An Introduction to Statistical Inference and Its Applications

Michael W. Trosset
Department of Mathematics
College of William & Mary

January 5, 2005

I of dice possess the science
and in numbers thus am skilled.

From *The Story of Nala*, the third book of the Indian epic *Mahábarata*.

This book is dedicated to

Richard A. Tapia,

my teacher, mentor, collaborator, and friend.

Contents

Chapter 1

Experiments

Statistical methods have proven enormously valuable in helping scientists interpret the results of their experiments—and in helping them design experiments that will produce interpretable results. In a quite general sense, the purpose of statistical analysis is to organize a data set in ways that reveal its structure. Sometimes this is so easy that one does not think that one is doing "statistics;" sometimes it is so difficult that one seeks the assistance of a professional statistician.

This is a book about how statisticians draw conclusions from experimental data. Its primary goal is to introduce the reader to an important type of reasoning that statisticians call "statistical inference." Rather than provide a superficial introduction to a wide variety of inferential methods, we will concentrate on fundamental concepts and study a few methods in depth.

Although statistics can be studied at many levels with varying degrees of sophistication, there is no escaping the simple fact that statistics is a mathematical discipline. Statistical inference rests on the mathematical foundation of probability. The better one desires to understand statistical inference, the more that one needs to know about probability. Accordingly, we will devote several chapters to probability before we begin our study of statistics. To motivate the reader to embark on this program of study, the present chapter describes the important role that probability plays in scientific investigation.

1.1 Examples

This section describes several scientific experiments. Each involves chance variation in a different way. The common theme is that chance variation

cannot be avoided in scientific experimentation.

1.1.1 Spinning a Penny

In August 1994, while attending the 15th International Symposium on Mathematical Programming in Ann Arbor, MI, I read an article in which the author asserted that spinning (as opposed to tossing/flipping) a typical penny is not fair, i.e., that `Heads` and `Tails` are not equally likely to result. Specifically, the author asserted that the chance of obtaining `Heads` by spinning a penny is about 30%.[1]

I was one of several people in a delegation from Rice University. That evening, we ended up at a local Subway restaurant for dinner and talk turned to whether or not spinning pennies is fair. Before long we were each spinning pennies and counting `Heads`. At first it seemed that about 70% of the spins were `Heads`, but this proved to be a temporary anomaly. By the time that we tired of our informal experiment, our results seemed to confirm the plausibility of the author's assertion.

I subsequently used penny-spinning as an example in introductory statistics courses, each time asserting that the chance of obtaining `Heads` by spinning a penny is about 30%. Students found this to be an interesting bit of trivia, but no one bothered to check it—until 2001. In the spring of 2001, three students at the College of William & Mary spun pennies, counted `Heads`, and obtained some intriguing results.

For example, Matt, James, and Sarah selected one penny that had been minted in the year 2000 and spun it 300 times, observing 145 `Heads`. This is very nearly 50% and the discrepancy might easily be explained by chance variation—perhaps spinning their penny is fair! They tried different pennies

[1] Years later, I have been unable to discover what I read or who wrote it. It seems to be widely believed that the chance is less than 50%. The most extreme assertion that I have discovered is by R. L. Graham, D. E. Knuth, and O. Patashnik (*Concrete Mathematics, Second Edition*, Addison-Wesley, 1994, page 401), who claimed that the chance is approximately 10% "when you spin a newly minted U.S. penny on a smooth table." A fairly comprehensive discussion of "Flipping, spinning, and tilting coins" can be found at

`http://www.dartmouth.edu/~chance/chance_news/recent_news/chance_news_11.02.html#item2`,

in which various individuals emphasize that the chance of `Heads` depends on such factors as the year in which the penny was minted, the surface on which the penny is spun, and the quality of the spin. For pennies minted in the 1960s, one individual reported 1878 `Heads` in 5520 spins, about 34%.

and obtained different percentages. Perhaps all pennies are not alike! (Pennies minted before 1982 are 95% copper and 5% zinc; pennies minted after 1982 are 97.5% zinc and 2.5% copper.) Or perhaps the differences were due to chance variation.

Were one to undertake a scientific study of penny spinning, there are many questions that one might ask. Here are several:

- Choose a penny. What is the chance of obtaining `Heads` by spinning that penny? (This question is the basis for Exercise 1 at the end of this chapter.)
- Choose two pennies. Are they equally likely to produce `Heads` when spun?
- Choose several pennies minted before 1982 and several pennies minted after 1982. As groups, are pre-1982 pennies and post-1982 pennies equally likely to produce `Heads` when spun?

1.1.2 The Speed of Light

According to Albert Einstein's special theory of relativity, the speed of light through a vacuum is a universal constant c. Since 1974, that speed has been given as $c = 299,792.458$ kilometers per second.[2] Long before Einstein, however, philosophers had debated whether or not light is transmitted instantaneously and, if not, at what speed it moved. In this section, we consider Albert Abraham Michelson's famous 1879 experiment to determine the speed of light.[3]

Aristotle believed that light "is not a movement" and therefore has no speed. Francis Bacon, Johannes Kepler, and René Descartes believed that light moved with infinite speed, whereas Galileo Galilei thought that its speed was finite. In 1638 Galileo proposed a terrestrial experiment to resolve the dispute, but two centuries would pass before this experiment became technologically practicable. Instead, early determinations of the speed of light were derived from astronomical data.

[2] Actually, a second is defined to be $9,192,631,770$ periods of radiation from cesium-133 and a kilometer is defined to be the distance travelled by light through a vacuum in 1/299792458 seconds!

[3] A. A. Michelson (1880). Experimental determination of the velocity of light made at the U.S. Naval Academy, Annapolis. *Astronomical Papers*, 1:109–145. The material in this section is taken from R. J. MacKay and R. W. Oldford (2000), Scientific method, statistical method and the speed of light, *Statistical Science*, 15:254–278.

The first empirical evidence that light is not transmitted instantaneously was presented by the Danish astronomer Ole Römer, who studied a series of eclipses of Io, Jupiter's largest moon. In September 1676, Römer correctly predicted a 10-minute discrepancy in the time of an impending eclipse. He argued that this discrepancy was due to the finite speed of light, which he estimated to be about $214,000$ kilometers per second. In 1729, James Bradley discovered an annual variation in stellar positions that could be explained by the earth's motion *if* the speed of light was finite. Bradley estimated that light from the sun took 8 minutes and 12 seconds to reach the earth and that the speed of light was 301,000 kilometers per second. In 1809, Jean-Baptiste Joseph Delambre used 150 years of data on eclipses of Jupiter's moons to estimate that light travels from sun to earth in 8 minutes and 13.2 seconds, at a speed of 300,267.64 kilometers per second.

In 1849, Hippolyte Fizeau became the first scientist to estimate the speed of light from a terrestrial experiment, a refinement of the one proposed by Galileo. An accurately machined toothed wheel was spun in front of a light source, automatically covering and uncovering it. The light emitted in the gaps between the teeth travelled 8633 meters to a fixed flat mirror, which reflected the light back to its source. The returning light struck either a tooth or a gap, depending on the wheel's speed of rotation. By varying the speed of rotation and observing the resulting image from reflected light beams, Fizeau was able to measure the speed of light.

In 1851, Leon Foucault further refined Galileo's experiment, replacing Fizeau's toothed wheel with a rotating mirror. Michelson further refined Foucault's experimental setup. A precise account of the experiment is beyond the scope of this book, but Mackay's and Oldford's account of how Michelson produced each of his 100 measurements of the speed of light provides some sense of what was involved. More importantly, their account reveals the multiple ways in which Michelson's measurements were subject to error.

1. The distance $|RM|$ from the rotating mirror to the fixed mirror was measured five times, each time allowing for temperature, and the average used as the "true distance" between the mirrors for all determinations.

2. The fire for the pump was started about a half hour before measurement began. After this time, there was sufficient pressure to begin the determinations.

3. The fixed mirror M was adjusted... and the heliostat placed and adjusted so that the Sun's image was directed at the slit.
4. The revolving mirror was adjusted on two different axes....
5. The distance $|SR|$ from the revolving mirror to the crosshair of the eyepiece was measured using the steel tape.
6. The vertical crosshair of the eyepiece of the micrometer was centred on the slit and its position recorded in terms of the position of the screw.
7. The electric tuning fork was started. The frequency of the fork was measured two or three times for each set of observations.
8. The temperature was recorded.
9. The revolving mirror was started. The eyepiece was set approximately to capture the displaced image. If the image did not appear in the eyepiece, the mirror was inclined forward or back until it came into sight.
10. The speed of rotation of the mirror was adjusted until the image of the revolving mirror came to rest.
11. The micrometer eyepiece was moved by turning the screw until its vertical crosshair was centred on the return image of the slit. The number of turns of the screw was recorded. The displacement is the difference in the two positions. To express this as the distance $|IS|$ in millimetres the measured number of turns was multiplied by the calibrated number of millimetres per turn of the screw.
12. Steps 10 and 11 were repeated until 10 measurements of the displacement $|IS|$ were made.
13. The rotating mirror was stopped, the temperature noted and the frequency of the electric fork was determined again.

Michelson used the procedure described above to obtain 100 measurements of the speed of light in air. Each measurement was computed using the average of the 10 measured displacements in Step 12. These measurements, reported in Table 1.1, subsequently were adjusted for temperature and corrected by a factor based on the refractive index of air. Michelson reported the speed of light in a vacuum as $299,944 \pm 51$ kilometers per second.

1.1.3 Termite Foraging Behavior

In the mid-1980s, Susan Jones was a USDA entomologist and a graduate student in the Department of Entomology at the University of Arizona. Her dissertation research concerned the foraging ecology of subterranean termites

50	−60	100	270	130	50	150	180	180	80
200	180	130	−150	−40	10	200	200	160	160
160	140	160	140	80	0	50	80	100	40
30	−10	10	80	80	30	0	−10	−40	0
80	80	80	60	−80	−80	−180	60	170	150
80	110	50	70	40	40	50	40	40	40
90	10	10	20	0	−30	−40	−60	−50	−40
110	120	90	60	80	−80	40	50	50	−20
90	40	−20	10	−40	10	−10	10	10	50
70	70	10	−60	10	140	150	0	10	70

Table 1.1: Michelson's 100 unadjusted measurements of the speed of light in air. Add 299,800 to obtain measurements in units of kilometers per second.

in the Sonoran Desert. Her field studies were conducted on the Santa Rita Experimental Range, about 40 kilometers south of Tucson, AZ:

> The foraging activity of *H. aureus*[4] was studied in 30 plots, each consisting of a grid (6 by 6 m) of 25 toilet-paper rolls which served as baits... Plots were selected on the basis of two criteria: the presence of *H. aureus* foragers in dead wood, and separation by at least 12 m from any other plot. A 6-by-6-m area was then marked off within the vicinity of infested wood, and toilet-paper rolls were aligned in five rows and five columns and spaced at 1.5-m intervals. The rolls were positioned on the soil surface and each was held in place with a heavy wire stake. All pieces of wood ca. 15 cm long and longer were removed from each plot and ca. 3 m around the periphery to minimize the availability of natural wood as an alternative food source. Before infested wood was removed from the site, termites were allowed to retreat into their galleries in the soil to avoid depleting the numbers of surface foragers. All plots were established within a 1-wk period during late June 1984.
>
> Plots were examined once a week during the first 5 wk after establishment, and then at least once monthly thereafter until August 1985.[5]

[4] *Heterotermes aureus* (Snyder) is the most common subterranean termite species in the Sonoran Desert. Haverty, Nutting, and LaFage (Density of colonies and spatial distribution of foraging territories of the desert subterranean termite, *Heterotermes aureues* (Snyder), *Environmental Entomology*, 4:105–109, 1975) estimated the population density of this species in the Santa Rita Experimental Range at 4.31×10^6 termites per hectare.

[5] Jones, Trosset, and Nutting. Biotic and abiotic influences on foraging of *Heteroter-*

An important objective of the above study was

> ... to investigate the relationship between food-source distance (on a scale 6 by 6 m) and foraging behavior. This was accomplished by analyzing the order in which different toilet-paper rolls in the same plot were attacked.... Specifically, a statistical methodology was developed to test the null hypothesis that any previously unattacked roll was equally likely to be the next roll attacked (random foraging). Alternative hypotheses supposed that the likelihood that a previously unattacked roll would be the next attacked roll decreased with increasing distance from previously attacked rolls (systematic foraging).[6]

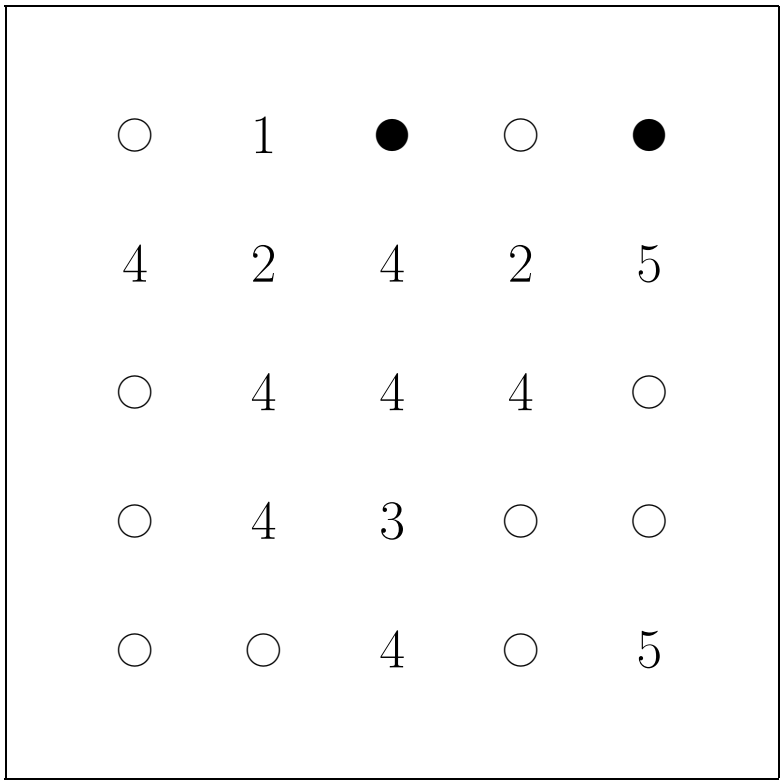

Figure 1.1: Order of *H. aureus* attack in Plot 20.

The order in which the toilet-paper rolls in Plot 20 were attacked is displayed in Figure 1.1. The unattacked rolls are denoted by ○, the initially

mes aureus (Snyder) (Isoptera: Rhinotermitidae), *Environmental Entomology*, 16:791–795, 1987.

[6]Ibid.

attacked rolls are denoted by •, and the subsequently attacked rolls are denoted (in order of attack) by 1, 2, 3, 4, and 5. Notice that these numbers do not specify a unique order of attack:

> ... because the plots were not observed continuously, a number of rolls seemed to have been attacked simultaneously. Therefore, it was not always possible to determine the exact order in which they were attacked. Accordingly, all permutations consistent with the observed ties in order were considered... [7]

In a subsequent chapter, we will return to the question of whether or not *H. aureus* forages randomly and describe the statistical methodology that was developed to answer it. Along the way, we will develop rigorous interpretations of the phrases that appear in the above passages, e.g., "permutations", "equally likely", "null hypothesis", "alternative hypotheses", etc.

1.2 Randomization

This section illustrates an important principle in the design of experiments. We begin by describing two famous studies that produced embarrassing results because they failed to respect this principle.

The Lanarkshire Milk Experiment A 1930 experiment in the schools of Lanarkshire attempted to ascertain the effect of milk supplements on Scottish children. For four months, 5000 children received a daily supplement of 3/4 pint of raw milk, 5000 children received a daily supplement of 3/4 pint of pasteurized milk, and 10,000 children received no daily milk supplement. Each child was weighed (while wearing indoor clothing) and measured for height before the study commenced (in February) and after it ended (in June). The final observations of the control group exceeded the final observations of the treatment groups by average amounts equivalent to 3 months growth in weight and 4 months growth in weight, thereby suggesting that the milk supplements actually retarded growth! What went wrong?

To explain the results of the Lanarkshire milk experiment, one must examine how the 20,000 children enrolled in the study were assigned to the study groups. An initial division into treatment versus control groups

[7] Ibid.

was made arbitrarily, e.g., using the alphabet. However, if the initial division appeared to produce groups with unbalanced numbers of well-fed or ill-nourished children, then teachers were allowed to swap children between the two groups in order to obtain (apparently) better balanced groups. It is thought that well-meaning teachers, concerned about the plight of ill-nourished children and "knowing" that milk supplements would be beneficial, consciously or subconsciously availed themselves of the opportunity to swap ill-nourished children into the treatment group. This resulted in a treatment group that was lighter and shorter than the control group. Furthermore, it is likely that differences in weight gains were confounded with a tendency for well-fed children to come from families that could afford warm (heavier) winter clothing, as opposed to a tendency for ill-nourished children to come from poor families that provided shabbier (lighter) clothing.[8]

The Pre-Election Polls of 1948 The 1948 presidential election pitted Harry Truman, the Democratic incumbent who had succeeded to the presidency when Franklin Roosevelt died in office, against Thomas Dewey, the Republican governor of New York.[9] Each of the three major polling organizations that covered the campaign predicted that Dewey would win: the Crossley poll predicted 50% of the popular vote for Dewey and 45% for Truman, the Gallup poll predicted 50% for Dewey and 44% for Truman, and the Roper poll predicted 53% for Dewey and 38% for Truman. Dewey was considered "as good as elected" until the votes were actually counted: in one of the great upsets in American politics, Truman received slightly less than 50% of the popular vote and Dewey received slightly more than 45%.[10] What went wrong?

Poll predictions are based on data collected from a sample of prospective

[8]For additional details and commentary, see Student (1931), The Lanarkshire milk experiment, *Biometrika*, 23:398, and Section 5.4 (Justification of Randomization) of Cox (1958), *Planning of Experiments*, John Wiley & Sons, New York.

[9]As a crusading district attorney in New York City, Dewey was a national hero in the 1930s. In late 1938, two Hollywood films attempted to capitalize on his popularity, RKO's *Smashing the Rackets* and Warner Brothers' *Racket Busters*. The special prosecutor in the latter film was played by Walter Abel, who bore a strong physical resemblance to Dewey.

[10]A famous photograph shows an exuberant Truman holding a copy of the *Chicago Tribune* with the headline `Dewey Defeats Truman`. On election night, Dewey confidently asked his wife, "How will it be to sleep with the president of the United States?" "A high honor, and quite frankly, darling, I'm looking forward to it," she replied. At breakfast next morning, having learned of Truman's upset victory, Frances playfully teased her husband: "Tell me, Tom, am I going to Washington or is Harry coming here?"

voters. For example, Gallup's prediction was based on 50,000 interviews. To assess the quality of Gallup's prediction, one must examine how his sample was selected. In 1948, all three polling organizations used a method called *quota sampling* that attempts to hand-pick a sample that is representative of the entire population. First, one attempts to identify several important characteristics that may be associated with different voting patterns, e.g., place of residence, sex, age, race, etc. Second, one attempts to obtain a sample that resembles the entire population with respect to those characteristics. For example, a Gallup interviewer in St. Louis was instructed to interview 13 subjects. Exactly 6 were to live in the suburbs, 7 in the city; exactly 7 were to be men, 6 women. Of the 7 men, exactly 3 were to be less than 40 years old and exactly 1 was to be black. Monthly rent categories for the 6 white men were specified. Et cetera, et cetera, et cetera.

Although the quotas used in quota sampling are reasonable, the method does not work especially well. The reason is that quota sampling does not specify how to choose the sample *within* the quotas—these choices are left to the discretion of the interviewer. Human choice is unpredictable and often subject to bias. In 1948, Republicans were more accessible than Democrats: they were more likely to have permanent addresses, own telephones, etc. Within their carefully prescribed quotas, Gallup interviewers were slightly more likely to find Republicans than Democrats. This unintentional bias toward Republicans had distorted previous polls; in 1948, the election was close enough that the polls picked the wrong candidate.[11]

In both the Lanarkshire milk experiment and the pre-election polls of 1948, subjective attempts to hand-pick representative samples resulted in embarrassing failures. Let us now exploit our knowledge of what *not* to do and design a simple experiment. An instructor—let's call him Ishmael—of one section of Math 106 (Elementary Statistics) has prepared two versions of a final exam. Ishmael hopes that the two versions are equivalent, but he recognizes that this will have to be determined experimentally. He therefore decides to divide his class of 40 students into two groups, each of which will receive a different version of the final. How should he proceed?

Ishmael recognizes that he requires two comparable groups if he hopes to draw conclusions about his two exams. For example, suppose that he

[11] For additional details and commentary, see Mosteller et al. (1949), *The Pre-Election Polls of 1948*, Social Science Research Council, New York, and Section 19.3 (The Year the Polls Elected Dewey) of Freedman, Pisani, and Purves (1998), *Statistics*, Third Edition, W. W. Norton & Company, New York.

administers one exam to the students who attained an `A` average on the midterms and the other exam to the other students. If the average score on exam `A` is 20 points higher than the average score on exam `B`, then what can he conclude? It might be that exam `A` is 20 points easier than exam `B`. Or it might be that the two exams are equally difficult, but that the `A` students are 20 points more capable than the `B` students. Or it might be that exam `A` is actually 10 points more difficult than exam `B`, but that the `A` students are 30 points more capable than the `B` students. There is no way to decide—exam version and student capability are *confounded* in this experiment.

The lesson of the Lanarkshire milk experiment and the pre-election polls of 1948 is that it is difficult to hand-pick representative samples. Accordingly, Ishmael decides to randomly assign the exams, relying on chance variation to produce balanced groups. This can be done in various ways, but a common principle prevails: each student is equally likely to receive exam `A` or `B`. Here are two possibilities:

1. Ishmael creates 40 identical slips of paper. He writes the name of each student on one slip, mixes the slips in a large jar, then draws 20 slips. (After each draw, the selected slip is set aside and the next draw uses only those slips that remain in the jar, i.e., sampling occurs *without replacement*.) The 20 students selected receive exam `A`; the remaining 20 students receive exam `B`. This is called *simple random sampling*.

2. Ishmael notices that his class comprises 30 freshmen and 10 non-freshman. Believing that it is essential to have 3/4 freshmen in each group, he assigns freshmen and non-freshmen separately. Again, Ishmael creates 40 identical slips of paper and writes the name of each student on one slip. This time he separates the 30 freshman slips from the 10 non-freshman slips. To assign the freshmen, he mixes the 30 freshman slips and draws 15 slips. The 15 freshmen selected receive exam `A`; the remaining 15 freshmen receive exam `B`. To assign the non-freshmen, he mixes the 10 non-freshman slips and draws 5 slips. The 5 non-freshmen selected receive exam `A`; the remaining 5 non-freshmen receive exam `B`. This is called *stratified random sampling*.

1.3 The Importance of Probability

Each of the experiments described in Sections 1.1 and 1.2 reveals something about the role of chance variation in scientific experimentation.

It is beyond our ability to predict with certainty if a spinning penny will come to rest with `Heads` facing up. Even if we believe that the outcome is completely determined, we cannot measure all the relevant variables with sufficient precision, nor can we perform the necessary calculations, to know what it will be. We express our inability to predict `Heads` versus `Tails` in the language of probability, e.g., "there is a 30% chance that `Heads` will result." (Section 3.1 discusses how such statements may be interpreted.) Thus, *even when studying allegedly deterministic phenomena*, probability models may be of enormous value.

When measuring the speed of light, it is not the phenomenon itself but the experiment that admits chance variation. Despite his excruciating precautions, Michelson was unable to remove chance variation from his experiment—his measurements differ. Adjusting the measurements for temperature removes one source of variation, but it is impossible to remove them all. Later experiments with more sophisticated equipment produced better measurements, but did not succeed in completely removing all sources of variation. Experiments are never perfect,[12] and probability models may be of enormous value in modelling errors that the experimenter is unable to remove or control.

Probability plays another, more subtle role in statistical inference. When studying termites, it is not clear whether or not one is observing a systematic foraging strategy. Probability was introduced as a hypothetical benchmark: *what if* termites forage randomly? Even if termites actually do forage deterministically, understanding how they would behave if they foraged randomly provides insights that inform our judgments about their behavior.

Thus, probability helps us answer questions that naturally arise when analyzing experimental data. Another example arose when we remarked that Matt, James, and Sarah observed *nearly* 50% `Heads`, specifically 145 `Heads` in 300 spins. What do we mean by "nearly"? Is this an important discrepancy or can chance variation account for it? To find out, we might study the behavior of penny spinning under the mathematical assumption that it is fair. If we learn that 300 spins of a fair penny rarely produce a discrepancy of 5 (or more) `Heads`, then we might conclude that penny spinning is not fair. If we learn that discrepancies of this magnitude are

[12] Another example is described by Freedman, Pisani, and Purves in Section 6.2 of *Statistics* (Third Edition, W. W. Norton & Company, 1998). The National Bureau of Standards repeatedly weighs the national prototype kilogram under carefully controlled conditions. The measurements are extremely precise, but nevertheless subject to small variations.

common, then we would be reluctant to draw this conclusion.

The ability to use the tools of probability to understand the behavior of inferential procedures is so powerful that good experiments are designed with this in mind. Besides avoiding the pitfalls of subjective methods, randomization allows us to answer questions about how well our methods work. For example, Ishmael might ask "How likely is simple random sampling to result in exactly 5 non-freshman receiving exam `A`?" Such questions derive meaning from the use of probability methods.

When a scientist performs an experiment, s/he observes a *sample* of possible experimental values. The set of all values that might have been observed is a *population*. Probability helps us describe the population and understand the data generating process that produced the sample. It also helps us understand the behavior of the statistical procedures used to analyze experimental data, e.g., averaging 100 measurements to produce an estimate. This linkage, of sample to population through probability, is the foundation on which statistical inference is based. Statistical inference is relatively new, but the linkage that we have described is wonderfully encapsulated in a remarkable passage from *The Book of Nala*, the third book of the ancient Indian epic *Mahábarata*.[13] Rtuparna examines a single twig of a spreading tree and accurately estimates the number of fruit on two great branches. Nala marvels at this ability, and Rtuparna rejoins:

> I of dice possess the science
> and in numbers thus am skilled.

1.4 Games of Chance

In *The Book of Nala*, Rtuparna's skill in estimation is connected with his prowess at dicing. Throughout history, probabilistic concepts have invariably been illustrated using simple games of chance. There are excellent reasons for us to embrace this pedagogical cliché. First, many fundamental probabilistic concepts were invented for the purpose of understanding certain games of chance; it is pleasant to incorporate a bit of this fascinating, centuries-old history into a modern program of study. Second, games of chance serve as idealized experiments that effectively reveal essential issues without the distraction of the many complicated nuances associated

[13] This passage is summarized in Ian Hacking's *The Emergence of Probability*, Cambridge University Press, 1975, pp. 6–7, which quotes H. H. Milman's 1860 translation.

with most scientific experiments. Third, as idealized experiments, games of chance provide canonical examples of various recurring experimental structures. For example, tossing a coin is a useful abstraction of such diverse experiments as observing whether a baby is male or female, observing whether an Alzheimer's patient does or does not know the day of the week, or observing whether a pond is or is not inhabited by geese. A scientist who is familiar with these idealized experiments will find it easier to diagnose the mathematical structure of an actual scientific experiment.

Many of the examples and exercises in subsequent chapters will refer to simple games of chance. The present section collects some facts and trivia about several of the most common.

Coins According to the *Encylopædia Britannica*,

> "Early cast-bronze animal shapes of known and readily identifiable weight, provided for the beam-balance scales of the Middle Eastern civilizations of the 7th millennium BC, are evidence of the first attempts to provide a medium of exchange. ... The first true coins, that is, cast disks of standard weight and value specifically designed as a medium of exchange, were probably produced by the Lydians of Anatolia in about 640 BC from a natural alloy of gold containing 20 to 35 percent silver."[14]

Despite (or perhaps because of) the simplicity of tossing a coin and observing which side (canonically identified as `Heads` or `Tails`) comes to lie facing up, it appears that coins did not play an important role in the early history of probability. Nevertheless, the use of coin tosses (or their equivalents) as randomizing agents is ubiquitous in modern times. In football, an official tosses a coin and a representative of one team calls `Heads` or `Tails`. If his call matches the outcome of the toss, then his team may choose whether to kick or receive (or, which goal to defend); otherwise, the opposing team chooses. A similar practice is popular in tennis, except that one player spins a racquet instead of tossing a coin. In each of these practices, it is presumed that the "coin" is *balanced* or *fair*, i.e., that each side is equally likely to turn up; see Section 1.1.1 for a discussion of whether or not spinning a penny is fair.

[14] "Coins and coinage," *The New Encyclopædia Britannica in 30 Volumes*, Macropædia, Volume 4, 1974, pp. 821–822.

Dice The noun *dice* is the plural form of the noun *die*.[15] A die is a small cube, marked on each of its six faces with a number of pips (spots, dots). To generate a random outcome, the die is cast (tossed, thrown, rolled) on a smooth surface and the number of pips on the uppermost face is observed. If each face is equally likely to be uppermost, then the die is *balanced* or *fair*; otherwise, it is *unbalanced* or *loaded*.

The casting of dice is an ancient practice. According to F. N. David,

> "The earliest dice so far found are described as being of well-fired buff pottery and date from the beginning of the third millenium. ... consecutive order of the pips must have continued for some time. It is still to be seen in dice of the late XVIIIth Dynasty (Egypt *c.* 1370 B.C.), but about that time, or soon after, the arrangement must have settled into the 2-partitions of 7 familiar to us at the present time. Out of some fifty dice of the classical period which I have seen, forty had the 'modern' arrangement of the pips."[16]

Today, pure dice games include craps, in which two dice are cast, and Yahtzee, in which five dice are cast. More commonly, the casting of dice is used as a randomizing agent in a variety of board games, e.g., backgammon and Monopoly™. Typically, two dice are cast and the outcome is defined to be the sum of the pips on the two uppermost faces.

Astragali Even more ancient than dice are *astragali*, the singular form of which is *astragalus*. The astragalus is a bone in the heel of many vertebrate animals; it lies directly above the talus, and is roughly symmetrical in hooved mammals, e.g., deer. Such astragali have been found in abundance in excavations of prehistoric man, who may have used them for counting. They were used for board games at least as early as the First Dynasty in Egypt (*c.* 3500 B.C.) and were the principal randomizing agent in classical Greece and Rome. According to F. N. David,

> "The astragalus has only four sides on which it will rest, since the other two are rounded... A favourite research of the scholars of

[15] In *The Devil's Dictionary*, Ambrose Bierce defined *die* as the singular of *dice*, remarking that "we seldom hear the word, because there is a prohibitory proverb, 'Never say die.' "

[16] F. N. David, *Games, Gods and Gambling: A History of Probability and Statistical Ideas*, 1962, p. 10 (Dover Publications).

> the Italian Renaissance was to try to deduce the scoring used. It was generally agreed from a close study of the writings of classical times that the upper side of the bone, broad and slightly convex, counted 4; the opposite side, broad and slightly concave, 3; the lateral side, flat and narrow, scored 1, and the opposite narrow lateral side, which is slightly hollow, 6. The numbers 2 and 5 were omitted."[17]

Accordingly, we can think of an astragalus as a 4-sided die with possible outcomes 1, 3, 4, and 6. An astragalus is not balanced. From tossing a modern sheep's astragalus, David estimated the chances of throwing a 1 or a 6 at roughly 10 percent each and the chances of throwing a 3 or a 4 at roughly 40 percent each.

The Greeks and Romans invariably cast four astragali. The most desirable result, the *venus*, occurred when the four uppermost sides were all different; the *dog*, which occurred when each uppermost side was a 1, was undesirable. In Asia Minor, five astragali were cast and different results were identified with the names of different gods, e.g., the throw of Saviour Zeus (one one, two threes, and two fours), the throw of child-eating Cronos (three fours and two sixes), etc. In addition to their use in gaming, astragali were cast for the purpose of divination, i.e., to ascertain if the gods favored a proposed undertaking.

In 1962, David reported that "it is not uncommon to see children in France and Italy playing games with them [astragali] today;" for the most part, however, unbalanced astragali have given way to balanced dice. A whimsical contemporary example of unbalanced dice that evoke astragali are the pig dice used in Pass the Pig™ (formerly Pigmania™).

Cards David estimated that playing cards "were not invented until *c.* A.D. 1350, but once in use, they slowly began to display dice both as instruments of play and for fortune-telling." By a *standard deck of playing cards*, we shall mean the familiar deck of 52 cards, organized into four *suits* (clubs, diamonds, hearts, spades) of thirteen *ranks* or *denominations* (2–10, jack, queen, king, ace). The diamonds and hearts are red; the clubs and spades are black. When we say that a deck has been shuffled, we mean that the order of the cards in the deck has been randomized. When we say that cards are dealt, we mean that they are removed from a shuffled deck in sequence,

[17] F. N. David, *Games, Gods and Gambling: A History of Probability and Statistical Ideas*, 1962, p. 7 (Dover Publications).

beginning with the top card. The cards received by a player constitute that player's *hand*. The quality of a hand depends on the game being played; however, unless otherwise specified, the order in which the player received the cards in her hand is irrelevant.

Poker involves hands of five cards. The following types of hands are arranged in order of decreasing value. An ace is counted as either the highest or the lowest rank, whichever results in the more valuable hand. Thus, every possible hand is of exactly one type.

1. A *straight flush* contains five cards of the same suit and of consecutive ranks.

2. A hand with *4 of a kind* contains cards of exactly two ranks, four cards of one rank and one of the other rank.

3. A *full house* contains cards of exactly two ranks, three cards of one rank and two cards of the other rank.

4. A *flush* contains five cards of the same suit, not of consecutive rank.

5. A *straight* contains five cards of consecutive rank, not all of the same suit.

6. A hand with *3 of a kind* contains cards of exactly three ranks, three cards of one rank and one card of each of the other two ranks.

7. A hand with *two pairs* contains contains cards of exactly three ranks, two cards of one rank, two cards of a second rank, and one card of a third rank.

8. A hand with *one pair* contains cards of exactly four ranks, two cards of one rank and one card each of a second, third, and fourth rank.

9. Any other hand contains *no pair*.

Urns For the purposes of this book, an urn is a container from which objects are drawn, e.g., a box of raffle tickets or a jar of marbles. Modern lotteries often select winning numbers by using air pressure to draw numbered ping pong balls from a clear plastic container. When an object is drawn from an urn, it is presumed that each object in the urn is equally likely to be selected.

That urn models have enormous explanatory power was first recognized by J. Bernoulli (1654–1705), who used them in *Ars Conjectandi*, his brilliant treatise on probability. It is not difficult to devise urn models that are equivalent to other randomizing agents considered in this section.

Example 1.1: Urn Model for Tossing a Fair Coin Imagine an urn that contains one red marble and one black marble. A marble is drawn from this urn. If it is red, then the outcome is `Heads`; if it is black, then the outcome is `Tails`. This is equivalent to tossing a fair coin *once*.

Example 1.2: Urn Model for Throwing a Fair Die Imagine an urn that contains six tickets, labelled 1 through 6. Drawing one ticket from this urn is equivalent to throwing a fair die *once*. If we want to throw the die a second time, then we return the selected ticket to the urn and repeat the procedure. This is an example of drawing *with replacement*.

Example 1.3: Urn Model for Throwing an Astragalus Imagine an urn that contains ten tickets, one labelled 1, four labelled 3, four labelled 4, and one labelled 6. Drawing one ticket from this urn is equivalent to throwing an astragalus *once*. If we want to throw four astragali, then we repeat this procedure four times, each time returning the selected ticket to the urn. This is another example of drawing *with replacement*.

Example 1.4: Urn Model for Drawing a Poker Hand Place a standard deck of playing cards in an urn. Draw one card, then a second, then a third, then a fourth, then a fifth. Because each card in the deck can only be dealt once, we do not return a card to the urn after drawing it. This is an example of drawing *without replacement*.

In the preceding examples, the statements about the equivalence of the urn model and another randomizing agent were intended to appeal to your intuition. Subsequent chapters will introduce mathematical tools that will allow us to validate these assertions.

1.5 Exercises

1. Select a penny minted in any year other than 1982. Find a smooth surface on which to spin it. Practice spinning the penny until you are

able to do so in a reasonably consistent manner. Develop an experimental protocol that specifies precisely how you spin your penny. Spin your penny 100 times in accordance with this protocol. Record the outcome of each spin, including aberrant events (e.g., the penny spun off the table and therefore neither `Heads` nor `Tails` was recorded). Your report of this experiment should include the following:

- The penny itself, taped to your report. Note any features of the penny that seem relevant, e.g., the year and city in which it was minted, its condition, etc.
- A description of the surface on which you spun it and of any possibly relevant environmental considerations.
- A description of your experimental protocol.
- The results of your 100 spins. This means a list, in order, of what happened on each spin.
- A summary of your results. This means (i) the total number of spins that resulted in either `Heads` or `Tails` (ideally, this number, n, will equal 100) and (ii) the number of spins that resulted in `Heads` (y).
- The observed frequency of heads, y/n.

2. The Department of Mathematics at the College of William & Mary is housed in Jones Hall. To find the department, one passes through the building's main entrance, into its lobby, and immediately turns left. In Jones 131, the department's seminar room, is a long rectangular wood table. Let L denote the length of this table. The purpose of this experiment is to measure L using a standard (12-inch) ruler.

   You will need a 12-inch ruler that is marked in increments of 1/16 inches. Groups of students may use the same ruler, but it is important that each student obtain his/her own measurement of L. Please do not attempt to obtain your measurement at a time when Jones 131 is being used for a seminar or faculty meeting!

   Your report of this experiment should include the following information:

   - A description of the ruler that you used. From what was it made? In what condition is it? Who owns it? What other students used the same ruler?

- A description of your measuring protocol. How did you position the ruler initially? How did you reposition it? How did you ensure that you were measuring along a straight line?
- An account of the experiment. When did you measure? How long did it take you? Please note any unusual circumstances that might bear on your results.
- Your estimate (in inches, to the nearest 1/16 inch) of L.

3. Statisticians say that a procedure that tends to either underestimate or overestimate the quantity that it is being used to determine is *biased*.

   (a) In the preceding problem, suppose that you tried to measure the length of the table with a ruler that—unbeknownst to you—was really 11.9 inches long instead of the nominal 12 inches. Would you tend to underestimate or overestimate the true length of the table? Explain.

   (b) In the Lanarkshire milk experiment, would a tendency for well-fed children to wear heavier winter clothing than ill-nourished children cause weight gains due to milk supplements to be underestimated or overestimated? Explain.

Chapter 2

Mathematical Preliminaries

This chapter collects some fundamental mathematical concepts that we will use in our study of probability and statistics. Most of these concepts should seem familiar, although our presentation of them may be a bit more formal than you have previously encountered. This formalism will be quite useful as we study probability, but it will tend to recede into the background as we progress to the study of statistics.

2.1 Sets

It is an interesting bit of trivia that "set" has the most different meanings of any word in the English language. To describe what we mean by a set, we suppose the existence of a designated *universe* of possible objects. In this book, we will often denote the universe by S. By a *set*, we mean a collection of objects with the property that each object in the universe either does or does not belong to the collection. We will tend to denote sets by uppercase Roman letters toward the beginning of the alphabet, e.g., A, B, C, etc. The set of objects that do not belong to a designated set A is called the *complement* of A. We will denote complements by A^c, B^c, C^c, etc. The complement of the universe is the *empty set*, denoted $S^c = \emptyset$.

An object that belongs to a designated set is called an *element* or *member* of that set. We will tend to denote elements by lower case Roman letters and write expressions such as $x \in A$, pronounced "x is an element of the set A." Sets with a small number of elements are often identified by simple enumeration, i.e., by writing down a list of elements. When we do so, we will enclose the list in braces and separate the elements by commas or semicolons.

For example, the set of all feature films directed by Sergio Leone is

$$\begin{array}{ll} \{ & \textit{A Fistful of Dollars}; \\ & \textit{For a Few Dollars More}; \\ & \textit{The Good, the Bad, and the Ugly}; \\ & \textit{Once Upon a Time in the West}; \\ & \textit{Duck, You Sucker!}; \\ & \textit{Once Upon a Time in America} \qquad \} \end{array}$$

In this book, of course, we usually will be concerned with sets defined by certain mathematical properties. Some familiar sets to which we will refer repeatedly include:

- The set of *natural numbers*, $\boldsymbol{N} = \{1, 2, 3, \ldots\}$.
- The set of *integers*, $\boldsymbol{Z} = \{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}$.
- The set of *real numbers*, $\Re = (-\infty, \infty)$.

If A and B are sets and each element of A is also an element of B, then we say that A is a *subset* of B and write $A \subset B$. For example,

$$\boldsymbol{N} \subset \boldsymbol{Z} \subset \Re.$$

Quite often, a set A is defined to be those elements of another set B that satisfy a specified mathematical property. In such cases, we often specify A by writing a generic element of B to the left of a colon, the property to the right of the colon, and enclosing this syntax in braces. For example,

$$A = \{x \in \boldsymbol{Z} \ : \ x^2 < 5\} = \{-2, -1, 0, 1, 2\},$$

is pronounced "A is the set of integers x such that x^2 is less than 5."

Given sets A and B, there are several important sets that can be constructed from them. The *union* of A and B is the set

$$A \cup B = \{x \in S \ : \ x \in A \text{ or } x \in B\}$$

and the *intersection* of A and B is the set

$$A \cap B = \{x \in S \ : \ x \in A \text{ and } x \in B\}.$$

For example, if A is as above and

$$B = \{x \in \boldsymbol{Z} \ : \ |x - 2| \leq 1\} = \{1, 2, 3\},$$

then $A \cup B = \{-2, -1, 0, 1, 2, 3\}$ and $A \cap B = \{1, 2\}$. Notice that unions and intersections are symmetric constructions, i.e., $A \cup B = B \cup A$ and $A \cap B = B \cap A$.

If $A \cap B = \emptyset$, i.e., if A and B have no elements in common, then A and B are *disjoint* or *mutually exclusive*. By convention, the empty set is a subset of every set, so

$$\emptyset \subset A \cap B \subset A \subset A \cup B \subset S$$

and

$$\emptyset \subset A \cap B \subset B \subset A \cup B \subset S.$$

These facts are illustrated by the *Venn diagram* in Figure 2.1, in which sets are qualitatively indicated by connected subsets of the plane. We will make frequent use of Venn diagrams as we develop basic facts about probabilities.

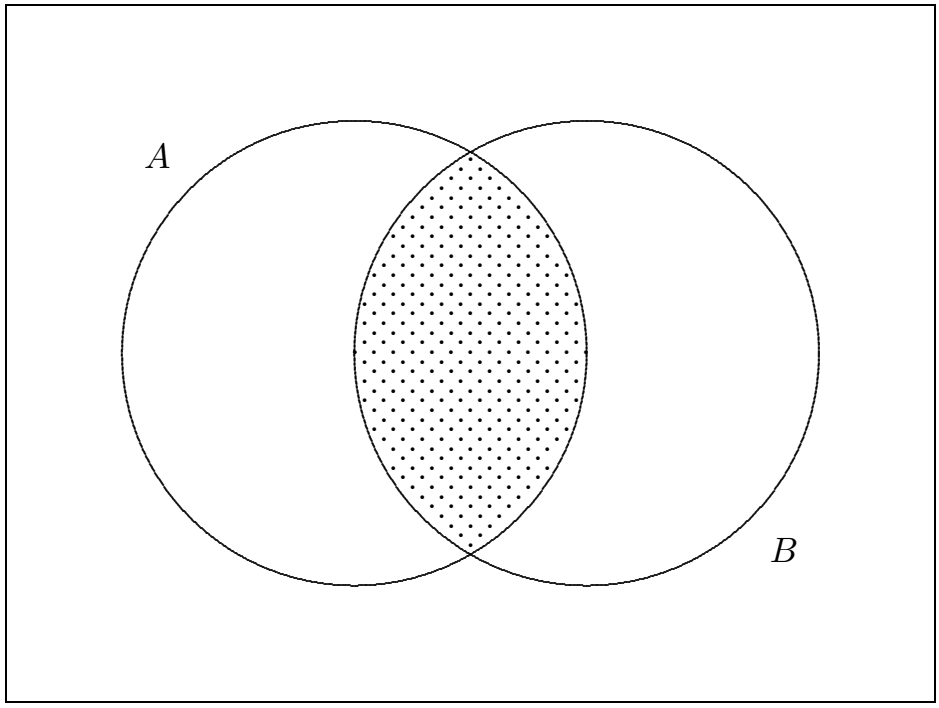

Figure 2.1: A Venn diagram. The shaded region represents the intersection of the nondisjoint sets A and B.

It is often useful to extend the concepts of union and intersection to more than two sets. Let $\{A_k\}$ denote an arbitrary collection of sets, where k is an index that identifies the set. Then $x \in S$ is an element of the union of $\{A_k\}$,

denoted

$$\bigcup_k A_k,$$

if and only if there exists some k_0 such that $x \in A_{k_0}$. Also, $x \in S$ is an element of the intersection of $\{A_k\}$, denoted

$$\bigcap_k A_k,$$

if and only if $x \in A_k$ for every k. For example, if $A_k = \{0, 1, \ldots, k\}$ for $k = 1, 2, 3, \ldots$, then

$$\bigcup_k A_k = \{0, 1, 2, 3, \ldots\}$$

and

$$\bigcap_k A_k = \{0, 1\}.$$

Furthermore, it will be important to distinguish collections of sets with the following property:

Definition 2.1 *A collection of sets is* pairwise disjoint *if and only if each pair of sets in the collection has an empty intersection.*

Unions and intersections are related to each other by two distributive laws:

$$B \cap \left(\bigcup_k A_k \right) = \bigcup_k (B \cap A_k)$$

and

$$B \cup \left(\bigcap_k A_k \right) = \bigcap_k (B \cup A_k) .$$

Furthermore, unions and intersections are related to complements by De-Morgan's laws:

$$\left(\bigcup_k A_k \right)^c = \bigcap_k A_k^c$$

and

$$\left(\bigcap_k A_k \right)^c = \bigcup_k A_k^c.$$

The first law states that an object is not in any of the sets in the collection if and only if it is in the complement of each set; the second law states that

an object is not in every set in the collection if it is in the complement of at least one set.

Finally, we consider another important set that can be constructed from A and B.

Definition 2.2 *The Cartesian product of two sets A and B, denoted $A \times B$, is the set of ordered pairs whose first component is an element of A and whose second component is an element of B, i.e.,*

$$A \times B = \{(a, b) \ : \ a \in A, \, b \in B\}.$$

For example, if $A = \{-2, -1, 0, 1, 2\}$ and $B = \{1, 2, 3\}$, then the set $A \times B$ contains the following elements:

$$\begin{array}{lllll} (-2,1) & (-1,1) & (0,1) & (1,1) & (2,1) \\ (-2,2) & (-1,2) & (0,2) & (1,2) & (2,2) \\ (-2,3) & (-1,3) & (0,3) & (1,3) & (2,3) \end{array}$$

A familiar example of a Cartesian product is the Cartesian coordinatization of the plane,

$$\Re^2 = \Re \times \Re = \{(x, y) \ : \ x, y \in \Re\}.$$

Of course, this construction can also be extended to more than two sets, e.g.,

$$\Re^3 = \{(x, y, z) \ : \ x, y, z \in \Re\}.$$

2.2 Counting

This section is concerned with determining the number of elements in a specified set. One of the fundamental concepts that we will exploit in our brief study of counting is the notion of a *one-to-one correspondence* between two sets. We begin by illustrating this notion with an elementary example.

Example 2.1 Define two sets,

$$A_1 = \{\text{diamond}, \text{emerald}, \text{ruby}, \text{sapphire}\}$$

and

$$B = \{\text{blue}, \text{green}, \text{red}, \text{white}\}.$$

The elements of these sets can be paired in such a way that to each element of A_1 there is assigned a unique element of B and to each element of B there

is assigned a unique element of A_1. Such a pairing can be accomplished in various ways; a natural assignment is the following:

$$\begin{aligned} \text{diamond} &\leftrightarrow \text{white} \\ \text{emerald} &\leftrightarrow \text{green} \\ \text{ruby} &\leftrightarrow \text{red} \\ \text{sapphire} &\leftrightarrow \text{blue} \end{aligned}$$

This assignment exemplifies a one-to-one correspondence.

Now suppose that we augment A_1 by forming

$$A_2 = A_1 \cup \{\text{aquamarine}\}.$$

Although we can still assign a color to each gemstone, we *cannot* do so in such a way that each gemstone corresponds to a different color. There does not exist a one-to-one correspondence between A_2 and B.

From Example 2.1, we abstract

Definition 2.3 *Two sets can be placed in* one-to-one correspondence *if their elements can be paired in such a way that each element of either set is associated with a unique element of the other set.*

The concept of one-to-one correspondence can then be exploited to obtain a formal definition of a familiar concept:

Definition 2.4 *A set A is* finite *if there exists a natural number N such that the elements of A can be placed in one-to-one correspondence with the elements of $\{1, 2, \ldots, N\}$.*

If A is finite, then the natural number N that appears in Definition 2.4 is unique. It is, in fact, the number of elements in A. We will denote this quantity, sometimes called the *cardinality* of A, by $\#(A)$. In Example 2.1 above, $\#(A_1) = \#(B) = 4$ and $\#(A_2) = 5$.

The Multiplication Principle Most of our counting arguments will rely on a fundamental principle, which we illustrate with an example.

Example 2.2 *Suppose that each gemstone in Example 2.1 has been mounted on a ring. You desire to wear one of these rings on your left hand and another on your right hand. How many ways can this be done?*

First, suppose that you wear the diamond ring on your left hand. Then there are three rings available for your right hand: emerald, ruby, sapphire.

Next, suppose that you wear the emerald ring on your left hand. Again there are three rings available for your right hand: diamond, ruby, sapphire.

Suppose that you wear the ruby ring on your left hand. Once again there are three rings available for your right hand: diamond, emerald, sapphire.

Finally, suppose that you wear the sapphire ring on your left hand. Once more there are three rings available for your right hand: diamond, emerald, ruby.

We have counted a total of $3 + 3 + 3 + 3 = 12$ ways to choose a ring for each hand. Enumerating each possibility is rather tedious, but it reveals a useful shortcut. There are 4 ways to choose a ring for the left hand and, for each such choice, there are three ways to choose a ring for the right hand. Hence, there are $4 \cdot 3 = 12$ ways to choose a ring for each hand. This is an instance of a general principle:

> *Suppose that two decisions are to be made and that there are n_1 possible outcomes of the first decision. If, for each outcome of the first decision, there are n_2 possible outcomes of the second decision, then there are $n_1 n_2$ possible outcomes of the pair of decisions.*

Permutations and Combinations We now consider two more concepts that are often employed when counting the elements of finite sets. We motivate these concepts with an example.

Example 2.3 *A fast-food restaurant offers a single entree that comes with a choice of* 3 *side dishes from a total of* 15*. To address the perception that it serves only one dinner, the restaurant conceives an advertisement that identifies each choice of side dishes as a distinct dinner. Assuming that each entree must be accompanied by* 3 *distinct side dishes, e.g.,* {stuffing, mashed potatoes, green beans} *is permitted but* {stuffing, stuffing, mashed potatoes} *is not, how many distinct dinners are available?*[1]

Answer 2.3a The restaurant reasons that a customer, asked to choose 3 side dishes, must first choose 1 side dish from a total of 15. There are

[1] This example is based on an actual incident involving the Boston Chicken (now Boston Market) restaurant chain and a high school math class in Denver, CO.

15 ways of making this choice. Having made it, the customer must then choose a second side dish that is different from the first. For each choice of the first side dish, there are 14 ways of choosing the second; hence 15×14 ways of choosing the pair. Finally, the customer must choose a third side dish that is different from the first two. For each choice of the first two, there are 13 ways of choosing the third; hence $15 \times 14 \times 13$ ways of choosing the triple. Accordingly, the restaurant advertises that it offers a total of $15 \times 14 \times 13 = 2730$ possible dinners.

Answer 2.3b A high school math class considers the restaurant's claim and notes that the restaurant has counted side dishes of

{	stuffing,	mashed potatoes,	green beans	},
{	stuffing,	green beans,	mashed potatoes	},
{	mashed potatoes,	stuffing,	green beans	},
{	mashed potatoes,	green beans,	stuffing	},
{	green beans,	stuffing,	mashed potatoes	}, and
{	green beans,	mashed potatoes,	stuffing	}

as distinct dinners. Thus, the restaurant has counted dinners that differ only with respect to the order in which the side dishes were chosen as distinct. Reasoning that what matters is what is on one's plate, not the order in which the choices were made, the math class concludes that the restaurant has overcounted. As illustrated above, each triple of side dishes can be ordered in 6 ways: the first side dish can be any of 3, the second side dish can be any of the remaining 2, and the third side dish must be the remaining 1 ($3 \times 2 \times 1 = 6$). The math class writes a letter to the restaurant, arguing that the restaurant has overcounted by a factor of 6 and that the correct count is $2730 \div 6 = 455$. The restaurant cheerfully agrees and donates \$1000 to the high school's math club.

From Example 2.3 we abstract the following definitions:

Definition 2.5 *The number of* permutations *(ordered choices) of r objects from n objects is*

$$P(n, r) = n \times (n - 1) \times \cdots \times (n - r + 1).$$

Definition 2.6 *The number of* combinations *(unordered choices) of r objects from n objects is*

$$C(n, r) = P(n, r) \div P(r, r).$$

In Example 2.3, the restaurant claimed that it offered $P(15, 3)$ dinners, while the math class argued that a more plausible count was $C(15, 3)$. There, as always, the distinction was made on the basis of whether the order of the choices is or is not relevant.

Permutations and combinations are often expressed using factorial notation. Let

$$0! = 1$$

and let k be a natural number. Then the expression $k!$, pronounced "k factorial" is defined recursively by the formula

$$k! = k \times (k-1)!.$$

For example,

$$3! = 3 \times 2! = 3 \times 2 \times 1! = 3 \times 2 \times 1 \times 0! = 3 \times 2 \times 1 \times 1 = 3 \times 2 \times 1 = 6.$$

Because

$$\begin{aligned} n! &= n \times (n-1) \times \cdots \times (n-r+1) \times (n-r) \times \cdots \times 1 \\ &= P(n, r) \times (n-r)!, \end{aligned}$$

we can write

$$P(n, r) = \frac{n!}{(n-r)!}$$

and

$$C(n, r) = P(n, r) \div P(r, r) = \frac{n!}{(n-r)!} \div \frac{r!}{(r-r)!} = \frac{n!}{r!(n-r)!}.$$

Finally, we note (and will sometimes use) the popular notation

$$C(n, r) = \binom{n}{r},$$

pronounced "n choose r".

Example 2.4 *A coin is tossed* 10 *times. How many sequences of* 10 *tosses result in a total of exactly* 2 `Heads`*?*

Answer A sequence of `Heads` and `Tails` is completely specified by knowing which tosses resulted in `Heads`. To count how many sequences result in 2 `Heads`, we simply count how many ways there are to choose the pair of tosses on which `Heads` result. This is choosing 2 tosses from 10, or

$$\binom{10}{2} = \frac{10!}{2!(10-2)!} = \frac{10 \cdot 9}{2 \cdot 1} = 45.$$

Example 2.5 *Consider the hypothetical example described in Section 1.2. In a class of* 40 *students, how many ways can one choose* 20 *students to receive exam* `A`*? Assuming that the class comprises* 30 *freshmen and* 10 *non-freshmen, how many ways can one choose* 15 *freshmen and* 5 *non-freshmen to receive exam* `A`*?*

Solution There are

$$\binom{40}{20} = \frac{40!}{20!(40-20)!} = \frac{40 \cdot 39 \cdots 22 \cdot 21}{20 \cdot 19 \cdots 2 \cdot 1} = 137,846,528,820$$

ways to choose 20 students from 40. There are

$$\binom{30}{15} = \frac{30!}{15!(30-15)!} = \frac{30 \cdot 29 \cdots 17 \cdot 16}{15 \cdot 14 \cdots 2 \cdot 1} = 155,117,520$$

ways to choose 15 freshmen from 30 and

$$\binom{10}{5} = \frac{10!}{5!(10-5)!} = \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 252$$

ways to choose 5 non-freshmen from 10; hence,

$$155,117,520 \cdot 252 = 39,089,615,040$$

ways to choose 15 freshmen and 5 non-freshmen to receive exam `A`. Notice that, of all the ways to choose 20 students to receive exam `A`, about 28% result in exactly 15 freshman and 5 non-freshman.

Countability Thus far, our study of counting has been concerned exclusively with finite sets. However, our subsequent study of probability will require us to consider sets that are not finite. Toward that end, we introduce the following definitions:

Definition 2.7 *A set is* infinite *if it is not finite.*

Definition 2.8 *A set is* denumerable *if its elements can be placed in one-to-one correspondence with the natural numbers.*

Definition 2.9 *A set is* countable *if it is either finite or denumerable.*

Definition 2.10 *A set is* uncountable *if it is not countable.*

Like Definition 2.4, Definition 2.8 depends on the notion of a one-to-one correspondence between sets. However, whereas this notion is completely straightforward when at least one of the sets is finite, it can be rather elusive when both sets are infinite. Accordingly, we provide some examples of denumerable sets. In each case, we superscript each element of the set in question with the corresponding natural number.

Example 2.6 Consider the set of even natural numbers, which excludes one of every two consecutive natural numbers It might seem that this set cannot be placed in one-to-one correspondence with the natural numbers in their entirety; however, infinite sets often possess counterintuitive properties. Here is a correspondence that demonstrates that this set is denumerable:

$$2^1, 4^2, 6^3, 8^4, 10^5, 12^6, 14^7, 16^8, 18^9, \ldots$$

Example 2.7 Consider the set of integers. It might seem that this set, which includes both a positive and a negative copy of each natural number, cannot be placed in one-to-one correspondence with the natural numbers; however, here is a correspondence that demonstrates that this set is denumerable:

$$\ldots, -4^9, -3^7, -2^5, -1^3, 0^1, 1^2, 2^4, 3^6, 4^8, \ldots$$

Example 2.8 Consider the Cartesian product of the set of natural numbers with itself. This set contains one copy of the entire set of natural numbers for each natural number—surely it cannot be placed in one-to-one correspondence with a single copy of the set of natural numbers! In fact, the

following correspondence demonstrates that this set is also denumerable:

$$\begin{array}{ccccccc}
(1,1)^1 & (1,2)^2 & (1,3)^6 & (1,4)^7 & (1,5)^{15} & \dots \\
(2,1)^3 & (2,2)^5 & (2,3)^8 & (2,4)^{14} & (2,5)^{17} & \dots \\
(3,1)^4 & (3,2)^9 & (3,3)^{13} & (3,4)^{18} & (3,5)^{26} & \dots \\
(4,1)^{10} & (4,2)^{12} & (4,3)^{19} & (4,4)^{25} & (4,5)^{32} & \dots \\
(5,1)^{11} & (5,2)^{20} & (5,3)^{24} & (5,4)^{33} & (5,5)^{41} & \dots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{array}$$

In light of Examples 2.6–2.8, the reader may wonder what is required to construct a set that is not countable. We conclude this section by remarking that the following intervals are uncountable sets, where $a, b \in \Re$ and $a < b$.

$$\begin{aligned}
(a,b) &= \{x \in \Re \,:\, a < x < b\} \\
[a,b) &= \{x \in \Re \,:\, a \leq x < b\} \\
(a,b] &= \{x \in \Re \,:\, a < x \leq b\} \\
[a,b] &= \{x \in \Re \,:\, a \leq x \leq b\}
\end{aligned}$$

We will make frequent use of such sets, often referring to (a, b) as an *open* interval and $[a, b]$ as a *closed* interval.

2.3 Functions

A function is a rule that assigns a unique element of a set B to each element of another set A. A familiar example is the rule that assigns to each real number x the real number $y = x^2$, e.g., that assigns $y = 4$ to $x = 2$. Notice that each real number has a unique square ($y = 4$ is the only number that this rule assigns to $x = 2$), but that more than one number may have the same square ($y = 4$ is also assigned to $x = -2$).

The set A is the function's *domain*. Notice that each element of A must be assigned some element of B, but that an element of B need not be assigned to any element of A. Thus, in the preceding example, every $x \in A = \Re$ has a squared value $y \in B = \Re$, but not every $y \in B$ is the square of some number $x \in A$. (For example, $y = -4$ is not the square of any real number.) The elements of B that are assigned to elements of A constitute the *image* of the function. In the preceding example, the image of $f(x) = x^2$ is $f(\Re) = [0, \infty)$.

We will use a variety of letters to denote various types of functions. Examples include $P, X, Y, f, g, F, G, \phi$. If ϕ is a function with domain A and

range B, then we write $\phi : A \to B$, often pronounced "ϕ maps A into B". If ϕ assigns $b \in B$ to $a \in A$, then we say that b is the value of ϕ at a and we write $b = \phi(a)$.

If $\phi : A \to B$, then for each $b \in B$ there is a subset (possibly empty) of A comprising those elements of A at which ϕ has value b. We denote this set by

$$\phi^{-1}(b) = \{a \in A \ : \ \phi(a) = b\}.$$

For example, if $\phi : \Re \to \Re$ is the function defined by $\phi(x) = x^2$, then

$$\phi^{-1}(4) = \{-2, 2\}.$$

More generally, if $B_0 \subset B$, then

$$\phi^{-1}(B_0) = \{a \in A \ : \ \phi(a) \in B_0\}.$$

Using the same example,

$$\phi^{-1}([4, 9]) = \left\{x \in \Re \ : \ x^2 \in [4, 9]\right\} = [-3, -2] \cup [2, 3].$$

The object ϕ^{-1} is called the *inverse* of ϕ and $\phi^{-1}(B_0)$ is called the inverse image of B_0.

2.4 Limits

In Section 2.2 we examined several examples of denumerable sets of real numbers. In each of these examples, we imposed an order on the set when we placed it in one-to-one correspondence with the natural numbers. Once an order has been specified, we can inquire how the set behaves as we progress through its values in the prescribed sequence. For example, the real numbers in the ordered denumerable set

$$\left\{1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \ldots\right\} \tag{2.1}$$

steadily decrease as one progresses through them. Furthermore, as in Zeno's famous paradoxes, the numbers seem to approach the value zero without ever actually attaining it. To describe such sets, it is helpful to introduce some specialized terminology and notation.

We begin with

Definition 2.11 *A sequence of real numbers is an ordered denumerable subset of* $\Re$.

Sequences are often denoted using a dummy variable that is specified or understood to index the natural numbers. For example, we might identify the sequence (2.1) by writing $\{1/n\}$ for $n = 1, 2, 3, \ldots$.

Next we consider the phenomenon that $1/n$ approaches 0 as n increases, although each $1/n > 0$. Let ϵ denote any strictly positive real number. What we have noticed is the fact that, no matter how small ϵ may be, eventually n becomes so large that $1/n < \epsilon$. We formalize this observation in

Definition 2.12 *Let* $\{y_n\}$ *denote a sequence of real numbers. We say that* $\{y_n\}$ *converges to a constant value* $c \in \Re$ *if, for every* $\epsilon > 0$*, there exists a natural number* N *such that* $y_n \in (c - \epsilon, c + \epsilon)$ *for each* $n \geq N$.

If the sequence of real numbers $\{y_n\}$ converges to c, then we say that c is the *limit* of $\{y_n\}$ and we write either $y_n \to c$ as $n \to \infty$ or $\lim_{n\to\infty} y_n = c$. In particular,

$$\lim_{n\to\infty} \frac{1}{n} = 0.$$

2.5 Exercises

1. A classic riddle inquires:

   As I was going to St. Ives,
   I met a man with seven wives.
   Each wife had seven sacks,
   Each sack had seven cats,
   Each cat had seven kits.
   Kits, cats, sacks, wives—
   How many were going to St. Ives?

   (a) How many creatures (human and feline) were in the entourage that the narrator encountered?

   (b) What is the answer to the riddle?

2. A well-known carol, "The Twelve Days of Christmas," describes a progression of gifts that the singer receives from her true love:

> On the first day of Christmas, my true love gave to me:
> A partridge in a pear tree.
> On the second day of Christmas, my true love gave to me:
> Two turtle doves, and a partridge in a pear tree.
> Et cetera.[2]

How many birds did the singer receive from her true love?

3. The throw of an astragalus (see Section 1.4) has four possible outcomes, $\{1, 3, 4, 6\}$. When throwing four astragali,

   (a) How many ways are there to obtain a dog, i.e., for each astragalus to produce a 1?

   (b) How many ways are there to obtain a venus, i.e., for each astragalus to produce a different outcome?

   Hint: Label each astragalus (e.g., antelope, bison, cow, deer) and keep track of the outcome of each distinct astragalus.

4. When throwing five astragali,

   (a) How many ways are there to obtain the throw of child-eating Cronos, i.e., to obtain three fours and two sixes?

   (b) How many ways are there to obtain the throw of Saviour Zeus, i.e., to obtain one one, two threes, and two fours?

5. The throw of one die has six possible outcomes, $\{1, 2, 3, 4, 5, 6\}$. A medieval poem, "The Chance of the Dyse," enumerates the fortunes that could be divined from casting three dice. Order does not matter, e.g., the fortune associated with 6-5-3 is also associated with 3-5-6. How many fortunes does the poem enumerate?

6. Suppose that five cards are dealt from a standard deck of playing cards.

   (a) How many hands are possible?

   (b) How many straight-flush hands are possible?

   (c) How many 4-of-a-kind hands are possible?

   (d) Why do you suppose that a straight flush beats 4-of-a-kind?

[2] You should be able to find the complete lyrics by doing a web search.

7. In the television reality game show *Survivor*, 16 contestants (the "castaways") compete for $1 million. The castaways are stranded in a remote location, e.g., an uninhabited island in the China Sea. Initially, the castaways are divided into two tribes. The tribes compete in a sequence of immunity challenges. After each challenge, the losing tribe must vote out one of its members and that person is eliminated from the game. Eventually, the tribes merge and the surviving castaways compete in a sequence of individual immunity challenges. The winner receives immunity and the merged tribe must then vote out one of its other members. After the merged tribe has been reduced to two members, a jury of the last 7 castaways to have been eliminated votes on who should be the Sole Survivor and win $1 million. (Technically, the jury votes *for* the Sole Survivor, but this is equivalent to eliminating one of the final two castaways.)

   (a) Suppose that we define an outcome of *Survivor* to be the name of the Sole Survivor. In any given game of *Survivor*, how many outcomes are possible?

   (b) Suppose that we define an outcome of *Survivor* to be a list of the castaways' names, arranged in the order in which they were eliminated. In any given game of *Survivor*, how many outcomes are possible?

8. The final eight castaways in *Survivor 2: Australian Outback* included four men (Colby, Keith, Nick, and Rodger) and four women (Amber, Elisabeth, Jerri, and Tina). They participated in a reward challenge that required them to form four teams of two persons, one male and one female. (The teams raced over an obstacle course, recording the time of the slower team member.) The castaways elected to pair off by drawing lots.

   (a) How many ways were there for the castaways to form four teams?

   (b) Jerri was opposed to drawing lots—she wanted to team with Colby. How many ways are there for the castaways to form four male-female teams if one of the teams is Colby-Jerri?

   (c) If all pairings (male-male, male-female, female-female) are allowed, then how many ways are there for the castaways to form four teams?

9. In Major League Baseball's World Series, the winners of the National (`N`) and American (`A`) League pennants play a sequence of games. The first team to win four games wins the Series. Thus, the Series must last at least four games and can last no more than seven games. Let us define an *outcome* of the World Series by identifying which League's pennant winner won each game. For example, the outcome of the 1975 World Series, in which the Cincinnati Reds represented the National League and the Boston Red Sox represented the American League, was `ANNANAN`. How many World Series outcomes are possible?

10. The following table defines a function that assigns to each feature film directed by Sergio Leone the year in which it was released.

A Fistful of Dollars	1964
For a Few Dollars More	1965
The Good, the Bad, and the Ugly	1966
Once Upon a Time in the West	1968
Duck, You Sucker!	1972
Once Upon a Time in America	1984

What is the inverse image of the set known as *The Sixties*?

11. For $n = 0, 1, 2, \ldots$, let

$$y_n = \sum_{k=0}^{n} 2^{-k} = 2^{-0} + 2^{-1} + \cdots + 2^{-n}.$$

(a) Compute y_0, y_1, y_2, y_3, and y_4.

(b) The sequence $\{y_0, y_1, y_2, \ldots\}$ is an example of a *sequence of partial sums*. Guess the value of its limit, usually written

$$\lim_{n\to\infty} y_n = \lim_{n\to\infty} \sum_{k=0}^{n} 2^{-k} = \sum_{k=0}^{\infty} 2^{-k}.$$

Chapter 3

Probability

The goal of statistical inference is to draw conclusions about a population from "representative information" about it. In future chapters, we will discover that a powerful way to obtain representative information about a population is through the planned introduction of chance. Thus, probability is the foundation of statistical inference—to study the latter, we must first study the former. Fortunately, the theory of probability is an especially beautiful branch of mathematics. Although our purpose in studying probability is to provide the reader with some tools that will be needed when we study statistics, we also hope to impart some of the beauty of those tools.

3.1 Interpretations of Probability

Probabilistic statements can be interpreted in different ways. For example, how would you interpret the following statement?

> There is a 40 percent chance of rain today.

Your interpretation is apt to vary depending on the context in which the statement is made. If the statement was made as part of a forecast by the National Weather Service, then something like the following interpretation might be appropriate:

> In the recent history of this locality, of all days on which present atmospheric conditions have been experienced, rain has occurred on approximately 40 percent of them.

This is an example of the *frequentist* interpretation of probability. With this interpretation, a probability is a long-run average proportion of occurence.

Suppose, however, that you had just peered out a window, wondering if you should carry an umbrella to school, and asked your roommate if she thought that it was going to rain. Unless your roommate is studying metereology, it is not plausible that she possesses the knowledge required to make a frequentist statement! If her response was a casual "I'd say that there's a 40 percent chance," then something like the following interpretation might be appropriate:

> I believe that it might very well rain, but that it's a little less likely to rain than not.

This is an example of the *subjectivist* interpretation of probability. With this interpretation, a probability expresses the strength of one's belief.

The philosopher I. Hacking has observed that dual notions of probability, one aleatory (frequentist) and one epistemological (subjectivist) have co-existed throughout history, and that "philosophers seem singularly unable to put [them] asunder..."[1] We shall not attempt so perilous an undertaking. But however we decide to interpret probabilities, we will need a formal mathematical description of probability to which we can appeal for insight and guidance. The remainder of this chapter provides an introduction to the most commonly adopted approach to *axiomatic probability*. The chapters that follow tend to emphasize a frequentist interpretation of probability, but the mathematical formalism can also be used with a subjectivist interpretation.

3.2 Axioms of Probability

The mathematical model that has dominated the study of probability was formalized by the Russian mathematician A. N. Kolmogorov in a monograph published in 1933. The central concept in this model is a *probability space*, which is assumed to have three components:

S A *sample space*, a universe of "possible" outcomes for the experiment in question.

[1] I. Hacking, *The Emergence of Probability*, Cambridge University Press, 1975, Chapter 2: Duality.

$\mathcal{C}$ A designated collection of "observable" subsets (called *events*) of the sample space.

P A *probability measure*, a function that assigns real numbers (called *probabilities*) to events.

We describe each of these components in turn.

The Sample Space The sample space is a set. Depending on the nature of the experiment in question, it may or may not be easy to decide upon an appropriate sample space.

Example 3.1 *A coin is tossed once.*
A plausible sample space for this experiment will comprise two outcomes, `Heads` and `Tails`. Denoting these outcomes by `H` and `T`, we have

$$S = \{\texttt{H}, \texttt{T}\}.$$

Remark: We have discounted the possibility that the coin will come to rest on edge. This is the first example of a theme that will recur throughout this text, that mathematical models are rarely—if ever—completely faithful representations of nature. As described by Mark Kac,

> "Models are, for the most part, caricatures of reality, but if they are good, then, like good caricatures, they portray, though perhaps in distorted manner, some of the features of the real world. The main role of models is not so much to explain and predict—though ultimately these are the main functions of science—as to polarize thinking and to pose sharp questions."[2]

In Example 3.1, and in most of the other elementary examples that we will use to illustrate the fundamental concepts of axiomatic probability, the fidelity of our mathematical descriptions to the physical phenomena described should be apparent. Practical applications of inferential statistics, however, often require imposing mathematical assumptions that may be suspect. Data analysts must constantly make judgments about the plausibility of their assumptions, not so much with a view to whether or not the assumptions are completely correct (they almost never are), but with a view to whether or not the assumptions are sufficient for the analysis to be meaningful.

[2]Mark Kac, "Some mathematical models in science," *Science*, 1969, 166:695–699.

Example 3.2 *A coin is tossed twice.*

A plausible sample space for this experiment will comprise four outcomes, two outcomes per toss. Here,

$$S = \left\{ \begin{array}{cc} \texttt{HH} & \texttt{TH} \\ \texttt{HT} & \texttt{TT} \end{array} \right\}.$$

Example 3.3 *An individual's height is measured.*

In this example, it is less clear what outcomes are possible. All human heights fall within certain bounds, but precisely what bounds should be specified? And what of the fact that heights are not measured exactly?

Only rarely would one address these issues when choosing a sample space. For this experiment, most statisticians would choose as the sample space the set of all real numbers, then worry about which real numbers were actually observed. Thus, the phrase "possible outcomes" refers to conceptual rather than practical possibility. The sample space is usually chosen to be mathematically convenient and all-encompassing.

The Collection of Events Events are subsets of the sample space, but how do we decide which subsets of S should be designated as events? If the outcome $s \in S$ was observed and $E \subset S$ is an event, then we say that E *occurred* if and only if $s \in E$. A subset of S is *observable* if it is always possible for the experimenter to determine whether or not it occurred. Our intent is that the collection of events should be the collection of observable subsets. This intent is often tempered by our desire for mathematical convenience and by our need for the collection to possess certain mathematical properties. In practice, the issue of observability is rarely considered and certain conventional choices are automatically adopted. For example, when S is a finite set, one usually designates *all* subsets of S to be events.

Whether or not we decide to grapple with the issue of observability, the collection of events *must* satisfy the following properties:

1. The sample space is an event.
2. If E is an event, then E^c is an event.
3. The union of any countable collection of events is an event.

A collection of subsets with these properties is sometimes called a *sigma-field*.

Taken together, the first two properties imply that both S and $\emptyset$ must be events. If S and $\emptyset$ are the only events, then the third property holds;

hence, the collection $\{S, \emptyset\}$ is a sigma-field. It is not, however, a very useful collection of events, as it describes a situation in which the experimental outcomes cannot be distinguished!

Example 3.1 (continued) To distinguish `Heads` from `Tails`, we must assume that each of these individual outcomes is an event. Thus, the only plausible collection of events for this experiment is the collection of all subsets of S, i.e.,

$$\mathcal{C} = \{S, \{\texttt{H}\}, \{\texttt{T}\}, \emptyset\}\,.$$

Example 3.2 (continued) If we designate all subsets of S as events, then we obtain the following collection:

$$\mathcal{C} = \left\{ \begin{array}{l} S, \\ \{\texttt{HH},\texttt{HT},\texttt{TH}\},\ \{\texttt{HH},\texttt{HT},\texttt{TT}\}, \\ \{\texttt{HH},\texttt{TH},\texttt{TT}\},\ \{\texttt{HT},\texttt{TH},\texttt{TT}\}, \\ \{\texttt{HH},\texttt{HT}\},\ \{\texttt{HH},\texttt{TH}\},\ \{\texttt{HH},\texttt{TT}\}, \\ \{\texttt{HT},\texttt{TH}\},\ \{\texttt{HT},\texttt{TT}\},\ \{\texttt{TH},\texttt{TT}\}, \\ \{\texttt{HH}\},\ \{\texttt{HT}\},\ \{\texttt{TH}\},\ \{\texttt{TT}\}, \\ \emptyset \end{array} \right\}.$$

This is perhaps the most plausible collection of events for this experiment, but others are also possible. For example, suppose that we were unable to distinguish the order of the tosses, so that we could not distinguish between the outcomes `HT` and `TH`. Then the collection of events should not include any subsets that contain one of these outcomes but not the other, e.g., $\{\texttt{HH},\texttt{TH},\texttt{TT}\}$. Thus, the following collection of events might be deemed appropriate:

$$\mathcal{C} = \left\{ \begin{array}{l} S, \\ \{\texttt{HH},\texttt{HT},\texttt{TH}\},\ \{\texttt{HT},\texttt{TH},\texttt{TT}\}, \\ \{\texttt{HH},\texttt{TT}\},\{\texttt{HT},\texttt{TH}\}, \\ \{\texttt{HH}\},\ \{\texttt{TT}\}, \\ \emptyset \end{array} \right\}.$$

The interested reader should verify that this collection is indeed a sigma-field.

The Probability Measure Once the collection of events has been designated, each event $E \in \mathcal{C}$ can be assigned a probability $P(E)$. This must

be done according to specific rules; in particular, the probability measure P *must* satisfy the following properties:

1. If E is an event, then $0 \leq P(E) \leq 1$.
2. $P(S) = 1$.
3. If $\{E_1, E_2, E_3, \ldots\}$ is a countable collection of pairwise disjoint events, then
$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i).$$

We discuss each of these properties in turn.

The first property states that probabilities are nonnegative and finite. Thus, neither the statement that "the probability that it will rain today is $-.5$" nor the statement that "the probability that it will rain today is infinity" are meaningful. These restrictions have certain mathematical consequences. The further restriction that probabilities are no greater than unity is actually a consequence of the second and third properties.

The second property states that the probability that an outcome occurs, that *something* happens, is unity. Thus, the statement that "the probability that it will rain today is 2" is not meaningful. This is a convention that simplifies formulae and facilitates interpretation.

The third property, called *countable additivity*, is the most interesting. Consider Example 3.2, supposing that $\{\texttt{HT}\}$ and $\{\texttt{TH}\}$ are events and that we want to compute the probability that exactly one `Head` is observed, i.e., the probability of
$$\{\texttt{HT}\} \cup \{\texttt{TH}\} = \{\texttt{HT}, \texttt{TH}\}.$$
Because $\{\texttt{HT}\}$ and $\{\texttt{TH}\}$ are events, their union is an event and therefore has a probability. Because they are mutually exclusive, we would like that probability to be
$$P\left(\{\texttt{HT}, \texttt{TH}\}\right) = P\left(\{\texttt{HT}\}\right) + P\left(\{\texttt{TH}\}\right).$$
We ensure this by requiring that the probability of the union of any two disjoint events is the sum of their respective probabilities.

Having assumed that
$$A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B), \tag{3.1}$$

it is easy to compute the probability of any finite union of pairwise disjoint events. For example, if A, B, C, and D are pairwise disjoint events, then

$$\begin{aligned} P(A \cup B \cup C \cup D) &= P(A \cup (B \cup C \cup D)) \\ &= P(A) + P(B \cup C \cup D) \\ &= P(A) + P(B \cup (C \cup D)) \\ &= P(A) + P(B) + P(C \cup D) \\ &= P(A) + P(B) + P(C) + P(D) \end{aligned}$$

Thus, from (3.1) can be deduced the following implication:

If $E_1, \ldots, E_n$ are pairwise disjoint events, then

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P(E_i).$$

This implication is known as *finite additivity*. Notice that the union of $E_1, \ldots, E_n$ must be an event (and hence have a probability) because each E_i is an event.

An extension of finite additivity, countable additivity is the following implication:

If $E_1, E_2, E_3, \ldots$ are pairwise disjoint events, then

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i).$$

The reason for insisting upon this extension has less to do with applications than with theory. Although some axiomatic theories of probability assume only finite additivity, it is generally felt that the stronger assumption of countable additivity results in a richer theory. Again, notice that the union of $E_1, E_2, \ldots$ must be an event (and hence have a probability) because each E_i is an event.

Finally, we emphasize that *probabilities are assigned to events*. It may or may not be that the individual experimental outcomes are events. If they are, then they will have probabilities. In some such cases (see Chapter 4), the probability of any event can be deduced from the probabilities of the individual outcomes; in other such cases (see Chapter 5), this is not possible.

All of the facts about probability that we will use in studying statistical inference are consequences of the assumptions of the Kolmogorov probability model. It is not the purpose of this book to present derivations of these facts; however, three elementary (and useful) propositions suggest how one might proceed along such lines. In each case, a Venn diagram helps to illustrate the proof.

Theorem 3.1 *If E is an event, then*

$$P\left(E^c\right) = 1 - P(E).$$

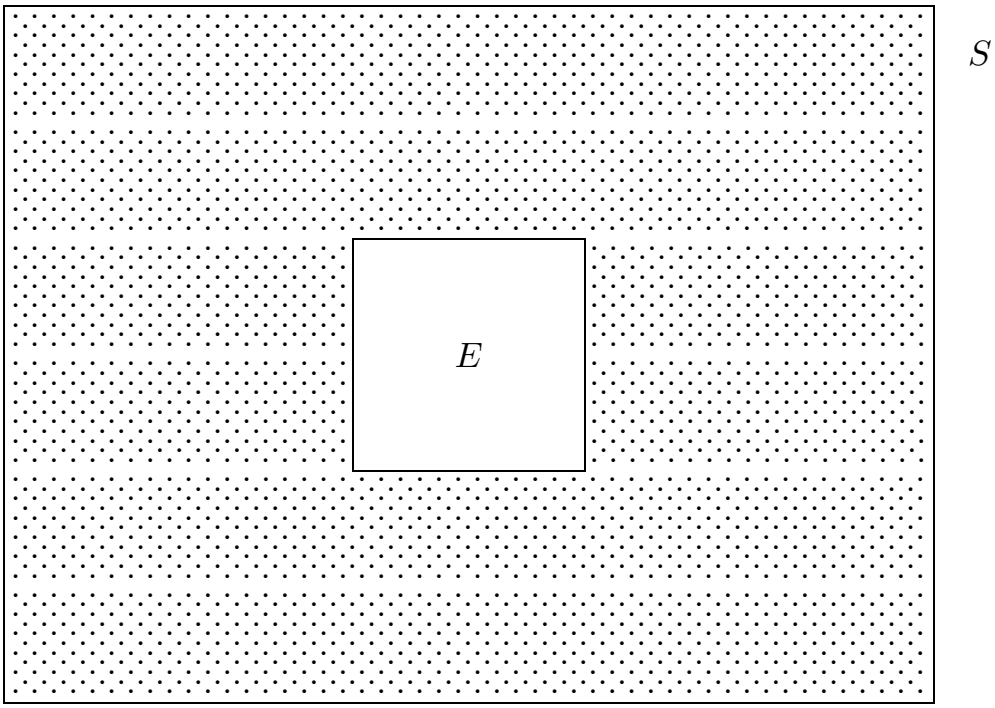

Figure 3.1: A Venn diagram for the probability of E^c.

Proof Refer to Figure 3.1. E^c is an event because E is an event. By definition, E and E^c are disjoint events whose union is S. Hence,

$$1 = P(S) = P\left(E \cup E^c\right) = P(E) + P\left(E^c\right)$$

and the theorem follows upon subtracting $P(E)$ from both sides. □

Theorem 3.2 *If A and B are events and $A \subset B$, then*

$$P(A) \leq P(B).$$

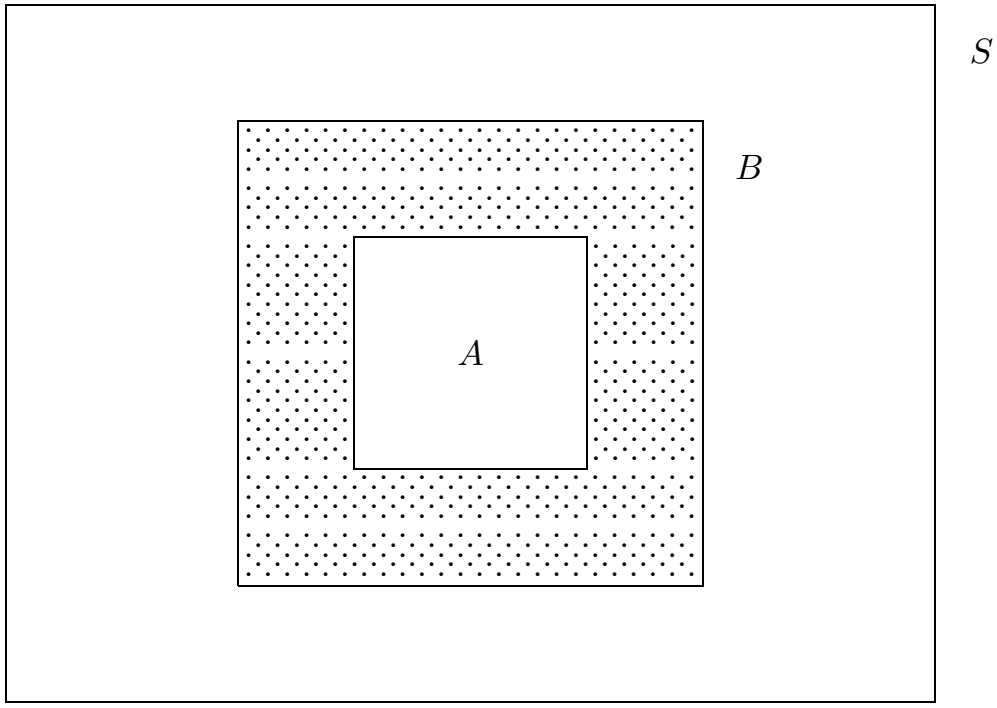

Figure 3.2: A Venn diagram for the probability of $A \subset B$.

Proof Refer to Figure 3.2. A^c is an event because A is an event. Hence, $B \cap A^c$ is an event and

$$B = A \cup (B \cap A^c).$$

Because A and $B \cap A^c$ are disjoint events,

$$P(B) = P(A) + P(B \cap A^c) \geq P(A),$$

as claimed. □

Theorem 3.3 *If A and B are events, then*

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

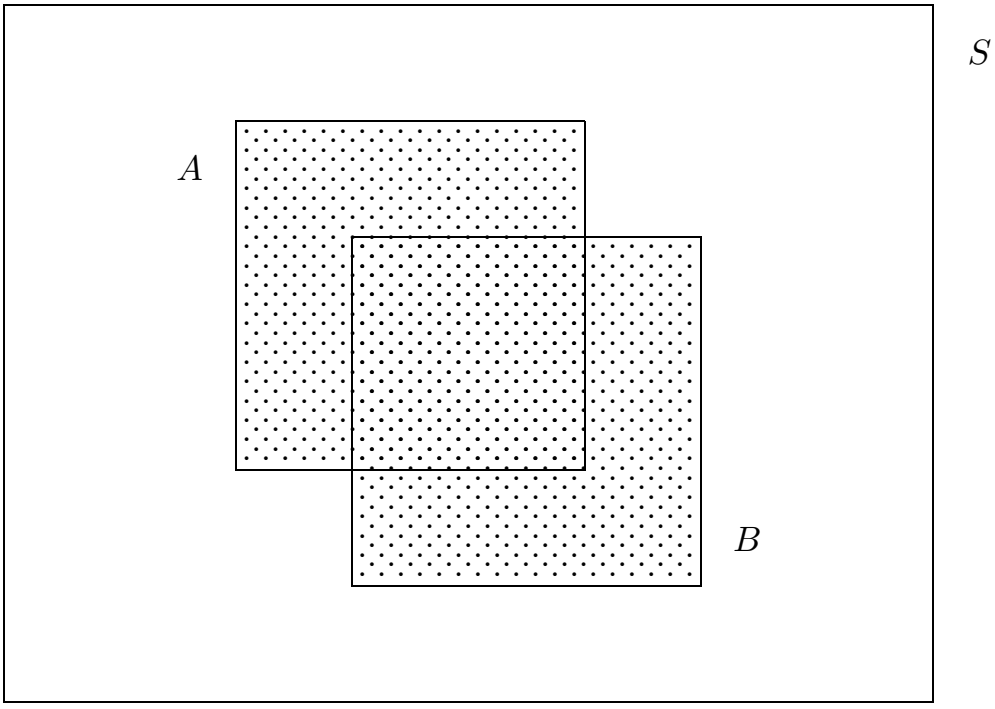

Figure 3.3: A Venn diagram for the probability of $A \cup B$.

Proof Refer to Figure 3.3. Both $A \cup B$ and $A \cap B = (A^c \cup B^c)^c$ are events because A and B are events. Similarly, $A \cap B^c$ and $B \cap A^c$ are also events.

Notice that $A \cap B^c$, $B \cap A^c$, and $A \cap B$ are pairwise disjoint events. Hence,

$$\begin{aligned}
&P(A) + P(B) - P(A \cap B) \\
&\quad = P\left((A \cap B^c) \cup (A \cap B)\right) + P\left((B \cap A^c) \cup (A \cap B)\right) - P(A \cap B) \\
&\quad = P\left(A \cap B^c\right) + P(A \cap B) + P\left(B \cap A^c\right) + P(A \cap B) - P(A \cap B) \\
&\quad = P\left(A \cap B^c\right) + P(A \cap B) + P\left(B \cap A^c\right) \\
&\quad = P\left((A \cap B^c) \cup (A \cap B) \cup (B \cap A^c)\right) \\
&\quad = P(A \cup B),
\end{aligned}$$

as claimed. □

Theorem 3.3 provides a general formula for computing the probability of the union of two sets. Notice that, if A and B are in fact disjoint, then

$$P\left(A \cap B\right) = P(\emptyset) = P\left(S^c\right) = 1 - P(S) = 1 - 1 = 0$$

and we recover our original formula for that case.

3.3 Finite Sample Spaces

Let

$$S = \{s_1, \ldots, s_N\}$$

denote a sample space that contains N outcomes and suppose that every subset of S is an event. For notational convenience, let

$$p_i = P(\{s_i\})$$

denote the probability of outcome i, for $i = 1, \ldots, N$. Then, for any event A, we can write

$$P(A) = P\left(\bigcup_{s_i \in A} \{s_i\}\right) = \sum_{s_i \in A} P(\{s_i\}) = \sum_{s_i \in A} p_i. \tag{3.2}$$

Thus, if the sample space is finite, then the probabilities of the individual outcomes determine the probability of any event. The same reasoning applies if the sample space is denumerable.

In this section, we focus on an important special case of finite probability spaces, the case of "equally likely" outcomes. By a fair coin, we mean a coin that when tossed is equally likely to produce `Heads` or `Tails`, i.e., the probability of each of the two possible outcomes is 1/2. By a fair die, we mean a die that when tossed is equally likely to produce any of six possible outcomes, i.e., the probability of each outcome is 1/6. In general, we say that the outcomes of a finite sample space are equally likely if

$$p_i = \frac{1}{N} \tag{3.3}$$

for $i = 1, \ldots, N$.

In the case of equally likely outcomes, we substitute (3.3) into (3.2) and obtain

$$P(A) = \sum_{s_i \in A} \frac{1}{N} = \frac{\sum_{s_i \in A} 1}{N} = \frac{\#(A)}{\#(S)}. \tag{3.4}$$

This equation reveals that, when the outcomes in a finite sample space are equally likely, calculating probabilities is just a matter of counting. The *counting* may be quite difficult, but the *probabilty* is trivial. We illustrate this point with some examples.

Example 3.4 *A fair coin is tossed twice. What is the probability of observing exactly one* `Head`*?*

The sample space for this experiment was described in Example 3.2. Because the coin is fair, each of the four outcomes in S is equally likely. Let A denote the event that exactly one `Head` is observed. Then $A = \{\texttt{HT}, \texttt{TH}\}$ and

$$P(A) = \frac{\#(A)}{\#(S)} = \frac{2}{4} = \frac{1}{2} = 0.5.$$

Example 3.5 *A fair die is tossed once. What is the probability that the number of dots on the top face of the die is a prime number?*

The sample space for this experiment is $S = \{1, 2, 3, 4, 5, 6\}$. Because the die is fair, each of the six outcomes in S is equally likely. Let A denote the event that a prime number is observed. If we agree to count 1 as a prime number, then $A = \{1, 2, 3, 5\}$ and

$$P(A) = \frac{\#(A)}{\#(S)} = \frac{4}{6} = \frac{2}{3}.$$

Example 3.6 *A deck of 40 cards, labelled 1,2,3,. . . ,40, is shuffled and cards are dealt as specified in each of the following scenarios.*

(a) *One hand of four cards is dealt to Arlen. What is the probability that Arlen's hand contains four even numbers?*

Let S denote the possible hands that might be dealt. Because the order in which the cards are dealt is not important,

$$\#(S) = \binom{40}{4}.$$

Let A denote the event that the hand contains four even numbers There are 20 even cards, so the number of ways of dealing 4 even cards is

$$\#(A) = \binom{20}{4}.$$

Substituting these expressions into (3.4), we obtain

$$P(A) = \frac{\#(A)}{\#(S)} = \frac{\binom{20}{4}}{\binom{40}{4}} = \frac{51}{962} \doteq 0.0530.$$

(b) *One hand of four cards is dealt to Arlen. What is the probability that this hand is a straight, i.e., that it contains four consecutive numbers?*

Let S denote the possible hands that might be dealt. Again,

$$\#(S) = \binom{40}{4}.$$

Let A denote the event that the hand is a straight. The possible straights are:

$$\begin{array}{c} \text{1-2-3-4} \\ \text{2-3-4-5} \\ \text{3-4-5-6} \\ \vdots \\ \text{37-38-39-40} \end{array}$$

By simple enumeration (just count the number of ways of choosing the smallest number in the straight), there are 37 such hands. Hence,

$$P(A) = \frac{\#(A)}{\#(S)} = \frac{37}{\binom{40}{4}} = \frac{1}{2470} \doteq 0.0004.$$

(c) *One hand of four cards is dealt to Arlen and a second hand of four cards is dealt to Mike. What is the probability that Arlen's hand is a straight and Mike's hand contains four even numbers?*

Let S denote the possible pairs of hands that might be dealt. Dealing the first hand requires choosing 4 cards from 40. After this hand has been dealt, the second hand requires choosing an additional 4 cards from the remaining 36. Hence,

$$\#(S) = \binom{40}{4} \cdot \binom{36}{4}.$$

Let A denote the event that Arlen's hand is a straight and Mike's hand contains four even numbers. There are 37 ways for Arlen's hand to be a straight. Each straight contains 2 even numbers, leaving 18 even numbers available for Mike's hand. Thus, for each way of dealing a straight to Arlen, there are $\binom{18}{4}$ ways of dealing 4 even numbers to Mike. Hence,

$$P(A) = \frac{\#(A)}{\#(S)} = \frac{37 \cdot \binom{18}{4}}{\binom{40}{4} \cdot \binom{36}{4}} \doteq 2.1032 \times 10^{-5}.$$

Example 3.7 *Five fair dice are tossed simultaneously.*

Let S denote the possible outcomes of this experiment. Each die has 6 possible outcomes, so

$$\#(S) = 6 \cdot 6 \cdot 6 \cdot 6 \cdot 6 = 6^5.$$

(a) *What is the probability that the top faces of the dice all show the same number of dots?*

Let A denote the specified event; then A comprises the following outcomes:

1-1-1-1-1
2-2-2-2-2
3-3-3-3-3
4-4-4-4-4
5-5-5-5-5
6-6-6-6-6

By simple enumeration, $\#(A) = 6$. (Another way to obtain $\#(A)$ is to observe that the first die might result in any of six numbers, after which only one number is possible for each of the four remaining dice. Hence, $\#(A) = 6 \cdot 1 \cdot 1 \cdot 1 \cdot 1 = 6$.) It follows that

$$P(A) = \frac{\#(A)}{\#(S)} = \frac{6}{6^5} = \frac{1}{1296} \doteq 0.0008.$$

(b) *What is the probability that the top faces of the dice show exactly four different numbers?*

Let A denote the specified event. If there are exactly 4 different numbers, then exactly 1 number must appear twice. There are 6 ways to choose the number that appears twice and $\binom{5}{2}$ ways to choose the two dice on which this number appears. There are $5 \cdot 4 \cdot 3$ ways to choose the 3 different numbers on the remaining dice. Hence,

$$P(A) = \frac{\#(A)}{\#(S)} = \frac{6 \cdot \binom{5}{2} \cdot 5 \cdot 4 \cdot 3}{6^5} = \frac{25}{54} \doteq 0.4630.$$

(c) *What is the probability that the top faces of the dice show exactly three 6's or exactly two 5's?*

Let A denote the event that exactly three 6's are observed and let B denote the event that exactly two 5's are observed. We must calculate

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{\#(A) + \#(B) - \#(A \cap B)}{\#(S)}.$$

There are $\binom{5}{3}$ ways of choosing the three dice on which a 6 appears and $5 \cdot 5$ ways of choosing a different number for each of the two remaining dice. Hence,

$$\#(A) = \binom{5}{3} \cdot 5^2.$$

There are $\binom{5}{2}$ ways of choosing the two dice on which a 5 appears and $5 \cdot 5 \cdot 5$ ways of choosing a different number for each of the three remaining dice. Hence,

$$\#(B) = \binom{5}{2} \cdot 5^3.$$

There are $\binom{5}{3}$ ways of choosing the three dice on which a 6 appears and only 1 way in which a 5 can then appear on the two remaining dice. Hence,

$$\#(A \cap B) = \binom{5}{3} \cdot 1.$$

Thus,

$$P(A \cup B) = \frac{\binom{5}{3} \cdot 5^2 + \binom{5}{2} \cdot 5^3 - \binom{5}{3}}{6^5} = \frac{1490}{6^5} \doteq 0.1916.$$

Example 3.8 (The Birthday Problem) *In a class of k students, what is the probability that at least two students share a common birthday?*

As is inevitably the case with constructing mathematical models of actual phenomena, some simplifying assumptions are required to make this problem tractable. We begin by assuming that there are 365 possible birthdays, i.e., we ignore February 29. Then the sample space, S, of possible birthdays for k students comprises 365^k outcomes.

Next we assume that each of the 365^k outcomes is equally likely. This is not literally correct, as slightly more babies are born in some seasons than

in others. Furthermore, if the class contains twins, then only certain pairs of birthdays are possible outcomes for those two students! In most situations, however, the assumption of equally likely outcomes is reasonably plausible.

Let A denote the event that at least two students in the class share a birthday. We might attempt to calculate

$$P(A) = \frac{\#(A)}{\#(S)},$$

but a moment's reflection should convince the reader that counting the number of outcomes in A is an extremely difficult undertaking. Instead, we invoke Theorem 3.1 and calculate

$$P(A) = 1 - P(A^c) = 1 - \frac{\#(A^c)}{\#(S)}.$$

This is considerably easier, because we count the number of outcomes in which each student has a different birthday by observing that 365 possible birthdays are available for the oldest student, after which 364 possible birthdays remain for the next oldest student, after which 363 possible birthdays remain for the next, etc. The formula is

$$\#\left(A^c\right) = 365 \cdot 364 \cdots (366 - k)$$

and so

$$P(A) = 1 - \frac{365 \cdot 364 \cdots (366 - k)}{365 \cdot 365 \cdots 365}.$$

The reader who computes $P(A)$ for several choices of k may be astonished to discover that a class of just $k = 23$ students is required to obtain $P(A) > 0.5$!

3.4 Conditional Probability

Consider a sample space with 10 equally likely outcomes, together with the events indicated in the Venn diagram that appears in Figure 3.4. Applying the methods of Section 3.3, we find that the (unconditional) probability of A is

$$P(A) = \frac{\#(A)}{\#(S)} = \frac{3}{10} = 0.3.$$

Suppose, however, that we know that we can restrict attention to the experimental outcomes that lie in B. Then the *conditional probability* of the

event A given the occurrence of the event B is

$$P(A|B) = \frac{\#(A \cap B)}{\#(S \cap B)} = \frac{1}{5} = 0.2.$$

Notice that (for this example) the conditional probability, $P(A|B)$, differs from the unconditional probability, $P(A)$.

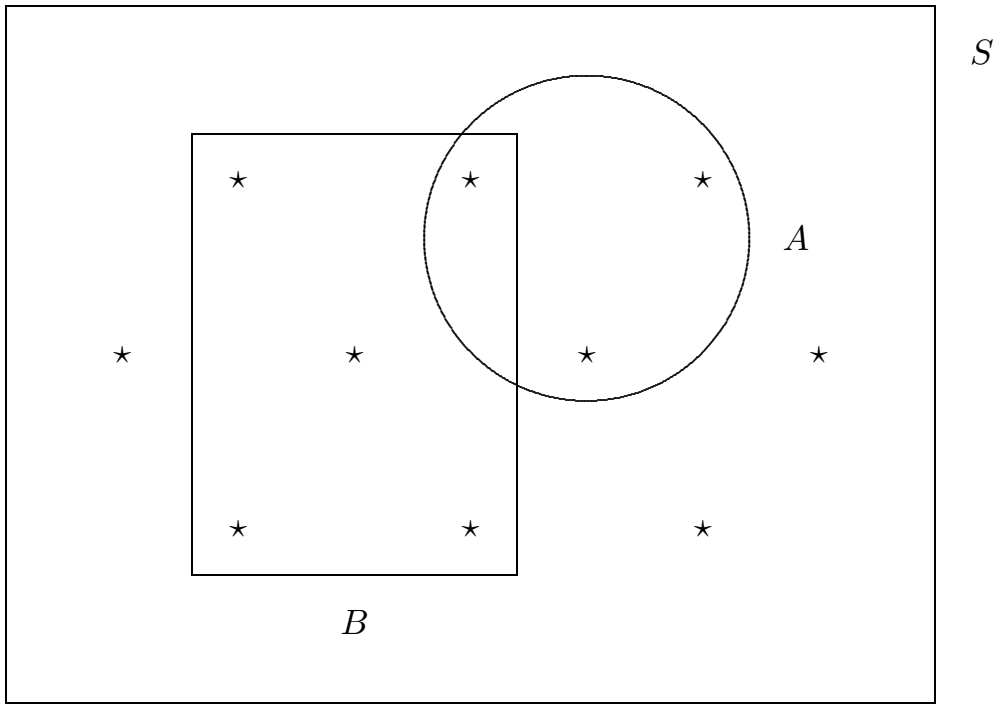

Figure 3.4: A Venn diagram that illustrates conditional probability. Each $\star$ represents an individual outcome.

To develop a definition of conditional probability that is not specific to finite sample spaces with equally likely outcomes, we now write

$$P(A|B) = \frac{\#(A \cap B)}{\#(S \cap B)} = \frac{\#(A \cap B)/\#(S)}{\#(B)/\#(S)} = \frac{P(A \cap B)}{P(B)}.$$

We take this as a definition:

Definition 3.1 *If A and B are events, and $P(B) > 0$, then*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{3.5}$$

The following consequence of Definition 3.1 is extremely useful. Upon multiplication of equation (3.5) by $P(B)$, we obtain

$$P(A \cap B) = P(B)P(A|B)$$

when $P(B) > 0$. Furthermore, upon interchanging the roles of A and B, we obtain

$$P(A \cap B) = P(B \cap A) = P(A)P(B|A)$$

when $P(A) > 0$. We will refer to these equations as the *multiplication rule* for conditional probability.

Used in conjunction with *tree diagrams*, the multiplication rule provides a powerful tool for analyzing situations that involve conditional probabilities.

Example 3.9 *Consider three fair coins, identical except that one coin (*`HH`*) is* `Heads` *on both sides, one coin (*`HT`*) is* `Heads` *on one side and* `Tails` *on the other, and one coin (*`TT`*) is* `Tails` *on both sides. A coin is selected at random and tossed. The face-up side of the coin is Heads. What is the probability that the face-down side of the coin is* `Heads`*?*

This problem was once considered by Marilyn vos Savant in her syndicated column, *Ask Marilyn*. As have many of the probability problems that she has considered, it generated a good deal of controversy. Many readers reasoned as follows:

1. The observation that the face-up side of the tossed coin is `Heads` means that the selected coin was not `TT`. Hence the selected coin was either `HH` or `HT`.

2. If `HH` was selected, then the face-down side is `Heads`; if `HT` was selected, then the face-down side is `Tails`.

3. Hence, there is a 1 in 2, or 50 percent, chance that the face-down side is `Heads`.

At first glance, this reasoning seems perfectly plausible and readers who advanced it were dismayed that Marilyn insisted that .5 is not the correct probability. How did these readers err?

A tree diagram of this experiment is depicted in Figure 3.5. The branches represent possible outcomes and the numbers associated with the branches are the respective probabilities of those outcomes. The initial triple of branches represents the initial selection of a coin—we have interpreted "at

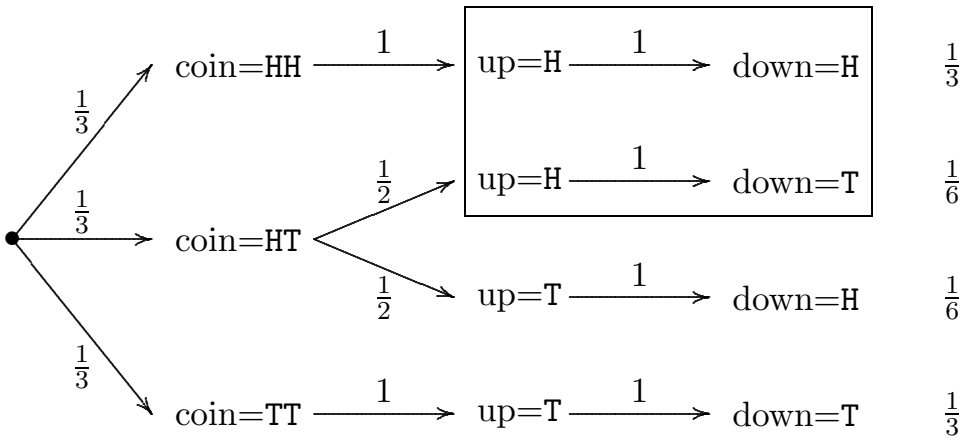

Figure 3.5: A tree diagram for Example 3.9.

random" to mean that each coin is equally likely to be selected. The second level of branches represents the toss of the coin by identifying its resulting up-side. For HH and TT, only one outcome is possible; for HT, there are two equally likely outcomes. Finally, the third level of branches represents the down-side of the tossed coin. In each case, this outcome is determined by the up-side.

The multiplication rule for conditional probability makes it easy to calculate the probabilities of the various paths through the tree. The probability that HT is selected and the up-side is Heads and the down-side is Tails is

$$\begin{aligned} P(\texttt{HT} \cap \text{up=}\texttt{H} \cap \text{down=}\texttt{T}) &= P(\texttt{HT} \cap \text{up=}\texttt{H}) \cdot P(\text{down=}\texttt{T} | \texttt{HT} \cap \text{up=}\texttt{H}) \\ &= P(\texttt{HT}) \cdot P(\text{up=}\texttt{H} | \texttt{HT}) \cdot 1 \\ &= (1/3) \cdot (1/2) \cdot 1 \\ &= 1/6 \end{aligned}$$

and the probability that HH is selected and the up-side is Heads and the down-side is Heads is

$$\begin{aligned} P(\texttt{HH} \cap \text{up=}\texttt{H} \cap \text{down=}\texttt{H}) &= P(\texttt{HH} \cap \text{up=}\texttt{H}) \cdot P(\text{down=}\texttt{H} | \texttt{HH} \cap \text{up=}\texttt{H}) \\ &= P(\texttt{HH}) \cdot P(\text{up=}\texttt{H} | \texttt{HH}) \cdot 1 \\ &= (1/3) \cdot 1 \cdot 1 \\ &= 1/3. \end{aligned}$$

Once these probabilities have been computed, it is easy to answer the original question:

$$P(\texttt{down=H}|\texttt{up=H}) = \frac{P(\texttt{down=H} \cap \texttt{up=H})}{P(\texttt{up=H})} = \frac{1/3}{(1/3)+(1/6)} = \frac{2}{3},$$

which was Marilyn's answer.

From the tree diagram, we can discern the fallacy in our first line of reasoning. Having narrowed the possible coins to `HH` and `HT`, we claimed that `HH` and `HT` were equally likely candidates to have produced the observed `Head`. In fact, `HH` was twice as likely as `HT`. Once this fact is noted it seems completely intuitive (`HH` has twice as many `Heads` as `HT`), but it is easily overlooked. This is an excellent example of how the use of tree diagrams may prevent subtle errors in reasoning.

Example 3.10 (Bayes Theorem) An important application of conditional probability can be illustrated by considering a population of patients at risk for contracting the HIV virus. The population can be partitioned into two sets: those who have contracted the virus and developed antibodies to it, and those who have not contracted the virus and lack antibodies to it. We denote the first set by D and the second set by D^c.

An ELISA test was designed to detect the presence of HIV antibodies in human blood. This test also partitions the population into two sets: those who test positive for HIV antibodies and those who test negative for HIV antibodies. We denote the first set by $+$ and the second set by $-$.

Together, the partitions induced by the true disease state and by the observed test outcome partition the population into four sets, as in the following Venn diagram:

$D \cap +$	$D \cap -$
$D^c \cap +$	$D^c \cap -$

(3.6)

In two of these cases, $D \cap +$ and $D^c \cap -$, the test provides the correct diagnosis; in the other two cases, $D^c \cap +$ and $D \cap -$, the test results in a diagnostic error. We call $D^c \cap +$ a *false positive* and $D \cap -$ a *false negative.*

In such situations, several quantities are likely to be known, at least approximately. The medical establishment is likely to have some notion of $P(D)$, the probability that a patient selected at random from the population is infected with HIV. This is the proportion of the population that is

infected—it is called the *prevalence* of the disease. For the calculations that follow, we will assume that $P(D) = .001$.

Because diagnostic procedures undergo extensive evaluation before they are approved for general use, the medical establishment is likely to have a fairly precise notion of the probabilities of false positive and false negative test results. These probabilities are conditional: a false positive is a positive test result within the set of patients who are not infected and a false negative is a negative test results within the set of patients who are infected. Thus, the probability of a false positive is $P(+|D^c)$ and the probability of a false negative is $P(-|D)$. For the calculations that follow, we will assume that $P(+|D^c) = .015$ and $P(-|D) = .003$.[3]

Now suppose that a randomly selected patient has a positive ELISA test result. Obviously, the patient has an extreme interest in properly assessing the chances that a diagnosis of HIV is correct. This can be expressed as $P(D|+)$, the conditional probability that a patient has HIV given a positive ELISA test. This quantity is called the *predictive value* of the test.

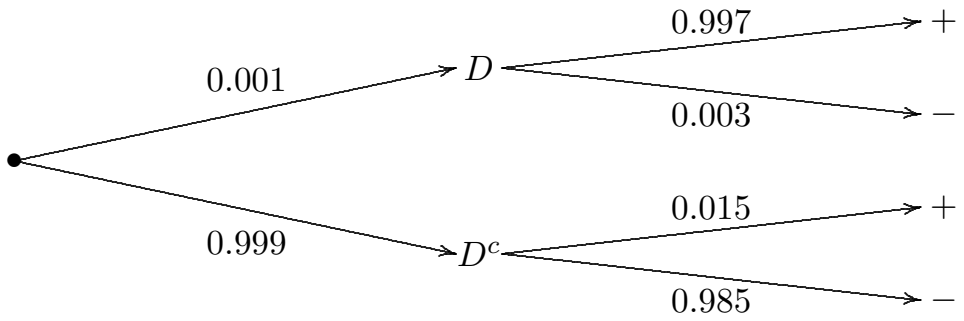

Figure 3.6: A tree diagram for Example 3.10.

To motivate our calculation of $P(D|+)$, it is again helpful to construct a tree diagram, as in Figure 3.6. This diagram was constructed so that the branches depicted in the tree have known probabilities, i.e., we first branch on the basis of disease state because $P(D)$ and $P(D^c)$ are known, then on the basis of test result because $P(+|D)$, $P(-|D)$, $P(+|D^c)$, and $P(-|D^c)$ are known. Notice that each of the four paths in the tree corresponds to exactly one of the four sets in (3.6). Furthermore, we can calculate the probability of

[3]See E.M. Sloan et al. (1991), "HIV Testing: State of the Art," *Journal of the American Medical Association*, 266:2861–2866.

each set by multiplying the probabilities that occur along its corresponding path:

$$\begin{aligned} P(D \cap +) &= P(D) \cdot P(+|D) = 0.001 \cdot 0.997, \\ P(D \cap -) &= P(D) \cdot P(-|D) = 0.001 \cdot 0.003, \\ P(D^c \cap +) &= P(D^c) \cdot P(+|D^c) = 0.999 \cdot 0.015, \\ P(D^c \cap -) &= P(D^c) \cdot P(-|D^c) = 0.999 \cdot 0.985. \end{aligned}$$

The predictive value of the test is now obtained by computing

$$\begin{aligned} P(D|+) &= \frac{P(D \cap +)}{P(+)} = \frac{P(D \cap +)}{P(D \cap +) + P(D^c \cap +)} \\ &= \frac{0.001 \cdot 0.997}{0.001 \cdot 0.997 + 0.999 \cdot 0.015} \doteq 0.0624. \end{aligned}$$

This probability may seem quite small, but consider that a positive test result can be obtained in two ways. If the person has the HIV virus, then a positive result is obtained with high probability, but very few people actually have the virus. If the person does not have the HIV virus, then a positive result is obtained with low probability, but so many people do not have the virus that the combined number of false positives is quite large relative to the number of true positives. This is a common phenomenon when screening for diseases.

The preceding calculations can be generalized and formalized in a formula known as Bayes Theorem; however, because such calculations will not play an important role in this book, we prefer to emphasize the use of tree diagrams to derive the appropriate calculations on a case-by-case basis.

Independence We now introduce a concept that is of fundamental importance in probability and statistics. The intuitive notion that we wish to formalize is the following:

> Two events are independent if the occurrence of either is unaffected by the occurrence of the other.

This notion can be expressed mathematically using the concept of conditional probability. Let A and B denote events and assume for the moment that the probability of each is strictly positive. If A and B are to be regarded as independent, then the occurrence of A is not affected by the occurrence of B. This can be expressed by writing

$$P(A|B) = P(A). \tag{3.7}$$

Similarly, the occurrence of B is not affected by the occurrence of A. This can be expressed by writing

$$P(B|A) = P(B). \tag{3.8}$$

Substituting the definition of conditional probability into (3.7) and multiplying by $P(B)$ leads to the equation

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

Substituting the definition of conditional probability into (3.8) and multiplying by $P(A)$ leads to the same equation. We take this equation, called the multiplication rule for independence, as a definition:

Definition 3.2 *Two events A and B are independent if and only if*

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

We proceed to explore some consequences of this definition.

Example 3.11 Notice that we did not require $P(A) > 0$ or $P(B) > 0$ in Definition 3.2. Suppose that $P(A) = 0$ or $P(B) = 0$, so that $P(A) \cdot P(B) = 0$. Because $A \cap B \subset A$, $P(A \cap B) \leq P(A)$; similarly, $P(A \cap B) \leq P(B)$. It follows that

$$0 \leq P(A \cap B) \leq \min(P(A), P(B)) = 0$$

and therefore that

$$P(A \cap B) = 0 = P(A) \cdot P(B).$$

Thus, if either of two events has probability zero, then the events are necessarily independent.

Example 3.12 Consider the disjoint events depicted in Figure 3.7 and suppose that $P(A) > 0$ and $P(B) > 0$. Are A and B independent? Many students instinctively answer that they are, but independence is very different from mutual exclusivity. In fact, if A occurs then B does not (and vice versa), so Figure 3.7 is actually a fairly extreme example of *dependent* events. This can also be deduced from Definition 3.2: $P(A) \cdot P(B) > 0$, but

$$P(A \cap B) = P(\emptyset) = 0$$

so A and B are not independent.

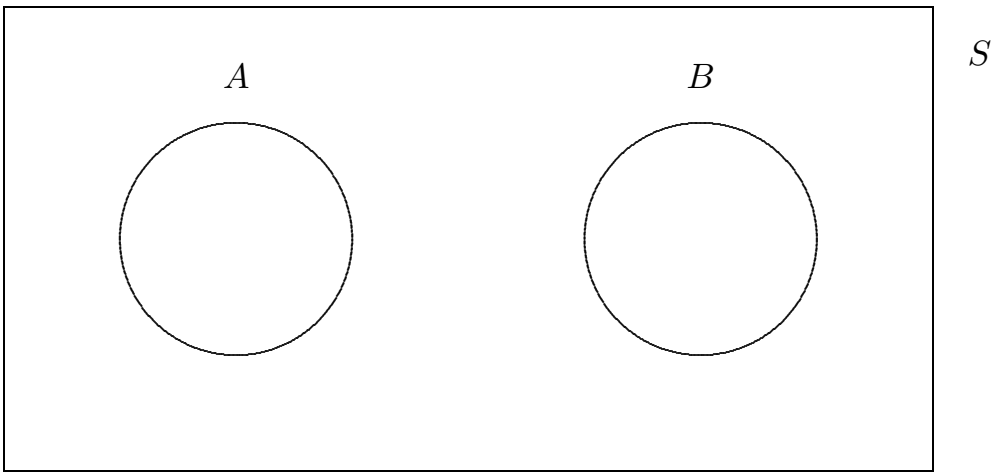

Figure 3.7: A Venn diagram for Example 3.12.

Example 3.13 *For each of the following, explain why the events A and B are or are not independent.*

(a) $P(A) = 0.4$, $P(B) = 0.5$, $P([A \cup B]^c) = 0.3$.

It follows that

$$P(A \cup B) = 1 - P\left([A \cup B]^c\right) = 1 - 0.3 = 0.7$$

and, because $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, that

$$P(A \cap B) = P(A) + P(B) - P(A \cup B) = 0.4 + 0.5 - 0.7 = 0.2.$$

Then, since

$$P(A) \cdot P(B) = 0.5 \cdot 0.4 = 0.2 = P(A \cap B),$$

it follows that A and B are independent events.

(b) $P(A \cap B^c) = 0.3$, $P(A^c \cap B) = 0.2$, $P(A^c \cap B^c) = 0.1$.

Refer to the Venn diagram in Figure 3.8 to see that

$$P(A) \cdot P(B) = 0.7 \cdot 0.6 = 0.42 \neq 0.40 = P(A \cap B)$$

and hence that A and B are dependent events.

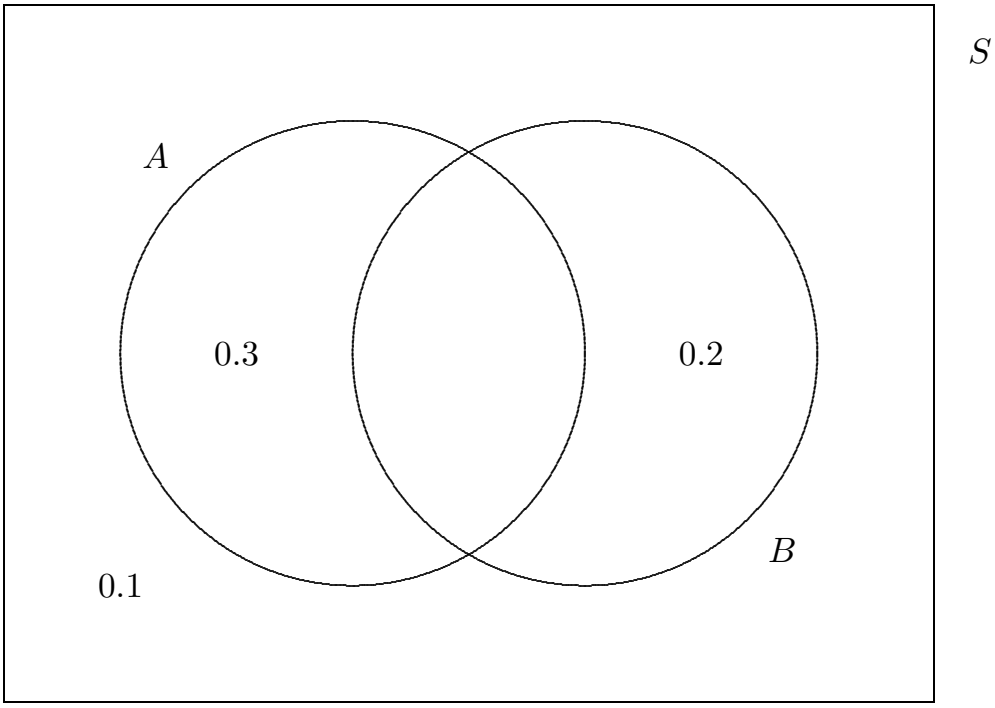

Figure 3.8: A Venn diagram for Example 3.13.

Thus far we have verified that two events are independent by verifying that the multiplication rule for independence holds. In applications, however, we usually reason somewhat differently. Using our *intuitive* notion of independence, we appeal to common sense, our knowledge of science, etc., to decide if independence is a property that we wish to incorporate into our mathematical model of the experiment in question. If it is, then we *assume* that two events are independent and the multiplication rule for independence becomes available to us for use as a computational formula.

Example 3.14 *Consider an experiment in which a typical penny is first tossed, then spun. Let A denote the event that the toss results in* `Heads` *and let B denote the event that the spin results in* `Heads`*. What is the probability of observing two* `Heads`*?*

We assume that, for a typical penny, $P(A) = 0.5$ and $P(B) = 0.3$ (see Section 1.1.1). Common sense tells us that the occurrence of either event is unaffected by the occurrence of the other. (Time is not reversible, so obviously the occurrence of A is not affected by the occurrence of B. One

might argue that tossing the penny so that A occurs results in wear that is slightly different than the wear that results if A^c occurs, thereby slightly affecting the subsequent probability that B occurs. However, this argument strikes most students as completely preposterous. Even if it has a modicum of validity, the effect is undoubtedly so slight that we can safely neglect it in constructing our mathematical model of the experiment.) Therefore, we *assume* that A and B are independent and calculate that

$$P(A \cap B) = P(A) \cdot P(B) = 0.5 \cdot 0.3 = 0.15.$$

Example 3.15 *For each of the following, explain why the events A and B are or are not independent.*

(a) *Consider the population of William & Mary undergraduate students, from which one student is selected at random. Let A denote the event that the student is female and let B denote the event that the student is concentrating in elementary education.*

I'm told that $P(A)$ is roughly 60 percent, while it appears to me that $P(A|B)$ exceeds 90 percent. Whatever the exact probabilities, it is evident that the probability that a random elementary education concentrator is female is considerably greater than the probability that a random student is female. Hence, A and B are dependent events.

(b) *Consider the population of registered voters, from which one voter is selected at random. Let A denote the event that the voter belongs to a country club and let B denote the event that the voter is a Republican.*

It is generally conceded that one finds a greater proportion of Republicans among the wealthy than in the general population. Since one tends to find a greater proportion of wealthy persons at country clubs than in the general population, it follows that the probability that a random country club member is a Republican is greater than the probability that a randomly selected voter is a Republican. Hence, A and B are dependent events.[4]

[4]This phenomenon may seem obvious, but it was overlooked by the respected *Literary Digest* poll. Their embarrassingly awful prediction of the 1936 presidential election resulted in the previously popular magazine going out of business. George Gallup's relatively accurate prediction of the outcome (and his uncannily accurate prediction of what the *Literary Digest* poll would predict) revolutionized polling practices.

Before progressing further, we ask what it should mean for A, B, and C to be three *mutually independent* events. Certainly each pair should comprise two independent events, but we would also like to write

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

It turns out that this equation cannot be deduced from the pairwise independence of A, B, and C, so we have to include it in our definition of mutual independence. Similar equations must be included when defining the mutual independence of more than three events. Here is a general definition:

Definition 3.3 *Let $\{A_\alpha\}$ be an arbitrary collection of events. These events are mutually independent if and only if, for every finite choice of events $A_{\alpha_1}, \ldots, A_{\alpha_k}$,*

$$P\left(A_{\alpha_1} \cap \cdots \cap A_{\alpha_k}\right) = P\left(A_{\alpha_1}\right) \cdots P\left(A_{\alpha_k}\right).$$

Example 3.16 In the preliminary hearing for the criminal trial of O.J. Simpson, the prosecution presented conventional blood-typing evidence that blood found at the murder scene possessed three characteristics also possessed by Simpson's blood. The prosecution also presented estimates of the prevalence of each characteristic in the general population, i.e., of the probabilities that a person selected at random from the general population would possess these characteristics. Then, to obtain the estimated probability that a randomly selected person would possess all three characteristics, the prosecution multiplied the three individual probabilities, resulting in an estimate of 0.005.

In response to this evidence, defense counsel Gerald Uehlman objected that the prosecution had not established that the three events in question were independent and therefore had not justified their use of the multiplication rule. The prosecution responded that it was standard practice to multiply such probabilities and Judge Kennedy-Powell admitted the 0.005 estimate on that basis. No attempt was made to assess whether or not the standard practice was proper; it was inferred from the fact that the practice was standard that it must be proper. In this example, science and law diverge. From a scientific perspective, Gerald Uehlman was absolutely correct in maintaining that an assumption of independence must be justified.

3.5 Random Variables

Informally, a *random variable* is a rule for assigning real numbers to experimental outcomes. By convention, random variables are usually denoted by upper case Roman letters near the end of the alphabet, e.g., X, Y, Z.

Example 3.17 *A coin is tossed once and* `Heads` *(*`H`*) or* `Tails` *(*`T`*) is observed.*

The sample space for this experiment is $S = \{\texttt{H}, \texttt{T}\}$. For reasons that will become apparent, it is often convenient to assign the real number 1 to `Heads` and the real number 0 to `Tails`. This assignment, which we denote by the random variable X, can be depicted as follows:

$$\boxed{\begin{array}{c}\texttt{H}\\ \texttt{T}\end{array}} \xrightarrow{X} \boxed{\begin{array}{c}1\\ 0\end{array}}$$

In functional notation, $X : S \to \Re$ and the rule of assignment is defined by

$$X(\texttt{H}) = 1,$$
$$X(\texttt{T}) = 0.$$

Example 3.18 *A coin is tossed twice and the number of* `Heads` *is counted.*

The sample space for this experiment is $S = \{\texttt{HH}, \texttt{HT}, \texttt{TH}, \texttt{TT}\}$. We want to assign the real number 2 to the outcome `HH`, the real number 1 to the outcomes `HT` and `TH`, and the real number 0 to the outcome `TT`. Several representations of this assignment are possible:

(a) Direct assignment, which we denote by the random variable Y, can be depicted as follows:

$$\boxed{\begin{array}{cc}\texttt{HH} & \texttt{HT}\\ \texttt{TH} & \texttt{TT}\end{array}} \xrightarrow{Y} \boxed{\begin{array}{cc}2 & 1\\ 1 & 0\end{array}}$$

In functional notation, $Y : S \to \Re$ and the rule of assignment is defined by

$$Y(\texttt{HH}) = 2,$$
$$Y(\texttt{HT}) = Y(\texttt{TH}) = 1,$$
$$Y(\texttt{TT}) = 0.$$

(b) Instead of directly assigning the counts, we might take the intermediate step of assigning an ordered pair of numbers to each outcome. As in

Example 3.17, we assign 1 to each occurence of `Heads` and 0 to each occurence of `Tails`. We denote this assignment by $X : S \to \Re^2$. In this context, $X = (X_1, X_2)$ is called a *random vector*. Each component of the random vector X is a random variable.

Next, we define a function $g : \Re^2 \to \Re$ by

$$g(x_1, x_2) = x_1 + x_2.$$

The composition $g(X)$ is equivalent to the random variable Y, as revealed by the following depiction:

$$\boxed{\begin{array}{cc} \texttt{HH} & \texttt{HT} \\ \texttt{TH} & \texttt{TT} \end{array}} \xrightarrow{X} \boxed{\begin{array}{cc} (1,1) & (1,0) \\ (0,1) & (0,0) \end{array}} \xrightarrow{g} \boxed{\begin{array}{cc} 2 & 1 \\ 1 & 0 \end{array}}$$

(c) The preceding representation suggests defining two random variables, X_1 and X_2, as in the following depiction:

$$\boxed{\begin{array}{cc} 1 & 1 \\ 0 & 0 \end{array}} \xleftarrow{X_1} \boxed{\begin{array}{cc} \texttt{HH} & \texttt{HT} \\ \texttt{TH} & \texttt{TT} \end{array}} \xrightarrow{X_2} \boxed{\begin{array}{cc} 1 & 0 \\ 1 & 0 \end{array}}$$

As in the preceding representation, the random variable X_1 counts the number of `Heads` observed on the first toss and the random variable X_2 counts the number of `Heads` observed on the second toss. The sum of these random variables, $X_1 + X_2$, is evidently equivalent to the random variable Y.

The primary reason that we construct a random variable, X, is to replace the probability space that is naturally suggested by the experiment in question with a familiar probability space in which the possible outcomes are real numbers. Thus, we replace the original sample space, S, with the familiar number line, $\Re$. To complete the transference, we must decide which subsets of $\Re$ will be designated as events and we must specify how the probabilities of these events are to be calculated.

It is an interesting fact that it is impossible to construct a probability space in which the set of outcomes is $\Re$ and every subset of $\Re$ is an event. For this reason, we define the collection of events to be the smallest collection of subsets that satisfies the assumptions of the Kolmogorov probability model and that contains every interval of the form $(-\infty, y]$. This collection is called the *Borel sets* and it is a very large collection of subsets of $\Re$. In particular, it contains every interval of real numbers and every set that can

be constructed by applying a countable number of set operations (union, intersection, complementation) to intervals. Most students will never see a set that is not a Borel set!

Finally, we must define a probability measure that assigns probabilities to Borel sets. Of course, we want to do so in a way that preserves the probability structure of the experiment in question. The only way to do so is to define the probability of each Borel set B to be the probability of the set of outcomes to which X assigns a value in B. This set of outcomes is denoted by

$$X^{-1}(B) = \{s \in S : X(s) \in B\}$$

and is depicted in Figure 3.9.

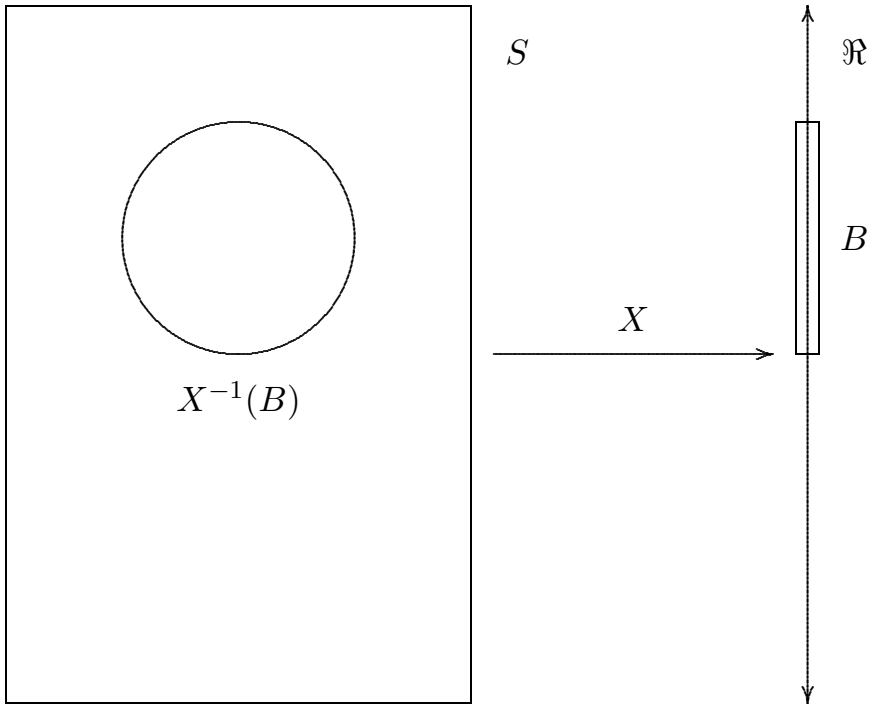

Figure 3.9: The inverse image of a Borel set.

How do we know that the set of outcomes to which X assigns a value in B is an event and therefore has a probability? We don't, so we guarantee that it is by including this requirement in our formal definition of a random variable.

Definition 3.4 *A function $X : S \to \Re$ is a random variable if and only if*

$$P\left(\{s \in S : X(s) \leq y\}\right)$$

exists for all choices of $y \in \Re$.

We will denote the probability measure induced by the random variable X by P_X. The following equation defines various representations of P_X:

$$\begin{aligned} P_X\left((-\infty, y]\right) &= P\left(X^{-1}\left((-\infty, y]\right)\right) \\ &= P\left(\{s \in S : X(s) \in (-\infty, y]\}\right) \\ &= P\left(-\infty < X \leq y\right) \\ &= P\left(X \leq y\right) \end{aligned}$$

A probability measure on the Borel sets is called a *probability distribution* and P_X is called the distribution of the random variable X. A hallmark feature of probability theory is that we study the distributions of random variables rather than arbitrary probability measures. One important reason for this emphasis is that many different experiments may result in identical distributions. For example, the random variable in Example 3.17 might have the same distribution as a random variable that assigns 1 to male newborns and 0 to female newborns.

Cumulative Distribution Functions Our construction of the probability measure induced by a random variable suggests that the following function will be useful in describing the properties of random variables.

Definition 3.5 *The cumulative distribution function (cdf) of a random variable X is the function $F : \Re \to \Re$ defined by*

$$F(y) = P(X \leq y).$$

Example 3.17 (continued) We consider two probability structures that might obtain in the case of a typical penny.

(a) *A typical penny is tossed.*

For this experiment, $P(\texttt{H}) = P(\texttt{T}) = 0.5$, and the following values of the cdf are easily determined:

- If $y < 0$, e.g., $y = -9.1185$ or $y = -0.3018$, then
$$F(y) = P(X \leq y) = P(\emptyset) = 0.$$
- $F(0) = P(X \leq 0) = P(\{\texttt{T}\}) = 0.5$.
- If $y \in (0, 1)$, e.g., $y = 0.6241$ or $y = 0.9365$, then
$$F(y) = P(X \leq y) = P(\{\texttt{T}\}) = 0.5.$$
- $F(1) = P(X \leq 1) = P(\{\texttt{T}, \texttt{H}\}) = 1$.
- If $y > 1$, e.g., $y = 1.5248$ or $y = 7.7397$, then
$$F(y) = P(X \leq y) = P(\{\texttt{T}, \texttt{H}\}) = 1.$$

The entire cdf is plotted in Figure 3.10.

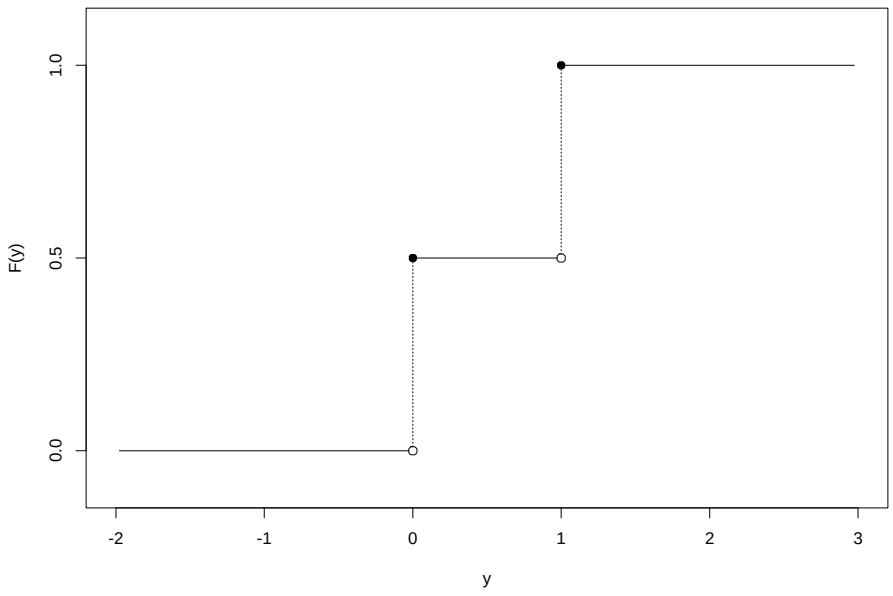

Figure 3.10: The cumulative distribution function for tossing a penny with $P(\texttt{Heads}) = 0.5$.

(b) *A typical penny is spun.*

For this experiment, we assume that $P(\texttt{H}) = 0.3$ and $P(\texttt{T}) = 0.7$ (see Section 1.1.1). Then the following values of the cdf are easily determined:

- If $y < 0$, e.g., $y = -1.6633$ or $y = -0.5485$, then

$$F(y) = P(X \leq y) = P(\emptyset) = 0.$$

- $F(0) = P(X \leq 0) = P(\{\texttt{T}\}) = 0.7$.
- If $y \in (0, 1)$, e.g., $y = 0.0685$ or $y = 0.4569$, then

$$F(y) = P(X \leq y) = P(\{\texttt{T}\}) = 0.7.$$

- $F(1) = P(X \leq 1) = P(\{\texttt{T}, \texttt{H}\}) = 1$.
- If $y > 1$, e.g., $y = 1.4789$ or $y = 2.6117$, then

$$F(y) = P(X \leq y) = P(\{\texttt{T}, \texttt{H}\}) = 1.$$

The entire cdf is plotted in Figure 3.11.

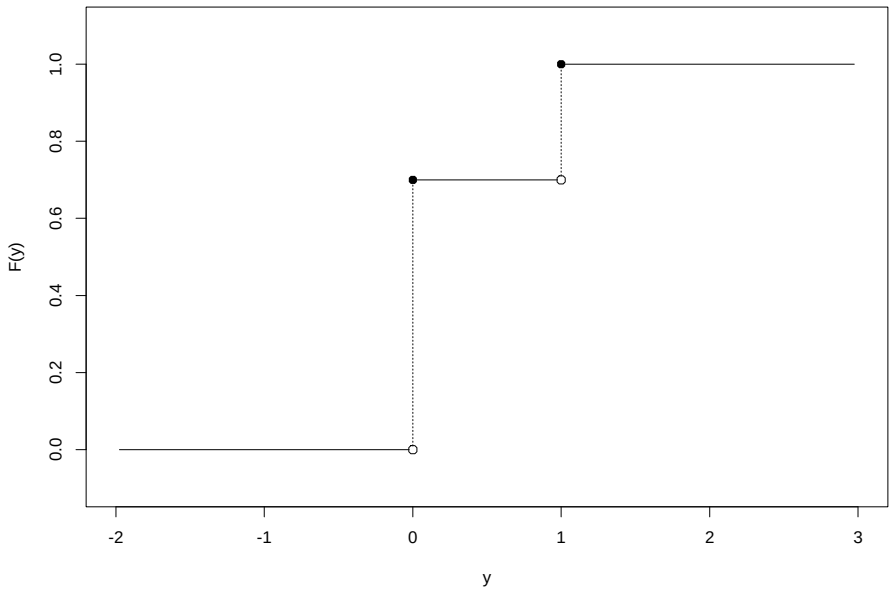

Figure 3.11: The cumulative distribution function for spinning a penny with $P(\texttt{Heads}) = 0.3$.

Example 3.18 (continued) Suppose that the coin is fair, so that each of the four possible outcomes in S is equally likely, i.e., has probability 0.25. Then the following values of the cdf are easily determined:

- If $y < 0$, e.g., $y = -4.2132$ or $y = -0.5615$, then
$$F(y) = P(X \leq y) = P(\emptyset) = 0.$$
- $F(0) = P(X \leq 0) = P(\{\texttt{TT}\}) = 0.25$.
- If $y \in (0, 1)$, e.g., $y = 0.3074$ or $y = 0.6924$, then
$$F(y) = P(X \leq y) = P(\{\texttt{TT}\}) = 0.25.$$
- $F(1) = P(X \leq 1) = P(\{\texttt{TT}, \texttt{HT}, \texttt{TH}\}) = 0.75$.
- If $y \in (1, 2)$, e.g., $y = 1.4629$ or $y = 1.5159$, then
$$F(y) = P(X \leq y) = P(\{\texttt{TT}, \texttt{HT}, \texttt{TH}\}) = 0.75.$$
- $F(2) = P(X \leq 2) = P(\{\texttt{TT}, \texttt{HT}, \texttt{TH}, \texttt{HH}\}) = 1$.
- If $y > 2$, e.g., $y = 2.1252$ or $y = 3.7790$, then
$$F(y) = P(X \leq y) = P(\{\texttt{TT}, \texttt{HT}, \texttt{TH}, \texttt{HH}\}) = 1.$$

The entire cdf is plotted in Figure 3.12.

Let us make some observations about the cdfs that we have plotted. First, each cdf assumes its values in the unit interval, $[0, 1]$. This is a general property of cdfs: each $F(y) = P(X \leq y)$, and probabilities necessarily assume values in $[0, 1]$.

Second, each cdf is nondecreasing; i.e., if $y_2 > y_1$, then $F(y_2) \geq F(y_1)$. This is also a general property of cdfs, for suppose that we observe an outcome s such that $X(s) \leq y_1$. Because $y_1 < y_2$, it follows that $X(s) \leq y_2$. Thus, $\{X \leq y_1\} \subset \{X \leq y_2\}$ and therefore

$$F(y_1) = P(X \leq y_1) \leq P(X \leq y_2) = F(y_2).$$

Finally, each cdf equals 1 for sufficiently large y and 0 for sufficiently small y. This is *not* a general property of cdfs—it occurs in our examples because $X(S)$ is a bounded set, i.e., there exist finite real numbers a and b such that every $x \in X(S)$ satisfies $a \leq x \leq b$. However, all cdfs do satisfy the following properties:

$$\lim_{y \to \infty} F(y) = 1 \quad \text{and} \quad \lim_{y \to -\infty} F(y) = 0.$$

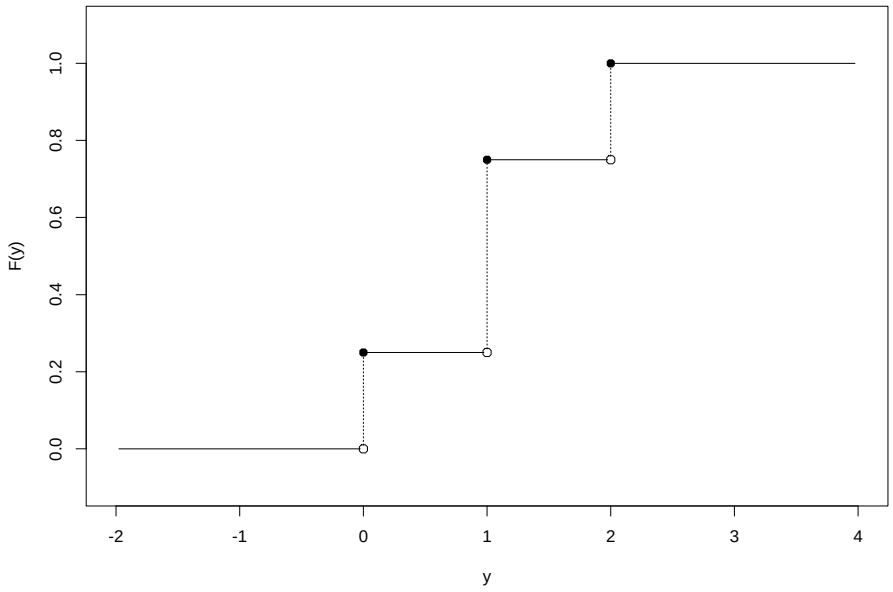

Figure 3.12: The cumulative distribution function for tossing two pennies with $P(\texttt{Heads}) = 0.5$ and counting the number of `Heads`.

Independence We say that two random variables, X_1 and X_2, are independent if each event defined by X_1 is independent of each event defined by X_2. More precisely,

Definition 3.6 *Let $X_1 : S \to \Re$ and $X_2 : S \to \Re$ be random variables. X_1 and X_2 are independent if and only if, for each $y_1 \in \Re$ and each $y_2 \in \Re$,*

$$P(X_1 \leq y_1, X_2 \leq y_2) = P(X_1 \leq y_1) \cdot P(X_2 \leq y_2).$$

This definition can be extended to mutually independent collections of random variables in precisely the same way that we extended Definition 3.2 to Definition 3.3.

Intuitively, two random variables are independent if the distribution of either does not depend on the value of the other. As we discussed in Section 3.4, in most applications we will appeal to common sense, our knowledge of science, etc., to decide if independence is a property that we wish to incorporate into our mathematical model of the experiment in question. If it is, then we will *assume* that the appropriate random variables are independent. This

assumption will allow us to apply many powerful theorems from probability and statistics that are only true of independent random variables.

3.6 Case Study: Padrolling in Milton Murayama's *All I asking for is my body*

The American dice game Craps evolved from the English dice game Hazard:

> "According to tradition, blacks living around New Orleans tried their hand at Hazard... In the course of time they modifed the rules and playing procedures so greatly that they ended up inventing the game of Craps (in the U.S. idiom known as Crapshooting or Shooting Craps and here identified as Private Craps to distinguish it from Open Craps and the more formalized variants offered in gambling casinos). ... The popularity of the private game of Craps with the U.S. military personnel during World Wars I and II helped to spread that game to many parts of the world."[5]

Craps is played with two fair dice, each marked in a specific way. According to Hoyle,

> Each face of [each] die is marked with one to six dots, opposite faces representing... numbers adding to seven; if the vertical face toward you is 5, and the horizontal face on top of the die is 6, [then] the 3 should be on the vertical face to your right."[6]

The *shooter* rolls the pair of dice, resulting in one of $6 \times 6 = 36$ possible outcomes. Of interest is the combined number of dots on the horizontal faces atop the two dice, a number that we denote by the random variable X. The possible values of X are displayed in Figure 3.13.

Let x denote the value of X produced by the first roll. The game ends immediately if $x \in \{2, 3, 7, 11, 12\}$. If $x \in \{7, 11\}$, then x is a *natural* and the shooter wins; if $x \in \{2, 3, 12\}$, then x is *craps* and the shooter loses; otherwise, x becomes the shooter's *point*. If the first roll is not decisive, then the shooter continues to roll until he either (a) again rolls x (*makes*

[5] "Dice and dice games," *The New Encyclopædia Britannica in 30 Volumes*, Macropædia, Volume 5, 1974, pp. 702–706.

[6] Richard L. Frey, *According to Hoyle*, Fawcett Publications, 1970, p. 266.

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Figure 3.13: The possible outcomes of rolling two standard dice.

his point), in which case he wins, or (b) rolls 7 (*craps out*), in which case he loses.

A game of craps is fair when each of the 36 outcomes in Figure 3.13 is equally likely. Fairness is usually ensured by tossing the dice from a cup, or, more crudely, by tossing them against a wall. In a fair game of craps, we have the following probabilities:

$$\begin{aligned} P(X=7) &= 6/36 \\ P(X=6)=P(X=8) &= 5/36 \\ P(X=5)=P(X=9) &= 4/36 \\ P(X=4)=P(X=10) &= 3/36 \\ P(X=3)=P(X=11) &= 2/36 \\ P(X=2)=P(X=12) &= 1/36 \end{aligned}$$

Let us begin by calculating the probability that the shooter wins a fair game of craps.

There are several ways for the shooter to win. We will calculate the probability of each, then sum these probabilities.

- Roll a natural.
$$P(X \in \{7,11\}) = \frac{6+2}{36} = \frac{2}{3^2}.$$
- Roll $x = 6$ or $x = 8$, then make point.

  First,
$$P(X \in \{6,8\}) = \frac{5+5}{36} = \frac{5}{18}.$$

Then, the shooter must roll x before rolling 7. Other outcomes are ignored. There are 5 ways to roll x versus 6 ways to roll 7, so the conditional probability of making point is 5/11. Hence, the probability of the shooter winning in this way is

$$\frac{5}{18} \cdot \frac{5}{11} = \frac{25}{2 \cdot 3^2 \cdot 11}.$$

- Roll $x = 5$ or $x = 9$, then make point.

  First,

  $$P(X \in \{5, 9\}) = \frac{4+4}{36} = \frac{2}{9}.$$

  Then, the shooter must roll x before rolling 7. Other outcomes are ignored. There are 4 ways to roll x versus 6 ways to roll 7, so the conditional probability of making point is 4/10. Hence, the probability of the shooter winning in this way is

  $$\frac{2}{9} \cdot \frac{4}{10} = \frac{4}{3^2 \cdot 5}.$$

- Roll $x = 4$ or $x = 10$, then make point.

  First,

  $$P(X \in \{4, 10\}) = \frac{3+3}{36} = \frac{1}{6}.$$

  Then, the shooter must roll x before rolling 7. Other outcomes are ignored. There are 3 ways to roll x versus 6 ways to roll 7, so the conditional probability of making point is 3/9. Hence, the probability of the shooter winning in this way is

  $$\frac{1}{6} \cdot \frac{3}{9} = \frac{1}{2 \cdot 3^2}.$$

The probability that the shooter wins is

$$\frac{2}{3^2} + \frac{25}{2 \cdot 3^2 \cdot 11} + \frac{4}{3^2 \cdot 5} + \frac{1}{2 \cdot 3^2} = \frac{244}{495} \doteq 0.4929.$$

Thus, the shooter is slightly more likely to lose than to win a fair game of craps.

Milton Murayama's 1959 novel, *All I asking for is my body*, is a brilliant evocation of *nisei* (second-generation Japanese American) life on Hawaiian

sugar plantations in the 1930s.[7] One of its central concerns is the concept of Japanese honor and its implicatons for the young protagonist/narrator, Kiyoshi, and his siblings. Years earlier, Kiyoshi's parents had sacrificed their future to pay Kiyoshi's grandfather's debts; now they owe the impossible sum of $6000 and they expect their children to do likewise. Toward the novel's end, Japan attacks Pearl Harbor and Kiyoshi subsequently volunteers for an all-*nisei* regiment that will fight in Europe. In the final chapter, he contrives to win $6000 by playing Craps.

Kiyoshi had watched a former classmate, Hiroshi Sakai, play Craps at the Citizens' Quarters in Kahana.

> "It was weird the way he kept winning. Whenever he rolled, the dice rolled in unison like the wheels of a cart, and even when one die rolled ahead of the other, neither flipped on its side. The Kahana players finally refused to fade [bet against] him, and he stopped coming."

We subsequently learn that Hiroshi's technique is called *padrolling*.

In the Army,

> "Everybody had money and every third guy was a crapshooter. The sight of all that money drove me mad. There was $25,000 at least floating around in the crap games.... Most of the games were played on blankets on barrack floors, the dice rolled by hand. There were a few guys who rolled the dice the way Hiroshi did at the Citizens' Quarters in Kahana. The dice didn't bounce but rolled out in unison like the wheels of a cart. There had to be an advantage to that."

Kiyoshi buys a pair of dice and examines them carefully. He realizes that, by rolling the dice "like the wheels of a cart," he can keep the sides of the dice that form the axis of the wheels from appearing. Then, by combining certain numbers to form the axis, he can improve his chance of winning.

Kiyoshi teaches himself to padroll and develops the following system for choosing the axis:

1. For the initial roll, use the 1-6 axis for each die.

   Padrolling this axis has the effect of eliminating the first and sixth rows and columns in Figure 3.13, resulting in the following set of possible

[7] I am indebted to M. Lynn Weiss for bringing this novel to my attention.

outcomes:

	2	3	4	5
2	4	5	6	7
3	5	6	7	8
4	6	7	8	9
5	7	8	9	10

Notice that this choice eliminates the possibility of crapping out! Furthermore, assuming that the 16 remaining outcomes are equally likely, it also improves the chance of rolling a natural from 4/18 to 4/16.

2. If $x \in \{6, 8\}$, then use the 1-6 axis on one die and the 2-5 axis on the other.

   Padrolling this axis results in the following set of possible outcomes:

	1	3	4	6
2	3	5	6	8
3	4	6	7	9
4	5	7	8	10
5	6	8	9	11

   With this choice, there are 3 ways to roll x versus 2 ways to roll 7. Again assuming that the 16 remaining outcomes are equally likely, this choice improves the conditional probability of making point from 5/11 to 3/5.

3. If $x \in \{4, 5, 9, 10\}$, then use the 1-6 axis on one die and the 3-4 axis on the other.

   Padrolling this axis results in the following set of possible outcomes:

	1	2	5	6
2	3	4	7	8
3	4	5	8	9
4	5	6	9	10
5	6	7	10	11

   With this choice, there are 2 ways to roll x versus 2 ways to roll 7. Again, assume that the 16 remaining outcomes are equally likely. If $x \in \{5, 9\}$, then this choice improves the conditional probability of making point from 4/10 to 2/4. If $x \in \{4, 10\}$, then this choice improves the conditional probability of making point from 3/9 to 2/4.

If a shooter padrolls successfully, then the probability that he will win using Kiyoshi's system is

$$\frac{4}{16} + \frac{6}{16} \cdot \frac{3}{5} + \frac{4}{16} \cdot \frac{2}{4} + \frac{2}{16} \cdot \frac{2}{4} = \frac{53}{80} = 0.6625,$$

a substantial improvement on his chance of winning a fair game. "And," Kiyoshi rationalizes, "it wasn't really cheating. The others had the option of stopping any of your rolls, or they could play with a cup, or have the roller bang the dice against the wall, or use a canvas or the bare floor instead of a blanket." So, Kiyoshi padrolls. I leave to my readers the pleasure of discovering whether or not he succeeds in winning the $6000 his family needs.

3.7 Exercises

1. Consider three events that might occur when a new mine is dug in the Cleveland National Forest in San Diego County, California:

   $$\begin{array}{lcl} \text{A} & = & \{ \text{ quartz specimens are found } \} \\ \text{B} & = & \{ \text{ tourmaline specimens are found } \} \\ \text{C} & = & \{ \text{ aquamarine specimens are found } \} \end{array}$$

   Assume the following probabilities: $P(A) = 0.80$, $P(B) = 0.36$, $P(C) = 0.28$, $P(A \cap B) = 0.29$, $P(A \cap C) = 0.24$, $P(B \cap C) = 0.16$, and $P(A \cap B \cap C) = 0.13$.

   (a) Draw a suitable Venn diagram for this situation.

   (b) Calculate the probability that both quartz and tourmaline will be found, but not aquamarine.

   (c) Calculate the probability that quartz will be found, but not tourmaline or aquamarine.

   (d) Calculate the probability that none of these types of specimens will be found.

   (e) Calculate the probability of $A^c \cap (B \cup C)$.

2. Consider two urns, one containing four tickets labelled $\{1, 3, 4, 6\}$; the other containing ten tickets, labelled $\{1, 3, 3, 3, 3, 4, 4, 4, 4, 6\}$.

   (a) What is the probability of drawing a 3 from the first urn?

(b) What is the probability of drawing a 3 from the second urn?

(c) Which urn is a better model for throwing an astragalus? Why?

3. Suppose that five cards are dealt from a standard deck of playing cards.

   (a) What is the probability of drawing a straight flush?

   (b) What is the probability of drawing 4 of a kind?

   Hint: Use the results of Exercise 2.5.6.

4. Suppose that four fair dice are thrown simultaneously.

   (a) How many outcomes are possible?

   (b) What is the probability that each top face shows a different number?

   (c) What is the probability that the top faces show four numbers that sum to five?

   (d) What is the probability that at least one of the top faces shows an odd number?

   (e) What is the probability that three of the top faces show the same odd number and the other top face shows an even number?

5. A *dreidl* is a four-sided top that contains a Hebrew letter on each side: nun, gimmel, heh, shin. These letters are an acronym for the Hebrew phrase *nes gadol hayah sham* (a great miracle happened there), which refers to the miracle of the temple light that burned for eight days with only one day's supply of oil—the miracle celebrated at Chanukah. Here we suppose that a fair dreidl (one that is equally likely to fall on each of its four sides) is to be spun ten times. Compute the probability of each of the following events:

   (a) Five gimmels and five hehs;

   (b) No nuns or shins;

   (c) Two letters are absent and two letters are present;

   (d) At least two letters are absent.

6. Suppose that $P(A) = 0.7$, $P(B) = 0.6$, and $P(A^c \cap B) = 0.2$.

   (a) Draw a Venn diagram that describes this experiment.

(b) Is it possible for A and B to be disjoint events? Why or why not?

(c) What is the probability of $A \cup B^c$?

(d) Is it possible for A and B to be independent events? Why or why not?

(e) What is the conditional probability of A given B?

7. Suppose that 20 percent of the adult population is hypertensive. Suppose that an automated blood-pressure machine diagnoses 84 percent of hypertensive adults as hypertensive and 23 percent of nonhypertensive adults as hypertensive. A person is selected at random from the adult population.

(a) Construct a tree diagram that describes this experiment.

(b) What is the probability that the automated blood-pressure machine will diagnose the selected person as hypertensive?

(c) Suppose that the automated blood-pressure machine does diagnose the selected person as hypertensive. What then is the probability that this person actually is hypertensive?

(d) The following passage appeared in a recent article (Bruce Bower, Roots of reason, *Science News*, 145:72–75, January 29, 1994) about how human beings think. Please comment on it in whatever way seems appropriate to you.

> And in a study slated to appear in COGNITION, Cosmides and Tooby confront a cognitive bias known as the "base-rate fallacy." As an illustration, they cite a 1978 study in which 60 staff and students at Harvard Medical School attempted to solve this problem: "If a test to detect a disease whose prevalence is 1/1,000 has a false positive rate of 5%, what is the chance that a person found to have a positive result actually has the disease, assuming you know nothing about the person's symptoms or signs?"
>
> Nearly half the sample estimated this probability as 95 percent; only 11 gave the correct response of 2 percent. Most participants neglected the base rate of the disease (it strikes 1 in 1,000 people) and formed a judgment solely from the characteristics of the test.

8. Mike owns a box that contains 6 pairs of 14-carat gold, cubic zirconia earrings. The earrings are of three sizes: 3mm, 4mm, and 5mm. There are 2 pairs of each size.

   Each time that Mike needs an inexpensive gift for a female friend, he randomly selects a pair of earrings from the box. If the selected pair is 4mm, then he buys an identical pair to replace it. If the selected pair is 3mm, then he does not replace it. If the selected pair is 5mm, then he tosses a fair coin. If he observes `Heads`, then he buys two identical pairs of earrings to replace the selected pair; if he observes `Tails`, then he does not replace the selected pair.

   (a) What is the probability that the second pair selected will be 4mm?

   (b) If the second pair was not 4mm, then what is the probability that the first pair was 5mm?

9. The following puzzle was presented on National Public Radio's *Car Talk*:

   > RAY: Three different numbers are chosen at random, and one is written on each of three slips of paper. The slips are then placed face down on the table. The objective is to choose the slip upon which is written the largest number.
   >
   > Here are the rules: You can turn over any slip of paper and look at the amount written on it. If for any reason you think this is the largest, you're done; you keep it. Otherwise you discard it and turn over a second slip. Again, if you think this is the one with the biggest number, you keep that one and the game is over. If you don't, you discard that one too.
   >
   > TOM: And you're stuck with the third. I get it.
   >
   > RAY: The chance of getting the highest number is one in three. Or is it? Is there a strategy by which you can improve the odds?

   Solve the puzzle, i.e., determine an optimal strategy for finding the highest number. What is the probability that your strategy will find the highest number? Explain your answer.

10. It is a curious fact that approximately 85% of all U.S. residents who are struck by lightning are men. Consider the population of U.S. residents,

from which a person is randomly selected. Let A denote the event that the person is male and let B denote the event that the person will be struck by lightning.

(a) Estimate $P(A|B)$ and $P(A^c|B)$.

(b) Compare $P(A|B)$ and $P(A)$. Are A and B independent events?

(c) Suggest reasons why $P(A|B)$ is so much larger than $P(A^c|B)$. It is tempting to joke that men don't know enough to come in out of the rain! Why might there be some truth to this possibility, i.e., why might men be more reluctant to take precautions than women? Can you suggest other explanations?

11. For each of the following pairs of events, explain why A and B are dependent or independent.

(a) Consider the population of U.S. citizens, from which a person is randomly selected. Let A denote the event that the person is a member of a chess club and let B denote the event that the person is a woman.

(b) Consider the population of male U.S. citizens who are 30 years of age. A man is selected at random from this population. Let A denote the event that he will be bald before reaching 40 years of age and let B denote the event that his father went bald before reaching 40 years of age.

(c) Consider the population of students who attend high school in the U.S. A student is selected at random from this population. Let A denote the event that the student speaks Spanish and let B denote the event that the student lives in Texas.

(d) Consider the population of months in the 20th century. A month is selected at random from this population. Let A denote the event that a hurricane crossed the North Carolina coastline during this month and let B denote the event that it snowed in Denver, Colorado, during this month.

(e) Consider the population of Hollywood feature films produced during the 20th century. A movie is selected at random from this population. Let A denote the event that the movie was filmed in color and let B denote the event that the movie is a western.

12. Two graduate students are renting a house. Before leaving town for winter break, each writes a check for her share of the rent. Emilie writes her check on December 16. By chance, it happens that the number of her check ends with the digits 16. Anne writes her check on December 18. By chance, it happens that the number of her check ends with the digits 18. What is the probability of such a coincidence, i.e., that both students would use checks with numbers that end in the same two digits as the date?

13. Suppose that X is a random variable with cdf

$$F(y) = \left\{ \begin{array}{cc} 0 & y \leq 0 \\ y/3 & y \in [0,1) \\ 2/3 & y \in [1,2] \\ y/3 & y \in [2,3] \\ 1 & y \geq 3 \end{array} \right\}.$$

Graph F and compute the following probabilities:

(a) $P(X > 0.5)$

(b) $P(2 < X \leq 3)$

(c) $P(0.5 < X \leq 2.5)$

(d) $P(X = 1)$

14. In Section 3.6, we calculated the probability that the shooter will win a fair game of craps. In so doing, we glossed a subtle point.

Suppose that the shooter's first roll results in $x = 8$. Now the shooter must roll until he rolls another 8, in which cases he makes his point and wins, or until he rolls a 7, in which case he craps out and loses. We argued that "there are 5 ways to roll 8 versus 6 ways to roll 7, so the conditional probability of making point is 5/11." This argument appears to ignore the possibility that the shooter might roll indefinitely, never rolling 8 or 7. The following calculations eliminate that possibility.

For $i = 1, 2, 3, \ldots$, let X_i denote the result of roll i in a fair game of craps. Assume that we have observed $X_1 = x = 8$.

(a) Calculate the probability that $X_2 \in \{7, 8\}$.

(b) Calculate the probability that $X_2 \in \{7, 8\}$ *and* that $X_3 \in \{7, 8\}$.

(c) Calculate the probability that $X_2 \in \{7, 8\}$ *and* that $X_3 \in \{7, 8\}$ *and* that $X_4 \in \{7, 8\}$.

(d) What is the probability that the shooter will never roll another 7 or 8?

15. In the final chapter of *All I asking for is my body*, Kiyoshi places an initial, double-or-nothing bet of \$200. If he wins, he will have \$400. If he then wins a second double-or-nothing bet of \$400, he will have \$800. And so on. If he wins five consecutive times, he will have \$6400, enough to pay his family's debt.

(a) Calculate the probability that the shooter will win five consecutive games of Craps if each of the games is fair.

(b) Calculate the probability that the shooter will win five consecutive games of Craps if the shooter is allowed to use Kiyoshi's padrolling system.

(c) Kiyoshi recalls that "Hiroshi never lost." Does this seem plausible?

Chapter 4

Discrete Random Variables

4.1 Basic Concepts

Our introduction of random variables in Section 3.5 was completely general, i.e., the principles that we discussed apply to *all* random variables. In this chapter, we will study an important special class of random variables, the *discrete* random variables. One of the advantages of restricting attention to discrete random variables is that the mathematics required to define various fundamental concepts for this class is fairly minimal.

We begin with a formal definition.

Definition 4.1 *A random variable X is* discrete *if $X(S)$, the set of possible values of X, is countable.*

Our primary interest will be in random variables for which $X(S)$ is finite; however, there are many important random variables for which $X(S)$ is denumerable. The methods described in this chapter apply to both possibilities.

In contrast to the cumulative distribution function (cdf) defined in Section 3.5, we now introduce the probability mass function (pmf).

Definition 4.2 *Let X be a discrete random variable. The* probability mass function *(pmf) of X is the function $f : \Re \to \Re$ defined by*

$$f(x) = P(X = x).$$

If f is the pmf of X, then f necessarily possesses several properties worth noting:

1. $f(x) \geq 0$ for every $x \in \Re$.

2. If $x \notin X(S)$, then $f(x) = 0$.

3. By the definition of $X(S)$,

$$\begin{aligned} \sum_{x \in X(S)} f(x) &= \sum_{x \in X(S)} P(X = x) = P\left(\bigcup_{x \in X(S)} \{x\} \right) \\ &= P\left(X \in X(S)\right) = 1. \end{aligned}$$

There is an important relation between the pmf and the cdf. For each $y \in \Re$, let

$$L(y) = \{x \in X(S) \ : \ x \leq y\}$$

denote the values of X that are less than or equal to y. Then

$$\begin{aligned} F(y) &= P(X \leq y) = P\left(X \in L(y)\right) \\ &= \sum_{x \in L(y)} P(X = x) = \sum_{x \in L(y)} f(x). \end{aligned} \tag{4.1}$$

Thus, the value of the cdf at y can be obtained by summing the values of the pmf at all values $x \leq y$.

More generally, we can compute the probability that X assumes its value in *any* set $B \subset \Re$ by summing the values of the pmf over all values of X that lie in B. Here is the formula:

$$P(X \in B) = \sum_{x \in X(S) \cap B} P(X = x) = \sum_{x \in X(S) \cap B} f(x). \tag{4.2}$$

We now turn to some elementary examples of discrete random variables and their pmfs.

4.2 Examples

Example 4.1 *A fair coin is tossed and the outcome is* `Heads` *or* `Tails`*. Define a random variable X by $X($*`Heads`*$) = 1$ and $X($*`Tails`*$) = 0$.*

The pmf of X is the function f defined by

$$\begin{aligned} f(0) &= P(X = 0) = 0.5, \\ f(1) &= P(X = 1) = 0.5, \end{aligned}$$

and $f(x) = 0$ for all $x \notin X(S) = \{0, 1\}$.

Example 4.2 *A typical penny is spun and the outcome is* `Heads` *or* `Tails`*. Define a random variable X by $X($*`Heads`*$) = 1$ and $X($*`Tails`*$) = 0$.*

Assuming that $P($`Heads`$) = 0.3$ (see Section 1.1.1), the pmf of X is the function f defined by

$$\begin{aligned} f(0) &= P(X = 0) = 0.7, \\ f(1) &= P(X = 1) = 0.3, \end{aligned}$$

and $f(x) = 0$ for all $x \notin X(S) = \{0, 1\}$.

Example 4.3 *A fair die is tossed and the number of dots on the upper face is observed. The sample space is $S = \{1, 2, 3, 4, 5, 6\}$. Define a random variable X by $X(s) = 1$ if s is a prime number and $X(s) = 0$ if s is not a prime number.*

The pmf of X is the function f defined by

$$\begin{aligned} f(0) &= P(X = 0) = P(\{4, 6\}) = 1/3, \\ f(1) &= P(X = 1) = P(\{1, 2, 3, 5\}) = 2/3, \end{aligned}$$

and $f(x) = 0$ for all $x \notin X(S) = \{0, 1\}$.

Examples 4.1–4.3 have a common structure that we proceed to generalize.

Definition 4.3 *A random variable X is a Bernoulli trial if $X(S) = \{0, 1\}$.*

Traditionally, we call $X = 1$ a "success" and $X = 0$ a "failure".

The family of probability distributions of Bernoulli trials is parametrized (indexed) by a real number $p \in [0, 1]$, usually by setting $p = P(X = 1)$. We communicate that X is a Bernoulli trial with success probability p by writing $X \sim \text{Bernoulli}(p)$. The pmf of such a random variable is the function f defined by

$$\begin{aligned} f(0) &= P(X = 0) = 1 - p, \\ f(1) &= P(X = 1) = p, \end{aligned}$$

and $f(x) = 0$ for all $x \notin X(S) = \{0, 1\}$.

Several important families of random variables can be derived from Bernoulli trials. Consider, for example, the familiar experiment of tossing a fair coin twice and counting the number of `Heads`. In Section 4.4, we will generalize this experiment and count the number of successes in n Bernoulli trials. This will lead to the family of *binomial* probability distributions.

Bernoulli trials are also a fundamental ingredient of the St. Petersburg Paradox, described in Example 4.14. In this experiment, a fair coin is tossed until `Heads` was observed and the number of `Tails` was counted. More generally, consider an experiment in which a sequence of independent Bernoulli trials, each with success probability p, is performed until the first success is observed. Let $X_1, X_2, X_3, \ldots$ denote the individual Bernoulli trials and let Y denote the number of failures that precede the first success. Then the possible values of Y are $Y(S) = \{0, 1, 2, \ldots\}$ and the pmf of Y is

$$\begin{aligned} f(j) = P(Y = j) &= P\left(X_1 = 0, \ldots, X_j = 0, X_{j+1} = 1\right) \\ &= P\left(X_1 = 0\right) \cdots P\left(X_j = 0\right) \cdot P\left(X_{j+1} = 1\right) \\ &= (1-p)^j p \end{aligned}$$

if $j \in Y(S)$ and $f(j) = 0$ if $j \notin Y(S)$. This family of probability distributions is also parametrized by a real number $p \in [0, 1]$. It is called the *geometric* family and a random variable with a geometric distribution is said to be a geometric random variable, written $Y \sim \text{Geometric}(p)$.

If $Y \sim \text{Geometric}(p)$ and $k \in Y(S)$, then

$$F(k) = P(Y \leq k) = 1 - P(Y > k) = 1 - P(Y \geq k+1).$$

Because the event $\{Y \geq k+1\}$ occurs if and only if $X_1 = \cdots X_{k+1} = 0$, we conclude that

$$F(k) = 1 - (1-p)^{k+1}.$$

Example 4.4 *Gary is a college student who is determined to have a date for an approaching formal. He believes that each woman he asks is twice as likely to decline his invitation as to accept it, but he resolves to extend invitations until one is accepted. However, each of his first ten invitations is declined. Assuming that Gary's assumptions about his own desirability are correct, what is the probability that he would encounter such a run of bad luck?*

Gary evidently believes that he can model his invitations as a sequence of independent Bernoulli trials, each with success probability $p = 1/3$. If so, then the number of unsuccessful invitations that he extends is a random variable $Y \sim \text{Geometric}(1/3)$ and

$$P(Y \geq 10) = 1 - P(Y \leq 9) = 1 - F(9) = 1 - \left[1 - \left(\frac{2}{3}\right)^{10}\right] \doteq 0.0173.$$

Either Gary is very unlucky or his assumptions are flawed. Perhaps his probability model is correct, but $p < 1/3$. Perhaps, as seems likely, the probability of success depends on who he asks. Or perhaps the trials were not really independent.[1] If Gary's invitations cannot be modelled as independent and identically distributed Bernoulli trials, then the geometric distribution cannot be used.

Another important family of random variables is often derived by considering an *urn model*. Imagine an urn that contains m red balls and n black balls. The experiment of present interest involves selecting k balls from the urn in such a way that each of the $\binom{m+n}{k}$ possible outcomes that might be obtained are equally likely. Let X denote the number of red balls selected in this manner. If we observe $X = x$, then x red balls were selected from a total of m red balls and $k - x$ black balls were selected from a total of n black balls. Evidently, $x \in X(S)$ if and only if x is an integer that satisfies $x \leq \min(m, k)$ and $k - x \leq \min(n, k)$. Furthermore, if $x \in X(S)$, then the pmf of X is

$$f(x) = P(X = x) = \frac{\#\{X = x\}}{\#S} = \frac{\binom{m}{x}\binom{n}{k-x}}{\binom{m+n}{k}}. \tag{4.3}$$

This family of probability distributions is parametrized by a triple of integers, (m, n, k), for which $m, n \geq 0$, $m + n \geq 1$, and $0 \leq k \leq m + n$. It is called the *hypergeometric* family and a random variable with a hypergeometric distribution is said to be a hypergeometric random variable, written $Y \sim \text{Hypergeometric}(m, n, k)$.

The trick to using the hypergeometric distribution in applications is to recognize a correspondence between the actual experiment and an idealized urn model, as in...

Example 4.5 *Consider the hypothetical example described in Section 1.2, in which* 30 *freshman and* 10 *non-freshmen are randomly assigned exam* `A` *or B. What is the probability that exactly* 15 *freshmen (and therefore exactly* 5 *non-freshmen) receive exam* `A`*?*

In Example 2.5 we calculated that the probability in question is

$$\frac{\binom{30}{15}\binom{10}{5}}{\binom{40}{20}} = \frac{39,089,615,040}{137,846,528,820} \doteq 0.28. \tag{4.4}$$

[1] In the actual incident on which this example is based, the women all lived in the *same residential college*. It seems doubtful that each woman was completely unaware of the invitation that preceded hers.

Let us re-examine this calculation. Suppose that we write each student's name on a slip of paper, mix the slips in a jar, then draw 20 slips without replacement. These 20 students receive exam `A`; the remaining 20 students receive exam `B`. Now drawing slips of paper from a jar is exactly like drawing balls from an urn. There are $m = 30$ slips with freshman names (red balls) and $n = 10$ slips with non-freshman names (black balls), of which we are drawing $k = 20$ without replacement. Using the hypergeometric pmf defined by (4.3), the probability of drawing exactly $x = 15$ freshman names is

$$\frac{\binom{m}{x}\binom{n}{k-x}}{\binom{m+n}{k}} = \frac{\binom{30}{15}\binom{10}{5}}{\binom{40}{20}},$$

the left-hand side of (4.4).

Example 4.6 (Adapted from an example analyzed by R.R. Sokal and F.J. Rohlf (1969), *Biometry: The Principles and Practice of Statistics in Biological Research*, W.H. Freeman and Company, San Francisco.)

All but 28 *acacia trees (of the same species) were cleared from a study area in Central America. The* 28 *remaining trees were freed from ants by one of two types of insecticide. The standard insecticide (A) was administered to* 15 *trees; an experimental insecticide (B) was administered to the other* 13 *trees. The assignment of insectides to trees was completely random. At issue was whether or not the experimental insecticide was more effective than the standard insecticide in inhibiting future ant infestations.*

Next, 16 *separate ant colonies were situated roughly equidistant from the acacia trees and permitted to invade them. Unless food is scarce, different colonies will not compete for the same resources; hence, it could be presumed that each colony would invade a different tree. In fact, the ants invaded* 13 *of the* 15 *trees treated with the standard insecticide and only* 3 *of the* 13 *trees treated with the experimental insecticide. If the two insecticides were equally effective in inhibiting future infestations, then what is the probability that no more than* 3 *ant colonies would have invaded trees treated with the experimental insecticide?*

This is a potentially confusing problem that is simplified by constructing an urn model for the experiment. There are $m = 13$ trees with the experimental insecticide (red balls) and $n = 15$ trees with the standard insecticide (black balls). The ants choose $k = 16$ trees (balls). Let X denote the number of experimental trees (red balls) invaded by the ants; then

$X \sim \text{Hypergeometric}(13, 15, 16)$ and its pmf is

$$f(x) = P(X = x) = \frac{\binom{13}{x}\binom{15}{16-x}}{\binom{28}{16}}.$$

Notice that there are not enough standard trees for each ant colony to invade one; hence, at least one ant colony *must* invade an experimental tree and $X = 0$ is impossible. Thus,

$$P(X \leq 3) = f(1) + f(2) + f(3) = \frac{\binom{13}{1}\binom{15}{15}}{\binom{28}{16}} + \frac{\binom{13}{2}\binom{15}{14}}{\binom{28}{16}} + \frac{\binom{13}{3}\binom{15}{13}}{\binom{28}{16}} \doteq 0.0010.$$

This reasoning illustrates the use of a statistical procedure called *Fisher's exact test*. The probability that we have calculated is an example of what we will later call a *significance probability*. In the present example, the fact that the significance probability is so small would lead us to challenge an assertion that the experimental insecticide is no better than the standard insecticide.

It is evident that calculations with the hypergeometric distribution can become rather tedious. Accordingly, this is a convenient moment to introduce computer software for the purpose of evaluating certain pmfs and cdfs. The statistical programming language `R` includes functions that evaluate pmfs and cdfs for a variety of distributions, including the geometric and hypergeometric.[2] For the geometric, these functions are `dgeom` and `pgeom`; for the hypergeometric, these functions are `dhyper` and `phyper`. We can calculate the probability in Example 4.4 as follows:

```
> 1-pgeom(q=9,prob=1/3)
[1] 0.01734153
```

Similarly, we can calculate the probability in Example 4.6 as follows:

```
> phyper(q=3,m=13,n=15,k=16)
[1] 0.001026009
```

[2] R is a free, open-source implementation of `S`, developed at AT&T Bell Laboratories. See Appendix R for information about obtaining, installing, and using `R`.

4.3 Expectation

Sometime in the early 1650s, the eminent theologian and amateur mathematician Blaise Pascal found himself in the company of the Chevalier de Méré.[3] De Méré posed to Pascal a famous problem: how to divide the pot of an interrupted dice game. Pascal communicated the problem to Pierre de Fermat in 1654, beginning a celebrated correspondence that established a foundation for the mathematics of probability.

Pascal and Fermat began by agreeing that the pot should be divided according to each player's chances of winning it. For example, suppose that each of two players has selected a number from the set $S = \{1, 2, 3, 4, 5, 6\}$. For each roll of a fair die that produces one of their respective numbers, the corresponding player receives a token. The first player to accumulate five tokens wins a pot of \$100. Suppose that the game is interrupted with Player A having accumulated four tokens and Player B having accumulated only one. The probability that Player B would have won the pot had the game been completed is the probability that B's number would have appeared four more times before A's number appeared one more time. Because we can ignore rolls that produce neither number, this is equivalent to the probability that a fair coin will have a run of four consecutive `Heads`, i.e., $0.5 \cdot 0.5 \cdot 0.5 \cdot 0.5 = 0.0625$. Hence, according to Pascal and Fermat, Player B is entitled to $0.0625 \cdot \$100 = \$6.25$ from the pot and Player A is entitled to the remaining \$93.75.

The crucial concept in Pascal's and Fermat's analysis is the notion that each prospect should be weighted by the chance of realizing that prospect. This notion motivates

Definition 4.4 *The expected value of a discrete random variable X, which we will denote $E(X)$ or simply EX, is the probability-weighted average of the possible values of X, i.e.,*

$$EX = \sum_{x \in X(S)} xP(X = x) = \sum_{x \in X(S)} xf(x).$$

Remark The expected value of X, EX, is often called the *population mean* and denoted μ.

[3] This account of the origins of modern probability can be found in Chapter 6 of David Bergamini's *Mathematics*, Life Science Library, Time Inc., New York, 1963.

Example 4.7 If $X \sim \text{Bernoulli}(p)$, then

$$\mu = EX = \sum_{x \in \{0,1\}} xP(X = x) = 0 \cdot P(X = 0) + 1 \cdot P(X = 1) = P(X = 1) = p.$$

Notice that, in general, the expected value of X is *not* the average of its possible values. In this example, the possible values are $X(S) = \{0, 1\}$ and the average of these values is (always) 0.5. In contrast, the expected value depends on the probabilities of the values.

Fair Value The expected payoff of a game of chance is sometimes called the *fair value* of the game. For example, suppose that you own a slot machine that pays a jackpot of \$1000 with probability $p = 0.0005$ and \$0 with probability $1 - p = 0.9995$. How much should you charge a customer to play this machine? Letting X denote the payoff (in dollars), the expected payoff per play is

$$EX = 1000 \cdot 0.0005 + 0 \cdot 0.9995 = 0.5;$$

hence, if you want to make a profit, then you should charge more than \$0.50 per play. Suppose, however, that a rival owner of an identical slot machine attempted to compete for the same customers. According to the theory of microeconomics, competition would cause each of you to try to undercut the other, eventually resulting in an equilibrium price of exactly \$0.50 per play, the fair value of the game.

We proceed to illustrate both the mathematics and the psychology of fair value by considering several lotteries. A *lottery* is a choice between receiving a certain payoff and playing a game of chance. In each of the following examples, we emphasize that the value accorded the game of chance by a rational person may be very different from the game's expected value. In this sense, the phrase "fair value" is often a misnomer.

Example 4.8a *You are offered the choice between receiving a certain \$5 and playing the following game: a fair coin is tossed and you receive \$10 or \$0 according to whether* `Heads` *or* `Tails` *is observed.*

The expected payoff from the game (in dollars) is

$$EX = 10 \cdot 0.5 + 0 \cdot 0.5 = 5,$$

so your options are equivalent with respect to expected earnings. One might therefore suppose that a rational person would be indifferent to which option

he or she selects. Indeed, in my experience, some students prefer to take the certain \$5 and some students prefer to gamble on perhaps winning \$10. For this example, the phrase "fair value" seems apt.

Example 4.8b *You are offered the choice between receiving a certain \$5000 and playing the following game: a fair coin is tossed and you receive \$10,000 or \$0 according to whether* `Heads` *or* `Tails` *is observed.*

The mathematical structure of this lottery is identical to that of the preceding lottery, except that the stakes are higher. Again, the options are equivalent with respect to expected earnings; again, one might suppose that a rational person would be indifferent to which option he or she selects. However, many students who opt to gamble on perhaps winning \$10 in Example 4.8a opt to take the certain \$5000 in Example 4.8b.

Example 4.8c *You are offered the choice between receiving a certain \$1 million and playing the following game: a fair coin is tossed and you receive \$2 million or \$0 according to whether* `Heads` *or* `Tails` *is observed.*

The mathematical structure of this lottery is identical to that of the preceding two lotteries, except that the stakes are now *much* higher. Again, the options are equivalent with respect to expected earnings; however, almost every student to whom I have presented this lottery has expressed a strong preference for taking the certain \$1 million.

Example 4.9 *You are offered the choice between receiving a certain \$1 million and playing the following game: a fair coin is tossed and you receive \$5 million or \$0 according to whether* `Heads` *or* `Tails` *is observed.*

The expected payoff from this game (in millions of dollars) is

$$EX = 5 \cdot 0.5 + 0 \cdot 0.5 = 2.5,$$

so playing the game is the more attractive option with respect to expected earnings. Nevertheless, most students opt to take the certain \$1 million. This should *not* be construed as an irrational decision. For example, the addition of \$1 million to my own modest estate would secure my eventual retirement. The addition of an extra \$4 million would be very pleasant indeed, allowing me to increase my current standard of living. However, I do not value the additional \$4 million nearly as much as I value the initial \$1 million. As Aesop observed, "A little thing in hand is worth more than a great thing in prospect." For this example, the phrase "fair value" introduces normative connotations that are not appropriate.

Example 4.10 Consider the following passage from a recent article about investing:

> "...it's human nature to overweight low probabilities that offer high returns. In one study, subjects were given a choice between a 1-in-1000 chance to win $5000 or a sure thing to win $5; or a 1-in-1000 chance of losing $5000 versus a sure loss of $5. In the first case, the expected value (mathematically speaking) is making $5. In the second case, it's losing $5. Yet in the first situation, which mimics a lottery, more than 70% of people asked chose to go for the $5000. In the second situation, more than 80% would take the $5 hit."[4]

The author evidently considered the reported preferences paradoxical, but are they really surprising? Plus or minus $5 will not appreciably alter the financial situations of most subjects, but plus or minus $5000 will. It is perfectly rational to risk a negligible amount on the chance of winning $5000 while declining to risk a negligible amount on the chance of losing $5000. The following examples further explicate this point.

Example 4.11 The same article advises, "To limit completely irrational risks, such as lottery tickets, try speculating only with money you would otherwise use for simple pleasures, such as your morning coffee."

Consider a hypothetical state lottery, in which 6 numbers are drawn (without replacement) from the set $\{1, 2, \ldots, 39, 40\}$. For $2, you can purchase a ticket that specifies 6 such numbers. If the numbers on your ticket match the numbers selected by the state, then you win $1 million; otherwise, you win nothing. (For the sake of simplicity, we ignore the possibility that you might have to split the jackpot with other winners and the possibility that you might win a lesser prize.) Is buying a lottery ticket "completely irrational"?

The probability of winning the lottery in question is

$$p = \frac{1}{\binom{40}{6}} = \frac{1}{3,838,380} \doteq 2.6053 \times 10^{-7},$$

so your expected prize (in dollars) is approximately

$$10^6 \cdot 2.6053 \times 10^{-7} \doteq 0.26,$$

[4]Robert Frick, "The 7 Deadly Sins of Investing," *Kiplinger's Personal Finance Magazine*, March 1998, p. 138.

which is considerably less than the cost of a ticket. Evidently, it is completely irrational to buy tickets for this lottery *as an investment strategy*. Suppose, however, that I buy one ticket per week and reason as follows: I will almost certainly lose \$2 per week, but that loss will have virtually no impact on my standard of living; however, if by some miracle I win, then gaining \$1 million will revolutionize my standard of living. This can hardly be construed as irrational behavior, although Robert Frick's advice to speculate only with funds earmarked for entertainment is well-taken.

In most state lotteries, the fair value of the game is less than the cost of a lottery ticket. This is only natural—lotteries exist because they generate revenue for the state that runs them! (By the same reasoning, gambling must favor the house because casinos make money for their owners.) However, on very rare occasions a jackpot is so large that the typical situation is reversed. Several years ago, an Australian syndicate noticed that the fair value of a Florida state lottery exceeded the price of a ticket and purchased a large number of tickets as an (ultimately successful) investment strategy. And Voltaire once purchased every ticket in a raffle upon noting that the prize was worth more than the total cost of the tickets being sold!

Example 4.12 If the first case described in Example 4.10 mimics a lottery, then the second case mimics insurance. Mindful that insurance companies (like casinos) make money, Ambrose Bierce offered the following definition:

> "INSURANCE, *n.* An ingenious modern game of chance in which the player is permitted to enjoy the comfortable conviction that he is beating the man who keeps the table."[5]

However, while it is certainly true that the fair value of an insurance policy is less than the premiums required to purchase it, it does not follow that buying insurance is irrational. I can easily afford to pay \$200 per year for homeowners insurance, but I would be ruined if all of my possessions were destroyed by fire and I received no compensation for them. My decision that a certain but affordable loss is preferable to an unlikely but catastrophic loss is an example of *risk-averse* behavior.

Before presenting our concluding example of fair value, we derive a useful formula. Suppose that $X : S \to \Re$ is a discrete random variable and $\phi : \Re \to$

[5] Ambrose Bierce, *The Devil's Dictionary*, 1881–1906. In *The Collected Writings of Ambrose Bierce*, Citadel Press, Secaucus, NJ, 1946.

$\Re$ is a function. Let $Y = \phi(X)$. Then $Y : \Re \to \Re$ is a random variable and

$$\begin{aligned}
E\phi(X) &= EY = \sum_{y \in Y(S)} yP(Y = y) \\
&= \sum_{y \in Y(S)} yP\left(\phi(X) = y\right) \\
&= \sum_{y \in Y(S)} yP\left(X \in \phi^{-1}(y)\right) \\
&= \sum_{y \in Y(S)} y \left(\sum_{x \in \phi^{-1}(y)} P(X = x) \right) \\
&= \sum_{y \in Y(S)} \sum_{x \in \phi^{-1}(y)} yP(X = x) \\
&= \sum_{y \in Y(S)} \sum_{x \in \phi^{-1}(y)} \phi(x)P(X = x) \\
&= \sum_{x \in X(S)} \phi(x)P(X = x) \\
&= \sum_{x \in X(S)} \phi(x)f(x). \qquad (4.5)
\end{aligned}$$

Example 4.13 *Consider a game in which the jackpot starts at \$1 and doubles each time that* `Tails` *is observed when a fair coin is tossed. The game terminates when* `Heads` *is observed for the first time. How much would you pay for the privilege of playing this game? How much would you charge if you were responsible for making the payoff?*

This is a curious game. With high probability, the payoff will be rather small; however, there is a small chance of a very large payoff. In response to the first question, most students discount the latter possibility and respond that they would only pay a small amount, rarely more than \$4. In response to the second question, most students recognize the possibility of a large payoff and demand payment of a considerably greater amount. Let us consider if the notion of fair value provides guidance in reconciling these perspectives.

Let X denote the number of `Tails` that are observed before the game terminates. Then $X(S) = \{0, 1, 2, \ldots\}$ and the geometric random variable X has pmf

$$f(x) = P(x \text{ consecutive } \texttt{Tails}) = 0.5^x.$$

The payoff from this game (in dollars) is $Y = 2^X$; hence, the expected

payoff is

$$E2^X = \sum_{x=0}^{\infty} 2^x \cdot 0.5^x = \sum_{x=0}^{\infty} 1 = \infty.$$

This is quite startling! The "fair value" of this game provides very little insight into the value that a rational person would place on playing it. This remarkable example is quite famous—it is known as the St. Petersburg Paradox.

Properties of Expectation We now state (and sometimes prove) some useful consequences of Definition 4.4 and Equation 4.5.

Theorem 4.1 *Let X denote a discrete random variable and suppose that $P(X = c) = 1$. Then $EX = c$.*

Theorem 4.1 states that, if a random variable always assumes the same value c, then the probability-weighted average of the values that it assumes is c. This should be obvious.

Theorem 4.2 *Let X denote a discrete random variable and suppose that $c \in \Re$ is constant. Then*

$$E\left[c\phi(X)\right] = \sum_{x \in X(S)} c\phi(x)f(x) = c \sum_{x \in X(S)} \phi(x)f(x) = cE\left[\phi(X)\right].$$

Theorem 4.2 states that we can interchange the order of multiplying by a constant and computing the expected value. Notice that this property of expectation follows directly from the analogous property for summation.

Theorem 4.3 *Let X denote a discrete random variable. Then*

$$\begin{aligned} E\left[\phi_1(X) + \phi_2(X)\right] &= \sum_{x \in X(S)} \left[\phi_1(x) + \phi_2(x)\right] f(x) \\ &= \sum_{x \in X(S)} \left[\phi_1(x)f(x) + \phi_2(x)f(x)\right] \\ &= \sum_{x \in X(S)} \phi_1(x)f(x) + \sum_{x \in X(S)} \phi_2(x)f(x) \\ &= E\left[\phi_1(X)\right] + E\left[\phi_2(X)\right]. \end{aligned}$$

Theorem 4.3 states that we can interchange the order of adding functions of a random variable and computing the expected value. Again, this property of expectation follows directly from the analogous property for summation.

Theorem 4.4 *Let X_1 and X_2 denote discrete random variables. Then*

$$E[X_1 + X_2] = EX_1 + EX_2.$$

Theorem 4.4 states that the expected value of a sum equals the sum of the expected values.

Variance Now suppose that X is a discrete random variable, let $\mu = EX$ denote its expected value, or population mean., and define a function $\phi : \Re \to \Re$ by

$$\phi(x) = (x - \mu)^2.$$

For any $x \in \Re$, $\phi(x)$ is the squared deviation of x from the expected value of X. If X always assumes the value μ, then $\phi(X)$ always assumes the value 0; if X tends to assume values near μ, then $\phi(X)$ will tend to assume small values; if X often assumes values far from μ, then $\phi(X)$ will often assume large values. Thus, $E\phi(X)$, the expected squared deviation of X from its expected value, is a measure of the variability of the population $X(S)$. We summarize this observation in

Definition 4.5 *The variance of a discrete random variable X, which we will denote Var(X) or simply Var$X$, is the probability-weighted average of the squared deviations of X from $EX = \mu$, i.e.,*

$$Var\, X = E(X - \mu)^2 = \sum_{x \in X(S)} (x - \mu)^2 f(x).$$

Remark The variance of X, Var X, is often called the *population variance* and denoted σ^2.

Denoting the population variance by σ^2 may strike the reader as awkward notation, but there is an excellent reason for it. Because the variance measures squared deviations from the population mean, it is measured in different units than either the random variable itself or its expected value. For example, if X measures length in meters, then so does EX, but Var X is measured in meters squared. To recover a measure of population variability in the original units of measurement, we take the square root of the variance and obtain σ.

Definition 4.6 *The standard deviation of a random variable is the square root of its variance.*

Remark The standard deviation of X, often denoted σ, is often called the *population standard deviation.*

Example 4.1 (continued) If $X \sim \text{Bernoulli}(p)$, then

$$\begin{aligned}
\sigma^2 = \text{Var}\, X &= E(X-\mu)^2 \\
&= (0-\mu)^2 \cdot P(X=0) + (1-\mu)^2 \cdot P(X=1) \\
&= (0-p)^2(1-p) + (1-p)^2 p \\
&= p(1-p)(p+1-p) \\
&= p(1-p).
\end{aligned}$$

Before turning to a more complicated example, we establish a useful fact.

Theorem 4.5 *If X is a discrete random variable, then*

$$\begin{aligned}
Var\, X &= E(X-\mu)^2 \\
&= E(X^2 - 2\mu X + \mu^2) \\
&= EX^2 + E(-2\mu X) + E\mu^2 \\
&= EX^2 - 2\mu EX + \mu^2 \\
&= EX^2 - 2\mu^2 + \mu^2 \\
&= EX^2 - (EX)^2.
\end{aligned}$$

A straightforward way to calculate the variance of a discrete random variable that assumes a fairly small number of values is to exploit Theorem 4.5 and organize one's calculations in the form of a table.

Example 4.14 *Suppose that X is a random variable whose possible values are $X(S) = \{2, 3, 5, 10\}$. Suppose that the probability of each of these values is given by the formula $f(x) = P(X = x) = x/20$.*

(a) Calculate the expected value of X.

(b) Calculate the variance of X.

(c) Calculate the standard deviation of X.

Solution

x	$f(x)$	$xf(x)$	x^2	$x^2f(x)$
2	0.10	0.20	4	0.40
3	0.15	0.45	9	1.35
5	0.25	1.25	25	6.25
10	0.50	5.00	100	50.00
		6.90		58.00

(a) $\mu = EX = 0.2 + 0.45 + 1.25 + 5 = 6.9.$

(b) $\sigma^2 = \operatorname{Var} X = EX^2 - (EX)^2 = (0.4 + 1.35 + 6.25 + 50) - 6.9^2 = 58 - 47.61 = 10.39.$

(c) $\sigma = \sqrt{10.39} \doteq 3.2234.$

Now suppose that $X : S \to \Re$ is a discrete random variable and $\phi : \Re \to \Re$ is a function. Let $Y = \phi(X)$. Then Y is a discrete random variable and

$$\operatorname{Var} \phi(X) = \operatorname{Var} Y = E\left[Y - EY\right]^2 = E\left[\phi(X) - E\phi(X)\right]^2. \tag{4.6}$$

We conclude this section by stating (and sometimes proving) some useful consequences of Definition 4.5 and Equation 4.6.

Theorem 4.6 *Let X denote a discrete random variable and suppose that $c \in \Re$ is constant. Then*

$$Var(X + c) = Var\, X.$$

Although possibly startling at first glance, this result is actually quite intuitive. The variance depends on the squared deviations of the values of X from the expected value of X. If we add a constant to each value of X, then we shift both the individual values of X and the expected value of X by the same amount, preserving the squared deviations. The *variability* of a population is not affected by shifting each of the values in the population by the same amount.

Theorem 4.7 *Let X denote a discrete random variable and suppose that $c \in \Re$ is constant. Then*

$$\begin{aligned} Var(cX) &= E\left[cX - E(cX)\right]^2 \\ &= E\left[cX - cEX\right]^2 \\ &= E\left[c(X - EX)\right]^2 \\ &= E\left[c^2(X - EX)^2\right] \\ &= c^2 E(X - EX)^2 \\ &= c^2\, Var\, X. \end{aligned}$$

To understand this result, recall that the variance is measured in the original units of measurement squared. If we take the square root of each expression in Theorem 4.7, then we see that one can interchange multiplying a random variable by a nonnegative constant with computing its *standard deviation.*

Theorem 4.8 *If the discrete random variables X_1 and X_2 are independent, then*

$$Var(X_1 + X_2) = Var\, X_1 \; + \; Var\, X_2.$$

Theorem 4.8 is analogous to Theorem 4.4. However, in order to ensure that the variance of a sum equals the sum of the variances, the random variables must be independent.

4.4 Binomial Distributions

Suppose that a fair coin is tossed twice and the number of `Heads` is counted. Let Y denote the total number of `Heads`. Because the sample space has four equally likely outcomes, viz.,

$$S = \{\texttt{HH}, \texttt{HT}, \texttt{TH}, \texttt{TT}\},$$

the pmf of Y is easily determined:

$$\begin{aligned} f(0) &= P(Y = 0) = P\left(\{\texttt{HH}\}\right) = 0.25, \\ f(1) &= P(Y = 1) = P\left(\{\texttt{HT}, \texttt{TH}\}\right) = 0.5, \\ f(2) &= P(Y = 2) = P\left(\{\texttt{TT}\}\right) = 0.25, \end{aligned}$$

and $f(y) = 0$ if $y \notin Y(S) = \{0, 1, 2\}$.

Referring to representation (c) of Example 3.18, the above experiment has the following characteristics:

- Let X_1 denote the number of `Heads` observed on the first toss and let X_2 denote the number of `Heads` observed on the second toss. Then the random variable of interest is $Y = X_1 + X_2$.
- The random variables X_1 and X_2 are independent.
- The random variables X_1 and X_2 have the same distribution, viz.

$$X_1, X_2 \sim \text{Bernoulli}(0.5).$$

We proceed to generalize this example in two ways:

1. We allow any finite number of trials.
2. We allow any success probability $p \in [0, 1]$.

Definition 4.7 *Let $X_1, \ldots, X_n$ be mutually independent Bernoulli trials, each with success probability p. Then*

$$Y = \sum_{i=1}^{n} X_i$$

is a binomial random variable, denoted

$$Y \sim \text{Binomial}(n; p).$$

Applying Theorem 4.4, we see that the expected value of a binomial random variable is the product of the number of trials and the probability of success:

$$EY = E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} EX_i = \sum_{i=1}^{n} p = np.$$

Furthermore, because the trials are independent, we can apply Theorem 4.8 to calculate the variance:

$$\text{Var}\, Y = \text{Var}\left(\sum_{i=1}^{n} X_i\right) = \left(\sum_{i=1}^{n} \text{Var}\, X_i\right) = \left(\sum_{i=1}^{n} p(1-p)\right) = np(1-p).$$

Because Y counts the total number of successes in n Bernoulli trials, it should be apparent that $Y(S) = \{0, 1, \ldots, n\}$. Let f denote the pmf of Y. For fixed n, p, and $j \in Y(S)$, we wish to determine

$$f(j) = P(Y = j).$$

To illustrate the reasoning required to make this determination, suppose that there are $n = 6$ trials, each with success probability $p = 0.3$, and that we wish to determine the probability of observing exactly $j = 2$ successes. Some examples of experimental outcomes for which $Y = 2$ include the following:

$$110000 \quad 000011 \quad 010010$$

Because the trials are mutually independent, we see that

$$\begin{aligned} P(110000) &= 0.3 \cdot 0.3 \cdot 0.7 \cdot 0.7 \cdot 0.7 \cdot 0.7 = 0.3^2 \cdot 0.7^4, \\ P(000011) &= 0.7 \cdot 0.7 \cdot 0.7 \cdot 0.7 \cdot 0.3 \cdot 0.3 = 0.3^2 \cdot 0.7^4, \\ P(010010) &= 0.7 \cdot 0.3 \cdot 0.7 \cdot 0.7 \cdot 0.3 \cdot 0.7 = 0.3^2 \cdot 0.7^4. \end{aligned}$$

It should be apparent that the probability of each outcome for which $Y = 2$ is the product of $j = 2$ factors of $p = 0.3$ and $n - j = 4$ factors of $1 - p = 0.7$. Furthermore, the number of such outcomes is the number of ways of choosing $j = 2$ successes from a total of $n = 6$ trials. Thus,

$$f(2) = P(Y = 2) = \binom{6}{2} 0.3^2 0.7^4$$

for the specific example in question and the general formula for the binomial pmf is

$$f(j) = P(Y = j) = \binom{n}{j} p^j (1 - p)^{n-j}.$$

It follows, of course, that the general formula for the binomial cdf is

$$\begin{aligned} F(k) = P(Y \leq k) &= \sum_{j=0}^{k} P(Y = j) = \sum_{j=0}^{k} f(j) \\ &= \sum_{j=0}^{k} \binom{n}{j} p^j (1 - p)^{n-j}. \end{aligned} \tag{4.7}$$

Except for very small numbers of trials, direct calculation of (4.7) is rather tedious. Fortunately, tables of the binomial cdf for selected values of

n and p are widely available, as is computer software for evaluating (4.7). In the examples that follow, we will evaluate (4.7) using the `R` function `pbinom`.

As the following examples should make clear, the trick to evaluating binomial probabilities is to write them in expressions that only involve probabilities of the form $P(Y \leq k)$.

Example 4.15 *In* 10 *trials with success probability* 0.5*, what is the probability that no more than* 4 *successes will be observed?*

Here, $n = 10$, $p = 0.5$, and we want to calculate

$$P(Y \leq 4) = F(4).$$

We do so in `R` as follows:

```
> pbinom(4,size=10,prob=.5)
[1] 0.3769531
```

Example 4.16 *In* 12 *trials with success probability* 0.3*, what is the probability that more than* 6 *successes will be observed?*

Here, $n = 12$, $p = 0.3$, and we want to calculate

$$P(Y > 6) = 1 - P(Y \leq 6) = 1 - F(6).$$

We do so in `R` as follows:

```
> 1-pbinom(6,12,.3)
[1] 0.03860084
```

Example 4.17 *In* 15 *trials with success probability* 0.6*, what is the probability that at least* 5 *but no more than* 10 *successes will be observed?*

Here, $n = 15$, $p = 0.6$, and we want to calculate

$$P(5 \leq Y \leq 10) = P(Y \leq 10) - P(Y \leq 4) = F(10) - F(4).$$

We do so in `R` as follows:

```
> pbinom(10,15,.6)-pbinom(4,15,.6)
[1] 0.7733746
```

Example 4.18 *In* 20 *trials with success probability* 0.9*, what is the probability that exactly* 16 *successes will be observed?*

Here, $n = 20$, $p = 0.9$, and we want to calculate

$$P(Y = 16) = P(Y \leq 16) - P(Y \leq 15) = F(16) - F(15).$$

We do so in `R` as follows:

```
> pbinom(16,20,.9)-pbinom(15,20,.9)
[1] 0.08977883
```

Example 4.19 *In* 81 *trials with success probability* 0.64*, what is the probability that the proportion of observed successes will be between* 60 *and* 70 *percent?*

Here, $n = 81$, $p = 0.64$, and we want to calculate

$$\begin{aligned} P(0.6 < Y/81 < 0.7) &= P(0.6 \cdot 81 < Y < 0.7 \cdot 81) \\ &= P(48.6 < Y < 56.7) \\ &= P(49 \leq Y \leq 56) \\ &= P(Y \leq 56) - P(Y \leq 48) \\ &= F(56) - F(48). \end{aligned}$$

We do so in `R` as follows:

```
> pbinom(56,81,.64)-pbinom(48,81,.64)
[1] 0.6416193
```

Many practical situations can be modelled using a binomial distribution. Doing so typically requires one to perform the following steps.

1. Identify what constitutes a Bernoulli trial and what constitutes a success. Verify or assume that the trials are mutually independent with a common probability of success.

2. Identify the number of trials (n) and the common probability of success (p).

3. Identify the event whose probability is to be calculated.

4. Calculate the probability of the event in question, e.g., by using the `pbinom` function in `R`.

Example 4.20 *RD Airlines flies planes that seat* 58 *passengers. Years of experience have revealed that* 20 *percent of the persons who purchase tickets fail to claim their seat. (Such persons are called "no-shows".) Because of this phenomenon, RD routinely overbooks its flights, i.e., RD typically sells more than* 58 *tickets per flight. If more than* 58 *passengers show, then the "extra" passengers are "bumped" to another flight. Suppose that RD sells* 64 *tickets for a certain flight from Washington to New York. How might RD estimate the probability that at least one passenger will have to be bumped?*

1. Each person who purchased a ticket must decide whether or not to claim his or her seat. This decision represents a Bernoulli trial, for which we will declare a decision to claim the seat a success. Strictly speaking, the Bernoulli trials in question are neither mutually independent nor identically distributed. Some individuals, e.g., families, travel together and make a common decision as to whether or not to claim their seats. Furthermore, some travellers are more likely to change their plans than others. Nevertheless, absent more detailed information, we should be able to compute an approximate answer by assuming that the total number of persons who claim their seats has a binomial distribution.

2. The problem specifies that $n = 64$ persons have purchased tickets. Appealing to past experience, we assume that the probability that each person will show is $p = 1 - 0.2 = 0.8$.

3. At least one passenger will have to be bumped if more than 58 passengers show, so the desired probability is

$$P(Y > 58) = 1 - P(Y \leq 58) = 1 - F(58).$$

4. The necessary calculation can be performed in `R` as follows:

```
> 1-pbinom(58,64,.8)
[1] 0.006730152
```

4.5 Exercises

1. Suppose that a weighted die is tossed. Let X denote the number of dots that appear on the upper face of the die, and suppose that $P(X = x) = (7 - x)/20$ for $x = 1, 2, 3, 4, 5$ and $P(X = 6) = 0$. Determine each of the following:

(a) The probability mass function of X.

(b) The cumulative distribution function of X.

(c) The expected value of X.

(d) The variance of X.

(e) The standard deviation of X.

2. Suppose that a jury of 12 persons is to be selected from a pool of 25 persons who were called for jury duty. The pool comprises 12 retired persons, 6 employed persons, 5 unemployed persons, and 2 students. Assuming that each person is equally likely to be selected, answer the following:

   (a) What is the probability that both students will be selected?

   (b) What is the probability that the jury will contain exactly twice as many retired persons as employed persons?

3. When casting four astragali, a throw that results in four different uppermost sides is called a *venus*. (See Section 1.4.) Suppose that four astragali, $\{A, B, C, D\}$ each have the following probabilities of producing the four possible uppermost faces: $P(1) = P(6) = 0.1$, $P(3) = P(4) = 0.4$.

   (a) Suppose that we write $A = 1$ to indicate the event that A produces side 1, etc. Compute $P(A = 1, B = 3, C = 4, D = 6)$.

   (b) Compute $P(A = 1, B = 6, C = 3, D = 4)$.

   (c) What is the probability that one throw of these four astragali will produce a venus?

   Hint: See Exercise 2.5.3.

   (d) For $k = 2$, $k = 3$, and $k = 100$, what is the probability that k throws of these four astragali will produce a run of k venuses?

4. Suppose that each of five astragali have the probabilities specified in the previous exercise. When throwing these five astagali,

   (a) What is the probability of obtaining the throw of child-eating Cronos, i.e., of obtaining three fours and two sixes?

   (b) What is the probability of obtaining the throw of Saviour Zeus, i.e., of obtaining one one, two threes, and two fours?

Hint: See Exercise 2.5.4.

5. Koko (a cat) is trying to catch a mouse who lives under Susan's house. The mouse has two exits, one outside and one inside, and randomly selects the outside exit 60% of the time. Each midnight, the mouse emerges for a constitutional. If Koko waits outside and the mouse chooses the outside exit, then Koko has a 20% chance of catching the mouse. If Koko waits inside, then there is a 30% chance that he will fall asleep. However, if he stays awake and the mouse chooses the inside exit, then Koko has a 40% chance of catching the mouse.

   (a) Is Koko more likely to catch the mouse if he waits inside or outside? Why?

   (b) If Koko decides to wait outside each midnight, then what is the probability that he will catch the mouse within a week (no more than 7 nights)?

6. Three urns each contain ten gems:

   - Urn 1 contains 6 rubies and 4 emeralds.
   - Urn 2r contains 8 rubies and 2 emeralds.
   - Urn 2e contains 4 rubies and 6 emeralds.

   The following procedure is used to select two gems. First, one gem is drawn at random from urn 1. If this first gem is a ruby, then a second gem is drawn at random from urn 2r; however, if the first gem is an emerald, then the second gem is drawn at random from urn 2e.

   (a) Construct a tree diagram that describes this procedure.

   (b) What is the probability that a ruby is obtained on the second draw?

   (c) Suppose that the second gem is a ruby. What then is the probability that the first gem was also a ruby?

   (d) Suppose that this procedure is independently replicated three times. What is the probability that a ruby is obtained on the second draw exactly once?

   (e) Suppose that this procedure is independently replicated three times and that a ruby is obtained on the second draw each time. What then is the probability that the first gem was a ruby each time?

7. Arlen is planning a dinner party at which he will be able to accommodate seven guests. From past experience, he knows that each person invited to the party will accept his invitation with probability 0.5. He also knows that each person who accepts will actually attend with probability 0.8. Suppose that Arlen invites twelve people. Assuming that they behave independently of one another, what is the probability that he will end up with more guests than he can accommodate?

8. Hotels that host conferences routinely overbook their rooms because some people who plan to attend conferences fail to arrive. A common assumption is that 10 percent of the hotel rooms reserved by conference attendees will not be claimed. In contrast, only 4 percent of the persons who reserve hotel rooms for the annual Joint Statistical Meetings (JSM) fail to claim them.

   Suppose that a certain hotel has 100 rooms. Incorrectly believing that statisticians behave like normal people, the hotel accepts 110 room reservations for JSM. What is the probability that the hotel will have to turn away statisticians who have reserved rooms?

9. A small liberal arts college receives applications for admission from 1000 high school seniors. The college has dormitory space for a freshman class of 95 students and will have to arrange for off-campus housing for any additional freshmen. In previous years, an average of 64 percent of the students that the college has accepted have elected to attend another school. Clearly the college should accept more than 95 students, but its administration does not want to take too big a chance that it will have to accommodate more than 95 students. After some delibration, the administrators decide to accept 225 students. Answer the following questions as well as you can with the information provided.

   (a) How many freshmen do you expect that the college will have to accommodate?

   (b) What is the the probability that the college will have to arrange for some freshmen to live off-campus?

10. In NCAA tennis matches, line calls are made by the players. If an umpire is observing the match, then a player can challenge an opponent's call. The umpire will either affirm or overrule the challenged call. In one of their recent team matches, the William & Mary women's tennis

team challenged 38 calls by their opponents. The umpires overruled 12 of the challenged calls. This struck Nina and Delphine as significant, as it is their impression that approximately 20 percent of challenged calls are overruled. Let us assume that their impression is correct.

(a) What is the probability that chance variation would result in at least 12 of 38 challenged calls being overruled?

(b) Suppose that the William & Mary women's tennis team plays 25 team matches next year and challenges exactly 38 calls in each match. (In fact, the number of challenged calls varies from match to match.) What is the probability that they will play at least one team match in which at least 12 challenged calls are overruled?

11. The Association for Research and Enlightenment (ARE) in Virginia Beach, VA, offers daily demonstrations of a standard technique for testing extrasensory perception (ESP). A "sender" is seated before a box on which one of five symbols (plus, square, star, circle, wave) can be illuminated. A random mechanism selects symbols in such a way that each symbol is equally likely to be illuminated. When a symbol is illuminated, the sender concentrates on it and a "receiver" attempts to identify which symbol has been selected. The receiver indicates a symbol on the receiver's box, which sends a signal to the sender's box that cues it to select and illuminate another symbol. This process of illuminating, sending, and receiving a symbol is repeated 25 times. Each selection of a symbol to be illuminated is independent of the others. The receiver's score (for a set of 25 trials) is the number of symbols that s/he correctly identifies. For the purpose of this exercise, please suppose that ESP does not exist.

(a) How many symbols should we expect the receiver to identify correctly?

(b) The ARE considers a score of more than 7 matches to be indicative of ESP. What is the probability that the receiver will provide such an indication?

(c) The ARE provides all audience members with scoring sheets and invites them to act as receivers. Suppose that, as on August 31, 2002, there are 21 people in attendance: 1 volunteer sender, 1 volunteer receiver, and 19 additional receivers in the audience. What is the probability that at least one of the 20 receivers will attain a score indicative of ESP?

12. Mike teaches two sections of Applied Statistics each year for thirty years, for a total of 1500 students. Each of his students spins a penny 89 times and counts the number of `Heads`. Assuming that each of these 1500 pennies has $P(\texttt{Heads}) = 0.3$ for a single spin, what is the probability that Mike will encounter at least one student who observes no more than two `Heads`?

Chapter 5

Continuous Random Variables

5.1 A Motivating Example

Some of the concepts that were introduced in Chapter 4 pose technical difficulties when the random variable is not discrete. In this section, we illustrate some of these difficulties by considering a random variable X whose set of possible values is the unit interval, i.e., $X(S) = [0, 1]$. Specifically, we ask the following question:

> *What probability distribution formalizes the notion of "equally likely" outcomes in the unit interval* $[0, 1]$*?*

When studying finite sample spaces in Section 3.3, we formalized the notion of "equally likely" by assigning the same probability to each individual outcome in the sample space. Thus, if $S = \{s_1, \ldots, s_N\}$, then $P(\{s_i\}) = 1/N$. This construction sufficed to define probabilities of events: if $E \subset S$, then

$$E = \{s_{i_1}, \ldots, s_{i_k}\};$$

and consequently

$$P(E) = P\left(\bigcup_{j=1}^{k} \left\{s_{i_j}\right\}\right) = \sum_{j=1}^{k} P\left(\left\{s_{i_j}\right\}\right) = \sum_{j=1}^{k} \frac{1}{N} = \frac{k}{N}.$$

Unfortunately, the present example does not work out quite so neatly.

How should we assign $P(X = 0.5)$? Of course, we must have $0 \leq P(X = 0.5) \leq 1$. If we try $P(X = 0.5) = \epsilon$ for any real number $\epsilon > 0$, then a difficulty arises. Because we are assuming that every value in the unit interval is equally likely, it must be that $P(X = x) = \epsilon$ for *every* $x \in [0, 1]$. Consider the event

$$E = \left\{ \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots \right\}. \tag{5.1}$$

Then we must have

$$P(E) = P\left(\bigcup_{j=2}^{\infty} \left\{ \frac{1}{j} \right\} \right) = \sum_{j=2}^{\infty} P\left(\left\{ \frac{1}{j} \right\} \right) = \sum_{j=2}^{\infty} \epsilon = \infty, \tag{5.2}$$

which we cannot allow. Hence, we *must* assign a probability of zero to the outcome $x = 0.5$ and, because all outcomes are equally likely, $P(X = x) = 0$ for every $x \in [0, 1]$.

Because every $x \in [0, 1]$ is a possible outcome, our conclusion that $P(X = x) = 0$ is initially somewhat startling. However, it is a mistake to identify impossibility with zero probability. In Section 3.2, we established that the impossible event (empty set) has probability zero, but we did *not* say that it is the only such event. To avoid confusion, we now emphasize:

> *If an event is impossible, then it necessarily has probability zero; however, having probability zero does not necessarily mean that an event is impossible.*

If $P(X = x) = \epsilon = 0$, then the calculation in (5.2) reveals that the event defined by (5.1) has probability zero. Furthermore, there is nothing special about this particular event—the probability of *any* countable event must be zero! Hence, to obtain positive probabilities, e.g., $P(X \in [0, 1]) = 1$, we must consider events whose cardinality is more than countable.

Consider the events $[0, 0.5]$ and $[0.5, 1]$. Because all outcomes are equally likely, these events must have the same probability, i.e.,

$$P\left(X \in [0, 0.5]\right) = P\left(X \in [0.5, 1]\right).$$

Because $[0, 0.5] \cup [0.5, 1] = [0, 1]$ and $P(X = 0.5) = 0$, we have

$$\begin{aligned} 1 = P\left(X \in [0, 1]\right) &= P\left(X \in [0, 0.5]\right) + P\left(X \in [0.5, 1]\right) - P\left(X = 0\right) \\ &= P\left(X \in [0, 0.5]\right) + P\left(X \in [0.5, 1]\right). \end{aligned}$$

Combining these equations, we deduce that each event has probability $1/2$. This is an intuitively pleasing conclusion: it says that, if outcomes are equally

likely, then the probability of each subinterval equals the proportion of the entire interval occupied by the subinterval. In mathematical notation, our conclusion can be expressed as follows:

> *Suppose that* $X(S) = [0, 1]$ *and each* $x \in [0, 1]$ *is equally likely. If* $0 \leq a \leq b \leq 1$, *then* $P(X \in [a, b]) = b - a$.

Notice that statements like $P(X \in [0, 0.5]) = 0.5$ cannot be deduced from knowledge that each $P(X = x) = 0$. To construct a probability distribution for this situation, it is necessary to assign probabilities to intervals, not just to individual points. This fact reveals the reason that, in Section 3.2, we introduced the concept of an event and insisted that probabilities be assigned to events rather than to outcomes.

The probability distribution that we have constructed is called the *continuous uniform distribution* on the interval $[0, 1]$, denoted Uniform$[0, 1]$. If $X \sim$ Uniform$[0, 1]$, then the cdf of X is easily computed:

- If $y < 0$, then

$$\begin{aligned} F(y) &= P(X \leq y) \\ &= P(X \in (-\infty, y]) \\ &= 0. \end{aligned}$$

- If $y \in [0, 1]$, then

$$\begin{aligned} F(y) &= P(X \leq y) \\ &= P(X \in (-\infty, 0)) + P(X \in [0, y]) \\ &= 0 + (y - 0) \\ &= y. \end{aligned}$$

- If $y > 1$, then

$$\begin{aligned} F(y) &= P(X \leq y) \\ &= P(X \in (-\infty, 0)) + P(X \in [0, 1]) + P(X \in (1, y)) \\ &= 0 + (1 - 0) + 0 \\ &= 1. \end{aligned}$$

This function is plotted in Figure 5.1.

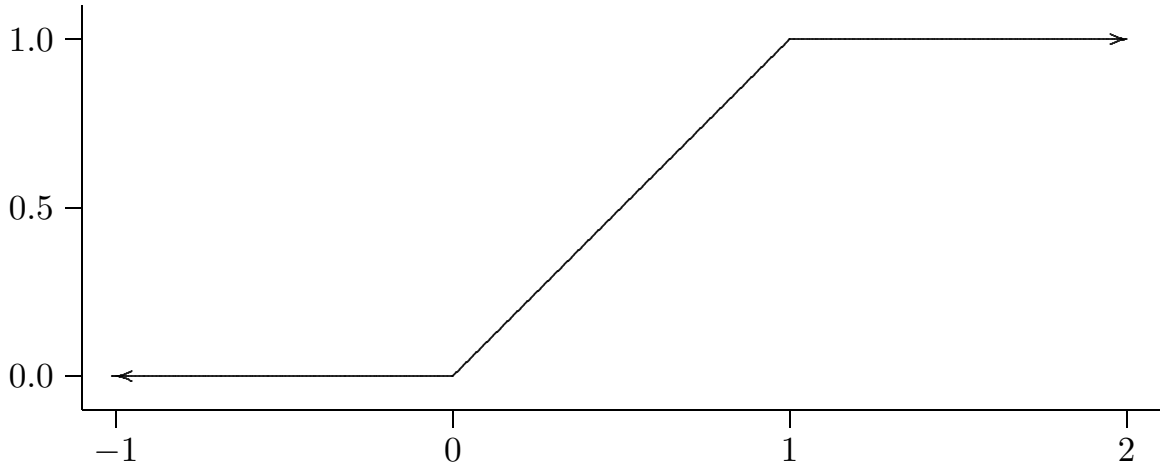

Figure 5.1: The cumulative distribution function of $X \sim \text{Uniform}(0, 1)$.

What about the pmf of X? In Section 4.1, we defined the pmf of a discrete random variable by $f(x) = P(X = x)$; we then used the pmf to calculate the probabilities of arbitrary events. In the present situation, $P(X = x) = 0$ for every x, so the pmf is not very useful. Instead of representing the probabilites of individual points, we need to represent the probabilities of intervals.

Consider the function

$$f(x) = \left\{ \begin{array}{ll} 0 & x \in (-\infty, 0) \\ 1 & x \in [0, 1] \\ 0 & x \in (1, \infty) \end{array} \right\}, \tag{5.3}$$

which is plotted in Figure 5.2. Notice that f is constant on $X(S) = [0, 1]$, the set of equally likely possible values, and vanishes elsewhere. If $0 \leq a \leq b \leq 1$, then the area under the graph of f between a and b is the area of a rectangle with sides $b - a$ (horizontal direction) and 1 (vertical direction). Hence, the area in question is

$$(b - a) \cdot 1 = b - a = P(X \in [a, b]),$$

so that the probabilities of intervals can be determined from f. In the next section, we will base our definition of continuous random variables on this observation.

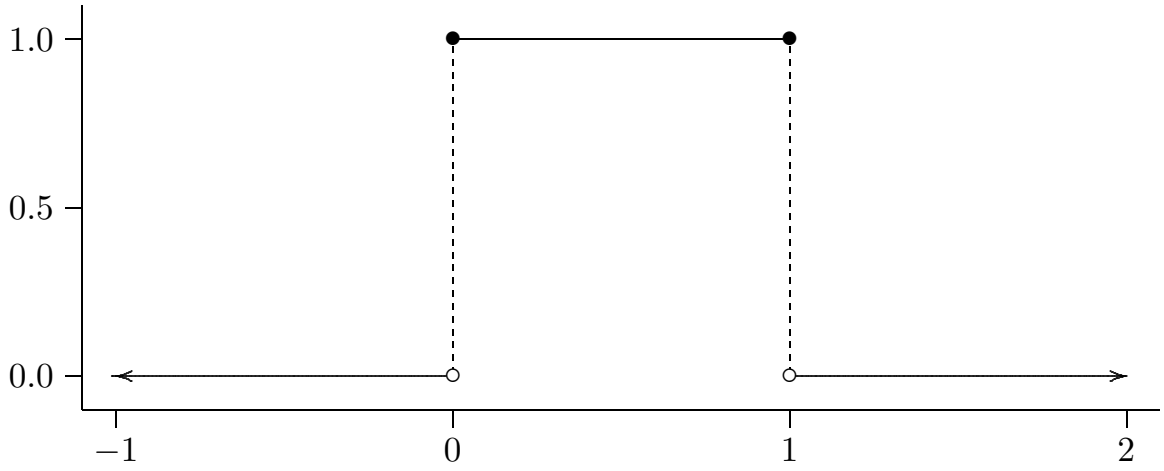

Figure 5.2: The probability density function of $X \sim \text{Uniform}(0, 1)$.

5.2 Basic Concepts

Consider the graph of a function $f : \Re \to \Re$, as depicted in Figure 5.3. Our interest is in the area of the shaded region. This region is bounded by the graph of f, the horizontal axis, and vertical lines at the specified endpoints a and b. We denote this area by $\text{Area}_{[a,b]}(f)$. Our intent is to identify such areas with the probabilities that random variables assume certain values.

For a very few functions, such as the one defined in (5.3), it is possible to determine $\text{Area}_{[a,b]}(f)$ by elementary geometric calculations. For most functions, some knowledge of calculus is required to determine $\text{Area}_{[a,b]}(f)$. Because we assume no previous knowledge of calculus, we will not be concerned with such calculations. Nevertheless, for the benefit of those readers who know some calculus, we find it helpful to borrow some notation and write

$$\text{Area}_{[a,b]}(f) = \int_a^b f(x)dx. \tag{5.4}$$

Readers who have no knowledge of calculus should interpret (5.4) as a definition of its right-hand side, which is pronounced "the integral of f from a to b". Readers who are familiar with the Riemann (or Lebesgue) integral should interpret this notation in its conventional sense.

We now introduce an alternative to the probability mass function.

Definition 5.1 *A probability density function (pdf) is a function $f : \Re \to \Re$ such that*

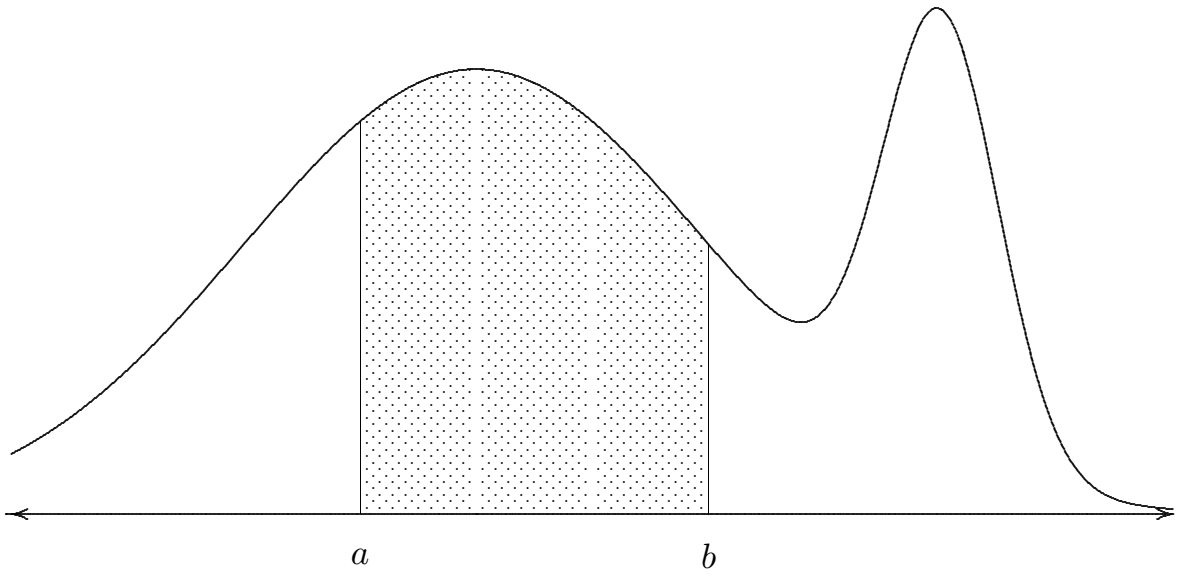

Figure 5.3: A continuous probability density function.

1. $f(x) \geq 0$ *for every* $x \in \Re$.

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

Notice that the definition of a pdf is analogous to the definition of a pmf. Each is nonnegative and assigns unit probability to the set of possible values. The only difference is that summation in the definition of a pmf is replaced with integration in the case of a pdf.

Definition 5.1 was made without reference to a random variable—we now use it to define a new class of random variables.

Definition 5.2 *A random variable X is continuous if there exists a probability density function f such that*

$$P\left(X \in [a,b]\right) = \int_a^b f(x)dx.$$

It is immediately apparent from this definition that the cdf of a continuous random variable X is

$$F(y) = P(X \leq y) = P\left(X \in (-\infty, y]\right) = \int_{-\infty}^{y} f(x)dx. \tag{5.5}$$

Equation (5.5) should be compared to equation (4.1). In both cases, the value of the cdf at y is represented as the accumulation of values of the pmf/pdf at $x \leq y$. The difference lies in the nature of the accumulating process: summation for the discrete case (pmf), integration for the continuous case (pdf).

> **Remark for Calculus Students:** By applying the Fundamental Theorem of Caluclus to (5.5), we deduce that the pdf of a continuous random variable is the derivative of its cdf:
>
> $$\frac{d}{dy}F(y) = \frac{d}{dy}\int_{-\infty}^{y} f(x)dx = f(y).$$

> **Remark on Notation:** It may strike the reader as curious that we have used f to denote both the pmf of a discrete random variable and the pdf of a continuous random variable. However, as our discussion of their relation to the cdf is intended to suggest, they play analogous roles. In advanced, *measure-theoretic* courses on probability, one learns that our pmf and pdf are actually two special cases of one general construction.

Likewise, the concept of expectation for continuous random variables is analogous to the concept of expectation for discrete random variables. Because $P(X = x) = 0$ if X is a continuous random variable, the notion of a probability-weighted average is not very useful in the continuous setting. However, if X is a discrete random variable, then $P(X = x) = f(x)$ and a probability-weighted average is identical to a pmf-weighted average. The notion of a pmf-weighted average *is* easily extended to the continuous setting: if X is a continuous random variable, then we introduce a pdf-weighted average of the possible values of X, where averaging is accomplished by replacing summation with integration.

Definition 5.3 *Suppose that X is a continuous random variable with probability density function f. Then the* expected value *of X is*

$$\mu = EX = \int_{-\infty}^{\infty} xf(x)dx,$$

assuming that this quantity exists.

If the function $g : \Re \to \Re$ is such that $Y = g(X)$ is a random variable, then it can be shown that

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

assuming that this quantity exists. In particular,

Definition 5.4 *If $\mu = EX$ exists and is finite, then the* variance *of X is*

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

Thus, for discrete *and* continuous random variables, the expected value is the pmf/pdf-weighted average of the possible values and the variance is the pmf/pdf-weighted average of the squared deviations of the possible values from the expected value.

Because calculus is required to compute the expected value and variance of most continuous random variables, our interest in these concepts lies not in computing them but in understanding what information they convey. We will return to this subject in Chapter 6.

5.3 Elementary Examples

In this section we consider some examples of continuous random variables for which probabilities can be calculated without recourse to calculus.

Example 5.1 *What is the probability that a battery-powered wristwatch will stop with its minute hand positioned between* 10 *and* 20 *minutes past the hour?*

To answer this question, let X denote the number of minutes past the hour to which the minute hand points when the watch stops. Then the possible values of X are $X(S) = [0, 60)$ and it is reasonable to assume that each value is equally likely. We must compute $P(X \in (10, 20))$. Because these values occupy one sixth of the possible values, it should be obvious that the answer is going to be $1/6$.

To obtain the answer using the formal methods of probability, we require a generalization of the Uniform$[0, 1]$ distribution that we studied in Section 5.1. The pdf that describes the notion of equally likely values in the interval

$[0, 60)$ is

$$f(x) = \left\{ \begin{array}{rl} 0 & x \in (-\infty, 0) \\ 1/60 & x \in [0, 60) \\ 0 & x \in [60, \infty) \end{array} \right\}. \tag{5.6}$$

To check that f is really a pdf, observe that $f(x) \geq 0$ for every $x \in \Re$ and that

$$\text{Area}_{[0,60)}(f) = (60 - 0)\frac{1}{60} = 1.$$

Notice the analogy between the pdfs (5.6) and (5.3). The present pdf defines the continuous uniform distribution on the interval $[0, 60)$; thus, we describe the present situation by writing $X \sim \text{Uniform}[0, 60)$. To calculate the specified probability, we must determine the area of the shaded region in Figure 5.4, i.e.,

$$P(X \in (10, 20)) = \text{Area}_{(10,20)}(f) = (20 - 10)\frac{1}{60} = \frac{1}{6}.$$

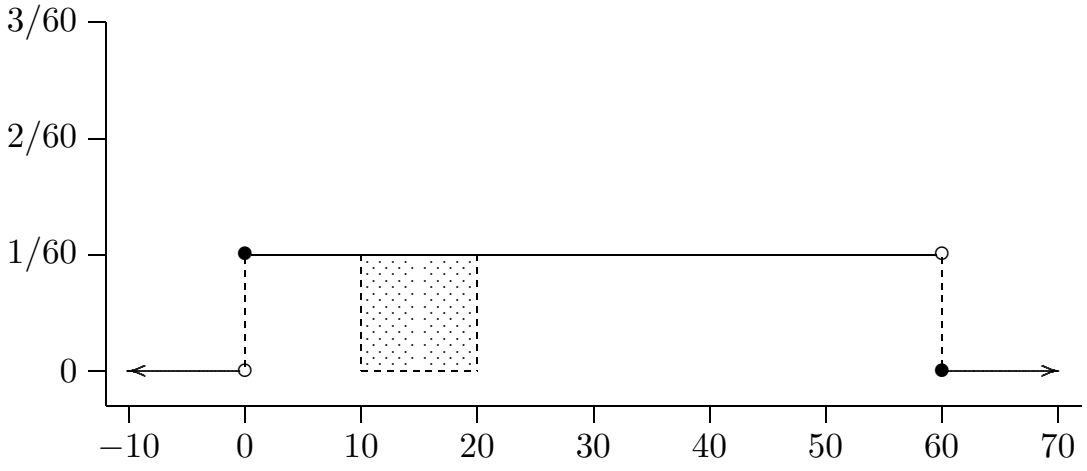

Figure 5.4: The probability density function of $X \sim \text{Uniform}[0, 60)$.

Example 5.2 *Consider two battery-powered watches. Let X_1 denote the number of minutes past the hour at which the first watch stops and let X_2 denote the number of minutes past the hour at which the second watch stops. What is the probability that the larger of X_1 and X_2 will be between* 30 *and* 50*?*

Here we have two independent random variables, each distributed as Uniform$[0, 60)$, and a third random variable,

$$Y = \max(X_1, X_2).$$

Let F denote the cdf of Y. We want to calculate

$$P(30 < Y < 50) = F(50) - F(30).$$

We proceed to derive the cdf of Y. It is evident that $Y(S) = [0, 60)$, so $F(y) = 0$ if $y < 0$ and $F(y) = 1$ if $y \geq 1$. If $y \in [0, 60)$, then (by the independence of X_1 and X_2)

$$\begin{aligned} F(y) = P(Y \leq y) &= P\left(\max(X_1, X_2) \leq y\right) = P\left(X_1 \leq y,\ X_2 \leq y\right) \\ &= P\left(X_1 \leq y\right) \cdot P\left(X_2 \leq y\right) = \frac{y-0}{60-0} \cdot \frac{y-0}{60-0} \\ &= \frac{y^2}{3600}. \end{aligned}$$

Thus, the desired probability is

$$P(30 < Y < 50) = F(50) - F(30) = \frac{50^2}{3600} - \frac{30^2}{3600} = \frac{4}{9}.$$

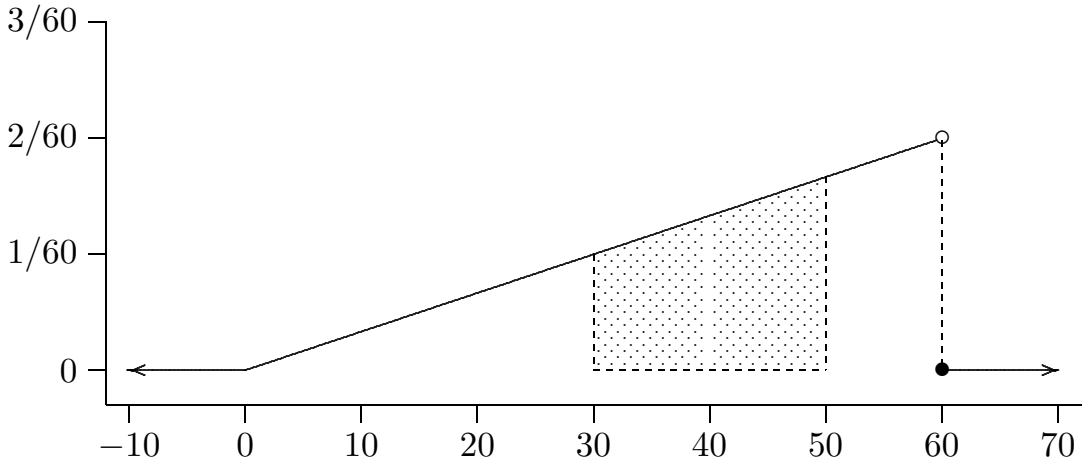

Figure 5.5: The probability density function for Example 5.2.

In preparation for Example 5.3, we claim that the pdf of Y is

$$f(y) = \left\{ \begin{array}{rl} 0 & y \in (-\infty, 0) \\ y/1800 & y \in [0, 60) \\ 0 & y \in [60, \infty) \end{array} \right\},$$

which is graphed in Figure 5.5. To check that f is really a pdf, observe that $f(y) \geq 0$ for every $y \in \Re$ and that

$$\text{Area}_{[0,60)}(f) = \frac{1}{2}(60 - 0)\frac{60}{1800} = 1.$$

To check that f is really the pdf of Y, observe that $f(y) = 0$ if $y \notin [0, 60)$ and that, if $y \in [0, 60)$, then

$$P\left(Y \in [0, y)\right) = P(Y \leq y) = F(y) = \frac{y^2}{3600} = \frac{1}{2}(y - 0)\frac{y}{1800} = \text{Area}_{[0,y)}(f).$$

If the pdf had been specified, then instead of deriving the cdf we would have simply calculated

$$P(30 < Y < 50) = \text{Area}_{(30,50)}(f)$$

by any of several convenient geometric arguments.

Example 5.3 *Consider two battery-powered watches. Let X_1 denote the number of minutes past the hour at which the first watch stops and let X_2 denote the number of minutes past the hour at which the second watch stops. What is the probability that the sum of X_1 and X_2 will be between 45 and 75?*

Again we have two independent random variables, each distributed as Uniform$[0, 60)$, and a third random variable,

$$Z = X_1 + X_2.$$

We want to calculate

$$P(45 < Z < 75) = P\left(Z \in (45, 75)\right).$$

It is apparent that $Z(S) = [0, 120)$. Although we omit the derivation, it can be determined mathematically that the pdf of Z is

$$f(z) = \left\{ \begin{array}{rl} 0 & z \in (-\infty, 0) \\ z/3600 & z \in [0, 60) \\ (120 - z)/3600 & z \in [60, 120) \\ 0 & z \in [120, \infty) \end{array} \right\}.$$

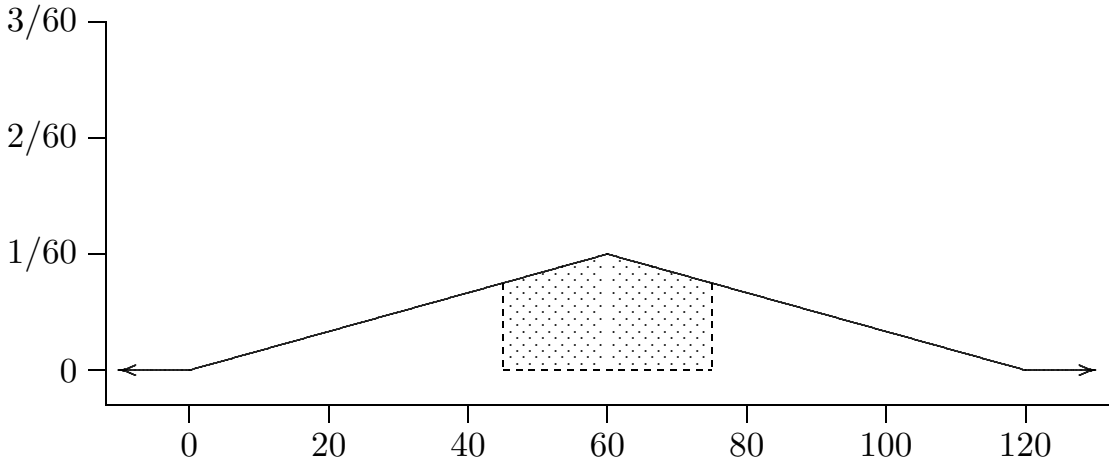

Figure 5.6: The probability density function for Example 5.3.

This pdf is graphed in Figure 5.6, in which it is apparent that the area of the shaded region is

$$\begin{aligned} P(45 < Z < 75) &= P\left(Z \in (45, 75)\right) = \text{Area}_{(45,75)}(f) \\ &= 1 - \frac{1}{2}(45 - 0)\frac{45}{3600} - \frac{1}{2}(120 - 75)\frac{120 - 75}{3600} \\ &= 1 - \frac{45^2}{60^2} = \frac{7}{16}. \end{aligned}$$

5.4 Normal Distributions

We now introduce the most important family of distributions in probability or statistics, the familiar *bell-shaped curve.*

Definition 5.5 *A continuous random variable X is normally distributed with mean μ and variance $\sigma^2 > 0$, denoted $X \sim Normal(\mu, \sigma^2)$, if the pdf of X is*

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x - \mu}{\sigma}\right)^2\right]. \tag{5.7}$$

Although we will not make extensive use of (5.7), a great many useful properties of normal distributions can be deduced directly from it. Most of the following properties can be discerned in Figure 5.7.

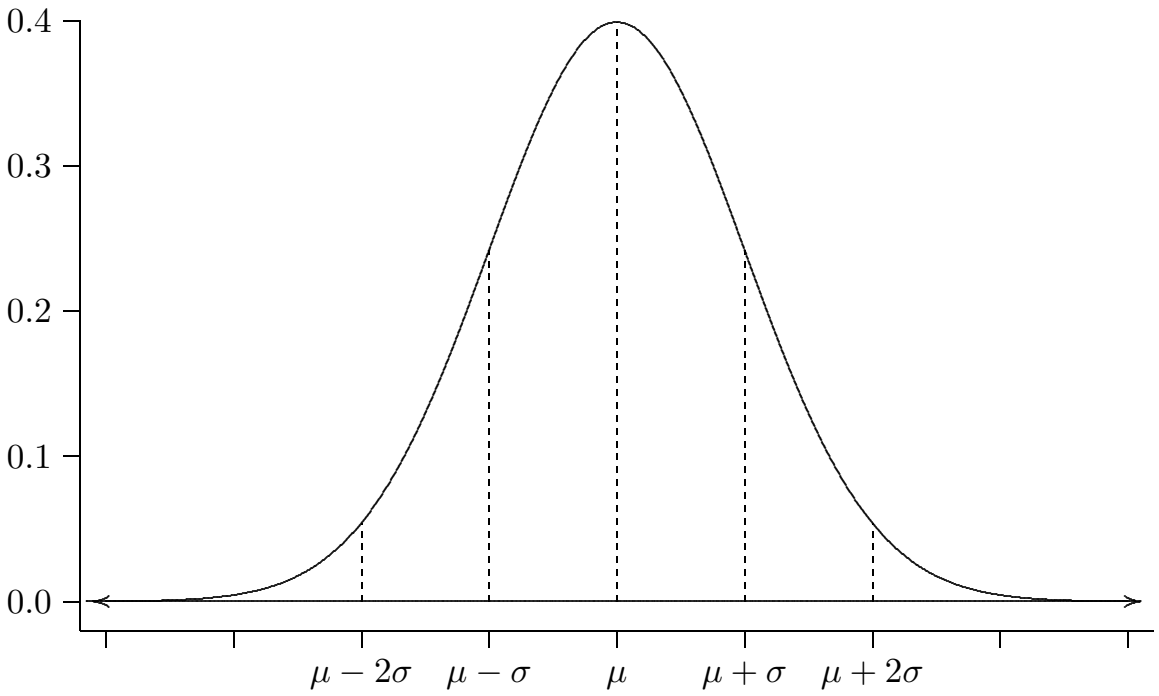

Figure 5.7: The probability density function of $X \sim \text{Normal}(\mu, \sigma^2)$.

1. $f(x) > 0$. It follows that, for any nonempty interval (a, b),

$$P\left(X \in (a, b)\right) = \text{Area}_{(a,b)}(f) > 0,$$

   and hence that $X(S) = (-\infty, +\infty)$.

2. f is symmetric about μ, i.e., $f(\mu + x) = f(\mu - x)$.

3. $f(x)$ decreases as $|x - \mu|$ increases. In fact, the decrease is very rapid. We express this by saying that f has very light tails.

4. $P(\mu - \sigma < X < \mu + \sigma) \doteq 0.683$.

5. $P(\mu - 2\sigma < X < \mu + 2\sigma) \doteq 0.954$.

6. $P(\mu - 3\sigma < X < \mu + 3\sigma) \doteq 0.997$.

Notice that there is no one normal distribution, but a 2-parameter family of uncountably many normal distributions. In fact, if we plot μ on a horizontal axis and $\sigma > 0$ on a vertical axis, then there is a distinct normal distribution for each point in the upper half-plane. However, Properties 4–6

above, which hold for *all* choices of μ and σ, suggest that there is a fundamental equivalence between different normal distributions. It turns out that, if one can compute probabilities for any one normal distribution, then one can compute probabilities for any other normal distribution. In anticipation of this fact, we distinguish one normal distribution to serve as a reference distribution:

Definition 5.6 *The standard normal distribution is Normal*$(0, 1)$.

The following result is of enormous practical value:

Theorem 5.1 *If* $X \sim Normal(\mu, \sigma^2)$, *then*

$$Z = \frac{X - \mu}{\sigma} \sim Normal(0, 1).$$

The transformation $Z = (X - \mu)/\sigma$ is called conversion to standard units.

Detailed tables of the standard normal cdf are widely available, as is computer software for calculating specified values. Combined with Theorem 5.1, this availability allows us to easily compute probabilities for arbitrary normal distributions. In the following examples, we let Φ denote the cdf of $Z \sim \text{Normal}(0, 1)$ and we make use of the `R` function `pnorm`.

Example 5.4a *If* $X \sim Normal(1, 4)$, *then what is the probability that* X *assumes a value no more than* 3*?*

Here, $\mu = 1$, $\sigma = 2$, and we want to calculate

$$P(X \leq 3) = P\left(\frac{X - \mu}{\sigma} \leq \frac{3 - \mu}{\sigma}\right) = P\left(Z \leq \frac{3 - 1}{2} = 1\right) = \Phi(1).$$

We do so in `R` as follows:

```
> pnorm(1)
[1] 0.8413447
```

Remark The `R` function `pnorm` accepts optional arguments that specify a mean and standard deviation. Thus, in Example 5.4a, we could directly evaluate $P(X \leq 3)$ as follows:

```
> pnorm(3,mean=1,sd=2)
[1] 0.8413447
```

This option, of course, is not available if one is using a table of the standard normal cdf. Because the transformation to standard units plays such a fundamental role in probability and statistics, we will emphasize computing normal probabilities via the standard normal distribution.

Example 5.4b *If $X \sim Normal(-1, 9)$, then what is the probability that X assumes a value of at least -7?*

Here, $\mu = -1$, $\sigma = 3$, and we want to calculate

$$\begin{aligned} P(X \geq -7) &= P\left(\frac{X-\mu}{\sigma} \geq \frac{-7-\mu}{\sigma}\right) \\ &= P\left(Z \geq \frac{-7+1}{3} = -2\right) \\ &= 1 - P(Z < -2) \\ &= 1 - \Phi(-2). \end{aligned}$$

We do so in `R` as follows:

```
> 1-pnorm(-2)
[1] 0.9772499
```

Example 5.4c *If $X \sim Normal(2, 16)$, then what is the probability that X assumes a value between 0 and 10?*

Here, $\mu = 2$, $\sigma = 4$, and we want to calculate

$$\begin{aligned} P(0 < X < 10) &= P\left(\frac{0-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{10-\mu}{\sigma}\right) \\ &= P\left(-0.5 = \frac{0-2}{4} < Z < \frac{10-2}{4} = 2\right) \\ &= P(Z < 2) - P(Z < -0.5) \\ &= \Phi(2) - \Phi(-0.5). \end{aligned}$$

We do so in `R` as follows:

```
> pnorm(2)-pnorm(-.5)
[1] 0.6687123
```

Example 5.4d *If $X \sim Normal(-3, 25)$, then what is the probability that $|X|$ assumes a value greater than 10?*

Here, $\mu = -3$, $\sigma = 5$, and we want to calculate

$$\begin{aligned} P(|X| > 10) &= P(X > 10 \text{ or } X < -10) \\ &= P(X > 10) + P(X < -10) \\ &= P\left(\frac{X-\mu}{\sigma} > \frac{10-\mu}{\sigma}\right) + P\left(\frac{X-\mu}{\sigma} < \frac{-10-\mu}{\sigma}\right) \\ &= P\left(Z > \frac{10+3}{5} = 2.6\right) + P\left(Z < \frac{-10+3}{5} = -1.2\right) \\ &= 1 - \Phi(2.6) + \Phi(-1.2). \end{aligned}$$

We do so in R as follows:

```
> 1-pnorm(2.6)+pnorm(-1.2)
[1] 0.1197309
```

Example 5.4e *If $X \sim Normal(4, 16)$, then what is the probability that X^2 assumes a value less than 36?*

Here, $\mu = 4$, $\sigma = 4$, and we want to calculate

$$\begin{aligned} P(X^2 < 36) &= P(-6 < X < 6) \\ &= P\left(\frac{-6-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{6-\mu}{\sigma}\right) \\ &= P\left(-2.5 = \frac{-6-4}{4} < Z < \frac{6-4}{4} = 0.5\right) \\ &= P(Z < 0.5) - P(Z < -2.5) \\ &= \Phi(0.5) - \Phi(-2.5). \end{aligned}$$

We do so in R as follows:

```
> pnorm(.5)-pnorm(-2.5)
[1] 0.6852528
```

We defer an explanation of why the family of normal distributions is so important until Section 8.3, concluding the present section with the following useful result:

Theorem 5.2 *If $X_1 \sim Normal(\mu_1, \sigma_1^2)$ and $X_2 \sim Normal(\mu_2, \sigma_2^2)$ are independent, then*

$$X_1 + X_2 \sim Normal(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

5.5 Normal Sampling Distributions

A number of important probability distributions can be derived by considering various functions of normal random variables. These distributions play important roles in statistical inference. They are rarely used to describe data; rather, they arise when analyzing data that is sampled from a normal distribution. For this reason, they are sometimes called *sampling distributions*.

This section collects some definitions of and facts about several important sampling distributions. It is not important to read this section until you encounter these distributions in later chapters; however, it is convenient to collect this material in one easy-to-find place.

Chi-Squared Distributions Suppose that $Z_1, \ldots, Z_n \sim \text{Normal}(0, 1)$ and consider the continuous random variable

$$Y = Z_1^2 + \cdots + Z_n^2.$$

Because each $Z_i^2 \geq 0$, the set of possible values of Y is $Y(S) = [0, \infty)$. We are interested in the distribution of Y.

The distribution of Y belongs to a family of probability distributions called the *chi-squared* family. This family is indexed by a single real-valued parameter, $\nu \in [1, \infty)$, called the *degrees of freedom* parameter. We will denote a chi-squared distribution with ν degrees of freedom by $\chi^2(\nu)$. Figure 5.8 displays the pdfs of several chi-squared distributions.

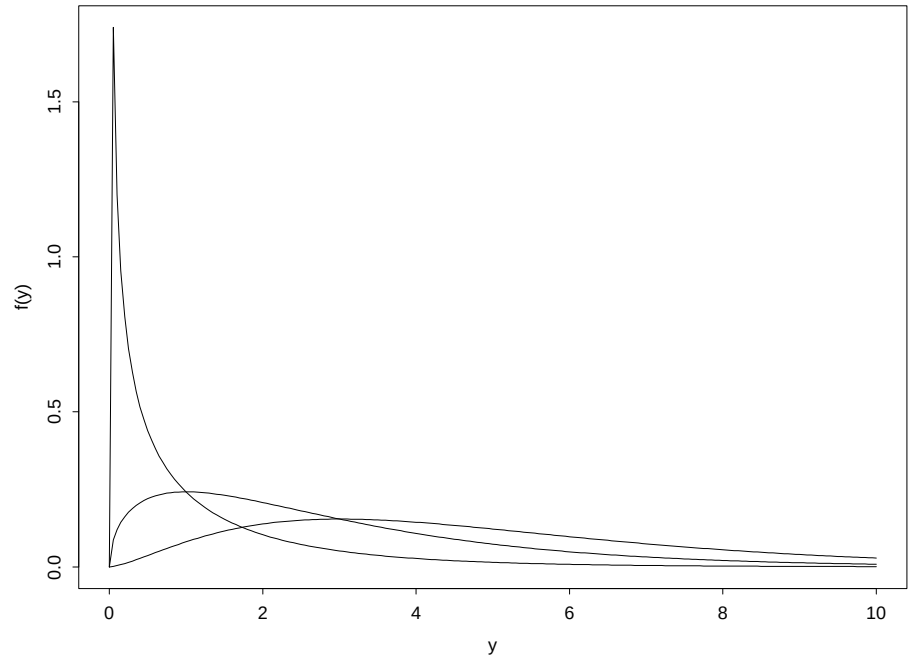

Figure 5.8: The probability density functions of $Y \sim \chi^2(\nu)$ for $\nu = 1, 3, 5$.

The following fact is quite useful:

Theorem 5.3 *If* $Z_1, \ldots, Z_n \sim$ *Normal*$(0, 1)$ *and* $Y = Z_1^2 + \cdots + Z_n^2$, *then* $Y \sim \chi^2(n)$.

In theory, this fact allows one to compute the probabilities of events defined by values of Y, e.g., $P(Y > 4.5)$. In practice, this requires evaluating the

cdf of $\chi^2(\nu)$, a function for which there is no simple formula. Fortunately, there exist efficient algorithms for numerically evaluating these cdfs. The `R` function `pchisq` returns values of the cdf of any specified chi-squared distribution. For example, if $Y \sim \chi^2(2)$, then $P(Y > 4.5)$ is

```
> 1-pchisq(4.5,df=2)
[1] 0.1053992
```

Finally, if $Z_i \sim \text{Normal}(0, 1)$, then

$$EZ_i^2 = \text{Var}\, Z_i + (EZ_i)^2 = 1.$$

It follows that

$$EY = E\left(\sum_{i=1}^{n} Z_i^2\right) = \sum_{i=1}^{n} EZ_i^2 = \sum_{i=1}^{n} 1 = n;$$

thus,

Corollary 5.1 *If $Y \sim \chi^2(n)$, then $EY = n$.*

Student's t Distributions Now let $Z \sim \text{Normal}(0, 1)$ and $Y \sim \chi^2(\nu)$ be independent random variables and consider the continuous random variable

$$T = \frac{Z}{\sqrt{Y/\nu}}.$$

The set of possible values of T is $T(S) = (-\infty, \infty)$. We are interested in the distribution of T.

Definition 5.7 *The distribution of T is called a t distribution with ν degrees of freedom. We will denote this distribution by $t(\nu)$.*

The standard normal distribution is symmetric about the origin; i.e., if $Z \sim \text{Normal}(0, 1)$, then $-Z \sim \text{Normal}(0, 1)$. It follows that $T = Z/\sqrt{Y/\nu}$ and $-T = -Z/\sqrt{Y/\nu}$ have the same distribution. Hence, if p is the pdf of T, then it must be that $p(t) = p(-t)$. Thus, t pdfs are symmetric about the origin, just like the standard normal pdf.

Figure 5.9 displays the pdfs of two t distributions. They can be distinguished by virtue of the fact that the variance of $t(\nu)$ decreases as ν increases. It may strike you that t pdfs closely resemble normal pdfs. In fact, the standard normal pdf is a limiting case of the t pdfs:

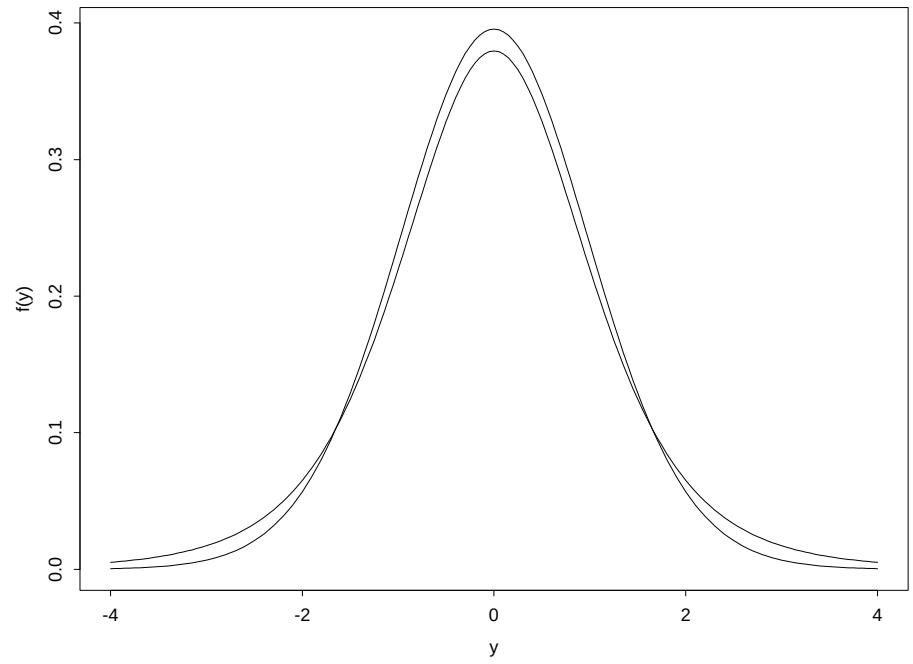

Figure 5.9: The probability density functions of $T \sim t(\nu)$ for $\nu = 5, 30$.

Theorem 5.4 *Let F_ν denote the cdf of $t(\nu)$ and let Φ denote the cdf of Normal$(0, 1)$. Then*

$$\lim_{\nu \to \infty} F_\nu(t) = \Phi(t)$$

for every $t \in (-\infty, \infty)$.

Thus, when ν is sufficiently large ($\nu > 40$ is a reasonable rule of thumb), $t(\nu)$ is approximately Normal$(0, 1)$ and probabilities involving the former can be approximated by probabilities involving the latter.

In `R`, it is just as easy to calculate $t(\nu)$ probabilities as it is to calculate Normal$(0, 1)$ probabilities. The `R` function `pt` returns values of the cdf of any specified t distribution. For example, if $T \sim t(14)$, then $P(T \leq -1.5)$ is

```
> pt(-1.5,df=14)
[1] 0.07791266
```

Fisher's F Distributions Finally, let $Y_1 \sim \chi^2(\nu_1)$ and $Y_2 \sim \chi^2(\nu_2)$ be independent random variables and consider the continuous random variable

$$F = \frac{Y_1/\nu_1}{Y_2/\nu_2}.$$

Because $Y_i \geq 0$, the set of possible values of F is $F(S) = [0, \infty)$. We are interested in the distribution of F.

Definition 5.8 *The distribution of F is called an F distribution with ν_1 and ν_2 degrees of freedom. We will denote this distribution by $F(\nu_1, \nu_2)$. It is customary to call ν_1 the "numerator" degrees of freedom and ν_2 the "denominator" degrees of freedom.*

Figure 5.10 displays the pdfs of several F distributions.

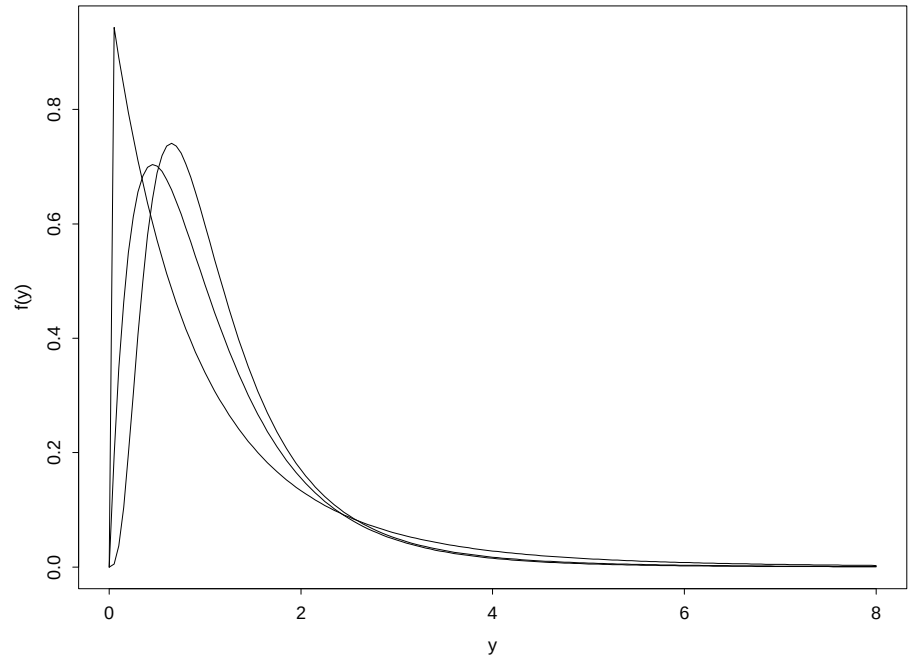

Figure 5.10: The probability density functions of $F \sim F(\nu_1, \nu_2)$ for $(\nu_1, \nu_2) = (2, 12), (4, 20), (9, 10)$.

There is an important relation between t and F distributions. To anticipate it, suppose that $Z \sim \text{Normal}(0, 1)$ and $Y_2 \sim \chi^2(\nu_2)$ are independent

random variables. Then $Y_1 = Z^2 \sim \chi^2(1)$, so

$$T = \frac{Z}{\sqrt{Y_2/\nu_2}} \sim t\left(\nu_2\right)$$

and

$$T^2 = \frac{Z^2}{Y_2/\nu_2} = \frac{Y_1/1}{Y_2/\nu_2} \sim F\left(1, \nu_2\right).$$

More generally,

Theorem 5.5 *If $T \sim t(\nu)$, then $T^2 \sim F(1, \nu)$.*

The `R` function `pf` returns values of the cdf of any specified F distribution. For example, if $F \sim F(2, 27)$, then $P(F > 2.5)$ is

```
> 1-pf(2.5,df1=2,df2=27)
[1] 0.1008988
```

5.6 Exercises

1. In this problem you will be asked to examine two equations. Several symbols from each equation will be identified. Your task will be to decide which symbols represent real numbers and which symbols represent functions. If a symbol represents a function, then you should state the domain and the range of that function.

   Recall: A function is a rule of assignment. The set of labels that the function might possibly assign is called the range of the function; the set of objects to which labels are assigned is called the domain. For example, when I grade your test, I assign a numeric value to your name. Grading is a function that assigns real numbers (the range) to students (the domain).

   (a) In the equation $p = P\left(Z > 1.96\right)$, please identify each of the following symbols as a real number or a function:

   i. p
   ii. P
   iii. Z

   (b) In the equation $\sigma^2 = E\left(X - \mu\right)^2$, please identify each of the following symbols as a real number or a function:

i. σ
ii. E
iii. X
iv. μ

2. Suppose that X is a continuous random variable with probability density function (pdf) f defined as follows:

$$f(x) = \left\{ \begin{array}{lll} 0 & \text{if} & x < 1 \\ 2(x-1) & \text{if} & 1 \leq x \leq 2 \\ 0 & \text{if} & x > 2 \end{array} \right\}.$$

   (a) Graph f.
   (b) Verify that f is a pdf.
   (c) Compute $P(1.50 < X < 1.75)$.

3. Consider the function $f : \Re \to \Re$ defined by

$$f(x) = \left\{ \begin{array}{ll} 0 & x < 0 \\ cx & 0 < x < 1.5 \\ c(3-x) & 1.5 < x < 3 \\ 0 & x > 3 \end{array} \right\},$$

   where c is an undetermined constant.

   (a) For what value of c is f a probability density function?
   (b) Suppose that a continuous random variable X has probability density function f. Compute EX. (Hint: Draw a picture of the pdf.)
   (c) Compute $P(X > 2)$.
   (d) Suppose that $Y \sim \text{Uniform}(0, 3)$. Which random variable has the larger variance, X or Y? (Hint: Draw a picture of the two pdfs.)
   (e) Determine and graph the cumulative distribution function of X.

4. Imagine throwing darts at a circular dart board, B. Let us measure the dart board in units for which the radius of B is 1, so that the area of B is π. Suppose that the darts are thrown in such a way that they are certain to hit a point in B, and that each point in B is equally

likely to be hit. Thus, if $A \subset B$, then the probability of hitting a point in A is

$$P(A) = \frac{\text{area}(A)}{\text{area}(B)} = \frac{\text{area}(A)}{\pi}.$$

Define the random variable X to be the distance from the center of B to the point that is hit.

(a) What are the possible values of X?

(b) Compute $P(X \leq 0.5)$.

(c) Compute $P(0.5 < X \leq 0.7)$.

(d) Determine and graph the cumulative distribution function of X.

(e) [Optional—for those who know a little calculus.] Determine and graph the probability density function of X.

5. Let X be a normal random variable with mean $\mu = -5$ and standard deviation $\sigma = 10$. Compute the following:

(a) $P(X < 0)$

(b) $P(X > 5)$

(c) $P(-3 < X < 7)$

(d) $P(|X + 5| < 10)$

(e) $P(|X - 3| > 2)$

Chapter 6

Quantifying Population Attributes

The distribution of a random variable is a mathematical abstraction of the possible outcomes of an experiment. Indeed, having identified a random variable of interest, we will often refer to its distribution as *the population*. If one's goal is to represent an entire population, then one can hardly do better than to display its entire probability mass or density function. Usually, however, one is interested in specific attributes of a population. This is true if only because it is through specific attributes that one comprehends the entire population, but it is also easier to draw inferences about a specific population attribute than about the entire population. Accordingly, this chapter examines several population attributes that are useful in statistics.

We will be especially concerned with measures of centrality and measures of dispersion. The former provide quantitative characterizations of where the "middle" of a population is located; the latter provide quantitative characterizations of how widely the population is spread. We have already introduced one important measure of centrality, the expected value of a random variable (the population mean, μ), and one important measure of dispersion, the standard deviation of a random variable (the population standard deviation, σ). This chapter discusses these measures in greater depth and introduces other, complementary measures.

6.1 Symmetry

We begin by considering the following question:

Where is the "middle" of a normal distribution?

It is quite evident from Figure 5.7 that there is only one plausible answer to this question: if $X \sim \text{Normal}(\mu, \sigma^2)$, then the "middle" of the distribution of X is μ.

Let f denote the pdf of X. To understand why μ is the only plausible middle of f, recall a property of f that we noted in Section 5.4: for any x, $f(\mu+x) = f(\mu-x)$. This property states that f is *symmetric* about μ. It is the property of symmetry that restricts the plausible locations of "middle" to the central value μ.

To generalize the above example of a measure of centrality, we introduce an important qualitative property that a population may or may not possess:

Definition 6.1 *Let X be a continuous random variable with probability density function f. If there exists a value $\theta \in \Re$ such that*

$$f(\theta + x) = f(\theta - x)$$

for every $x \in \Re$, then X is a symmetric random variable and θ is its center of symmetry.

We have already noted that $X \sim \text{Normal}(\mu, \sigma^2)$ has center of symmetry μ. Another example of symmetry is illustrated in Figure 6.1: $X \sim \text{Uniform}[a, b]$ has center of symmetry $(a + b)/2$.

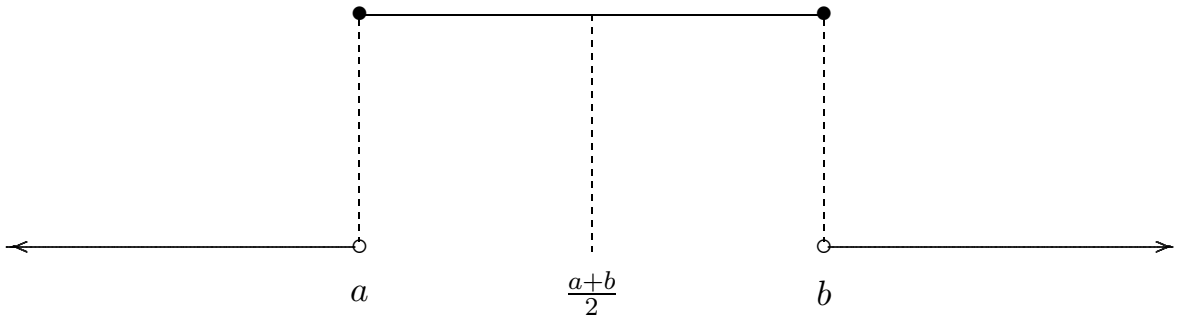

Figure 6.1: $X \sim \text{Uniform}[a, b]$ has center of symmetry $(a + b)/2$.

For symmetric random variables, the center of symmetry is the only plausible measure of centrality—of where the "middle" of the distribution is located. Symmetry will play an important role in our study of statistical

inference. Our primary concern will be with continuous random variables, but the concept of symmetry can be used with other random variables as well. Here is a general definition:

Definition 6.2 *Let X be a random variable. If there exists a value $\theta \in \Re$ such that the random variables $X - \theta$ and $\theta - X$ have the same distribution, then X is a symmetric random variable and θ is its center of symmetry.*

Suppose that we attempt to compute the expected value of a symmetric random variable X with center of symmetry θ. Thinking of the expected value as a weighted average, we see that each $\theta + x$ will be weighted precisely as much as the corresponding $\theta - x$. Thus, if the expect value exists (there are a few pathological random variables for which the expected value is undefined), then it must equal the center of symmetry, i.e., $EX = \theta$. Of course, we have already seen that this is the case for $X \sim \text{Normal}(\mu, \sigma^2)$ and for $X \sim \text{Uniform}[a, b]$.

6.2 Quantiles

In this section we introduce population quantities that can be used for a variety of purposes. As in Section 6.1, these quantities are most easily understood in the case of continuous random variables:

Definition 6.3 *Let X be a continuous random variable and let $\alpha \in (0, 1)$. If $q = q(X; \alpha)$ is such that $P(X < q) = \alpha$ and $P(X > q) = 1 - \alpha$, then q is called an α* quantile *of X.*

If we express the probabilities in Definition 6.3 as percentages, then we see that q is the 100α percentile of the distribution of X.

Example 6.1 Suppose that $X \sim \text{Uniform}[a, b]$ has pdf f, depicted in Figure 6.2. Then q is the value in (a, b) for which

$$\alpha = P(X < q) = \text{Area}_{[a,q]}(f) = (q - a) \cdot \frac{1}{b - a},$$

i.e., $q = a + \alpha(b - a)$. This expression is easily interpreted: to the lower endpoint a, add $100\alpha\%$ of the distance $b - a$ to obtain the 100α percentile.

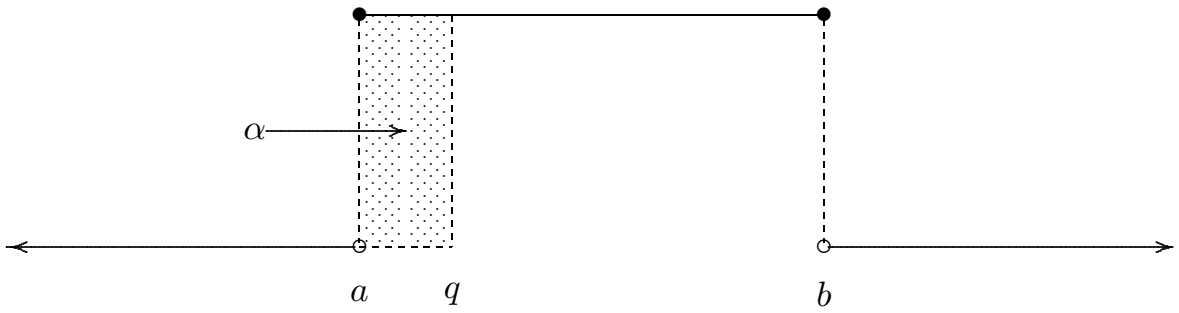

Figure 6.2: A quantile of a Uniform distribution.

Example 6.2 Suppose that X has pdf

$$f(x) = \left\{ \begin{array}{cc} x/2 & x \in [0,2] \\ 0 & \text{otherwise} \end{array} \right\},$$

depicted in Figure 6.3. Then q is the value in $(0,2)$ for which

$$\alpha = P(X < q) = \text{Area}_{[a,q]}(f) = \frac{1}{2} \cdot (q-0) \cdot \left(\frac{q}{2} - 0 \right) = \frac{q^2}{4},$$

i.e., $q = 2\sqrt{\alpha}$.

Example 6.3 Suppose that $X \sim \text{Normal}(0,1)$ has cdf Φ. Then q is the value in $(-\infty, \infty)$ for which $\alpha = P(X < q) = \Phi(q)$, i.e., $q = \Phi^{-1}(\alpha)$. Unlike the previous examples, we cannot compute q by elementary calculations. Fortunately, the `R` function `qnorm` computes quantiles of normal distributions. For example, we compute the $\alpha = 0.95$ quantile of X as follows:

```
> qnorm(.95)
[1] 1.644854
```

Example 6.4 Suppose that X has pdf

$$f(x) = \left\{ \begin{array}{cc} 1/2 & x \in [0,1] \cup [2,3] \\ 0 & \text{otherwise} \end{array} \right\},$$

depicted in Figure 6.4. Notice that $P(X \in [0,1]) = 0.5$ and $P(X \in [2,3]) = 0.5$. If $\alpha \in (0, 0.5)$, then we can use the same reasoning that we employed

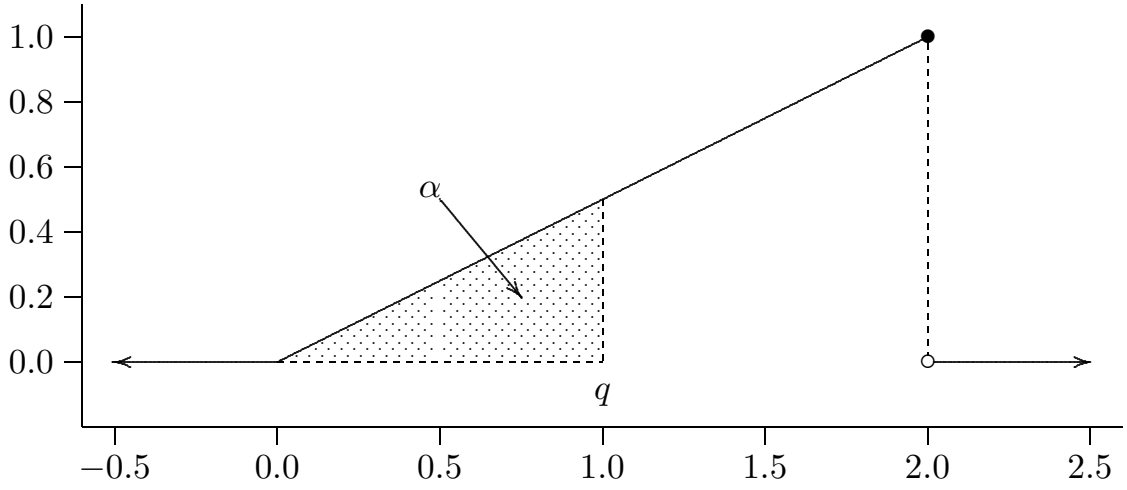

Figure 6.3: A quantile of another distribution.

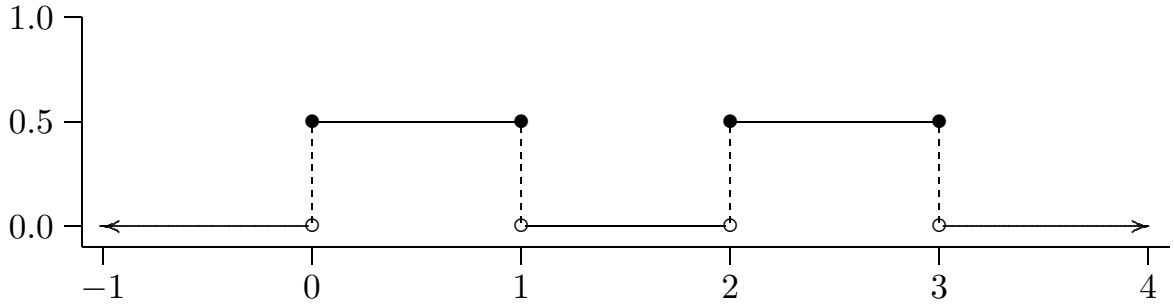

Figure 6.4: A distribution for which the $\alpha = 0.5$ quantile is not unique.

in Example 6.1 to deduce that $q = 2\alpha$. Similarly, if $\alpha \in (0.5, 1)$, then $q = 2 + 2(\alpha - 0.5) = 2\alpha + 1$. However, if $\alpha = 0.5$, then we encounter an ambiguity: the equalities $P(X < q) = 0.5$ and $P(X > q) = 0.5$ hold for *any* $q \in [1, 2]$. Accordingly, any $q \in [1, 2]$ is an $\alpha = 0.5$ quantile of X. Thus, quantiles are not always unique.

To avoid confusion when a quantile is not unique, it is nice to have a convention for selecting one of the possible quantile values. In the case that $\alpha = 0.5$, there is a universal convention:

Definition 6.4 *The midpoint of the interval of all values of the* $\alpha = 0.5$ *quantile is called the* population median.

In Example 6.4, the population median is $q = 1.5$.

Working with the quantiles of a continuous random variable X is straightforward because $P(X = q) = 0$ for any choice of q. This means that $P(X < q) + P(X > q) = 1$; hence, if $P(X < q) = \alpha$, then $P(X > q) = 1 - \alpha$. Furthermore, it is always possible to find a q for which $P(X < q) = \alpha$. This is not the case if X is discrete.

Example 6.5 *Let X be a discrete random variable that assumes values in the set $\{1, 2, 3\}$ with probabilities $p(1) = 0.4$, $p(2) = 0.4$, and $p(3) = 0.2$. What is the median of X?*

Imagine accumulating probability as we move from $-\infty$ to ∞. At what point do we find that we have acquired half of the total probability? The answer is that we pass from having 40% of the probability to having 80% of the probability as we occupy the point $q = 2$. It makes sense to declare this value to be the median of X.

Here is another argument that appeals to Definition 6.3. If $q < 2$, then $P(X > q) = 0.6 > 0.5$. Hence, it would seem that the population median should not be less than 2. Similarly, if $q > 2$, then $P(X < q) = 0.8 > 0.5$. Hence, it would seem that the population median should not be greater than 2. We conclude that the population median should equal 2. But notice that $P(X < 2) = 0.4 < 0.5$ and $P(X > 2) = 0.2 < 0.5$! We conclude that Definition 6.3 will not suffice for discrete random variables. However, we can generalize the reasoning that we have just employed as follows:

Definition 6.5 *Let X be a random variable and let $\alpha \in (0, 1)$. If $q = q(X; \alpha)$ is such that $P(X < q) \leq \alpha$ and $P(X > q) \leq 1 - \alpha$, then q is called an* α quantile *of X.*

The remainder of this section describes how quantiles are often used to measure centrality and dispersion. The following three quantiles will be of particular interest:

Definition 6.6 *Let X be a random variable. The first, second, and third* quartiles *of X, denoted $q_1(X)$, $q_2(X)$, and $q_3(X)$, are the $\alpha = 0.25$, $\alpha = 0.50$, and $\alpha = 0.75$ quantiles of X. The second quartile is also called the median of X.*

6.2.1 The Median of a Population

If X is a symmetric random variable with center of symmetry θ, then

$$P(X < \theta) = P(X > \theta) = \frac{1 - P(X = \theta)}{2} \leq \frac{1}{2}$$

and $q_2(X) = \theta$. Even if X is not symmetric, the median of X is an excellent way to define the "middle" of the population. Many statistical procedures use the median as a measure of centrality.

Example 6.6 One useful property of the median is that it is rather insensitive to the influence of extreme values that occur with small probability. For example, let X_k denote a discrete random variable that assumes values in $\{-1, 0, 1, 10^k\}$ for $k = 1, 2, 3, \ldots$. Suppose that X_k has the following pmf:

x	$p_k(x)$
-1	0.19
0	0.60
1	0.19
10^k	0.02

Most of the probability (98%) is concentrated on the values $\{-1, 0, 1\}$. This probability is centered at $x = 0$. A small amount of probability is concentrated at a large value, $x = 10, 100, 1000, \ldots$. If we want to treat these large values as aberrations (perhaps our experiment produces a physically meaningful value $x \in \{-1, 0, 1\}$ with probability 0.98, but our equipment malfunctions and produces a physically meaningless value $x = 10^k$ with probability 0.02), then we might prefer to declare that $x = 0$ is the central value of X. In fact, no matter how large we choose k, the median refuses to be distracted by the aberrant value: $P(X < 0) = 0.19$ and $P(X > 0) = 0.21$, so the median of X is $q_2(X) = 0$.

6.2.2 The Interquartile Range of a Population

Now we turn our attention from the problem of measuring centrality to the problem of measuring dispersion. Can we use quantiles to quantify how widely spread are the values of a random variable? A natural approach is to choose two values of α and compute the corresponding quantiles. The distance between these quantiles is a measure of dispersion.

To avoid comparing apples and oranges, let us agree on which two values of α we will choose. Statisticians have developed a preference for $\alpha = 0.25$ and $\alpha = 0.75$, in which case the corresponding quantiles are the first and third quartiles.

Definition 6.7 *Let X be a random variable with first and third quartiles q_1 and q_3. The* interquartile range *of X is the quantity*

$$iqr(X) = q_3 - q_1.$$

If X is a continuous random variable, then $P(q_1 < X < q_3) = 0.5$, so the interquartile range is the interval of values on which is concentrated the central 50% of the probability.

Like the median, the interquartile range is rather insensitive to the influence of extreme values that occur with small probability. In Example 6.6, the central 50% of the probability is concentrated on the single value $x = 0$. Hence, the interquartile range is $0 - 0 = 0$, regardless of where the aberrant 2% of the probability is located.

6.3 The Method of Least Squares

Let us return to the case of a symmetric random variable X, in which case the "middle" of the distribution is unambiguously the center of symmetry θ. Given this measure of centrality, how might we construct a measure of dispersion? One possibility is to measure how far a "typical" value of X lies from its central value, i.e., to compute $E|X - \theta|$. This possibility leads to several remarkably fertile approaches to describing both dispersion and centrality.

Given a designated central value c and another value x, we say that the *absolute deviation* of x from c is $|x - c|$ and that the *squared deviation* of x from c is $(x - c)^2$. The magnitude of a typical absolute deviation is $E|X - c|$ and the magnitude of a typical squared deviation is $E(X - c)^2$. A natural approach to measuring centrality is to choose a value of c that typically results in small deviations, i.e., to choose c either to minimize $E|X - c|$ or to minimize $E(X - c)^2$. The second possibility is a simple example of the *method of least squares*.

Measuring centrality by minimizing the magnitude of a typical absolute or squared deviation results in two familiar quantities:

Theorem 6.1 *Let X be a random variable with population median q_2 and population mean $\mu = EX$. Then*

1. *The value of c that minimizes $E|X - c|$ is $c = q_2$.*

2. *The value of c that minimizes $E(X - c)^2$ is $c = \mu$.*

It follows that medians are naturally associated with absolute deviations and that means are naturally associated with squared deviations. Having discussed the former in Section 6.2.1, we now turn to the latter.

6.3.1 The Mean of a Population

Imagine creating a physical model of a probability distribution by distributing weights along the length of a board. The location of the weights are the values of the random variable and the weights represent the probabilities of those values. After gluing the weights in place, we position the board atop a fulcrum. How must the fulcrum be positioned in order that the board be perfectly balanced? It turns out that one should position the fulcrum at the mean of the probability distribution. For this reason, the expected value of a random variable is sometimes called its *center of mass*.

Thus, like the population median, the population mean has an appealing interpretation that commends its use as a measure of centrality. If X is a symmetric random variable with center of symmetry θ, then $\mu = EX = \theta$ and $q_2 = q_2(X) = \theta$, so the population mean and the population median agree. In general, this is not the case. If X is not symmetric, then one should think carefully about whether one is interested in the population mean and the population median. Of course, computing both measures and examining the discrepancy between them may be highly instructive. In particular, if $EX \neq q_2(X)$, then X is not a symmetric random variable.

In Section 6.2.1 we noted that the median is rather insensitive to the influence of extreme values that occur with small probability. The mean lacks this property. In Example 6,

$$EX_k = -0.19 + 0.00 + 0.19 + 10^k \cdot 0.02 = 2 \cdot 10^{k-2},$$

which equals 0.2 if $k = 1$, 2 if $k = 2$, 20 if $k = 3$, 200 if $k = 4$, and so on. No matter how reluctantly, the population mean follows the aberrant value toward infinity as k increases.

6.3.2 The Standard Deviation of a Population

Suppose that X is a random variable with $EX = \mu$ and $\operatorname{Var} X = \sigma^2$. If we adopt the method of least squares, then we obtain $c = \mu$ as our measure

of centrality, in which case the magnitude of a typical squared deviation is $E(X-\mu)^2 = \sigma^2$, the population variance. The variance measures dispersion in squared units. For example, if X measures length in meters, then $\operatorname{Var} X$ is measured in meters squared. If, as in Section 6.2.2, we prefer to measure dispersion in the original units of measurement, then we must take the square root of the variance. Accordingly, we will emphasize the population standard deviation, σ, as a measure of dispersion.

Just as it is natural to use the median and the interquartile range together, so is it natural to use the mean and the standard deviation together. In the case of a symmetric random variable, the median and the mean agree. However, the interquartile range and the standard deviation measure dispersion in two fundamentally different ways. To gain insight into their relation to each other, suppose that $X \sim \text{Normal}(0,1)$, in which case the population standard deviation is $\sigma = 1$. We use `R` to compute $\text{iqr}(X)$:

```
> qnorm(.75)-qnorm(.25)
[1] 1.348980
```

We have derived a useful fact: *the interquartile range of a normal random variable is approximately* 1.35 *standard deviations.* If we encounter a random variable for which this is not the case, then that random variable is not normally distributed.

Like the mean, the standard deviation is sensitive to the influence of extreme values that occur with small probability. Consider Example 6. The variance of X_k is

$$\begin{aligned}\sigma_k^2 &= EX_k^2 - (EX_k)^2 = \left(0.19 + 0.00 + 0.19 + 100^k \cdot 0.02\right) - \left(2 \cdot 10^{k-2}\right)^2 \\ &= 0.38 + 2 \cdot 100^{k-1} - 4 \cdot 100^{k-2} = 0.38 + 196 \cdot 100^{k-2},\end{aligned}$$

so $\sigma_1 = \sqrt{2.34}$, $\sigma_2 = \sqrt{196.38}$, $\sigma_3 = \sqrt{19600.38}$, and so on. The population standard deviation tends toward infinity as the aberrant value tends toward infinity.

6.4 Exercises

1. Refer to the random variable X defined in Exercise 2 of Chapter 5. Compute the following two quantities: $q_2(X)$, the population median; and $\text{iqr}(X)$, the population interquartile range.

2. Consider the function $g : \Re \to \Re$ defined by

$$g(x) = \left\{ \begin{array}{cc} 0 & x < 0 \\ x & x \in [0,1] \\ 1 & x \in [1,2] \\ 3-x & x \in [2,3] \\ 0 & x > 3 \end{array} \right\}.$$

Let $f(x) = cg(x)$, where c is an undetermined constant.

   (a) For what value of c is f a probability density function?

   (b) Suppose that a continuous random variable X has probability density function f. Compute $P(1.5 < X < 2.5)$.

   (c) Compute EX.

   (d) Let F denote the cumulative distribution function of X. Compute $F(1)$.

   (e) Determine the 0.90 quantile of f.

3. Suppose that X is a continuous random variable with probability density function

$$f(x) = \left\{ \begin{array}{cc} 0 & x < 0 \\ x & x \in (0,1) \\ (3-x)/4 & x \in (1,3) \\ 0 & x > 3 \end{array} \right\}.$$

   (a) Compute $q_2(X)$, the population median.

   (b) Which is greater, $q_2(X)$ or EX? Explain your reasoning.

   (c) Compute $P(0.5 < X < 1.5)$.

   (d) Compute $\text{iqr}(X)$, the population interquartile range.

4. Lynn claims that Lulu is the cutest dog in the world. Slightly more circumspect, Michael allows that Lulu is "one in a million." Seizing the opportunity to revel in Lulu's charm, Lynn devises a procedure for measuring CCQ (canine cuteness quotient), which she calibrates so that CCQ $\sim$ Normal$(100, 400)$. Assuming that Michael is correct, what is Lulu's CCQ score?

5. Identify each of the following statements as *True* or *False*. Briefly explain each of your answers.

(a) For every symmetric random variable X, the median of X equals the average of the first and third quartiles of X.

(b) For every random variable X, the interquartile range of X is greater than the standard deviation of X.

(c) For every random variable X, the expected value of X lies between the first and third quartile of X.

(d) If the standard deviation of a random variable equals zero, then so does its interquartile range.

(e) If the median of a random variable equals its expected value, then the random variable is symmetric.

6. For each of the following random variables, discuss whether the median or the mean would be a more useful measure of centrality:

   (a) The annual income of U.S. households.

   (b) The lifetime of 75-watt light bulbs.

7. The `R` function `qbinom` returns quantiles of the binomial distribution. For example, quartiles of $X \sim \text{Binomial}(n = 3; p = 0.5)$ can be computed as follows:

```
> alpha <- c(.25,.5,.75)
> qbinom(alpha,size=3,prob=.5)
[1] 1 1 2
```

   Notice that X is a symmetric random variable with center of symmetry $\theta = 1.5$, but `qbinom` computes $q_2(X) = 1$. This reveals that `R` may produce unexpected results when it computes the quantiles of discrete random variables. By experimenting with various choices of n and p, try to discover a rule according to which `qbinom` computes quartiles of the binomial distribution.

Chapter 7

Data

Chapters 3–6 developed mathematical tools for studying populations. Experiments are performed for the purpose of obtaining information about a population that is imperfectly understood. Experiments produce data, the raw material from which statistical procedures draw inferences about the population under investigation.

The probability distribution of a random variable X is a mathematical abstraction of an experimental procedure for sampling from a population. When we perform the experiment, we observe one of the possible values of X. To distinguish an observed value of a random variable from the random variable itself, we designate random variables by uppercase letters and observed values by corresponding lowercase letters.

Example 7.1 A coin is tossed and `Heads` is observed. The mathematical abstraction of this experiment is $X \sim \text{Bernoulli}(p)$ and the observed value of X is $x = 1$.

We will be concerned with experiments that are replicated a fixed number of times. By replication, we mean that each repetition of the experiment is performed under identical conditions and that the repetitions are mutually independent. Mathematically, we write $X_1, \ldots, X_n \sim P$. Let x_i denote the observed value of X_i. The set of observed values, $\vec{x} = \{x_1, \ldots, x_n\}$, is called a sample.

This chapter introduces several useful techniques for extracting information from samples. This information will be used to draw inferences about populations (for example, to guess the value of the population mean) and to assess assumptions about populations (for example, to decide whether

or not the population can plausibly be modelled by a normal distribution). Drawing inferences about population attributes (especially means) is the primary subject of subsequent chapters, which will describe specific procedures for drawing specific types of inferences. However, deciding which procedure is appropriate often involves assessing the validity of certain statistical assumptions. The methods described in this chapter will be our primary tools for making such assessments.

To assess whether or not an assumption is plausible, one must be able to investigate what happens when the assumption holds. For example, if a scientist needs to decide whether or not it is plausible that her sample was drawn from a normal distribution, then she needs to be able to recognize normally distributed data. For this reason, the samples studied in this chapter were generated under carefully controlled conditions, by computer simulation. This allows us to investigate how samples drawn from specified distributions *should* behave, thereby providing a standard against which to compare experimental data for which the true distribution can never be known. Fortunately, `R` provides several convenient functions for simulating random sampling.

Example 7.2 Consider the experiment of tossing a fair die $n = 20$ times. We can simulate this experiment as follows:

```
> SampleSpace <- c(1,2,3,4,5,6)
> sample(x=SampleSpace,size=20,replace=T)
[1] 1 6 3 2 2 3 5 3 6 4 3 2 5 3 2 2 3 2 4 2
```

Example 7.3 Consider the experiment of drawing a sample of size $n = 5$ from Normal$(2, 3)$. We can simulate this experiment as follows:

```
> rnorm(5,mean=2,sd=sqrt(3))
[1] 1.3274812 0.5901923 2.5881013 1.2222812 3.4748139
```

7.1 The Plug-In Principle

We will employ a general methodology for relating samples to populations. In Chapters 3–6 we developed a formidable apparatus for studying populations (probability distributions). We would like to exploit this apparatus fully. Given a sample, we will pretend that the sample is a finite population (discrete probability distribution) and then we will use methods for studying

finite populations to learn about the sample. This approach is sometimes called the Plug-In Principle.

The Plug-In Principle employs a fundamental construction:

Definition 7.1 *Let $\vec{x} = (x_1, \ldots, x_n)$ be a sample. The* empirical probability distribution *associated with $\vec{x}$, denoted $\hat{P}_n$, is the discrete probability distribution defined by assigning probability $1/n$ to each $\{x_i\}$.*

Notice that, if a sample contains several copies of the same numerical value, then *each copy* is assigned probability $1/n$. This is illustrated in the following example.

Example 7.2 (continued) A fair die is rolled $n = 20$ times, resulting in the sample

$$\vec{x} = \{1, 6, 3, 2, 2, 3, 5, 3, 6, 4, 3, 2, 5, 3, 2, 2, 3, 2, 4, 2\}. \tag{7.1}$$

The empirical distribution $\hat{P}_{20}$ is the discrete distribution that assigns the following probabilities:

x_i	$\#\{x_i\}$	$\hat{P}_{20}(\{x_i\})$
1	1	0.05
2	7	0.35
3	6	0.30
4	2	0.10
5	2	0.10
6	2	0.10

Notice that, although the true probabilities are $P(\{x_i\}) = 1/6$, the empirical probabilities range from 0.05 to 0.35. The fact that $\hat{P}_{20}$ differs from P is an example of sampling variation. Statistical inference is concerned with determining what the empirical distribution (the sample) tells us about the true distribution (the population).

The empirical distribution, $\hat{P}_n$, is an intuitively appealing approximation of the actual probability distribution, P, from which the sample was drawn. Notice that the empirical probability of any event A is just

$$\hat{P}_n(A) = \#\{x_i \in A\} \cdot \frac{1}{n},$$

the observed frequency with which A occurs in the sample. Because the empirical distribution is an authentic probability distribution, all of the methods that we developed for studying (discrete) distributions are available for studying samples. For example,

Definition 7.2 *The empirical cdf, usually denoted $\hat{F}_n$, is the cdf associated with $\hat{P}_n$, i.e.*

$$\hat{F}_n(y) = \hat{P}_n(X \leq y) = \frac{\#\{x_i \leq y\}}{n}.$$

The empirical cdf of sample (7.1) is graphed in Figure 7.1.

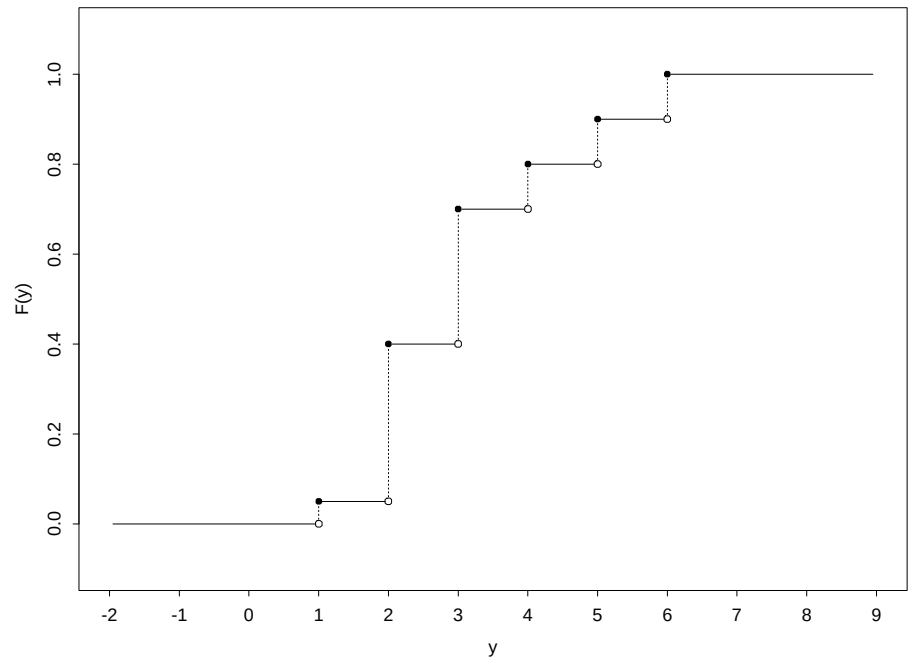

Figure 7.1: An empirical cdf.

In `R`, one can graph the empirical cdf of a sample `x` with the following command:

```
> plot.ecdf(x)
```

7.2 Plug-In Estimates of Mean and Variance

Population quantities defined by expected values are easily estimated by the plug-in principle. For example, suppose that $X_1, \ldots, X_n \sim P$ and that we observe a sample $\vec{x} = \{x_1, \ldots, x_n\}$. Let $\mu = EX_i$ denote the population mean. Then

Definition 7.3 *The plug-in estimate of μ, denoted $\hat{\mu}_n$, is the mean of the empirical distibution:*

$$\hat{\mu}_n = \sum_{i=1}^{n} x_i \cdot \frac{1}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}_n.$$

This quantity is called the sample mean.

Example 7.2 (continued) The population mean is

$$\mu = EX_i = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{1+2+3+4+5+6}{6} = 3.5.$$

The sample mean of sample (7.1) is

$$\begin{aligned} \hat{\mu}_{20} = \bar{x}_{20} &= 1 \cdot \frac{1}{20} + 6 \cdot \frac{1}{20} + \cdots + 4 \cdot \frac{1}{20} + 2 \cdot \frac{1}{20} \\ &= 1 \times 0.05 + 2 \times 0.35 + 3 \times 0.30 + 4 \times 0.10 + \\ &\quad 5 \times 0.10 + 6 \times 0.10 \\ &= 3.15. \end{aligned}$$

Notice that $\hat{\mu}_{20} \neq \mu$. This is another example of sampling variation.

The variance can be estimated in the same way. Let $\sigma^2 = \operatorname{Var} X_i$ denote the population variance; then

Definition 7.4 *The plug-in estimate of σ^2, denoted $\widehat{\sigma_n^2}$, is the variance of the empirical distribution:*

$$\widehat{\sigma_n^2} = \sum_{i=1}^{n} (x_i - \hat{\mu}_n)^2 \cdot \frac{1}{n} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right)^2.$$

Notice that we do not refer to $\widehat{\sigma_n^2}$ as the sample variance. As will be discussed in Section 9.2.2, most authors designate another, equally plausible estimate of the population variance as *the* sample variance.

Example 7.2 (continued) The population variance is

$$\sigma^2 = EX_i^2 - (EX_i)^2 = \frac{1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2}{6} - 3.5^2 = \frac{35}{12} \doteq 2.9167.$$

The plug-in estimate of the variance is

$$\begin{aligned} \widehat{\sigma^2_{20}} &= \Big(1^2 \times 0.05 + 2^2 \times 0.35 + 3^2 \times 0.30 + \\ &\quad 4^2 \times 0.10 + 5^2 \times 0.10 + 6^2 \times 0.10\Big) - 3.15^2 \\ &= 1.9275. \end{aligned}$$

Again, notice that $\widehat{\sigma^2_{20}} \neq \sigma^2$, yet another example of sampling variation.

There are many ways to compute the preceding plug-in estimates using R. Assuming that x contains the sample, here are two possibilities:

```
> n <- length(x)
> plug.mean <- sum(x)/n
> plug.var <- sum(x^2)/n - plug.mean^2

> plug.mean <- mean(x)
> plug.var <- mean(x^2) - plug.mean^2
```

7.3 Plug-In Estimates of Quantiles

Population quantities defined by quantiles can also be estimated by the plug-in principle. Again, suppose that $X_1, \ldots, X_n \sim P$ and that we observe a sample $\vec{x} = \{x_1, \ldots, x_n\}$. Then

Definition 7.5 *The plug-in estimate of a population quantile is the corresponding quantile of the empirical distribution. In particular, the* sample median *is the median of the empirical distribution. The* sample interquartile range *is the interquartile range of the empirical distribution.*

Example 7.4 Consider the experiment of drawing a sample of size $n = 20$ from Uniform$(1, 5)$. This probability distribution has a population median of 3 and a population interquartile range of $4 - 2 = 2$. I simulated this experiment (and listed the sample in increasing order) with the following R command:

```
> x <- sort(runif(20,min=1,max=5))
```

This resulted in the following sample:

1.124600	1.161286	1.445538	1.828181	1.853359
1.934939	1.943951	2.107977	2.372500	2.448152
2.708874	3.297806	3.418913	3.437485	3.474940
3.698471	3.740666	4.039637	4.073617	4.195613

The sample median is

$$\frac{2.448152 + 2.708874}{2} = 2.578513,$$

which also can be computed with the following R command:

```
> median(x)
[1] 2.578513
```

Notice that the sample median does not exactly equal the population median. This is another example of sampling variation.

To compute the sample interquartile range, we require the first and third sample quartiles, i.e., the $\alpha = 0.25$ and $\alpha = 0.75$ sample quantiles. We must now confront the fact that Definition 6.5 may not specify unique quantile values. For the empirical distribution of the sample above, any number in $[1.853359, 1.934939]$ is a sample first quartile and any number in $[3.474940, 3.698471]$ is a sample third quartile.

The statistical community has not agreed on a convention for resolving the ambiguity in the definition of quartiles. One natural and popular possibility is to use the central value in each interval of possible quartiles. If we adopt that convention here, then the sample interquartile range is

$$\frac{3.474940 + 3.698471}{2} - \frac{1.853359 + 1.934939}{2} = 1.692556.$$

R adopts a slightly different convention, illustrated below. The following command computes the 0.25 and 0.75 quantiles:

```
> quantile(x,probs=c(.25,.75))
      25%       75%
1.914544 3.530823
```

The following command computes several useful sample quantities:

```
> summary(x)
     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
1.124600 1.914544 2.578513 2.715325 3.530823 4.195613
```

If we use the R definition of quantile, then the sample interquartile range is $3.530823 - 1.914544 = 1.616279$. Rather than typing the quartiles into R, we can compute the sample interquartile range as follows:

```
> q <- as.vector(quantile(x,probs=c(.25,.75)))
> q[2]-q[1]
[1] 1.616279
```

This is sufficiently complicated that we might prefer to create a function that computes the interquartile range of a sample:

```
> iqr <- function(x) {
+ q <- as.vector(quantile(x,probs=c(.25,.75)))
+ return(q[2]-q[1])
+ }
> iqr(x)
[1] 1.616279
```

Notice that the sample quantities do not exactly equal the population quantities that they estimate, regardless of which convention we adopt for defining quartiles. This is another example of sampling variation.

Used judiciously, sample quantiles can be extremely useful when trying to discern various features of the population from which the sample was drawn. The remainder of this section describes two graphical techniques for assimilating and displaying sample quantile information.

7.3.1 Box Plots

Information about sample quartiles is often displayed visually, in the form of a *box plot*. A box plot of a sample consists of a rectangle that extends from the first to the third sample quartile, thereby drawing attention to the central 50% of the data. Thus, the length of the rectangle equals the sample interquartile range. The location of the sample median is also identified, and its location within the rectangle often provides insight into whether or not the population from which the sample was drawn is symmetric. Whiskers extend from the ends of the rectangle, either to the extreme values of the data or to 1.5 times the sample interquartile range, whichever is less. Values that lie beyond the whiskers are called *outliers* and are individually identified.

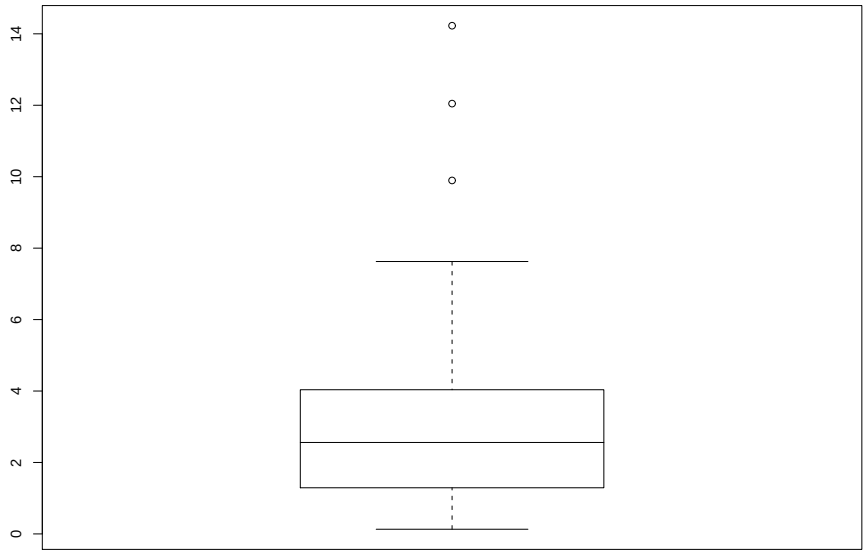

Figure 7.2: A box plot of a sample from $\chi^2(3)$.

Example 7.5 The pdf of the asymmetric distribution $\chi^2(3)$ was graphed in Figure 5.8. The following `R` commands draw a random sample of $n = 100$ observed values from this population, then construct a box plot of the sample:

```
> x <- rchisq(100,df=3)
> boxplot(x)
```

An example of a box plot produced by these commands is displayed in Figure 7.2. In this box plot, the numerical values in the sample are represented by the *vertical* axis.

The third quartile of the box plot in Figure 7.2 is farther above the median than the first quartile is below it. The short lower whisker extends from the first quartile to the minimal value in the sample, whereas the long upper whisker extends 1.5 interquartile ranges beyond the third quartile. Furthermore, there are 3 outliers beyond the upper whisker. Once we learn to discern these key features of the box plot, we can easily recognize that the population from which the sample was drawn is not symmetric.

The frequency of outliers in a sample often provides useful diagnostic information. Recall that, in Section 6.3, we computed that the interquartile range of a normal distribution is 1.34898 standard deviations. A value is an outlier if it lies more than

$$z = \frac{1.34898}{2} + 1.5 \cdot 1.34898 = 2.69796$$

standard deviations from the mean. Hence, the probability that an observation drawn from a normal distribution is an outlier is

```
> 2*pnorm(-2.69796)
[1] 0.006976582
```

and we would expect a sample drawn from a normal distribution to contain approximately 7 outliers per 1000 observations. A sample that contains a dramatically different proportion of outliers, as in Example 7.5, is not likely to have been drawn from a normal distribution.

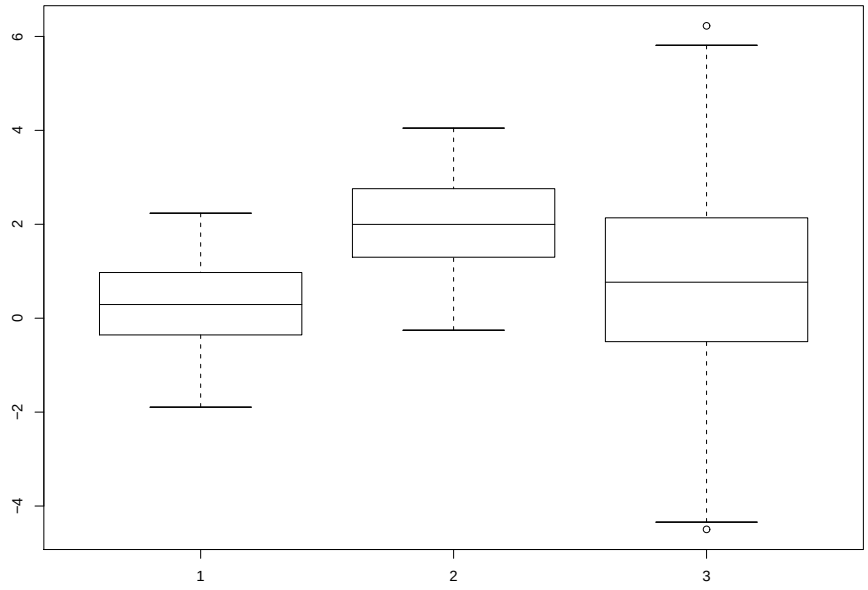

Figure 7.3: Box plots of samples from three normal distributions.

Box plots are especially useful for comparing several populations.

Example 7.6 We drew samples of 100 observations from three normal populations: Normal$(0, 1)$, Normal$(2, 1)$, and Normal$(1, 4)$. To attempt to discern in the samples the various differences in population mean and standard deviation, we examined side-by-side box plots. This was accomplished by the following R commands:

```
> z1 <- rnorm(100)
> z2 <- rnorm(100,mean=2,sd=1)
> z3 <- rnorm(100,mean=1,sd=2)
> boxplot(z1,z2,z3)
```

An example of the output of these commands is displayed in Figure 7.3.

7.3.2 Normal Probability Plots

Another powerful graphical technique that relies on quantiles are quantile-quantile (QQ) plots, which plot the quantiles of one distribution against the quantiles of another. QQ plots are used to compare the shapes of two distributions, most commonly by plotting the observed quantiles of an empirical distribution against the corresponding quantiles of a theoretical normal distribution. In this case, a QQ plot is often called a normal probability plot. If the shape of the empirical distribution resembles a normal distribution, then the points in a normal probability plot should tend to fall on a straight line. If they do not, then we should be skeptical that the sample was drawn from a normal distribution. Extracting useful information from normal probability plots requires some practice, but the patient data analyst will be richly rewarded.

Example 7.4 (continued) A normal probability plot of the sample generated in Example 7.5 against a theoretical normal distribution is displayed in Figure 7.4. This plot was created using the following R command:

```
> qqnorm(x)
```

Notice the systematic and asymmetric bending away from linearity in this plot. In particular, the smaller quantiles are much closer to the central values than should be the case for a normal distribution. This suggests that this sample was drawn from a nonnormal distribution that is skewed to the right. Of course, we know that this sample was drawn from $\chi^2(3)$, which is in fact skewed to the right.

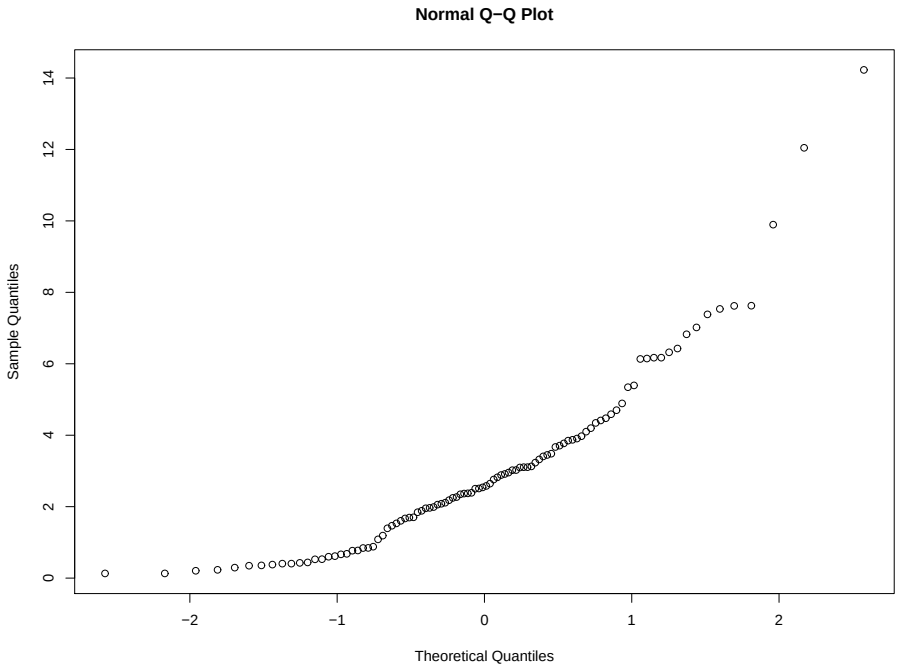

Figure 7.4: A normal probability plot of a sample from $\chi^2(3)$.

When using normal probability plots, one must guard against overinterpreting slight departures from linearity. Remember: *some departures from linearity will result from sampling variation.* Consequently, before drawing definitive conclusions, the wise data analyst will generate several random samples from the theoretical distribution of interest in order to learn how much sampling variation is to be expected. Before dismissing the possibility that the sample in Example 7.5 was drawn from a normal distribution, one should generate several normal samples of the same size for comparison. The normal probability plots of four such samples are displayed in Figure 7.5. In none of these plots did the points fall exactly on a straight line. However, upon comparing the normal probability plot in Figure 7.4 to the normal probability plots in Figure 7.5, it is abundantly clear that the sample in Example 7.5 was not drawn from a normal distribution.

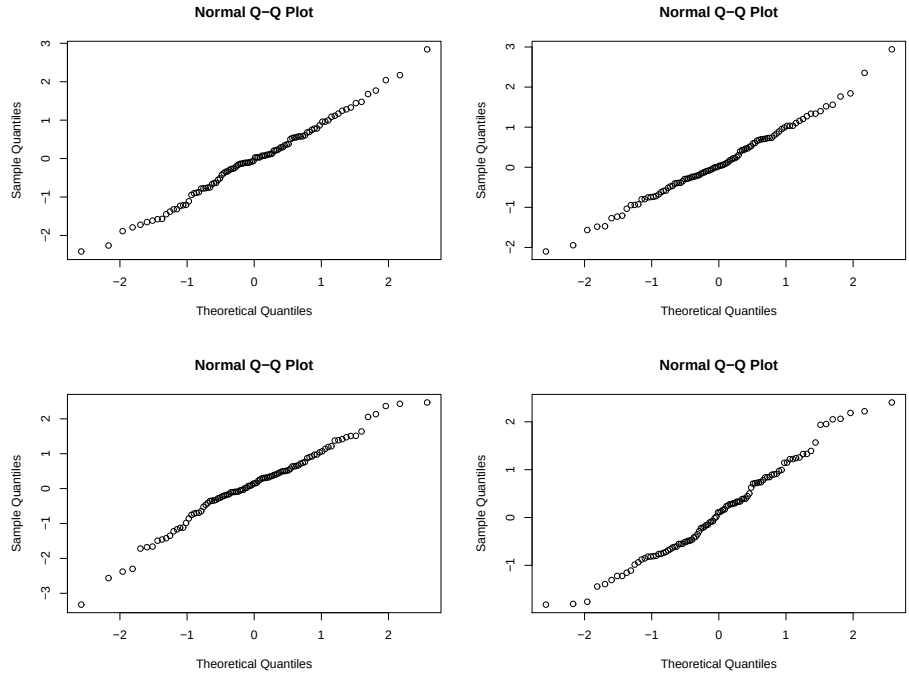

Figure 7.5: Normal probability plots of four samples from Normal$(0, 1)$.

7.4 Density Estimates

Suppose that $\vec{x} = \{x_1, \ldots, x_n\}$ is a sample drawn from an unknown pdf f. Box plots and normal probability plots are extremely useful graphical techniques for discerning in $\vec{x}$ certain important attributes of f, e.g., centrality, dispersion, asymmetry, nonnormality. To discern more subtle features of f, we now ask if it is possible to reconstruct from $\vec{x}$ a pdf $\hat{f}_n$ that approximates f. This is a difficult problem, one that remains a vibrant topic of research and about which little is said in introductory courses. However, using the concept of the empirical distribution, one can easily motivate one of the most popular techniques for *nonparametric probability density estimation*.

The logic of the empirical distribution is this: by assigning probability $1/n$ to each x_i, one accumulates more probability in regions that produced more observed values. However, because the entire amount $1/n$ is placed exactly on the value x_i, the resulting empirical distribution is necessarily discrete. If the population from which the sample was drawn is discrete, then the empirical distribution estimates the probability mass function. However,

if the population from which the sample was drawn is continuous, then *all* possible values occur with zero probability. In this case, there is nothing special about the precise values that were observed—what is important are the regions in which they occurred.

Instead of placing all of the probability $1/n$ assigned to x_i exactly on the value x_i, we now imagine distributing it in a neighborhood of x_i according to some probability density function. This construction will also result in more probability accumulating in regions that produced more values, but it will produce a pdf instead of a pmf. Here is a general description of this approach, usually called *kernel density estimation*:

1. Choose a probability density function K, the *kernel*. Typically, K is a symmetric pdf centered at the origin. Common choices of K include the Normal$(0, 1)$ and Uniform$[-0.5, 0.5]$ pdfs.

2. At each x_i, center a rescaled copy of the kernel. This pdf,
$$\frac{1}{h} K\left(\frac{x - x_i}{h}\right), \tag{7.2}$$
will control the distribution of the $1/n$ probability assigned to x_i. The parameter h is variously called the *smoothing parameter*, the *window width*, or the *bandwidth*.

3. The difficult decision in constructing a kernel density estimate is the choice of h. The technical details of this issue are beyond the scope of this book, but the underlying principles are quite simple:

   - Small values of h mean that the standard deviation of (7.2) will be small, so that the $1/n$ probability assigned to x_i will be distributed close to x_i. This is appropriate when n is large and the x_i are tightly packed.
   - Large values of h mean that the standard deviation of (7.2) will be large, so that the $1/n$ probability assigned to x_i will be widely distributed in the general vicinity of x_i. This is appropriate when n is small and the x_i are sparse.

4. After choosing K and h, the kernel density estimate of f is
$$\hat{f}_n(x) = \sum_{i=1}^{n} \frac{1}{n} \frac{1}{h} K\left(\frac{x - x_i}{h}\right) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right).$$
Such estimates are easily computed and graphed using the `R` functions `density` and `plot`.

Example 7.7 *Consider the probability density function f displayed in Figure 7.6. The most striking feature of f is that it is* bimodal. *Can we detect this feature using a sample drawn from f?*

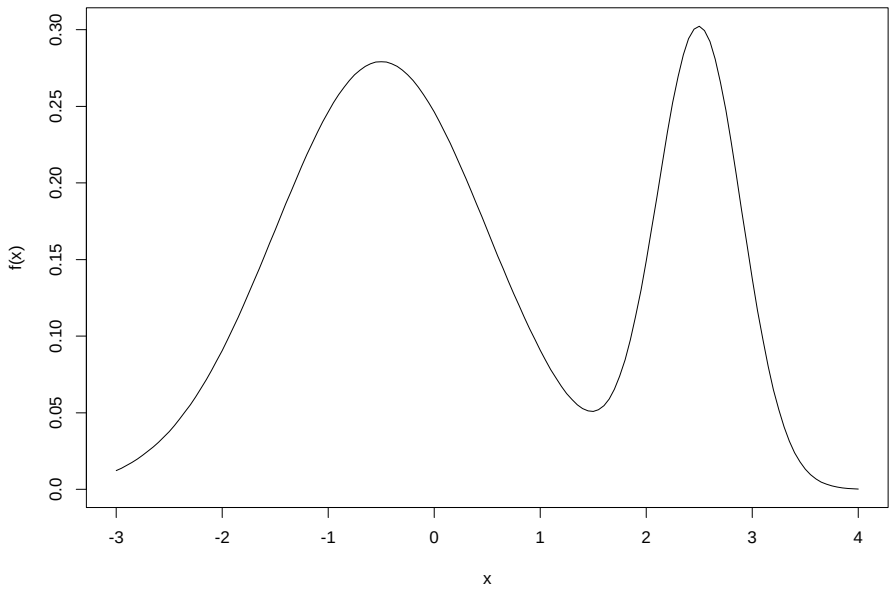

Figure 7.6: A bimodal probability density function.

We drew a sample of size $n = 100$ from f. A box plot and a normal probability plot of this sample are displayed in Figure 7.7. It is difficult to discern anything unusual from the box plot. The normal probability plot contains all of the information in the sample, but it is encoded in such a way that the feature of interest is not easily extracted. In contrast, the kernel density estimate displayed in Figure 7.8 clearly reveals that the sample was drawn from a bimodal population. After storing the sample in the vector `x`, this estimate was computed and plotted using the following `R` command:

```
> plot(density(x))
```

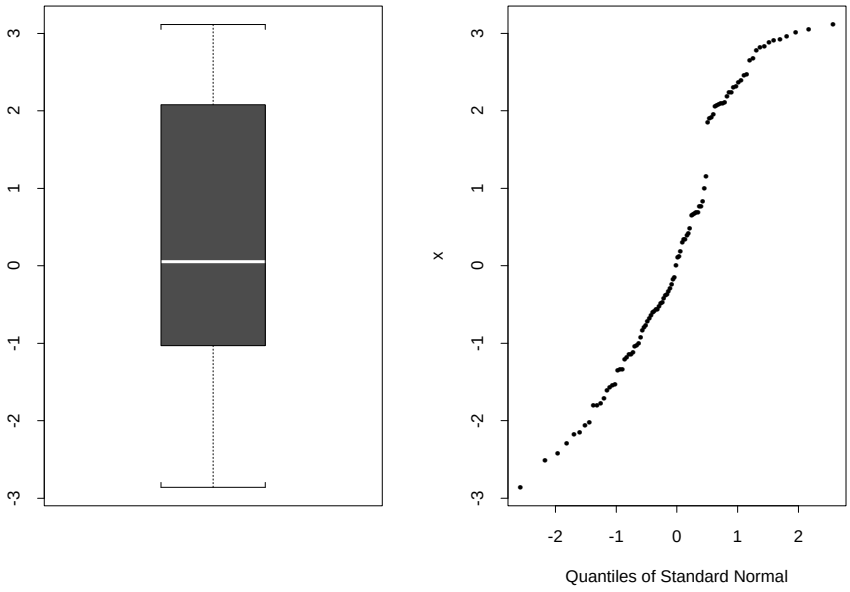

Figure 7.7: A box plot and a normal probability plot for Example 7.7.

7.5 Exercises

1. The following independent samples were drawn from four populations:

Sample 1	Sample 2	Sample 3	Sample 4
5.098	4.627	3.021	7.390
2.739	5.061	6.173	5.666
2.146	2.787	7.602	6.616
5.006	4.181	6.250	7.868
4.016	3.617	1.875	2.428
9.026	3.605	6.996	6.740
4.965	6.036	4.850	7.605
5.016	4.745	6.661	10.868
6.195	2.340	6.360	1.739
4.523	6.934	7.052	1.996

   (a) Use the `boxplot` function to create side-by-side box plots of these samples. Does it appear that these samples were all drawn from the same population? Why or why not?

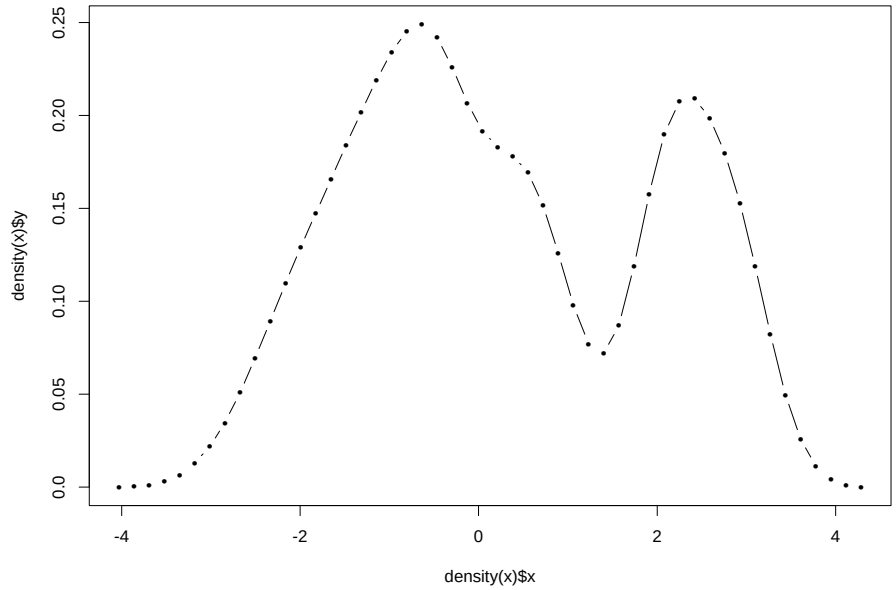

Figure 7.8: A kernel density estimate for Example 7.7.

(b) Use the `rnorm` function to draw four independent samples, each of size $n = 10$, from one normal distribution. Examine box plots of these samples. Is it possible that Samples 1–4 were all drawn from the same normal distribution?

2. The following sample, $\vec{x}$, was collected and sorted:

0.246	0.327	0.423	0.425	0.434
0.530	0.583	0.613	0.641	1.054
1.098	1.158	1.163	1.439	1.464
2.063	2.105	2.106	4.363	7.517

(a) Graph the empirical cdf of $\vec{x}$.

(b) Calculate the plug-in estimates of the mean, the variance, the median, and the interquartile range.

(c) Take the square root of the plug-in estimate of the variance and compare it to the plug-in estimate of the interquartile range. Do you think that $\vec{x}$ was drawn from a normal distribution? Why or why not?

(d) Use the `qqnorm` function to create a normal probability plot. Do you think that $\vec{x}$ was drawn from a normal distribution? Why or why not?

(e) Now consider the transformed sample $\vec{y}$ produced by replacing each x_i with its natural logarithm. If $\vec{x}$ is stored in the vector `x`, then $\vec{y}$ can be computed by the following `R` command:

```
> y <- log(x)
```

Do you think that $\vec{y}$ was drawn from a normal distribution? Why or why not?

3. In January 2002, twelve students enrolled in Math 351 (Applied Statistics) at the College of William & Mary reported the following results for the experiment described in Exercise 1.4.2. (Two students reported more than one measurement, but only one measurement per student is reported here.)

$$\begin{array}{cccccc} 143\frac{3}{16} & 144\frac{4}{16} & 140\frac{14}{16} & 144\frac{7}{16} & 143\frac{12}{16} & 153\frac{13}{16} \\ 119\frac{10}{16} & 143\frac{1}{16} & 143\frac{14}{16} & 144\frac{3}{16} & 144\frac{7}{16} & 148\frac{3}{16} \end{array}$$

(a) Do these measurements appear to be a sample from a normal distribution? Why or why not?

(b) Suggest possible explanations for the surprising amount of variation in these measurements.

(c) Use these measurements to estimate the true length of the table. Justify your estimation procedure.

4. Forty-one students taking Math 351 (Applied Statistics) at the College of William & Mary were administered a test. The following test scores were observed and sorted:

90	90	89	88	85	85	84	82	82	82	
81	81	81	80	79	79	78	76	75	74	
72	71	70	66	65	63	62	62	61	59	
58	58	57	56	56	53	48	44	40	35	33

(a) Do these numbers appear to be a random sample from a normal distribution?

(b) Does this list of numbers have any interesting anomalies?

5. Do the numbers in Table 1.1 (Michelson's measurements of the speed of light) appear to be a random sample from a normal distribution?

6. Experiment with using `R` to generate simulated random samples of various sizes. Use the `summary` function to compute the quartiles of these samples. Try to discern the convention that this function uses to define sample quartiles.

Chapter 8

Lots of Data

Throughout Chapter 7 we emphasized that, because of sampling variation, the plug-in estimate of a population quantity rarely equals the actual value of the population quantity. The present chapter explores this phenomenon in greater depth.

Suppose that $X_1, \ldots, X_n \sim P$ and that an experimental scientist wants to estimate the population mean, $\mu = EX_i$. To do so, she observes values $x_1, \ldots, x_n$ of $X_1, \ldots, X_n$, then computes

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

the plug-in estimate of μ. Mathematically, this is equivalent to first defining a new random variable,

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

then observing the value $\bar{x}_n$ of $\bar{X}_n$. The random variable $\bar{X}_n$ is the average of the random variables $X_1, \ldots, X_n$. Both the random variable $\bar{X}_n$ and the observed value $\bar{x}_n$ are called the *sample mean*. This is potentially confusing, but the convention of using uppercase letters for random variables and lowercase letters for observed values allows us to be clear about which concept we have in mind when we use the phrase "sample mean." In this chapter, we study the behavior of $\bar{X}_n$.

We begin with an example. Suppose that, unbeknownst to the scientist, P is the asymmetric probability distribution $\chi^2(3)$, with pdf depicted in Figure 5.8. Because of Corollary 5.1, it follows that $\mu = 3$. Hence, we can

assess the quality of the scientist's estimates of μ by comparing the estimates to the correct value, $\mu = 3$. We will use simulation to explore what might occur in this situation.

First, consider drawing a small sample of $n = 5$ observations. Here is what happened when I performed that experiment three times:

```
> x <- rchisq(5,df=3)
> mean(x)
[1] 3.650077

> x <- rchisq(5,df=3)
> mean(x)
[1] 2.963841

> x <- rchisq(5,df=3)
> mean(x)
[1] 2.063129
```

Due to sampling variation, the first estimate is too high, the second estimate is just about right, and the third estimate is too low. These results suggest that small samples may be unreliable. Of course, if we admit the possibility that small samples are unreliable, then it might be wise to perform the simulation more than three times! So, I performed the same simulation 1000 times, each time observing values of $X_1, \ldots, X_5 \sim \chi^2(3)$ and then computing $\bar{x}_5$, the observed value of $\bar{X}_5$. To display the results, I applied the method described in Section 7.4 to the 1000 observed values of $\bar{X}_5$. This produced a kernel density estimate, displayed in Figure 8.1, of the pdf of $\bar{X}_5$. Notice the considerable variation in the observed values of $\bar{X}_5$.

Next, consider drawing a moderate sample of $n = 20$ observations. I did this 1000 times, each time observing values of $X_1, \ldots, X_{20} \sim \chi^2(3)$ and then computing $\bar{x}_{20}$, the observed value of $\bar{X}_{20}$. From these 1000 observed values of $\bar{X}_{20}$, I constructed a kernel density estimate of the pdf of $\bar{X}_{20}$. This estimated pdf is displayed in Figure 8.2. Notice that the observed values of $\bar{X}_{20}$ tend to be more tightly clustered around $\mu = 3$ than do the observed values of $\bar{X}_5$, suggesting that moderate samples are more reliable than small samples.

Finally, consider drawing a large sample of $n = 80$ observations. I did this 1000 times, each time observing values of $X_1, \ldots, X_{80} \sim \chi^2(3)$ and then computing $\bar{x}_{80}$, the observed value of $\bar{X}_{80}$. From these 1000 observed values of $\bar{X}_{80}$, I constructed a kernel density estimate of the pdf of $\bar{X}_{80}$. This

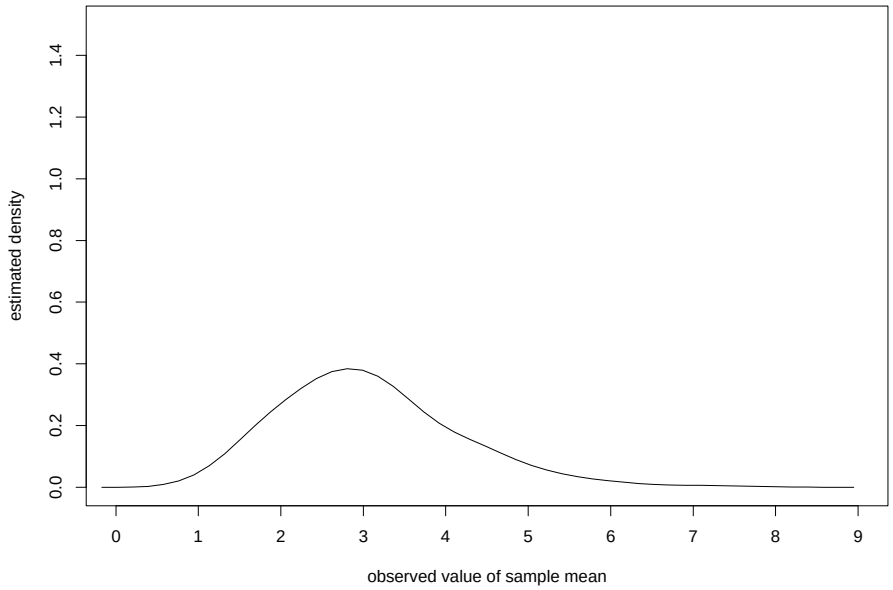

Figure 8.1: Kernel density estimate constructed from 1000 observed values of $\bar{X}_n$ for $n = 5$. $X_1, \ldots, X_n \sim \chi^2(3)$ and $\mu = EX_i = 3$.

estimated pdf is displayed in Figure 8.3. Notice that the observed values of $\bar{X}_{80}$ tend to be more tightly clustered around $\mu = 3$ than do the observed values of $\bar{X}_{20}$, suggesting that largee samples are more reliable than moderate samples.

The sections in this chapter generalize the preceding observations. We consider any experiment that can be performed, independently and identically, as many times as we please. We describe this situation by supposing the existence of a sequence of independent and identically distributed random variables, $X_1, X_2, \ldots$, and we assume that these random variables have a finite mean $\mu = EX_i$ and a finite variance $\sigma^2 = \operatorname{Var} X_i$. Under these assumptions, we study the behavior of the sample mean, $\bar{X}_n$, as n increases.

8.1 Averaging Decreases Variation

By definition, $EX_i = \mu$. Thus, the population mean is the average value assumed by the random variable X_i. This statement is also true of the

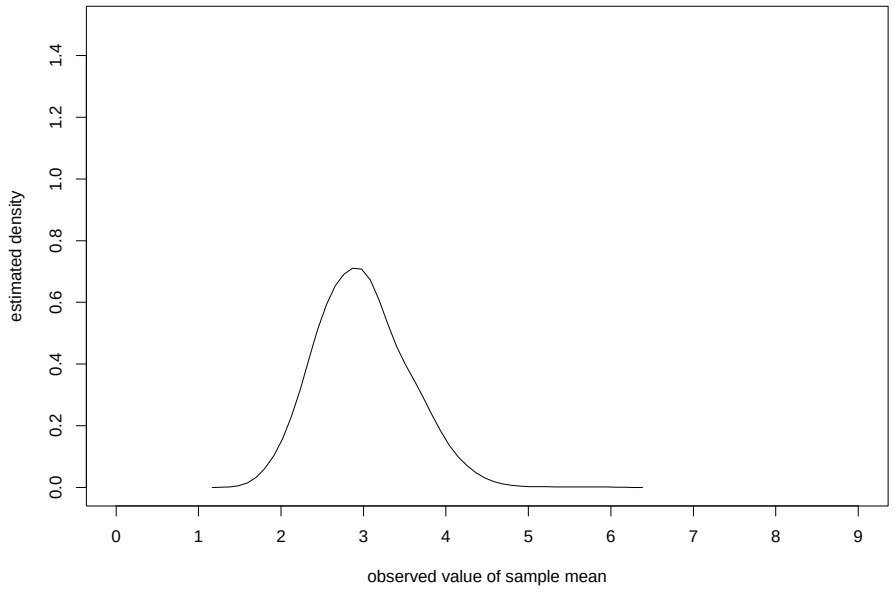

Figure 8.2: Kernel density estimate constructed from 1000 observed values of $\bar{X}_n$ for $n = 20$. $X_1, \ldots, X_n \sim \chi^2(3)$ and $\mu = EX_i = 3$.

sample mean:

$$E\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} EX_i = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu;$$

however, there is a crucial distinction between X_i and $\bar{X}_n$.

The tendency of a random variable to assume a value that is close to its expected value is quantified by computing its variance. By definition, $\operatorname{Var} X_i = \sigma^2$, but

$$\operatorname{Var} \bar{X}_n = \operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \operatorname{Var} X_i = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}.$$

Hence, the sample mean has less variability than any of the individual random variables that are being averaged. *Averaging decreases variation.* Furthermore, as $n \to \infty$, $\operatorname{Var} \bar{X}_n \to 0$. Thus, by repeating our experiment enough times, we can make the variation in the sample mean as small as we please.

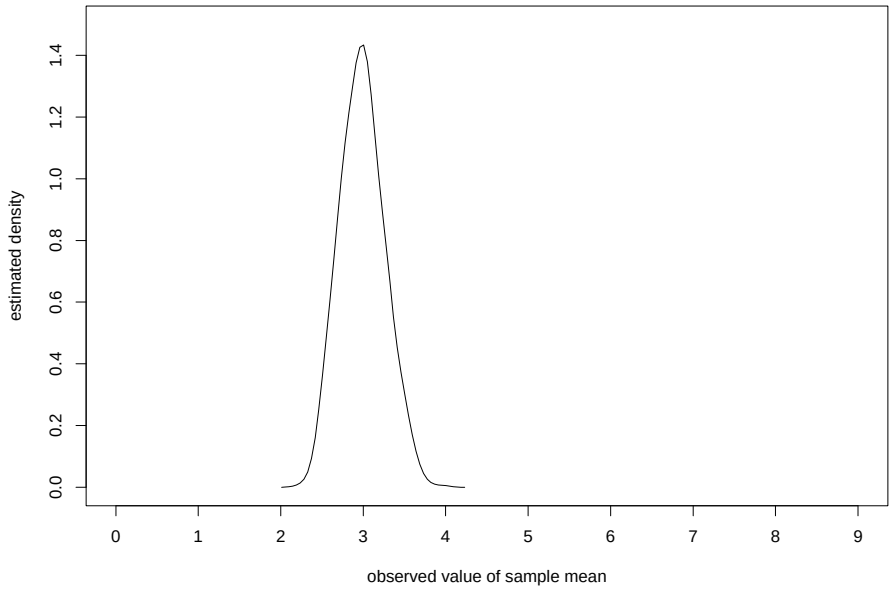

Figure 8.3: Kernel density estimate constructed from 1000 observed values of $\bar{X}_n$ for $n = 80$. $X_1, \ldots, X_n \sim \chi^2(3)$ and $\mu = EX_i = 3$.

The preceding remarks suggest that, if the population mean is unknown, then we can draw inferences about it by observing the behavior of the sample mean. This fundamental insight is the basis for a considerable portion of this book. The remainder of this chapter refines the relation between the population mean and the behavior of the sample mean.

8.2 The Weak Law of Large Numbers

Recall Definition 2.12 from Section 2.4: a sequence of real numbers $\{y_n\}$ converges to a limit $c \in \Re$ if and only if, for every $\epsilon > 0$, there exists a natural number N such that $y_n \in (c - \epsilon, c + \epsilon)$ for each $n \geq N$. Our first task is to generalize from convergence of a sequence of real numbers to convergence of a sequence of random variables.

If we replace $\{y_n\}$, a sequence of real numbers, with $\{Y_n\}$, a sequence of random variables, then the event that $Y_n \in (c-\epsilon, c+\epsilon)$ is uncertain. Rather than demand that this event *must* occur for n sufficiently large, we ask only

that the probability of this event tend to unity as n tends to infinity. This results in

Definition 8.1 *A sequence of random variables $\{Y_n\}$ converges in probability to a constant c, written $Y_n \xrightarrow{P} c$, if and only if, for every $\epsilon > 0$,*

$$\lim_{n\to\infty} P\left(Y_n \in (c-\epsilon, c+\epsilon)\right) = 1.$$

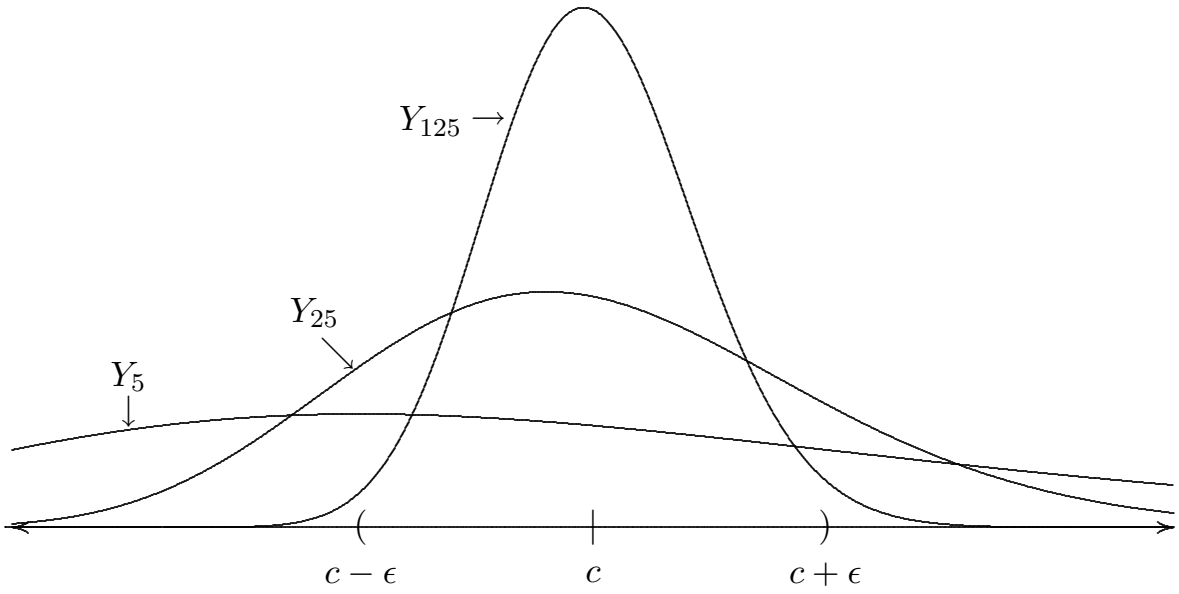

Figure 8.4: An example of convergence in probability.

Convergence in probability is depicted in Figure 8.4 using the pdfs f_n of continuous random variables Y_n. (One could also use the pmfs of discrete random variables.) We see that

$$p_n = P\left(Y_n \in (c-\epsilon, c+\epsilon)\right) = \int_{c-\epsilon}^{c+\epsilon} f_n(x)\,dx$$

is tending to unity as n increases. Notice, however, that each $p_n < 1$.

The concept of convergence in probability allows us to state an important result.

Theorem 8.1 *(Weak Law of Large Numbers) Let $X_1, X_2, \ldots$ be any sequence of independent and identically distributed random variables having*

finite mean μ and finite variance σ^2. Then

$$\bar{X}_n \xrightarrow{P} \mu.$$

This result is of considerable consequence. It states that, as we average more and more X_i, the average values that we observe tend to be distributed closer and closer to the theoretical average of the X_i. This property of the sample mean strengthens our contention that the behavior of $\bar{X}_n$ provides more and more information about the value of μ as n increases.

The Weak Law of Large Numbers (WLLN) has an important special case.

Corollary 8.1 *(Law of Averages) Let A be any event and consider a sequence of independent and identical experiments in which we observe whether or not A occurs. Let $p = P(A)$ and define independent and identically distributed random variables by*

$$X_i = \left\{ \begin{array}{ll} 1 & A \text{ occurs} \\ 0 & A^c \text{ occurs} \end{array} \right\}.$$

Then $X_i \sim$ Bernoulli(p), $\bar{X}_n$ is the observed frequency with which A occurs in n trials, and $\mu = EX_i = p = P(A)$ is the theoretical probability of A. The WLLN states that the former tends to the latter as the number of trials increases.

The Law of Averages formalizes our common experience that "things tend to average out in the long run." For example, we might be surprised if we tossed a fair coin $n = 10$ times and observed $\bar{X}_{10} = 0.9$; however, if we knew that the coin was indeed fair ($p = 0.5$), then we would remain confident that, as n increased, $\bar{X}_n$ would eventually tend to 0.5.

Notice that the *conclusion* of the Law of Averages is the frequentist *interpretation* of probability. Instead of defining probability via the notion of long-run frequency, we defined probability via the Kolmogorov axioms. Although our approach does not require us to interpret probabilities in any one way, the Law of Averages states that probability necessarily behaves in the manner specified by frequentists.

Finally, recall from Section 7.1 that the empirical probability of an event A is the observed frequency with which A occurs in the sample:

$$\hat{P}_n(A) = \#\{x_i \in A\} \cdot \frac{1}{n},$$

By the Law of Averages, this quantity tends to the true probability of A as the size of the sample increases. Thus, the theory of probability provides a mathematical justification for approximating P with $\hat{P}_n$ when P is unknown.

8.3 The Central Limit Theorem

The Weak Law of Large Numbers states a precise sense in which the distribution of values of the sample mean collapses to the population mean as the size of the sample increases. As interesting and useful as this fact is, it leaves several obvious questions unanswered:

1. How rapidly does the sample mean tend toward the population mean?
2. How does the shape of the sample mean's distribution change as the sample mean tends toward the population mean?

To answer these questions, we convert the random variables in which we are interested to standard units.

We have supposed the existence of a sequence of independent and identically distributed random variables, $X_1, X_2, \ldots$, with finite mean $\mu = EX_i$ and finite variance $\sigma^2 = \text{Var}\, X_i$. We are interested in the sum and/or the average of $X_1, \ldots, X_n$. It will be helpful to identify several crucial pieces of information for each random variable of interest:

random variable	expected value	standard deviation	standard units
X_i	μ	σ	$(X_i - \mu)/\sigma$
$\sum_{i=1}^n X_i$	$n\mu$	$\sqrt{n}\,\sigma$	$(\sum_{i=1}^n X_i - n\mu) \div (\sqrt{n}\,\sigma)$
$\bar{X}_n$	μ	$\sigma/\sqrt{n}$	$(\bar{X}_n - \mu) \div (\sigma/\sqrt{n})$

First we consider X_i. Notice that converting to standard units does *not* change the *shape* of the distribution of X_i. For example, if $X_i \sim$ Bernoulli(0.5), then the distribution of X_i assigns equal probability to each of two values, $x = 0$ and $x = 1$. If we convert to standard units, then the distribution of

$$Z_1 = \frac{X_i - \mu}{\sigma} = \frac{X_i - 0.5}{0.5}$$

also assigns equal probability to each of two values, $z_1 = -1$ and $z_1 = 1$. In particular, notice that converting X_i to standard units does *not* automatically result in a normally distributed random variable.

Next we consider the sum and the average of $X_1, \ldots, X_n$. Notice that, after converting to standard units, these quantities are identical:

$$Z_n = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma} = \frac{(1/n)}{(1/n)} \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\,\sigma} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}.$$

It is this new random variable on which we shall focus our attention.

We begin by observing that

$$\text{Var}\left[\sqrt{n}\,(\bar{X}_n - \mu)\right] = \text{Var}\,(\sigma Z_n) = \sigma^2\,\text{Var}\,(Z_n) = \sigma^2$$

is constant. The WLLN states that

$$(\bar{X}_n - \mu) \xrightarrow{P} 0,$$

so $\sqrt{n}$ is a "magnification factor" that maintains random variables with a constant positive variance. We conclude that $1/\sqrt{n}$ measures how rapidly the sample mean tends toward the population mean.

Now we turn to the more refined question of how the distribution of the sample mean changes as the sample mean tends toward the population mean. By converting to standard units, we are able to distinguish changes in the shape of the distribution from changes in its mean and variance. Despite our inability to make general statements about the behavior of Z_1, it turns out that we can say quite a bit about the behavior of Z_n as n becomes large. The following theorem is one of the most remarkable and useful results in all of mathematics. It is fundamental to the study of both probability and statistics.

Theorem 8.2 *(Central Limit Theorem) Let $X_1, X_2, \ldots$ be any sequence of independent and identically distributed random variables having finite mean μ and finite variance σ^2. Let*

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}},$$

let F_n denote the cdf of Z_n, and let Φ denote the cdf of the standard normal distribution. Then, for any fixed value $z \in \Re$,

$$P\,(Z_n \leq z) = F_n(z) \to \Phi(z)$$

as $n \to \infty$.

The Central Limit Theorem (CLT) states that the behavior of the average (or, equivalently, the sum) of a large number of independent and identically distributed random variables will resemble the behavior of a standard normal random variable. *This is true regardless of the distribution of the random variables that are being averaged.* Thus, the CLT allows us to approximate a variety of probabilities that otherwise would be intractable. Of course, we require some sense of how many random variables must be averaged in order for the normal approximation to be reasonably accurate. This *does* depend on the distribution of the random variables, but a popular rule of thumb is that the normal approximation can be used if $n \geq 30$. Often, the normal approximation works quite well with even smaller n.

Example 8.1 *A chemistry professor is attempting to determine the conformation of a certain molecule. To measure the distance between a pair of nearby hydrogen atoms, she uses NMR spectroscopy. She knows that this measurement procedure has an expected value equal to the actual distance and a standard deviation of* 0.5 *angstroms. If she replicates the experiment* 36 *times, then what is the probability that the average measured value will fall within* 0.1 *angstroms of the true value?*

Let X_i denote the measurement obtained from replication i, for $i = 1, \ldots, 36$. We are told that $\mu = EX_i$ is the actual distance between the atoms and that $\sigma^2 = \text{Var}\, X_i = 0.5^2$. Let $Z \sim \text{Normal}(0, 1)$. Then, applying the CLT,

$$\begin{aligned} P\left(\mu - 0.1 < \bar{X}_{36} < \mu + 0.1\right) &= P\left(\mu - 0.1 - \mu < \bar{X}_{36} - \mu < \mu + 0.1 - \mu\right) \\ &= P\left(\frac{-0.1}{0.5/6} < \frac{\bar{X}_{36} - \mu}{0.5/6} < \frac{0.1}{0.5/6}\right) \\ &= P\left(-1.2 < Z_n < 1.2\right) \\ &\doteq P\left(-1.2 < Z < 1.2\right) \\ &= \Phi(1.2) - \Phi(-1.2). \end{aligned}$$

Now we use R:

```
> pnorm(1.2)-pnorm(-1.2)
[1] 0.7698607
```

We conclude that there is a chance of approximately 77% that the average of the measured values will fall with 0.1 angstroms of the true value.

Notice that it is not possible to compute the exact probability. To do so would require knowledge of the distribution of the X_i.

It is sometimes useful to rewrite the normal approximations derived from the CLT as statements of the approximate distributions of the sum and the average. For the sum we obtain

$$\sum_{i=1}^{n} X_i \dot{\sim} \text{Normal}\left(n\mu, n\sigma^2\right) \tag{8.1}$$

and for the average we obtain

$$\bar{X}_n \dot{\sim} \text{Normal}\left(\mu, \frac{\sigma^2}{n}\right). \tag{8.2}$$

These approximations are especially useful when combined with Theorem 5.2.

Example 8.2 *The chemistry professor in Example 8.1 asks her graduate student to replicate the experiment that she performed an additional* 64 *times. What is the probability that the averages of their respective measured values will fall within* 0.1 *angstroms of each other?*

The professor's measurements are

$$X_1, \ldots, X_{36} \sim \left(\mu, 0.5^2\right).$$

Applying (8.2), we obtain

$$\bar{X}_{36} \dot{\sim} \text{Normal}\left(\mu, \frac{0.25}{36}\right).$$

Similarly, the student's measurements are

$$Y_1, \ldots, Y_{64} \sim \left(\mu, 0.5^2\right).$$

Applying (8.2), we obtain

$$\bar{Y}_{64} \dot{\sim} \text{Normal}\left(\mu, \frac{0.25}{64}\right) \quad \text{or} \quad -\bar{Y}_{64} \dot{\sim} \text{Normal}\left(-\mu, \frac{0.25}{64}\right).$$

Now we apply Theorem 5.2 to conclude that

$$\bar{X}_{36} - \bar{Y}_{64} = \bar{X}_{36} + (-\bar{Y}_{64}) \dot{\sim} \text{Normal}\left(0, \frac{0.25}{36} + \frac{0.25}{64} = \frac{5^2}{48^2}\right).$$

Converting to standard units, it follows that

$$\begin{aligned} P\left(-0.1 < \bar{X}_{36} - \bar{Y}_{64} < 0.1\right) &= P\left(\frac{-0.1}{5/48} < \frac{\bar{X}_{36} - \bar{Y}_{64}}{5/48} < \frac{0.1}{5/48}\right) \\ &\doteq P\left(-0.96 < Z < 0.96\right) \\ &= \Phi(0.96) - \Phi(-0.96). \end{aligned}$$

Now we use `R`:

```
> pnorm(.96)-pnorm(-.96)
[1] 0.6629448
```

We conclude that there is a chance of approximately 66% that the two averages will fall with 0.1 angstroms of each other.

The CLT has a long history. For the special case of $X_i \sim \text{Bernoulli}(p)$, a version of the CLT was obtained by De Moivre in the 1730s. The first attempt at a more general CLT was made by Laplace in 1810, but definitive results were not obtained until the second quarter of the 20th century. Theorem 8.2 is actually a very special case of far more general results established during that period. However, with one exception to which we now turn, it is sufficiently general for our purposes.

The astute reader may have noted that, in Examples 8.1 and 8.2, we assumed that the population mean μ was unknown but that the population variance σ^2 was known. Is this plausible? In Examples 8.1 and 8.2, it might be that the nature of the instrumentation is sufficiently well understood that the population variance may be considered known. In general, however, it seems somewhat implausible that we would know the population variance and not know the population mean.

The normal approximations employed in Examples 8.1 and 8.2 require knowledge of the population variance. If the variance is not known, then it must be estimated from the measured values. Chapters 7 and 9 will introduce procedures for doing so. In anticipation of those procedures, we state the following generalization of Theorem 8.2:

Theorem 8.3 *Let $X_1, X_2, \ldots$ be any sequence of independent and identically distributed random variables having finite mean μ and finite variance σ^2. Suppose that $D_1, D_2, \ldots$ is a sequence of random variables with the property that $D_n^2 \xrightarrow{P} \sigma^2$ and let*

$$T_n = \frac{\bar{X}_n - \mu}{D_n/\sqrt{n}}.$$

Let F_n denote the cdf of T_n, and let Φ denote the cdf of the standard normal distribution. Then, for any fixed value $t \in \Re$,

$$P(T_n \leq t) = F_n(t) \to \Phi(t)$$

as $n \to \infty$.

We conclude this section with a warning. Statisticians usually invoke the CLT in order to approximate the distribution of a sum or an average of random variables $X_1, \ldots, X_n$ that are observed in the course of an experiment. The X_i need not be normally distributed themselves—indeed, the grandeur of the CLT is that it does *not* assume normality of the X_i. Nevertheless, we will discover that many important statistical procedures do assume that the X_i are normally distributed. Researchers who hope to use these procedures naturally want to believe that their X_i are normally distributed. Often, they look to the CLT for reassurance. Many think that, if only they replicate their experiment enough times, then somehow their observations will be drawn from a normal distribution. This is absurd! Suppose that a fair coin is tossed once. Let X_1 denote the number of `Heads`, so that $X_1 \sim$ Bernoulli(0.5). The Bernoulli distribution is not at all like a normal distribution. If we toss the coin one million times, then each $X_i \sim$ Bernoulli(0.5). The Bernoulli distribution does not miraculously become a normal distribution. Remember,

> *The Central Limit Theorem does* not *say that a large sample was necessarily drawn from a normal distribution!*

On some occasions, it is possible to invoke the CLT to anticipate that the random variable to be observed will behave like a normal random variable. This involves recognizing that the observed random variable is the sum or the average of lots of independent and identically distributed random variables that are not observed.

Example 8.3 *To study the effect of an insect growth regulator (IGR) on termite appetite, an entomologist plans an experiment. Each replication of the experiment will involve placing* 100 *ravenous termites in a container with a dried block of wood. The block of wood will be weighed before the experiment begins and after a fixed number of days. The random variable of interest is the decrease in weight, the amount of wood consumed by the termites. Can we anticipate the distribution of this random variable?*

The total amount of wood consumed is the sum of the amounts consumed by each termite. Assuming that the termites behave independently and identically, the CLT suggests that this sum should be approximately normally distributed.

When reasoning as in Example 8.3, one should construe the CLT as no more than suggestive. Most natural processes are far too complicated to be modelled so simplistically with any guarantee of accuracy. One should *always* examine the observed values to see if they are consistent with one's theorizing. The next chapter will introduce several techniques for doing precisely that.

8.4 Exercises

1. Suppose that I toss a fair coin 100 times and observe 60 `Heads`. Now I decide to toss the same coin another 100 times. Does the Law of Averages imply that I should expect to observe another 40 `Heads`?

2. In Example 7.7, we observed a sample of size $n = 100$. A normal probability plot and kernel density estimate constructed from this sample suggested that the observations had been drawn from a nonnormal distribution. True or False: *It follows from the Central Limit Theorem that a kernel density estimate constructed from a much larger sample would more closely resemble a normal distribution.*

3. Suppose that an astragalus has the following probabilities of producing the four possible uppermost faces: $P(1) = P(6) = 0.1$, $P(3) = P(4) = 0.4$. This astragalus is to be thrown 100 times. Let X_i denote the value of the uppermost face that results from throw i.

   (a) Compute the expected value and the variance of X_i.

   (b) Compute the probability that the average value of the 100 throws will exceed 3.6.

4. Chris owns a laser pointer that is powered by two `AAAA` batteries. A pair of batteries will power the pointer for an average of five hours use, with a standard deviation of 30 minutes. Chris decides to take advantage of a sale and buys 20 2-packs of `AAAA` batteries. What is the probability that he will get to use his laser pointer for at least 105 hours before he needs to buy more batteries?

5. A certain financial theory posits that daily fluctuations in stock prices are independent random variables. Suppose that the daily price fluctuations (in dollars) of a certain blue-chip stock are independent and identically distributed random variables $X_1, X_2, X_3, \ldots$, with $EX_i = 0.01$ and $\operatorname{Var} X_i = 0.01$. (Thus, if today's price of this stock is \$50, then tomorrow's price is $\$50 + X_1$, etc.) Suppose that the daily price fluctuations (in dollars) of a certain internet stock are independent and identically distributed random variables $Y_1, Y_2, Y_3, \ldots$, with $EY_j = 0$ and $\operatorname{Var} Y_j = 0.25$.

   Now suppose that both stocks are currently selling for \$50 per share and you wish to invest \$50 in one of these two stocks for a period of 400 market days. Assume that the costs of purchasing and selling a share of either stock are zero.

   (a) Approximate the probability that you will make a profit on your investment if you purchase a share of the blue-chip stock.

   (b) Approximate the probability that you will make a profit on your investment if you purchase a share of the internet stock.

   (c) Approximate the probability that you will make a profit of at least \$20 if you purchase a share of the blue-chip stock.

   (d) Approximate the probability that you will make a profit of at least \$20 if you purchase a share of the internet stock.

   (e) Assuming that the internet stock fluctuations and the blue-chip stock fluctuations are independent, approximate the probability that, after 400 days, the price of the internet stock will exceed the price of the blue-chip stock.

Chapter 9

Inference

In Chapters 3–8 we developed methods for studying the behavior of random variables. Given a specific probability distribution, we can calculate the probabilities of various events. For example, knowing that $Y \sim \text{Binomial}(n = 100; p = 0.5)$, we can calculate $P(40 \leq Y \leq 60)$. Roughly speaking, statistics is concerned with the opposite sort of problem. For example, knowing that $Y \sim \text{Binomial}(n = 100; p)$, *where the value of* p *is unknown*, and having observed $Y = y$ (say $y = 32$), what can we say about p? The phrase *statistical inference* describes any procedure for extracting information about a probability distribution from an observed sample.

The present chapter introduces the fundamental principles of statistical inference. We will discuss three types of statistical inference—point estimation, hypothesis testing, and set estimation—in the context of drawing inferences about a single population mean. More precisely, we will consider the following situation:

1. $X_1, \ldots, X_n$ are independent and identically distributed random variables. We observe a sample, $\vec{x} = \{x_1, \ldots, x_n\}$.

2. Both $EX_i = \mu$ and $\text{Var}\, X_i = \sigma^2$ exist and are finite. We are interested in drawing inferences about the population mean μ, a quantity that is fixed but unknown.

3. The sample size, n, is sufficiently large that we can use the normal approximation provided by the Central Limit Theorem.

We begin, in Section 9.1, by examining a narrative that is sufficiently nuanced to motivate each type of inferential technique. We then proceed to

discuss point estimation (Section 9.2), hypothesis testing (Sections 9.3 and 9.4), and set estimation (Section 9.5). Although we are concerned exclusively with large-sample inferences about a single population mean, it should be appreciated that this concern often arises in practice. More importantly, the fundamental concepts that we introduce in this context are common to virtually all problems that involve statistical inference.

9.1 A Motivating Example

We consider an artificial example that permits us to scrutinize the precise nature of statistical reasoning. Two siblings, a magician (Arlen) and an attorney (Robin) agree to resolve their disputed ownership of an Erté painting by tossing a penny. Arlen produces a penny and, just as Robin is about to toss it in the air, Arlen smoothly suggests that spinning the penny on a table might ensure better randomization. Robin assents and spins the penny. As it spins, Arlen calls "Tails!" The penny comes to rest with `Tails` facing up and Arlen takes possession of the Erté. Robin is left with the penny.

That evening, Robin wonders if she has been had. She decides to perform an experiment. She spins the same penny on the same table 100 times and observes 68 `Tails`. It occurs to Robin that perhaps spinning this penny was not entirely fair, but she is reluctant to accuse her brother of impropriety until she is convinced that the results of her experiment cannot be dismissed as coincidence. How should she proceed?

It is easy to devise a mathematical model of Robin's experiment: each spin of the penny is a Bernoulli trial and the experiment is a sequence of $n = 100$ trials. Let X_i denote the outcome of spin i, where $X_i = 1$ if `Heads` is observed and $X_i = 0$ if `Tails` is observed. Then $X_1, \ldots, X_{100} \sim \text{Bernoulli}(p)$, where p is the fixed but unknown (to Robin!) probability that a single spin will result in `Heads`. The probability distribution Bernoulli(p) is our mathematical abstraction of a population and the population parameter of interest is $\mu = EX_i = p$, the population mean.

Let

$$Y = \sum_{i=1}^{100} X_i,$$

the total number of `Heads` obtained in $n = 100$ spins. Under the mathematical model that we have proposed, $Y \sim \text{Binomial}(p)$. In performing her

experiment, Robin observes a sample $\vec{x} = \{x_1, \ldots, x_{100}\}$ and computes

$$y = \sum_{i=1}^{100} x_i,$$

the total number of `Heads` in her sample. In our narrative, $y = 32$.

We emphasize that $p \in [0, 1]$ is fixed but unknown. Robin's goal is to draw inferences about this fixed but unknown quantity. We consider three questions that she might ask:

1. What is the true value of p? More precisely, what is a reasonable guess as to the true value of p?
2. Is $p = 0.5$? Specifically, is the evidence that $p \neq 0.5$ so compelling that Robin can comfortably accuse Arlen of impropriety?
3. What are plausible values of p? In particular, is there a subset of $[0, 1]$ that Robin can confidently claim contains the true value of p?

The first set of questions introduces a type of inference that statisticians call *point estimation*. We have already encountered (in Chapter 7) a natural approach to point estimation, the plug-in principle. In the present case, the plug-in principle suggests estimating the theoretical probability of success, p, by computing the observed proportion of successes,

$$\hat{p} = \frac{y}{n} = \frac{32}{100} = 0.32.$$

The second set of questions introduces a type of inference that statisticians call *hypothesis testing*. Having calculated $\hat{p} = 0.32 \neq 0.5$, Robin is inclined to guess that $p \neq 0.5$. But how compelling is the evidence that $p \neq 0.5$? Let us play devil's advocate: perhaps $p = 0.5$, but chance produced "only" $y = 32$ instead of a value nearer $EY = np = 100 \times 0.5 = 50$. This is a possibility that we can quantify. If $Y \sim \text{Binomial}(n = 100; p = 0.5)$, then the probability that Y will deviate from its expected value by at least $|50 - 32| = 18$ is

$$\begin{aligned}
\mathbf{p} &= P\left(|Y - 50| \geq 18\right) \\
&= P(Y \leq 32 \text{ or } Y \geq 68) \\
&= P(Y \leq 32) + P(Y \geq 68) \\
&= P(Y \leq 32) + 1 - P(Y \leq 67) \\
&= \texttt{pbinom(32,100,.5)+1-pbinom(67,100,.5)} \\
&= 0.0004087772.
\end{aligned}$$

This *significance probability* seems fairly small—perhaps small enough to convince Robin that in fact $p \neq 0.5$.

The third set of questions introduces a type of inference that statisticians call *set estimation*. We have just tested the possibility that $p = p_0$ in the special case $p_0 = 0.5$. Now, imagine testing the possibility that $p = p_0$ for each $p_0 \in [0, 1]$. Those p_0 that are not rejected as inconsistent with the observed data, $y = 32$, will constitute a set of plausible values of p.

To implement this procedure, Robin will have to adopt a standard of implausibility. Perhaps she decides to reject p_0 as implausible when the corresponding significance probability,

$$\begin{aligned} \mathbf{p} &= P\left(|Y - 100p_0| \geq |32 - 100p_0|\right) \\ &= P\left(Y - 100p_0 \geq |32 - 100p_0|\right) + P\left(Y - 100p_0 \leq -|32 - 100p_0|\right) \\ &= P\left(Y \geq 100p_0 + |32 - 100p_0|\right) + P\left(Y \leq 100p_0 - |32 - 100p_0|\right), \end{aligned}$$

satisfies $\mathbf{p} \leq 0.1$. Recalling that $Y \sim \text{Binomial}(100; p_0)$ and using the `R` function `pbinom`, some trial and error reveals that $\mathbf{p} > 0.1$ if p_0 lies in the interval $[0.245, 0.404]$. (The endpoints of this interval are included.) Notice that this interval does *not* contain $p_0 = 0.5$, which we had already rejected as implausible.

9.2 Point Estimation

The goal of point estimation is to make a reasonable guess of the unknown value of a designated population quantity, e.g., the population mean. The quantity that we hope to guess is called the *estimand*.

9.2.1 Estimating a Population Mean

Suppose that the estimand is μ, the population mean. The plug-in principle suggests estimating μ by computing the mean of the empirical distribution. This leads to the plug-in estimate of μ, $\hat{\mu} = \bar{x}_n$. Thus, we estimate the mean of the population by computing the mean of the sample, which is certainly a natural thing to do.

We will distinguish between

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

a real number that is calculated from the sample $\vec{x} = \{x_1, \ldots, x_n\}$, and

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

a random variable that is a function of the random variables $X_1, \ldots, X_n$. (Such a random variable is called a *statistic*.) The latter is our rule for guessing, an *estimation procedure* or *estimator*. The former is the guess itself, the result of applying our rule for guessing to the sample that we observed, an *estimate*.

The quality of an individual estimate depends on the individual sample from which it was computed and is therefore affected by chance variation. Furthermore, it is rarely possible to assess how close to correct an individual estimate may be. For these reasons, we study estimation procedures and identify the statistical properties that these random variables possess. In the present case, two properties are worth noting:

1. We know that $E\bar{X}_n = \mu$. Thus, on the average, our procedure for guessing the population mean produces the correct value. We express this property by saying that $\bar{X}_n$ is an *unbiased* estimator of μ.

   The property of unbiasedness is intuitively appealing and sometimes is quite useful. However, many excellent estimation procedures are biased and some unbiased estimators are unattractive. For example, $EX_1 = \mu$ by definition, so X_1 is also an unbiased estimator of μ; but most researchers would find the prospect of estimating a population mean with a single observation to be rather unappetizing. Indeed,

   $$\operatorname{Var} \bar{X}_n = \frac{\sigma^2}{n} < \sigma^2 = \operatorname{Var} X_1,$$

   so the unbiased estimator $\bar{X}_n$ has smaller variance than the unbiased estimator X_1.

2. The Weak Law of Large Numbers states that $\bar{X}_n \xrightarrow{P} \mu$. Thus, as the sample size increases, the estimator $\bar{X}_n$ converges in probability to the estimand μ. We express this property by saying that $\bar{X}_n$ is a *consistent* estimator of μ.

   The property of consistency is essential—it is difficult to conceive a circumstance in which one would be willing to use an estimation procedure that might fail regardless of how much data one collected. Notice that the unbiased estimator X_1 is not consistent.

9.2.2 Estimating a Population Variance

Now suppose that the estimand is σ^2, the population variance. Although we are concerned with drawing inferences about the population mean, we will discover that hypothesis testing and set estimation may require knowing the population variance. If the population variance is not known, then it must be estimated from the sample.

The plug-in principle suggests estimating σ^2 by computing the variance of the empirical distribution. This leads to the plug-in estimate of σ^2,

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 .$$

The plug-in estimator of σ^2 is *biased*; in fact,

$$E\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2\right] = \frac{n-1}{n}\sigma^2 < \sigma^2.$$

This does not present any particular difficulties; however, if we desire an unbiased estimator, then we simply multiply the plug-in estimator by the factor $(n-1)/n$, obtaining

$$S_n^2 = \frac{n}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2\right] = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2 . \qquad (9.1)$$

The statistic S_n^2 is the most popular estimator of σ^2 and many books refer to the estimate

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$$

as *the* sample variance. (For example, the `R` command `var` computes s_n^2.) In fact, both estimators are perfectly reasonable, consistent estimators of σ^2. We will prefer S_n^2 for the rather mundane reason that using it will simplify some of the formulas that we will encounter.

9.3 Heuristics of Hypothesis Testing

Hypothesis testing is appropriate for situations in which one wants to guess which of two possible statements about a population is correct. For example, in Section 9.1 we considered the possibility that spinning a penny is fair ($p = 0.5$) versus the possibility that spinning a penny is not fair ($p \neq 0.5$). The logic of hypothesis testing is of a familar sort:

> *If an alleged coincidence seems too implausible, then we tend to believe that it wasn't really a coincidence.*

Man has engaged in this kind of reasoning for millenia. In Cicero's *De Divinatione*, Quintus exclaims:

> "They are entirely fortuitous you say? Come! Come! Do you really mean that? ... When the four dice [astragali] produce the venus-throw you may talk of accident: but suppose you made a hundred casts and the venus-throw appeared a hundred times; could you call that accidental?"[1]

The essence of hypothesis testing is captured by the familiar saying, "Where there's smoke, there's fire." In this section we formalize such reasoning, appealing to three prototypical examples:

1. Assessing circumstantial evidence in a criminal trial.

   For simplicity, suppose that the defendant has been charged with a single count of pre-meditated murder and that the jury has been instructed to either convict of murder in the first degree or acquit. The defendant had motive, means, and opportunity. Furthermore, two types of blood were found at the crime scene. One type was evidently the victim's. Laboratory tests demonstrated that the other type was not the victim's, but failed to demonstrate that it was not the defendant's. What should the jury do?

   The evidence used by the prosecution to try to establish a connection between the blood of the defendant and blood found at the crime scene is probabilistic, i.e., circumstantial. It will likely be presented to the jury in the language of mathematics, e.g., "Both blood samples have characteristics x, y and z; yet only 0.5% of the population has such blood." The defense will argue that this is merely an unfortunate coincidence. The jury must evaluate the evidence and decide whether or not such a coincidence is too extraordinary to be believed, i.e., they must decide if their assent to the proposition that the defendant committed the murder rises to a level of certainty sufficient to convict.

[1] Cicero rejected the conclusion that a run of one hundred venus-throws is so improbable that it must have been caused by divine intervention; however, Cicero was castigating the practice of divination. Quintus was entirely correct in suggesting that a run of one hundred venus-throws should not be rationalized as "entirely fortuitous." A modern scientist might conclude that an unusual set of astragali had been used to produce this remarkable result.

If the combined weight of the evidence against the defendant is a chance of one in ten, then the jury is likely to acquit; if it is a chance of one in a million, then the jury is likely to convict.

2. Assessing data from a scientific experiment.

   A study[2] of termite foraging behavior reached the controversial conclusion that two species of termites compete for scarce food resources. In this study, a site in the Sonoran desert was cleared of dead wood and toilet paper rolls were set out as food sources. The rolls were examined regularly over a period of many weeks and it was observed that only very rarely was a roll infested with both species of termites. Was this just a coincidence or were the two species competing for food?

   The scientists constructed a mathematical model of termite foraging behavior under the assumption that the two species forage independently of each other. This model was then used to quantify the probability that infestation patterns such as the one observed arise due to chance. This probability turned out to be just one in many billions—a coincidence far too extraordinary to be dismissed as such—and the researchers concluded that the two species were competing.

3. Assessing the results of Robin's penny-spinning experiment.

   In Section 9.1, we noted that Robin observed only $y = 32$ `Heads` when she would expect $EY = 50$ `Heads` if indeed $p = 0.5$. This is a discrepancy of $|32 - 50| = 18$, and we considered that possibility that such a large discrepancy might have been produced by chance. More precisely, we calculated $\mathbf{p} = P(|Y - EY| \geq 18)$ under the assumption that $p = 0.5$, obtaining $\mathbf{p} \doteq 0.0004$. On this basis, we speculated that Robin might be persuaded to accuse her brother of cheating.

In each of the preceding examples, a binary decision was based on a level of assent to probabilistic evidence. At least conceptually, this level can be quantified as a *significance probability*, which we loosely interpret to mean the probability that chance would produce a coincidence at least as extraordinary as the phenomenon observed. This begs an obvious question, which we pose now for subsequent consideration: how small should a significance probability be for one to conclude that a phenomenon is not a coincidence?

[2] S.C. Jones and M.W. Trosset (1991). Interference competition in desert subterranean termites. *Entomologia Experimentalis et Applicata*, 61:83–90.

We now proceed to explicate a formal model for statistical hypothesis testing that was proposed by J. Neyman and E. S. Pearson in the late 1920s and 1930s. Our presentation relies heavily on drawing simple analogies to criminal law, which we suppose is a more familiar topic than statistics to most students.

The States of Nature

The states of nature are the possible mechanisms that might have produced the observed phenomenon. Mathematically, they are the possible probability distributions under consideration. Thus, in the penny-spinning example, the states of nature are the Bernoulli trials indexed by $p \in [0, 1]$. In hypothesis testing, the states of nature are partitioned into two sets or *hypotheses*. In the penny-spinning example, the hypotheses that we formulated were $p = 0.5$ (penny-spinning is fair) and $p \neq 0.5$ (penny-spinning is not fair); in the legal example, the hypotheses are that the defendant did commit the murder (the defendant is factually guilty) and that the defendant did not commit the murder (the defendant is factually innocent).

The goal of hypothesis testing is to decide which hypothesis is correct, i.e., which hypothesis contains the true state of nature. In the penny-spinning example, Robin wants to determine whether or not penny-spinning is fair. In the termite example, Jones and Trosset wanted to determine whether or not termites were foraging independently. More generally, scientists usually partition the states of nature into a hypothesis that corresponds to a theory that the experiment is designed to investigate and a hypothesis that corresponds to a chance explanation; the goal of hypothesis testing is to decide which explanation is correct. In a criminal trial, the jury would like to determine whether the defendant is factually innocent of factually guilty—in the words of the United States Supreme Court in *Bullington v. Missouri* (1981):

> Underlying the question of guilt or innocence is an objective truth: the defendant did or did not commit the crime. From the time an accused is first suspected to the time the decision on guilt or innocence is made, our system is designed to enable the trier of fact to discover that truth.

Formulating appropriate hypotheses can be a delicate business. In the penny-spinning example, we formulated hypotheses $p = 0.5$ and $p \neq 0.5$. These hypotheses are appropriate if Robin wants to determine whether or

not penny-spinning is fair. However, one can easily imagine that Robin is not interested in whether or not penny-spinning is fair, but rather in whether or not her brother gained an advantage by using the procedure. If so, then appropriate hypotheses would be $p < 0.5$ (penny-spinning favored Arlen) and $p \geq 0.5$ (penny-spinning did not favor Arlen).

The Actor

The states of nature having been partitioned into two hypotheses, it is necessary for a decisionmaker (the actor) to choose between them. In the penny-spinning example, the actor is Robin; in the termite example, the actor is the team of researchers; in the legal example, the actor is the jury.

Statisticians often describe hypothesis testing as a game that they play against Nature. To study this game in greater detail, it becomes necessary to distinguish between the two hypotheses under consideration. In each example, we declare one hypothesis to be the *null hypothesis* (H_0) and the other to be the *alternative hypothesis* (H_1). Roughly speaking, the logic for determining which hypothesis is H_0 and which is H_1 is the following: H_0 should be the hypothesis to which one defaults if the evidence is equivocal and H_1 should be the hypothesis that one requires compelling evidence to embrace.

We shall have a great deal more to say about distinguishing null and alternative hypotheses, but for now suppose that we have declared the following: (1) H_0: the defendant did not commit the murder, (2) H_0: the termites are foraging independently, and (3) H_0: spinning the penny is fair. Having done so, the game takes the following form:

		State of Nature	
		H_0	H_1
Actor's	H_0		Type II error
Choice	H_1	Type I error	

There are four possible outcomes to this game, two of which are favorable and two of which are unfavorable. If the actor chooses H_1 when in fact H_0 is true, then we say that a Type I error has been committed. If the actor chooses H_0 when in fact H_1 is true, then we say that a Type II error has been committed. In a criminal trial, a Type I error occurs when a jury convicts a factually innocent defendant and a Type II error occurs when a jury acquits a factually guilty defendant.

Innocent Until Proven Guilty

Because we are concerned with probabilistic evidence, any decision procedure that we devise will occasionally result in error. Obviously, we would like to devise procedures that minimize the probabilities of committing errors. Unfortunately, there is an inevitable tradeoff between Type I and Type II error that precludes simultaneously minimizing the probabilities of both types. To appreciate this, consider two juries. The first jury always acquits and the second jury always convicts. Then the first jury *never* commits a Type I error and the second jury *never* commits a Type II error. The only way to simultaneously better both juries is to never commit an error of either type, which is impossible with probabilistic evidence.

The distinguishing feature of hypothesis testing (and Anglo-American criminal law) is the manner in which it addresses the tradeoff between Type I and Type II error. The Neyman-Pearson formulation of hypothesis testing accords the null hypothesis a privileged status: H_0 will be maintained unless there is compelling evidence against it. It is instructive to contrast the asymmetry of this formulation with situations in which neither hypothesis is privileged. In statistics, this is the problem of determining which hypothesis better explains the data. This is *discrimination*, not hypothesis testing. In law, this is the problem of determining whether the defendant or the plaintiff has the stronger case. This is the criterion in civil suits, not in criminal trials.

In the penny-spinning example, Robin required compelling evidence against the privileged null hypothesis that penny-spinning is fair to overcome her scruples about accusing her brother of impropriety. In the termite example, Jones and Trosset required compelling evidence against the privileged null hypothesis that two termite species forage independently in order to write a credible article claiming that two species were competing with each other. In a criminal trial, the principle of according the null hypothesis a privileged status has a familiar characterization: the defendant is "innocent until proven guilty."

According the null hypothesis a privileged status is equivalent to declaring Type I errors to be more egregious than Type II errors. This connection was eloquently articulated by Justice John Harlan in a 1970 Supreme Court decision: "If, for example, the standard of proof for a criminal trial were a preponderance of the evidence rather than proof beyond a reasonable doubt, there would be a smaller risk of factual errors that result in freeing guilty persons, but a far greater risk of factual errors that result in convicting the innocent."

A preference for Type II errors instead of Type I errors can often be glimpsed in scientific applications. For example, because science is conservative, it is generally considered better to wrongly accept than to wrongly reject the prevailing wisdom that termite species forage independently. Moreover, just as this preference is the foundation of statistical hypothesis testing, so is it a fundamental principle of criminal law. In his famous *Commentaries*, William Blackstone opined that "it is better that ten guilty persons escape, than that one innocent man suffer;" and in his influential *Practical Treatise on the Law of Evidence* (1824), Thomas Starkie suggested that "The maxim of the law... is that it is better that ninety-nine... offenders shall escape than that one innocent man be condemned." In *Reasonable Doubts* (1996), Alan Dershowitz quotes both maxims and notes anecdotal evidence that jurors actually do prefer committing Type II to Type I errors: on *Prime Time Live* (October 4, 1995), O.J. Simpson juror Anise Aschenbach stated, "If we made a mistake, I would rather it be a mistake on the side of a person's innocence than the other way."

Beyond a Reasonable Doubt

To actualize its antipathy to Type I errors, the Neyman-Pearson formulation imposes an upper bound on the maximal probability of Type I error that will be tolerated. This bound is the *significance level*, conventionally denoted α. The significance level is specified (prior to examining the data) and only decision rules for which the probability of Type I error is no greater than α are considered. Such tests are called *level α tests*.

To fix ideas, we consider the penny-spinning example and specify a significance level of α. Let $\mathbf{p}$ denote the significance probability that results from performing the analysis in Section 9.1 and consider a rule that rejects the null hypothesis $H_0 : p = 0.5$ if and only if $\mathbf{p} \leq \alpha$. Then a Type I error occurs if and only if $p = 0.5$ and we observe y such that $\mathbf{p} = P(|Y - 50| \geq |y - 50|) \leq \alpha$. We claim that the probability of observing such a y is just α, in which case we have constructed a level α test.

To see why this is the case, let $W = |Y - 50|$ denote the *test statistic*. The decision to accept or reject the null hypothesis H_0 depends on the observed value, w, of this random variable. Let

$$\mathbf{p}(w) = P_{H_0}(W \geq w)$$

denote the significance probability associated with w. Notice that w is the $1 - \mathbf{p}(w)$ quantile of the random variable W under H_0. Let q denote the

$1 - \alpha$ quantile of W under H_0, i.e.,

$$\alpha = P_{H_0}(W \geq q).$$

We reject H_0 if and only if we observe

$$P_{H_0}(W \geq w) = \mathbf{p}(w) \leq \alpha = P_{H_0}(W \geq q),$$

i.e., if and only $w \geq q$. If H_0 is true, then the probability of committing a Type I error is precisely

$$P_{H_0}(W \geq q) = \alpha,$$

as claimed above. We conclude that α quantifies the level of assent that we require to risk rejecting H_0, i.e., the significance level specifies how small a significance probability is required in order to conclude that a phenomenon is not a coincidence.

In statistics, the significance level α is a number in the interval $[0, 1]$. It is not possible to quantitatively specify the level of assent required for a jury to risk convicting an innocent defendant, but the legal principle is identical: in a criminal trial, the operative significance level is *beyond a reasonable doubt*. Starkie (1824) described the possible interpretations of this phrase in language derived from British empirical philosopher John Locke:

> Evidence which satisfied the minds of the jury of the truth of the fact in dispute, to the entire exclusion of every reasonable doubt, constitute full proof of the fact. ... Even the most direct evidence can produce nothing more than such a high degree of probability as amounts to moral certainty. From the highest it may decline, by an infinite number of gradations, until it produces in the mind nothing more than a preponderance of assent in favour of the particular fact.

The gradations that Starkie described are not intrinsically numeric, but it is evident that the problem of defining reasonable doubt in criminal law is the problem of specifying a significance level in statistical hypothesis testing.

In both criminal law and statistical hypothesis testing, actions typically are described in language that acknowledges the privileged status of the null hypothesis and emphasizes that the decision criterion is based on the probability of committing a Type I error. In describing the action of choosing H_0, many statisticians prefer the phrase "fail to reject the null hypothesis" to the less awkward "accept the null hypothesis" because choosing H_0 does

not imply an affirmation that H_0 is correct, only that the level of evidence against H_0 is not sufficiently compelling to warrant its rejection at significance level α. In precise analogy, juries render verdicts of "not guilty" rather than "innocent" because acquital does not imply an affirmation that the defendant did not commit the crime, only that the level of evidence against the defendant's innocence was not beyond a reasonable doubt.[3]

And To a Moral Certainty

The Neyman-Pearson formulation of statistical hypothesis testing is a mathematical abstraction. Part of its generality derives from its ability to accommodate *any* specified significance level. As a practical matter, however, α must be specified and we now ask how to do so.

In the penny-spinning example, Robin is making a personal decision and is free to choose α as she pleases. In the termite example, the researchers were guided by decades of scientific convention. In 1925, in his extremely influential *Statistical Methods for Research Workers*, Ronald Fisher[4] suggested that $\alpha = 0.05$ and $\alpha = 0.01$ are often appropriate significance levels. These suggestions were intended as practical guidelines, but they have become enshrined (especially $\alpha = 0.05$) in the minds of many scientists as a sort of Delphic determination of whether or not a hypothesized theory is true. While some degree of conformity is desirable (it inhibits a researcher from choosing—after the fact—a significance level that will permit rejecting the null hypothesis in favor of the alternative in which s/he may be invested), many statisticians are disturbed by the scientific community's slavish devotion to a single standard and by its often uncritical interpretation of the resulting conclusions.[5]

The imposition of an arbitrary standard like $\alpha = 0.05$ is possible because of the precision with which mathematics allows hypothesis testing to be formulated. Applying this precision to legal paradigms reveals the issues

[3]In contrast, Scottish law permits a jury to return a verdict of "not proven," thereby reserving a verdict of "not guilty" to affirm a defendant's innocence.

[4]Sir Ronald Fisher is properly regarded as the single most important figure in the history of statistics. It should be noted that he did not subscribe to all of the particulars of the Neyman-Pearson formulation of hypothesis testing. His fundamental objection to it, that it may not be possible to fully specify the alternative hypothesis, does not impact our development, since we are concerned with situations in which both hypotheses are fully specified.

[5]See, for example, J. Cohen (1994). The world is round ($p < .05$). *American Psychologist*, 49:997–1003.

with great clarity, but is of little practical value when specifying a significance level, i.e., when trying to define the meaning of "beyond a reasonable doubt." Nevertheless, legal scholars have endeavored for centuries to position "beyond a reasonable doubt" along the infinite gradations of assent that correspond to the continuum $[0, 1]$ from which α is selected. The phrase "beyond a reasonable doubt" is still often connected to the archaic phrase "to a moral certainty." This connection survived because moral certainty was actually a significance level, intended to invoke an enormous body of scholarly writings and specify a level of assent:

> Throughout this development two ideas to be conveyed to the jury have been central. The first idea is that there are two realms of human knowledge. In one it is possible to obtain the absolute certainty of mathematical demonstration, as when we say that the square of the hypotenuse is equal to the sum of the squares of the other two sides of a right triangle. In the other, which is the empirical realm of events, absolute certainty of this kind is not possible. The second idea is that, in this realm of events, just because absolute certainty is not possible, we ought not to treat everything as merely a guess or a matter of opinion. Instead, in this realm there are levels of certainty, and we reach higher levels of certainty as the quantity and quality of the evidence available to us increase. The highest level of certainty in this empirical realm in which no absolute certainty is possible is what traditionally was called "moral certainty," a certainty which there was no reason to doubt.[6]

Although it is rarely (if ever) possible to quantify a juror's level of assent, those comfortable with statistical hypothesis testing may be inclined to wonder what values of α correspond to conventional interpretations of reasonable doubt. If a juror believes that there is a 5 percent probability that chance alone could have produced the circumstantial evidence presented against a defendant accused of pre-meditated murder, is the juror's level of assent beyond a reasonable doubt and to a moral certainty? We hope not. We may be willing to tolerate a 5 percent probability of a Type I error when studying termite foraging behavior, but the analogous prospect of a 5

[6]Barbara J. Shapiro (1991). *"Beyond Reasonable Doubt" and "Probable Cause": Historical Perspectives on the Anglo-American Law of Evidence*, University of California Press, Berkeley, p. 41.

percent probability of wrongly convicting a factually innocent defendant is abhorrent.[7]

In fact, little is known about how anyone in the legal system quantifies reasonable doubt. Mary Gray cites a 1962 Swedish case in which a judge trying an overtime parking case explicitly ruled that a significance probability of 1/20736 was beyond reasonable doubt but that a significance probability of 1/144 was not.[8] In contrast, Alan Dershowitz relates a provocative classroom exercise in which his students preferred to acquit in one scenario with a significance probability of 10 percent and to convict in an analogous scenario with a significance probability of 15 percent.[9]

9.4 Testing Hypotheses About a Population Mean

We now apply the heuristic reasoning described in Section 9.3 to the problem of testing hypotheses about a population mean. Initially, we consider testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$.

The intuition that we are seeking to formalize is fairly straightfoward. By virtue of the Weak Law of Large Numbers, the observed sample mean ought to be fairly close to the true population mean. Hence, if the null hypothesis is true, then $\bar{x}_n$ ought to be fairly close to the hypothesized mean, μ_0. If we observe $\bar{X}_n = \bar{x}_n$ far from μ_0, then we guess that $\mu \neq \mu_0$, i.e., we reject H_0.

Given a significance level α, we want to calculate a significance probability $\mathbf{p}$. The significance level is a real number that is fixed by and known to the researcher, e.g., $\alpha = 0.05$. The significance probability is a real number that is determined by the sample, e.g., $\mathbf{p} \doteq 0.0004$ in Section 9.1. We will reject H_0 if and only if $\mathbf{p} \leq \alpha$.

In Section 9.3, we interpreted the significance probability as the probability that chance would produce a coincidence at least as extraordinary as the phenomenon observed. Our first challenge is to make this notion mathematically precise; how we do so depends on the hypotheses that we

[7] This discrepancy illustrates that the consequences of committing a Type I error influence the choice of a significance level. The consequences of Jones and Trosset wrongly concluding that termite species compete are not commensurate with the consequences of wrongly imprisoning a factually innocent citizen.

[8] M.W. Gray (1983). Statistics and the law. *Mathematics Magazine*, 56:67–81. As a graduate of Rice University, I cannot resist quoting another of Gray's examples of statistics-as-evidence: "In another case, that of millionaire W. M. Rice, the signature on his will was disputed, and the will was declared a forgery on the basis of probability evidence. As a result, the fortune of Rice went to found Rice Institute."

[9] A.M. Dershowitz (1996). *Reasonable Doubts*, Simon & Schuster, New York, p. 40.

want to test. In the present situation, we submit that a natural significance probability is

$$\mathbf{p} = P_{\mu_0}\left(\left|\bar{X}_n - \mu_0\right| \geq \left|\bar{x}_n - \mu_0\right|\right). \tag{9.2}$$

To understand why this is the case, it is essential to appreciate the following details:

1. The hypothesized mean, μ_0, is a real number that is fixed by and known to the researcher.

2. The estimated mean, $\bar{x}_n$, is a real number that is calculated from the observed sample and known to the researcher; hence, the quantity $|\bar{x}_n - \mu_0|$ is a fixed real number.

3. The estimator, $\bar{X}_n$, is a random variable. Hence, the inequality

$$\left|\bar{X}_n - \mu_0\right| \geq \left|\bar{x}_n - \mu_0\right| \tag{9.3}$$

   defines an event that may or may not occur each time the experiment is performed. Specifically, (9.3) is the event that the sample mean assumes a value at least as far from the hypothesized mean as the researcher observed.

4. The significance probability, $\mathbf{p}$, is the probability that (9.3) occurs. The notation P_{μ_0} reminds us that we are interested in the probability that this event occurs *under the assumption that the null hypothesis is true*, i.e., under the assumption that $\mu = \mu_0$.

Having formulated an appropriate significance probability for testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$, our second challenge is to find a way to compute $\mathbf{p}$. We remind the reader that we have assumed that n is large.

Case 1: The population variance is known or specified by the null hypothesis.

We define two new quantities, the random variable

$$Z_n = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}}$$

and the real number

$$z = \frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}}.$$

Under the null hypothesis $H_0 : \mu = \mu_0$, $Z_n \dot{\sim} \text{Normal}(0,1)$ by the Central Limit Theorem; hence,

$$\begin{aligned}
\mathbf{p} &= P_{\mu_0}\left(|\bar{X}_n - \mu_0| \geq |\bar{x}_n - \mu_0|\right) \\
&= 1 - P_{\mu_0}\left(-|\bar{x}_n - \mu_0| < \bar{X}_n - \mu_0 < |\bar{x}_n - \mu_0|\right) \\
&= 1 - P_{\mu_0}\left(-\frac{|\bar{x}_n - \mu_0|}{\sigma/\sqrt{n}} < \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} < \frac{|\bar{x}_n - \mu_0|}{\sigma/\sqrt{n}}\right) \\
&= 1 - P_{\mu_0}\left(-|z| < Z_n < |z|\right) \\
&\doteq 1 - [\Phi(|z|) - \Phi(-|z|)] \\
&= 2\Phi(-|z|),
\end{aligned}$$

which can be computed by the `R` command

```
> 2*pnorm(-abs(z))
```

or by consulting a table. An illustration of the normal probability of interest is sketched in Figure 9.1.

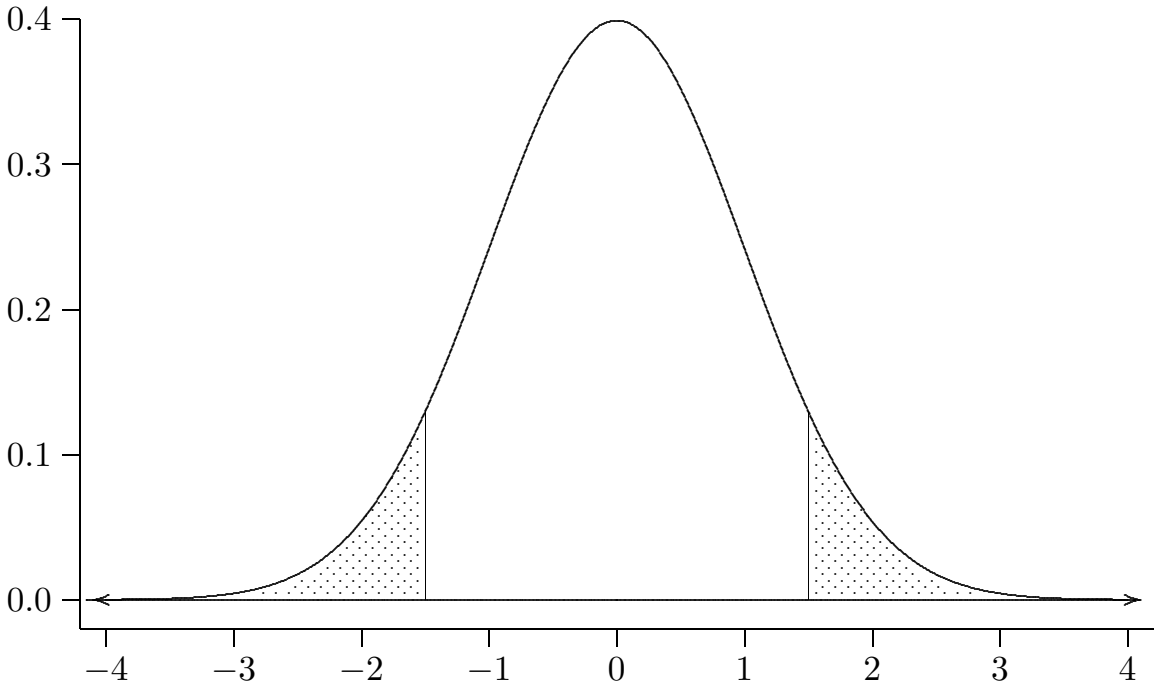

Figure 9.1: $P(|Z| \geq |z| = 1.5)$

An important example of Case 1 occurs when $X_i \sim \text{Bernoulli}(\mu)$. In this case, $\sigma^2 = \text{Var}\, X_i = \mu(1-\mu)$; hence, under the null hypothesis that $\mu = \mu_0$,

$\sigma^2 = \mu_0(1-\mu_0)$ and

$$z = \frac{\bar{x}_n - \mu_0}{\sqrt{\mu_0(1-\mu_0)/n}}.$$

Example 9.1 *To test $H_0 : \mu = 0.5$ versus $H_1 : \mu \neq 0.5$ at significance level $\alpha = 0.05$, we perform $n = 2500$ trials and observe 1200 successes. Should H_0 be rejected?*

The observed proportion of successes is $\bar{x}_n = 1200/2500 = 0.48$, so the value of the test statistic is

$$z = \frac{0.48 - 0.50}{\sqrt{0.5(1-0.5)/2500}} = \frac{-0.02}{0.5/50} = -2$$

and the significance probability is

$$\mathbf{p} \doteq 2\Phi(-2) \doteq 0.0456 < 0.05 = \alpha.$$

Because $\mathbf{p} \leq \alpha$, we reject H_0.

Case 2: The population variance is unknown.

Because σ^2 is unknown, we must estimate it from the sample. We will use the estimator introduced in Section 9.2,

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2,$$

and define

$$T_n = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}}.$$

Because S_n^2 is a consistent estimator of σ^2, i.e., $S_n^2 \xrightarrow{P} \sigma^2$, it follows from Theorem 8.3 that

$$\lim_{n\to\infty} P\left(T_n \leq z\right) = \Phi(z).$$

Just as we could use a normal approximation to compute probabilities involving Z_n, so can we use a normal approximation to compute probabilities involving T_n. The fact that we must estimate σ^2 slightly degrades the quality of the approximation; however, because n is large, we should observe an accurate estimate of σ^2 and the approximation should not suffer much. Accordingly, we proceed as in Case 1, using

$$t = \frac{\bar{x}_n - \mu_0}{s_n/\sqrt{n}}$$

instead of z.

Example 9.2 *To test $H_0 : \mu = 20$ versus $H_1 : \mu \neq 20$ at significance level $\alpha = 0.05$, we collect $n = 400$ observations, observing $\bar{x}_n = 21.82935$ and $s_n = 24.70037$. Should H_0 be rejected?*

The value of the test statistic is

$$t = \frac{21.82935 - 20}{24.70037/20} = 1.481234$$

and the significance probability is

$$\mathbf{p} \doteq 2\Phi(-1.481234) = 0.1385441 > 0.05 = \alpha.$$

Because $\mathbf{p} > \alpha$, we decline to reject H_0.

9.4.1 One-Sided Hypotheses

In Section 9.3 we suggested that, if Robin is not interested in whether or not penny-spinning is fair but rather in whether or not it favors her brother, then appropriate hypotheses would be $p < 0.5$ (penny-spinning favors Arlen) and $p \geq 0.5$ (penny-spinning does not favor Arlen). These are examples of one-sided (as opposed to two-sided) hypotheses.

More generally, we will consider two canonical cases:

$$\begin{array}{lll} H_0 : \mu \leq \mu_0 & \text{versus} & H_1 : \mu > \mu_0 \\ H_0 : \mu \geq \mu_0 & \text{versus} & H_1 : \mu < \mu_0 \end{array}$$

Notice that the possibility of equality, $\mu = \mu_0$, belongs to the null hypothesis in both cases. This is a technical necessity that arises because we compute significance probabilities using the μ in H_0 that is nearest H_1. For such a μ to exist, the boundary between H_0 and H_1 must belong to H_0. We will return to this necessity later in this section.

Instead of memorizing different formulas for different situations, we will endeavor to understand which values of our test statistic tend to undermine the null hypothesis in question. Such reasoning can be used on a case-by-case basis to determine the relevant significance probability. In so doing, sketching crude pictures can be quite helpful!

Consider testing each of the following:

$$\begin{array}{lllll} \text{(a)} & H_0 : \mu = \mu_0 & \text{versus} & H_1 : \mu \neq \mu_0 \\ \text{(b)} & H_0 : \mu \leq \mu_0 & \text{versus} & H_1 : \mu > \mu_0 \\ \text{(c)} & H_0 : \mu \geq \mu_0 & \text{versus} & H_1 : \mu < \mu_0 \end{array}$$

Qualitatively, we will be inclined to reject the null hypothesis if

(a) We observe $\bar{x}_n \ll \mu_0$ or $\bar{x}_n \gg \mu_0$, i.e., if we observe $|\bar{x}_n - \mu_0| \gg 0$.

This is equivalent to observing $|t| \gg 0$, so the significance probability is

$$\mathbf{p}_a = P_{\mu_0}(|T_n| \geq |t|).$$

(b) We observe $\bar{x}_n \gg \mu_0$, i.e., if we observe $\bar{x}_n - \mu_0 \gg 0$.

This is equivalent to observing $t \gg 0$, so the significance probability is

$$\mathbf{p}_b = P_{\mu_0}(T_n \geq t).$$

(c) We observe $\bar{x}_n \ll \mu_0$, i.e., if we observe $\bar{x}_n - \mu_0 \ll 0$.

This is equivalent to observing $t \ll 0$, so the significance probability is

$$\mathbf{p}_c = P_{\mu_0}(T_n \leq t).$$

Example 9.2 (continued) Applying the above reasoning, we obtain the significance probabilities sketched in Figure 9.2. Notice that $\mathbf{p}_b = \mathbf{p}_a/2$ and that $\mathbf{p}_b + \mathbf{p}_c = 1$. The probability $\mathbf{p}_b$ is fairly small, about 7%. This makes sense: we observed $\bar{x}_n \doteq 21.8 > 20 = \mu_0$, so the sample does contain *some* evidence that $\mu > 20$. However, the statistical test reveals that the strength of this evidence is not sufficiently compelling to reject $H_0 : \mu \leq 20$.

In contrast, the probability of $\mathbf{p}_c$ is quite large, about 93%. This also makes sense, because the sample contains *no* evidence that $\mu < 20$. In such instances, performing a statistical test only confirms that which is transparent from comparing the sample and hypothesized means.

9.4.2 Formulating Suitable Hypotheses

Examples 9.1 and 9.2 illustrated the mechanics of hypothesis testing. Once understood, the above techniques for calculating significance probabilities are fairly straightforward and can be applied routinely to a wide variety of problems. In contrast, determining suitable hypotheses to be tested requires one to carefully consider each situation presented. These determinations cannot be reduced to formulas. To make them requires good judgment, which can only be acquired through practice.

We now consider some examples that illustrate some important issues that arise when formulating hypotheses. In each case, there are certain key questions that must be answered: *Why was the experiment performed? Who needs to be convinced of what? Is one type of error perceived as more important than the other?*

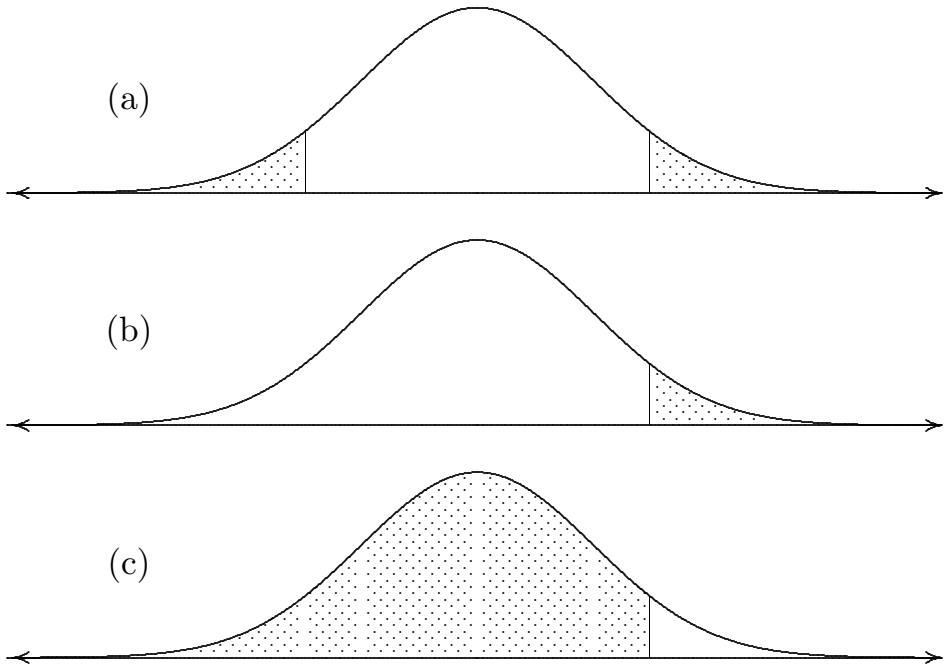

Figure 9.2: Significance probabilities for Example 9.2. Each significance probability is the area of the corresponding shaded region.

Example 9.3 *A group of concerned parents wants speed humps installed in front of a local elementary school, but the city traffic office is reluctant to allocate funds for this purpose. Both parties agree that humps should be installed if the average speed of all motorists who pass the school while it is in session exceeds the posted speed limit of* 15 *miles per hour (mph). Let* μ *denote the average speed of the motorists in question. A random sample of* $n = 150$ *of these motorists was observed to have a sample mean of* $\bar{x} = 15.3$ *mph with a sample standard deviation of* $s = 2.5$ *mph.*

(a) State null and alternative hypotheses that are appropriate from the parents' perspective.

(b) State null and alternative hypotheses that are appropriate from the city traffic office's perspective.

(c) Compute the value of an appropriate test statistic.

(d) Adopting the parents' perspective and assuming that they are willing to risk a 1% *chance of committing a Type I error, what action should be taken? Why?*

(e) Adopting the city traffic office's perspective and assuming that they are willing to risk a 10% chance of committing a Type I error, what action should be taken? Why?

Solution

(a) The parents would prefer to err on the side of protecting their children, so they would rather build unnecessary speed humps than forego necessary speed humps. Hence, they would like to see the hypotheses formulated so that foregoing necessary speed humps is a Type I error. Since speed humps will be built if it is concluded that $\mu > 15$ and will not be built if it is concluded that $\mu < 15$, the parents would prefer a null hypothesis of $H_0 : \mu \geq 15$ and an alternative hypothesis of $H_1 : \mu < 15$.

Equivalently, if we suppose that the purpose of the experiment is to provide evidence to the parents, then it is clear that the parents need to be persuaded that speed humps are unnecessary. The null hypothesis to which they will default in the absence of compelling evidence is $H_0 : \mu \geq 15$. They will require compelling evidence to the contrary, $H_1 : \mu < 15$.

(b) The city traffic office would prefer to err on the side of conserving their budget for important public works, so they would rather forego necessary speed humps than build unnecessary speed humps. Hence, they would like to see the hypotheses formulated so that building unnecessary speed humps is a Type I error. Since speed humps will be built if it is concluded that $\mu > 15$ and will not be built if it is concluded that $\mu < 15$, the city traffic office would prefer a null hypothesis of $H_0 : \mu \leq 15$ and an alternative hypothesis of $H_1 : \mu > 15$.

Equivalently, if we suppose that the purpose of the experiment is to provide evidence to the city traffic, then it is clear that the office needs to be persuaded that speed humps are necessary. The null hypothesis to which it will default in the absence of compelling evidence is $H_0 : \mu \leq 15$. It will require compelling evidence to the contrary, $H_1 : \mu > 15$.

(c) Because the population variance is unknown, the appropriate test statistic is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{15.3 - 15}{2.5/\sqrt{150}} \doteq 1.47.$$

(d) We would reject the null hypothesis in (a) if $\bar{x}$ is sufficiently smaller than $\mu_0 = 15$. Since $\bar{x} = 15.3 > 15$, there is no evidence against $H_0 : \mu \geq 15$. The null hypothesis is retained and speed humps are installed.

(e) We would reject the null hypothesis in (b) if $\bar{x}$ is sufficiently larger than $\mu_0 = 15$, i.e., for sufficiently large positive values of t. Hence, the significance probability is

$$\mathbf{p} = P\left(T_n \geq t\right) \doteq P(Z \geq 1.47) = 1 - \Phi(1.47) \doteq 0.071 < 0.10 = \alpha.$$

Because $\mathbf{p} \leq \alpha$, the traffic office should reject $H_0 : \mu \leq 15$ and install speed humps.

Example 9.4 *Imagine a variant of the Lanarkshire milk experiment described in Section 1.2. Suppose that it is known that 10-year-old Scottish schoolchildren gain an average of* 0.5 *pounds per month. To study the effect of daily milk supplements, a random sample of* $n = 1000$ *such children is drawn. Each child receives a daily supplement of 3/4 cups pasteurized milk. The study continues for four months and the weight gained by each student during the study period is recorded. Formulate suitable null and alternative hypotheses for testing the effect of daily milk supplements.*

Solution Let $X_1, \ldots, X_n$ denote the weight gains and let $\mu = EX_i$. Then milk supplements are effective if $\mu > 2$ and ineffective if $\mu < 2$. One of these possibilities will be declared the null hypothesis, the other will be declared the alternative hypothesis. The possibility $\mu = 2$ will be incorporated into the null hypothesis.

The alternative hypothesis should be the one for which compelling evidence is desired. Who needs to be convinced of what? The parents and teachers already believe that daily milk supplements are beneficial and would have to be convinced otherwise. But this is not the purpose of the study! The study is performed for the purpose of obtaining objective scientific evidence that supports prevailing popular wisdom. It is performed to convince government bureaucrats that spending money on daily milk supplements for schoolchildren will actually have a beneficial effect. The parents and teachers hope that the study will provide compelling evidence of this effect. Thus, the appropriate alternative hypothesis is $H_1 : \mu > 2$ and the appropriate null hypothesis is $H_0 : \mu \leq 2$.

9.4.3 Statistical Significance and Material Significance

The significance probability is the probability that a coincidence at least as extraordinary as the phenomenon observed can be produced by chance. The smaller the significance probability, the more confidently we reject the null hypothesis. However, it is one thing to be convinced that the null hypothesis is incorrect—it is something else to assert that the true state of nature is very different from the state(s) specified by the null hypothesis.

Example 9.5 A government agency requires prospective advertisers to provide statistical evidence that documents their claims. In order to claim that a gasoline additive increases mileage, an advertiser must fund an independent study in which n vehicles are tested to see how far they can drive, first without and then with the additive. Let X_i denote the increase in miles per gallon (mpg with the additive minus mpg without the additive) observed for vehicle i and let $\mu = EX_i$. The null hypothesis $H_0 : \mu \leq 1$ is tested against the alternative hypothesis $H_1 : \mu > 1$ and advertising is authorized if H_0 is rejected at a significance level of $\alpha = 0.05$.

Consider the experiences of two prospective advertisers:

1. A large corporation manufactures an additive that increases mileage by an average of $\mu = 1.01$ miles per gallon. The corporation funds a large study of $n = 900$ vehicles in which $\bar{x} = 1.01$ and $s = 0.1$ are observed. This results in a test statistic of

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{1.01 - 1.00}{0.1/\sqrt{900}} = 3$$

and a significance probability of

$$\mathbf{p} = P\left(T_n \geq t\right) \doteq P(Z \geq 3) = 1 - \Phi(3) \doteq 0.00135 < 0.05 = \alpha.$$

The null hypothesis is decisively rejected and advertising is authorized.

2. An amateur automotive mechanic invents an additive that increases mileage by an average of $\mu = 1.21$ miles per gallon. The mechanic funds a small study of $n = 9$ vehicles in which $\bar{x} = 1.21$ and $s = 0.4$ are observed. This results in a test statistic of

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{1.21 - 1.00}{0.4/\sqrt{9}} = 1.575$$

and (assuming that the normal approximation remains valid) a significance probability of

$$\mathbf{p} = P\left(T_n \geq t\right) \doteq P(Z \geq 1.575) = 1 - \Phi(1.575) \doteq 0.05763 > 0.05 = \alpha.$$

The null hypothesis is not rejected and advertising is not authorized.

These experiences are highly illuminating. Although the corporation's mean increase of $\mu = 1.01$ mpg is much closer to the null hypothesis than the mechanic's mean increase of $\mu = 1.21$ mpg, the corporation's study resulted in a much smaller significance probability. This occurred because of the smaller standard deviation and larger sample size in the corporation's study. As a result, the government could be more confident that the corporation's product had a mean increase of more than 1.0 mpg than they could be that the mechanic's product had a mean increase of more than 1.0 mpg.

The preceding example illustrates that a small significance probability does not imply a large physical effect and that a large physical effect does not imply a small significance probability. To avoid confusing these two concepts, statisticians distinguish between statistical significance and *material significance* (importance). To properly interpret the results of hypothesis testing, it is essential that one remember:

Statistical significance is not the same as material significance.

9.5 Set Estimation

Hypothesis testing is concerned with situations that demand a binary decision, e.g., whether or not to install speed humps in front of an elementary school. The relevance of hypothesis testing in situations that do not demand a binary decision is somewhat less clear. For example, many statisticians feel that the scientific community overuses hypothesis testing and that other types of statistical inference are often more appropriate. As we have discussed, a typical application of hypothesis testing in science partitions the states of nature into two sets, one that corresponds to a theory and one that corresponds to chance. Usually the theory encompasses a great many possible states of nature and the mere conclusion that the theory is true only begs the question of which states of nature are actually plausible. Furthermore, it is a rather fanciful conceit to imagine that a single scientific article should attempt to decide whether a theory is or is not true. A more

sensible enterprise for the authors to undertake is simply to set forth the evidence that they have discovered and allow evidence to accumulate until the scientific community reaches a consensus. One way to accomplish this is for each article to identify what its authors consider a set of plausible values for the population quantity in question.

To construct a set of plausible values of μ, we imagine testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ for *every* $\mu_0 \in (-\infty, \infty)$ and eliminating those μ_0 for which $H_0 : \mu = \mu_0$ is rejected. To see where this leads, let us examine our decision criterion in the case that σ is known: we reject $H_0 : \mu = \mu_0$ if and only if

$$\mathbf{p} = P_{\mu_0}\left(\left|\bar{X}_n - \mu_0\right| \geq |\bar{x}_n - \mu_0|\right) \doteq 2\Phi\left(-|z|\right) \leq \alpha, \tag{9.4}$$

where $z = (\bar{x}_n - \mu_0)/(\sigma/\sqrt{n})$. Using the symmetry of the normal distribution, we can rewrite condition (9.4) as

$$\alpha/2 \geq \Phi\left(-|z|\right) = P\left(Z < -|z|\right) = P\left(Z > |z|\right),$$

which in turn is equivalent to the condition

$$\Phi\left(|z|\right) = P\left(Z < |z|\right) = 1 - P\left(Z > |z|\right) \geq 1 - \alpha/2, \tag{9.5}$$

where $Z \sim \text{Normal}(0, 1)$.

Now let q denote the $1 - \alpha/2$ quantile of Normal$(0, 1)$, so that

$$\Phi(q) = 1 - \alpha/2.$$

Then condition (9.5) obtains if and only if $|z| \geq q$. We express this by saying that q is the *critical value* of the test statistic $|Z_n|$, where $Z_n = (\bar{X}_n - \mu_0)/(\sigma/\sqrt{n})$. For example, suppose that $\alpha = 0.05$, so that $1 - \alpha/2 = 0.975$. Then the critical value is computed in `R` as follows:

```
> qnorm(.975)
[1] 1.959964
```

Given a significance level α and the corresponding q, we have determined that q is the critical value of $|Z_n|$ for testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ at significance level α. Thus, we reject $H_0 : \mu = \mu_0$ if and only if (iff)

$$\begin{aligned} & \left|\frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}}\right| = |z| \geq q \\ \text{iff} \quad & |\bar{x}_n - \mu_0| \geq q\sigma/\sqrt{n} \\ \text{iff} \quad & \mu_0 \notin \left(\bar{x}_n - q\sigma/\sqrt{n},\ \bar{x}_n + q\sigma/\sqrt{n}\right). \end{aligned}$$

Thus, the desired set of plausible values is the interval

$$\left(\bar{x}_n - q\frac{\sigma}{\sqrt{n}},\ \bar{x}_n + q\frac{\sigma}{\sqrt{n}}\right). \tag{9.6}$$

If σ is unknown, then the argument is identical except that we estimate σ^2 as

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2,$$

obtaining as the set of plausible values the interval

$$\left(\bar{x}_n - q\frac{s_n}{\sqrt{n}},\ \bar{x}_n + q\frac{s_n}{\sqrt{n}}\right). \tag{9.7}$$

Example 9.2 (continued) *A random sample of $n = 400$ observations is drawn from a population with unknown mean μ and unknown variance σ^2, resulting in $\bar{x}_n = 21.82935$ and $s_n = 24.70037$. Using a significance level of $\alpha = 0.05$, determine a set of plausible values of μ.*

First, because $\alpha = 0.05$ is the significance level, $q = 1.959964$ is the critical value. From (9.7), an interval of plausible values is

$$21.82935 \pm 1.959964 \cdot 24.70037/\sqrt{400} = (19.40876, 24.24994).$$

Notice that $20 \in (19.40876, 24.24994)$, meaning that (as we discovered in Section 9.4) we would accept $H_0 : \mu = 20$ at significance level $\alpha = 0.05$.

Now consider the random interval I, defined in Case 1 (population variance known) by

$$I = \left(\bar{X}_n - q\frac{\sigma}{\sqrt{n}},\ \bar{X}_n - q\frac{\sigma}{\sqrt{n}}\right)$$

and in Case 2 (population variance unknown) by

$$I = \left(\bar{X}_n - q\frac{S_n}{\sqrt{n}},\ \bar{X}_n - q\frac{S_n}{\sqrt{n}}\right).$$

The probability that this random interval covers the real number μ_0 is

$$P_\mu\left(I \supset \mu_0\right) = 1 - P_\mu\left(\mu_0 \notin I\right) = 1 - P_\mu\left(\text{reject } H_0 : \mu = \mu_0\right).$$

If $\mu = \mu_0$, then the probability of coverage is

$$1 - P_{\mu_0}\left(\text{reject } H_0 : \mu = \mu_0\right) = 1 - P_{\mu_0}(\text{Type I error}) \geq 1 - \alpha.$$

Thus, the probability that I covers the true value of the population mean is at least $1-\alpha$, which we express by saying that I is a $(1-\alpha)$-*level confidence interval* for μ. The level of confidence, $1-\alpha$, is also called the *confidence coefficient*.

We emphasize that the confidence interval I is random and the population mean μ is fixed, albeit unknown. Each time that the experiment in question is performed, a random sample is observed and an interval is constructed from it. As the sample varies, so does the interval. Any one such interval, constructed from a single sample, either does or does not contain the population mean. However, if this procedure is repeated a great many times, then the proportion of such intervals that contain μ will be at least $1-\alpha$. Actually observing one sample and constructing one interval from it amounts to randomly selecting one of the many intervals that might or might not contain μ. Because most (at least $1-\alpha$) of the intervals do, we can be "confident" that the interval that was actually constructed does contain the unknown population mean.

9.5.1 Sample Size

Confidence intervals are often used to determine sample sizes for future experiments. Typically, the researcher specifies a desired confidence level, $1-\alpha$, and a desired interval length, L. After determining the appropriate critical value, q, one equates L with $2q\sigma/\sqrt{n}$ and solves for n, obtaining

$$n = (2q\sigma/L)^2 . \tag{9.8}$$

Of course, this formula presupposes knowledge of the population variance. In practice, it is usually necessary to replace σ with an estimate—which may be easier said than done if the experiment has not yet been performed. This is one reason to perform a pilot study: to obtain a preliminary estimate of the population variance and use it to design a better study.

Several useful relations can be deduced from equation (9.8):

1. Higher levels of confidence $(1-\alpha)$ correspond to larger critical values (q), which result in larger sample sizes (n).
2. Smaller interval lengths (L) result in larger sample sizes (n).
3. Larger variances (σ^2) result in larger sample sizes (n).

In summary, if a researcher desires high confidence that the true mean of a highly variable population is covered by a small interval, then s/he should plan on collecting a great deal of data!

Example 9.5 (continued) *A rival corporation purchases the rights to the amateur mechanic's additive. How large a study is required to determine this additive's mean increase in mileage to within* 0.05 *mpg with a confidence coefficient of* $1 - \alpha = 0.99$*?*

The desired interval length is $L = 2 \cdot 0.05 = 0.1$ and the critical value that corresponds to $\alpha = 0.01$ is computed in `R` as follows:

```
> qnorm(1-.01/2)
[1] 2.575829
```

From the mechanic's small pilot study, we estimate σ to be $s = 0.4$. Then

$$n = (2 \cdot 2.575829 \cdot 0.4/0.1)^2 \doteq 424.6,$$

so the desired study will require $n = 425$ vehicles.

9.5.2 One-Sided Confidence Intervals

The set of μ_0 for which we would accept the null hypothesis $H_0 : \mu = \mu_0$ when tested against the two-sided alternative hypothesis $H_1 : \mu \neq \mu_0$ is a traditional, 2-sided confidence interval. In situations where 1-sided alternatives are appropriate, we can construct corresponding 1-sided confidence intervals by determining the set of μ_0 for which the appropriate null hypothesis would be accepted.

Example 9.5 (continued) The government test has a significance level of $\alpha = 0.05$. It rejects the null hypothesis $H_0 : \mu \leq \mu_0$ if and only if (iff)

$$\begin{aligned} & \mathbf{p} = P(Z \geq t) \leq 0.05 \\ \text{iff} \quad & P(Z < t) \geq 0.95 \\ \text{iff} \quad & t \geq \texttt{qnorm(0.95)} \doteq 1.645. \end{aligned}$$

Equivalently, the null hypothesis $H_0 : \mu \leq \mu_0$ is accepted if and only if

$$\begin{aligned} & t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} < 1.645 \\ \text{iff} \quad & \bar{x} < \mu_0 + 1.645 \cdot \frac{s}{\sqrt{n}} \\ \text{iff} \quad & \mu_0 > \bar{x} - 1.645 \cdot \frac{s}{\sqrt{n}}. \end{aligned}$$

1. In the case of the large corporation, the null hypothesis $H_0 : \mu \leq \mu_0$ is accepted if and only if

$$\mu_0 > 1.01 - 1.645 \cdot \frac{0.1}{\sqrt{900}} \doteq 1.0045,$$

so the 1-sided confidence interval with confidence coefficient $1 - \alpha = 0.95$ is $(1.0045, \infty)$.

2. In the case of the amateur mechanic, the null hypothesis $H_0 : \mu \leq \mu_0$ is accepted if and only if

$$\mu_0 > 1.21 - 1.645 \cdot \frac{0.4}{\sqrt{9}} \doteq 0.9967,$$

so the 1-sided confidence interval with confidence coefficient $1 - \alpha = 0.95$ is $(0.9967, \infty)$.

9.6 Exercises

1. According to *The Justice Project*, "John Spirko was sentenced to death on the testimony of a witness who was '70 percent certain' of his identification." Formulate this case as a problem in hypothesis testing. What can be deduced about the significance level used to convict Spirko? Does this choice of significance level strike you as suitable for a capital murder trial?

2. Blaise Pascal, the French theologian and mathematician, argued that we cannot know whether or not God exists, but that we must behave as though we do. He submitted that the consequences of wrongly behaving as though God does not exist are greater than the consequences of wrongly behaving as though God does exist, concluding that it is better to err on the side of caution and act as though God exists. This argument is known as Pascal's Wager. Formulate Pascal's Wager as a hypothesis testing problem. What are the Type I and Type II errors? On whom did Pascal place the burden of proof, believers or nonbelievers?

3. Dorothy owns a lovely glass dreidl. Curious as to whether or not it is fairly balanced, she spins her dreidl ten times, observing five gimels and five hehs. Surprised by these results, Dorothy decides to compute

the probability that a fair dreidl would produce such aberrant results. Which of the probabilities specified in Exercise 3.6.5 is the most appropriate choice of a significance probability for this investigation? Why?

4. It is thought that human influenza viruses originate in birds. It is quite possible that, several years ago, a human influenza pandemic was averted by slaughtering 1.5 million chickens brought to market in Hong Kong. Because it is impossible to test each chicken individually, such decisions are based on samples. Suppose that a boy has already died of a bird flu virus apparently contracted from a chicken. Several diseased chickens have already been identified. The health officials would prefer to err on the side of caution and destroy all chickens that might be infected; the farmers do not want this to happen unless it is absolutely necessary. Suppose that both the farmers and the health officals agree that all chickens should be destroyed if more than 2 percent of them are diseased. A random sample of $n = 1000$ chickens reveals 40 diseased chickens.

   (a) Let $X_i = 1$ if chicken i is diseased and $X_i = 0$ if it is not. Assume that $X_1, \ldots, X_n \sim P$. To what family of probability distributions does P belong? What population parameter indexes this family? Use this parameter to state formulas for $\mu = EX_i$ and $\sigma^2 = \operatorname{Var} X_i$.

   (b) State appropriate null and alternative hypotheses from the perspective of the health officials.

   (c) State appropriate null and alternative hypotheses from the perspective of the farmers.

   (d) Use the value of μ_0 in the above hypotheses to compute the value of σ^2 under H_0. Then compute the value of the test statistic z.

   (e) Adopting the health officials' perspective, and assuming that they are willing to risk a 0.1% chance of committing a Type I error, what action should be taken? Why?

   (f) Adopting the farmers' perspective, and assuming that they are willing to risk a 10% chance of committing a Type I error, what action should be taken? Why?

5. A company that manufactures light bulbs has advertised that its 75-watt bulbs burn an average of 800 hours before failing. In reaction

to the company's advertising campaign, several dissatisfied customers have complained to a consumer watchdog organization that they believe the company's claim to be exaggerated. The consumer organization must decide whether or not to allocate some of its financial resources to countering the company's advertising campaign. So that it can make an informed decision, it begins by purchasing and testing 100 of the disputed light bulbs. In this experiment, the 100 light bulbs burned an average of $\bar{x} = 745.1$ hours before failing, with a sample standard deviation of $s = 238.0$ hours. Formulate null and alternative hypotheses that are appropriate for this situation. Calculate a significance probability. Do these results warrant rejecting the null hypothesis at a significance level of $\alpha = 0.05$?

6. To study the effects of Alzheimer's disease (AD) on cognition, a scientist administers two batteries of neuropsychological tasks to 60 mildly demented AD patients. One battery is administered in the morning, the other in the afternoon. Each battery includes a task in which discourse is elicited by showing the patient a picture and asking the patient to describe it. The quality of the discourse is measured by counting the number of "information units" conveyed by the patient. The scientist wonders if asking a patient to describe Picture A in the morning is equivalent to asking the same patient to describe Picture B in the afternoon, after having described Picture A several hours earlier. To investigate, she computes the number of information units for Picture A minus the number of information units for Picture B for each patient. She finds an average difference of $\bar{x} = -0.1833$, with a sample standard deviation of $s = 5.18633$. Formulate null and alternative hypotheses that are appropriate for this situation. Calculate a significance probability. Do these results warrant rejecting the null hypothesis at a significance level of $\alpha = 0.05$?

7. Each student in a large statistics class of 600 students is asked to toss a fair coin 100 times, count the resulting number of `Heads`, and construct a 0.95-level confidence interval for the probability of `Heads`. Assume that each student uses a fair coin and constructs the confidence interval correctly. True of False: *We would expect approximately* 570 *of the confidence intervals to contain the number* 0.5.

8. The USGS decides to use a laser altimeter to measure the height μ of Mt. Wrightson, the highest point in Pima County, Arizona. It is

known that measurements made by the laser altimeter have an expected value equal to μ and a standard deviation of 1 meter. How many measurements should be made if the USGS wants to construct a 0.90-level confidence interval for μ that has a length of 20 centimeters?

9. Professor Johnson is interested in the probability that a certain type of randomly generated matrix has a positive determinant. His student attempts to calculate the probability exactly, but runs into difficulty because the problem requires her to evaluate an integral in 9 dimensions. Professor Johnson therefore decides to obtain an approximate probability by simulation, i.e., by randomly generating some matrices and observing the proportion that have positive determinants. His preliminary investigation reveals that the probability is roughly 0.05. At this point, Professor Park decides to undertake a more comprehensive simulation experiment that will, with 0.95-level confidence, correctly determine the probability of interest to within ± 0.00001. How many random matrices should he generate to achieve the desired accuracy?

10. In September 2003, Lena spun a penny 89 times and observed 2 `Heads`. Let p denote the true probability that one spin of her penny will result in `Heads`.

    (a) The significance probability for testing $H_0 : p \geq 0.3$ versus $H_1 : p < 0.3$ is $\mathbf{p} = P(Y \leq 2)$, where $Y \sim \text{Binomial}(89; 0.3)$.

        i. Compute $\mathbf{p}$ as in Section 9.1, using the binomial distribution and `pbinom`.

        ii. Approximate $\mathbf{p}$ as in Section 9.4, using the normal distribution and `pnorm`. How good is this approximation?

    (b) Construct a 1-sided confidence interval for p by determining for which values of p_0 the null hypothesis $H_0 : p \geq p_0$ would be accepted at a significance level of (approximately) $\alpha = 0.05$.

Chapter 10

1-Sample Location Problems

The basic ideas associated with statistical inference were introduced in Chapter 9. We developed these ideas in the context of drawing inferences about a single population mean, and we assumed that the sample was large enough to justify appeals to the Central Limit Theorem for normal approximations. The population mean is a natural measure of centrality, but it is not the only one. Furthermore, even if we are interested in the population mean, our sample may be too small to justify the use of a large-sample normal approximation. The purpose of the next several chapters is to explore more thoroughly how statisticians draw inferences about measures of centrality.

Measures of centrality are sometimes called location parameters. The title of this chapter indicates an interest in a location parameter of a *single* population. More specifically, we assume that $X_1, \ldots, X_n \sim P$ are independently and identically distributed, we observe a random sample $\vec{x} = \{x_1, \ldots, x_n\}$, and we attempt to draw an inference about a location parameter of P. Because it is not always easy to identify the relevant population in a particular experiment, we begin with some examples. Our analysis of these examples is clarified by posing the following four questions:

1. What are the experimental units, i.e., what are the objects that are being measured?

2. From what population (or populations) were the experimental units drawn?

3. What measurements were taken on each experimental unit?

4. What random variables are relevant to the specified inference?

For the sake of specificity, we assume that the location parameter of interest in the following examples is the population median, $q_2(P)$.

Example 10.1 A machine is supposed to produce ball bearings that are 1 millimeter in diameter. To determine if the machine was correctly calibrated, a sample of ball bearings is drawn and the diameter of each ball bearing is measured. For this experiment:

1. An experimental unit is a ball bearing. Notice that we are distinguishing between experimental units, the objects being measured (ball bearings), and units of measurement (e.g., millimeters).

2. There is one population, viz., all ball bearings that might be produced by the designated machine.

3. One measurement (diameter) is taken on each experimental unit.

4. Let X_i denote the diameter of ball bearing i. Then $X_1, \ldots, X_n \sim P$ and we are interested in drawing inferences about $q_2(P)$, the population median diameter. For example, we might test $H_0 : q_2(P) = 1$ against $H_1 : q_2(P) \neq 1$.

Example 10.2 A drug is supposed to lower blood pressure. To determine if it does, a sample of hypertensive patients are administered the drug for two months. Each person's blood pressure is measured before and after the two month period. For this experiment:

1. An experimental unit is a patient.

2. There is one population of hypertensive patients. (It may be difficult to discern the precise population that was actually sampled. All hypertensive patients? All Hispanic male hypertensive patients who live in Houston, TX? All Hispanic male hypertensive patients who live in Houston, TX, and who are sufficiently well-disposed to the medical establishment to participate in the study? In published journal articles, scientists are often rather vague about just what population was actually sampled.)

3. Two measurements (blood pressure before and after treatment) are taken on each experimental unit. Let B_i and A_i denote the blood pressures of patient i before and after treatment.

4. Let $X_i = B_i - A_i$, the decrease in blood pressure for patient i. Then $X_1, \ldots, X_n \sim P$ and we are interested in drawing inferences about $q_2(P)$, the population median decrease. For example, we might test $H_0 : q_2(P) \leq 0$ against $H_1 : q_2(P) > 0$.

Example 10.3 A graduate student investigated the effect of Parkinson's disease (PD) on speech breathing. She recruited 16 PD patients to participate in her study. She also recruited 16 normal control (NC) subjects. Each NC subject was carefully matched to one PD patient with respect to sex, age, height, and weight. The lung volume of each study participant was measured. For this experiment:

1. An experimental unit was a matched PD-NC pair.

2. The population comprises all possible PD-NC pairs that satisfy the study criteria.

3. Two measurements (PD and NC lung volume) were taken on each experimental unit. Let D_i and C_i denote the PD and NC lung volumes of pair i.

4. Let $X_i = \log(D_i/C_i) = \log D_i - \log C_i$, the logarithm of the PD proportion of NC lung volume. (This is not the only way of comparing D_i and C_i, but it worked well in this investigation. Ratios can be difficult to analyze and logarithms convert ratios to differences. Furthermore, lung volume data tend to be skewed to the right. As in Exercise 2 of Section 7.5, logarithmic transformations of such data often have a symmetrizing effect.) Then $X_1, \ldots, X_n \sim P$ and we are interested in drawing inferences about $q_2(P)$. For example, to test the theory that PD restricts lung volume, we might test $H_0 : q_2(P) \geq 0$ against $H_1 : q_2(P) < 0$.

This chapter is divided into sections according to distributional assumptions about the X_i:

10.1 If the data are assumed to be normally distributed, then we will be interested in inferences about the population's center of symmetry, which we will identify as the population mean.

10.3 If the data are only assumed to be symmetrically distributed, then we will also be interested in inferences about the population's center of symmetry, but we will identify it as the population median.

10.2 If the data are only assumed to be continuously distributed, then we will be interested in inferences about the population median.

Each section is subdivided into subsections, according to the type of inference (point estimation, hypothesis testing, set estimation) at issue.

10.1 The Normal 1-Sample Location Problem

In this section we assume that $P = \text{Normal}(\mu, \sigma^2)$. As necessary, we will distinguish between cases in which σ is known and cases in which σ is unknown.

10.1.1 Point Estimation

Because normal distributions are symmetric, the location parameter μ is the center of symmetry and therefore both the population mean and the population median. Hence, there are (at least) two natural estimators of μ, the sample mean $\bar{X}_n$ and the sample median $q_2(\hat{P}_n)$. Both are consistent, unbiased estimators of μ. We will compare them by considering their *asymptotic relative efficiency* (ARE). A rigorous definition of ARE is beyond the scope of this book, but the concept is easily interpreted.

If the true distribution is $P = N(\mu, \sigma^2)$, then the ARE of the sample median to the sample mean for estimating μ is

$$e(P) = \frac{2}{\pi} \doteq 0.64.$$

This statement has the following interpretation: for large samples, using the sample median to estimate a normal population mean is equivalent to randomly discarding approximately 36% of the observations and calculating the sample mean of the remaining 64%. Thus, the sample mean is substantially more efficient than is the sample median at extracting location information from a normal sample.

In fact, if $P = \text{Normal}(\mu, \sigma^2)$, then the ARE of *any* estimator of μ to the sample mean is ≤ 1. This is sometimes expressed by saying that the sample mean is *asymptotically efficient* for estimating a normal mean. The sample mean also enjoys a number of other optimal properties in this case. The sample mean is unquestionably the preferred estimator for the normal 1-sample location problem.

10.1.2 Hypothesis Testing

If σ is known, then the possible distributions of X_i are

$$\left\{\text{Normal}(\mu, \sigma^2) : -\infty < \mu < \infty\right\}.$$

If σ is unknown, then the possible distributions of X_i are

$$\left\{\text{Normal}(\mu, \sigma^2) : -\infty < \mu < \infty, \sigma > 0\right\}.$$

We partition the possible distributions into two subsets, the null and alternative hypotheses. For example, if σ is known then we might specify

$$H_0 = \left\{\text{Normal}(0, \sigma^2)\right\} \quad \text{and} \quad H_1 = \left\{\text{Normal}(\mu, \sigma^2) : \mu \neq 0\right\},$$

which we would typically abbreviate as $H_0 : \mu = 0$ and $H_1 : \mu \neq 0$. Analogously, if σ is unknown then we might specify

$$H_0 = \left\{\text{Normal}(0, \sigma^2) : \sigma > 0\right\}$$

and

$$H_1 = \left\{\text{Normal}(\mu, \sigma^2) : \mu \neq 0, \sigma > 0\right\},$$

which we would also abbreviate as $H_0 : \mu = 0$ and $H_1 : \mu \neq 0$.

More generally, for any real number μ_0 we might specify

$$H_0 = \left\{\text{Normal}(\mu_0, \sigma^2)\right\} \quad \text{and} \quad H_1 = \left\{\text{Normal}(\mu, \sigma^2) : \mu \neq \mu_0\right\}$$

if σ is known, or

$$H_0 = \left\{\text{Normal}(\mu_0, \sigma^2) : \sigma > 0\right\}$$

and

$$H_1 = \left\{\text{Normal}(\mu, \sigma^2) : \mu \neq \mu_0, \sigma > 0\right\}$$

if σ in unknown. In both cases, we would typically abbreviate these hypotheses as $H_0 : \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$.

The preceding examples involve two-sided alternative hypotheses. Of course, as in Section 9.4, we might also specify one-sided hypotheses. However, the material in the present section is so similar to the material in Section 9.4 that we will only discuss two-sided hypotheses.

The intuition that underlies testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ was discussed in Section 9.4:

- If H_0 is true, then we would expect the sample mean to be close to the population mean μ_0.
- Hence, if $\bar{X}_n = \bar{x}_n$ is observed far from μ_n, then we are inclined to reject H_0.

To make this reasoning precise, we reject H_0 if and only if the significance probability

$$\mathbf{p} = P_{\mu_0}\left(|\bar{X}_n - \mu_0| \geq |\bar{x}_n - \mu_0|\right) \leq \alpha. \tag{10.1}$$

The first equation in (10.1) is a formula for a significance probability. Notice that this formula is identical to equation (9.2). The one difference between the material in Section 9.4 and the present material lies in how one computes **p**. For emphasis, we recall the following:

1. The hypothesized mean μ_0 is a fixed number specified by the null hypothesis.
2. The estimated mean, $\bar{x}_n$, is a fixed number computed from the sample. Therefore, so is $|\bar{x}_n - \mu_0|$, the difference between the estimated mean and the hypothesized mean.
3. The estimator, $\bar{X}_n$, is a random variable.
4. The subscript in P_{μ_0} reminds us to compute the probability under $H_0 : \mu = \mu_0$.
5. The significance level α is a fixed number specified by the researcher, preferably before the experiment was performed.

To apply (10.1), we must compute **p**. In Section 9.4, we overcame that technical difficulty by appealing to the Central Limit Theorem. This allowed us to approximate **p** even when we did not know the distribution of the X_i, but only for reasonably large sample sizes. However, if we know that $X_1, \ldots, X_n$ are normally distributed, then it turns out that we can calculate **p** exactly, even when n is small.

Case 1: The Population Variance is Known

Under the null hypothesis that $\mu = \mu_0$, $X_1, \ldots, X_n \sim \text{Normal}(\mu_0, \sigma^2)$ and

$$\bar{X}_n \sim \text{Normal}\left(\mu_0, \frac{\sigma^2}{n}\right).$$

This is the exact distribution of $\bar{X}_n$, not an asymptotic approximation. We convert $\bar{X}_n$ to standard units, obtaining

$$Z = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \sim \text{Normal}\,(0,1)\,. \tag{10.2}$$

The observed value of Z is

$$z = \frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}}.$$

The significance probability is

$$\begin{aligned} \mathbf{p} &= P_{\mu_0}\left(|\bar{X}_n - \mu_0| \geq |\bar{x}_n - \mu_0|\right) \\ &= P_{\mu_0}\left(\left|\frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}}\right| \geq \left|\frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}}\right|\right) \\ &= P\left(|Z| \geq |z|\right) \\ &= 2P\left(Z \geq |z|\right). \end{aligned}$$

In this case, the test that rejects H_0 if and only if $\mathbf{p} \leq \alpha$ is sometimes called the 1-*sample z-test*. The random variable Z is the *test statistic*.

Before considering the case of an unknown population variance, we remark that it is possible to derive point estimators from hypothesis tests. For testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$, the test statistics are

$$Z(\mu_0) = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}}.$$

If we observe $\bar{X}_n = \bar{x}_n$, then what value of μ_0 minimizes $|z(\mu_0)|$? Clearly, the answer is $\mu_0 = \bar{x}_n$. Thus, our preferred point estimate of μ is the μ_0 for which it is most difficult to reject $H_0 : \mu = \mu_0$. This type of reasoning will be extremely useful for analyzing situations in which we know how to test but don't know how to estimate.

Case 2: The Population Variance is Unknown

Statement (10.2) remains true if σ is unknown, but it is no longer possible to compute z. Therefore, we require a different test statistic for this case. A natural approach is to modify Z by replacing the unknown σ with an estimator of it. Toward that end, we introduce the test statistic

$$T_n = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}},$$

where S_n^2 is the unbiased estimator of the population variance defined by equation (9.1). Because T_n and Z are different random variables, they have different probability distributions and our first order of business is to determine the distribution of T_n.

We begin by stating a useful fact:

Theorem 10.1 *If $X_1, \ldots, X_n \sim$ Normal(μ, σ^2), then*

$$\frac{(n-1)S_n^2}{\sigma^2} = \sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2 / \sigma^2 \sim \chi^2(n-1).$$

The χ^2 (chi-squared) distribution was described in Section 5.5 and Theorem 10.1 is closely related to Theorem 5.3.

Next we write

$$\begin{aligned} T_n &= \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}} = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \cdot \frac{\sigma/\sqrt{n}}{S_n/\sqrt{n}} \\ &= Z \cdot \frac{\sigma}{S_n} = Z/\sqrt{S_n^2/\sigma^2} \\ &= Z/\sqrt{\left[(n-1)S_n^2/\sigma^2\right]/(n-1)}. \end{aligned}$$

Using Theorem 10.1, we see that T_n can be written in the form

$$T_n = \frac{Z}{\sqrt{Y/\nu}},$$

where $Z \sim \text{Normal}(0,1)$ and $Y \sim \chi^2(\nu)$. If Z and Y are independent random variables, then it follows from Definition 5.7 that $T_n \sim t(n-1)$.

Both Z and $Y = (n-1)S_n^2/\sigma^2$ depend on $X_1, \ldots, X_n$, so one would be inclined to think that Z and Y are dependent. This is usually the case, but it turns out that they are independent if $X_1, \ldots, X_n \sim \text{Normal}(\mu, \sigma^2)$. This is another remarkable property of normal distributions, usually stated as follows:

Theorem 10.2 *If $X_1, \ldots, X_n \sim$ Normal(μ, σ^2), then $\bar{X}_n$ and S_n^2 are independent random variables.*

The result that interests us can then be summarized as follows:

Corollary 10.1 *If $X_1, \ldots, X_n \sim$ Normal(μ_0, σ^2), then*

$$T_n = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}} \sim t(n-1).$$

Now let

$$t_n = \frac{\bar{x}_n - \mu_0}{s_n/\sqrt{n}},$$

the observed value of the test statistic T_n. The significance probability is

$$\mathbf{p} = P_{\mu_0}\left(|T_n| \geq |t_n|\right) = 2P_{\mu_0}\left(T_n \geq |t_n|\right).$$

In this case, the test that rejects H_0 if and only if $\mathbf{p} \leq \alpha$ is called *Student's 1-sample t-test*. Because it is rarely the case that the population variance is known when the population mean is not, Student's 1-sample t-test is used much more frequently than the 1-sample z-test. We will use the `R` function `pt` to compute significance probabilities for Student's 1-sample t-test, as illustrated in the following examples.

Example 10.4 Suppose that, to test $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$ (a 2-sided alternative), we draw a sample of size $n = 25$ and observe $\bar{x} = 1$ and $s = 3$. Then $t = (1-0)/(3/\sqrt{25}) \doteq 1.67$ and the 2-tailed significance probability is computed using both tails of the $t(24)$ distribution, i.e., $\mathbf{p} = 2 * \texttt{pt}(-1.67, \texttt{df} = 24) \doteq 0.054$.

Example 10.5 Suppose that, to test $H_0 : \mu \leq 0$ versus $H_1 : \mu > 0$ (a 1-sided alternative), we draw a sample of size $n = 25$ and observe $\bar{x} = 2$ and $s = 5$. Then $t = (2-0)/(5/\sqrt{25}) = 2.00$ and the 1-tailed significance probability is computed using one tail of the $t(24)$ distribution, i.e., $\mathbf{p} = 1 - \texttt{pt}(2.00, \texttt{df} = 24) \doteq 0.028$.

10.1.3 Interval Estimation

As in Section 9.5, we will derive confidence intervals from tests. We imagine testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ for every $\mu_0 \in (-\infty, \infty)$. The μ_0 for which $H_0 : \mu = \mu_0$ is rejected are implausible values of μ; the μ_0 for which $H_0 : \mu = \mu_0$ is accepted constitute the confidence interval. To accomplish this, we will have to derive the critical values of our tests. A significance level of α will result in a confidence coefficient of $1 - \alpha$.

Case 1: The Population Variance is Known

If σ is known, then we reject $H_0 : \mu = \mu_0$ if and only if

$$\mathbf{p} = P_{\mu_0}\left(\left|\bar{X}_n - \mu_0\right| \geq |\bar{x}_n - \mu_0|\right) = 2\Phi\left(-|z_n|\right) \leq \alpha,$$

where $z_n = (\bar{x}_n - \mu_0)/(\sigma/\sqrt{n})$. By the symmetry of the normal distribution, this condition is equivalent to the condition

$$1 - \Phi(-|z_n|) = P(Z > -|z_n|) = P(Z < |z_n|) = \Phi(|z_n|) \geq 1 - \alpha/2,$$

where $Z \sim \text{Normal}(0, 1)$, and therefore to the condition $|z_n| \geq q_z$, where q_z denotes the $1 - \alpha/2$ quantile of Normal$(0, 1)$. The quantile q_z is the critical value of the two-sided 1-sample z-test. Thus, given a significance level α and a corresponding critical value q_z, we reject $H_0 : \mu = \mu_0$ if and only if (iff)

$$\begin{aligned} & \left| \frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}} \right| = |z_n| \geq q_z \\ \text{iff} \quad & |\bar{x}_n - \mu_0| \geq q_z \sigma/\sqrt{n} \\ \text{iff} \quad & \mu_0 \notin (\bar{x}_n - q_z\sigma/\sqrt{n},\ \bar{x}_n + q_z\sigma/\sqrt{n}) \end{aligned}$$

and we conclude that the desired set of plausible values is the interval

$$\left(\bar{x}_n - q_z \frac{\sigma}{\sqrt{n}},\ \bar{x}_n + q_z \frac{\sigma}{\sqrt{n}} \right).$$

Notice that both the preceding derivation and the resulting confidence interval are identical to the derivation and confidence interval in Section 9.5. The only difference is that, because we are now assuming that $X_1, \ldots, X_n \sim \text{Normal}(\mu, \sigma^2)$ instead of relying on the Central Limit Theorem, no approximation is required.

Example 10.6 Suppose that we desire 90% confidence about μ and $\sigma = 3$ is known. Then $\alpha = 0.10$ and $q_z \doteq 1.645$. Suppose that we draw $n = 25$ observations and observe $\bar{x}_n = 1$. Then

$$1 \pm 1.645 \frac{3}{\sqrt{25}} = 1 \pm 0.987 = (0.013, 1.987)$$

is a 0.90-level confidence interval for μ.

Case 2: The Population Variance is Unknown

If σ is unknown, then it must be estimated from the sample. The reasoning in this case is the same, except that we rely on Student's 1-sample t-test.

As before, we use S_n^2 to estimate σ^2. The critical value of the 2-sided 1-sample t-test is q_t, the $1-\alpha/2$ quantile of a t distribution with $n-1$ degrees of freedom, and the confidence interval is

$$\left(\bar{x}_n - q_t \frac{s_n}{\sqrt{n}},\ \bar{x}_n + q_t \frac{s_n}{\sqrt{n}} \right).$$

Example 10.7 Suppose that we desire 90% confidence about μ and σ is unknown. Suppose that we draw $n = 25$ observations and observe $\bar{x}_n = 1$ and $s = 3$. Then $q_t = \texttt{qt}(.95, \texttt{df} = 24) \doteq 1.711$ and

$$1 \pm 1.711 \times 3/\sqrt{25} = 1 \pm 1.027 = (-0.027, 2.027)$$

is a 90% confidence interval for μ. Notice that the confidence interval is larger when we use $s = 3$ instead of $\sigma = 3$.

10.2 The General 1-Sample Location Problem

In Section 10.1 we assumed that $X_1, \ldots, X_n \sim P$ and $P = \text{Normal}(\mu, \sigma^2)$. In this section, we again assume that $X_1, \ldots, X_n \sim P$, but now we assume only that the X_i are continuous random variables.

Because P is not assumed to be symmetric, we must decide which location parameter to study. The population median, $q_2(P)$, enjoys several advantages. Unlike the population mean, the population median always exists and is not sensitive to the influence of outliers. Furthermore, it turns out that one can develop fairly elementary ways to study medians, even when little is known about the probability distribution P. For simplicity, we will denote the population median by θ.

10.2.1 Hypothesis Testing

It is convenient to begin our study of the general 1-sample location problem with a discussion of hypothesis testing. As is Section 10.1, we initially consider testing a 2-sided alternative, $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. We will explicate a procedure known as the *sign test*.

The intuition that underlies the sign test is elementary. If the population median is $\theta = \theta_0$, then when we sample P we should observe roughly half the x_i above θ_0 and half the x_i below θ_0. Hence, if we observe proportions of x_i above/below θ_0 that are very different from one half, then we are inclined to reject the possibility that $\theta = \theta_0$.

More formally, let $p_+ = P_{H_0}(X_i > \theta_0)$ and $p_- = P_{H_0}(X_i < \theta_0)$. Because the X_i are continuous, $P_{H_0}(X_i = \theta_0) = 0$ and therefore $p_+ = p_- = 0.5$. Hence, under H_0, observing whether $X_i > \theta_0$ or $X_i < \theta_0$ is equivalent to tossing a fair coin, i.e., to observing a Bernoulli trial with success probability $p = 0.5$. The sign test is the following procedure:

1. Let $\vec{x} = \{x_1, \ldots, x_n\}$ denote the observed sample. If the X_i are continuous random variables, then $P(X_i = \theta_0) = 0$ and it should be that each $x_i \neq \theta_0$. In practice, of course, it may happen that we do observe one or more $x_i = \theta_0$. For the moment, we assume that $\vec{x}$ contains no such values.

2. Let
$$Y = \#\{X_i > \theta_0\} = \#\{X_i - \theta_0 > 0\}$$
be the test statistic. Under $H_0 : \theta = \theta_0$, $Y \sim \text{Binomial}(n; p = 0.5)$. The observed value of the test statistic is
$$y = \#\{x_i > \theta_0\} = \#\{x_i - \theta_0 > 0\}.$$

3. Notice that $EY = n/2$. The significance probability is
$$\mathbf{p} = P_{\theta_0}\left(\left|Y - \frac{n}{2}\right| \geq \left|y - \frac{n}{2}\right|\right).$$
The sign test rejects $H_0 : \theta = \theta_0$ if and only if $\mathbf{p} \leq \alpha$.

4. To compute $\mathbf{p}$, we first note that
$$\left|Y - \frac{n}{2}\right| \geq \left|y - \frac{n}{2}\right|$$
is equivalent to the event
   (a) $\{Y \leq y \text{ or } Y \geq n - y\}$ if $y \leq n/2$;
   (b) $\{Y \geq y \text{ or } Y \leq n - y\}$ if $y \geq n/2$.

   To accomodate both cases, let $c = \min(y, n - y)$. Then
$$\mathbf{p} = P_{\theta_0}(Y \leq c) + P_{\theta_0}(Y \geq c) = 2P_{\theta_0}(Y \leq c) = \texttt{2*pbinom(c,n,.5)}.$$

Example 10.8(a) Suppose that we want to test $H_0 : \theta = 100$ versus $H_1 : \theta \neq 100$ at significance level $\alpha = 0.05$, having observed the sample
$$\vec{x} = \{98.73, 97.17, 100.17, 101.26, 94.47, 96.39, 99.67, 97.77, 97.46, 97.41\}.$$
Here $n = 10$, $y = \#\{x_i > 100\} = 2$, and $c = \min(2, 10 - 2) = 2$, so
$$\mathbf{p} = \texttt{2*pbinom(2,10,.5)} = 0.109375 > 0.05$$
and we decline to reject H_0.

Example 10.8(b) Now suppose that we want to test $H_0 : \theta \leq 97$ versus $H_1 : \theta > 97$ at significance level $\alpha = 0.05$, using the same data. Here $n = 10$, $y = \#\{x_i > 97\} = 8$, and $c = \min(8, 10 - 8) = 2$. Because large values of Y are evidence against $H_0 : \theta \leq 97$,

$$\begin{aligned} \mathbf{p} &= P_{\theta_0}(Y \geq y) = P_{\theta_0}(Y \geq 8) = 1 - P_{\theta_0}(Y \leq 7) \\ &= \texttt{1-pbinom(7,10,.5)} = 0.0546875 > 0.05 \end{aligned}$$

and we decline to reject H_0.

Thus far we have assumed that the sample contains no values for which $x_i = \theta_0$. In practice, we may well observe such values. For example, if the measurements in Example 10.8(a) were made less precisely, then we might have observed the following sample:

$$\vec{x} = \{99, 97, 100, 101, 94, 96, 100, 98, 97, 97\}. \tag{10.3}$$

If we want to test $H_0 : \theta = 100$ versus $H_1 : \theta \neq 100$, then we have two values that equal θ_0 and the sign test requires modification.

We assume that $\#\{x_i = \theta_0\}$ is fairly small; otherwise, the assumption that the X_i are continuous is questionable. We consider two possible ways to proceed:

1. Perhaps the most satisfying solution is to compute all of the significance probabilities that correspond to different ways of counting the $x_i = \theta_0$ as larger or smaller than θ_0. If there are k observations $x_i = \theta_0$, then this will produce 2^k significance probabilities, which we might average to obtain a single $\mathbf{p}$.

2. Alternatively, let $\mathbf{p}_0$ denote the significance probability obtained by counting in the way that is most favorable to H_0 (least favorable to H_1). This is the largest of the possible significance probabilities, so if $\mathbf{p}_0 \leq \alpha$ then we reject H_0. Similarly, let $\mathbf{p}_1$ denote the significance probability obtained by counting in the way that is least favorable to H_0 (most favorable to H_1). This is the smallest of the possible significance probabilities, so if $\mathbf{p}_1 > \alpha$ then we decline to reject H_0. If $\mathbf{p}_0 > \alpha \geq \mathbf{p}_1$, then we simply declare the results to be equivocal.

Example 10.8(c) Suppose that we want to test $H_0 : \theta = 100$ versus $H_1 : \theta \neq 100$ at significance level $\alpha = 0.05$, having observed the sample

(10.3). Here $n = 10$ and $y = \#\{x_i > 100\}$ depends on how we count the observations $x_3 = x_7 = 100$. There are $2^2 = 4$ possibilities:

possibility	$y = \#\{x_i > 100\}$	$c = \min(y, 10 - y)$	$\mathbf{p}$
$y_3 < 100, y_7 < 100$	1	1	0.021484
$y_3 < 100, y_7 > 100$	2	2	0.109375
$y_3 > 100, y_7 < 100$	2	2	0.109375
$y_3 > 100, y_7 > 100$	3	3	0.343750

Noting that $\mathbf{p}_0 \doteq 0.344 > 0.05 > 0.021 \doteq \mathbf{p}_1$, we might declare the results to be equivocal. However, noting that 3 of the 4 possibilities lead us to accept H_0 (and that the average $\mathbf{p} \doteq 0.146$), we might conclude—somewhat more decisively—that there is insufficient evidence to reject H_0. The distinction between these two interpretations is largely rhetorical, as the fundamental logic of hypothesis testing requires that we decline to reject H_0 unless there is compelling evidence against it.

10.2.2 Point Estimation

Next we consider the problem of estimating the population median. A natural estimate is the plug-in estimate, the sample median. Another approach begins by posing the following question: For what value of θ_0 is the sign test least inclined to reject $H_0 : \theta = \theta_0$ in favor of $H_1 : \theta \neq \theta_0$? The answer to this question is also a natural estimate of the population median.

In fact, the plug-in and sign-test approaches lead to the same estimation procedure. To understand why, we focus on the case that n is even, in which case $n/2$ is a possible value of $Y = \#\{X_i > \theta_0\}$. If $|y - n/2| = 0$, then

$$\mathbf{p} = P\left(\left|Y - \frac{n}{2}\right| \geq 0\right) = 1.$$

We see that the sign test produces the maximal significance probability of $\mathbf{p} = 1$ when $y = n/2$, i.e., when θ_0 is chosen so that precisely half the observations exceed θ_0. This means that the sign test is least likely to reject $H_0 : \theta = \theta_0$ when θ_0 is the sample median. (A similar argument leads to the same conclusion when n is odd.)

Thus, using the sign test to test hypotheses about population medians corresponds to using the sample median to estimate population medians, just as using Student's t-test to test hypotheses about population means corresponds to using the sample mean to estimate population means. One

consequence of this remark is that, when the population mean and median are identical, the "Pitman efficiency" of the sign test to Student's t-test equals the asymptotic relative efficiency of the sample median to the sample median. For example, using the sign test on normal data is asymptotically equivalent to randomly discarding 36% of the observations, then using Student's t-test on the remaining 64%.

10.2.3 Interval Estimation

Finally, we consider the problem of constructing a $(1-\alpha)$-level confidence interval for the population median. Again we rely on the sign test, determining for which θ_0 the level-α sign test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ will accept H_0.

The sign test will reject $H_0 : \theta = \theta_0$ if and only if

$$y(\theta_0) = \#\{x_i > \theta_0\}$$

is either too large or too small. Equivalently, H_0 will be accepted if θ_0 is such that the numbers of observations above and below θ_0 are roughly equal.

To determine the critical value for the desired sign test, we suppose that $Y \sim \text{Binomial}(n; 0.5)$. We would like to find k such that $\alpha = 2P(Y \leq k)$, or $\alpha/2 = \texttt{pbinom}(k, n, 0.5)$. In practice, we won't be able to solve this equation exactly. We will use the `qbinom` function plus trial-and-error to solve it approximately, then modify our choice of α accordingly.

Having determined an acceptable (α, k), the sign test rejects $H_0 : \theta = \theta_0$ at level α if and only if either $y(\theta_0) \leq k$ or $y(\theta_0) \geq n - k$. We need to translate these inequalities into an interval of plausible values of θ_0. To do so, it is helpful to sort the values observed in the sample.

Definition 10.1 *The* order statistics *of* $\vec{x} = \{x_1, \ldots, x_n\}$ *are any permutation of the* x_i *such that*

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n-1)} \leq x_{(n)}.$$

If $\vec{x}$ *contains* n *distinct values, then there is a unique set of order statistics and the above inequalities are strict; otherwise, we say that* $\vec{x}$ *contains* ties.

Thus, $x_{(1)}$ is the smallest value in $\vec{x}$ and $x_{(n)}$ is the largest. If $n = 2m+1$ (n is odd), then the sample median is $x_{(m+1)}$; if $n = 2m$ (n is even), then the sample median is $[x_{(m)} + x_{(m+1)}]/2$.

For simplicity we assume that $\vec{x}$ contains no ties. If $\theta_0 < x_{(k+1)}$, then at least $n-k$ observations exceed θ_0 and the sign test rejects $H_0 : \theta = \theta_0$. Similarly, if $\theta_0 > x_{(k+1)}$, then no more than k observations exceed θ_0 and the sign test rejects $H_0 : \theta = \theta_0$. We conclude that the sign test accepts $H_0 : \theta = \theta_0$ if and only if θ_0 lies in the $(1-\alpha)$-level confidence interval

$$\left(x_{(k+1)}, x_{(n-k)}\right).$$

Example 10.8(d) Using the $n = 10$ observations from Example 10.8(a), we endeavor to construct a 0.90-level confidence interval for the population median. We begin by determining a suitable choice of (α, k). If $1-\alpha = 0.90$, then $\alpha/2 = 0.05$. `R` returns `qbinom(.05,10,.5)` $= 2$. Next we experiment:

k	`pbinom`$(k, 10, 0.5)$
2	0.0546875
1	0.01074219

We choose $k = 2$, resulting in a confidence level of

$$1 - \alpha = 1 - 2 \cdot 0.0546875 = 0.890625 \doteq 0.89,$$

nearly equal to the requested level of 0.90. Now, upon sorting the data (the `sort` function in `R` may be useful), we quickly discern that the desired confidence interval is

$$\left(x_{(3)}, x_{(8)}\right) = (97.17, 99.67).$$

10.3 The Symmetric 1-Sample Location Problem

10.4 A Case Study from Neuropsychology

10.5 Exercises

Problem Set A The following data are from Darwin (1876), *The Effect of Cross- and Self-Fertilization in the Vegetable Kingdom, Second Edition*, London: John Murray. Pairs of seedlings of the same age (one produced by cross-fertilization, the other by self-fertilization) were grown together so that the members of each pair were reared under nearly identical conditions. The aim was to demonstrate the greater vigour of the cross-fertilized plants. The data are the final heights (in inches) of each plant after a fixed period of time. Darwin consulted Francis Galton about the analysis of these data, and they were discussed further in Ronald Fisher's *Design of Experiments*.

Pair	Fertilized	
	Cross	Self
1	23.5	17.4
2	12.0	20.4
3	21.0	20.0
4	22.0	20.0
5	19.1	18.4
6	21.5	18.6
7	22.1	18.6
8	20.4	15.3
9	18.3	16.5
10	21.6	18.0
11	23.3	16.3
12	21.0	18.0
13	22.1	12.8
14	23.0	15.5
15	12.0	18.0

1. Show that this problem can be formulated as a 1-sample location problem. To do so, you should:

   (a) Identify the experimental units and the measurement(s) taken on each unit.

   (b) Define appropriate random variables $X_1, \ldots, X_n \sim P$. Remember that the statistical procedures that we will employ assume that these random variables are independent and identically distributed.

(c) Let θ denote the location parameter (measure of centrality) of interest. Depending on which statistical procedure we decide to use, either $\theta = EX_i = \mu$ or $\theta = q_2(X_i)$. State appropriate null and alternative hypotheses about θ.

2. Does it seem reasonable to assume that the sample $\vec{x} = (x_1, \ldots, x_n)$, the observed values of $X_1, \ldots, X_n$, were drawn from:

   (a) a normal distribution? Why or why not?

   (b) a symmetric distribution? Why or why not?

3. Assume that $X_1, \ldots, X_n$ are normally distributed and let $\theta = EX_i = \mu$.

   (a) Test the null hypothesis derived above using Student's 1-sample t-test. What is the significance probability? If we adopt a significance level of $\alpha = 0.05$, should we reject the null hypothesis?

   (b) Construct a (2-sided) confidence interval for θ with a confidence coefficient of approximately 0.90.

4. Now we drop the assumption of normality. Assume that $X_1, \ldots, X_n$ are symmetric (but not necessarily normal), continuous random variables and let $\theta = q_2(X_i)$.

   (a) Test the null hypothesis derived above using the Wilcoxon signed rank test. What is the significance probability? If we adopt a significance level of $\alpha = 0.05$, should we reject the null hypothesis?

   (b) Estimate θ by computing the median of the Walsh averages.

   (c) Construct a (2-sided) confidence interval for θ with a confidence coefficient of approximately 0.90.

5. Finally we drop the assumption of symmetry, assuming only that $X_1, \ldots, X_n$ are continuous random variables, and let $\theta = q_2(X_i)$.

   (a) Test the null hypothesis derived above using the sign test. What is the significance probability? If we adopt a significance level of $\alpha = 0.05$, should we reject the null hypothesis?

   (b) Estimate θ by computing the sample median.

   (c) Construct a (2-sided) confidence interval for θ with a confidence coefficient of approximately 0.90.

Problem Set B The ancient Greeks greatly admired rectangles with a height-to-width ratio of

$$1 : \frac{1+\sqrt{5}}{2} = 0.618034.$$

They called this number the "golden ratio" and used it repeatedly in their art and architecture, e.g. in building the Parthenon. Furthermore, golden rectangles are often found in the art of later western cultures.

A cultural anthropologist wondered if the Shoshoni, a native American civilization, also used golden rectangles. The following measurements, which appear as Data Set 150 in *A Handbook of Small Data Sets*, are height-to-width ratios of beaded rectangles used by the Shoshoni in decorating various leather goods:

0.693	0.662	0.690	0.606	0.570
0.749	0.672	0.628	0.609	0.844
0.654	0.615	0.668	0.601	0.576
0.670	0.606	0.611	0.553	0.933

We will analyze the Shoshoni rectangles as a 1-sample location problem.

1. There are two natural scales that we might use in analyzing these data. One possibility is to analyze the ratios themselves; the other is to analyze the (natural) logarithms of the ratios. For which of these possibilities would an assumption of normality seem more plausible? Please justify your answer.

2. Choose the possibility (ratios or logarithms of ratios) for which an assumption of normality seems more plausible. Formulate suitable null and alternative hypotheses for testing the possibility that the Shoshoni were using golden rectangles. Using Student's 1-sample t-test, compute a significance probability for testing these hypotheses. Would you reject or accept the null hypothesis using a significance level of 0.05?

3. Suppose that we are unwilling to assume that either the ratios or the log-ratios were drawn from a normal distribution. Use the sign test to construct a 0.90-level confidence interval for the population median of the ratios.

Problem Set C Researchers studied the effect of the drug caprotil on essential hypertension, reporting their findings in the *British Medical Journal*. They measured the supine systolic and diastolic blood pressures of 15

patients with moderate essential hypertension, immediately before and two hours after administering caprotil. The following measurements are Data Set 72 in *A Handbook of Small Data Sets*:

Patient	Systolic		Diastolic	
	before	after	before	after
1	210	201	130	125
2	169	165	122	121
3	187	166	124	121
4	160	157	104	106
5	167	147	112	101
6	176	145	101	85
7	185	168	121	98
8	206	180	124	105
9	173	147	115	103
10	146	136	102	98
11	174	151	98	90
12	201	168	119	98
13	198	179	106	110
14	148	129	107	103
15	154	131	100	82

We will consider the question of whether or not caprotil affects systolic and diastolic blood pressure differently.

1. Let SB and SA denote before and after systolic blood pressure; let DB and DB denote before and after diastolic blood pressure. There are several random variables that might be of interest:

$$X_i = (SB_i - SA_i) - (DB_i - DA_i) \tag{10.4}$$

$$X_i = \frac{SB_i - SA_i}{SB_i} - \frac{DB_i - DA_i}{DB_i} \tag{10.5}$$

$$X_i = \frac{SB_i - SA_i}{SB_i} \div \frac{DB_i - DA_i}{DB_i} \tag{10.6}$$

$$X_i = \log\left(\frac{SB_i - SA_i}{SB_i} \div \frac{DB_i - DA_i}{DB_i}\right) \tag{10.7}$$

   Suggest rationales for considering each of these possibilities.

2. Which (if any) of the above random variables appear to be normally distrbuted? Which appear to be symmetrically distrbuted?

3. Does caprotil affect systolic and diastolic blood pressure differently? Write a brief report that summarizes your investigation and presents your conclusion(s).

Problem Set D

1. A device counts the number of ions that arrive in a given time interval, unless too many arrive. An experiment that relies on this device produces the following counts, where `Big` means that the count exceeded 255.

251	238	249	`Big`	243	248	229	`Big`	235	244
254	251	252	244	230	222	224	246	`Big`	239

   Use these data to construct a 0.95-level confidence interval for the population median number of ions.

Chapter 11

2-Sample Location Problems

Thus far, in Chapters 9 and 10, we have studied inferences about a single population. In contrast, the present chapter is concerned with comparing *two* populations with respect to some measure of centrality, typically the population mean or the population median. Specifically, we assume the following:

1. $X_1, \ldots, X_{n_1} \sim P_1$ and $Y_1, \ldots, Y_{n_2} \sim P_2$ are continuous random variables. The X_i and the Y_j are mutually independent. In particular, there is no natural pairing of X_1 with Y_1, X_2 with Y_2, etc.

2. P_1 has location parameter θ_1 and P_2 has location parameter θ_2. We assume that comparisons of θ_1 and θ_2 are meaningful. For example, we might compare population means, $\theta_1 = \mu_1 = EX_i$ and $\theta_2 = \mu_2 = EY_j$, or population medians, $\theta_1 = q_2(X_i)$ and $\theta_2 = q_2(Y_j)$, but we would not compare the mean of one population and the median of another population. The *shift parameter*, $\Delta = \theta_1 - \theta_2$, measures the difference in population location.

3. We observe random samples $\vec{x} = \{x_1, \ldots, x_{n_1}\}$ and $\vec{y} = \{y_1, \ldots, y_{n_2}\}$, from which we attempt to draw inferences about Δ. Notice that we do *not* assume that $n_1 = n_2$.

The same four questions that we posed at the beginning of Chapter 10 can be asked here. What distinguishes 2-sample problems from 1-sample problems is the number of populations from which the experimental units were drawn. The prototypical case of a 2-sample problem is the case of a treatment population and a control population. We begin by considering some examples.

Example 11.1 A researcher investigated the effect of Alzheimer's disease (AD) on ability to perform a confrontation naming task. She recruited 60 mildly demented AD patients and 60 normal elderly control subjects. The control subjects resembled the AD patients in that the two groups had comparable mean ages, years of education, and (estimated) IQ scores; however, the control subjects were not individually matched to the AD patients. Each person was administered the Boston Naming Test (BNT), on which higher scores represent better performance. For this experiment:

1. An experimental unit is a person.

2. The experimental units belong to one of two populations: AD patients or normal elderly persons.

3. One measurement (score on BNT) is taken on each experimental unit.

4. Let X_i denote the BNT score for AD patient i. Let Y_j denote the BNT score for control subject j. Then $X_1, \ldots, X_{n_1} \sim P_1$, $Y_1, \ldots, Y_{n_2} \sim P_2$, and we are interested in drawing inferences about $\Delta = \theta_1 - \theta_2$. Notice that $\Delta < 0$ if and only if $\theta_1 < \theta_2$. Thus, to document that AD compromises confrontation naming ability, we might test $H_0 : \Delta \geq 0$ against $H_1 : \Delta < 0$.

Example 11.2 A drug is supposed to lower blood pressure. To determine if it does, $n_1 + n_2$ hypertensive patients are recruited to participate in a *double-blind* study. The patients are randomly assigned to a treatment group of n_1 patients and a control group of n_2 patients. Each patient in the treatment group receives the drug for two months; each patient in the control group receives a *placebo* for the same period. Each patient's blood pressure is measured before and after the two month period, and neither the patient nor the technician know to which group the patient was assigned. For this experiment:

1. An experimental unit is a patient.

2. The experimental units belong to one of two populations: hypertensive patients who receive the drug and hypertensive patients who receive the placebo. Notice that there are two populations despite the fact that all $n_1 + n_2$ patients were initially recruited from a single population. *Different treatment protocols create different populations.*

3. Two measurements (blood pressure before and after treatment) are taken on each experimental unit.

4. Let B_{1i} and A_{1i} denote the before and after blood pressures of patient i in the treatment group. Similarly, let B_{2j} and A_{2j} denote the before and after blood pressures of patient j in the control group. Let $X_i = B_{1i} - A_{1i}$, the decrease in blood pressure for patient i in the treatment group, and let $Y_j = B_{2j} - A_{2j}$, the decrease in blood pressure for patient j in the control group. Then $X_1, \ldots, X_{n_1} \sim P_1$, $Y_1, \ldots, Y_{n_2} \sim P_2$, and we are interested in drawing inferences about $\Delta = \theta_1 - \theta_2$. Notice that $\Delta > 0$ if and only if $\theta_1 > \theta_2$, i.e., if the decrease in blood pressure is greater for the treatment group than for the control group. Thus, a drug company required to produce compelling evidence of the drug's efficacy might test $H_0 : \Delta \leq 0$ against $H_1 : \Delta > 0$.

This chapter is divided into three sections:

11.1 If the data are assumed to be normally distributed, then we will be interested in inferences about the difference in population means. We will distinguish three cases, corresponding to what is known about the population variances.

11.2 If the data are only assumed to be continuously distributed, then we will be interested in inferences about the difference in population medians. We will assume a *shift model*, i.e., we will assume that P_1 and P_2 only differ with respect to location.

11.3 If the data are also assumed to be symmetrically distributed, then we will be interested in inferences about the difference in population centers of symmetry. If we assume symmetry, then we need not assume a shift model.

11.1 The Normal 2-Sample Location Problem

In this section we assume that

$$P_1 = \text{Normal}\left(\mu_1, \sigma_1^2\right) \quad \text{and} \quad P_2 = \text{Normal}\left(\mu_2, \sigma_2^2\right).$$

In describing inferential methods for $\Delta = \mu_1 - \mu_2$, we emphasize connections with material in Chapter 9 and Section 10.1. For example, the natural

estimator of a single normal population mean μ is the plug-in estimator $\hat{\mu}$, the sample mean, an unbiased, consistent, asymptotically efficient estimator of μ. In precise analogy, the natural estimator of $\Delta = \mu_1 - \mu_2$, the difference in populations means, is $\hat{\Delta} = \hat{\mu}_1 - \hat{\mu}_2 = \bar{X} - \bar{Y}$, the difference in sample means. Because

$$E\hat{\Delta} = E\bar{X} - E\bar{Y} = \mu_1 - \mu_2 = \Delta,$$

$\hat{\Delta}$ is an unbiased estimator of Δ. It is also consistent and asymptotically efficient.

In Chapter 9 and Section 10.1, hypothesis testing and set estimation for a single population mean were based on knowing the distribution of the standardized natural estimator, a random variable of the form

$$\frac{\begin{array}{c}\text{sample} \\ \text{mean}\end{array} \; - \; \begin{array}{c}\text{hypothesized} \\ \text{mean}\end{array}}{\begin{array}{c}\text{standard deviation} \\ \text{of sample mean}\end{array}}.$$

The denominator of this random variable, often called the *standard error*, was either known or estimated, depending on our knowledge of the population variance σ^2. For σ^2 known, we learned that

$$Z = \frac{\bar{X} - \mu_0}{\sqrt{\sigma^2/n}} \qquad \left\{ \begin{array}{lll} \sim & \text{Normal}(0,1) & \text{if} \quad X_1, \ldots, X_n \sim \text{Normal}\left(\mu_0, \sigma^2\right) \\ \dot{\sim} & \text{Normal}(0,1) & \text{if} \quad n \text{ large} \end{array} \right\}.$$

For σ^2 unknown and estimated by S^2, we learned that

$$T = \frac{\bar{X} - \mu_0}{\sqrt{S^2/n}} \qquad \left\{ \begin{array}{lll} \sim & t(n-1) & \text{if} \quad X_1, \ldots, X_n \sim \text{Normal}\left(\mu_0, \sigma^2\right) \\ \dot{\sim} & \text{Normal}(0,1) & \text{if} \quad n \text{ large} \end{array} \right\}.$$

These facts allowed us to construct confidence intervals for and test hypotheses about the population mean. The confidence intervals were of the form

$$\left(\begin{array}{c}\text{sample} \\ \text{mean}\end{array} \right) \quad \pm \quad q \cdot \left(\begin{array}{c}\text{standard} \\ \text{error}\end{array} \right),$$

where the critical value q is the appropriate quantile of the distribution of Z or T. The tests also were based on Z or T, and the significance probabilities were computed using the corresponding distribution.

The logic for drawing inferences about two populations means is identical to the logic for drawing inferences about one population mean—we simply

replace "mean" with "difference in means" and base inferences about Δ on the distribution of

$$\frac{\begin{array}{c}\text{sample} \\ \text{difference}\end{array} - \begin{array}{c}\text{hypothesized} \\ \text{difference}\end{array}}{\begin{array}{c}\text{standard deviation} \\ \text{of sample difference}\end{array}} = \frac{\hat{\Delta} - \Delta_0}{\text{standard error}}.$$

Because $X_i \sim \text{Normal}(\mu_1, \sigma_1^2)$ and $Y_j \sim \text{Normal}(\mu_2, \sigma_2^2)$,

$$\bar{X} \sim \text{Normal}\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \quad \text{and} \quad \bar{Y} \sim \text{Normal}\left(\mu_2, \frac{\sigma_2^2}{n_2}\right).$$

Because $\bar{X}$ and $\bar{Y}$ are independent, it follows from Theorem 5.2 that

$$\hat{\Delta} = \bar{X} - \bar{Y} \sim \text{Normal}\left(\Delta = \mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

We now distinguish three cases:

1. Both σ_i are known (and possibly unequal). The inferential theory for this case is easy; unfortunately, population variances are rarely known.

2. The σ_i are unknown, but necessarily equal ($\sigma_1 = \sigma_2 = \sigma$). This case should strike the student as somewhat implausible. If the population variances are not known, then under what circumstances might we reasonably assume that they are equal? Although such circumstances do exist, the primary importance of this case is that the corresponding theory is elementary. Nevertheless, it is important to study this case because the methods derived from the assumption of an unknown common variance are widely used—and abused.

3. The σ_i are unknown and possibly unequal. This is clearly the case of greatest practical importance, but the corresponding theory is somewhat unsatisfying. The problem of drawing inferences when the population variances are unknown and possibly unequal is sufficiently notorious that it has a name: the *Behrens-Fisher problem*.

11.1.1 Known Variances

If $\Delta = \Delta_0$, then

$$Z = \frac{\hat{\Delta} - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \text{Normal}(0, 1).$$

Given $\alpha \in (0,1)$, let q_z denote the $1-\alpha/2$ quantile of Normal$(0,1)$. We construct a $(1-\alpha)$-level confidence interval for Δ by writing

$$\begin{aligned} 1-\alpha &= P\left(|Z| < q_z\right) \\ &= P\left(|\hat{\Delta} - \Delta| < q_z\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) \\ &= P\left(\hat{\Delta} - q_z\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \Delta < \hat{\Delta} - q_z\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) \end{aligned}$$

The desired confidence interval is

$$\hat{\Delta} \pm q_z\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Example 11.3 For the first population, suppose that we know that the population standard deviation is $\sigma_1 = 5$ and that we observe a sample of size $n_1 = 60$ with sample mean $\bar{x} = 7.6$. For the second population, suppose that we know that the population standard deviation is $\sigma_2 = 2.5$ and that we observe a sample of size $n_2 = 15$ with sample mean $\bar{y} = 5.2$. To construct a 0.95-level confidence interval for Δ, we first compute

$$q_z = \texttt{qnorm}(.975) = 1.959964 \doteq 1.96,$$

then

$$(7.6 - 5.2) \pm 1.96\sqrt{\frac{5^2}{60} + \frac{2.5^2}{15}} \doteq 2.4 \pm 1.79 = (0.61, 4.21).$$

Example 11.4 For the first population, suppose that we know that the population variance is $\sigma_1^2 = 8$ and that we observe a sample of size $n_1 = 10$ with sample mean $\bar{x} = 9.7$. For the second population, suppose that we know that the population variance is $\sigma_2^2 = 96$ and that we observe a sample of size $n_2 = 5$ with sample mean $\bar{y} = 2.6$. To construct a 0.95-level confidence interval for Δ, we first compute

$$q_z = \texttt{qnorm}(.975) = 1.959964 \doteq 1.96,$$

then

$$(9.7 - 2.6) \pm 1.96\sqrt{\frac{8}{10} + \frac{96}{5}} \doteq 7.1 \pm 8.765 = (-1.665, 15.865).$$

To test $H_0 : \Delta = \Delta_0$ versus $H_1 : \Delta \neq \Delta_0$, we exploit the fact that $Z \sim \text{Normal}(0, 1)$ under H_0. Let z denote the observed value of Z. Then a natural level-α test is the test that rejects H_0 if and only if

$$\mathbf{p} = P_{\Delta_0}(|Z| \geq |z|) \leq \alpha,$$

which is equivalent to rejecting H_0 if and only if $|z| \geq q_z$. This test is sometimes called the 2-sample z-test.

Example 11.3 (continued) To test $H_0 : \Delta = 0$ versus $H_1 : \Delta \neq 0$, we compute

$$z = \frac{(7.6 - 5.2) - 0}{\sqrt{5^2/60 + 2.5^2/15}} \doteq 2.629.$$

Because $|2.629| > 1.96$, we reject H_0 at significance level $\alpha = 0.05$. The significance probability is

$$\mathbf{p} = P_{\Delta_0}(|Z| \geq |2.629|) = 2 * \texttt{pnorm}(-2.629) \doteq 0.008562.$$

Example 11.4 (continued) To test $H_0 : \Delta = 0$ versus $H_1 : \Delta \neq 0$, we compute

$$z = \frac{(9.7 - 2.6) - 0}{\sqrt{8/10 + 96/5}} \doteq 1.5876.$$

Because $|1.5876| < 1.96$, we decline to reject H_0 at significance level $\alpha = 0.05$. The significance probability is

$$\mathbf{p} = P_{\Delta_0}(|Z| \geq |1.5876|) = 2 * \texttt{pnorm}(-1.5876) \doteq 0.1124.$$

11.1.2 Unknown Common Variance

Now we assume that $\sigma_1 = \sigma_2 = \sigma$, but that the common variance σ^2 is unknown. Because σ^2 is unknown, we must estimate it. Let

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2$$

denote the sample variance for the X_i and let

$$S_2^2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (Y_j - \bar{Y})^2$$

denote the sample variance for the Y_j. If we only sampled the first population, then we would use S_1^2 to estimate the first population variance, σ^2. Likewise, if we only sampled the second population, then we would use S_2^2 to estimate the second population variance, σ^2. Neither is appropriate in the present situation, as S_1^2 does not use the second sample and S_2^2 does not use the first sample. Therefore, we create a weighted average of the separate sample variances,

$$\begin{aligned} S_P^2 &= \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{(n_1-1)+(n_2-1)} \\ &= \frac{1}{n_1+n_2-2}\left[\sum_{i=1}^{n_1}(X_i-\bar{X})^2+\sum_{j=1}^{n_2}(Y_j-\bar{Y})^2\right], \end{aligned}$$

the *pooled sample variance*. Then

$$ES_P^2 = \frac{(n_1-1)ES_1^2+(n_2-1)ES_2^2}{(n_1-1)+(n_2-1)} = \frac{(n_1-1)\sigma^2+(n_2-1)\sigma^2}{(n_1-1)+(n_2-1)} = \sigma^2,$$

so the pooled sample variance is an unbiased estimator of a common population variance. It is also consistent and asymptotically efficient for estimating a common normal variance.

Instead of

$$Z = \frac{\hat{\Delta}-\Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} = \frac{\hat{\Delta}-\Delta_0}{\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\sigma^2}},$$

we now rely on

$$T = \frac{\hat{\Delta}-\Delta_0}{\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)S_P^2}}.$$

The following result allows us to construct confidence intervals and test hypotheses about the shift parameter $\Delta = \mu_1 - \mu_2$.

Theorem 11.1 *If $\Delta = \Delta_0$, then $T \sim t(n_1+n_2-2)$.*

Given $\alpha \in (0,1)$, let q_t denote the $1-\alpha/2$ quantile of $t(n_1+n_2-2)$. Exploiting Theorem 11.1, a $(1-\alpha)$-level confidence interval for Δ is

$$\hat{\Delta} \pm q_t\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)S_P^2}.$$

Example 11.3 (continued) Now suppose that, instead of knowing population standard deviations $\sigma_1 = 5$ and $\sigma_2 = 2.5$, we observe sample standard deviations $s_1 = 5$ and $s_2 = 2.5$. The ratio of sample variances, $s_1^2/s_2^2 = 4 \neq 1$, strongly suggests that the population variances are unequal. We proceed under the assumption that $\sigma_1 = \sigma_2$ for the purpose of illustration. The pooled sample variance is

$$S_P^2 = \sqrt{\frac{59 \cdot 5^2 + 14 \cdot 2.5^2}{59 + 14}} = 21.40411.$$

To construct a 0.95-level confidence interval for Δ, we first compute

$$q_t = \texttt{qt}(.975, 73) = 1.992997 \doteq 1.993,$$

then

$$(7.6 - 5.2) \pm 1.993\sqrt{\left(\frac{1}{60} + \frac{1}{15}\right) \cdot 21.40411} \doteq 2.4 \pm 2.66 = (-0.26, 5.06).$$

Example 11.4 (continued) Now suppose that, instead of knowing population variances $\sigma_1^2 = 8$ and $\sigma_2^2 = 96$, we observe sample variances $s_1^2 = 8$ and $s_2^2 = 96$. Again, the ratio of sample variances, $s_2^2/s_1^2 = 12 \neq 1$, strongly suggests that the population variances are unequal. We proceed under the assumption that $\sigma_1 = \sigma_2$ for the purpose of illustration. The pooled sample variance is

$$S_P^2 = \sqrt{\frac{9 \cdot 8 + 4 \cdot 96}{9 + 4}} = 35.07692.$$

To construct a 0.95-level confidence interval for Δ, we first compute

$$q_t = \texttt{qt}(.975, 13) = 2.160369 \doteq 2.16,$$

then

$$(9.7 - 2.6) \pm 2.16\sqrt{\left(\frac{1}{10} + \frac{1}{5}\right) \cdot 35.07692} \doteq 7.1 \pm 7.01 = (0.09, 14.11).$$

To test $H_0 : \Delta = \Delta_0$ versus $H_1 : \Delta \neq \Delta_0$, we exploit the fact that $T \sim t(n_1 + n_2 - 2)$ under H_0. Let t denote the observed value of T. Then a natural level-α test is the test that rejects H_0 if and only if

$$\mathbf{p} = P_{\Delta_0}(|T| \geq |t|) \leq \alpha,$$

which is equivalent to rejecting H_0 if and only if $|t| \geq q_t$. This test is called *Student's 2-sample t-test.*

Example 11.3 (continued) To test $H_0 : \Delta = 0$ versus $H_1 : \Delta \neq 0$, we compute

$$t = \frac{(7.6 - 5.2) - 0}{\sqrt{(1/60 + 1/15) \cdot 21.40411}} \doteq 1.797.$$

Because $|1.797| < 1.993$, we decline to reject H_0 at significance level $\alpha = .05$. The significance probability is

$$\mathbf{p} = P_{\Delta_0}(|T| \geq |1.797|) = 2 * \texttt{pt}(-1.797, 73) \doteq 0.0764684.$$

Example 11.4 (continued) To test $H_0 : \Delta = 0$ versus $H_1 : \Delta \neq 0$, we compute

$$t = \frac{(9.7 - 2.6) - 0}{\sqrt{(1/10 + 1/5) \cdot 35.07692}} \doteq 2.19.$$

Because $|2.19| > 2.16$, we reject H_0 at significance level $\alpha = .05$. The significance probability is

$$\mathbf{p} = P_{\Delta_0}(|T| \geq |2.19|) = 2 * \texttt{pt}(-2.19, 13) \doteq 0.04747.$$

11.1.3 Unknown Variances

Now we drop the assumption that $\sigma_1 = \sigma_2$. We must then estimate each population variance separately, σ_1^2 with S_1^2 and σ_2^2 with S_2^2. Instead of

$$Z = \frac{\hat{\Delta} - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

we now rely on

$$T_W = \frac{\hat{\Delta} - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}.$$

Unfortunately, there is no analogue of Theorem 11.1—the exact distribution of T_W is not known.

The exact distribution of T_W appears to be intractable, but Welch (1937, 1947) argued that $T_W \dot{\sim} t(\nu)$, with

$$\nu = \frac{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^2}{\frac{(\sigma_1^2/n_1)^2}{n_1 - 1} + \frac{(\sigma_2^2/n_2)^2}{n_2 - 1}}.$$

Because σ_1^2 and σ_2^2 are unknown, we estimate ν by

$$\hat{\nu} = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1-1} + \frac{(S_2^2/n_2)^2}{n_2-1}}.$$

Simulation studies have revealed that the approximation $T_W \dot{\sim} t(\hat{\nu})$ works well in practice.

Given $\alpha \in (0,1)$, let q_t denote the $1-\alpha/2$ quantile of $t(\hat{\nu})$. Using Welch's approximation, an approximate $(1-\alpha)$-level confidence interval for Δ is

$$\hat{\Delta} \pm q_t \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}.$$

Example 11.3 (continued) Now we estimate the unknown population variances separately, σ_1^2 by $s_1^2 = 5^2$ and σ_2^2 by $s_2^2 = 2.5^2$. Welch's approximation involves

$$\hat{\nu} = \frac{\left(\frac{5^2}{60} + \frac{2.5^2}{15}\right)^2}{\frac{(5^2/60)^2}{60-1} + \frac{(2.5^2/15)^2}{15-1}} = 45.26027 \doteq 45.26$$

degrees of freedom. To construct a 0.95-level confidence interval for Δ, we first compute

$$q_t = \texttt{qt}(.975, 45.26) \doteq 2.014,$$

then

$$(7.6 - 5.2) \pm 2.014\sqrt{5^2/60 + 2.5^2/15} \doteq 2.4 \pm 1.84 = (0.56, 4.24).$$

Example 11.4 (continued) Now we estimate the unknown population variances separately, σ_1^2 by $s_1^2 = 8$ and σ_2^2 by $s_2^2 = 96$. Welch's approximation involves

$$\hat{\nu} = \frac{\left(\frac{8}{10} + \frac{96}{5}\right)^2}{\frac{(8/10)^2}{10-1} + \frac{(96/5)^2}{5-1}} = 4.336931 \doteq 4.337$$

degrees of freedom. To construct a 0.95-level confidence interval for Δ, we first compute

$$q_t = \texttt{qt}(.975, 4.337) \doteq 2.6934,$$

then

$$(9.7 - 2.6) \pm 2.6934\sqrt{8/10 + 96/5} \doteq 7.1 \pm 13.413 = (-6.313, 20.513).$$

To test $H_0 : \Delta = \Delta_0$ versus $H_1 : \Delta \neq \Delta_0$, we exploit the approximation $T_W \dot{\sim} t(\hat{\nu})$ under H_0. Let t_W denote the observed value of T_W. Then a natural approximate level-α test is the test that rejects H_0 if and only if

$$\mathbf{p} = P_{\Delta_0}\left(|T_W| \geq |t_W|\right) \leq \alpha,$$

which is equivalent to rejecting H_0 if and only if $|t_W| \geq q_t$. This test is sometimes called Welch's approximate t-test.

Example 11.3 (continued) To test $H_0 : \Delta = 0$ versus $H_1 : \Delta \neq 0$, we compute

$$t_W = \frac{(7.6 - 5.2) - 0}{\sqrt{5^2/60 + 2.5^2/15}} \doteq 2.629.$$

Because $|2.629| > 2.014$, we reject H_0 at significance level $\alpha = 0.05$. The significance probability is

$$\mathbf{p} = P_{\Delta_0}\left(|T_W| \geq |2.629|\right) = 2 * \texttt{pt}(-2.629, 45.26) \doteq 0.011655.$$

Example 11.4 (continued) To test $H_0 : \Delta = 0$ versus $H_1 : \Delta \neq 0$, we compute

$$t_W = \frac{(9.7 - 2.6) - 0}{\sqrt{8/10 + 96/5}} \doteq 1.4257.$$

Because $|1.4257| < 2.6934$, we decline to reject H_0 at significance level $\alpha = 0.05$. The significance probability is

$$\mathbf{p} = P_{\Delta_0}\left(|T_W| \geq |1.4257|\right) = 2 * \texttt{pt}(-1.4257, 4.337) \doteq 0.2218.$$

Examples 11.3 and 11.4 were carefully constructed to reveal the sensitivity of Student's 2-sample t-test to the assumption of equal population variances. Welch's approximation is good enough that we can use it to benchmark Student's test when variances are unequal. In Example 11.3, Welch's approximate t-test produced a significance probability of $\mathbf{p} \doteq 0.012$, leading us to reject the null hypothesis at $\alpha = 0.05$. Student's 2-sample t-test produced a misleading significance probability of $\mathbf{p} \doteq 0.076$, leading

us to commit a Type II error. In Example 11.4, Welch's approximate t-test produced a significance probability of $\mathbf{p} \doteq 0.222$, leading us to accept the null hypothesis at $\alpha = 0.05$. Student's 2-sample t-test produced a misleading significance probability of $\mathbf{p} \doteq 0.047$, leading us to commit a Type I error.

Evidently, Student's 2-sample t-test (and the corresponding procedure for constructing confidence intervals) should not be used unless one is convinced that the population variances are identical. The consequences of using Student's test when the population variances are unequal may be exacerbated when the sample sizes are unequal. In general:

- If $n_1 = n_2$, then $t = t_W$.
- If the population variances are (approximately) equal, then t and t_W tend to be (approximately) equal.
- If the larger sample is drawn from the population with the larger variance, then t will tend to be less than t_W. All else equal, this means that Student's test will tend to produce significance probabilities that are too large.
- If the larger sample is drawn from the population with the smaller variance, then t will tend to be greater than t_W. All else equal, this means that Student's test will tend to produce significance probabilities that are too small.
- If the population variances are (approximately) equal, then $\hat{\nu}$ will be (approximately) $n_1 + n_2 - 2$.
- It will *always* be the case that $\hat{\nu} \leq n_1 + n_2 - 2$. All else equal, this means that Student's test will tend to produce significance probabilities that are too large.

From these observations we draw the following conclusions:

1. If the population variances are unequal, then Student's 2-sample t-test may produce misleading significance probabilities.
2. If the population variances are equal, then Welch's approximate t-test is approximately equivalent to Student's 2-sample t-test. Thus, if one uses Welch's test in the situation for which Student's test is appropriate, one is not likely to be led astray.

3. *Don't use Student's 2-sample t-test!* I remember how shocked I was when I first heard this advice as a first-year graduate student in a course devoted to the theory of hypothesis testing. The instructor, Erich Lehmann, one of the great statisticians of the 20th century and the author of a famous book on hypothesis testing, told us: "If you get just one thing out of this course, I'd like it to be that you should *never* use Student's 2-sample t-test."

11.2 The Case of a General Shift Family

11.3 The Symmetric Behrens-Fisher Problem

11.4 Exercises

Problem Set A

1. We have been using various mathematical symbols in our study of 1- and 2-sample location problems. Each of the symbols listed below is used to represent a real number. State which of the following statements applies to each symbol:

   i. The real number represented by this symbol is an unknown population parameter.

   ii. The real number represented by this symbol is calculated from the observed data.

   iii. The real number represented by this symbol is specified by the experimenter.

   Here are the symbols:

   $$\mu \quad \mu_0 \quad \bar{x} \quad s^2 \quad t \quad \alpha \quad \Delta \quad \Delta_0 \quad \mathbf{p} \quad \hat{\nu}$$

2. Assume that $X_1, \ldots, X_{10} \sim \text{Normal}(\mu_1, \sigma_1^2)$ and that $Y_1, \ldots, Y_{20} \sim \text{Normal}(\mu_2, \sigma_2^2)$. None of the population parameters are known. Let $\Delta = \mu_1 - \mu_2$. To test $H_0 : \Delta \geq 0$ versus $H_1 : \Delta < 0$ at significance level $\alpha = 0.05$, we observe samples $\vec{x}$ and $\vec{y}$.

   (a) What test should be used in this situation? If we observe $\vec{x}$ and $\vec{y}$ that result in $\bar{x} = -0.82$, $s_1 = 4.09$, $\bar{y} = 1.39$, and $s_2 = 1.22$, then what is the value of the test statistic?

   (b) If we observe $\vec{x}$ and $\vec{y}$ that result in $s_1 = 4.09$, $s_2 = 1.22$,, and a test statistic value of 1.76, then which of the following `R` expressions best approximates the significance probability?

      i. `2*pnorm(-1.76)`

      ii. `pt(-1.76,df=28)`

      iii. `pt(1.76,df=10)`

      iv. `pt(-1.76,df=10)`

      v. `2*pt(1.76,df=28)`

   (c) True of False: if we observe $\vec{x}$ and $\vec{y}$ that result in a significance probability of $\mathbf{p} = 0.96$, then we should reject the null hypothesis.

Problem Set B Each of the following scenarios can be modelled as a 1- or 2-sample location problem. For 1-sample problems, let X_i denote the random variables of interest and let $\mu = EX_i$. For 2-sample problems, let X_i and Y_j denote the random variables of interest; let $\mu_1 = EX_i$, $\mu_2 = EY_j$, and $\Delta = \mu_1 - \mu_2$. For each scenario, you should answer/do the following:

(a) What is the experimental unit?

(b) From how many populations were the experimental units drawn? Identify the population(s). How many units were drawn from each population? Is this a 1- or a 2-sample problem?

(c) How many measurements were taken on each experimental unit? Identify them.

(d) Define the parameter(s) of interest for this problem. For 1-sample problems, this should be μ; for 2-sample problems, this should be Δ.

(e) State appropriate null and alternative hypotheses.

Here are the scenarios:

1. A mathematics/education concentrator theorizes that learning mathematics and statistics is sometimes impeded by the widespread use of odd symbols like α, χ, and ω. She reasons that, if her theory is correct, then students who belong to sororities and fraternities—who she presumes are more familiar with Greek letters—should have an easier time learning the mathematical subjects that use such symbols. To investigate, she obtains a list of all William & Mary students who are enrolled in Math 111 (calculus) and a list of all William & Mary students who belong to a sorority or fraternity. She uses this information to choose (at random) 20 calculus students who do belong to a sorority or fraternity and 20 calculus students who do not. She persuades each of these students to take a calculus quiz, specially designed to use lots of Greek letters. How might she use the resulting data to test her theory? (Respond to (a)–(e) above.)

2. Umberto theorizes that living with a dog diminishes depression in the elderly, here defined as more than 70 years of age. To investigate his theory, he recruits 15 single elderly men who own dogs and 15 single elderly men who do not own any pets. The Hamilton instrument for

measuring depressive tendency is administered to each subject. High scores indicate depression. How might Umberto use the resulting data to test his theory? (Respond to (a)–(e) above.)

3. The William & Mary women's tennis team uses championship balls in their matches and less expensive practice balls in their team practices. The players have formed a strong impression that the practice balls do not wear as well as the championship balls, i.e., that the practice balls lose their bounce more quickly than the championship balls. To investigate this perception, Nina and Delphine conceive the following experiment. Before one practice, the team opens new cans of championship balls and practice balls, which they then use for that day's practice. After practice, Nina and Delphine randomly select 10 of the used championship balls and 10 of the used practice balls. They drop each ball from a height of 1 meter and measure the height of its first bounce. How might Nina and Delphine test the team's impression that practice balls do not wear as well as championship balls? (Respond to (a)–(e) above.)

4. A political scientist theorizes that women tend to be more opposed to military intervention than do men. To investigate this theory, he devises an instrument on which a subject responds to several recent U.S. military interventions on a 5-point Likert scale (1="strongly support,"...,5="strongly oppose"). A subject's score on this instrument is the sum of his/her individual responses. The scientist randomly selects 50 married couples in which neither spouse has a registered party affiliation and administers the instrument to each of the 100 individuals so selected. How might he use his results to determine if his theory is correct? (Respond to (a)–(e) above.)

5. A shoe company claims that wearing its racing flats will typically improve one's time in a 10K road race by more than 30 seconds. A running magazine sponsors an event to test this claim. It arranges for 120 runners to enter two road races, held two weeks apart on the same course. For the second race, each of these runners is supplied with the new racing flat. How might the race results be used to determine the validity of the shoe company's claim? (Respond to (a)–(e) above.)

6. Susan theorizes that impregnating wood with an IGR (insect growth regulator) will reduce wood consumption by termites. To investigate

this theory, she impregnates 60 wood blocks with a solvent containing the IGR and 60 wood blocks with just the solvent. Each block is weighed, then placed in a separate container with 100 ravenous termites. After two weeks, she removes the blocks and weighs them again to determine how much wood has been consumed. How might Susan use her results to determine if her theory is correct? (Respond to (a)–(e) above.)

7. According to an article in *Newsweek* (May 10, 2004, page 89), recent "studies have shown consistently that women are better than men at reading and responding to subtle cues about mood and temperament." Some psychologists believe that such differences can be explained in part by biological differences between male and female brains. One such psychologist conducts a study in which day-old babies are shown three human faces and three mechanical objects. The time that the baby stares at each face/object is recorded. Of interest is how much time the baby spends staring at faces versus how much time the baby spends staring at objects. The psychologist's theory predicts that this comparison will differ by sex, with female babies preferring faces to objects to a greater extent than do male babies. How might the psychologist use his results to determine if his theory is correct? (Respond to (a)–(e) above.)

8. To investigate the effect of swing dancing on cardiovascular fitness, an exercise physiologist recruits 20 couples enrolled in introductory swing dance classes. Each class meets once a week for ten weeks. Participants are encouraged to go out dancing on at least two additional occasions each week. In general, lower resting pulses are associated with greater cardiovascular fitness. Accordingly, each participant's resting pulse is measured at the beginning and at the end of the ten-week class. How might the resulting data be used to determine if swing dancing improves cardiovascular fitness? (Respond to (a)–(e) above.)

9. It is thought that Alzheimer's disease (AD) impairs short-term memory more than it impairs long-term memory. To test this theory, a psychologist studied 60 mildly demented AD patients and 60 normal elderly control subjects. Each subject was administered a short-term and a long-term memory task. On each task, high scores are better than low scores. How might the psychologist use the resulting task scores to determine if the theory is correct? (Respond to (a)–(e) above.)

Problem Set C In the early 1960s, the Western Collaborative Group Study investigated the relation between behavior and risk of coronary heart disease in middle-aged men. Type A behavior is characterized by urgency, aggression and ambition; Type B behavior is noncompetitive, more relaxed and less hurried. The following data, which appear in Table 2.1 of Selvin (1991) and Data Set 47 in *A Handbook of Small Data Sets*, are the cholesterol measurements of 20 heavy men of each behavior type. (In fact, these 40 men were the heaviest in the study. Each weighed at least 225 pounds.) We consider whether or not they provide evidence that heavy Type A men have higher cholesterol levels than heavy Type B men.

Cholesterol Levels for Heavy Type A Men									
233	291	312	250	246	197	268	224	239	239
254	276	234	181	248	252	202	218	212	325

Cholesterol Levels for Heavy Type B Men									
344	185	263	246	224	212	188	250	148	169
226	175	242	252	153	183	137	202	194	213

1. Respond to (a)–(e) in Problem Set A.

2. Does it seem reasonable to assume that the samples $\vec{x}$ and $\vec{y}$, the observed values of $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$, were drawn from normal distributions? Why or why not?

3. Assume that the X_i and the Y_j are normally distributed.

   (a) Test the null hypothesis derived above using Welch's approximate t-test. What is the significance probability? If we adopt a significance level of $\alpha = 0.05$, should we reject the null hypothesis?

   (b) Construct a (2-sided) confidence interval for Δ with a confidence coefficient of approximately 0.90.

Problem Set D Researchers measured urinary β-thromboglobulin excretion in 12 diabetic patients and 12 normal control subjects, reporting their findings in *Thrombosis and Haemostasis*. The following measurements are Data Set 313 in *A Handbook of Small Data Sets*:

Normal	4.1	6.3	7.8	8.5	8.9	10.4
	11.5	12.0	13.8	17.6	24.3	37.2
Diabetic	11.5	12.1	16.1	17.8	24.0	28.8
	33.9	40.7	51.3	56.2	61.7	69.2

1. Do these measurements appear to be samples from symmetric distributions? Why or why not?

2. Both samples of positive real numbers appear to be drawn from distributions that are skewed to the right, i.e., the upper tail of the distribution is longer than the lower tail of the distribution. Often, such distributions can be symmetrized by applying a suitable data transformation. Two popular candidates are:

   (a) The natural logarithm: $u_i = \log(x_i)$ and $v_j = \log(y_j)$.

   (b) The square root: $u_i = \sqrt{x_i}$ and $v_j = \sqrt{y_j}$.

   Investigate the effect of each of these transformations on the above measurements. Do the transformed measurements appear to be samples from symmetric distributions? Which transformation do you prefer?

3. Do the transformed measurements appear to be samples from normal distributions? Why or why not?

4. The researchers claimed that diabetic patients have increased urinary β-thromboglobulin excretion. Assuming that the transformed measurements are samples from normal distributions, how convincing do you find the evidence for their claim?

Problem Set E

1. Chemistry lab partners Arlen and Stuart collaborated on an experiment in which they measured the melting points of 20 specimens of two types of sealing wax. Twelve of the specimens were of one type (A); eight were of the other type (B). Each student then used Welch's approximate t-test to test the null hypothesis of no difference in mean melting point between the two methods:

   - Arlen applied Welch's approximate t-test to the original melting points, which were measured in degrees Fahrenheit.
   - Stuart first converted each melting point to degrees Celsius (by subtracting 32, then multiplying by 5/9), then applied Welch's approximate t-test to the converted melting points.

Comment on the potential differences between these two analyses. In particular, is it *True* or *False* that (ignoring round-off error) Arlen and Stuart will obtain identical significance probabilities? Please justify your comments.

2. A graduate student in ornithology would like to determine if created marshes differ from natural marches in their appeal to avian communities. He plans to observe $n_1 = 9$ natural marshes and $n_2 = 9$ created marshes, counting the number of red-winged blackbirds per acre that inhabit each marsh. His thesis committee wants to know how much he thinks he will be able to learn from this experiment.

   Let X_i denote the number of blackbirds per acre in natural marsh i and let Y_j denote the number of blackbirds per acre in created marsh j. In order to respond to his committee, the student makes the simplifying assumptions that $X_i \sim \text{Normal}(\mu_1, \sigma^2)$ and $Y_j \sim \text{Normal}(\mu_2, \sigma^2)$. He estimates that $\text{iqr}(X_i) = \text{iqr}(Y_j) = 10$. Calculate L, the length of the 0.90-level confidence interval for $\Delta = \mu_1 - \mu_2$ that he can expect to construct.

3. A film buff has formed the vague impression that movies tend to be longer than they used to be. Are they really longer? Or do they just *seem* longer? To investigate, he randomly samples U.S. feature films made in 1956 and U.S. feature films made in 1996, obtaining the following results:

Year	Title	Minutes
1956	*Accused of Murder*	74
	Away All Boats	114
	Baby Doll	114
	The Bold and the Brave	87
	Come Next Spring	92
	The Flaming Teen-Age	55
	Gun Girls	67
	Helen of Troy	118
	The Houston Story	79
	Patterns	83
	The Price of Fear	79
	The Revolt of Mamie Stover	92
	Written on the Wind	99
	The Young Guns	87
1996	*$40,000*	70
	Barb Wire	98
	Breathing Room	90
	Daddy's Girl	95
	Ed's Next Move	88
	From Dusk to Dawn	108
	Galgameth	110
	The Glass Cage	96
	Kissing a Dream	91
	Love & Sex etc.	88
	Love is All There Is	120
	Making the Rules	96
	Spirit Lost	90
	Work	90

Do these data provide convincing evidence that 1996 movies are longer than 1956 movies? Compute a significance probability that may be used to encourage or discourage the film buff's impression. Explain how this number should be interpreted. Identify and defend any assumptions that you made in your calculations.

Chapter 12

k-Sample Location Problems

Now we generalize our study of location problems from two to $k \geq 3$ populations. Again we are concerned with comparing the populations with respect to some measure of centrality, typically the population mean or the population median. We designate the populations by $P_1, \ldots, P_k$ and the corresponding sample sizes by $n_1, \ldots, n_k$. Our bookkeeping will be facilitated by the use of double subscripts, e.g.,

$$\begin{aligned} X_{11}, \ldots, X_{1n_1} &\sim P_1, \\ X_{21}, \ldots, X_{2n_2} &\sim P_2, \\ &\vdots \\ X_{k1}, \ldots, X_{kn_k} &\sim P_k. \end{aligned}$$

These expressions can be summarized succinctly by writing

$$X_{ij} \sim P_i.$$

We assume the following:

1. The X_{ij} are mutually independent continuous random variables.
2. P_i has location parameter θ_i, e.g., $\theta_i = \mu_i = EX_{ij}$ or $\theta_i = q_2(X_{ij})$.
3. We observe random samples $\vec{x}_i = \{x_{i1}, \ldots, x_{in_i}\}$, from which we attempt to draw inferences about $(\theta_1, \ldots, \theta_k)$. In general, we do *not* assume that $n_1 = \cdots = n_k$. However, certain procedures do require equal sample sizes. Furthermore, certain procedures that can be used with unequal sample sizes are greatly simplified when the sample sizes are equal.

The same four questions that we posed at the beginning of Chapter 10 and asked in Chapters 10–11 can be asked here. What distinguishes k-sample problems from 1-sample and 2-sample problems is the number of populations from which the experimental units were drawn. The prototypical case of a k-sample problem is the case of several treatment populations.

One may wonder why we distinguish between $k = 2$ and $k \geq 3$ populations. In fact, many methods for k-sample problems can be applied to 2-sample problems, in which case they often simplify to methods studied in Chapter 11. However, many issues arise with $k \geq 3$ populations that do not arise with two populations, so the problem of comparing more than two location parameters is considerably more complicated than the problem of comparing only two. For this reason, our study of k-sample location problems will be less comprehensive than our previous studies of 1-sample and 2-sample location problems.

12.1 The Case of a Normal Shift Family

In this section we assume that $P = \text{Normal}(\mu_i, \sigma^2)$. This is sometimes called the fixed effects model for the oneway analysis of variance (ANOVA). Notice that we are assuming that each normal population has the same variance. Recall that we criticized the assumption of equal variances for the normal 2-sample problem. In that setting, however, Welch's approximate t-test provides a viable alternative that is available in many popular statistical software packages. In the more complicated setting of k normal populations, the assumption of equal variances (sometimes called the assumption of *homoscedasticity*) is fairly standard, if only because it is less clear how to proceed when the variances are unequal. The problem of unequal variances is discussed in Section 12.3.

12.1.1 The Fundamental Null Hypothesis

The fundamental problem of the analysis of variance is the problem of testing the null hypothesis that all of the population means are the same, i.e.,

$$H_0 : \mu_1 = \cdots = \mu_k, \tag{12.1}$$

against the alternative hypothesis that they are not all the same. Notice that the statement that the population means are not identical does *not* imply that each population mean is distinct. For example, if $\mu_1 = \mu_2 = 1.5$

and $\mu_3 = 2.2$, then H_0 is false. We stress that the analysis of variance is concerned with inferences about means, not variances.

To motivate our test of H_0, we formulate another null hypothesis that is equivalent to H_0. First, let

$$N = \sum_{i=1}^{k} n_i$$

denote the sum of the sample sizes and let

$$\bar{\mu}_\cdot = \sum_{i=1}^{k} \frac{n_i}{N} \mu_i$$

denote the *population grand mean*. The population grand mean is a weighted average of the individual population means, each population weighted in proportion to how many of the observations were drawn from it. If H_0 is true, then $\mu_1 = \cdots = \mu_k$ have a common value, say μ, and the population grand mean equals that common value:

$$\bar{\mu}_\cdot = \sum_{i=1}^{k} \frac{n_i}{N} \mu = \frac{\mu}{N} \sum_{i=1}^{k} n_i = \mu.$$

Next we introduce a quantity that measures how nearly the individual population means equal the population grand mean. Let

$$\gamma = \sum_{i=1}^{k} n_i \left(\mu_i - \bar{\mu}_\cdot\right)^2. \tag{12.2}$$

Notice that $\gamma \geq 0$ and that $\gamma = 0$ if and only if each $\mu_i = \bar{\mu}_\cdot$. But each $\mu_i = \bar{\mu}_\cdot$ if and only if each individual mean assumes a common value, which occurs if and only if the individual means are identical. Thus, H_0 is equivalent to the null hypothesis

$$H_0' : \gamma = 0,$$

which is to be tested against the alternative hypothesis

$$H_1' : \gamma > 0.$$

12.1.2 Testing the Fundamental Null Hypothesis

The idea that underlies our test is to estimate γ and reject H_0' when the estimate is sufficiently larger than zero. To estimate γ, we need only estimate

the population means that appear in (12.2). The individual sample means,

$$\bar{X}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij},$$

are unbiased estimators of the individual population means, and the sample grand mean,

$$\bar{X}_{\cdot\cdot} = \sum_{i=1}^{k} \frac{n_i}{N} \bar{X}_{i\cdot} = \sum_{i=1}^{k} \frac{n_i}{N} \left(\frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij} \right) = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij}$$

is an unbiased estimator of the population grand mean. Hence, a natural estimator of γ is the *between-groups* or *treatment* sum of squares,

$$SS_B = \sum_{i=1}^{k} n_i \left(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot} \right)^2,$$

the variation of the individual sample means about the sample grand mean. A useful formula for computing the observed value of SS_B from the observed values of the individual sample means is

$$ss_B = \sum_{i=1}^{k} n_i \bar{x}_{i\cdot}^2 - \frac{1}{N} \left(\sum_{i=1}^{k} n_i \bar{x}_{i\cdot} \right)^2.$$

What remains is to determine when SS_B is "sufficiently larger than zero." We consider two cases, depending on whether or not the common population variance σ^2 is known.

Known Population Variance

Situations in which σ^2 is known are rarely encountered, but it is useful to consider how to proceed in this case. Here is the key fact that we require:

Theorem 12.1 *Under the fundamental null hypothesis (12.1), the random variable*

$$SS_B/\sigma^2 \sim \chi^2(k-1),$$

where $\chi^2(\nu)$ denotes the chi-squared distribution with ν degrees of freedom, introduced in Section 5.5. The quantity $k-1$ is the between-groups degrees of freedom.

Theorem 12.1 suggests a way to determine whether or not SS_B is "sufficiently larger than zero." Under H_0,

$$P\left(SS_B \geq q\right) = P\left(SS_B/\sigma^2 \geq q/\sigma^2\right) = P\left(Y \geq q/\sigma^2\right),$$

where $Y \sim \chi^2(k-1)$; hence, we can use the chi-squared distribution to compute significance probabilities and/or critical values.

Example 12.1 Suppose that we draw samples of $n_1 = 20$, $n_2 = 25$, and $n_3 = 30$ observations from normal populations with unknown means and common variance $\sigma^2 = 9$, obtaining sample means of $\bar{x}_1 = 1.489$, $\bar{x}_2 = 1.712$, and $\bar{x}_3 = 3.082$. To test the fundamental null hypothesis that the individual population means are identical, we first compute $N = 20 + 25 + 30 = 75$ and evaluate SS_B, obtaining

$$\begin{aligned} ss_B &= \left(20 \cdot 1.489^2 + 25 \cdot 1.712^2 + 30 \cdot 3.082^2\right) - \\ &\quad \left(20 \cdot 1.489 + 25 \cdot 1.712 + 30 \cdot 3.082\right)^2 \\ &\doteq 39.402. \end{aligned}$$

Now we use the `R` function `pchisq` to compute a significance probability **p**:

```
> 1-pchisq(39.402/9,df=2)
[1] 0.1120287
```

For conventional levels of significance, $\mathbf{p} > 0.10$ is too large to warrant rejecting the null hypothesis.

Unknown Population Variance

Now we consider the more realistic case of an unknown population variance. Our development will mimic the case of a known population variance, but it is complicated by the need to estimate σ^2. Recall that, in Section 11.1.2, we estimated the unknown common population variance of $k = 2$ normal populations with the pooled sample variance,

$$S_P^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)},$$

where S_i^2 is the sample variance for sample i. This procedure is easily extended to the present case of $k \geq 3$ by defining the pooled sample variance

as

$$\begin{aligned} S_P^2 &= \frac{(n_1-1)S_1^2+\cdots+(n_k-1)S_k^2}{(n_1-1)+\cdots+(n_2-1)} \\ &= \frac{1}{n_1+\cdots+n_k-k}\sum_{i=1}^{k}(n_i-1)\,S_i^2 \\ &= \frac{1}{N-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}_{i\cdot}\right)^2. \end{aligned}$$

As in the case of $k=2$,

$$\begin{aligned} ES_P^2 &= \frac{(n_1-1)ES_1^2+\cdots+(n_k-1)ES_k^2}{(n_1-1)+\cdots+(n_k-1)} \\ &= \frac{(n_1-1)\sigma^2+\cdots(n_k-1)\sigma^2}{(n_1-1)+\cdots+(n_k-1)}=\sigma^2, \end{aligned}$$

so the pooled sample variance is an unbiased estimator of a common population variance. It is also consistent and asymptotically efficient for estimating a common normal variance.

In the previous case of a known population variance, our statistic for testing the fundamental null hypothesis was SS_B/σ^2. In the present case of an unknown population variance, we estimate σ^2 with S_P^2. Our test statistic will turn out to be SS_B/S_P^2 multiplied by a constant.

In order to simplify the formulas that follow, we multiply S_P^2 by $N-k$, obtaining the *within-groups* or *error* sum of squares

$$SS_W=(N-k)S_P^2=\sum_{i=1}^{k}(n_i-1)\,S_i^2=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}_{i\cdot}\right)^2.$$

In contrast to SS_B, which measures the variation of the individual sample means about the sample grand mean, SS_W measures the variations of the individual observations about the corresponding sample means. For completeness, we also define the *total* sum of squares,

$$SS_T=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}_{\cdot\cdot}\right)^2,$$

which measures the variation of the individual observations about the sample grand mean.

There is a beautiful relationship between SS_B, SS_W, and SS_T, viz.,

Theorem 12.2 $SS_B + SS_W = SS_T$

This formula turns out to be a corollary of the Pythagorean Theorem in N-dimensional Euclidean space! (In Section 14.2, we will explore a similar formula in greater detail.) The reason that our method for testing the fundamental null hypothesis is called the analysis of variance is that the method relies on decomposing total squared error into squared error between groups and squared error within groups. This elegant—and extremely useful—decomposition is only possible when we use *squared* error.

The quantities SS_B, SS_W, and SS_T are random variables. The following facts, which subsume Theorem 12.1, summarize the statistical behavior of these random variables.

Theorem 12.3 *The random variable*

$$SS_T/\sigma^2 \sim \chi^2(N-1).$$

The quantity $N-1$ is the total degrees of freedom.

Under the fundamental null hypothesis (12.1), SS_B and SS_W are independent random variables and

$$\begin{aligned} SS_B/\sigma^2 &\sim \chi^2(k-1), \\ SS_W/\sigma^2 &\sim \chi^2(N-k). \end{aligned}$$

The quantity $k-1$ is the between-groups degrees of freedom and the quantity $N-k$ is the within-groups degrees of freedom.

We have already remarked that the random variable

$$\frac{SS_B}{S_P^2} = \frac{SS_B}{SS_W/(N-k)}$$

would seem to be a natural statistic for testing the fundamental null hypothesis. Although sound in theory, this approach fails in practice because the distribution of SS_B/S_P^2 is not tractable. Fortunately, this approach can be salvaged by a trivial modification. Applying the definition of Fisher's F distribution in Section 5.5 to the independent χ^2 random variables SS_B/σ^2 and SS_W/σ^2, we discover

Corollary 12.1 *Under the fundamental null hypothesis (12.1),*

$$F = \frac{\frac{SS_B}{\sigma^2}/(k-1)}{\frac{SS_W}{\sigma^2}/(N-k)} = \frac{SS_B/(k-1)}{SS_W/(N-k)} \sim F(k-1, N-k),$$

where $F(\nu_1, \nu_2)$ denotes Fisher's F distribution with ν_1 and ν_2 degrees of freedom.

The random variable F is the desired test statistic; notice that

$$F = \frac{SS_B/(k-1)}{SS_W/(N-k)} = \frac{1}{k-1}\frac{SS_B}{S_P^2}.$$

Appealing to Corollary 12.1, we see that the ANOVA F-test of the fundamental null hypothesis of equal population means is to reject H_0 at significance level α if and only if the significance probability

$$\mathbf{p} = P(Y \geq f) \leq \alpha,$$

where f denotes the observed value of F and $Y \sim F(k-1, N-k)$. Of course, we can also formulate the test using critical values instead of significance probabilities, in which case we reject H_0 at significance level α if and only if $f \geq q$, where q is the $1-\alpha$ quantile of the $F(k-1, N-k)$ distribution.

Example 12.2 Suppose that we draw samples of $n_1 = 25$, $n_2 = 20$, and $n_3 = 20$ observations from normal populations with unknown means and unknown common variance, obtaining the following sample quantities:

	$i=1$	$i=2$	$i=3$
n_i	25	20	20
$\bar{x}_{i\cdot}$	9.783685	10.908170	15.002820
s_i^2	29.89214	18.75800	51.41654

To test the null hypothesis of equal population means at significance level $\alpha = 0.05$, we begin by computing the observed values of SS_B and SS_W, obtaining $ss_B \doteq 322.4366$ and

$$ss_W = (25-1)\cdot 29.89214 + (20-1)\cdot 18.75800 + (20-1)\cdot 51.41654 \doteq 2050.7280.$$

It follows that the observed value of the test statistic is

$$f = \frac{ss_B/(k-1)}{ss_W/(N-k)} \doteq \frac{322.4366/2}{2050.7280/62} \doteq 4.874141.$$

Now we use the `R` function `pf` to compute a significance probability $\mathbf{p}$:

```
> 1-pf(4.874141,df1=2,df2=62)
[1] 0.01081398
```

Because $\mathbf{p} < \alpha$, we reject the null hypothesis. Equivalently, we might use the `R` function `qf` to compute a critical value q:

```
> qf(1-.05,df1=2,df2=62)
[1] 3.145258
```

Because $f > q$, we reject the null hypothesis.

The information related to an ANOVA F-test is usually collected in an ANOVA table:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	Test Statistic	Significance Probability
Between	SS_B	$k-1$	MS_B	F	$\mathbf{p}$
Within	SS_W	$N-k$	$MS_W = S_P^2$		
Total	SS_T	$N-1$			

Note that we have introduced new notation for the *mean squares*, $MS_B = SS_B/(k-1)$ and $MS_W = SS_W/(N-k)$, allowing us to write $F = MS_B/MS_W$. It is also helpful to examine $R^2 = SS_B/SS_T$, the proportion of total variation "explained" by differences in the sample means.

Example 12.2 (continued) For the ANOVA performed in Example 12.2, the ANOVA table is

Source	SS	df	MS	F	$\mathbf{p}$
Between	322.4366	2	161.21830	4.874141	0.01081398
Within	2050.7280	62	33.07625		
Total	2373.1640	64			

The proportion of total variation explained by differences in the sample means is $322.4366/2373.1640 \doteq 0.1358678$. Thus, although there is sufficient variation between the sample means for us to infer that the population means are not identical, this variation accounts for a fairly small proportion of the total variation in the data.

12.1.3 Planned Comparisons

- Rejecting $H_0 : \mu_1 = \cdots = \mu_k$ leaves numerous alternatives. Typically, the investigator would like to say more than simply "H_0 is false." Often, one can determine specific comparisons of interest *in advance of the experiment.*

- Example: Heyl (1930) attempted to determine the gravitational constant using $k = 3$ different materials—gold, platinum, and glass. It seems natural to ask not just if the three materials lead to identical determinations of the gravitational constant, by testing $H_0 : \mu_1 = \mu_2 = \mu_3$, but also to ask:

  1. If glass differs from the two heavy metals, by testing

  $$H_0 : \frac{\mu_1 + \mu_2}{2} = \mu_3 \quad \text{vs.} \quad H_1 : \frac{\mu_1 + \mu_2}{2} \neq \mu_3,$$

  or, equivalently,

  $$H_0 : \mu_1 + \mu_2 = 2\mu_3 \quad \text{vs.} \quad H_1 : \mu_1 + \mu_2 \neq 2\mu_3,$$

  or, equivalently,

  $$H_0 : \mu_1 + \mu_2 - 2\mu_3 = 0 \quad \text{vs.} \quad H_1 : \mu_1 + \mu_2 - 2\mu_3 \neq 0,$$

  or, equivalently,

  $$H_0 : \theta_1 = 0 \quad \text{vs.} \quad H_1 : \theta_1 \neq 0,$$

  where $\theta_1 = \mu_1 + \mu_2 - 2\mu_3$.

  2. If the two heavy metals differ from each other, by testing

  $$H_0 : \mu_1 = \mu_2 \quad \text{vs.} \quad H_1 : \mu_1 \neq \mu_2,$$

  or, equivalently,

  $$H_0 : \mu_1 - \mu_2 = 0 \quad \text{vs.} \quad H_1 : \mu_1 - \mu_2 \neq 0,$$

  or, equivalently,

  $$H_0 : \theta_2 = 0 \quad \text{vs.} \quad H_1 : \theta_2 \neq 0,$$

  where $\theta_2 = \mu_1 - \mu_2$.

- **Definition 12.1** *A contrast is a linear combination of the k population means,*

$$\theta = \sum_{i=1}^{k} c_i \mu_i,$$

for which $\sum_{i=1}^{k} c_i = 0$.

- For example, in the contrasts suggested above,

  1. $\theta_1 = 1 \cdot \mu_1 + 1 \cdot \mu_2 + (-2) \cdot \mu_3$ and $1 + 1 - 2 = 0$; and
  2. $\theta_2 = 1 \cdot \mu_1 + (-1) \cdot \mu_2 + 0 \cdot \mu_3$ and $1 - 1 + 0 = 0$.

  We usually identify different contrasts by their coefficients, e.g., $c = (1, 1, -2)$.

Orthogonal Contrasts

- We want to test $H_0 : \theta = 0$ vs. $H_1 : \theta \neq 0$. An unbiased estimator of θ is

$$\hat{\theta} = \sum_{i=1}^{k} c_i \bar{X}_{i\cdot};$$

  we will reject H_0 if $\hat{\theta}$ is observed sufficiently far from zero.

- The quantity $(\hat{\theta})^2$ is not a satisfactory measure of departure from $H_0 : \theta = 0$ because it depends on the magnitude of the coefficients in the contrast. Accordingly, we define the sum of squares associated with the contrast θ to be

$$SS_\theta = \frac{\left(\sum_{i=1}^{k} c_i \bar{X}_{i\cdot}\right)^2}{\sum_{i=1}^{k} c_i^2 / n_i}.$$

- Fact: Under $H_0 : \mu_1 = \cdots = \mu_k$, SS_θ is independent of SS_W and

$$SS_\theta / \sigma^2 \sim \chi^2(1).$$

- Fact: Under $H_0 : \mu_1 = \cdots = \mu_k$,

$$F(\theta) = \frac{\frac{SS_\theta}{\sigma^2}/1}{\frac{SS_W}{\sigma^2}/(N-k)} = \frac{SS_\theta}{SS_W/(N-k)} \sim F(1, N-k).$$

- The F-test of $H_0 : \theta = 0$ is to reject if and only if

$$\mathbf{p} = P_{H_0}\left(F(\theta) \geq f(\theta)\right) \leq \alpha,$$

  i.e., if and only if

$$f(\theta) \geq q = \texttt{qf(1-}\alpha\texttt{,df1=1,df2=N-k)},$$

  where $f(\theta)$ denotes the observed value of $F(\theta)$.

- **Definition 12.2** *Two contrasts with coefficient vectors* $(c_1, \ldots, c_k)$ *and* $(d_1, \ldots, d_k)$ *are orthogonal if*

$$\sum_{i=1}^{k} \frac{c_i d_i}{n_i} = 0.$$

- Notice that, if $n_1 = \cdots = n_k$, then the orthogonality condition simplifies to

$$\sum_{i=1}^{k} c_i d_i = 0.$$

- In the Heyl (1930) example:
  - If $n_1 = n_2 = n_3$, then θ_1 and θ_2 are orthogonal because

$$1 \cdot 1 + 1 \cdot (-1) + (-2) \cdot 0 = 0.$$

  - If $n_1 = 6$ and $n_2 = n_3 = 5$, then θ_1 and θ_2 are not orthogonal because

$$\frac{1 \cdot 1}{6} + \frac{1 \cdot (-1)}{5} + \frac{(-2) \cdot 0}{5} = \frac{1}{6} - \frac{1}{5} \neq 0.$$

    However, θ_1 is orthogonal to $\theta_3 = 18\mu_1 - 17\mu_2 - \mu_3$ because

$$\frac{1 \cdot 18}{6} + \frac{1 \cdot (-17)}{5} + \frac{(-2) \cdot (-1)}{5} = 3 - 3.2 + 0.2 = 0.$$

- One can construct families of up to $k-1$ mutually orthogonal contrasts. Such families have several very pleasant properties.

- First, any family of $k-1$ mutually orthogonal contrasts partitions SS_B into $k-1$ separate components,

$$SS_B = SS_{\theta_1} + \cdots + SS_{\theta_{k-1}},$$

  each with one degree of freedom.

- For example, Heyl (1930) collected the following data:

Gold	83	81	76	78	79	72
Platinum	61	61	67	67	64	
Glass	78	71	75	72	74	

This results in the following ANOVA table:

Source	SS	df	MS	F	**p**
Between	565.1	2	282.6	26.1	0.000028
θ_1	29.2	1	29.2	2.7	0.124793
θ_3	535.9	1	535.9	49.5	0.000009
Within	140.8	13	10.8		
Total	705.9	15			

- **Definition 12.3** *Given a family of contrasts, the family rate α' of Type I error is the probability under $H_0 : \mu_1 = \cdots = \mu_k$ of falsely rejecting at least one null hypothesis.*

- A second pleasant property of mutually orthogonal contrasts is that the tests of the contrasts are mutually independent. This allows us to deduce the relation between the significance level(s) of the individual tests and the family rate of Type I error.

  - Let E_r denote the event that $H_0 : \theta_r = 0$ is falsely rejected. Then $P(E_r) = \alpha$ is the rate of Type I error for an individual test.
  - Let E denote the event that at least one Type I error is committed, i.e.,
  $$E = \bigcup_{r=1}^{k-1} E_r.$$
  The family rate of Type I error is $\alpha' = P(E)$.
  - The event that no Type I errors are committed and
  $$E^c = \bigcap_{r=1}^{k-1} E_r^c$$
  and the probability of this event is $P(E^c) = 1 - \alpha'$.
  - By independence,
  $$1 - \alpha' = P(E^c) = P(E_1^c) \times \cdots \times P(E_{k-1}^c) = (1 - \alpha)^{k-1};$$
  hence,
  $$\alpha' = 1 - (1 - \alpha)^{k-1}.$$

- Notice that $\alpha' > \alpha$, i.e., the family rate of Type I error is greater than the rate for an individual test. For example, if $k = 3$ and $\alpha = .05$, then

$$\alpha' = 1 - (1 - .05)^2 = 0.0975.$$

This phenomenon is sometimes called "alpha slippage." To protect against alpha slippage, we usually prefer to specify the family rate of Type I error that will be tolerated and compute a significance level that will ensure the specified family rate. For example, if $k = 3$ and $\alpha' = .05$, then we solve

$$.05 = 1 - (1 - \alpha)^2$$

to obtain a significance level of

$$\alpha = 1 - \sqrt{.95} \doteq 0.0253.$$

Bonferroni t-Tests

- Now suppose that we plan m pairwise comparisons. These comparisons are defined by contrasts $\theta_1, \ldots, \theta_m$ of the form $\mu_i - \mu_j$, not necessarily mutually orthogonal. Notice that each $H_0 : \theta_r = 0$ vs. $H_1 : \theta_r \neq 0$ is a normal 2-sample location problem with equal variances.

- Fact: Under $H_0 : \mu_1 = \cdots = \mu_k$,

$$Z = \frac{\bar{X}_{i\cdot} - \bar{X}_{j\cdot}}{\sqrt{\left(\frac{1}{n_i} + \frac{1}{n_j}\right)\sigma^2}} \sim N(0, 1)$$

and

$$T(\theta_r) = \frac{\bar{X}_{i\cdot} - \bar{X}_{j\cdot}}{\sqrt{\left(\frac{1}{n_i} + \frac{1}{n_j}\right) MS_W}} \sim t(N - k).$$

- The t-test of $H_0 : \theta_r = 0$ is to reject if and only if

$$\mathbf{p} = P_{H_0}\left(|T(\theta_r)| \geq |t(\theta_r)|\right) \leq \alpha,$$

i.e., if and only if

$$|t(\theta_r)| \geq q = \texttt{qt(1-}\alpha\texttt{/2,df=N-k)},$$

where $t(\theta_r)$ denotes the observed value of $T(\theta_r)$.

- Unless the contrasts are mutually orthogonal, we cannot use the multiplication rule for independent events to compute the family rate of Type I error. However, it follows from the *Bonferroni inequality* that

$$\alpha' = P(E) = P\left(\bigcup_{r=1}^{m} E_r\right) \leq \sum_{r=1}^{m} P(E_r) = m\alpha;$$

hence, we can ensure that the family rate of Type I error is no greater than a specified α' by testing each contrast at significance level $\alpha = \alpha'/m$.

12.1.4 Post Hoc Comparisons

- We now consider situations in which we determine that a comparison is of interest *after* inspecting the data. For example, after inspecting Heyl's (1930) data, we might decide to define $\theta_4 = \mu_1 - \mu_3$ and test $H_0 : \theta_4 = 0$ vs. $H_1 : \theta_4 \neq 0$.

Bonferroni t-Tests

- Suppose that only pairwise comparisons are of interest. Because we are testing *after* we have had the opportunity to inspect the data (and therefore to construct the contrasts that appear to be nonzero), we suppose that *all* pairwise contrasts were of interest *a priori*.

- Hence, whatever the number of pairwise contrasts actually tested *a posteriori*, we set

$$m = \binom{k}{2} = k(k-1)/2$$

and proceed as before.

Scheffé F-Tests

- The most conservative of all multiple comparison procedures, Scheffé's procedure is predicated on the assumption that *all possible* contrasts were of interest *a priori*.

- Scheffé's F-test of $H_0 : \theta_r = 0$ vs. $H_1 : \theta_r \neq 0$ is to reject H_0 if and only if

$$f(\theta_r)/(k-1) \geq q = \texttt{qf(1-}\alpha\texttt{,k-1,N-k)},$$

where $f(\theta_r)$ denotes the observed value of the $F(\theta_r)$ defined for the method of planned orthogonal contrasts.

- Fact: No matter how many $H_0 : \theta_r = 0$ are tested by Scheffé's F-test, the family rate of Type I error is no greater than α.

- Example: For Heyl's (1930) data, Scheffé's F-test produces

Source	F	**p**
θ_1	1.3	0.294217
θ_2	25.3	0.000033
θ_3	24.7	0.000037
θ_4	2.2	0.151995

  For the first three comparisons, our conclusions are not appreciably affected by whether the contrasts were constructed before or after examining the data. However, if θ_4 had been planned, we would have obtained $f(\theta_4) = 4.4$ and $\mathbf{p} = 0.056772$.

12.2 The Case of a General Shift Family

12.2.1 The Kruskal-Wallis Test

12.3 The Behrens-Fisher Problem

12.4 Exercises

1. Jean Kerr devoted an entire chapter of *Please Don't Eat the Daisies* (1959) to the subject of dieting, observing that...

   > "Today, with the science of nutrition advancing so rapidly, there is plenty of food for conversation, if for nothing else. We have the Rockefeller diet, the Mayo diet, high-protein diets, low-protein diets, "blitz" diets which feature cottage cheese and something that tastes like very thin sandpaper, and—finally—a liquid diet that duplicates all the rich, nourishing goodness of mother's milk. I have no way of knowing which of these is the most efficacious for losing weight, but there's no question in my mind that as a conversation-stopper the "mother's milk diet" is quite a ways out ahead."

   For her master's thesis, a nutrition student at the University of Arizona decides to compare several weight loss strategies. She recruits 140 moderately obese adult women and randomly assigns each woman to one of the following diets: Rockefeller, Mayo, Atkins (high-protein), a low-protein diet, a blitz diet, a liquid diet, and—as a control—Aunt Jean's marshmallow fudge diet. Each woman is weighed before dieting, asked to follow the prescribed diet for eight weeks, then weighed again. The resulting data will be analyzed using the analysis of variance and related statistical techniques.

   (a) This is a k-sample problem. What is the value of k?

   (b) What null hypothesis is tested by an analysis of variance? (Your answer should specify relations between certain population parameters. Be sure to define these parameters!)

   (c) How many pairwise comparisons are possible?

   (d) The student is especially interested in three pairwise comparisons: Atkins versus low-protein, low-protein versus fudge, and fudge versus liquid. Specify contrasts that correspond to each of these comparisons.

   (e) Are the preceding contrasts orthogonal? Why or why not?

2. As part of her senior thesis, a William & Mary physics major decides to repeat Heyl's (1930) experiment for determining the gravitational

constant using 4 different materials: silver, copper, topaz, and quartz. She plans to test 10 specimens of each material.

(a) Three comparisons are planned:

i. Metal (silver & copper) versus Gem (topaz & quartz)

ii. Silver versus Copper

iii. Topaz versus Quartz

What contrasts correspond to these comparisons? Are they orthogonal? Why or why not? If the desired family rate of Type I error is 0.05, then what significance level should be used for testing the null hypotheses $H_0 : \theta_r = 0$?

(b) After analyzing the data, an ANOVA table is constructed. Complete the table from the information provided.

Source	SS	df	MS	F	**p**
Between					
θ_1					0.001399
θ_2					0.815450
θ_3					0.188776
Within			9.418349		
Total					

(c) Referring to the above table, explain what conclusion the student should draw about each of her planned comparisons.

(d) Assuming that the ANOVA assumption of homoscedasticity is warranted, use the above table to estimate the common population variance.

3. R. R. Sokal observed 25 females of each of three genetic lines (RS, SS, NS) of the fruitfly *Drosophila melanogaster* and recorded the number of eggs laid per day by each female for the first 14 days of her life. The lines labelled RS and SS were selectively bred for resistance and for susceptibility to the insecticide DDT. A nonselected control line is labelled NS. The purpose of the experiment was to investigate the following research questions:

- Do the two selected lines (RS and SS) differ in fecundity from the nonselected line (NS)?

- Does the line selected for resistance (RS) differ in fecundity from the line selected for susceptibility (SS)?

The data are presented in Table 12.1.

RS	12.8	21.6	14.8	23.1	34.6	19.7	22.6	29.6	16.4	20.3
	29.3	14.9	27.3	22.4	27.5	20.3	38.7	26.4	23.7	26.1
	29.5	38.6	44.4	23.2	23.6					
SS	38.4	32.9	48.5	20.9	11.6	22.3	30.2	33.4	26.7	39.0
	12.8	14.6	12.2	23.1	29.4	16.0	20.1	23.3	22.9	22.5
	15.1	31.0	16.9	16.1	10.8					
NS	35.4	27.4	19.3	41.8	20.3	37.6	36.9	37.3	28.2	23.4
	33.7	29.2	41.7	22.6	40.4	34.4	30.4	14.9	51.8	33.8
	37.9	29.5	42.4	36.6	47.4					

Table 12.1: Fecundity of Female Fruitflies

(a) Use side-by-side boxplots and normal probability plots to investigate the ANOVA assumptions of normality and homoscedasticity. Do these assumptions seem plausible? Why or why not?

(b) Construct constrasts that correspond to the research questions framed above. Verify that these constrasts are orthogonal. At what significance level should the contrasts be tested in order to maintain a family rate of Type I error equal to 5%?

(c) Use ANOVA and the method of orthogonal contrasts to construct an ANOVA table. State the null and alternative hypotheses that are tested by these methods. For each null hypothesis, state whether or not it should be rejected. (Use $\alpha = 0.05$ for the ANOVA hypothesis and the significance level calculated above for the contrast hypotheses.)

4. A number of Byzantine coins were discovered in Cyprus. These coins were minted during the reign of King Manuel I, Comnenus (1143–1180). It was determined that $n_1 = 9$ of these coins were minted in an early coinage, $n_2 = 7$ were minted several years later, $n_3 = 4$ were minted in a third coinage, and $n_4 = 7$ were minted in a fourth coinage.

   The silver content (percentage) of each coin was measured, with the results presented in Table 12.2.

1	5.9	6.8	6.4	7.0	6.6	7.7	7.2	6.9	6.2
2	6.9	9.0	6.6	8.1	9.3	9.2	8.6		
3	4.9	5.5	4.6	4.5					
4	5.3	5.6	5.5	5.1	6.2	5.8	5.8		

Table 12.2: Silver Content of Byzantine Coins

(a) Investigate the ANOVA assumptions of normality and homoscedasticity. Do these assumptions seem plausible? Why or why not?

(b) Construct an ANOVA table. State the null and alternative hypotheses tested by this method. Should the null hypothesis be rejected at the $\alpha = 0.10$ level?

(c) Examining the data, it appears that coins minted early in King Manuel's reign (the first two coinages) tended to contain more silver than coins minted later in his reign (the last two coinages). Construct a contrast that is suitable for investigating if this is the case. State appropriate null and alternative hypotheses and test them using Scheffé's F-test for multiple comparisons with a significance level of 5%.

5. R. E. Dolkart and colleagues compared antibody responses in normal and alloxan diabetic mice. Three groups of mice were studied: normal, alloxan diabetic, and alloxan diabetic treated with insulin. Several comparisons are of interest:

   - Does the antibody response of alloxan diabetic mice differ from the antibody response of normal mice?
   - Does the antibody response of alloxan diabetic mice treated with insulin differ from the antibody response of normal mice?
   - Does treating alloxan diabetic mice with insulin affect their antibody response?

   Table 12.3 contains the measured amounts of nitrogen-bound bovine serum albumen produced by the mice.

   (a) Using the above data, investigate the ANOVA assumptions of normality and homoscedasticity. Do these assumptions seem plausible for these data? Why or why not?

Normal	156	282	197	297	116	127	119	29	253	122
	349	110	143	64	26	86	122	455	655	14
Alloxan	391	46	469	86	174	133	13	499	168	62
	127	276	176	146	108	276	50	73		
Alloxan	82	100	98	150	243	68	228	131	73	18
+insulin	20	100	72	133	465	40	46	34	44	

Table 12.3: Antibody Responses of Diabetic Mice

(b) Now transform the data by taking the square root of each measurement. Using the transformed data, investigate the ANOVA assumptions of normality and homoscedasticity. Do these assumptions seem plausible for the transformed data? Why or why not?

(c) Using the transformed data, construct an ANOVA table. State the null and alternative hypotheses tested by this method. Should the null hypothesis be rejected at the $\alpha = 0.05$ level?

(d) Using the transformed data, construct suitable contrasts for investigating the research questions framed above. State appropriate null and alternative hypotheses and test them using the method of Bonferroni t-tests. At what significance level should these hypotheses be tested in order to maintain a family rate of Type I error equal to 5%? Which null hypotheses should be rejected?

Chapter 13

Association

13.1 Categorical Random Variables

13.2 Normal Random Variables

The continuous random variables (X, Y) define a function that assigns a pair of real numbers to each experimental outcome. Let

$$B = [a, b] \times [c, d] \subset \Re^2$$

be a rectangular set of such pairs and suppose that we want to compute

$$P\left((X, Y) \in B\right) = P\left(X \in [a, b], Y \in [c, d]\right).$$

Just as we compute $P(X \in [a, b])$ using the pdf of X, so we compute $P((X, Y) \in B)$ using the *joint probability density function* of (X, Y). To do so, we must extend the concept of area under the graph of a function of one variable to the concept of volume under the graph of a function of two variables.

Theorem 13.1 *Let X be a continuous random variable with pdf f_x and let Y be a continuous random variable with pdf f_y. In this context, f_x and f_y are called the* marginal pdfs *of (X, Y). Then there exists a function $f : \Re^2 \to \Re$, the* joint pdf *of (X, Y), such that*

$$P\left((X, Y) \in B\right) = \text{Volume}_B(f) = \int_a^b \int_c^d f(x, y)\, dy\, dx \tag{13.1}$$

for all rectangular subsets B. If X and Y are independent, then

$$f(x, y) = f_x(x) f_y(y).$$

Remark: If (13.1) is true for all rectangular subsets of $\Re^2$, then it is true for all subsets in the sigma-field generated by the rectangular subsets.

We can think of the joint pdf as a function that assigns an elevation to a point identified by two coordinates, longitude (x) and latitude (y). Noting that topographic maps display elevations via contours of constant elevation, we can describe a joint pdf by identifying certain of its contours, i.e., subsets of $\Re^2$ on which $f(x, y)$ is contant.

Definition 13.1 *Let f denote the joint pdf of (X, Y) and fix $c > 0$. Then*

$$\Big\{(x, y) \in \Re^2 \ : \ f(x, y) = c\Big\}$$

is a contour *of f.*

13.2.1 Bivariate Normal Distributions

Suppose that $X \sim \text{Normal}(0, 1)$ and $Y \sim \text{Normal}(0, 1)$, not necessarily independent. To measure the degree of dependence between X and Y, we consider the quantity $E(XY)$.

- If there is a *positive association* between X and Y, then experimental outcomes that have...
  - positive values of X will tend to have positive values of Y, so XY will tend to be positive;
  - negative values of X will tend to have negative values of Y, so XY will tend to be positive.

  Hence, $E(XY) > 0$ indicates positive association.

- If there is a *negative association* between X and Y, then experimental outcomes that have...
  - positive values of X will tend to have negative values of Y, so XY will tend to be negative;
  - negative values of X will tend to have positive values of Y, so XY will tend to be negative.

  Hence, $E(XY) < 0$ indicates negative association.

If $X \sim \text{Normal}(\mu_x, \sigma_x^2)$ and $Y \sim \text{Normal}(\mu_y, \sigma_y^2)$, then we measure dependence after converting to standard units:

Definition 13.2 *Let $\mu_x = EX$ and $\sigma_x^2 = \text{Var}\, X < \infty$. Let $\mu_y = EY$ and $\sigma_y^2 = \text{Var}\, Y < \infty$. The* population product-moment correlation coefficient *of X and Y is*

$$\rho = \rho(X, Y) = E\left[\left(\frac{X - \mu_x}{\sigma_x}\right)\left(\frac{Y - \mu_y}{\sigma_y}\right)\right].$$

The product-moment correlation coefficient has the following properties:

Theorem 13.2 *If X and Y have finite variances, then*

1. $-1 \leq \rho \leq 1$
2. $\rho = \pm 1$ *if and only if*

$$\frac{Y - \mu_y}{\sigma_y} = \pm \frac{X - \mu_x}{\sigma_x},$$

   in which case Y is completely determined by X.

3. *If X and Y are independent, then $\rho = 0$.*
4. *If X and Y are normal random variables for which $\rho = 0$, then X and Y are independent.*

If $\rho = \pm 1$, then the values of (X, Y) fall on a straight line. If $|\rho| < 1$, then the five population parameters $(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$ determine a unique bivariate normal pdf. The contours of this joint pdf are concentric ellipses centered at (μ_x, μ_y). We use one of these ellipses to display the basic features of the bivariate normal pdf in question.

Definition 13.3 *Let f denote a nondegenerate ($|\rho| < 1$) bivariate normal pdf. The* population concentration ellipse *is the contour of f that contains the four points*

$$(\mu_x \pm \sigma_x, \mu_y \pm \sigma_y).$$

It is not difficult to create an `R` function that plots concentration ellipses. The function `binorm.ellipse` is described in Appendix R and/or can be obtained from the web page for this book/course.

Example 13.1 The following `R` commands produce the population concentration ellipse for a bivariate normal distribution with parameters $\mu_x = 10$, $\mu_y = 20$, $\sigma_x^2 = 4$, $\sigma_y^2 = 16$ and $\rho = 0.5$:

```
> pop <- c(10,20,4,16,.5)
> binorm.ellipse(pop)
```

The ellipse plotted by these commands is displayed in Figure 13.1.

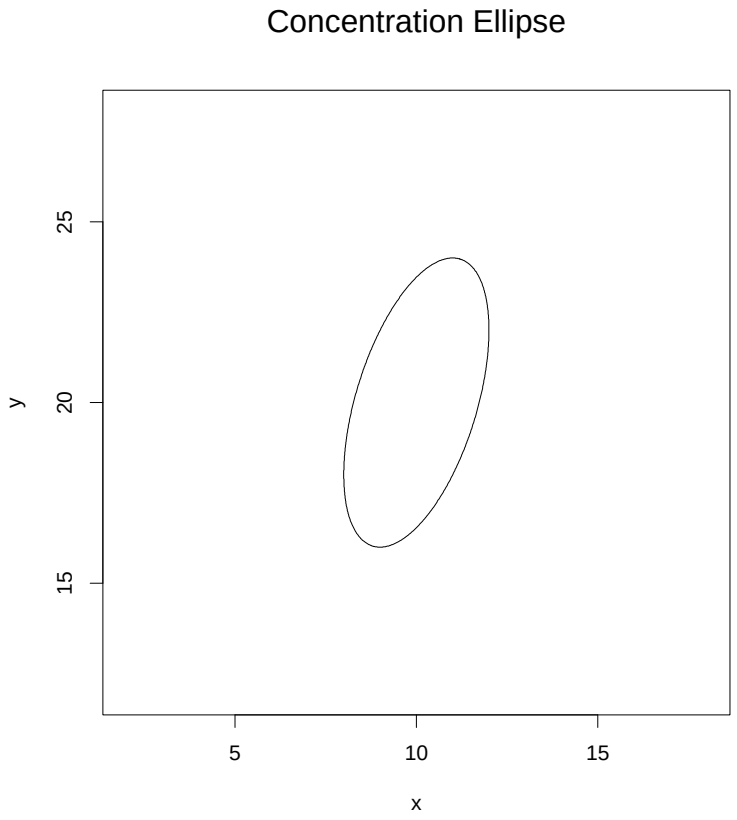

Figure 13.1: The population concentration ellipse for a bivariate normal distribution with parameters $(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho) = (10, 20, 4, 16, 0.5)$.

Unless the population concentration ellipse is circular, it has a unique major axis. The line that coincides with this axis is the *first principal component* of the population and plays an important role in multivariate statistics. We will encounter this line again in Chapter 14.

13.2.2 Bivariate Normal Samples

A bivariate sample is a set of paired observations:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n).$$

We assume that each pair (x_i, y_i) was independently drawn from the same bivariate distribution. Bivariate samples are usually stored in an $n \times 2$ *data matrix*,

$$\begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \\ x_n & y_n \end{bmatrix},$$

and are often displayed by plotting each (x_i, y_i) in the Cartesian plane. The resulting figure is called a *scatter diagram*.

Example 13.2 Twenty students enrolled in Math 351 (Applied Statistics) at the College of William & Mary produced the following scores on two midterm tests:

x	y
87	87
25	57
76	91
84	67
91	67
82	66
94	86
89	74
92	92
76	85
84	75
99	92
92	55
74	74
84	74
94	69
99	98
63	81
82	80
91	85

A scatter diagram of these data is displayed in Figure 13.2. Typically, it is easier to discern patterns by inspecting a scatter diagram than by inspecting a table of numbers. In particular, note the presence of an apaprent outlier.

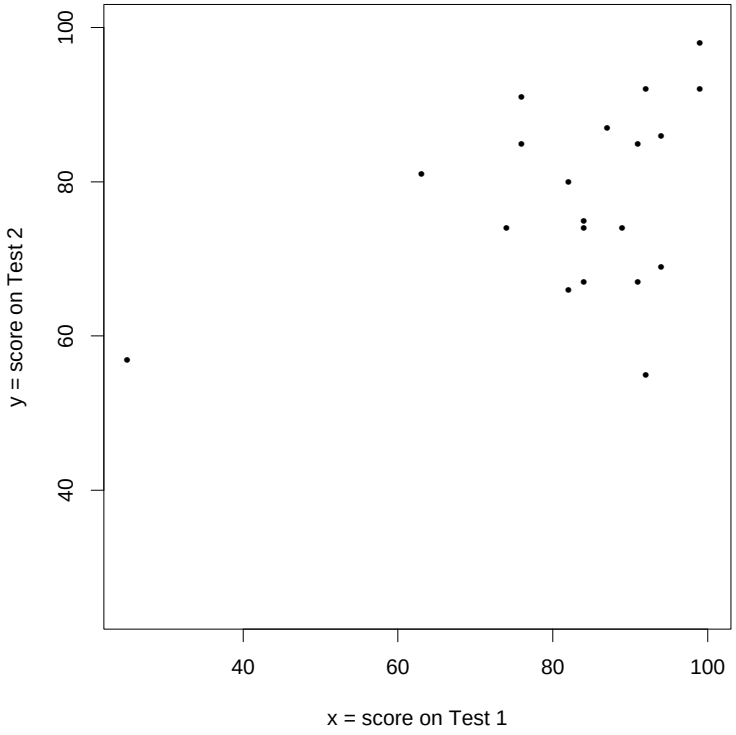

Figure 13.2: A scatter diagram of a bivariate sample. Each point corresponds to a student. The horizontal position of the point represents the student's score on the first midterm test; the vertical position of the point represents the student's score on the second midterm test.

The population from which the bivariate sample in Example 13.2 was drawn is not known, so this sample should not be interpreted as a typical example of a bivariate normal sample. However, it is not difficult to create an `R` function that simulates sampling from a specified bivariate normal population. The function `binorm.sample` is described in Appendix R and/or can be obtained from the web page for this book/course.

Example 13.1 (continued) The following `R` command draws $n = 5$ observations from the previously specified bivariate normal distribution:

```
> binorm.sample(pop,5)
          [,1]     [,2]
[1,] 12.293160 24.07643
[2,] 11.819520 24.13076
[3,] 11.529582 17.28637
[4,]  6.912459 23.39430
[5,] 11.043991 18.12538
```

Notice that `binorm.sample` returns the sample in the form of a data matrix.

Having observed a bivariate normal sample, we inquire how to estimate the five population parameters $(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$. We have already discussed how to estimate the population means (μ_x, μ_y) with the sample means $(\bar{x}, \bar{y})$ and the population variances (σ_x^2, σ_y^2) with the sample variances (s_x^2, s_y^2). The plug-in estimate of ρ is

$$\begin{aligned}
\hat{\rho} &= \frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{x_i - \hat{\mu}_x}{\hat{\sigma}_x}\right)\left(\frac{y_i - \hat{\mu}_y}{\hat{\sigma}_y}\right)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{x_i - \bar{x}}{\sqrt{(n-1)s_x^2/n}}\right)\left(\frac{y_i - \bar{y}}{\sqrt{(n-1)s_y^2/n}}\right)\right] \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left[\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)\right],
\end{aligned}$$

where

$$\hat{\sigma}_x = \sqrt{\widehat{\sigma_x^2}} \quad \text{and} \quad \hat{\sigma}_y = \sqrt{\widehat{\sigma_y^2}}.$$

This quantity is *Pearson's product-moment correlation coefficient coefficient*, usually denoted r.

It is not difficult to create an `R` function that computes the estimates $(\bar{x}, \bar{y}, s_x^2, s_y^2, r)$ from a bivariate data matrix. The function `binorm.estimate` is described in Appendix R and/or can be obtained from the web page for this book/course.

Example 13.1 (continued) The following R commands draw $n = 100$ observations from a bivariate normal distribution with parameters $\mu_x = 10$, $\mu_y = 20$, $\sigma_x^2 = 4$, $\sigma_y^2 = 16$ and $\rho = 0.5$, then estimate the parameters from the sample:

```
> Data <- binorm.sample(pop,100)
> binorm.estimate(Data)
[1]  9.8213430 20.3553502  4.2331147 16.7276819  0.5632622
```

Naturally, the estimates do not equal the estimands because of sampling variation.

Finally, it is not difficult to create an R function that plots a scatter diagram and overlays the *sample concentration ellipse*, i.e., the concentration ellipse constructed using the computed sample quantities $(\bar{x}, \bar{y}, s_x^2, s_y^2, r)$ instead of the unknown population quantities $(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$. The function `binorm.scatter` is described in Appendix R and/or can be obtained from the web page for this book/course.

Example 13.1 (continued) The following R command creates the overlaid scatter diagram displayed in Figure 13.3:

```
> binorm.scatter(Data)
```

When analyzing bivariate data, it is good practice to examine both the scatter diagram and the sample concentration ellipse in order to ascertain how well the latter summarizes the former. A poor summary suggests that the sample may not have been drawn from a bivariate normal distribution, as in Figure 13.4.

13.2.3 Inferences about Correlation

We have already observed that $\hat{\rho} = r$ is the plug-in estimate of ρ. In this section, we consider how to test hypotheses about and construct confidence intervals for ρ.

Given normal random variables X and Y, an obvious question is whether or not they are uncorrelated. To answer this question, we test the null hypothesis $H_0 : \rho = 0$ against the alternative hypothesis $H_1 : \rho \neq 0$. (One might also be interested in one-sided hypotheses and ask, for example, whether or not there is convincing evidence of positive correlation.) We can derive a test from the following fact about the plug-in estimator of ρ.

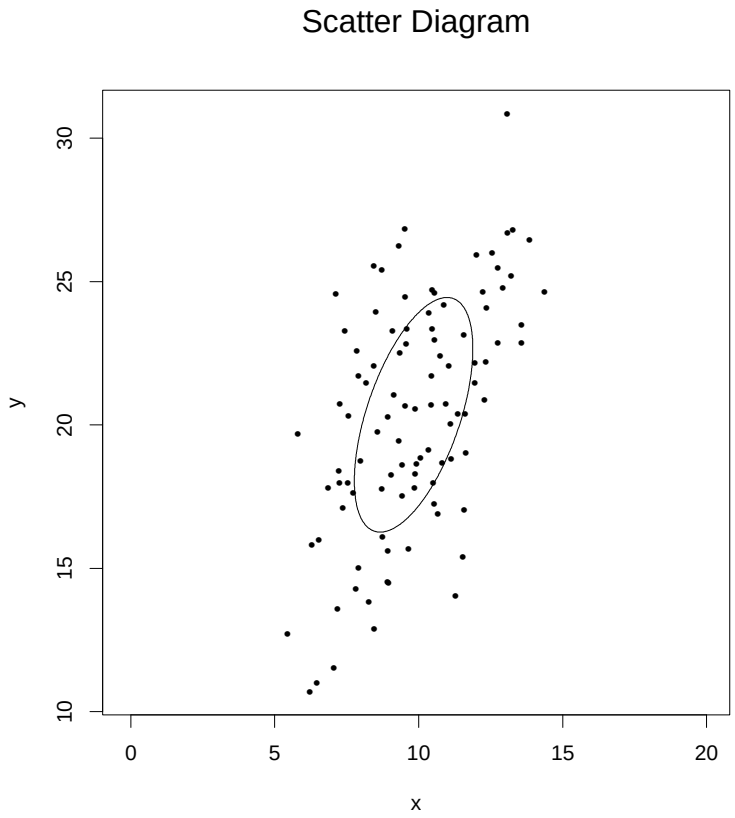

Figure 13.3: A scatter diagram of a bivariate normal sample, with the sample concentration ellipse overlaid.

Theorem 13.3 *Suppose that* (X_i, Y_i), $i = 1, \ldots, n$, *are independent pairs of random variables with a bivariate normal distribution. Let* $\hat{\rho}$ *denote the plug-in estimator of* ρ. *If* X_i *and* Y_i *are uncorrelated, i.e.,* $\rho = 0$, *then*

$$\frac{\hat{\rho}\sqrt{n-2}}{\sqrt{1-\hat{\rho}^2}} \sim t(n-2).$$

Assuming that (X_i, Y_i) have a bivariate normal distribution, Theorem 13.3 allows us to compute a significance probability for testing $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$. Let $T \sim t(n-2)$. Then the probability of observing $|\hat{\rho}| \geq |r|$ under H_0 is

$$\mathbf{p} = P\left(|T| \geq \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}\right)$$

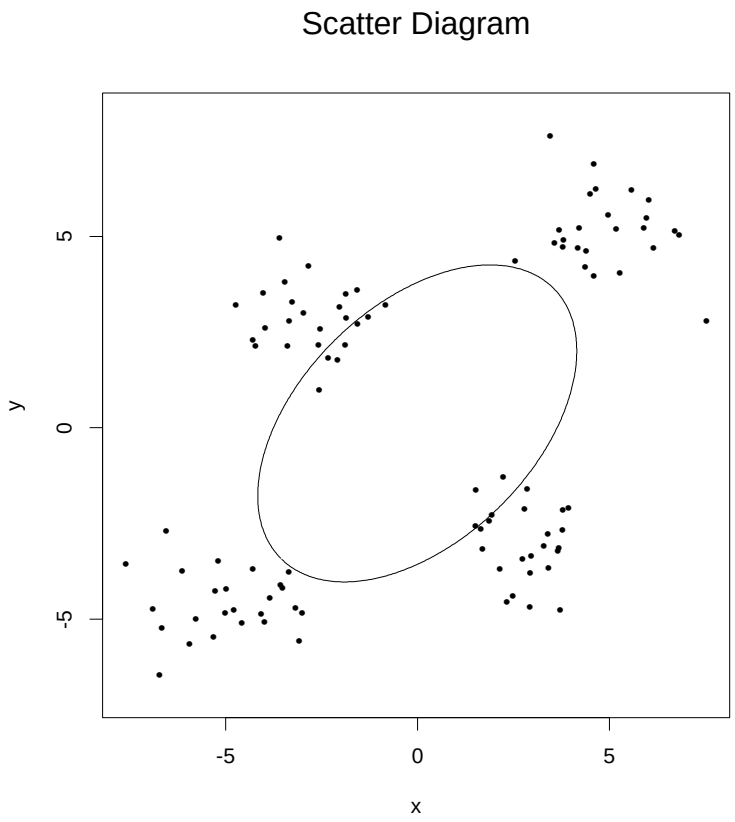

Figure 13.4: A scatter diagram for which the sample concentration ellipse is a poor summary. These data were not drawn from a bivariate normal distribution.

and we reject H_0 if and only if $\mathbf{p} \leq \alpha$. Equivalently, we reject H_0 if and only if (iff)

$$\left| \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \right| \geq q_t \quad \text{iff} \quad \frac{r^2(n-2)}{1-r^2} \geq q_t^2 \quad \text{iff} \quad r^2 \geq \frac{q_t^2}{n-2+q_t^2},$$

where $q_t = \texttt{qt}(1-\alpha/2, n-2)$.

When testing hypotheses about correlation, it is important to appreciate the distinction between statistical significance and material significance. *Strong evidence that an association exists is not the same as evidence of a strong association.* The following examples illustrate the distinction.

Example 13.3 I used `binorm.sample` to draw a sample of $n = 300$ observations from a bivariate normal distribution with a population correlation coefficient of $\rho = 0.1$. This is a rather weak association. I then used `binorm.estimate` to compute a sample correlation coefficient of $r = 0.16225689$. The test statistic is

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = 2.838604$$

and the significance probability is

$$\underline{\mathrm{p}} = 2 * \mathtt{pt}(-2.838604, 298) = 0.004842441.$$

This is fairly decisive evidence that $\rho \neq 0$, but concluding that X and Y are correlated does not warrant concluding that X and Y are strongly correlated.

Example 13.4 I used `binorm.sample` to draw a sample of $n = 10$ observations from a bivariate normal distribution with a population correlation coefficient of $\rho = 0.8$. This is a fairly strong association. I then used `binorm.estimate` to compute a sample correlation coefficient of $r = 0.3759933$. The test statistic is

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = 1.147684$$

and the significance probability is

$$\mathbf{p} = 2 * \mathtt{pt}(-1.147684, 8) = 0.2842594.$$

There is scant evidence that $\rho \neq 0$, despite the fact that X and Y are strongly correlated.

Although testing whether of not $\rho = 0$ is an important decision, it is not the only inference of interest. For example, if we want to construct confidence intervals for ρ, then we need to test $H_0 : \rho = \rho_0$ versus $H_1 : \rho \neq \rho_0$. To do so, we rely on an approximation due to Ronald Fisher. Let

$$\zeta = \frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right)$$

and rewrite the hypotheses as $H_0 : \zeta = \zeta_0$ versus $H_1 : \zeta \neq \zeta_0$. This is sometimes called Fisher's z-transformation. Fisher discovered that

$$\hat{\zeta} = \frac{1}{2}\log\left(\frac{1+\hat{\rho}}{1-\hat{\rho}}\right) \stackrel{\cdot}{\sim} \text{Normal}\left(\zeta, \frac{1}{n-3}\right),$$

which allows us to compute an approximate significance probability. Let $Z \sim \text{Normal}(0,1)$ and set

$$z = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right).$$

Then

$$\mathbf{p} \doteq P\left(|Z| \geq |z-\zeta_0|\sqrt{n-3}\right)$$

and we reject $H_0 : \zeta = \zeta_0$ if and only if $\mathbf{p} \leq \alpha$. Equivalently, we reject $H_0 : \zeta = \zeta_0$ if and only if

$$|z-\zeta_0|\sqrt{n-3} \geq q_z,$$

where $q_z = \texttt{qnorm}(1-\alpha/2)$.

To construct an approximate $(1-\alpha)$-level confidence interval for ρ, we first observe that

$$z \pm \frac{q_z}{\sqrt{n-3}} \tag{13.2}$$

is an approximate $(1-\alpha)$-level confidence interval for ζ. We then use the inverse of Fisher's z-transformation,

$$\rho = \frac{e^{2\zeta}-1}{e^{2\zeta}+1},$$

to transform (13.2) to a confidence interval for ρ.

Example 13.5 Suppose that we draw $n = 100$ observations from a bivariate normal distribution and observe $r = 0.5$. To construct a 0.95-level confidence interval, we use $q_z \doteq 1.96$. First we compute

$$z = \frac{1}{2}\log\left(\frac{1+0.5}{1-0.5}\right) = 0.5493061$$

and

$$z \pm \frac{q_z}{\sqrt{n-3}} \doteq 0.5493061 \pm \frac{1.96}{\sqrt{97}} = (0.350302, 0.7483103)$$

to obtain a confidence interval (a,b) for ζ. The corresponding confidence interval for ρ is

$$\left(\frac{e^{2a}-1}{e^{2a}+1}, \frac{e^{2b}-1}{e^{2b}+1}\right) = (0.3366433, 0.6341398).$$

Notice that the plug-in estimate $\hat{\rho} = r = 0.5$ is *not* the midpoint of this interval.

13.3 Monotonic Association

13.4 Spurious Association

13.5 Exercises

1. Consider the following data matrix:

4.81310497776088	5.50546805210632
3.20790912734096	3.23537831017746
2.03360531141548	1.57466192734915
3.80353555823225	4.0777212868518
3.44874039566775	3.57596515608872
4.02513467455476	4.39110976256498
4.18921274133904	4.62315118989928
1.57765999081644	0.929857871257454
2.55801286069007	2.31628619574412
3.30197349607145	3.36840541617217
3.49344457748324	3.63918641630698
3.84773963203205	4.14023528753161
1.6571339655711	1.04225104421118
2.01676932918443	1.55085225294214
3.26802020797819	3.32038821566353
3.21119453633111	3.24002458012926
3.98834405943784	4.33907997569859
3.39396169865743	3.49849637984759
3.98470335590536	4.33393124338638
2.92484672005844	2.83506761480053
3.24990948234283	2.98840952533401
4.48210022495756	1.24582866569767
2.49246311350902	4.05960045290903
2.5490793094774	3.97953306072058
3.56806772786439	2.53846581953658
2.58341332552653	3.93097742957316
3.00614070448958	3.33315063705718
3.59845899773574	2.49548607350678
3.24798603840268	2.99112968584062
3.27071210738312	2.95899017086906
3.61265049129421	2.47541627084607
3.98487089689919	1.94901712504748
2.92139406397179	3.453000485443
2.10733672639563	4.60425141279254
3.20304499253985	3.05468592240708
1.84295811639769	4.97813922865297
3.11571443259585	3.17818998468951
3.5505950180758	2.56317596269101
3.41454250084746	2.75558327775034
2.6505463184044	3.83603704056258

(a) Do the x values appear to have been drawn from a normal distribution? Why or why not?

(b) Do the y values appear to have been drawn from a normal distribution? Why or why not?

(c) Do the (x, y) values appear to have been drawn from a bivariate normal distribution? Why or why not?

(d) Suggest an explanation for the phenomena observed in (a)–(c). Is this a paradox? How do you think that these (x, y) pairs were obtained?

Hint: Do *not* try to type these data into `R`! They are available electronically. Assuming that the data matrix is stored in a text file named `ex131.dat`, located in the root directory of a diskette, the following command reads the data into the Windows version of `R`:

```
> Data <- matrix(scan("a:\\ex131.dat"),byrow=T,ncol=2)
```

The following `R` commands then create vectors of x and y values:

```
> x <- Data[,1]
> y <- Data[,2]
```

2. Consider the test score data reported in Example 13.2.

(a) Quantify the association between midterm test scores by computing Pearson's product-moment correlation coefficient. Is the association positive or negative?

(b) Examining the scatter diagram displaying in Figure 13.2, one student appears to be an outlier. Omitting the corresponding row of the data matrix, re-compute Pearson's product-moment correlation coefficient. How does the outlier affect the value of r?

Hint: If `Data` is a complete data matrix, then `Data[-17,]` is the same data matrix without row 17.

3. Pearson and Lee reported the following heights (in inches) of eleven pairs of siblings:

sister	brother
69	71
64	68
65	66
63	67
65	70
62	71
65	70
64	73
66	72
59	65
62	66

Assuming that these pairs were drawn from a bivariate normal population, construct a confidence interval for ρ, the population product-moment correlation coefficient, that has a confidence level of approximately 0.90.

Hint: If `x` is the vector of sister heights and `y` is the vector of brother heights (in the same order), then the following `R` command creates the above data matrix:

```
> Data <- cbind(x,y)
```

4. Let $\alpha = 0.05$.

   (a) Suppose that we sample from a bivariate normal distribution with $\rho = 0.5$. Assuming that we observe $r = 0.5$, how large a sample will be needed to reject $H_0 : \rho = 0$ in favor of $H_0 : \rho \neq 0$?

   (b) Suppose that we sample from a bivariate normal distribution with $\rho = 0.1$. Assuming that we observe $r = 0.1$, how large a sample will be needed to reject $H_0 : \rho = 0$ in favor of $H_0 : \rho \neq 0$?

Chapter 14

Simple Linear Regression

One way to quantify the association between two random variables, X and Y, is to quantify the extent to which knowledge of X allows one to predict values of Y. Notice that this approach to association is asymmetric: one variable (conventionally denoted X) is the *predictor variable* and the other variable (conventionally denoted Y) is the *predicted variable*. The predictor variable is often called the *independent variable* and the predicted variable is often called the *dependent variable*. We will eschew this terminology, as it has nothing to do with the probabilistic (in)dependence of events and random variables.

14.1 The Regression Line

Suppose that $Y \sim \text{Normal}(\mu_y, \sigma_y^2)$ and that we want to predict the outcome of an experiment in which we observe Y. If we know μ_y, then the obvious value of Y to predict is $EY = \mu_y$. The expected value of the squared error of this prediction is $E(Y - \mu_y)^2 = \text{Var}\, Y = \sigma_y^2$.

Now suppose that $X \sim \text{Normal}(\mu_x, \sigma_x^2)$ and that we observe $X = x$. Again we want to predict Y. Does knowing $X = x$ allow us to predict Y more accurately? The answer depends on the association between X and Y. If X and Y are independent, then knowing $X = x$ will not help us predict Y. If X and Y are dependent, then knowing $X = x$ should help us predict Y.

Example 14.1 Suppose that we want to predict the adult height to which a male baby will grow. Knowing only that adult male heights are normally distributed, we would predict the average height of this population.

However, if we knew that the baby's father had attained a height of $6'$-$11''$, then we surely would be inclined to revise our prediction and predict that the baby will grow to a greater-than-average height.

When X and Y are normally distributed, the key to predicting Y from $X = x$ is the following result.

Theorem 14.1 *Suppose that (X, Y) have a bivariate normal distribution with parameters $(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$. Then the conditional distribution of Y given $X = x$ is*

$$Y|X = x \sim \text{Normal}\left(\mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x), \left(1 - \rho^2\right)\sigma_y^2\right).$$

Because $Y|X = x$ is normally distributed, the obvious value of Y to predict when $X = x$ is

$$\hat{y}(x) = E(Y|X = x) = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x). \tag{14.1}$$

Interpreting (14.1) as a function that assigns a predicted value of Y to each value of x, we see that the prediction function (14.1) corresponds to a line that passes through the point (μ_x, μ_y) with slope $\rho\,\sigma_y/\sigma_x$. The prediction function (14.1) is the *population regression function* and the corresponding line is the *population regression line*.

The expected squared error of the prediction (14.1) is

$$\text{Var}(Y|X = x) = (1 - \rho^2)\sigma_y^2.$$

Notice that this quantity does not depend on the value of x. If X and Y are strongly correlated, then $\rho \approx \pm 1$, $(1-\rho^2)\sigma_y^2 \approx 0$, and prediction is extremely accurate. If X and Y are uncorrelated, then $\rho = 0$, $(1 - \rho^2)\sigma_y^2 = \sigma_y^2$, and the accuracy of prediction is not improved by knowing $X = x$. These remarks suggest a natural way of interpreting what ρ actually measures: the proportion by which the expected squared error of prediction is reduced by virtue of knowing $X = x$ is

$$\frac{\sigma_y^2 - \left(1 - \rho^2\right)\sigma_y^2}{\sigma_y^2} = \rho^2,$$

the *population coefficient of determination*. Statisticians often express this interpretation by saying that ρ^2 is "the proportion of variation explained by linear regression." Of course, as we emphasized in Section 13.4, this is not an explanation in the sense of articulating a causal mechanism.

Example 14.2 Suppose that $(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho) = (10, 20, 2^2, 4^2, 0.5)$. Then

$$\hat{y}(x) = 20 + 0.5 \cdot \frac{4}{2}(x - 10) = x + 10$$

and $\rho^2 = 0.25$.

Rewriting (14.1), the equation for the population regression line, as

$$\frac{\hat{y}(x) - \mu_y}{\sigma_y} = \rho \, \frac{x - \mu_x}{\sigma_x},$$

we discern an important fact:

Corollary 14.1 *Suppose that (x, y) lies on the population regression line. If x lies z standard deviations above μ_x, then y lies ρz standard deviations above μ_y.*

Example 14.2 (continued) The value $x = 12$ lies $(12 - 10)/2 = 1$ standard deviations above the X-population mean, $\mu_x = 10$. The predicted y-value that corresponds to $x = 12$, $\hat{y}(12) = 12 + 10 = 22$, lies $(22 - 20)/4 = 0.5$ standard deviations above the Y-population mean, $\mu_y = 20$.

Example 14.2 (continued) The 0.90 quantile of X is

$$x = \texttt{qnorm(.9,mean=10,sd=2)} = 12.5631.$$

The predicted y-value that corresponds to $x = 12.5631$ is $\hat{y}(12.5631) = 22.5631$. At what quantile of Y does the predicted y-value lie? The answer is

$$P\left(Y \leq \hat{y}(x)\right) = \texttt{pnorm(22.5631,mean=20,sd=4)} = 0.7391658.$$

At first, most students find the preceding example counterintuitive. If x lies at the 0.90 quantile of X, then should we not predict $\hat{y}(x)$ to lie at the 0.90 quantile of Y? This is a natural first impression, but one that must be dispelled. We begin by considering two familiar situations:

1. Consider the case of a young boy whose father is extremely tall, at the 0.995 quantile of adult male heights. We surely would predict that the boy will grow to be quite tall. But precisely how tall? A father's height does not completely determine his son's height. Height is also

affected by myriad other factors, considered here as chance variation. Statistically speaking, it's more likely that the boy will grow to an adult height slightly shorter than his extremely tall father than that he will grow to be even taller.

2. Consider the case of two college freshman, William and Mary, who are enrolled in an introductory chemistry class of 250 students. On the first midterm examination, Mary attains the 5th highest score and William obtains the 245th highest (5th lowest) score. How should we predict their respective performances on the second midterm examonation? There is undoubtedly a strong, positive correlation between scores on the two tests. We surely will predict that Mary will do quite well on the second test and that William will do rather badly. But how well and how badly? One test score does not completely determine another—if it did, then computing semester grades would be easy! Mary can't do much better on the second test than she did on the first, but she might easily do worse. Statistically speaking, it's likely that she'll rank slightly below 5th on the second test. Likewise, William can't do much worse on the second test than he did on the first. Statistically speaking, it's likely that he'll rank slightly above 245th on the second test.

The phenomenon that we have just described, that experimental units with extreme X quantiles will tend to have less extreme Y quantiles, is purely statistical. It was first discerned by Sir Francis Galton, who called it "regression to mediocrity." Modern statisticians call it *regression to the mean*, or simply *the regression effect*.

Having refined our intuition, we can now explain the regression effect by examining the population concentration ellipse in Figure 14.1. In this bivariate normal population, $\mu_x = \mu_y$ and $\sigma_x^2 = \sigma_y^2$. Given $X = x$, it may seem tempting to predict $Y = x$. But this would be a mistake! Here, $x < \mu_x$ and clearly

$$P\left(Y > x | X = x\right) > \frac{1}{2},$$

so x underpredicts $Y|X = x$. Similarly, if $x > \mu_x$, then x overpredicts $Y|X = x$. The population regression line is the line of conditional expected values, $y = E(Y|X = x)$. Given x, let (x, a) and (x, b) denote the lower and upper points at which the line $X = x$ intersects the population concentration ellipse. As one might guess, it turns out that

$$\hat{y}(x) = E(Y|X = x) = \frac{a + b}{2}.$$

Figure 14.1: The Regression Effect

However, the midpoint of the vertical line segment that connects (x, a) and (x, b) is *not* (x, x).

The correlation coefficient ρ mediates the strength of the regression effect. If $\rho = \pm 1$, then

$$\frac{Y - \mu_x}{\sigma_y} = \pm \frac{X - \mu_x}{\sigma_x}$$

and Y is completely determined by X. In this case there is no regression effect: if x lies z standard deviations above μ_x, then we know that y lies z standard deviations above μ_y. At the other extreme, if $\rho = 0$, then knowing $X = x$ does not reduce the expected squared error of prediction at all. In this case, we regress all the way to the mean: regardless of where x lies, we predict $\hat{y} = \mu_y$.

Yet another important fact can be gleaned from Figure 14.1. Notice that the regression line does *not* coincide with the first principal component. Both lines pass through (μ_x, μ_y), the center of the concentration ellipse. The first principal component coincides with the major axis of the concentration ellipse. In contrast, it can be shown that the regression line is parallel to the line that is tangent to the concentration ellipse at the point $(\mu_x, \mu_y + \sigma_y)$. Except in very special circumstances, the regression line and the first principal component have different slopes. Both are important, as we further

discuss in Section 14.2.

Thus far, we have focussed on predicting Y from $X = x$ in the case that the population concentration ellipse is known. We have done so in order to emphasize that the regression effect is an inherent property of prediction, not a statistical anomaly caused by chance variation. In practice, however, the population concentration ellipse typically is not known and we must rely on the sample concentration ellipse, estimated from bivariate data. This means that we must substitute $(\bar{x}, \bar{y}, s_x^2, s_y^2, r)$ for $(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$. The *sample regression function* is

$$\hat{y}(x) = \bar{y} + r\frac{s_y}{s_x}(x - \bar{x}) \tag{14.2}$$

and the corresponding line is the *sample regression line*. Notice that the slope of the sample regression line does not depend on whether we use plug-in or unbiased estimates of the population variances. The variances affect the regression line through the (square root of) their ratio,

$$\frac{\widehat{\sigma_y^2}}{\widehat{\sigma_x^2}} = \frac{\frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2}{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})^2}{\frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{s_y^2}{s_x^2},$$

which is not affected by the choice of plug-in or unbiased.

Example 14.2 (continued) I used `binorm.sample` to draw a sample of $n = 100$ observations from a bivariate normal distribution with parameters

$$\texttt{pop} = \left(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho\right) = \left(10, 20, 2^2, 4^2, 0.5\right).$$

I then used `binorm.estimate` to compute sample estimates of `pop`, obtaining

$$\begin{aligned} \texttt{est} &= \left(\bar{x}, \bar{y}, s_x^2, s_y^2, r\right) \\ &= (10.0006837, 19.3985929, 4.4512393, 14.1754248, 0.4707309). \end{aligned}$$

The resulting formula for the sample regression line is

$$\hat{y}(x) = \bar{y} + r\frac{s_y}{s_x}(x - \bar{x}) = \bar{y} + 1.784545\,(x - \bar{x}) = 1.55192 + 1.784545x.$$

It is not difficult to create an `R` function that plots a scatter diagram of the sample and overlays both the sample concentration ellipse and the sample regression line. The function `binorm.regress` is described in Appendix R and/or can be obtained from the web page for this book/course. The commands used in this example are as follows:

```
> pop <- c(10,20,4,16,.5)
> Data <- binorm.sample(pop,100)
> est <- binorm.estimate(Data)
> binorm.regress(Data)
```

The scatter diagram created by `binorm.regress` is displayed in Figure 14.2.

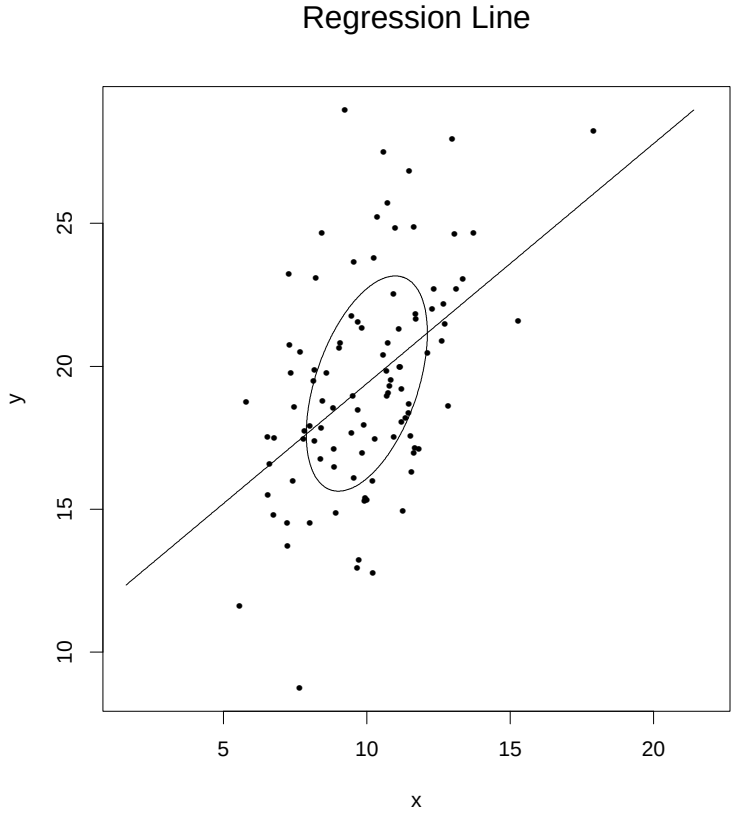

Figure 14.2: Scatter diagram, sample concentration ellipse, and sample regression line of $n = 100$ observations sampled from a bivariate normal distribution. Notice that the sample regression line is *not* the major axis of the sample concentration ellipse.

14.2 The Method of Least Squares

In Section 14.1 we derived the regression line from properties of bivariate normal distributions. Having derived it, we now note that the sample re-

gression line can be computed from *any* set of $n \geq 2$ points $(x_i, y_i) \in \Re^2$ for which the x_i assume more than one distinct value (and therefore $s_x > 0$). In this section, we derive the regression line in this more general setting.

Given points $(x_i, y_i) \in \Re^2$, $i = 1, \ldots, n$, we ask two conceptually distinct questions:

1. What line best *summarizes* the (x, y) pairs?
2. What line best *predicts* values of y from values of x?

We will answer each of these questions by applying the method of least squares. The possible lines are of the form $y = a + bx$. Given a candidate line, we measure the error between the line and each (x_i, y_i), then sum the squared errors from $i = 1, \ldots, n$. The best line is the one that minimizes this sum of squared errors:

$$\min_{a,b} \sum_{i=1}^{n} \left[\text{error} \left(\begin{array}{c} (x_i, y_i) \\ y = a + bx \end{array} \right) \right]^2 \tag{14.3}$$

The distinction between (1) summary) and (2) prediction lies in how we define error.

Figure 14.3: Perpendicular Errors for Summary

To define the line that best summarizes the (x, y) pairs, it is natural to define the error between a point and a line as the Euclidean distance from the

point to the line. This is found by measuring the length of the perpendicular line segment that connects them, as in Figure 14.3. Thus,

$$\begin{array}{c}\text{summary}\\ \text{error}\end{array}\left(\begin{array}{c}(x_i, y_i)\\ y = a + bx\end{array}\right) = \begin{array}{c}\text{perpendicular}\\ \text{distance}\end{array}\left(\begin{array}{c}(x_i, y_i)\\ y = a + bx\end{array}\right).$$

Using this definition of error, the solution of Problem 14.3 is the major axis of the sample concentration ellipse, the first principal component of the sample. We emphasize: *the first principal component is used for summary, not prediction.*

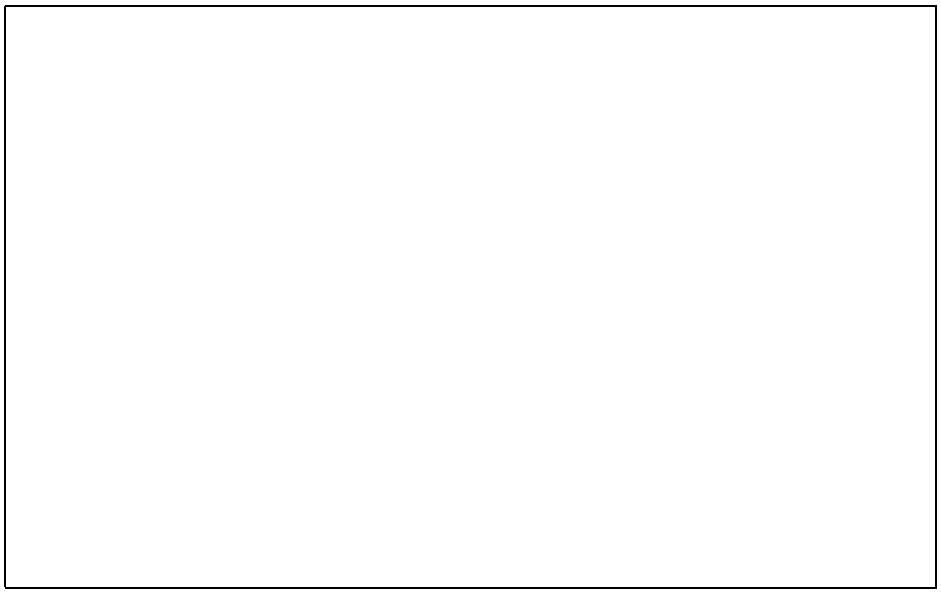

Figure 14.4: Vertical Errors for Prediction

In contrast, to define the line that best predicts y values from x values, it is natural to define the error between a point (x_i, y_i) and a line $y = a + bx$ as the difference between the observed value $y = y_i$ and the predicted value

$$y = \hat{y}\,(x_i) = a + bx_i.$$

The difference $y_i - \hat{y}(x_i)$ is a *residual error* and the absolute difference $|y_i - \hat{y}(x_i)|$ is the length of the vertical line segment that connects (x_i, y_i) and $y = a + bx$, as in Figure 14.4. Using this definition of error, the solution of Problem 14.3 is the sample regression line. We emphasize: *the regression line is used for prediction, not summary.*

The remainder of this section provides a more detailed exposition of the squared error approach to prediction. Let

$$\mathrm{SS}(a,b) = \sum_{i=1}^{n} (y_i - a - bx_i)^2 ,$$

the sum of the squared residual errors that result from the prediction function $\hat{y}(x) = a + bx$. The method of least squares chooses (a, b) to minimize $\mathrm{SS}(a, b)$. Before analyzing this problem, we first consider an easier problem. If we knew $\{y_1, \ldots, y_n\}$ but not the corresponding $\{x_1, \ldots, x_n\}$, then it would be impossible to measure errors associated with prediction functions that involve x. In this situation we would be forced to restrict attention to prediction functions of the form $\hat{y} = a$, which corresponds to restricting attention to lines with zero slope. The method of least squares then chooses a to minimize

$$\sum_{i=1}^{n} (y_i - a)^2 = \mathrm{SS}(a, 0).$$

Theorem 14.2 *The value of a that minimizes $\mathrm{SS}(a, 0)$ is $a = \bar{y}$.*

Proof We can conclude that $\mathrm{SS}(a, 0)/n$ is minimal when $a = \bar{y}$ by applying part (2) of Theorem 6.1 to the empirical distribution of $\{y_1, \ldots, y_n\}$; however, it is instructive to verify this conclusion by direct calculation:

$$\begin{aligned}
\mathrm{SS}(a,0) &= \sum_{i=1}^{n} (y_i - a)^2 = \sum_{i=1}^{n} (y_i - \bar{y} + \bar{y} - a)^2 \\
&= \sum_{i=1}^{n} (y_i - \bar{y})^2 + \sum_{i=1}^{n} 2\,(y_i - \bar{y})\,(\bar{y} - a) + \sum_{i=1}^{n} (\bar{y} - a)^2 \\
&= (n-1)s_y^2 + 2\,(\bar{y} - a)\left[\sum_{i=1}^{n} y_i - n\bar{y}\right] + n\,(\bar{y} - a)^2 \\
&= (n-1)s_y^2 + n\,(\bar{y} - a)^2
\end{aligned}$$

The second term in this expression is the only term that involves a. It achieves its minimal value of zero when $a = \bar{y}$. □

For future reference, we define the *total sum of squares* to be

$$\mathrm{SS}_T = \mathrm{SS}\,(\bar{y}, 0) = \sum_{i=1}^{n} (y_i - \bar{y})^2 = (n-1)s_y^2.$$

This is the smallest squared error possible when predicting y without information about x.

Now we consider the problem of finding the line $y = a + bx$ that best predicts values of y from values of x. The method of least squares chooses (a, b) to minimize $\mathrm{SS}(a, b)$. Let (a^*, b^*) denote the minimizing values of (a, b) and define the *error sum of squares* to be

$$\mathrm{SS}_E = \mathrm{SS}\left(a^*, b^*\right).$$

Because we have not restricted attention to $b = 0$, $\hat{y}(x) = a^* + b^*x$ must predict at least as well as $\hat{y} = \bar{y}$. Thus,

$$\mathrm{SS}_E = \mathrm{SS}\left(a^*, b^*\right) \leq \mathrm{SS}\left(\bar{y}, 0\right) = \mathrm{SS}_T.$$

We have already stated that $y = a^* + b^*x$ is the sample regression line. We can verify that statement by a calculation that resembles the proof of Theorem 14.2.

Theorem 14.3 *Let $(x_i, y_i) \in \Re^2$, $i = 1, \ldots, n$, be a set of (x, y) pairs with at least two distinct values of x. Let*

$$b^* = r\frac{s_y}{s_x} \quad \textit{and} \quad a^* = \bar{y} - b^*\bar{x}.$$

Then

$$\mathrm{SS}\left(a^*, b^*\right) \leq \mathrm{SS}(a, b)$$

for all choices of (a, b).

Proof First, write

$$\begin{aligned} \mathrm{SS}(a, b) &= \sum_{i=1}^{n} \left(y_i - a - bx_i\right)^2 = \sum_{i=1}^{n} \left(y_i - \bar{y} + \bar{y} - b\bar{x} + b\bar{x} - a - bx_i\right)^2 \\ &= \sum_{i=1}^{n} \left[\left(y_i - \bar{y}\right) + \left(\bar{y} - b\bar{x} - a\right) - b\left(x_i - \bar{x}\right)\right]^2. \end{aligned}$$

Expanding the square in this expression results in six terms. The three squared terms are:

$$\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2 = (n-1)s_y^2,$$

$$\begin{aligned}\sum_{i=1}^{n}(\bar{y}-b\bar{x}-a)^2 &= n(\bar{y}-b\bar{x}-a)^2,\\ \sum_{i=1}^{n}(-b)^2(x_i-\bar{x})^2 &= b^2\sum_{i=1}^{n}(x_i-\bar{x})^2 = b^2(n-1)s_x^2.\end{aligned}$$

The three cross-product terms are:

$$\begin{aligned}&\sum_{i=1}^{n}2(y_i-\bar{y})(\bar{y}-b\bar{x}-a) = 2(\bar{y}-b\bar{x}-a)\sum_{i=1}^{n}(y_i-\bar{y})\\ &\quad = 2(\bar{y}-b\bar{x}-a)\left[\sum_{i=1}^{n}y_i-n\bar{y}\right]=0,\\ &\sum_{i=1}^{n}2(y_i-\bar{y})(-b)(x_i-\bar{x}) = -2b\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\\ &\quad = -2b(n-1)s_xs_y\frac{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{s_xs_y} = -2b(n-1)s_xs_yr,\\ &\sum_{i=1}^{n}2(\bar{y}-b\bar{x}-a)(-b)(x_i-\bar{x}) = -2b(\bar{y}-b\bar{x}-a)\sum_{i=1}^{n}(x_i-\bar{x})=0.\end{aligned}$$

Hence,

$$\begin{aligned}\mathrm{SS}(a,b) &= (n-1)s_y^2+n(\bar{y}-b\bar{x}-a)^2+b^2(n-1)s_x^2-2b(n-1)s_xs_yr\\ &= n(\bar{y}-b\bar{x}-a)^2+(n-1)\left[b^2s_x^2-2bs_xrs_y+r^2s_y^2\right]\\ &\quad -(n-1)r^2s_y^2+(n-1)s_y^2\\ &= n(\bar{y}-b\bar{x}-a)^2+(n-1)\left[bs_x-rs_y\right]^2+\left(1-r^2\right)(n-1)s_y^2.\end{aligned}$$

The third term in this expression does not involve b or a. The second term achieves its minimal value of zero when $b = rs_y/s_x = b^*$. The first term is the only term that involves a. Whatever the value of b, the first term achieves its minimal value of zero when $a = \bar{y} - b\bar{x}$. Hence, for $b = b^*$, the minimizing value of a is $a = \bar{y} - b^*\bar{x} = a^*$. □

The total sum of squares, SS_T, measures the prediction error from $\hat{y} = \bar{y}$. The error sum of squares,

$$\mathrm{SS}_E = \mathrm{SS}(a^*,b^*) = \sum_{i=1}^{n}\left[y_i-(\bar{y}-b^*\bar{x})-b^*x_i\right]^2$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} \left[y_i - \bar{y} - b^* \left(x_i - \bar{x} \right) \right]^2 = \sum_{i=1}^{n} \left[\left(y_i - \bar{y} \right) - r \frac{s_y}{s_x} \left(x_i - \bar{x} \right) \right]^2 \\
&= \sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2 - 2r \frac{s_y}{s_x} \sum_{i=1}^{n} \left(x_i - \bar{x} \right) \left(y_i - \bar{y} \right) + r^2 \frac{s_y^2}{s_x^2} \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2 \\
&= (n-1) s_y^2 - 2r s_y^2 (n-1) \frac{\frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \bar{x} \right) \left(y_i - \bar{y} \right)}{s_x s_y} + r^2 s_y^2 (n-1) \\
&= (n-1) s_y^2 - 2(n-1) s_y^2 r^2 + r^2 (n-1) s_y^2 \\
&= (n-1) s_y^2 \left(1 - r^2 \right) \\
&= \left(1 - r^2 \right) \mathrm{SS}_T,
\end{aligned}
$$

measures the prediction error from the sample regression line. Now we define the *regression sum of squares* to be the sum of the squared differences between the two predictions,

$$
\begin{aligned}
\mathrm{SS}_R &= \sum_{i=1}^{n} \left[\hat{y} - \hat{y} \left(x_i \right) \right]^2 = \sum_{i=1}^{n} \left[\bar{y} - \bar{y} - r \frac{s_y}{s_x} \left(x_i - \bar{x} \right) \right]^2 \\
&= r^2 \frac{s_y^2}{s_x^2} \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2 = r^2 s_y^2 (n-1) = r^2 \mathrm{SS}_T.
\end{aligned}
$$

The three sums of squares $(\mathrm{SS}_R, \mathrm{SS}_E, \mathrm{SS}_T)$ are precisely analogous to the three sums of squares $(\mathrm{SS}_B, \mathrm{SS}_W, \mathrm{SS}_T)$ that arise in the analysis of variance and they enjoy an identical property:

$$
\mathrm{SS}_R + \mathrm{SS}_E = r^2 \mathrm{SS}_T + \left(1 - r^2 \right) \mathrm{SS}_T = \mathrm{SS}_T
$$

This is the Pythagorean Theorem in n-dimensional Euclidean space! The points

$$
A = \begin{bmatrix} \bar{y} \\ \vdots \\ \bar{y} \end{bmatrix}, \quad B = \begin{bmatrix} \bar{y} - r \frac{s_y}{s_x} \left(x_1 - \bar{x} \right) \\ \vdots \\ \bar{y} - r \frac{s_y}{s_x} \left(x_n - \bar{x} \right) \end{bmatrix}, \quad C = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}
$$

are the vertices of a right triangle in $\Re^n$. The right angle occurs at vertex B. The squared Euclidean distances of the sides that meet at B are

$$
d^2(A, B) = \mathrm{SS}_R \quad \text{and} \quad d^2(B, C) = \mathrm{SS}_E
$$

and the squared Euclidean distance of the hypotenuse is

$$d^2(A, C) = \mathrm{SS}_T,$$

so

$$d^2(A, B) + d^2(B, C) = \mathrm{SS}_R + \mathrm{SS}_E = \mathrm{SS}_T = d^2(A, C).$$

To quantify the extent to which knowledge of x improves our ability to predict y, we measure the proportion by which the squared error of prediction is reduced when we use the sample regression line instead of the constant prediction $\hat{y} = \bar{y}$. This proportion is just

$$\frac{\mathrm{SS}\,(\bar{y}, 0) - \mathrm{SS}(a, b)}{\mathrm{SS}\,(\bar{y}, 0)} = \frac{\mathrm{SS}_T - \mathrm{SS}_E}{\mathrm{SS}_T} = \frac{\mathrm{SS}_R}{\mathrm{SS}_T} = \frac{r^2 \mathrm{SS}_T}{\mathrm{SS}_T} = r^2,$$

the sample coefficient of determination. Again, we conclude that the square of Pearson's product-moment correlation coefficient measures the proportion of variation "explained" by simple linear regression.

Example 14.2 (continued) For the bivariate sample displayed in Figure 14.2, the total sum of squares is

$$\mathrm{SS}_T = (n-1)s_y^2 = 99 \cdot 14.1754248 = 1403.3671$$

and the coefficient of determination is

$$r^2 = 0.4707309^2 = 0.2215876.$$

Hence, the regression sum of squares is

$$\mathrm{SS}_R = r^2 \mathrm{SS}_T = 0.2215876 \cdot 1403.367 = 310.9688$$

and the error sum of squares is

$$\mathrm{SS}_E = \mathrm{SS}_T - \mathrm{SS}_R = 1403.3671 - 310.9688 = 1092.3983.$$

14.3 Computation

A bivariate sample consists of $2n$ numbers. However, all of the quantities used in the preceding sections can be computed from just six fundamental quantities:

$$n \quad \sum_{i=1}^n x_i \quad \sum_{i=1}^n y_i \quad \sum_{i=1}^n x_i^2 \quad \sum_{i=1}^n y_i^2 \quad \sum_{i=1}^n x_i y_i$$

These quantities are used by many calculators. One reason that they are so convenient is that they are easily incremented as new (x, y) pairs are observed.

Example 14.2 (continued) For the bivariate sample displayed in Figure 14.2, the six fundamental quantities are as follows:

$$n = 100 \quad \sum_{i=1}^{n} x_i = 1000.068 \quad \sum_{i=1}^{n} y_i = 1939.859$$
$$\sum_{i=1}^{n} x_i^2 = 10442.04 \quad \sum_{i=1}^{n} y_i^2 = 39033.91 \quad \sum_{i=1}^{n} x_i y_i = 19770.1$$

Now suppose that we draw another (x, y) pair from the same population, say $(8.9, 13.5)$. Then the new sample has the following fundamental quantities:

$$n = 100 + 1 \qquad \sum_{i=1}^{n} x_i^2 = 10442.04 + 8.9^2$$
$$\sum_{i=1}^{n} x_i = 1000.068 + 8.9 \qquad \sum_{i=1}^{n} y_i^2 = 39033.91 + 13.5^2$$
$$\sum_{i=1}^{n} y_i = 1939.859 + 13.5 \qquad \sum_{i=1}^{n} x_i y_i = 19770.1 + 8.9 \cdot 13.5$$

Three useful quantities are easily computed from the six fundamental quantities:

$$\begin{aligned}
t_{xx} &= \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x}) = \sum_{i=1}^{n} \left(x_i^2 - 2\bar{x}x_i + \bar{x}^2 \right) \\
&= \sum_{i=1}^{n} x_i^2 - 2\bar{x} \sum_{i=1}^{n} x_i + n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 \\
&= \sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right)^2 \\
t_{yy} &= \sum_{i=1}^{n} (y_i - \bar{y})(y_i - \bar{y}) = \sum_{i=1}^{n} y_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} y_i \right)^2 \\
t_{xy} &= \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} \left(y_i^2 - \bar{y}x_i - \bar{x}y_i + \bar{x}\bar{y} \right) \\
&= \sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i - \bar{x} \sum_{i=1}^{n} y_i + n\bar{x}\bar{y} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} \\
&= \sum_{i=1}^{n} x_i y_i - \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right) \left(\sum_{i=1}^{n} y_i \right)
\end{aligned}$$

These quantities are useful because all of the important quantities derived in the preceding sections are easily computed from them. Here are the formulas:

1. Sample variances:

$$s_x^2 = \frac{t_{xx}}{n-1} \qquad s_y^2 = \frac{t_{yy}}{n-1}$$

2. Pearson's correlation coefficient:

$$\begin{aligned} r &= \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{s_x s_y} = \frac{t_{xy}}{\sqrt{t_{xx}}\sqrt{t_{yy}}} \\ r^2 &= \frac{t_{xy}^2}{t_{xx}t_{yy}} \end{aligned}$$

3. Sample regression coefficients:

$$\begin{aligned} b^* &= r\frac{s_y}{s_x} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{s_x^2} = \frac{t_{xy}}{t_{xx}} \\ a^* &= \bar{y} - b^*\bar{x} = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{t_{xy}}{t_{xx}}\frac{1}{n}\sum_{i=1}^{n} x_i \end{aligned}$$

4. Sums of squares:

$$\begin{aligned} \mathrm{SS}_T &= \sum_{i=1}^{n}(y_i-\bar{y})^2 = t_{yy} \\ \mathrm{SS}_R &= r^2\mathrm{SS}_T = \frac{t_{xy}^2}{t_{xx}t_{yy}}t_{yy} = \frac{t_{xy}^2}{t_{xx}} \\ \mathrm{SS}_E &= \mathrm{SS}_T - \mathrm{SS}_R = t_{yy} - \frac{t_{xy}^2}{t_{xx}} \end{aligned}$$

14.4 The Simple Linear Regression Model

Let $x_1, \ldots, x_n$ be a list of real numbers for which $s_x > 0$. Suppose that:

1. Associated with each x_i is a random variable

$$Y_i \sim \text{Normal}\left(\mu_i, \sigma^2\right).$$

Notice that the Y_i have a common population variance $\sigma^2 > 0$. This is analogous to the homoscedasticity assumption of the analysis of variance.

2. The population means μ_i satisfy the linear relation

$$\mu_i = \beta_0 + \beta_1 x_i$$

for some $\beta_0, \beta_1 \in \Re$. The population parameters (β_0, β_1) are called the population regression coeffcients.

These assumptions define the *simple linear regression model*. Suppose that we sample from a bivariate normal distribution, then condition on the observed values $x_1, \ldots, x_n$. It follows from Theorem 14.1 that this is a special case of the simple linear regression model in which

$$\begin{aligned} \beta_1 &= \rho\frac{\sigma_y}{\sigma_x}, \\ \beta_0 &= \mu_y - \rho\frac{\sigma_y}{\sigma_x}\mu_x = \mu_y - \beta_1\mu_x, \\ \sigma^2 &= \left(1-\rho^2\right)\sigma_y^2. \end{aligned}$$

The simple linear regression model has three unknown parameters. The method of least squares estimates (β_0, β_1) by

$$\begin{aligned} \hat{\beta}_1 &= b^* = r\frac{s_y}{s_x} = \frac{t_{xy}}{t_{xx}}, \\ \hat{\beta}_0 &= a^* = \bar{y} - \hat{\beta}_1\bar{x}. \end{aligned}$$

These are also the plug-in estimates of (β_0, β_1), and the plug-in estimate of σ^2 is

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 = \frac{1}{n}\mathrm{SS}_E.$$

We proceed to explore some properties of the corresponding estimators. These properties are consequences of the following key facts:

Theorem 14.4 *Under the assumptions of the simple linear regression model, the random variables $\hat{\beta}_1$ and SS_E are independent and satisfy*

$$\hat{\beta}_1 \sim \mathrm{Normal}\left(\beta_1, \frac{\sigma^2}{t_{xx}}\right) \tag{14.4}$$

$$\frac{SS_E}{\sigma^2} \sim \chi^2(n-2). \tag{14.5}$$

It follows from (14.4) that $E\hat{\beta}_1 = \beta_1$, and consequently that

$$\begin{aligned} E\hat{\beta}_0 &= E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i - \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(Y_i - \hat{\beta}_1 x_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} (\beta_0 + \beta_1 x_i - \beta_1 x_i) = \frac{1}{n}\sum_{i=1}^{n} \beta_0 = \beta_0. \end{aligned}$$

Thus, $(\hat{\beta}_0, \hat{\beta}_1)$ are unbiased estimators of (β_0, β_1). Furthermore, it follows from (14.5) and Corollary 5.1 that $E(\mathrm{SS}_E/\sigma^2) = n-2$. Hence, $E[\mathrm{SS}_E/(n-2)] = \sigma^2$ and

$$\mathrm{MS}_E = \frac{1}{n-2}\mathrm{SS}_E$$

is an unbiased estimator of σ^2.

Converting (14.4) to standard units results in

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/t_{xx}}} \sim \mathrm{Normal}(0,1). \tag{14.6}$$

Dividing (14.6) by (14.5), it follows from Definition 5.7 that

$$\frac{\left(\hat{\beta}_1 - \beta_1\right)/\sqrt{\sigma^2/t_{xx}}}{\sqrt{\frac{\mathrm{SS}_E}{\sigma^2}/(n-2)}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\mathrm{MS}_E/t_{xx}}} \sim t(n-2).$$

This fact allows us to construct confidence intervals for β_1. Given α, we first compute the critical value

$$q_t = \mathtt{qt}(1-\alpha/2, n-2).$$

Then

$$\hat{\beta}_1 \pm q_t\sqrt{\frac{\mathrm{MS}_E}{t_{xx}}}$$

is a $(1-\alpha)$-level confidence interval for β_1.

Remark: It may be helpful to write

$$\begin{aligned} \frac{\mathrm{MS}_E}{t_{xx}} &= \frac{(1-r^2)\,\mathrm{SS}_T/(n-2)}{(n-1)s_x^2} = \frac{(1-r^2)\,(n-1)s_y^2/(n-2)}{(n-1)s_x^2} \\ &= \left(1-r^2\right)\frac{s_y^2}{s_x^2}/(n-2). \end{aligned}$$

Example Need an example here.

Next we consider how to test $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$. This is an important decision because rejecting $H_0 : \beta_1 = 0$ means that we are convinced that values of x help us to predict values of y. Furthermore, if we sampled from a bivariate normal population, then

$$\beta_1 = \rho \frac{\sigma_y}{\sigma_x} = 0$$

if and only if $\rho = 0$. Because normal random variables X and Y are independent if and only if they are uncorrelated, the null hypothesis $H_0 : \beta_1 = 0$ is equivalent to the null hypothesis that X and Y are independent.

If $\beta_1 = 0$, then

$$\frac{\hat{\beta}_1}{\sqrt{\mathrm{MS}_E / t_{xx}}} \sim t(n-2).$$

Hence, the significance probability for testing $H_0 : \beta_1 = 0$ is

$$\mathbf{p} = P\left(|T| \geq \left| \frac{\hat{\beta}_1}{\sqrt{\mathrm{MS}_E / t_{xx}}} \right| \right),$$

where the random variable $T \sim t(n-2)$, and we reject $H_0 : \beta_1 = 0$ if and only if $\mathbf{p} \leq \alpha$. Equivalently, we reject $H_0 : \beta_1 = 0$ if and only if we observe

$$\left| \frac{\hat{\beta}_1}{\sqrt{\mathrm{MS}_E / t_{xx}}} \right| \geq q_t,$$

where q_t is the critical value defined above. Notice that

$$\begin{aligned}
\frac{\hat{\beta}_1}{\sqrt{\mathrm{MS}_E / t_{xx}}} &= \frac{t_{xy}/t_{xx}}{\sqrt{\mathrm{MS}_E / t_{xx}}} = \frac{t_{xy}}{\sqrt{t_{xx}}} \frac{1}{\sqrt{\mathrm{SS}_E/(n-2)}} \\
&= \frac{t_{xy}}{\sqrt{t_{xx}}\sqrt{t_{yy}}} \frac{\sqrt{t_{yy}}\sqrt{n-2}}{\sqrt{t_{yy} - t_{xy}^2/t_{xx}}} \\
&= r \frac{\sqrt{n-2}}{\sqrt{1 - t_{xy}^2/(t_{xx} t_{yy})}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}},
\end{aligned}$$

so this is the same t-test that we described in Section 13.2.3 for testing $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$.

It follows from Theorem 5.5 that

$$\left(\frac{\hat{\beta}_1}{\sqrt{\mathrm{MS}_E/t_{xx}}}\right)^2 \sim F(1, n-2).$$

Hence, an F-test that is equivalent to the t-test derived in the preceding paragraph rejects $H_0 : \beta_1 = 0$ if and only if we observe

$$(n-2)\frac{r^2}{1-r^2} \geq q_F,$$

where the critical value q_F is defined by

$$q_F = \texttt{qf}(1-\alpha, 1, n-2).$$

Equivalently, we reject $H_0 : \beta_1 = 0$ if and only if the significance probability

$$\mathbf{p} = P\left(F \geq (n-2)\frac{r^2}{1-r^2}\right) \leq \alpha,$$

where the random variable $F \sim F(1, n-2)$. The results of the F-test of $H_0 : \beta_1 = 0$ are traditionally presented in the form of an ANOVA table:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F-Test Statistic	**p**-Value
Regression	$r^2\mathrm{SS}_T$	1	$r^2\mathrm{SS}_T$	$(n-2)\frac{r^2}{1-r^2}$	**p**
Error	$(1-r^2)\mathrm{SS}_T$	$n-2$	$\frac{1-r^2}{n-2}\mathrm{SS}_T$		
Total	SS_T				

Example Need an example here.

Although equivalent, the t-test and F-test of $H_0 : \beta_1 = 0$ each enjoy certain advantages. The former is more flexible, as it is easily adapted to test 1-sided hypotheses. The F-test is more readily generalized to testing a variety of hypotheses that naturally arise when studying more complicated regression models.

14.5 Regression Diagnostics

14.6 Exercises

1. According to Stanford University Professor Claude M. Steele (Not just a test, *The Nation*, May 3, 2004, page 40),

   > "The SAT, for example, correlates .42 with freshman grades... This means that it measures about 18 percent of the characteristics, whatever they are, that determine freshman grades."

   Comment on this passage. Do you agree with Professor Steele's interpretation of what $r = 0.42$ means?

2. Suppose that (X, Y) have a bivariate normal distribution with parameters $(5, 3, 1, 4, 0.5)$. Compute the following quantities:

   (a) $P(Y > 6)$
   (b) $E(Y|X = 6.5)$
   (c) $P(Y > 6|X = 6.5)$

3. Assume that the population of all sister-brother heights has a bivariate normal distribution and that the data in Exercise 13.5.3 were sampled from this population. Use these data in the following:

   (a) Consider the population of all sister-brother heights. Estimate the proportion of all brothers who are at least $5'\ 10''$.
   (b) Suppose that Carol is $5'\ 1''$. Predict her brother's height.
   (c) Consider the population of all sister-brother heights for which the sister is $5'\ 1''$. Estimate the proportion of these brothers who are at least $5'\ 10''$.

4. Assume that the population of all sister-brother heights has a bivariate normal distribution and that the data in Exercise 13.5.2 were sampled from this population. Use these data in the following:

   (a) Compute the sample coefficient of determination, the proportion of variation "explained" by simple linear regression.
   (b) Let $\alpha = 0.05$. Do these data provide convincing evidence that knowing a sister's height (x) helps one predict her brother's height (y)?
   (c) Construct a 0.90-level confidence interval for the slope of the population regression line for predicting y from x.

(d) Suppose that you are planning to conduct a more comprehensive study of sibling heights. Your goal is to better estimate the slope of the population regression line for predicting y from x. If you want to construct a 0.95-level confidence interval of length 0.1, then how many sister-brother pairs should you plan to observe?

Hint:

$$\frac{\mathrm{MS}_E}{t_{xx}} = \left(1 - r^2\right) \frac{s_y^2}{s_x^2}/(n-2).$$

5. A class of 35 students took two midterm tests. Jack missed the first test and Jill missed the second test. The 33 students who took both tests scored an average of 75 points on the first test, with a standard deviation of 10 points, and an average of 64 points on the second test, with a standard deviation of 12 points. The scatter diagram of their scores is roughly ellipsoidal, with a correlation coefficient of $r = 0.5$.

   Because Jack and Jill each missed one of the tests, their professor needs to guess how each would have performed on the missing test in order to compute their semester grades.

   (a) Jill scored 80 points on Test 1. She suggests that her missing score on Test 2 be replaced with her score on Test 1, 80 points. What do you think of this suggestion? What score would you advise the professor to assign?

   (b) Jack scored 76 points on Test 2, precisely one standard deviation above the Test 2 mean. He suggests that his missing score on Test 1 be replaced with a score of 85 points, precisely one standard deviation above the Test 1 mean. What do you think of this suggestion? What score would you advise the professor to assign?

6. In the athletics event known as the shot put, male competitors "put" the "shot," a 16-pound metal ball. (Female competitors use a smaller shot.) In the United States, high school male competitors put a 12-pound shot, then graduate to the 16-pound shot used in NCAA, US-ATF, and IAAF competition. In its August 2002 "Stat Corner," the respected athletics periodical *Track & Field News* proclaimed an "Inverse Relationship Between 12 & 16lb Shots:"

   > "A look at the accompanying all-time Top 11 lists for high schoolers with the 12lb shot—11 because there have been 11

of them over 70 [feet]—and for U.S. men with the 16 sends two messages to aspiring prep putters:

- If you're not very good in high school, don't worry about it; few of the big guys were either.
- If you're great in high school, that may be about as good as you'll ever get.

"The numbers are astounding. We'll leave it to a technical expert to figure out why..."

The numbers follow.[1] Do you agree with T&FN's two messages?

ALL-TIME HIGH SCHOOL 70-FOOTERS			
	12	*16*	*16–12*
1. Michael Carter '79	81-3.5	71-4.75	–9-10.75
2. Brent Noon '90	76-2	70-5.75	–5-8.25
3. Arnold Campbell '84	74-10.5	64-3	–10-7.5
4. Charles Moye '87	72-8	57-1	–15-7
5. Sam Walker '68	72-3.25	66-9.5	–5-5.75
6. Jesse Stuart '70	71-11i	68-11.5i	–2-11.5
7. Roger Roesler '96	71-2	61-6.25	–11-7.75
8. Kevin Bookout '02	71-1.5	(too early still)	
9. Doug Lane '68	70-11	66-11.25	–3-11.75
10. Dennis Black '91	70-7	68-10	–1-9
11. Ron Semkiw '72	70-1.75	70-0.5	–0-1.25

[1] Perhaps the most astounding number is Michael Carter's prodigious heave of 81-3.50, arguably the most formidable record in all of track and field. Carter broke an 11-year-old record by *nine feet*! He went on to a sensational college career at SMU, winning the NCAA championship and a silver medal at the 1984 Olympic Games. He then opted for a career in professional football, becoming an All-Pro defensive lineman for the NFL Champion San Francisco 49er's.

ALL-TIME U.S. TOP 11			
	16	*12*	*16–12*
1. Randy Barnes '90	75-10.25	66-9.5	+9-0.75
2. Brian Oldfield '75	75-0	58-10	+16-2
3. John Brenner '87	73-10.75	64-5.5	+9-5.25
4. Adam Nelson '02	73-10.25	63-2.25	+10-8
5. Kevin Toth '02	72-9.75	58-11	+13-10.75
6. George Woods '74	72-3i	60-11	+11-4
6. Dave Laut '82	72-3	65-9	+6-6
6. John Godina '99	72-3	64-1.25	+8-1.75
9. Gregg Trafalis '92	72-1.5	57-0	+15-1.5
10. Terry Albritton '76	71-8.5	67-9	+3-11.5
11. Andy Bloom '00	71-7.25	64-2.5	+7-4.74

Chapter 15

Simulation-Based Inference

Appendix R

A Statistical Programming Language

R.1 Introduction

R.1.1 What is R?

In the 1970s, researchers at AT&T Bell Laboratories developed `S`, a high-level statistical programming language that became popular with academic statisticians. Bell Labs subsequently licensed `S` to a company that added a variety of capabilities, creating the commercial product `S-Plus`. `R` is yet another implementation of `S`. The R Project for Statistical Computing is an ongoing effort by a group of statisticians to extend and improve `R`.

`R` is free, Open Source software, that can be downloaded in compiled or source code form. It runs on a variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows, and MacOS. The primary web site for information about `R` is:

`http://www.r-project.org/`

R.1.2 Why Use R?

This question encompasses several issues. First, there is the question of what role statistical software is to play in the course. Introductory statistics courses may use software in different ways. Once upon a time, many instructors (myself included) avoided using software in the first semester. The rationale for this approach is that one should begin one's study of statistics by focussing on basic concepts and learn what the computer is doing

before one uses the computer to do it. Unfortunately, this approach condemns one to analyzing fairly trivial data sets, and even then calculating by hand and/or calculator quickly becomes extremely tedious. As a result, this approach has fallen from favor.

At the other end of the spectrum, many introductory statistics courses use statistics packages like Minitab, SPSS, or SAS to analyze data. Such packages are extremely useful and every statistician should have some familiarity with at least one such package. However, if one begins to rely on such packages too quickly, the package may be viewed as a black box and the student may never really learn what that black box is doing.

There are many different ways to introduce the subject of statistics, and no one way is best for all students. This book is intended for students who want to *understand* what is going on inside the black box procedures available in so many statistics packages. This intention determines the use that we shall make of the computer. We will strive for an intermediate approach, in which the computer is used to relieve the tedium of calculation, but in which the student is obliged to tell the computer what intermediate steps need to be performed in order to obtain the desired output. Such an approach requires a high-level, interactive programming language. Several such languages are available, but `S-Plus` and `R` have achieved the greatest popularity within the statistics community. Acquiring some familiarity with `S-Plus` and/or `R` will benefit students who continue to study statistics and/or analyze data in the future.

Why `R` instead of `S-Plus`? For most of the examples in this book, `R` and `S-Plus` are interchangeable—the same commands work for both. But `R` has two compelling advantages. First, `R` is available for certain operating systems for which `S-Plus` is not, e.g., MacOS. Second, `R` is free! As a result, students who begin using `R` in this course can be confident that they will always have access to `R`.

R.1.3 Installing R

To efficiently download software, documentation, etc., you should use a nearby CRAN (Comprehensive R Archive Network) mirror site, e.g., Statlib at Carnegie Mellon University:

`http://lib.stat.cmu.edu/R/CRAN/`

Most students will want to install `R` in compiled form by downloading executable binary files. On-line documentation and several manuals are in-

cluded, although you may find it easier to get started using the examples provided in this book.

R.1.4 Learning About R

R is far too complicated to learn in one (or even several) lessons. I doubt that any one person—including the R developers—knows everything about R! But don't be intimidated: *the best way to learn* R *is to just start using* R*.* And, the best time to use R is when you're trying to accomplish a specific task. Try to learn bits and pieces of R as they're introduced in the text and/or you develop an interest in a specific capability.

Of course, it's hard to learn anything without documentation. The material in this book, both the examples scattered throughout various chapters to illustrate various statistical methods and the tutorial material in this appendix, is a good way to get started. Once you know the name of one R function, you can learn more about it and discover related functions using various utilities included in your R installation. If you're using the Windows version of R then you can start by exploring the Help menu in RGui, which will lead you to manuals, search utilities, and web pages. I tend to use R `functions (text)` for help on specific functions.

R.2 Using R

R is an interpreted language, designed to be used interactively. The user is prompted to issue a command as follows:

```
>
```

The cursor-up key allows the user to recall previous commands.

Except for a few standard arithmetic operations, R accomplishes things by executing various functions. For example, to exit R one executes the `quit` function:

```
> q()
```

When you quit, R will inquire if you want to "Save workspace image?" If you answer `yes` (`y`), then all of the objects in your current workspace, e.g., any data sets and functions that you created, will be saved and restored the next time that you start R.

R.2.1 Vectors

R can store and manipulate a variety of data objects, the most basic of which is a vector. In R a vector is an ordered list of numbers, i.e., a list of numbers with a designated first element, second element, etc. Vectors can be created in various ways. In each of the following examples, the created vector is assigned the name x.

Note that R has a large number of built-in functions. Assigning their names to user-created objects will mask the built-in functions. For this reason certain simple names, e.g., c and t, should be avoided.

Example R.1 To enter a list of numbers from the keyboard, use the concatenate function:

```
> x <- c(20,5,15,18,5,13,1)
```

Notice that this can be done recursively, e.g.,

```
> x <- c(20,5,15)
> x <- c(x,18,5)
> x <- c(x,13,1)
```

To display the vector, type its name:

```
> x
```

Just typing

```
> c(20,5,15,18,5,13,1)
```

causes R to display the vector without saving it for future use.

Example R.2 To read a list of numbers from an ascii text file, say data.txt, use the scan function. In most situations, you will need to specify the complete path of data.txt. How one does this depends on which operating system your computer uses.

For example, suppose that you are using the Windows version of R and data.txt resides in the directory c:\Courses\Math351. Then the following command will read the contents of data.txt into the vector x:

```
> x <- scan("c:\\Courses\\Math351\\data.txt")
```

Notice that the single slashes in the path name must be entered as double slashes in R.

Example R.3 Several functions are useful for creating sequences of numbers, e.g.,

```
> x <- seq(from=1,to=15,by=2)
> x <- rep(1,times=10)
```

Consecutive integers are especially easy, e.g.,

```
x <- 11:20
```

Example R.4 `R` has a variety of functions for generating pseudorandom samples.[1]

To draw 10 numbers from a uniform distribution on $(0, \pi)$:

```
> x <- runif(10,min=0,max=pi)
```

To draw 20 numbers from a normal distribution with mean 5 and standard deviation 1.5:

```
> x <- rnorm(20,mean=5,sd=1.5)
```

To simulate rolling a fair die 30 times:

```
> die <- 1:6
> x <- sample(x=die,size=30,replace=T)
```

A subset of a vector can be identified by a vector of index values. For example, to extract the 2nd, 3rd, and 5th elements of the vector x, one might type:

```
> k <- c(2,3,5)
> x[k]
```

To extract the other elements, just type:

```
> x[-k]
```

One may wish to rearrange the elements, e.g.,

```
> y <- sort(x)
```

The preceding command is equivalent to

```
> y <- x[order(x)]
```

[1]The precise meanings of the phrases that follow are explained in Chapters 3–5.

R.2.2 R is a Calculator!

R provides a variety of arithmetical operations and mathematical functions. These operations/functions have been vectorized, i.e., they work on entire vectors, not just individual numbers. Several examples follow.

First, let's create two vectors:

```
> x <- 10:20
> y <- seq(from=1.8,to=2.2,length=length(x))
```

Now, each of the following is a valid R command:

```
> x+100
> x-20
> x*10
> x/10
> x^2
> sqrt(x)
> exp(x)
> log(x)
> x+y
> x-y
> x*y
> x/y
> x^y
```

R.2.3 Some Statistics Functions

R provides hundreds of functions that perform or facilitate a variety of statistical analyses. Most R functions are not used in this book. (You may enjoy discovering and using some of them on your own initiative.) Tables R.1 and R.2 list some of the R functions that are used.

R.2.4 Creating New Functions

The full power of R emerges when one writes one's own functions. To illustrate, I've written a short function named `Edist` that computes the Euclidean distance between two vectors. When I type `Edist`, R displays the function:

```
> Edist
```

Function	Distribution	Section
`pgeom`	Geometric	4.2
`phyper`	Hypergeometric	4.2
`pbinom`	Binomial	4.4
`punif`	Uniform	5.3
`pnorm`	Normal	5.4
`pchisq`	Chi-Squared	5.5
`pt`	Student's t	5.5
`pf`	Fisher's F	5.5

Table R.1: Some `R` functions that evaluate the cumulative distribution function (cdf) for various families of probability distributions. The prefix `p` designates a cdf function; the remainder of the function name specifies the distribution. For the analogous quantile functions, use the prefix `q`, e.g., `qnorm`. To evaluate the analogous probability mass function (pmf) or probability density function (pdf), use the prefix `d`, e.g., `dnorm`. To generate a pseudorandom sample, use the prefix `r`, e.g., `rnorm`.

```
function(u,v){
return(sqrt(sum((u-v)^2)))
}
>
```

`Edist` has two arguments, `u` and `v`, which it interprets as vectors of equal length. `Edist` computes the vector of differences, squares each difference, sums the squares, then takes the square root of the sum to obtain the distance. Finally, it returns the computed distance. I could have written `Edist` as a sequence of intermediate steps, but there's no need to do so.

I might have created `Edist` in any of the following ways:

Example R.5

```
> Edist <- function(u,v){ return(sqrt(sum((u-v)^2))) }
>
```

Example R.6

```
> Edist <- function(u,v){
```

Function	Used to Compute/Display
`sum`	sample sum
`mean`	sample mean
`median`	sample median
`var`	sample variance
`quantile`	sample quantile(s)
`summary`	several useful quantities
`plot.ecdf`	empirical cdf
`boxplot`	box plot(s)
`qqnorm`	normal probability plot
`plot`, `density`	kernel estimate of pdf

Table R.2: Some `R` functions that compute or display useful information about one or more univariate samples. See Chapter 7.

```
+ return(sqrt(sum((u-v)^2)))
+ }
>
```

Notice that `R` recognizes that the command creating `Edist` is not complete and provides continuation prompts (`+`) until it is.

Examples R.5 and R.6 are useful for very short functions, but not for anything complicated. Be warned: ff you mistype and `R` cannot interpret what you did type, then `R` ignores the command and you have to retype it. Using the cursor-up key to recall what you typed may help, but for anything complicated it is best to create a permanent file that you can edit. This can be done within `R` or outside of `R`.

Example R.7 To create moderately complicated functions in `R`, use the `edit` function. For example, I might start by typing

```
> Edist <- function(u,v){u-v}
```

This creates an R object called Edist, but not the Edist that we want—this Edist returns the vector of differences.[2] So, I use edit to modify Edist.[3] This process is initiated with the command

```
> Edist <- edit(Edist)
```

After making and saving the desired changes to Edist, I close the editor, thereby returning control to R. R checks the edited version of Edist: if R can interpret the edited version, then R replaces the previous version with the edited version; if R cannot interpret the edited version, e.g., because of typographical errors, then R issues an error message and retains the previous version. Fortunately, R also retains a temporary version of whatever modifications I attempted to make, so I have another chance at getting it right. To access the temporary version, I type

```
> Edist <- edit()
```

Note that I should *not* retype

```
> Edist <- edit(Edist)
```

as this command returns to the original unedited version and discards whatever changes I attempted to make.

Example R.8 Objects created in R can be lost, e.g., if one forgets to save one's workspace image when one quits R. For this reason, I prefer to create my R functions outside of R. To accomplish this, I first use a text editor to create an ascii text file that contains whatever R commands I want to execute, e.g., the command that creates Edist. For example, I might use the Windows notepad editor to create an ascii text file that contains the following:

```
Edist <- function(u,v)
{
return(sqrt(sum((u-v)^2)))
}
```

[2]Using the return function is good practice, but often unnecessary. An R function will automatically return the last quantity that it computes.

[3]Each installation has a default editor. For the Windows operating system, the default editor is the Windows notepad editor.

Let's suppose that I call this file `myRfcns.txt` and save it in the directory `c:\Courses\Math351`. Then, I can start `R` and use the `source` function to execute the commands in `myRfcns.txt`:

```
> source("c:\\Courses\\Math351\\myRfcns.txt")
```

To check that I succeeded in creating `Edist`, I can produce a list of all the objects in my workspace by typing

```
> objects()
```

R.2.5 Simulating Termite Foraging

R.2.6 Exploring Bivariate Normal Data

In Sections 13.2 and 14.1, we explored the structure of bivariate normal data using five `R` functions:

```
binorm.ellipse
binorm.sample
binorm.estimate
binorm.scatter
binorm.regress
```

These functions are not part of your `R` installation—I created them for this book/course. To use them, download the ascii text file `binorm.R` from the web page for this book/course, then `source` its contents into your `R` workspace. For example, suppose that you have a Windows operating system and that you save `binorm.R` in the directory `c:\Courses\Math351`. Then the following command instructs `R` to execute the commands in `binorm.R` that create the five `binorm` functions:

```
> source("c:\\Courses\\Math351\\binorm.R")
```

Tables R.3–R.7 reproduce the commands in `binorm.R`. Notice that the `#` symbol is used to insert comments, as `R` ignores lines that begin with `#`.

```
binorm.ellipse <- function(pop) {
#
#  This function plots the concentration ellipse of a bivariate
#  normal distribution.  The 5 bivariate normal parameters are
#  specified in the vector pop in the following order:
#    mean of X, mean of Y, variance of X, variance of Y,
#    correlation of (X,Y).
#  For example:  pop <- c(0,0,1,4,.5)
#
n <- 628
m <- matrix(pop[1:2],nrow=2)
off <- pop[5] * sqrt(pop[3]*pop[4])
C <- matrix(c(pop[3],off,off,pop[4]),nrow=2)
E <- eigen(C,symmetric=T)
a <- 0:n/100
X <- cbind(cos(a),sin(a))
X <- X %*% diag(sqrt(E$values)) %*% t(E$vectors)
X <- X + matrix(rep(1,n+1),ncol=1) %*% t(m)
xmin <- min(X[,1])
xmax <- max(X[,1])
ymin <- min(X[,2])
ymax <- max(X[,2])
dif <- max(xmax-xmin,ymax-ymin)
xlim <- c(m[1]-dif,m[1]+dif)
ylim <- c(m[2]-dif,m[2]+dif)
par(pty="s")
plot(X,type="l",xlab="x",ylab="y",xlim=xlim,ylim=ylim)
title("Concentration Ellipse")
}
```

Table R.3: The command that creates the `R` function `binorm.ellipse`, described in Section 13.2. This command is included in the file `binorm.R`.

```
binorm.sample <- function(pop,n) {
#
#  This function returns a sample of n observations drawn from a
#  bivariate normal distribution.  The 5 bivariate normal
#  parameters are specified in the vector pop in the following
#  order: mean of X, mean of Y, variance of X, variance of Y,
#  correlation of (X,Y).  For example:  pop <- c(0,0,1,4,.5)
#  The sample is returned in the form of an n-by-2 data matrix,
#  each row of which is an observed value of (X,Y).
#
m <- matrix(pop[1:2],nrow=2)
off <- pop[5] * sqrt(pop[3]*pop[4])
C <- matrix(c(pop[3],off,off,pop[4]),nrow=2)
E <- eigen(C,symmetric=T)
Data <- matrix(rnorm(2*n),nrow=n)
Data <- Data %*% diag(sqrt(E$values)) %*% t(E$vectors)
Data + matrix(rep(1,n),nrow=n) %*% t(m)
}
```

Table R.4: The command that creates the `R` function `binorm.sample`, described in Section 13.2. This command is included in the file `binorm.R`.

```
binorm.estimate <- function(Data) {
#
#  This function estimates bivariate normal parameters from a
#  bivariate data matrix.  Each row of the n-by-2 matrix Data
#  contains a single observation of (X,Y).  The function returns
#  a vector of 5 estimated parameters: mean of X, mean of Y,
#  variance of X, variance of Y, correlation of (X,Y).
#
n <- nrow(Data)
m <- c(sum(Data[,1]),sum(Data[,2]))/n
v <- c(var(Data[,1]),var(Data[,2]))
z1 <- (Data[,1]-m[1])/sqrt(v[1])
z2 <- (Data[,2]-m[2])/sqrt(v[2])
r <- sum(z1*z2)/(n-1)
c(m,v,r)
}
```

Table R.5: The command that creates the `R` function `binorm.estimate`, described in Section 13.2. This command is included in the file `binorm.R`.

```
binorm.scatter <- function(Data) {
#
#  This function produces a scatter diagram of the bivariate data
#  contained in the n-by-2 data matrix Data.  It also superimposes
#  the sample concentration ellipse.
#
n <- 628
xmin <- min(Data[,1])
xmax <- max(Data[,1])
xmid <- (xmin+xmax)/2
ymin <- min(Data[,2])
ymax <- max(Data[,2])
ymid <- (ymin+ymax)/2
dif <- max(xmax-xmin,ymax-ymin)/2
xlim <- c(xmid-dif,xmid+dif)
ylim <- c(ymid-dif,ymid+dif)
par(pty="s")
plot(Data,xlab="x",ylab="y",xlim=xlim,ylim=ylim)
title("Scatter Diagram")
v <- binorm.estimate(Data)
m <- matrix(v[1:2],nrow=2)
off <- v[5] * sqrt(v[3]*v[4])
C <- matrix(c(v[3],off,off,v[4]),nrow=2)
E <- eigen(C,symmetric=T)
a <- 1:n/100
Y <- cbind(cos(a),sin(a))
Y <- Y %*% diag(sqrt(E$values)) %*% t(E$vectors)
Y <- Y + matrix(rep(1,n),nrow=n) %*% t(m)
lines(Y)
}
```

Table R.6: The command that creates the `R` function `binorm.scatter`, described in Section 13.2. This command is included in the file `binorm.R`.

```
binorm.regress <- function(Data) {
#
#  This function produces a scatter diagram of the bivariate data
#  contained in the n-by-2 data matrix Data.  It also superimposes
#  the sample concentration ellipse and the regression line.
#
n <- 628
xmin <- min(Data[,1])
xmax <- max(Data[,1])
xmid <- (xmin+xmax)/2
ymin <- min(Data[,2])
ymax <- max(Data[,2])
ymid <- (ymin+ymax)/2
dif <- max(xmax-xmin,ymax-ymin)/2
xlim <- c(xmid-dif,xmid+dif)
ylim <- c(ymid-dif,ymid+dif)
par(pty="s")
plot(Data,xlab="x",ylab="y",xlim=xlim,ylim=ylim)
title("Regression Line")
v <- binorm.estimate(Data)
m <- matrix(v[1:2],nrow=2)
off <- v[5] * sqrt(v[3]*v[4])
C <- matrix(c(v[3],off,off,v[4]),nrow=2)
E <- eigen(C,symmetric=T)
a <- 0:n/100
Y <- cbind(cos(a),sin(a))
Y <- Y %*% diag(sqrt(E$values)) %*% t(E$vectors)
Y <- Y + matrix(rep(1,n+1),ncol=1) %*% t(m)
lines(Y)
x <- xlim[1] + (2*dif*(0:n))/n
slope <- v[5] * sqrt(v[4]/v[3])
y <- v[2] + slope*(x-v[1])
Y <- cbind(x,y)
Y <- Y[Y[,2] < ymax,]
Y <- Y[Y[,2] > ymin,]
lines(Y)
}
```

Table R.7: The command that creates the R function `binorm.regress`, described in Section 14.1. This command is included in the file `binorm.R`.