

An Introduction to Statistical Inference and Its Applications

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

August 11, 2006

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

1	Experiments	7
1.1	Examples	7
1.1.1	Spinning a Penny	8
1.1.2	The Speed of Light	9
1.1.3	Termite Foraging Behavior	11
1.2	Randomization	14
1.3	The Importance of Probability	17
1.4	Games of Chance	19
1.5	Exercises	24
2	Mathematical Preliminaries	27
2.1	Sets	27
2.2	Counting	31
2.3	Functions	38
2.4	Limits	39
2.5	Exercises	40
3	Probability	45
3.1	Interpretations of Probability	45
3.2	Axioms of Probability	46
3.3	Finite Sample Spaces	55
3.4	Conditional Probability	60
3.5	Random Variables	72
3.6	Case Study: Padrolling in Milton Murayama's <i>All I asking for is my body</i>	80
3.7	Exercises	85

4	Discrete Random Variables	93
4.1	Basic Concepts	93
4.2	Examples	94
4.3	Expectation	100
4.4	Binomial Distributions	110
4.5	Exercises	115
5	Continuous Random Variables	121
5.1	A Motivating Example	121
5.2	Basic Concepts	125
5.3	Elementary Examples	128
5.4	Normal Distributions	132
5.5	Normal Sampling Distributions	136
5.6	Exercises	141
6	Quantifying Population Attributes	145
6.1	Symmetry	145
6.2	Quantiles	147
6.2.1	The Median of a Population	151
6.2.2	The Interquartile Range of a Population	151
6.3	The Method of Least Squares	152
6.3.1	The Mean of a Population	153
6.3.2	The Standard Deviation of a Population	153
6.4	Exercises	154
7	Data	159
7.1	The Plug-In Principle	160
7.2	Plug-In Estimates of Mean and Variance	163
7.3	Plug-In Estimates of Quantiles	164
7.3.1	Box Plots	166
7.3.2	Normal Probability Plots	168
7.4	Kernel Density Estimates	171
7.5	Case Study: Are Forearm Lengths Normally Distributed? . .	174
7.6	Exercises	180
8	Lots of Data	183
8.1	Averaging Decreases Variation	185
8.2	The Weak Law of Large Numbers	187
8.3	The Central Limit Theorem	190

8.4 Exercises	196
9 Inference	199
9.1 A Motivating Example	200
9.2 Point Estimation	202
9.2.1 Estimating a Population Mean	202
9.2.2 Estimating a Population Variance	204
9.3 Heuristics of Hypothesis Testing	204
9.4 Testing Hypotheses About a Population Mean	214
9.4.1 One-Sided Hypotheses	218
9.4.2 Formulating Suitable Hypotheses	219
9.4.3 Statistical Significance and Material Significance	223
9.5 Set Estimation	224
9.5.1 Sample Size	227
9.5.2 One-Sided Confidence Intervals	228
9.6 Exercises	229
10 1-Sample Location Problems	235
10.1 The Normal 1-Sample Location Problem	238
10.1.1 Point Estimation	238
10.1.2 Hypothesis Testing	239
10.1.3 Interval Estimation	243
10.2 The General 1-Sample Location Problem	245
10.2.1 Hypothesis Testing	245
10.2.2 Point Estimation	248
10.2.3 Interval Estimation	249
10.3 The Symmetric 1-Sample Location Problem	250
10.4 A Case Study from Neuropsychology	250
10.5 Exercises	251
11 2-Sample Location Problems	257
11.1 The Normal 2-Sample Location Problem	259
11.1.1 Known Variances	261
11.1.2 Unknown Common Variance	263
11.1.3 Unknown Variances	266
11.2 The Case of a General Shift Family	270
11.3 The Symmetric Behrens-Fisher Problem	270
11.4 Case Study: Etruscan versus Italian Head Breadth	270
11.5 Exercises	274

12 k-Sample Location Problems	283
12.1 The Case of a Normal Shift Family	284
12.1.1 The Fundamental Null Hypothesis	284
12.1.2 Testing the Fundamental Null Hypothesis	285
12.1.3 Planned Comparisons	291
12.1.4 Post Hoc Comparisons	301
12.2 The Case of a General Shift Family	303
12.2.1 The Kruskal-Wallis Test	303
12.3 The Behrens-Fisher Problem	303
12.4 Exercises	304
13 Association	309
13.1 Categorical Random Variables	309
13.2 Normal Random Variables	309
13.2.1 Bivariate Normal Distributions	310
13.2.2 Bivariate Normal Samples	313
13.2.3 Inferences about Correlation	316
13.3 Monotonic Association	321
13.4 Spurious Association	321
13.5 Exercises	322
14 Simple Linear Regression	325
14.1 The Regression Line	325
14.2 The Method of Least Squares	331
14.3 Computation	339
14.4 The Simple Linear Regression Model	341
14.5 Regression Diagnostics	346
14.6 Exercises	347
15 Simulation-Based Inference	353
15.1 Termite Foraging Revisited	353
R A Statistical Programming Language	355
R.1 Introduction	355
R.1.1 What is R?	355
R.1.2 Why Use R?	355
R.1.3 Installing R	356
R.1.4 Learning About R	357
R.2 Using R	357

R.2.1	Vectors	358
R.2.2	R is a Calculator!	360
R.2.3	Some Statistics Functions	360
R.2.4	Creating New Functions	360
R.2.5	Exploring Bivariate Normal Data	364
R.2.6	Simulating Termite Foraging	364

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

¹Years 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.

²Actually, 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!

³A. 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 aureus* (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.

⁵Jones, 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

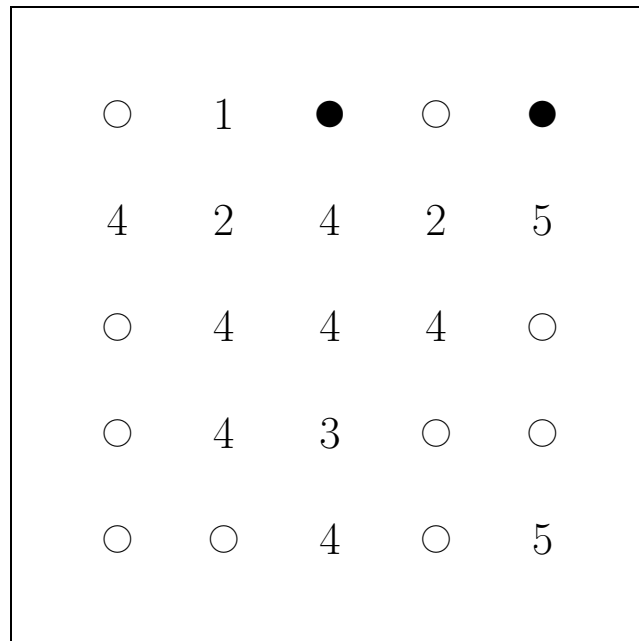


Figure 1.1: Order of *H. aureus* attack in Plot 20.

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.

⁶Ibid.

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 height, 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

⁷Ibid.

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

⁸For 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.

⁹As 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.

¹⁰A 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

¹¹For 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

¹²Another 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

¹³This 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 *Encyclopædia 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 MonopolyTM. 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

¹⁵In *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.’ ”

¹⁶F. 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 PigTM (formerly PigmaniaTM).

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,

¹⁷F. 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

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

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*, $\mathbf{N} = \{1, 2, 3, \dots\}$.
- The set of *integers*, $\mathbf{Z} = \{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}$.
- The set of *real numbers*, $\mathbb{R} = (-\infty, \infty)$.

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,

$$\mathbf{N} \subset \mathbf{Z} \subset \mathbb{R}.$$

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,

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

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

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

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.

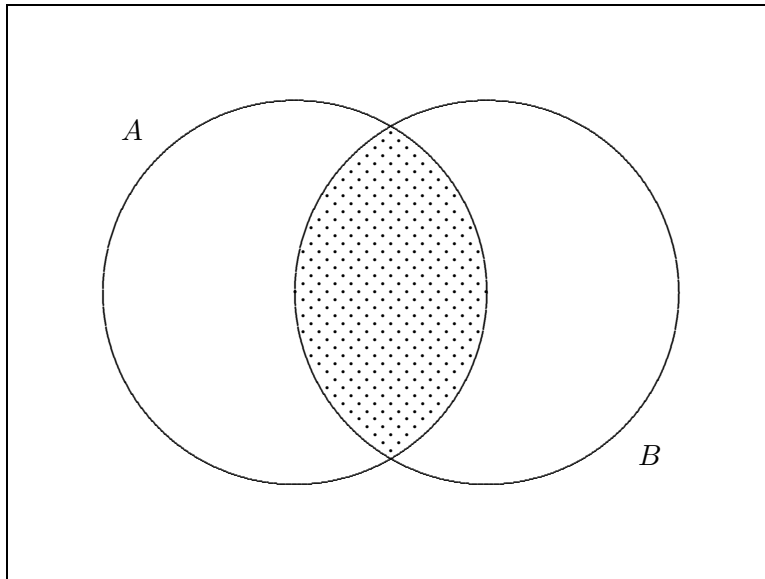


Figure 2.1: A Venn diagram. The shaded region represents the intersection of the nondisjoint sets A and B .

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 $A_k = \{0, 1, \dots, k\}$ for $k = 1, 2, 3, \dots$, then

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

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:

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

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

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

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

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

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,

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

and

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

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:

diamond	\leftrightarrow	white
emerald	\leftrightarrow	green
ruby	\leftrightarrow	red
sapphire	\leftrightarrow	blue

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 $\{1, 2, \dots, N\}$.*

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

¹This 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

$$\begin{array}{l} \{ \text{stuffing, mashed potatoes, green beans} \}, \\ \{ \text{stuffing, green beans, mashed potatoes} \}, \\ \{ \text{mashed potatoes, stuffing, green beans} \}, \\ \{ \text{mashed potatoes, green beans, stuffing} \}, \\ \{ \text{green beans, stuffing, mashed potatoes} \}, \text{ and} \\ \{ \text{green beans, mashed potatoes, stuffing} \} \end{array}$$

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*

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

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:

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

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:

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

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:

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

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 $a, b \in \mathfrak{R}$ and $a < b$.

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

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 $x \in A = \mathfrak{R}$ has a squared value $y \in B = \mathfrak{R}$, 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 $f(\mathfrak{R}) = [0, \infty)$.

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 $\phi : A \rightarrow B$, 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 $\phi : A \rightarrow B$, 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 $\phi : \mathbb{R} \rightarrow \mathbb{R}$ 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,

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

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

$$\left\{1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \dots\right\} \quad (2.1)$$

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 \mathbb{R} .*

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 $n = 1, 2, 3, \dots$

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 $c \in \mathbb{R}$ 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 $y_n \rightarrow c$ as $n \rightarrow \infty$ or $\lim_{n \rightarrow \infty} y_n = c$. In particular,

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

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?

²You 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.

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

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

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

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

- (a) Compute y_0, y_1, y_2, y_3 , and y_4 .
- (b) The sequence $\{y_0, y_1, y_2, \dots\}$ is an example of a *sequence of partial sums*. Guess the value of its limit, usually written

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

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 occurrence.

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 meteorology, 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.

¹I. 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

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

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.

²Mark 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,

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

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.,

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

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

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

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., $\{HH, TH, TT\}$. Thus, the following collection of events might be deemed appropriate:

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

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 $\{E_1, E_2, E_3, \dots\}$ 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 -0.5 ” 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 $\{\text{HT}\}$ and $\{\text{TH}\}$ are events and that we want to compute the probability that exactly one **Head** is observed, i.e., the probability of

$$\{\text{HT}\} \cup \{\text{TH}\} = \{\text{HT}, \text{TH}\}.$$

Because $\{\text{HT}\}$ and $\{\text{TH}\}$ are events, their union is an event and therefore has a probability. Because they are mutually exclusive, we would like that probability to be

$$P(\{\text{HT}, \text{TH}\}) = P(\{\text{HT}\}) + P(\{\text{TH}\}).$$

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

$$A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B), \quad (3.1)$$

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, \dots, 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, \dots, 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, \dots 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, \dots 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*

$$P(E^c) = 1 - P(E).$$

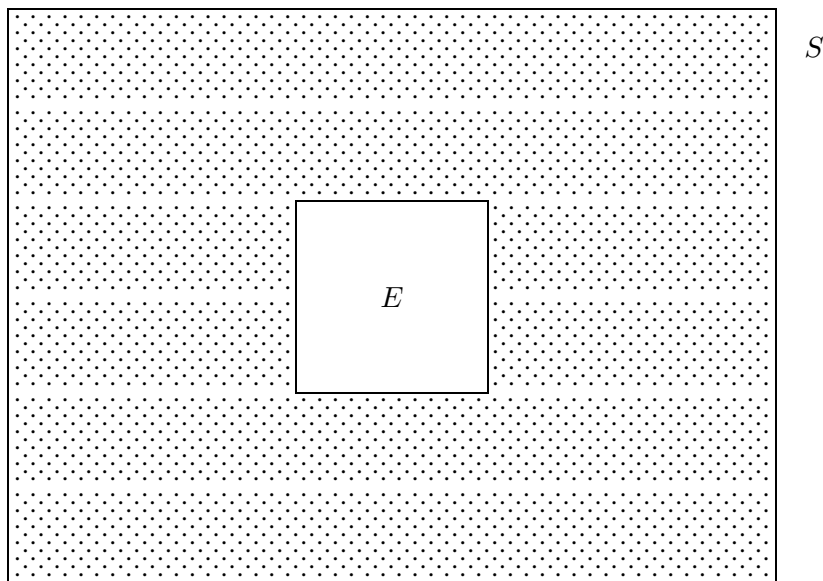


Figure 3.1: A Venn diagram for the probability of E^c .

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,

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

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

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

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

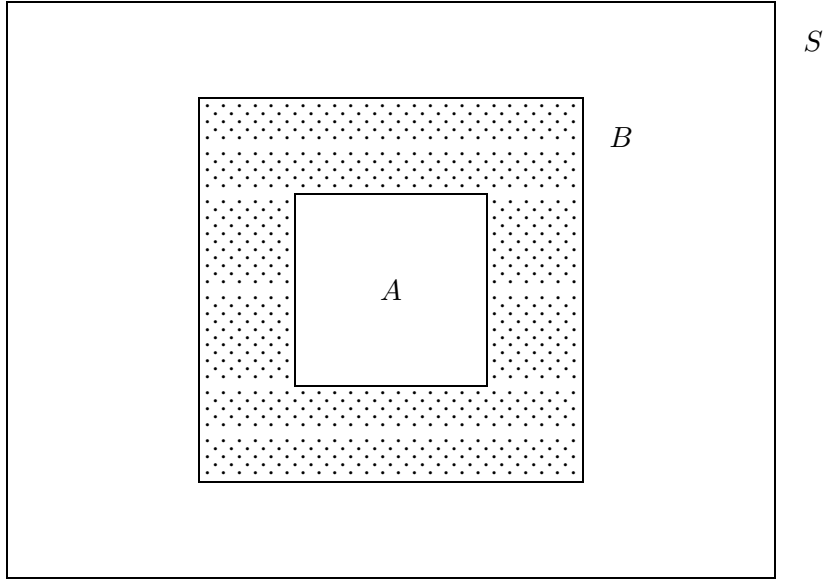


Figure 3.2: A Venn diagram for the probability of $A \subset B$.

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. \square

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

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

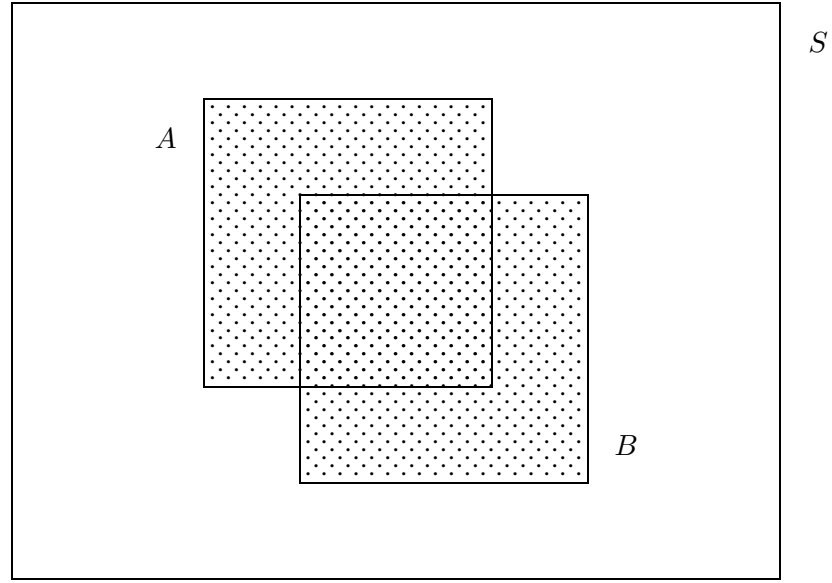


Figure 3.3: A Venn diagram for the probability of $A \cup B$.

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,

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

as claimed. \square

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

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

and we recover our original formula for that case.

3.3 Finite Sample Spaces

Let

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

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 $i = 1, \dots, N$. Then, for any event A , we can write

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

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

$$p_i = \frac{1}{N} \quad (3.3)$$

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

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

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

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 *probability* 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 $A = \{\text{HT}, \text{TH}\}$ 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 $1, 2, 3, \dots, 40$, 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:

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

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

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

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, $P(A)$.

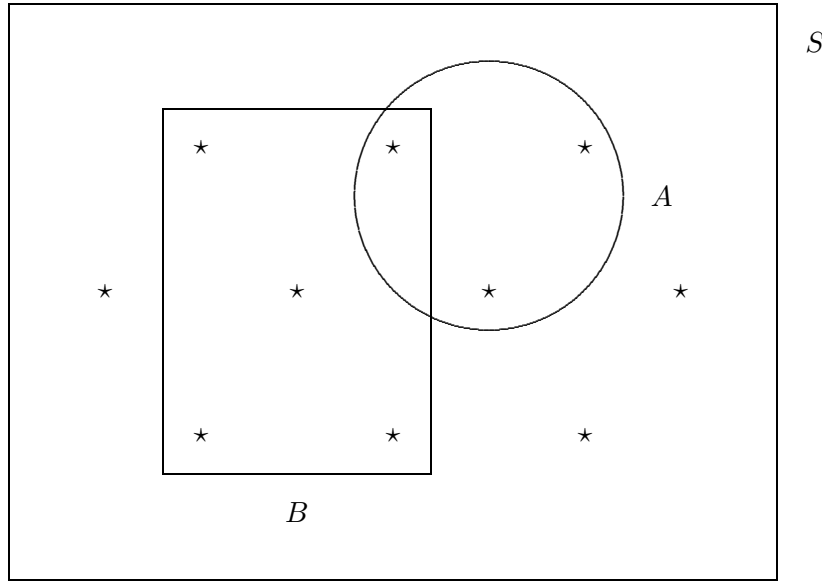


Figure 3.4: A Venn diagram that illustrates conditional probability. Each \star represents an individual outcome.

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*

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

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

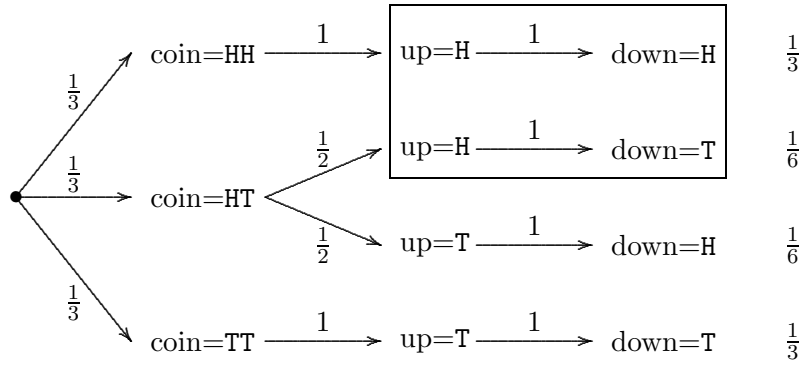


Figure 3.5: A tree diagram for Example 3.9.

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

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

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

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

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

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

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.

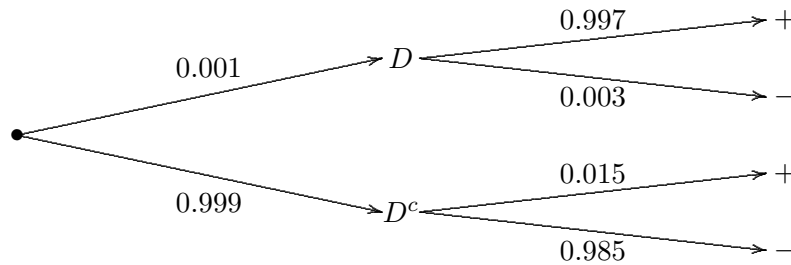


Figure 3.6: A tree diagram for Example 3.10.

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

³See 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

$$P(B|A) = P(B). \quad (3.8)$$

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.

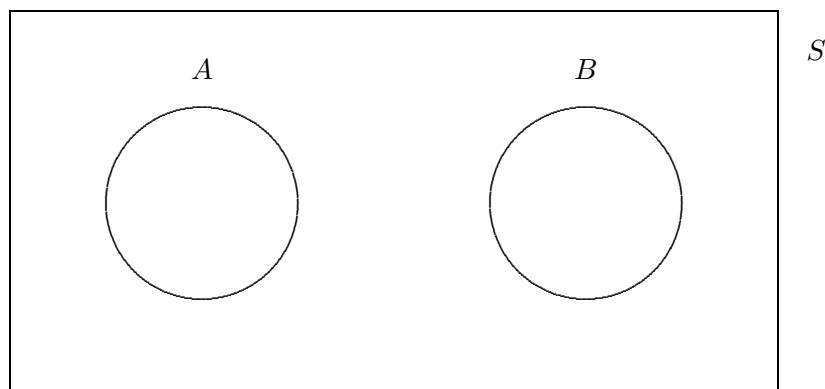


Figure 3.7: A Venn diagram for Example 3.12.

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

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

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.

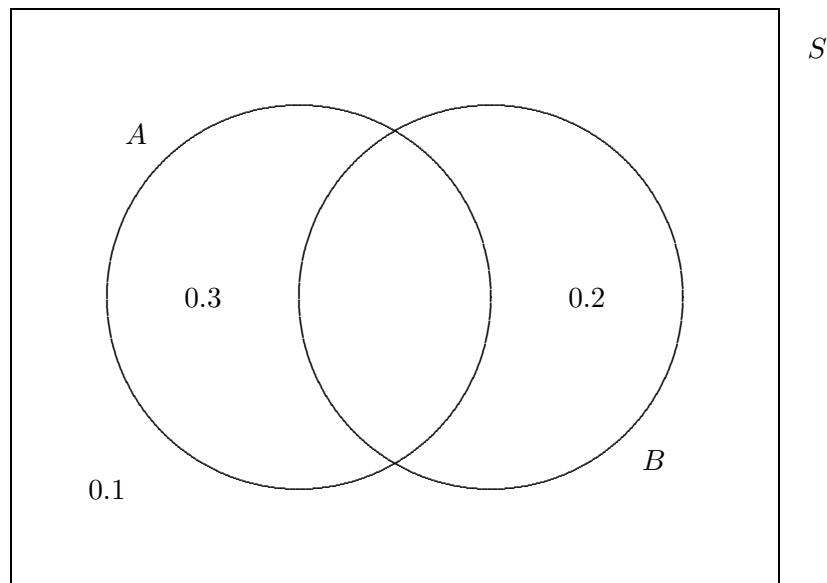


Figure 3.8: A Venn diagram for Example 3.13.

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

⁴This 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 $A_{\alpha_1}, \dots, A_{\alpha_k}$,*

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

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

$$\begin{array}{|c|} \hline H \\ \hline T \\ \hline \end{array} \xrightarrow{X} \begin{array}{|c|} \hline 1 \\ \hline 0 \\ \hline \end{array}$$

In functional notation, $X : S \rightarrow \mathfrak{R}$ and the rule of assignment is defined by

$$\begin{aligned} X(H) &= 1, \\ X(T) &= 0. \end{aligned}$$

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

The sample space for this experiment is $S = \{HH, HT, TH, TT\}$. 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:

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

In functional notation, $Y : S \rightarrow \mathfrak{R}$ and the rule of assignment is defined by

$$\begin{aligned} Y(HH) &= 2, \\ Y(HT) &= Y(TH) = 1, \\ Y(TT) &= 0. \end{aligned}$$

- (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 occurrence of **Heads** and 0 to each occurrence of **Tails**. We denote this assignment by $X : S \rightarrow \mathfrak{R}^2$. 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 $g : \mathfrak{R}^2 \rightarrow \mathfrak{R}$ 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:

$$\begin{array}{|c|c|} \hline \text{HH} & \text{HT} \\ \hline \text{TH} & \text{TT} \\ \hline \end{array} \xrightarrow{X} \begin{array}{|c|c|} \hline (1, 1) & (1, 0) \\ \hline (0, 1) & (0, 0) \\ \hline \end{array} \xrightarrow{g} \begin{array}{|c|c|} \hline 2 & 1 \\ \hline 1 & 0 \\ \hline \end{array}$$

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

$$\begin{array}{|c|c|} \hline 1 & 1 \\ \hline 0 & 0 \\ \hline \end{array} \xleftarrow{X_1} \begin{array}{|c|c|} \hline \text{HH} & \text{HT} \\ \hline \text{TH} & \text{TT} \\ \hline \end{array} \xrightarrow{X_2} \begin{array}{|c|c|} \hline 1 & 0 \\ \hline 1 & 0 \\ \hline \end{array}$$

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, \mathfrak{R} . To complete the transference, we must decide which subsets of \mathfrak{R} 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 \mathfrak{R} and every subset of \mathfrak{R} 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 \mathfrak{R} . 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.

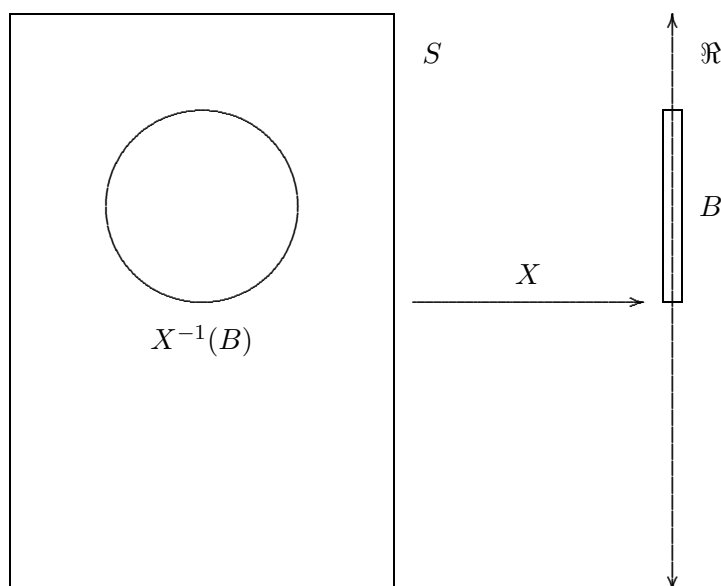


Figure 3.9: The inverse image of a Borel set.

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 $X : S \rightarrow \mathfrak{R}$ is a random variable if and only if

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

exists for all choices of $y \in \mathfrak{R}$.

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

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

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 $F : \mathfrak{R} \rightarrow \mathfrak{R}$ 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, $P(\text{H}) = P(\text{T}) = 0.5$, and the following values of the cdf are easily determined:

- If $y < 0$, e.g., $y = -9.1185$ or $y = -0.3018$, then

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

- $F(0) = P(X \leq 0) = P(\{\text{T}\}) = 0.5$.

- If $y \in (0, 1)$, e.g., $y = 0.6241$ or $y = 0.9365$, then

$$F(y) = P(X \leq y) = P(\{\text{T}\}) = 0.5.$$

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

- If $y > 1$, e.g., $y = 1.5248$ or $y = 7.7397$, then

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

The entire cdf is plotted in Figure 3.10.

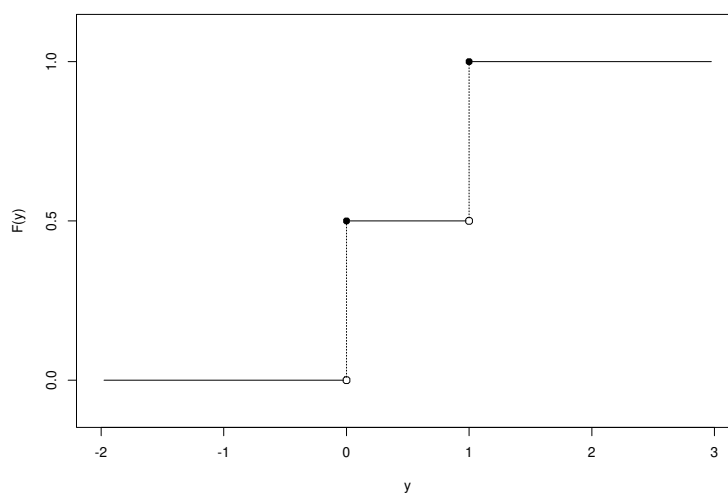


Figure 3.10: The cumulative distribution function for tossing a penny with $P(\text{Heads}) = 0.5$.

(b) *A typical penny is spun.*

For this experiment, we assume that $P(\text{H}) = 0.3$ and $P(\text{T}) = 0.7$ (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

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

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

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

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

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

The entire cdf is plotted in Figure 3.11.

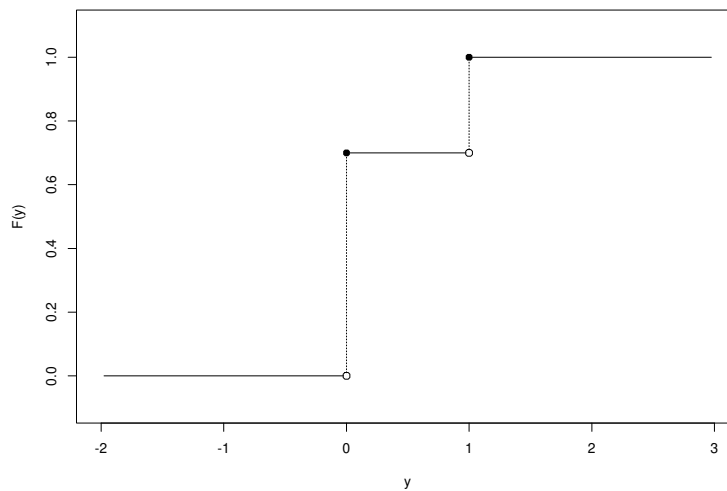


Figure 3.11: The cumulative distribution function for spinning a penny with $P(\text{Heads}) = 0.3$.

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

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

- $F(0) = P(X \leq 0) = P(\{\text{TT}\}) = 0.25$.

- If $y \in (0, 1)$, e.g., $y = 0.3074$ or $y = 0.6924$, then

$$F(y) = P(X \leq y) = P(\{\text{TT}\}) = 0.25.$$

- $F(1) = P(X \leq 1) = P(\{\text{TT}, \text{HT}, \text{TH}\}) = 0.75$.

- If $y \in (1, 2)$, e.g., $y = 1.4629$ or $y = 1.5159$, then

$$F(y) = P(X \leq y) = P(\{\text{TT}, \text{HT}, \text{TH}\}) = 0.75.$$

- $F(2) = P(X \leq 2) = P(\{\text{TT}, \text{HT}, \text{TH}, \text{HH}\}) = 1$.

- If $y > 2$, e.g., $y = 2.1252$ or $y = 3.7790$, then

$$F(y) = P(X \leq y) = P(\{\text{TT}, \text{HT}, \text{TH}, \text{HH}\}) = 1.$$

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:

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

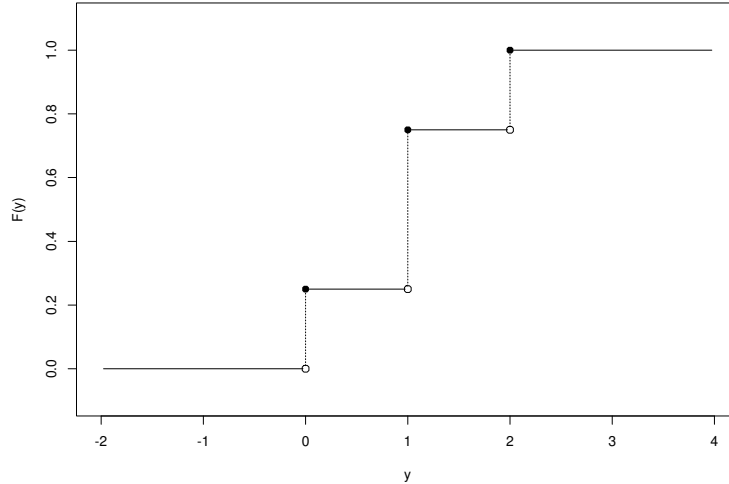


Figure 3.12: The cumulative distribution function for tossing two pennies with $P(\text{Heads}) = 0.5$ and counting the number of **Heads**.

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 $X_1 : S \rightarrow \mathfrak{R}$ and $X_2 : S \rightarrow \mathfrak{R}$ be random variables. X_1 and X_2 are independent if and only if, for each $y_1 \in \mathfrak{R}$ and each $y_2 \in \mathfrak{R}$,

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

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 modified 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.

⁶Richard 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 implications 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

⁷I 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 $4/16$.

- 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 $3/5$.

- 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:

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

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.

- (f) Consider the population of U.S. college freshmen, from which a student is randomly selected. Let A denote the event that the student attends the College of William & Mary, and let B denote the event that the student graduated from high school in Virginia.
 - (g) Consider the population of all persons (living or dead) who have earned a Ph.D. from an American university, from which one is randomly selected. Let A denote the event that the person's Ph.D. was earned before 1950 and let B denote the event that the person is female.
 - (h) Consider the population of persons who resided in New Orleans before Hurricane Katrina. A person is selected at random from this population. Let A denote the event that the person left New Orleans before Katrina arrived, and let B denote the event that the person belonged to a household whose 2004 income was below the federal poverty line.
 - (i) Consider the population of all couples who married in the United States in 1995. A couple is selected at random from this population. Let A denote the event that the couple cohabited (lived together) before marrying, and let B denote the event that the couple had divorced by 2005.
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

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

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 $i = 1, 2, 3, \dots$, 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 $f : \mathfrak{R} \rightarrow \mathfrak{R}$ 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 $x \in \mathfrak{R}$.

2. If $x \notin X(S)$, then $f(x) = 0$.
3. By the definition of $X(S)$,

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

There is an important relation between the pmf and the cdf. For each $y \in \mathfrak{R}$, let

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

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

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

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 $B \subset \mathfrak{R}$ by summing the values of the pmf over all values of X that lie in B . Here is the formula:

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

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 $X(\text{Heads}) = 1$ and $X(\text{Tails}) = 0$.*

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 $X(\text{Heads}) = 1$ and $X(\text{Tails}) = 0$.

Assuming that $P(\text{Heads}) = 0.3$ (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 4.13. 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, \dots 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 $Y(S) = \{0, 1, 2, \dots\}$ and the pmf of Y is

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

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

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

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

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

¹In 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 insecticides 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 .

³This 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 $\{1, 2, \dots, 39, 40\}$. 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

$$10^6 \cdot 2.6053 \times 10^{-7} \doteq 0.26,$$

⁴Robert 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 $X : S \rightarrow \mathfrak{R}$ is a discrete random variable and $\phi : \mathfrak{R} \rightarrow$

⁵Ambrose 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 $Y : \Re \rightarrow \Re$ is a random variable and

$$\begin{aligned}
 E\phi(X) &= EY = \sum_{y \in Y(S)} yP(Y = y) \\
 &= \sum_{y \in Y(S)} yP(\phi(X) = y) \\
 &= \sum_{y \in Y(S)} yP(X \in \phi^{-1}(y)) \\
 &= \sum_{y \in Y(S)} y \left(\sum_{x \in \phi^{-1}(y)} P(X = x) \right) \\
 &= \sum_{y \in Y(S)} \sum_{x \in \phi^{-1}(y)} yP(X = x) \\
 &= \sum_{y \in Y(S)} \sum_{x \in \phi^{-1}(y)} \phi(x)P(X = x) \\
 &= \sum_{x \in X(S)} \phi(x)P(X = x) \\
 &= \sum_{x \in X(S)} \phi(x)f(x). \tag{4.5}
 \end{aligned}$$

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 $X(S) = \{0, 1, 2, \dots\}$ and the geometric random variable X has pmf

$$f(x) = P(x \text{ consecutive Tails}) = 0.5^x.$$

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*

$$E[c\phi(X)] = \sum_{x \in X(S)} c\phi(x)f(x) = c \sum_{x \in X(S)} \phi(x)f(x) = cE[\phi(X)].$$

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*

$$\begin{aligned} E[\phi_1(X) + \phi_2(X)] &= \sum_{x \in X(S)} [\phi_1(x) + \phi_2(x)]f(x) \\ &= \sum_{x \in X(S)} [\phi_1(x)f(x) + \phi_2(x)f(x)] \\ &= \sum_{x \in X(S)} \phi_1(x)f(x) + \sum_{x \in X(S)} \phi_2(x)f(x) \\ &= E[\phi_1(X)] + E[\phi_2(X)]. \end{aligned}$$

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 $\phi : \Re \rightarrow \Re$ 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 $\text{Var}(X)$ or simply $\text{Var} X$, is the probability-weighted average of the squared deviations of X from $EX = \mu$, i.e.,*

$$\text{Var} X = E(X - \mu)^2 = \sum_{x \in X(S)} (x - \mu)^2 f(x).$$

Remark The variance of X , $\text{Var} X$, 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 $\text{Var} X$ 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*

$$\begin{aligned}\text{Var } X &= E(X - \mu)^2 \\ &= E(X^2 - 2\mu X + \mu^2) \\ &= EX^2 + E(-2\mu X) + E\mu^2 \\ &= EX^2 - 2\mu EX + \mu^2 \\ &= EX^2 - 2\mu^2 + \mu^2 \\ &= EX^2 - (EX)^2.\end{aligned}$$

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) $\sigma^2 = \text{Var } X = EX^2 - (EX)^2 = (0.4 + 1.35 + 6.25 + 50) - 6.9^2 = 58 - 47.61 = 10.39.$

(c) $\sigma = \sqrt{10.39} \doteq 3.2234.$

Now suppose that $X : S \rightarrow \Re$ is a discrete random variable and $\phi : \Re \rightarrow \Re$ is a function. Let $Y = \phi(X)$. Then Y is a discrete random variable and

$$\text{Var } \phi(X) = \text{Var } Y = E[Y - EY]^2 = E[\phi(X) - E\phi(X)]^2. \quad (4.6)$$

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*

$$\text{Var}(X + c) = \text{Var } X.$$

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*

$$\begin{aligned}
 \text{Var}(cX) &= E[cX - E(cX)]^2 \\
 &= E[cX - cEX]^2 \\
 &= E[c(X - EX)]^2 \\
 &= E[c^2(X - EX)^2] \\
 &= c^2 E(X - EX)^2 \\
 &= c^2 \text{Var } X.
 \end{aligned}$$

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*

$$\text{Var}(X_1 + X_2) = \text{Var } X_1 + \text{Var } X_2.$$

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.,

$$S = \{\text{HH}, \text{HT}, \text{TH}, \text{TT}\},$$

the pmf of Y is easily determined:

$$\begin{aligned}
 f(0) &= P(Y = 0) = P(\{\text{HH}\}) = 0.25, \\
 f(1) &= P(Y = 1) = P(\{\text{HT}, \text{TH}\}) = 0.5, \\
 f(2) &= P(Y = 2) = P(\{\text{TT}\}) = 0.25,
 \end{aligned}$$

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, \dots, 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 $Y(S) = \{0, 1, \dots, n\}$. 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 000011 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 astragali,
- (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 deliberation, 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 all challenged calls in NCAA tennis matches 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 $P(\text{Heads}) = 0.3$ 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 $S = \{s_1, \dots, s_N\}$, then $P(\{s_i\}) = 1/N$. This construction sufficed to define probabilities of events: if $E \subset S$, then

$$E = \{s_{i_1}, \dots, s_{i_k}\};$$

and consequently

$$P(E) = P\left(\bigcup_{j=1}^k \{s_{i_j}\}\right) = \sum_{j=1}^k P(\{s_{i_j}\}) = \sum_{j=1}^k \frac{1}{N} = \frac{k}{N}.$$

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

$$E = \left\{ \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \dots \right\}. \quad (5.1)$$

Then we must have

$$P(E) = P\left(\bigcup_{j=2}^{\infty} \left\{ \frac{1}{j} \right\}\right) = \sum_{j=2}^{\infty} P\left(\left\{ \frac{1}{j} \right\}\right) = \sum_{j=2}^{\infty} \epsilon = \infty, \quad (5.2)$$

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.,

$$P(X \in [0, 0.5]) = P(X \in [0.5, 1]).$$

Because $[0, 0.5] \cup [0.5, 1] = [0, 1]$ and $P(X = 0.5) = 0$, we have

$$\begin{aligned} 1 = P(X \in [0, 1]) &= P(X \in [0, 0.5]) + P(X \in [0.5, 1]) - P(X = 0.5) \\ &= P(X \in [0, 0.5]) + P(X \in [0.5, 1]). \end{aligned}$$

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 $\text{Uniform}[0, 1]$. If $X \sim \text{Uniform}[0, 1]$, 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.

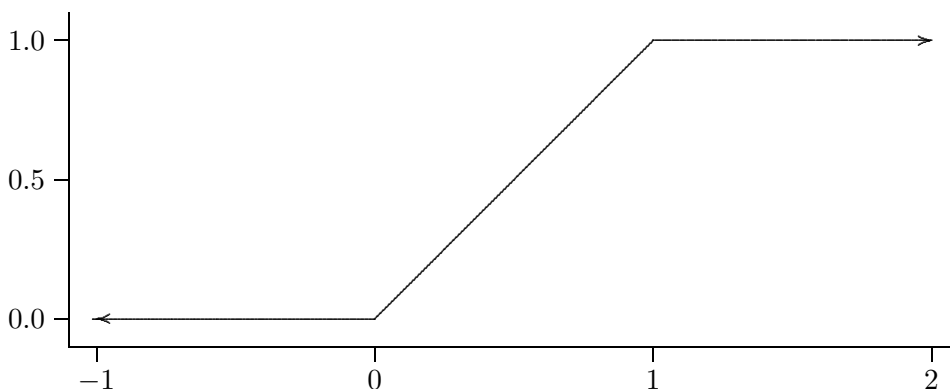


Figure 5.1: The cumulative distribution function of $X \sim \text{Uniform}(0, 1)$.

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 probabilities of individual points, we need to represent the probabilities of intervals.

Consider the function

$$f(x) = \begin{cases} 0 & x \in (-\infty, 0) \\ 1 & x \in [0, 1] \\ 0 & x \in (1, \infty) \end{cases}, \quad (5.3)$$

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.

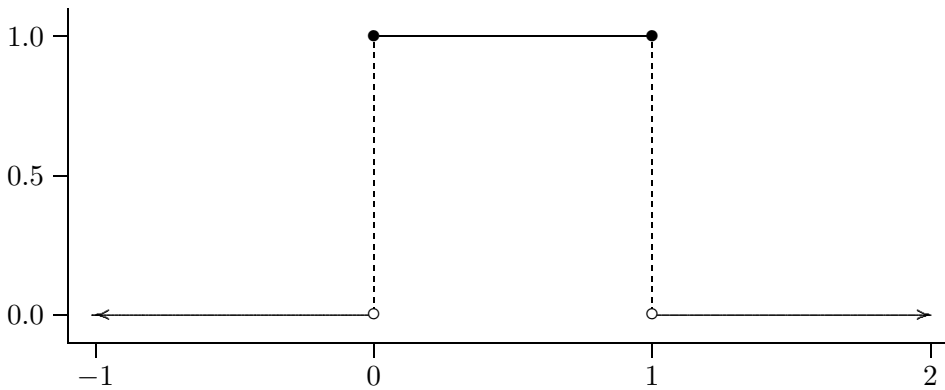


Figure 5.2: The probability density function of $X \sim \text{Uniform}(0, 1)$.

5.2 Basic Concepts

Consider the graph of a function $f : \mathfrak{R} \rightarrow \mathfrak{R}$, 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

$$\text{Area}_{[a,b]}(f) = \int_a^b f(x)dx. \quad (5.4)$$

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 $f : \mathfrak{R} \rightarrow \mathfrak{R}$ such that*

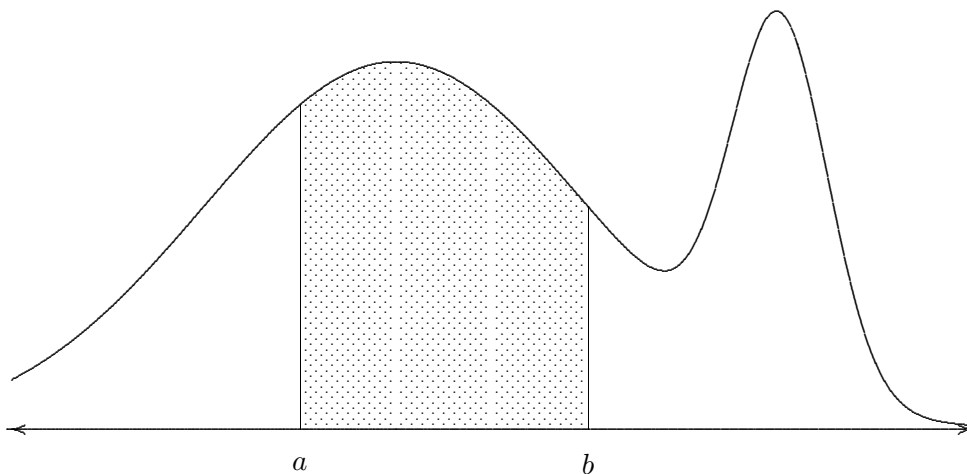


Figure 5.3: A continuous probability density function.

1. $f(x) \geq 0$ for every $x \in \mathbb{R}$.
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*

$$P(X \in [a, b]) = \int_a^b f(x)dx.$$

It is immediately apparent from this definition that the cdf of a continuous random variable X is

$$F(y) = P(X \leq y) = P(X \in (-\infty, y]) = \int_{-\infty}^y f(x)dx. \quad (5.5)$$

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 Calculus 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 $g : \Re \rightarrow \Re$ 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*

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

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

$$f(x) = \begin{cases} 0 & x \in (-\infty, 0) \\ 1/60 & x \in [0, 60) \\ 0 & x \in [60, \infty) \end{cases}. \quad (5.6)$$

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.,

$$P(X \in (10, 20)) = \text{Area}_{(10,20)}(f) = (20 - 10) \frac{1}{60} = \frac{1}{6}.$$

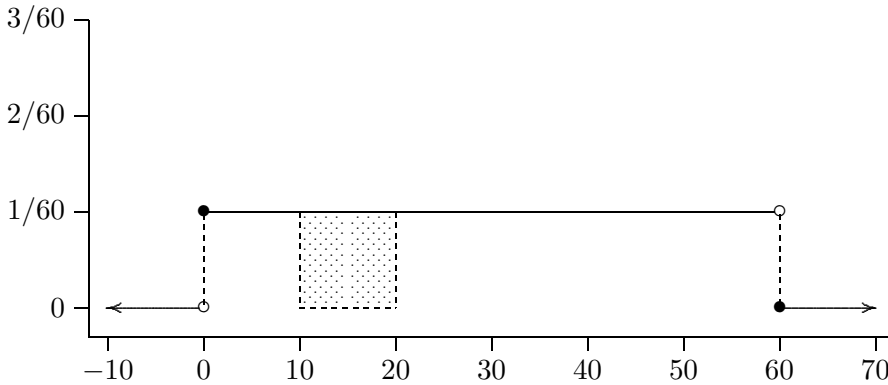


Figure 5.4: The probability density function of $X \sim \text{Uniform}[0, 60)$.

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 $\text{Uniform}[0, 60)$, 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 $y \geq 60$. If $y \in [0, 60)$, then (by the independence of X_1 and X_2)

$$\begin{aligned} F(y) = P(Y \leq y) &= P(\max(X_1, X_2) \leq y) = P(X_1 \leq y, X_2 \leq y) \\ &= P(X_1 \leq y) \cdot P(X_2 \leq y) = \frac{y-0}{60-0} \cdot \frac{y-0}{60-0} \\ &= \frac{y^2}{3600}. \end{aligned}$$

Thus, the desired probability is

$$P(30 < Y < 50) = F(50) - F(30) = \frac{50^2}{3600} - \frac{30^2}{3600} = \frac{4}{9}.$$

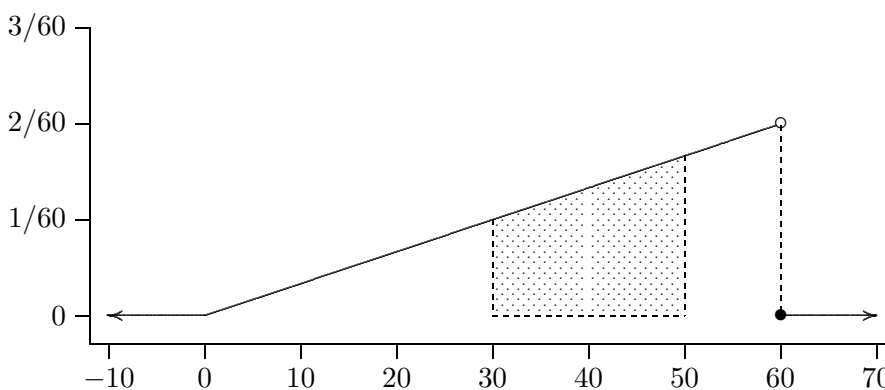


Figure 5.5: The probability density function for Example 5.2.

In preparation for Example 5.3, we claim that the pdf of Y is

$$f(y) = \begin{cases} 0 & y \in (-\infty, 0) \\ y/1800 & y \in [0, 60) \\ 0 & y \in [60, \infty) \end{cases},$$

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

$$P(Y \in [0, y)) = P(Y \leq y) = F(y) = \frac{y^2}{3600} = \frac{1}{2}(y - 0)\frac{y}{1800} = \text{Area}_{[0,y)}(f).$$

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

$$P(45 < Z < 75) = P(Z \in (45, 75)).$$

It is apparent that $Z(S) = [0, 120)$. Although we omit the derivation, it can be determined mathematically that the pdf of Z is

$$f(z) = \begin{cases} 0 & z \in (-\infty, 0) \\ z/3600 & z \in [0, 60) \\ (120 - z)/3600 & z \in [60, 120) \\ 0 & z \in [120, \infty) \end{cases}.$$

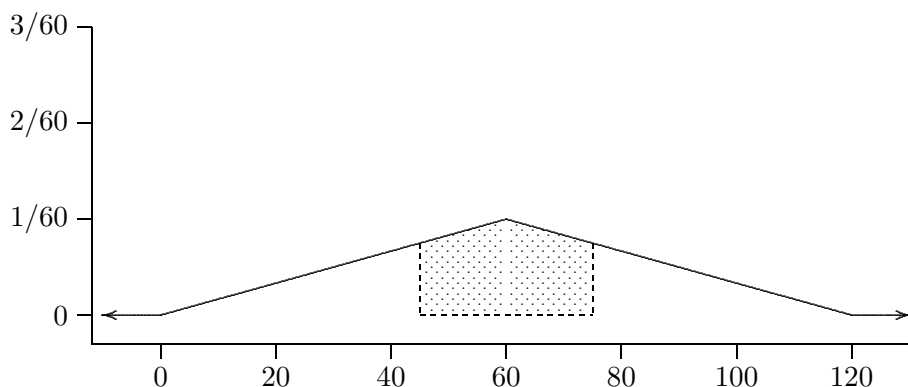


Figure 5.6: The probability density function for Example 5.3.

This pdf is graphed in Figure 5.6, in which it is apparent that the area of the shaded region is

$$\begin{aligned}
 P(45 < Z < 75) &= P(Z \in (45, 75)) = \text{Area}_{(45,75)}(f) \\
 &= 1 - \frac{1}{2}(45 - 0)\frac{45}{3600} - \frac{1}{2}(120 - 75)\frac{120 - 75}{3600} \\
 &= 1 - \frac{45^2}{60^2} = \frac{7}{16}.
 \end{aligned}$$

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 $X \sim \text{Normal}(\mu, \sigma^2)$, if the pdf of X is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left[-\frac{1}{2} \left(\frac{x - \mu}{\sigma} \right)^2 \right]. \quad (5.7)$$

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.

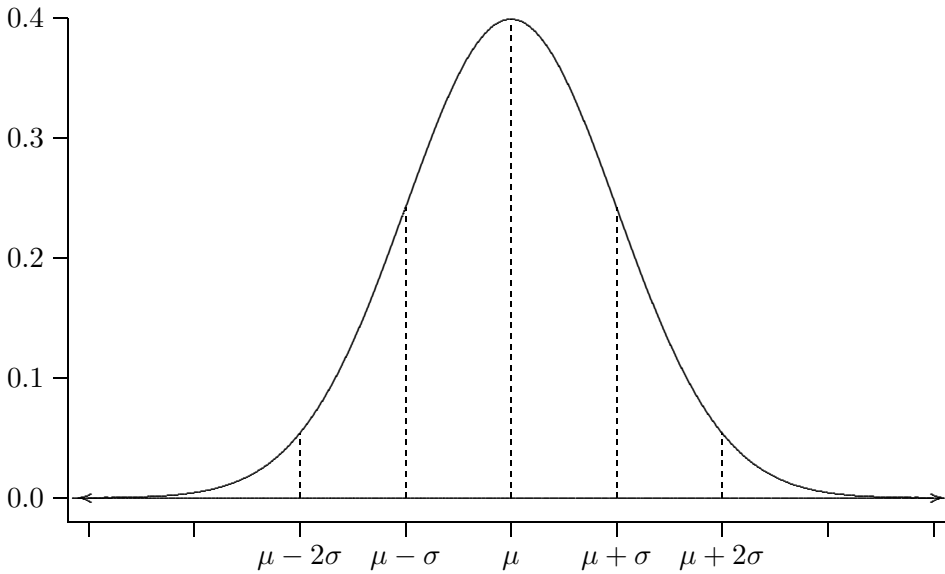


Figure 5.7: The probability density function of $X \sim \text{Normal}(\mu, \sigma^2)$.

1. $f(x) > 0$. It follows that, for any nonempty interval (a, b) ,

$$P(X \in (a, b)) = \text{Area}_{(a,b)}(f) > 0,$$

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 $\text{Normal}(0, 1)$.*

The following result is of enormous practical value:

Theorem 5.1 *If $X \sim \text{Normal}(\mu, \sigma^2)$, then*

$$Z = \frac{X - \mu}{\sigma} \sim \text{Normal}(0, 1).$$

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 $X \sim \text{Normal}(1, 4)$, 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 $X \sim \text{Normal}(-1, 9)$, 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 $X \sim \text{Normal}(2, 16)$, 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 $X \sim \text{Normal}(-3, 25)$, 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 $X \sim \text{Normal}(4, 16)$, 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 $X_1 \sim \text{Normal}(\mu_1, \sigma_1^2)$ and $X_2 \sim \text{Normal}(\mu_2, \sigma_2^2)$ are independent, then

$$X_1 + X_2 \sim \text{Normal}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

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 $Z_1, \dots, Z_n \sim \text{Normal}(0, 1)$ and consider the continuous random variable

$$Y = Z_1^2 + \dots + Z_n^2.$$

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.

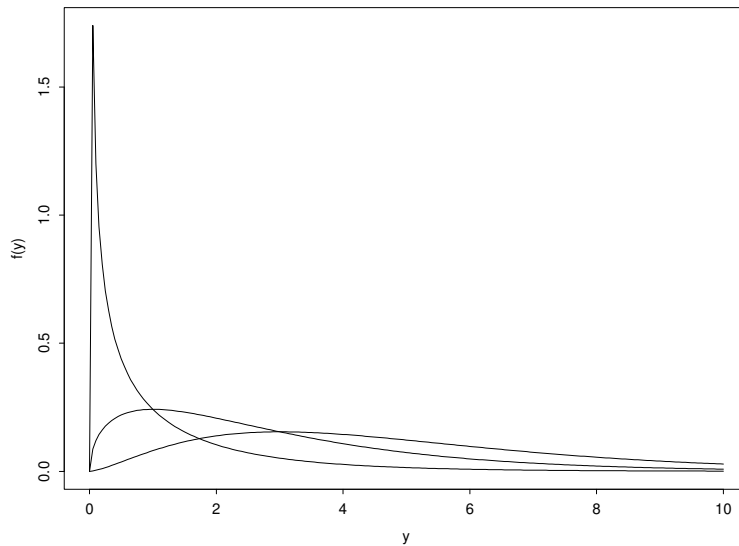


Figure 5.8: The probability density functions of $Y \sim \chi^2(\nu)$ for $\nu = 1, 3, 5$.

The following fact is quite useful:

Theorem 5.3 If $Z_1, \dots, Z_n \sim \text{Normal}(0, 1)$ and $Y = Z_1^2 + \dots + Z_n^2$, 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:

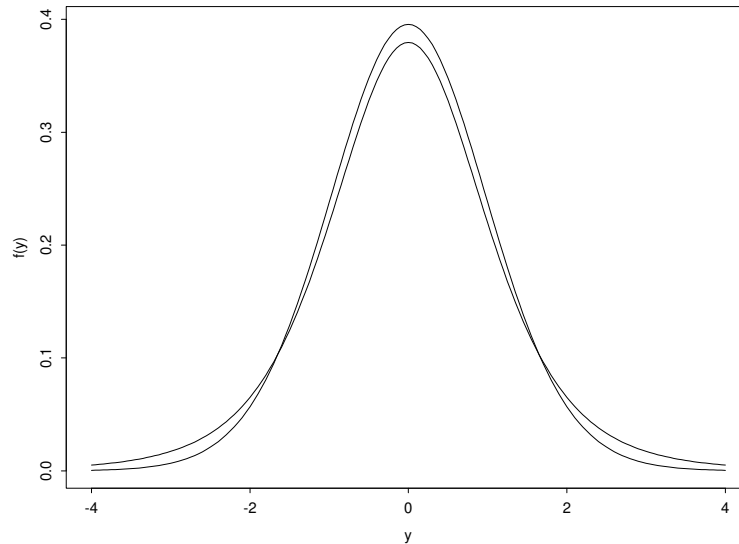


Figure 5.9: The probability density functions of $T \sim t(\nu)$ for $\nu = 5, 30$.

Theorem 5.4 Let F_ν denote the cdf of $t(\nu)$ and let Φ denote the cdf of $\text{Normal}(0, 1)$. Then

$$\lim_{\nu \rightarrow \infty} F_\nu(t) = \Phi(t)$$

for every $t \in (-\infty, \infty)$.

Thus, when ν is sufficiently large ($\nu > 40$ is a reasonable rule of thumb), $t(\nu)$ is approximately $\text{Normal}(0, 1)$ 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 $\text{Normal}(0, 1)$ 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.

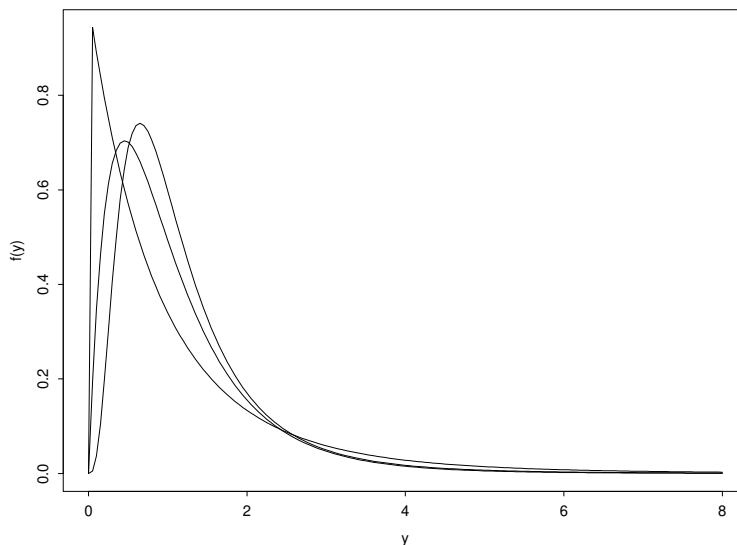


Figure 5.10: The probability density functions of $F \sim F(\nu_1, \nu_2)$ for $(\nu_1, \nu_2) = (2, 12), (4, 20), (9, 10)$.

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

$$T = \frac{Z}{\sqrt{Y_2/\nu_2}} \sim t(\nu_2)$$

and

$$T^2 = \frac{Z^2}{Y_2/\nu_2} = \frac{Y_1/1}{Y_2/\nu_2} \sim F(1, \nu_2).$$

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 $p = P(Z > 1.96)$, please identify each of the following symbols as a real number or a function:
 - i. p
 - ii. P
 - iii. Z
- (b) In the equation $\sigma^2 = E(X - \mu)^2$, 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:

$$f(x) = \begin{cases} 0 & \text{if } x < 1 \\ 2(x-1) & \text{if } 1 \leq x \leq 2 \\ 0 & \text{if } x > 2 \end{cases}.$$

- (a) Graph f .
 - (b) Verify that f is a pdf.
 - (c) Compute $P(1.50 < X < 1.75)$.
3. Consider the function $f : \Re \rightarrow \Re$ defined by

$$f(x) = \begin{cases} 0 & x < 0 \\ cx & 0 < x < 1.5 \\ c(3-x) & 1.5 < x < 3 \\ 0 & x > 3 \end{cases},$$

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. Imagine throwing darts at a triangular dart board,

$$B = \{(x, y) : 0 \leq y \leq x \leq 1\}.$$

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. Define the random variable X to be the value of the x -coordinate of the point that is hit, and define the random variable Y to be the value of the y -coordinate of the point that is hit.

- (a) Draw a picture of B .
 - (b) Compute $P(X \leq 0.5)$.
 - (c) Determine and graph the cumulative distribution function of X .
 - (d) Are X and Y independent?
6. 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 $\theta \in \mathbb{R}$ such that*

$$f(\theta + x) = f(\theta - x)$$

for every $x \in \mathbb{R}$, 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 $(a + b)/2$.

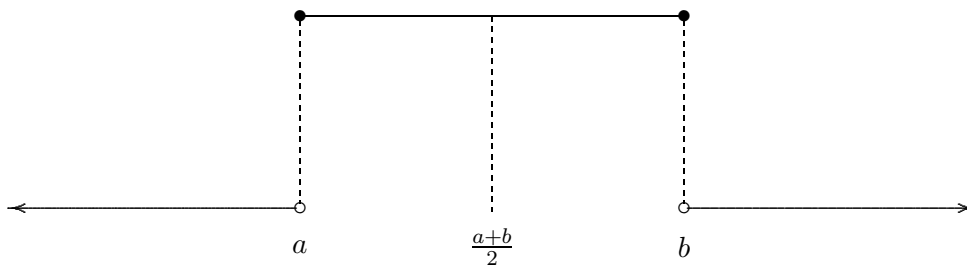


Figure 6.1: $X \sim \text{Uniform}[a, b]$ has center of symmetry $(a + b)/2$.

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 $\theta \in \mathfrak{R}$ 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 expected 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.

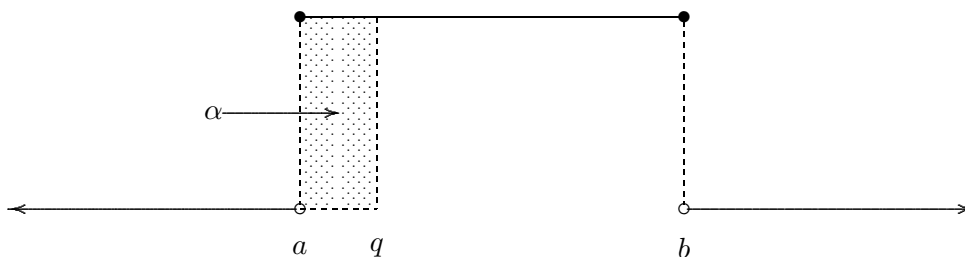


Figure 6.2: A quantile of a Uniform distribution.

Example 6.2 Suppose that X has pdf

$$f(x) = \begin{cases} x/2 & x \in [0, 2] \\ 0 & \text{otherwise} \end{cases},$$

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

$$f(x) = \begin{cases} 1/2 & x \in [0, 1] \cup [2, 3] \\ 0 & \text{otherwise} \end{cases},$$

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

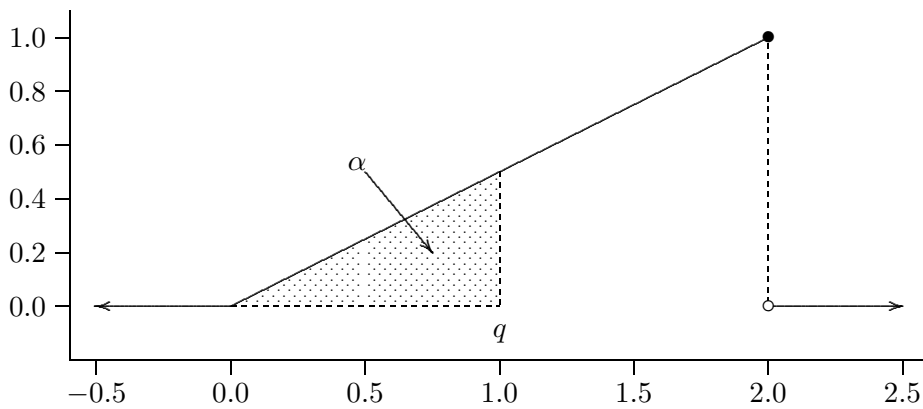
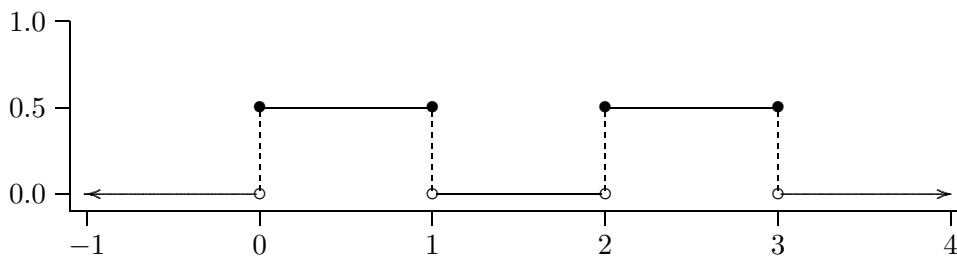


Figure 6.3: A quantile of another distribution.

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 $k = 1, 2, 3, \dots$. 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, $x = 10, 100, 1000, \dots$. 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 $\text{Var } X = \sigma^2$. 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 $\text{Var } X$ 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

$$\begin{aligned}\sigma_k^2 &= EX_k^2 - (EX_k)^2 = (0.19 + 0.00 + 0.19 + 100^k \cdot 0.02) - (2 \cdot 10^{k-2})^2 \\ &= 0.38 + 2 \cdot 100^{k-1} - 4 \cdot 100^{k-2} = 0.38 + 196 \cdot 100^{k-2},\end{aligned}$$

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 $g : \Re \rightarrow \Re$ defined by

$$g(x) = \begin{cases} 0 & x < 0 \\ x & x \in [0, 1] \\ 1 & x \in [1, 2] \\ 3 - x & x \in [2, 3] \\ 0 & x > 3 \end{cases}.$$

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

$$f(x) = \begin{cases} 0 & x < 0 \\ x & x \in (0, 1) \\ (3 - x)/4 & x \in (1, 3) \\ 0 & x > 3 \end{cases}.$$

- (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. Consider the dart-throwing experiment described in Exercise 5.6.5 and compute the following quantities:
- (a) $q_2(X)$
 - (b) $q_2(Y)$
 - (c) $\text{iqr}(X)$
 - (d) $\text{iqr}(Y)$

5. 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 $\text{CCQ} \sim \text{Normal}(100, 400)$. Assuming that Michael is correct, what is Lulu’s CCQ score?
6. A random variable $X \sim \text{Uniform}(5, 15)$ has population mean $\mu = EX = 10$ and population variance $\sigma^2 = \text{Var } X = 225$. Let Y denote a normal random variable with the same mean and variance.
 - (a) Consider X . What is the ratio of its interquartile range to its standard deviation?
 - (b) Consider Y . What is the ratio of its interquartile range to its standard deviation?
7. 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.
8. 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.
9. 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 $X_1, \dots, X_n \sim P$. Let x_i denote the observed value of X_i . The set of observed values, $\vec{x} = \{x_1, \dots, x_n\}$, 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 $\text{Normal}(2, 3)$. 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 $\vec{x} = (x_1, \dots, x_n)$ 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

$$\vec{x} = \{1, 6, 3, 2, 2, 3, 5, 3, 6, 4, 3, 2, 5, 3, 2, 2, 3, 2, 4, 2\}. \quad (7.1)$$

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.

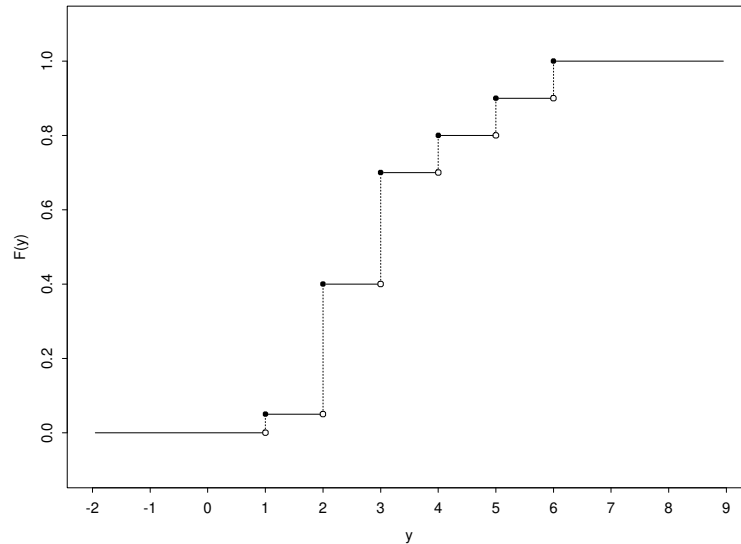


Figure 7.1: An empirical cdf.

In R, one can graph the empirical cdf of a sample \mathbf{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 $X_1, \dots, X_n \sim P$ and that we observe a sample $\vec{x} = \{x_1, \dots, x_n\}$. 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 distribution:*

$$\hat{\mu}_n = \sum_{i=1}^n x_i \cdot \frac{1}{n} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}_n.$$

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 $\sigma^2 = \text{Var } X_i$ 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

$$\begin{aligned}\widehat{\sigma_{20}^2} &= \left(1^2 \times 0.05 + 2^2 \times 0.35 + 3^2 \times 0.30 + \right. \\ &\quad \left. 4^2 \times 0.10 + 5^2 \times 0.10 + 6^2 \times 0.10\right) - 3.15^2 \\ &= 1.9275.\end{aligned}$$

Again, notice that $\widehat{\sigma_{20}^2} \neq \sigma^2$, 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 $X_1, \dots, X_n \sim P$ and that we observe a sample $\vec{x} = \{x_1, \dots, x_n\}$. 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 $\text{Uniform}(1, 5)$. 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.

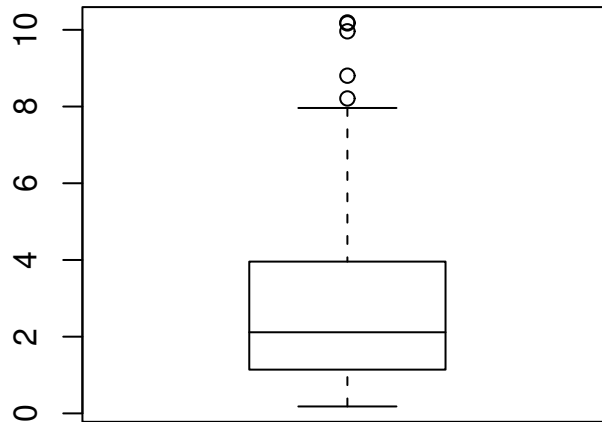


Figure 7.2: A box plot of a sample from $\chi^2(3)$.

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 4 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.

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

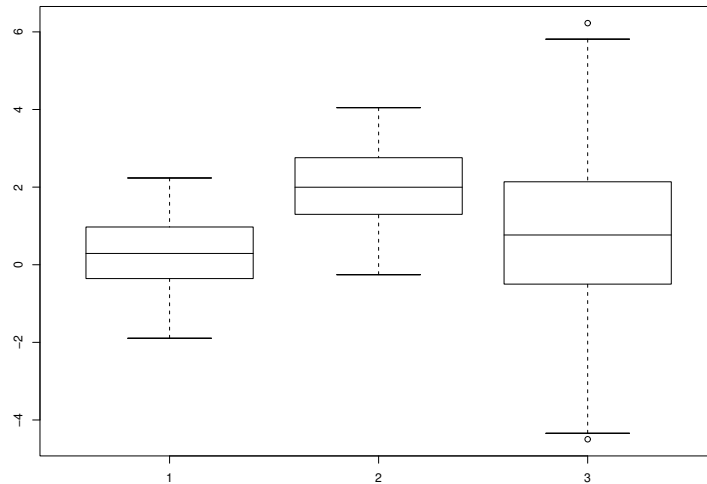


Figure 7.3: Box plots of samples from three normal distributions.

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

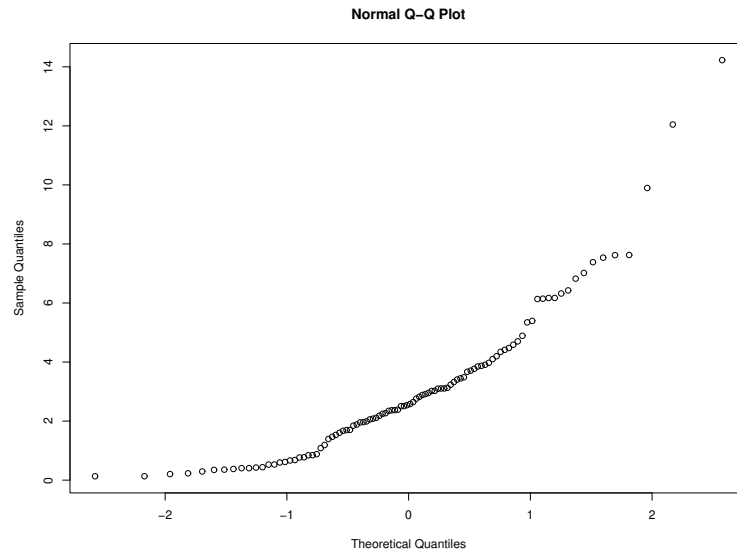


Figure 7.4: A normal probability plot of a sample from $\chi^2(3)$.

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.

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.

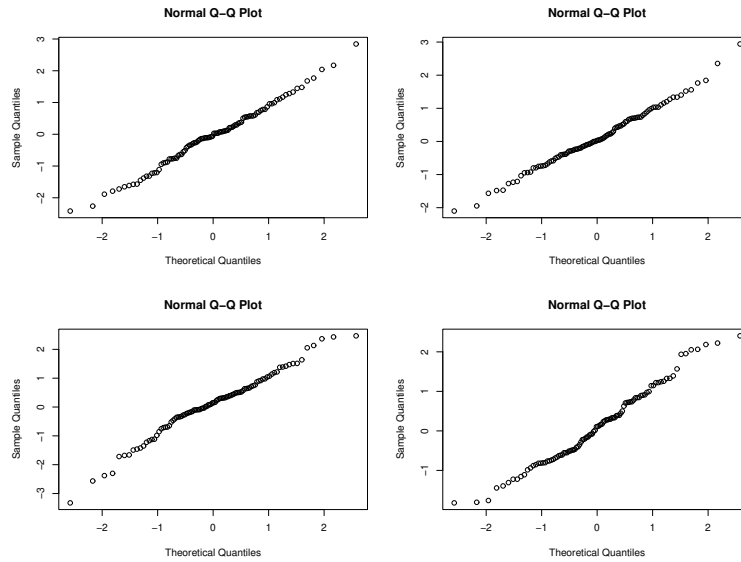


Figure 7.5: Normal probability plots of four samples from $\text{Normal}(0, 1)$.

7.4 Kernel Density Estimates

Suppose that $\vec{x} = \{x_1, \dots, x_n\}$ 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,

$$\frac{1}{h} K\left(\frac{x - x_i}{h}\right), \quad (7.2)$$

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 ?

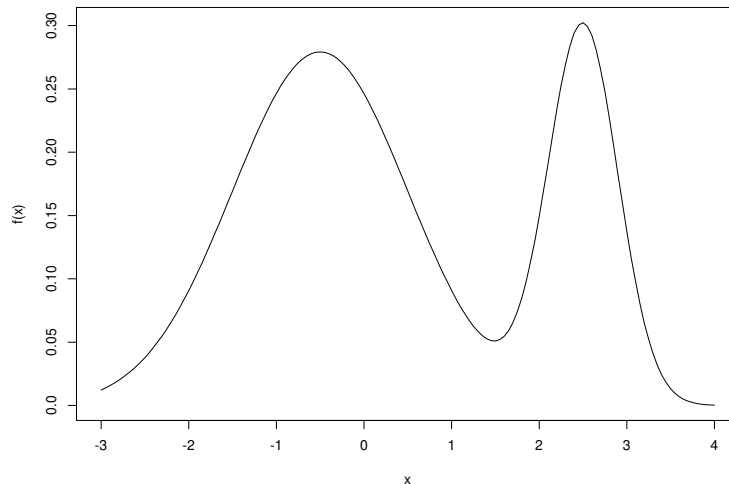


Figure 7.6: A bimodal probability density function.

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 \mathbf{x} , this estimate was computed and plotted using the following R command:

```
> plot(density(x))
```

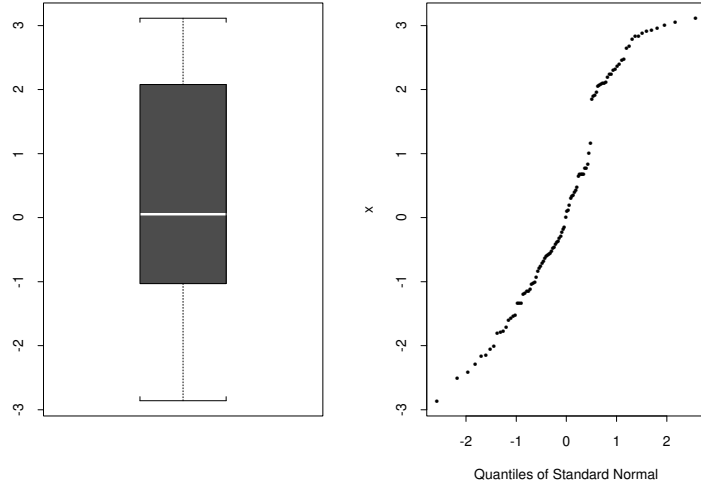


Figure 7.7: A box plot and a normal probability plot for Example 7.7.

7.5 Case Study: Are Forearm Lengths Normally Distributed?

Many of the inferential procedures that statisticians have developed assume that the data to be analyzed were drawn from one (or more) normal distributions. These procedures are often the most elegant and powerful methods available to the data analyst, but they can easily mislead when applied to nonnormal data. Conveniently, the normality assumption is often quite plausible; just as often, however, it is not. It is therefore essential that the data analyst be able to make informed decisions about whether or not to assume normality. One of our primary uses for the methods introduced in the present chapter will be to assist us in making such decisions. Be warned: because we cannot *know* the true distributions of (most) random variables encountered in scientific experimentation, we cannot *know* whether or not they are in fact normally distributed. Such ignorance should humble and can intimidate, but it should not paralyze. To analyze data, one must proceed somehow, and it is best to do so with as much information as possible.

It is often (but not universally) the case that measurements of linear

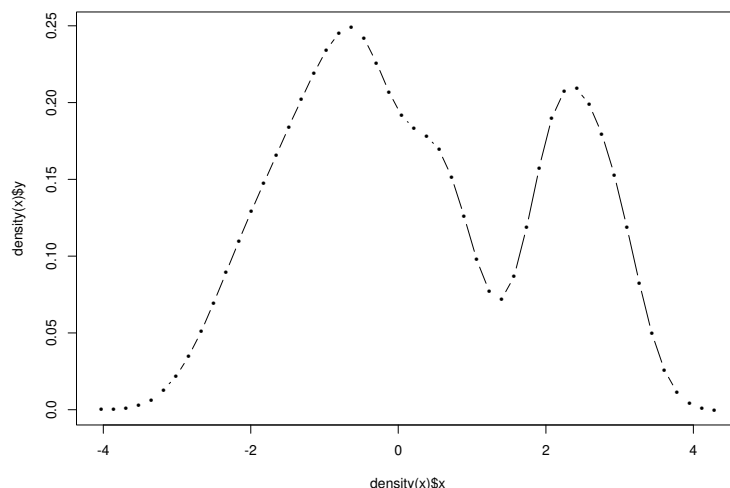


Figure 7.8: A kernel density estimate for Example 7.7.

dimension (height, length, width, depth, breadth, etc.) are normally distributed. To further illustrate the methods introduced in the present chapter, we apply them to a famous data set, measurements of forearm length made on $n = 140$ adult males, inquiring whether or not it appears plausible to assume that forearm lengths are normally distributed. These data, displayed in Table 7.1, were studied by K. Pearson and A. Lee¹ and subsequently reproduced as Data Set 139 in *A Handbook of Small Data Sets*.

Examining the numbers in Table 7.1, we note that the measurements were made with a precision of 0.1 inches and that many values occur several times. For example, 9 of the 140 men had forearms with a measured length of 18.5 inches. Because the probability that any two continuous random variables will be equal is zero, the existence of equal values in the sample should cause one to consider whether or not these measurements should be modelled as observed values of continuous random variables. In this case, it makes sense to proceed. Actual (as opposed to measured) forearm

¹K. Pearson and A. Lee (1903). On the laws of inheritance in man. I. Inheritance of physical characters. *Biometrika*, 2:357–462.

17.3	18.4	20.9	16.8	18.7	20.5	17.9	20.4	18.3	20.5
19.0	17.5	18.1	17.1	18.8	20.0	19.1	19.1	17.9	18.3
18.2	18.9	19.4	18.9	19.4	20.8	17.3	18.5	18.3	19.4
19.0	19.0	20.5	19.7	18.5	17.7	19.4	18.3	19.6	21.4
19.0	20.5	20.4	19.7	18.6	19.9	18.3	19.8	19.6	19.0
20.4	17.3	16.1	19.2	19.6	18.8	19.3	19.1	21.0	18.6
18.3	18.3	18.7	20.6	18.5	16.4	17.2	17.5	18.0	19.5
19.9	18.4	18.8	20.1	20.0	18.5	17.5	18.5	17.9	17.4
18.7	18.6	17.3	18.8	17.8	19.0	19.6	19.3	18.1	18.5
20.9	19.8	18.1	17.1	19.8	20.6	17.6	19.1	19.5	18.4
17.7	20.2	19.9	18.6	16.6	19.2	20.0	17.4	17.1	18.3
19.1	18.5	19.6	18.0	19.4	17.1	19.9	16.3	18.9	20.7
19.7	18.5	18.4	18.7	19.3	16.3	16.9	18.2	18.5	19.3
18.1	18.0	19.5	20.3	20.1	17.2	19.5	18.8	19.2	17.7

Table 7.1: Forearm lengths (in inches) of 140 adult males, studied by K. Pearson and A. Lee (1903).

length is surely continuous, and there are 47 distinct values in Table 7.1. To preserve important numerical relations, e.g., $19.5 - 18.5 = 2(18.5 - 18)$, we can accomplish far more with continuous random variables than we might with discrete random variables. We proceed to investigate the plausibility of assuming that the continuous random variables are normal random variables.

Figure 7.9 displays a box plot, a normal probability plot, and a kernel density estimate, constructed from the 140 forearm measurements in Table 7.1 by the following R commands:

```
> par(mfrow=c(1,3))
> boxplot(forearms,main="Box Plot")
> qqnorm(forearms)
> plot(density(forearms),type="l",main="PDF Estimate")
```

Examining the box plot, we first note that the sample median lies roughly halfway between the first and third sample quartiles, and that the whiskers are of roughly equal length. This is precisely what we would expect to observe if the data were drawn from a symmetric distribution. We also note that these data contain no outliers. These features are consistent with the

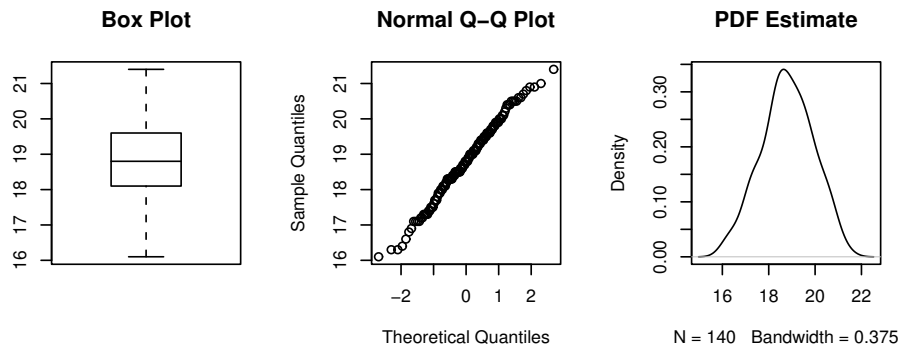


Figure 7.9: Three displays of 140 forearm measurements.

possibility that these data were drawn from a normal distribution, but they do not preclude other symmetric distributions.

Both normal probability plots and kernel density estimates reveal far more about the data than do box plots. More information is generally desirable, but seeing too much creates the danger that patterns created by chance variation will be overinterpreted by the too-eager data analyst. Key to the proper use of normal probability plots and kernel density estimates is mature judgment about which features reflect on the population and which features are due to chance variation.

The normal probability plot of the forearm data is generally straight, but should we worry about the kink at the lower end? The kernel density estimate of the forearm data is unimodal and nearly symmetric, but should we be concerned by its apparent lack of inflection points at ± 1 standard deviations? The best way to investigate such concerns is to generate pseudorandom normal samples, each of the same size as the observed sample (here $n = 140$), and consider what—if anything—distinguishes the observed sample from the normal samples. I generated three pseudorandom normal samples using the `rnorm` function. The four normal probability plots are displayed in Figure 7.10 and the four kernel density estimates are displayed in Figure 7.11. I am unable to advance a credible argument that the forearm sample looks any less normal than the three normal samples.

In addition to the admittedly subjective comparison of normal probabil-

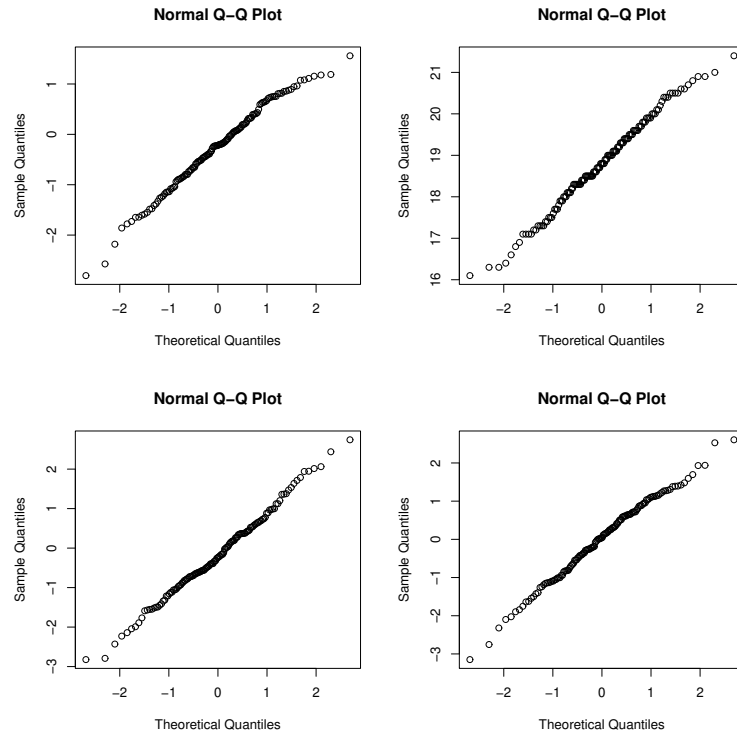


Figure 7.10: Normal probability plots of the forearm data and three pseudorandom samples from $\text{Normal}(0, 1)$.

ity plots and kernel density estimates, it may be helpful to compare certain quantitative attributes of the sample to known quantitative attributes of normal distributions. In Section 6.3, for example, we noted that the ratio of population interquartile range to population standard deviation is $1.34898 \doteq 1.35$ for a normal distribution. The analogous ratio of sample interquartile range to sample standard deviation can be quite helpful in deciding whether or not the sample was drawn from a normal distribution. It should be noted, however, that not all distributions with this ratio are normal; thus, although a ratio substantially different from 1.35 may suggest that the sample was not drawn from a normal distribution, a ratio close to 1.35 does not prove that it was.

7.5. CASE STUDY: ARE FOREARM LENGTHS NORMALLY DISTRIBUTED? 179

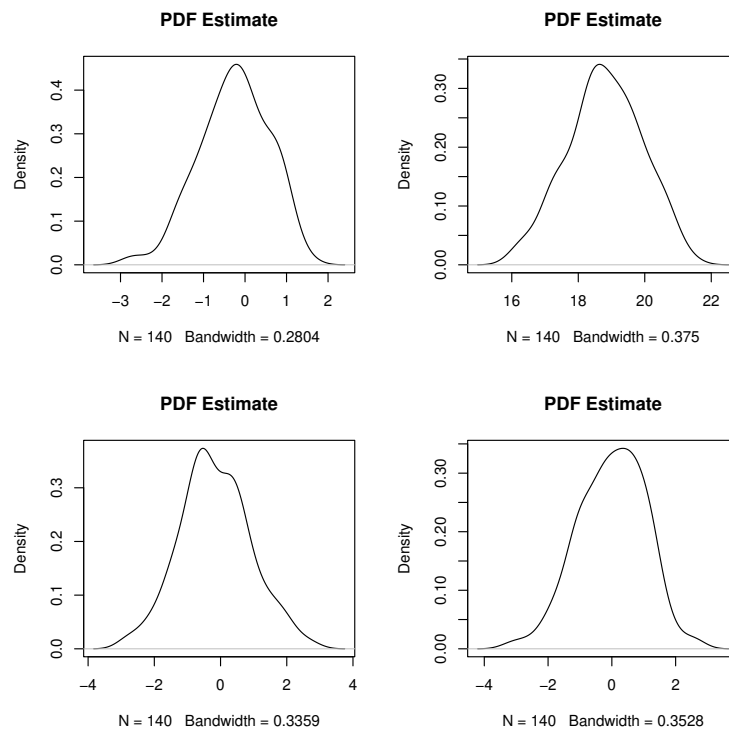


Figure 7.11: Kernel density estimates constructed from the forearm data and three pseudorandom samples from $\text{Normal}(0, 1)$.

To facilitate the calculation of `iqr:sd` ratios, we define an R function that performs the necessary operations. Here are the R commands that define our new function, `iqr:sd`:

```
> iqr:sd <- function(x) {
+ x.mean <- mean(x)
+ x.var <- mean(x^2)-x.mean^2
+ q <- as.vector(quantile(x,probs=c(.25,.75)))
+ x.iqr <- q[2]-q[1]
+ return(x.iqr/sqrt(x.var))
+ }
```

I generated 10 pseudorandom normal samples, each of size $n = 140$, using the `rnorm` function, then applied the new `iqrstd` function to each sample. The resulting ratios ranged from a minimum of 1.178 to a maximum of 1.545. The ratio for the forearm data is 1.344, so one can hardly object to assuming normality on the basis of this quantity.

Overall, the forearm data look about as normal as one could ever hope to encounter with actual experimental data. If one is hoping to use an inferential procedure that assumes normality, then this is the ideal case. Unfortunately, one rarely encounters situations in which one can so comfortably assume normality.

7.6 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?
 - (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 1.5.2. (Two students reported more than one measurement, but only one measurement per student is reported here.)

$143\frac{3}{16}$	$144\frac{4}{16}$	$140\frac{14}{16}$	$144\frac{7}{16}$	$143\frac{12}{16}$	$153\frac{13}{16}$
$119\frac{10}{16}$	$143\frac{1}{16}$	$143\frac{14}{16}$	$144\frac{3}{16}$	$144\frac{7}{16}$	$148\frac{3}{16}$

- (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. Consider a box that contains 10 tickets, labelled

$$\{1, 1, 1, 1, 2, 5, 5, 10, 10, 10\}.$$

From this box, I propose to draw (with replacement) $n = 40$ tickets. I am interested in the sum, Y , of the 40 ticket values that I draw. Write an R function named `box.model` that simulates this experiment, i.e., evaluating `box.model` is like observing a value, y , of the random variable Y .

7. 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 $X_1, \dots, X_n \sim P$ and that an experimental scientist wants to estimate the population mean, $\mu = EX_i$. To do so, she observes values x_1, \dots, x_n of X_1, \dots, 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, \dots, 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 $X_1, \dots, X_5 \sim \chi^2(3)$ 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 $X_1, \dots, X_{20} \sim \chi^2(3)$ 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 $X_1, \dots, X_{80} \sim \chi^2(3)$ 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

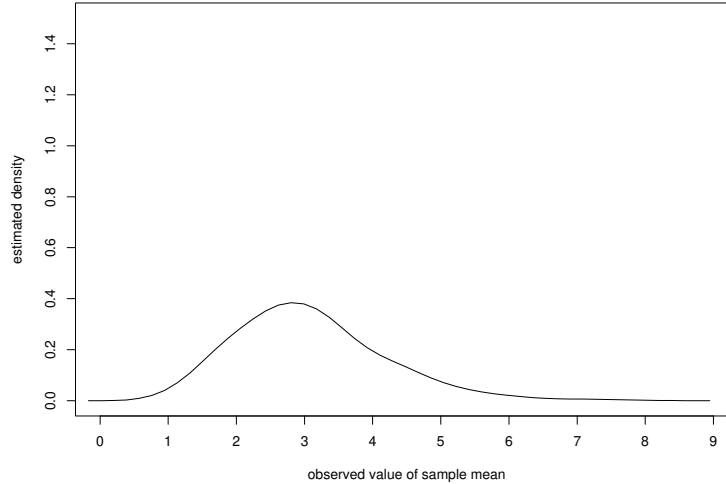


Figure 8.1: Kernel density estimate constructed from 1000 observed values of \bar{X}_n for $n = 5$. $X_1, \dots, X_n \sim \chi^2(3)$ and $\mu = EX_i = 3$.

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 large 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, \dots , and we assume that these random variables have a finite mean $\mu = EX_i$ and a finite variance $\sigma^2 = \text{Var } X_i$. 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

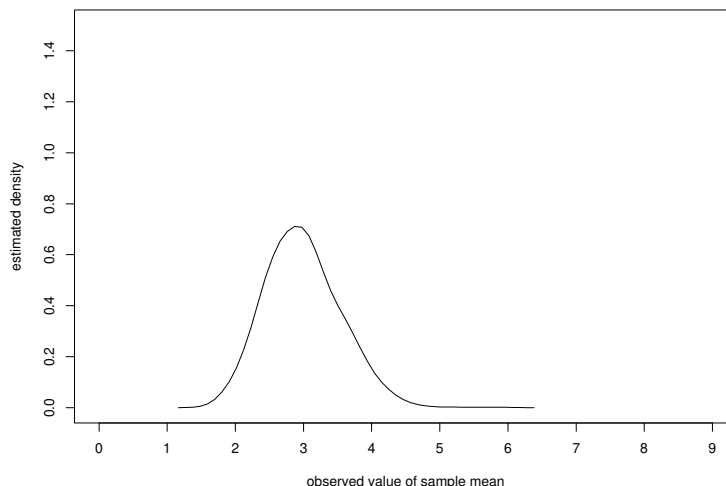


Figure 8.2: Kernel density estimate constructed from 1000 observed values of \bar{X}_n for $n = 20$. $X_1, \dots, X_n \sim \chi^2(3)$ and $\mu = EX_i = 3$.

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, $\text{Var } X_i = \sigma^2$, but

$$\text{Var } \bar{X}_n = \text{Var} \left(\frac{1}{n} \sum_{i=1}^n X_i \right) = \frac{1}{n^2} \sum_{i=1}^n \text{Var } X_i = \frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \frac{\sigma^2}{n}.$$

Hence, the sample mean has less variability than any of the individual random variables that are being averaged. *Averaging decreases variation.* Furthermore, as $n \rightarrow \infty$, $\text{Var } \bar{X}_n \rightarrow 0$. Thus, by repeating our experiment enough times, we can make the variation in the sample mean as small as we please.

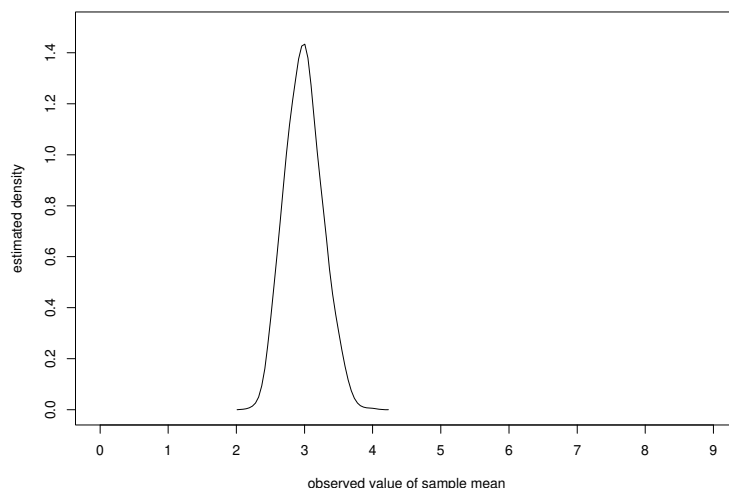


Figure 8.3: Kernel density estimate constructed from 1000 observed values of \bar{X}_n for $n = 80$. $X_1, \dots, X_n \sim \chi^2(3)$ and $\mu = EX_i = 3$.

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 $c \in \mathfrak{R}$ 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 $\epsilon > 0$,

$$\lim_{n \rightarrow \infty} P(Y_n \in (c - \epsilon, c + \epsilon)) = 1.$$

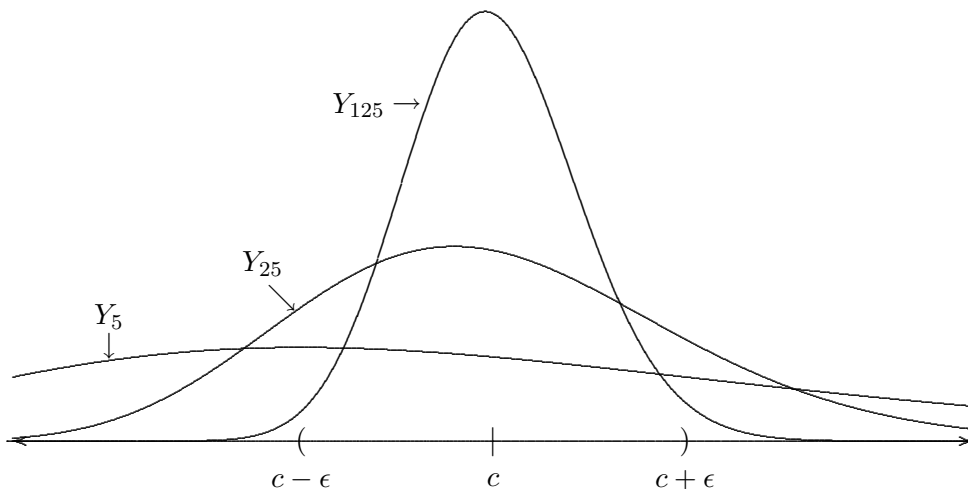


Figure 8.4: An example of convergence in probability.

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

$$p_n = P(Y_n \in (c - \epsilon, c + \epsilon)) = \int_{c-\epsilon}^{c+\epsilon} f_n(x) dx$$

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, \dots 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*

$$X_i = \begin{cases} 1 & A \text{ occurs} \\ 0 & A^c \text{ occurs} \end{cases}.$$

Then $X_i \sim \text{Bernoulli}(p)$, \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, \dots , 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, \dots, 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	σ / \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 $X_i \sim \text{Bernoulli}(0.5)$, 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, \dots, X_n . Notice that, after converting to standard units, these quantities are identical:

$$Z_n = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma} = \frac{(1/n) \sum_{i=1}^n X_i - n\mu}{(1/n) \sqrt{n}\sigma} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}.$$

It is this new random variable on which we shall focus our attention.

We begin by observing that

$$\text{Var} [\sqrt{n} (\bar{X}_n - \mu)] = \text{Var} (\sigma Z_n) = \sigma^2 \text{Var} (Z_n) = \sigma^2$$

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, \dots 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 $z \in \mathbb{R}$,

$$P(Z_n \leq z) = F_n(z) \rightarrow \Phi(z)$$

as $n \rightarrow \infty$.

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 $i = 1, \dots, 36$. 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,

$$\begin{aligned} P(\mu - 0.1 < \bar{X}_{36} < \mu + 0.1) &= P(\mu - 0.1 - \mu < \bar{X}_{36} - \mu < \mu + 0.1 - \mu) \\ &= P\left(\frac{-0.1}{0.5/6} < \frac{\bar{X}_{36} - \mu}{0.5/6} < \frac{0.1}{0.5/6}\right) \\ &= P(-1.2 < Z_n < 1.2) \\ &\doteq P(-1.2 < Z < 1.2) \\ &= \Phi(1.2) - \Phi(-1.2). \end{aligned}$$

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 within 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

$$\sum_{i=1}^n X_i \sim \text{Normal}(n\mu, n\sigma^2) \quad (8.1)$$

and for the average we obtain

$$\bar{X}_n \sim \text{Normal}\left(\mu, \frac{\sigma^2}{n}\right). \quad (8.2)$$

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

$$X_1, \dots, X_{36} \sim (\mu, 0.5^2).$$

Applying (8.2), we obtain

$$\bar{X}_{36} \sim \text{Normal}\left(\mu, \frac{0.25}{36}\right).$$

Similarly, the student's measurements are

$$Y_1, \dots, Y_{64} \sim (\mu, 0.5^2).$$

Applying (8.2), we obtain

$$\bar{Y}_{64} \sim \text{Normal}\left(\mu, \frac{0.25}{64}\right) \quad \text{or} \quad -\bar{Y}_{64} \sim \text{Normal}\left(-\mu, \frac{0.25}{64}\right).$$

Now we apply Theorem 5.2 to conclude that

$$\bar{X}_{36} - \bar{Y}_{64} = \bar{X}_{36} + (-\bar{Y}_{64}) \sim \text{Normal}\left(0, \frac{0.25}{36} + \frac{0.25}{64} = \frac{5^2}{48^2}\right).$$

Converting to standard units, it follows that

$$\begin{aligned} P(-0.1 < \bar{X}_{36} - \bar{Y}_{64} < 0.1) &= P\left(\frac{-0.1}{5/48} < \frac{\bar{X}_{36} - \bar{Y}_{64}}{5/48} < \frac{0.1}{5/48}\right) \\ &\doteq P(-0.96 < Z < 0.96) \\ &= \Phi(0.96) - \Phi(-0.96). \end{aligned}$$

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 within 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, \dots 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, \dots 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 $t \in \mathbb{R}$,

$$P(T_n \leq t) = F_n(t) \rightarrow \Phi(t)$$

as $n \rightarrow \infty$.

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, \dots, 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 $X_1 \sim \text{Bernoulli}(0.5)$. The Bernoulli distribution is not at all like a normal distribution. If we toss the coin one million times, then each $X_i \sim \text{Bernoulli}(0.5)$. 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.

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. Consider a box that contains 10 tickets, labelled

$$\{1, 1, 1, 1, 2, 5, 5, 10, 10, 10\}.$$

From this box, I propose to draw (with replacement) $n = 40$ tickets. Let Y denote the sum of the values on the tickets that are drawn.

- (a) To approximate $P(170.5 < Y < 199.5)$, one Math 351 student writes an R function `box.model` that simulates the proposed experiment. Evaluating `box.model` is like observing a value, y , of the random variable Y . Then she writes a loop that repeatedly evaluates `box.model` and computes the proportion of times that `box.model` produces $y \in (170.5, 199.5)$. She reasons that, if she evaluates `box.model` a large number of times, then the observed proportion of $y \in (170.5, 199.5)$ should approximate $P(170.5 < Y < 199.5)$. Is her reasoning justified? Why or why not?
- (b) Another student suggests that $P(170.5 < Y < 199.5)$ can be approximated by performing the following R commands:

```
> se <- sqrt(585.6)
> pnorm(199.5, mean=184, sd=se) -
+ pnorm(170.5, mean=184, sd=se)
```

Do you agree? Why or why not?

- (c) Which approach will produce the more accurate approximation of $P(170.5 < Y < 199.5)$? Explain your reasoning.
6. 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, \dots , with $EX_i = 0.01$ and $\text{Var } X_i = 0.01$. (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, \dots , with $EY_j = 0$ and $\text{Var } Y_j = 0.25$.

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.