δ-close to x_0. That is,

$$\text{if } |x - x_0| < \delta \text{ then } |f(x) - L| < \varepsilon.$$

A function $f(x)$ has a **limit** L at $+\infty$ if $f(x)$ approaches L when x goes to $+\infty$. Rigorously, for any ε there exists such N that $f(x)$ is ε-close to L when x gets beyond N, i.e.,

$$\text{if } x > N \text{ then } |f(x) - L| < \varepsilon.$$

Similarly, $f(x)$ has a **limit** L at $-\infty$ if for any ε there exists such N that $f(x)$ is ε-close to L when x gets below $(-N)$, i.e.,

$$\text{if } x < -N \text{ then } |f(x) - L| < \varepsilon.$$

NOTATION
$$\left|\begin{array}{lll} \lim\limits_{x\to x_0} f(x) = L, & \text{or} & f(x) \to L \text{ as } x \to x_0 \\ \lim\limits_{x\to \infty} f(x) = L, & \text{or} & f(x) \to L \text{ as } x \to \infty \\ \lim\limits_{x\to -\infty} f(x) = L, & \text{or} & f(x) \to L \text{ as } x \to -\infty \end{array}\right|$$

Function f is **continuous** at a point x_0 if

$$\lim_{x\to x_0} f(x) = f(x_0).$$

Function f is **continuous** if it is continuous at every point.

11.3.3 Sequences and series

A **sequence** is a function of a positive integer argument,

$$f(n) \text{ where } n = 1, 2, 3, \ldots.$$

Sequence $f(n)$ **converges** to L if

$$\lim_{n\to\infty} f(n) = L$$

and **diverges** to infinity if for any M there exists N such that

$$f(n) > M \text{ when } n > N.$$

A **series** is a sequence of partial sums,

$$f(n) = \sum_{k=1}^{n} a_k = a_1 + a_2 + \ldots + a_n.$$

Geometric series is defined by

$$a_n = Cr^n,$$

where r is called the **ratio** of the series. In general,

$$\sum_{k=m}^{n} Cr^n = f(n) - f(m-1) = C\frac{r^m - r^{n+1}}{1-r}.$$

For $m = 0$, we get

$$\sum_{k=0}^{n} Cr^n = C\frac{1 - r^{n+1}}{1-r}.$$

A geometric series converges if and only if $|r| < 1$. In this case,

$$\lim_{n\to\infty} \sum_{k=m}^{\infty} Cr^n = \frac{Cr^m}{1-r} \quad \text{and} \quad \sum_{k=0}^{\infty} Cr^n = \frac{C}{1-r}.$$

A geometric series diverges to ∞ if $r \geq 1$.

11.3.4 Derivatives, minimum, and maximum

Derivative of a function f at a point x is the limit

$$f'(x) = \lim_{y \to x} \frac{f(y) - f(x)}{y - x}$$

provided that this limit exists. Taking derivative is called **differentiation**. A function that has derivatives is called **differentiable**.

NOTATION $\quad f'(x) \quad \text{or} \quad \dfrac{d}{dx} f(x)$

Differentiating a function of several variables, we take **partial derivatives** denoted as

$$\frac{\partial}{\partial x_1} f(x_1, x_2, \ldots).$$

The most important derivatives are:

Derivatives

$$\boxed{\begin{array}{lcl}
(x^m)' & = & m\, x^{m-1} \\
(e^x)' & = & e^x \\
(\ln x)' & = & 1/x \\
C' & = & 0 \\
(f+g)'(x) & = & f'(x) + g'(x) \\
(Cf)'(x) & = & Cf'(x) \\
(f(x)g(x))' & = & f'(x)g(x) + f(x)g'(x) \\
\left(\dfrac{f(x)}{g(x)}\right)' & = & \dfrac{f'(x)g(x) - f(x)g'(x)}{g^2(x)} \\
\multicolumn{3}{l}{\text{for any functions } f,\ g \text{ and any number } C}
\end{array}}$$

To differentiate a composite function

$$f(x) = g(h(x)),$$

we use a chain rule,

Chain rule

$$\boxed{\frac{d}{dx} g(h(x)) = g'(h(x))h'(x)}$$

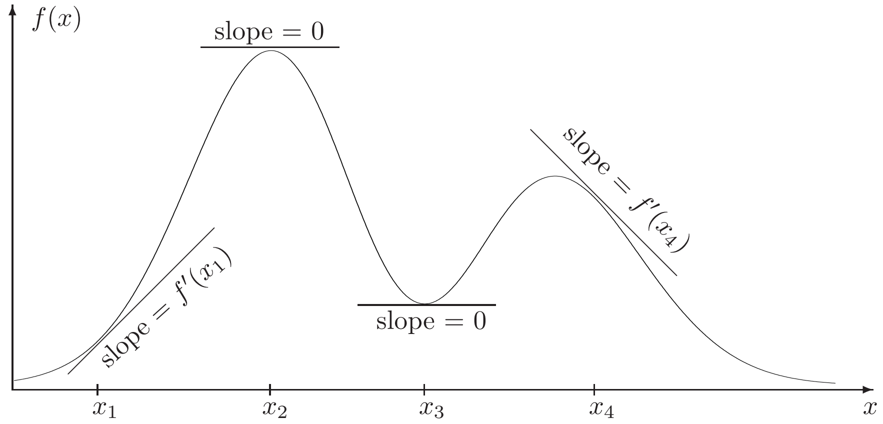


Figure 11.1 *Derivative is the slope of a tangent line.*

For example,

$$\frac{d}{dx}\ln^3(x) = 3\ln^2(x)\,\frac{1}{x}.$$

Geometrically, derivative $f'(x)$ equals the slope of a tangent line at point x, see Figure 11.1.

Computing maxima and minima

At the points where a differentiable function reaches its minimum or maximum, the tangent line is always flat, see points x_2 and x_3 on Figure 11.1. The slope of a horizontal line is 0, and thus,

$$f'(x) = 0$$

at these points.

To find out where a function is maximized or minimized, we consider

– solutions of the equation $f'(x) = 0$,

– points x where $f'(x)$ fails to exist,

– endpoints.

The highest and the lowest values of the function can only be attained at these points.

11.3.5 Integrals

Integration is an action opposite to differentiation.

A function $F(x)$ is an **antiderivative** (**indefinite integral**) of a function $f(x)$ if

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

Indefinite integrals are defined up to a constant C because when we take derivatives, $C' = 0$.

<u>Notation</u> $F(x) = \int f(x)\,dx$

An **integral** (**definite integral**) of a function $f(x)$ from point a to point b is the difference of antiderivatives,

$$\int_a^b f(x)\,dx = F(b) - F(a).$$

Improper integrals

$$\int_a^{\infty} f(x)\,dx, \quad \int_{-\infty}^{b} f(x)\,dx, \quad \int_{-\infty}^{\infty} f(x)\,dx$$

are defined as limits. For example,

$$\int_a^{\infty} f(x)\,dx = \lim_{b\to\infty} \int_a^b f(x)\,dx.$$

The most important integrals are:

Indefinite integrals

$$\boxed{\begin{array}{lcl}
\int x^m\,dx & = & \dfrac{x^{m+1}}{m+1} \quad \text{for } m \neq -1 \\
\int x^{-1}\,dx & = & \ln(x) \\
\int e^x\,dx & = & e^x \\
\int (f(x) + g(x))dx & = & \int f(x)dx + \int g(x)dx \\
\int Cf(x)dx & = & C \int f(x)dx \\
\multicolumn{3}{c}{\text{for any functions } f,\, g \text{ and any number } C}
\end{array}}$$

For example, to evaluate a definite integral $\int_0^2 x^3 dx$, we find an antiderivative

$F(x) = x^4/4$ and compute $F(2) - F(0) = 4 - 0 = 4$. A standard way to write this solution is

$$\int_0^2 x^3 dx = \left. \frac{x^4}{4} \right|_{x=0}^{x=2} = \frac{2^4}{4} - \frac{0^4}{4} = 4.$$

Two important integration skills are *integration by substitution* and *integration by parts.*

Integration by substitution

An integral often simplifies when we can denote a part of the function as a new variable (y). The limits of integration a and b are then recomputed in terms of y, and dx is replaced by

$$dx = \frac{dy}{dy/dx} \quad \text{or} \quad dx = \frac{dx}{dy}\, dy,$$

whichever is easier to find. Notice that dx/dy is the derivative of the *inverse function* $x(y)$.

Integration by substitution

$$\boxed{\int f(x)\, dx \quad = \int f(x(y)) \frac{dx}{dy}\, dy}$$

For example,

$$\int_{-1}^{2} e^{3x} dx = \int_{-3}^{6} e^y \left(\frac{1}{3}\right) dy = \left. \frac{1}{3}\, e^y \right|_{y=-3}^{y=6} = \frac{e^6 - e^{-3}}{3} = 134.5.$$

Here we substituted $y = 3x$, recomputed the limits of integration, found the inverse function $x = y/3$ and its derivative $dx/dy = 1/3$.

In the next example, we substitute $y = x^2$. Derivative of this substitution is $dy/dx = 2x$:

$$\int_0^2 x\, e^{x^2} dx = \int_0^4 x\, e^y \frac{dy}{2x} = \frac{1}{2} \int_0^4 e^y dy = \left. \frac{1}{2} e^y \right|_{y=0}^{y=4} = \frac{e^4 - 1}{2} = 26.8.$$

Integration by parts

This technique often helps to integrate a *product* of two functions. One of the parts is integrated, the other is differentiated.

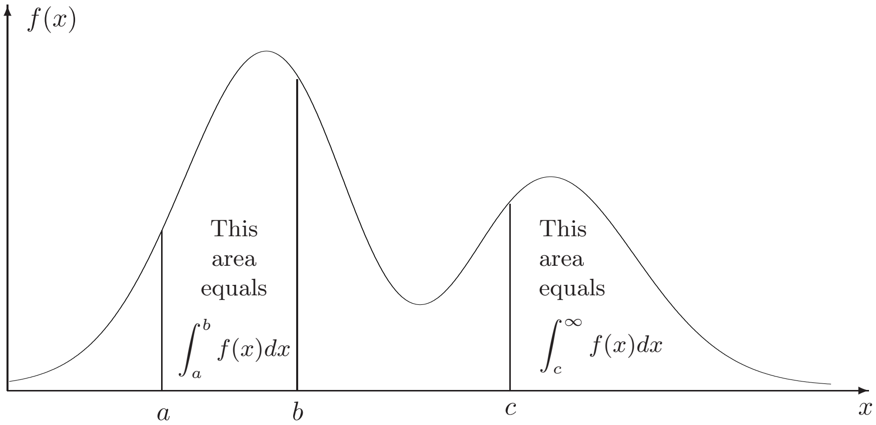


Figure 11.2 *Integrals are areas under the graph of $f(x)$.*

Integration by parts

$$\boxed{\int f'(x)g(x)dx = f(x)g(x) - \int f(x)g'(x)dx}$$

Applying this method is reasonable only when function (fg') is simpler than the initial function $(f'g)$.

In the following example, we let $f'(x) = e^x$ be one part and $g(x) = x$ be the other. Then $f'(x)$ is integrated, and its antiderivative $f(x) = e^x$. The other part $g(x)$ is differentiated, and $g'(x) = x' = 1$. The integral simplifies, and we can evaluate it,

$$\int x\,e^x dx = x\,e^x - \int (1)(e^x)dx = x\,e^x - e^x.$$

Computing areas

Area under the graph of a positive function $f(x)$ and above the interval $[a, b]$ equals the integral,

$$(\text{area from } a \text{ to } b) = \int_a^b f(x)dx.$$

Here a and b may be finite or infinite, see Figure 11.2.

Gamma function and factorial

Gamma function is defined as

$$\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx \qquad \text{for} \quad t > 0.$$

Taking this integral by parts, we obtain two important properties of a Gamma function,

$$\begin{array}{rcl} \Gamma(t+1) & = & t\Gamma(t) \quad \text{for} \quad \text{any } t > 0, \\ \Gamma(t+1) & = & t! = 1 \cdot 2 \cdot \ldots \cdot t \quad \text{for} \quad \text{integer } t. \end{array}$$

11.4 Matrices and linear systems

A **matrix** is a rectangular chart with numbers written in rows and columns,

$$A = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1c} \\ A_{21} & A_{22} & \cdots & A_{2c} \\ \cdots & \cdots & \cdots & \cdots \\ A_{r1} & A_{r2} & \cdots & A_{rc} \end{pmatrix}$$

where r is the number of rows and c is the number of columns. Every element of matrix A is denoted by A_{ij}, where $i \in [1, r]$ is the row number and $j \in [1, c]$ is the column number. It is referred to as an "$r \times c$ matrix."

Multiplying a row by a column

A row can only be multiplied by a column of the same length. The product of a row A and a column B is a number computed as

$$(A_1, \ldots, A_n) \begin{pmatrix} B_1 \\ \vdots \\ B_n \end{pmatrix} = \sum_{i=1}^{n} A_i B_i.$$

Example 11.1 (Measurement conversion). To convert, say, 3 hours 25 minutes 45 seconds into seconds, one may use formula

$$(3\ 25\ 45) \begin{pmatrix} 3600 \\ 60 \\ 1 \end{pmatrix} = 12345 \text{ (sec)}.$$

◇

Multiplying matrices

Matrix A may be multiplied by matrix B only if the number of columns in A equals the number of rows in B.

If A is a $k \times m$ matrix and B is an $m \times n$ matrix, then their product $AB = C$ is a $k \times n$ matrix. Each element of C is computed as

$$C_{ij} = \sum_{s=1}^{m} A_{is} B_{sj} = \left(\begin{array}{c} i^{\text{th}} \text{ row} \\ \text{of } A \end{array} \right) \left(\begin{array}{c} j^{\text{th}} \text{ column} \\ \text{of } B \end{array} \right).$$

Each element of AB is obtained as a product of the corresponding row of A and column of B.

Example 11.2. The following product of two matrices is computed as

$$\begin{pmatrix} 2 & 6 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} 9 & -3 \\ -3 & 1 \end{pmatrix} = \begin{pmatrix} (2)(9)+(6)(-3), & (2)(-3)+(6)(1) \\ (1)(9)+(3)(-3), & (1)(-3)+(3)(1) \end{pmatrix}$$
$$= \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

◇

In the last example, the result was a zero matrix "accidentally." This is not always the case. However, we can notice that matrices do not always obey the usual rules of arithmetics. In particular, a product of two non-zero matrices may equal a 0 matrix.

Also, in this regard, matrices *do not commute*, that is, $AB \neq BA$, in general.

Transposition

Transposition is reflecting the entire matrix about its main diagonal.

NOTATION | A^T = transposed matrix A |

Rows become columns, and columns become rows. That is,

$$A^T_{ij} = A_{ji}.$$

For example,

$$\begin{pmatrix} 1 & 2 & 3 \\ 7 & 8 & 9 \end{pmatrix}^T = \begin{pmatrix} 1 & 7 \\ 2 & 8 \\ 3 & 9 \end{pmatrix}.$$

The transposed product of matrices is

$$\boxed{(AB)^T = B^T A^T}$$

Solving systems of equations

In Chapters 6 and 7, we often solve systems of n linear equations with n unknowns and find a *steady-state distribution*. There are several ways of doing so.

One method to solve such a system is **by variable elimination**. Express one variable in terms of the others from one equation, then substitute it into the unused equations. You will get a system of $(n-1)$ equations with $(n-1)$ unknowns. Proceeding in the same way, we reduce the number of unknowns until we end up with 1 equation and 1 unknown. We find this unknown, then go back and find all the other unknowns.

Example 11.3 (Linear system). Solve the system

$$\left\{\begin{array}{rcrcrcr} 2x & + & 2y & + & 5z & = & 12 \\ & & 3y & - & z & = & 0 \\ 4x & - & 7y & - & z & = & 2 \end{array}\right.$$

We don't have to start solving from the first equation. Start with the one that seems simple. From the second equation, we see that

$$z = 3y.$$

Substituting $(3y)$ for z in the other equations, we obtain

$$\left\{\begin{array}{rcrcr} 2x & + & 17y & = & 12 \\ 4x & - & 10y & = & 2 \end{array}\right.$$

We are down by one equation and one unknown. Next, express x from the first equation,

$$x = \frac{12 - 17y}{2} = 6 - 8.5y$$

and substitute into the last equation,

$$4(6 - 8.5y) - 10y = 2.$$

Simplifying, we get $44y = 22$, hence $y = 0.5$. Now, go back and recover the other variables,

$$x = 6 - 8.5y = 6 - (8.5)(0.5) = 1.75; \qquad z = 3y = 1.5.$$

The answer is $\underline{x = 1.75,\ y = 0.5,\ z = 1.5}$.

We can check the answer by substituting the result into the initial system,

$$\left\{\begin{array}{rcrcrcr} 2(1.75) & + & 2(0.5) & + & 5(1.5) & = & 12 \\ & & 3(0.5) & - & 1.5 & = & 0 \\ 4(1.75) & - & 7(0.5) & - & 1.5 & = & 2 \end{array}\right.$$

◇

We can also eliminate variables by multiplying entire equations by suitable coefficients, adding and subtracting them.

Example 11.4 (Another method). Here is a shorter solution of Example 11.3. Double the first equation,

$$4x + 4y + 10z = 24,$$

and subtract the third equation from it,

$$11y + 11z = 22, \quad \text{or} \quad y + z = 2.$$

This way, we eliminated x. Then, adding $(y + z = 2)$ and $(3y - z = 0)$, we get $4y = 2$, and again, $y = 0.5$. Other variables, x and z, can now be obtained from y, as in Example 11.3. ◇

The system of equations in this example can be written in a matrix form as

$$\begin{pmatrix} x & y & z \end{pmatrix} \begin{pmatrix} 2 & 0 & 4 \\ 2 & 3 & -7 \\ 5 & -1 & -1 \end{pmatrix} = \begin{pmatrix} 12 & 0 & 2 \end{pmatrix},$$

or, equivalently,

$$\begin{pmatrix} 2 & 2 & 5 \\ 0 & 3 & -1 \\ 4 & -7 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 12 & 0 & 2 \end{pmatrix}.$$

Inverse matrix

Matrix B is the **inverse matrix** of A if

$$AB = BA = I = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix},$$

where I is the *identity matrix*. It has 1s on the diagonal and 0s elsewhere. Matrices A and B have to have the same number of rows and columns.

Notation | A^{-1} = inverse of matrix A |

Inverse of a product can be computed as

$$\boxed{(AB)^{-1} = B^{-1}A^{-1}}$$

To find the inverse matrix A^{-1} by hand, write matrices A and I next to each other. Multiplying rows of A by constant coefficients, adding and interchanging them, convert matrix A to the identity matrix I. The same operations convert matrix I to A^{-1},

$$\left(\begin{array}{c|c} A & I \end{array}\right) \longrightarrow \left(\begin{array}{c|c} I & A^{-1} \end{array}\right).$$

Example 11.5. Linear system in Example 11.3 is given by matrix

$$A = \begin{pmatrix} 2 & 2 & 5 \\ 0 & 3 & -1 \\ 4 & -7 & -1 \end{pmatrix}.$$

Repeating the row operations from this example, we can find the inverse matrix A^{-1},

$$\left(\begin{array}{ccc|ccc} 2 & 2 & 5 & 1 & 0 & 0 \\ 0 & 3 & -1 & 0 & 1 & 0 \\ 4 & -7 & -1 & 0 & 0 & 1 \end{array}\right) \longrightarrow \left(\begin{array}{ccc|ccc} 4 & 4 & 10 & 2 & 0 & 0 \\ 0 & 3 & -1 & 0 & 1 & 0 \\ 4 & -7 & -1 & 0 & 0 & 1 \end{array}\right)$$

$$\longrightarrow \left(\begin{array}{ccc|ccc} 0 & 11 & 11 & 2 & 0 & -1 \\ 0 & 3 & -1 & 0 & 1 & 0 \\ 4 & -7 & -1 & 0 & 0 & 1 \end{array}\right) \longrightarrow \left(\begin{array}{ccc|ccc} 0 & 1 & 1 & 2/11 & 0 & -1/11 \\ 0 & 3 & -1 & 0 & 1 & 0 \\ 4 & -7 & -1 & 0 & 0 & 1 \end{array}\right)$$

$$\longrightarrow \left(\begin{array}{ccc|ccc} 0 & 4 & 0 & 2/11 & 1 & -1/11 \\ 0 & 3 & -1 & 0 & 1 & 0 \\ 4 & -7 & -1 & 0 & 0 & 1 \end{array}\right) \longrightarrow \left(\begin{array}{ccc|ccc} 0 & 1 & 0 & 1/22 & 1/4 & -1/44 \\ 0 & 3 & -1 & 0 & 1 & 0 \\ 4 & -10 & 0 & 0 & -1 & 1 \end{array}\right)$$

$$\longrightarrow \left(\begin{array}{ccc|ccc} 0 & 1 & 0 & 1/22 & 1/4 & -1/44 \\ 0 & 0 & -1 & -3/22 & 1/4 & 3/44 \\ 4 & 0 & 0 & 10/22 & 3/2 & 34/44 \end{array}\right) \longrightarrow \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 5/44 & 3/8 & 17/88 \\ 0 & 1 & 0 & 1/22 & 1/4 & -1/44 \\ 0 & 0 & 1 & 3/22 & -1/4 & -3/44 \end{array}\right)$$

The inverse matrix is found,

$$A^{-1} = \begin{pmatrix} 5/44 & 3/8 & 17/88 \\ 1/22 & 1/4 & -1/44 \\ 3/22 & -1/4 & -3/44 \end{pmatrix}.$$

You can verify the result by multiplying $A^{-1}A$ or AA^{-1}. ◇

For a 2×2 matrix, the formula for the inverse is

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

Matrix operations in Matlab

The following Matlab commands perform basic matrix operations. Certainly, the sizes of participating matrices have to match, otherwise the program will return an error.

```
A = [1 3 5; 8 3 0; 0 -3 -1];   %  Entering a matrix
B  =  [  3 9 8
         0 0 2          %  Another way to define a matrix
         9 2 1 ];
A+B                     %  Addition
A*B                     %  Matrix multiplication
C=A.*B                  %  Multiplying element by element,
                        %  $C_{ij} = A_{ij}B_{ij}$
A^n                     %  Power of a matrix, $A^n = \underbrace{A \cdot \ldots \cdot A}_{n \text{ times}}$
A'                      %  transposed matrix
inv(A)   }
A^(-1)   }              %  inverse matrix
eye(n)                  %  $n \times n$ identity matrix
zeros(m,n)              %  $m \times n$ matrix of 0s
ones(m,n)               %  $m \times n$ matrix of 1s
rand(m,n)               %  matrix of Uniform(0,1) random numbers
randn(m,n)              %  matrix of Normal(0,1) random numbers
```

11.5 Answers to selected exercises

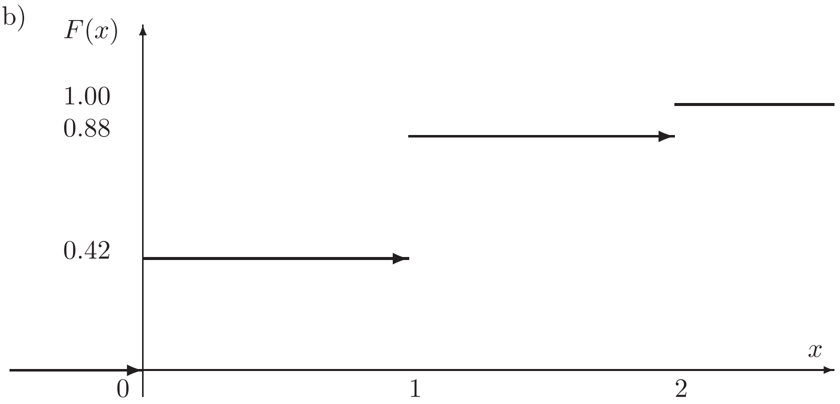
Chapter 2

2.1. 1/15. **2.3**. 0.45. **2.5**. 0.72. **2.7**. 0.66. **2.8**. 0.9508. **2.9**. 0.1792. **2.12**. 0.9744. **2.13**. 0.992. **2.15**. 0.1364. **2.16**. (a) 0.049. (b) 0.510. **2.18**. 0.0847. **2.20**. 0.00534. **2.21**. 0.8854. **2.24**. (a) 5/21 or 0.238. (b) 10/41 or 0.244.

Chapter 3

3.1. (a) P(0)=0.42, P(1)=0.46, P(2)=0.12.

(b)



3.2. $\mathbf{E}(Y) = 200$ dollars, $\text{Var}(Y) = 110,000$ squared dollars. **3.3**. $\mathbf{E}(X) = 0.6$, $\text{Var}(X) = 0.24$. **3.4**. $\mathbf{E}(X) = 3.5$, $\text{Var}(X) = 2.9167$. **3.7**. $\mathbf{E}(Y) = 1.6$, $\text{Var}(Y) = 1.12$. **3.9**. This probability does not exceed 1/16. **3.10**. 0.28. **3.11**. (a) The joint pmf is

		y						
$P_{(X,Y)}(x,y)$		1	2	3	4	5	6	$P_X(x)$
x	1	1/36	1/18	1/18	1/18	1/18	1/18	11/36
	2	0	1/36	1/18	1/18	1/18	1/18	9/36
	3	0	0	1/36	1/18	1/18	1/18	7/36
	4	0	0	0	1/36	1/18	1/18	5/36
	5	0	0	0	0	1/36	1/18	3/36
	6	0	0	0	0	0	1/36	1/36
	$P_Y(y)$	1/36	3/36	5/36	7/36	9/36	11/36	1

(b) They are dependent. (c) $P_X(1) = 11/36$, $P_X(2) = 9/36$, $P_X(3) = 7/36$, $P_X(4) = 5/36$, $P_X(5) = 3/36$, $P_X(6) = 1/36$. (d) $\boldsymbol{P}\{Y = 5 \mid X = 2\} = 2/9$. **3.12**. (a) Dependent. (b) Dependent. **3.15**. (a) 0.48. (b) Dependent. **3.17**. Third portfolio. **3.18**. (a) **E**(Profit)=6, Var(Profit)=684. (b) **E**(Profit)=6, Var(Profit)=387. (c) **E**(Profit)=6, Var(Profit)=864. The least risky portfolio is (b); the most risky portfolio is (c). **3.20**. (a) 0.0596. (b) 0.9860. **3.21**. 0.2447. **3.22**. 0.0070. **3.23**. (a) 0.0055. (b) 0.00314. **3.24**. (a) 0.0328, (b) 0.4096. **3.26**. (a) 0.3704. (b) 0.0579. **3.27**. (a) 0.945. (b) 0.061. **3.29**. 0.0923. **3.31**. (a) 0.968. (b) 0.018. **3.36**. 0.827.

Chapter 4

4.1. 3; $1-1/x^3$; 0.125. **4.3**. 4/3; 31/48 or 0.6458. **4.5**. (a) 0.2. (b) 0.75. (c) $3\frac{1}{3}$ or 3.3333. **4.7**. 0.875. **4.8**. 0.264. **4.10**. 0.4764. **4.14**. (a) 4 and 0.2. (b) 0.353. **4.15**. (a) 0.062. (b) 0.655. **4.17**. 0.2033. **4.21**. 0.1151. **4.22**. 0.567.

Chapter 5

5.1. $X = U^{2/3} = 0.01$. **5.3**. $X = 2\sqrt{U} - 1$. **5.5**. $X = (9U-1)^{1/3} = 1.8371$. **5.7**. (d) $n = 38,416$ is sufficient. **5.15**. $p = \frac{1}{(b-a)c}$.

Most exercises in this chapter use computer simulations and don't have a unique answer.

Chapter 6

6.1. (a)

$$P = \begin{pmatrix} 1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/2 \\ 0 & 1/2 & 1/2 \end{pmatrix}$$

(b) Irregular. (c) 0.125. **6.2**. (a)

$$P^{(2)} = \begin{pmatrix} 0.52 & 0.48 \\ 0.48 & 0.52 \end{pmatrix}.$$

(b) 0.496. **6.3**. (a)

$$P = \begin{pmatrix} 0.6 & 0.4 \\ 0.2 & 0.8 \end{pmatrix}$$

(b) 0.28. **6.4**. (a)

$$P = \begin{pmatrix} 0.6 & 0.4 \\ 0.3 & 0.7 \end{pmatrix}$$

(b) 0.52. (c) 4/7 or 0.5714. **6.5**. 0.333. **6.9**. (a) 0.0579. (b) practically 0. **6.11**. (a) 0.05 min or 3 sec. (b) **E**(X) =120 jobs; Std$(X) = \sqrt{108}$ or 10.39 jobs. **6.13**. **E**(T) = 12 sec; Var(T) = 120 sec^2. **6.15**. (a) 0.75 seconds. (b) **E**(T) = 5 seconds; Std(T) = 4.6 seconds. **6.18**. 0.0162. **6.19**. 0.8843. **6.21**. (a) 0.735. (b) 7500 dollars; 3354 dollars 10 cents. **6.22**. (a) 0.0028. (b) **E**$(W_3) = 3/5$, Var$(W_3) = 3/25$.

Chapter 7

7.1.

$$P = \begin{pmatrix} 35/36 & 1/36 & 0 & 0 & \dots \\ 35/432 & 386/432 & 11/432 & 0 & \dots \\ 0 & 35/432 & 386/432 & 11/432 & \dots \\ 0 & 0 & 35/432 & 386/432 & \dots \\ \dots & \dots & \dots & \dots & \dots \end{pmatrix}$$

7.2. $\pi_0 = 0.5070$, $\pi_1 = 0.4225$, $\pi_2 = 0.0704$. **7.3**. 2/75 or 0.02667. **7.5**. $\pi_0 = 49/334 = 0.1467$, $\pi_1 = 105/334 = 0.3144$, $\pi_2 = 180/334 = 0.5389$. **7.7**. (a) 1/16 min or 3.75 sec. (b) There is no indication that the arrival rate is on the increase. **7.9**. (a) 2/27 or 0.074. (b) 6 sec. **7.10**. (a) The job is expected to be printed at 12:03:20. (b) 0.064 or 6.4% of the time. **7.11**. (a) 4 min. (b) 1/3. **7.12**. (a) 4 and 3.2. (b) 0.8. (c) 0.262. **7.13**. (a) 7.5 min. (b) 0.64. (c) 0.6. **7.15**. (a) 3/7 or 0.4286. (b) 0.7 or 70% of time. (c) 30/7 or 4.2857 min. **7.17**. (a) $\pi_0 = 29/89$, $\pi_1 = 60/89$. (b) 0.0026.

Chapter 8

8.1. (b) The five-point summaries are (37, 43, 50, 56, 60) and (21, 35, 39, 46, 53). **8.2**. (a) 17.95, 9.97, 3.16. (b) (11.9, 15.8, 17.55, 19.9, 24.1). (c) IQR=4.1. No outliers. **8.4**. 0.007. **8.6**. (a) 13.21 mln, 12.8 mln, 82.51 mln^2. **8.7**. (a) 0.2298, 0.21, 0.0099. (c) There is a rather strong negative correlation, $r = -0.7683$. **8.8**. (a) left-skewed, symmetric, right-skewed. (b) Set 1: 14.97, 15.5. Set 2: 20.83, 21.0. Set 3: 41.3, 39.5.

Chapter 9

9.1. (a) 0.625. (b) 0.625. **9.4**. (a) 2. (b) 2.1766. **9.8**. (a) [36.1866, 39.2134]. (b) Yes, there is a sufficient evidence. **9.10**. (a) 50 ± 33.7 or [16.3; 83.7]. (b) At this level of significance, the data does not provide a significant evidence against the hypothesis that the average salary of all entry-level computer engineers equals $80,000. **9.11**. (a) [0.073; 0.167]. (b) No. No. **9.12**. There is no significant evidence. P-value is 0.1515. **9.18**. (a) [4.5, 15.1]. (b) Each test has a p-value between 0.0005 and 0.001. There is a significant reduction.

Chapter 10

10.2. (a) $y = -9.71 + 0.56x$. (b) $SS_{\text{REG}} = 149.85$ with 1 d.f., $SS_{\text{ERR}} = 57.35$ with 73 d.f., $SS_{\text{TOT}} = 207.2$ with 74 d.f.; 72.32% is explained. (c) [0.455, 0.670]. The slope is significant with a p-value less than 0.0002. **10.4**. (a) 7.39. (b) There is a significant evidence at the 0.01 level that the investment increases by more than $ 1,800 every year, on the average (i.e., $\beta > 1.8$). (c) $[50.48, 67.10]$. **10.5**. (a) $b_0 = 32.21, b_1 = 2.36, b_2 = 4.09$. (b) 52.8. (c) Reduces by 4093 dollars. (d) $SS_{\text{TOT}} = 841.6$ with 10 d.f., $SS_{\text{REG}} = 808.0$ with 2 d.f., $SS_{\text{ERR}} = 33.6$ with 8 d.f. $R^2 = 0.96$. Significance of the model: $F = 96.14$, significant at the 0.01 level. (e) The new variable explains additional 3.9% of the total variation. It is significant at the 0.05 level, but not at the 0.01 level. **10.10**. (a) Population $= -27.9 + 1.29(\text{year} - 1800)$. (b) $SS_{\text{TOT}} = 159,893$ with 21 d.f., $SS_{\text{REG}} = 147,085$ with 1 d.f., $SS_{\text{ERR}} = 12,808$ with 20 d.f., $R^2 = 0.9199$. (c) 242.9 mln., 249.4 mln., 255.8 mln. **10.11**. (a) Population $= 5.622 + 0.020(\text{year} - 1800) + 0.0067(\text{year} - 1800)^2$. (b) 305.3 mln., 319.6 mln., 333.2 mln. (c) $SS_{\text{TOT}} = 159,893$ with 21 d.f., $SS_{\text{REG}} = 159,723$ with 2 d.f., $SS_{\text{ERR}} = 170$ with 19 d.f. $R^2 = 0.9989$. Quadratic term explains additional 7.9% of the total variation. (d) Adjusted R-square is 0.9159 for the linear model, 0.9988 for the full model, and 0.9989 for the reduced model with only a quadratic term. This last model is the best, according to the adjusted R-square criterion.

Index