Serious Stats

A Guide to Advanced Statistics for the Behavioral Sciences

Thom Baguley

SERIOUS STATS

SERIOUS STATS

A GUIDE TO ADVANCED STATISTICS FOR THE BEHAVIORAL SCIENCES

THOM BAGULEY
Professor of Experimental Psychology, Nottingham Trent University, UK

macmillan international HIGHER EDUCATION
RED GLOBE PRESS

First published 2012 by
RED GLOBE PRESS

Red Globe Press in the UK is an imprint of Springer Nature Limited,
registered in England, company number 785998, of 4 Crinan Street,
London, N1 9XW.

Red Globe Press® is a registered trademark in the United States,
the United Kingdom, Europe and other countries.

ISBN 978–0–230–57718–3 ISBN 978–0–230–36355–7 (eBook)

This book is printed on paper suitable for recycling and made from fully
managed and sustained forest sources. Logging, pulping and manufacturing
processes are expected to conform to the environmental regulations
of the country of origin.

A catalogue record for this book is available from the British Library.

A catalog record for this book is available from the Library of Congress.

Contents overview

Contents

List of tables

List of figures

List of boxes

List of key concepts

Preface

About this book

This book is a bridging text for students and researchers in the human and behavioral sciences. The ideal reader will have some familiarity with inferential statistics – perhaps as part of an undergraduate degree in a discipline such as psychology, cognitive science or ergonomics – and be interested in deepening their understanding or learning new material. This book aims to bridge the gap between a reader's existing understanding of statistics and that required to apply and interpret more advanced statistical procedures.

I have also tried to make the book a helpful resource for experienced researchers who wish to refresh their statistical knowledge or who have good understanding of a 'narrow' but fairly advanced topic such as analysis of variance. I hope it will allow these readers to expand from islands of existing expertise to new territory.

The book starts with a review of basic inferential statistics, beginning with descriptive statistics, probability distributions and statistical inference (in the form of confidence intervals and significance tests). If you are already familiar with these topics I would encourage you to look through these chapters to refresh your understanding. In addition, this material may be presented in a slightly different way (e.g., from a different perspective or in greater depth).

Later chapters introduce core topics such as, regression, correlation and covariance, effect size, and statistical power. Unless you have advanced training in statistics it is likely that you will benefit from looking closely at this material – it is fundamental to an appreciation of later content. Two further chapters consider the messiness inherent in working with real data (particularly data from human participants). The approach I adopt is to give a taster of some methods for exploring and dealing with messy data, rather than provide a comprehensive recipe for checking and solving every possible problem. This is both for practical reasons (as each of these chapters could be a book in its own right) and because the best approach in any particular situation depends on what you are trying to do and the context from which the data are drawn.

Later chapters cover what I consider to be advanced material: multiple regression, analysis of variance, analysis of covariance, and the general linear model. Before covering these topics I review alternatives to classical, frequentist inference (and significance tests in particular). In order to get the most out of the more advanced material in the book, you will need to understand the problems inherent in relying (solely) on a p value from a significance test for inference. I also think it important to go beyond criticism of the p value approach and present viable alternatives. Three are presented here: Bayesian, likelihood, and information theoretic approaches to inference. There are important connections (and distinctions) between these three approaches. In this chapter, I sacrifice depth for breadth (though there is sufficient material to run a range of analyses using each approach).

The final chapters explore the most challenging topics. Also included are chapters on interaction effects and contrasts. These topics are extremely important for researchers in the human and behavioral sciences, but are often covered only briefly (if at all) in introductory classes. My goal here is to remedy this deficit. The final two chapters introduce generalized linear models

(for discrete outcomes) and multilevel models (with emphasis on repeated measures models). I have tried to emphasize the links between these advanced topics and the general linear model and to demonstrate what they offer over and above simpler models.

If you want to learn and understand what is covered, it is essential that you have a go at applying it. Each chapter contains worked examples. Many of these use real data sets. These are necessarily a bit messy and don't always lead to clear answers (and several data sets are chosen because they have interesting or unusual quirks). My aim is to illustrate some of the challenges of working with real data sets (and the importance of data exploration and model checking). In other cases I have resorted to creating artificial data sets to illustrate a particular point, or to make it easier to conduct calculations by hand. These data sets are carefully constructed to meet the requirements of the example – though you will sometimes encounter real data sets with similar properties. In general, the early examples use hand calculation while later examples require you to use a computer.

Hand calculation can sometimes help you to understand how an equation works or demystify a (supposedly) complex technique. This will depend on your confidence and ability with basic mathematical operations. This doesn't work with every procedure, and in many examples I explain how to use a computer package to provide intermediate values that, when put together in the right way, illustrate what is going on. From time to time the mathematics is sufficiently challenging that I merely describe the gist of what is happening (and rely on the computer to provide a complete solution). Where necessary, I refer interested readers to a more detailed mathematical account of what is taking place.

The contents of the book differ from the coverage of a typical introductory or intermediate statistics course in the behavioral or human sciences. One difference is the breadth of coverage, which runs from descriptive statistics to generalized linear and multilevel models. Another is the reduced emphasis on null hypothesis significance tests and increased emphasis on confidence intervals or other inferential tools. Several topics have more prominence than you might expect: graphical methods; effect size; contrasts and interactions. Other topics have less emphasis or are presented differently: psychometrics, multivariate analysis of variance, non-parametric statistics; and pairwise comparisons. I have chosen to focus on univariate methods – methods for the analysis of a single outcome measure (though there may be many predictor variables). Covering multivariate statistics (in the sense of modeling multiple outcomes) and psychometrics would probably have doubled the page count. Nonparametric statistics are covered, but in an atypical way. Several methods, often considered to be 'nonparametric' (e.g., bootstrapping, kernel density estimation, the rank transformation and robust regression), are integrated into the text at appropriate points. I have, in particular, avoided describing a large number of rank transformation tests in detail. My preference is to emphasize the link between parametric and so-called non-parametric approaches and to encourage consideration of robust methods as alternatives to the usual (e.g., least squares) models.

If there is a single message to take away from this book, it is that statistical modeling is not a set of recipes or instructions. It is the search for a model or set of models that capture the regularities and uncertainties in data, and help us to understand what is going on.

Software

Many of the statistical tools described in the book require specialist software to run them. Nearly all the examples were implemented in the free, open source statistical programming

environment R (R Core Development Team, 2011). This will run on PC, Mac and Linux operating systems. Installation on a Mac or PC is generally very easy. For details on downloading and installing R see: http://cran.r-project.org/

R works slightly differently on PC, Mac and Linux machines and for this reason I have, for the most part, avoided referring to platform-specific features of R (e.g., resizing windows, printing, opening or saving files). There are many online guides and books that run through the basics of installing and running R (in addition to the information available when you download R for the first time).

At the end of each chapter, I provide a detailed description of the R code used to reproduce the examples in that chapter. Data sets and R scripts for each chapter are available with the online resources for this book. I assume no previous knowledge of R (and only limited familiarity with statistical computing). Although I am not trying to teach R *per se*, I have tried to include enough explanation of how R works for readers, if they wish, to learn it as they go along. To this end, the complexity of the R code being used increases gradually from Chapter 1 through to Chapter 18. In some cases I have glossed over fine grain technical details about how R works (e.g., the difference between 'modes' of vector) and used generic terms such as 'string' alongside R-specific terms such as 'data frame'.

As well as R code, I provide very brief notes on relevant SPSS syntax at the end of each chapter (where relevant). SPSS is the most widely used statistics package in the human and behavioral sciences (in the UK at least). These notes are included for two reasons: (1) to reveal some of the hidden features and capabilities of SPSS, and (2) to highlight the advantages of using R alongside or in place of SPSS.

Mathematics

To get the most from this book you will need to have basic mathematical competence. I assume readers will have mastered basic arithmetic (e.g., addition, subtraction, division and multiplication) and be familiar with concepts such as, fractions, decimals, rounding, negative numbers, squares and square roots, and cubes and cubed roots. Knowing the order in which to apply arithmetic operations (e.g., PEMDAS or BODMAS) is also necessary. You should also have some understanding of percentages, probabilities and ratios (and perhaps reciprocals, exponents, logarithms and factorials) and simple algebra. If you are rusty on any of these topics don't worry too much – as there will be some 'refresher' material in each chapter.

I would expect readers to be able to answer the following arithmetic problems without much difficulty:

$$4+3\times 5=? \quad 6(3-1)=? \quad \sqrt{9}=?$$
$$10^2=? \quad \frac{9}{2}=? \quad 0.5\times 0.5=?$$

I would also expect readers to understand what the following equations mean, and (perhaps with a little help) be able to solve them:

$$3x+1=10\Rightarrow x=? \quad 4!=? \quad \sqrt{2}\times\sqrt{2}=?$$
$$P(A)+P(\sim A)=? \quad \sqrt[3]{27}=? \quad -2\times -3=?$$

In addition to mathematical competence, I anticipate some familiarity with data collection, exploration and analysis. This may include experimental design, descriptive statistics and simple graphical methods such as line graphs or scatter plots (and technical terms such as *x*-axis and *y*-axis). Again, the text contains refresher material on most of these topics.

Many of the examples refer to 'hand calculation'. This is a fuzzy concept, but I take it to mean reproducing the calculations step-by-step as if doing them by hand. It therefore includes using paper and pencil, mental arithmetic, pocket calculators or spreadsheets (if used in the correct way). In fact, a spreadsheet is one of the best ways to organize hand calculations to understand what they are doing (provided you know how to use one and are careful to set out all the intermediate steps).

Boxed sections and online supplements

As well as learning features (such as examples, sections on R code, and SPSS notes), there are two types of boxed section used throughout the book. One type covers key concepts or important ideas that are referred to in several different chapters. These are referred to as 'key concepts' and numbered by chapter and serial position within chapter (e.g., Key Concept 2.1 is the first key concept box in Chapter 2). The other is a more traditional boxed section that is used to improve the flow of the text and contains material that is relatively self-contained (and that is generally referred to again only within that chapter). These are referred to as 'boxes' and numbered by chapter and serial position within chapter (e.g., Box 1.2 is the second boxed section in Chapter 1).

In addition to the boxed sections there are five online supplements. Supplements 1, 2 and 5 cover advanced topics that are not central to the text, but will be very useful for some readers (meta-analysis, dealing with missing data and loglinear models). Supplements 3 and 4 provide more detail on peripheral topics that are mentioned in the main text (replication probabilities and pseudo-R^2 measures).

Acknowledgments

I would like to thank the following people who have contributed suggestions for the book, read chapters or provided encouragement: Jaime Marshall; Andrew Dunn; Danny Kaye; Mark Andrews; Stanislav Kolenikov; Dave Atkins; Tim Wells; Tom Dunn; Mark Torrance; James Houston and Zoltan Dienes.

THOM BAGULEY

1
Data, samples and statistics

Contents

1.1 Chapter overview

The aim of this chapter is to review some basic statistical ideas, with particular emphasis on the use of descriptive statistics to explore data. The main focus is on statistics for summarizing the central tendency and dispersion of a sample. Key ideas introduced here include the distinction between sample statistics and population parameters, and between inferential or descriptive statistics.

1.2 What are data?

To understand what data are, it helps to consider the distinction between *numbers* and *data*. Numbers are abstract tokens or symbols used for counting or measuring. Data are numbers that represent 'real world' entities.[1] The crucial feature that distinguishes data from numbers is that they are connected to a particular context and acquire meaning from that connection.

Example 1.1 Take a look at this set of numbers (labeled D_1 for later reference):

12 14 9 11 15 11 D_1

Their interpretation would be very different if they described the ages (in years) of six children than if they described the number of words remembered by a participant from a series of six 20-word lists. Not only is the context vital in understanding what the numbers describe, it also has profound implications for what you might want to do with them (and on the subsequent findings of a statistical analysis). Knowing that 11 is an age (in years) makes it reasonable to represent it as $11 \times 12 = 132$ months. Alternatively, knowing that the six numbers represent repeated memory measures from the same individual makes it likely that the numbers can be considered as some combination of the participant's learning ability, improvement with practice and chance factors (influences that are potential components of what, in statistics, is usually termed *error*).

The context, in turn, depends on the process that generated the data. This process can for many sciences be characterized as collecting a subset (a *sample*) of observations from a larger set (the *population*) of observations that are of interest. The idea of taking a sample from a population is central to understanding statistics, and at the heart of most statistical procedures.

Working with samples is attractive to researchers because the populations themselves are usually considered to be infinitely large (and so beyond reach). Even where a population might reasonably be considered finite it is rarely possible, in practice, to sample the whole population. This presents a fundamental difficulty for researchers. A sample, being a subset of the whole population, won't necessarily resemble it. Therefore, the information the sample provides about the population is inherently uncertain. Statistics involves finding ways to deal with this uncertainty. For example, the uncertainty can be quantified and expressed in terms of *probability* (see Box 1.1).

Box 1.1 Probability

There are many ways to represent the uncertainty of an event numerically, but probability is the most common. A probability is a number between zero and one, where one indicates that the event is certain to occur and zero that it is certain not to occur. The probability of an event x can be written as $P(x)$ or $Pr(x)$.

A probability can be interpreted in several ways, but a reasonable place to start is to consider the probability as the relative frequency with which an event such as x occurs in the long run. For instance, if the event H was the occurrence of heads on tossing a fair coin then, in the long run, equal numbers of heads and tails would be observed, and $Pr(H)$ would be 0.5. For example, if a fair coin were tossed 1 million times you'd expect to see 500,000 heads out of 1 million coin tosses and so $Pr(H) = 500,000/1,000,000 = 0.5$.

The problem of dealing with the uncertainty inherent in taking a sample from a population is fundamental to understanding even the simplest statistical tools – the descriptive statistics that are the focus of this chapter. The next section considers these issues in a little more detail, before turning to consider a range of tools for describing and summarizing data.

1.3 Samples and populations

An important point to understand about the concept of a population in statistics is that it is an abstraction. Rarely, if ever, does it refer to a particular set of things (e.g., objects or people). The customary assumption is that samples are drawn from an infinitely large, hypothetical population defined by the *sampling procedure*. A well-designed study will use (or attempt) a sampling procedure that draws from a population that is relevant to the aims of the research. For most (and perhaps all) research, the sampling procedure is imperfect and introduces potential bias into the sample (e.g., because not every member of the population has an equal chance of being chosen). Therefore, the sample will almost never match the intended population exactly. A good study is one that minimizes these problems and thus limits their impact on the statistical model and on the research findings.

Treating a sample as drawn from an infinite population may at first seem unreasonable. However, it represents a fairly cautious position for a researcher to take in practice. Before we examine why, we need to introduce a few technical terms. The first term is the *sample size* – the number of observations (data points) in a sample – usually abbreviated to n. The sample size can, in theory, vary from one to ∞ (infinity). The larger the population, the less information (proportionately) a sample of size n provides about the population of size N.[2] The *sampling fraction* is the ratio of sample to population size: n/N. In theory, the larger the sampling fraction the more closely the sample matches the population and the more likely that characteristics of the sample are also true of the population. Therefore, treating the population as infinitely large is a very cautious option, one that regards the sampling fraction as negligible. The consequence of this cautious position is that conclusions drawn from looking at the sample are assessed more carefully before deciding that they are likely to generalize to the population. The practical limits on generalization from a statistical model depend on the adequacy of the sampling procedure in relation to the objective of the research.

For example, if a researcher collects data from an opportunity sample of 100 healthy people from the city of Nottingham this limits the generalizability of any findings (e.g., to people broadly similar to those making up the sample). For some research questions, this restriction would be severely limiting, but for others it would not. Assessing the adequacy of the sample depends on extra-statistical factors – notably an understanding of the research domain. This is why it helps greatly if researchers have routinely obtained information about the context of the data (e.g., demographic data about human participants). An opportunity sample from Nottingham might be adequate to assess the impact of caffeine on simple reaction time, but not to determine the relative popularity of different UK soccer teams. For the caffeine example it is probably reasonable to assume a certain degree of similarity in physiological response to caffeine among healthy adults. For the football example the sample is likely to be biased (e.g., by support for local teams).

1.3.1 Exploring data

There are many good reasons to explore data, but a very important one is to understand the relationship between a sample and the population from which it is drawn. In order to extrapolate from information in any sample it is necessary to have at least some knowledge of that population. This process of extrapolation from sampled data is known as statistical generalization. The methods used are termed *statistical inference* or *statistical modeling* depending on whether the primary interest is in testing a specific hypothesis or in understanding the process that generated the observed data (e.g., by predicting new observations). This book looks broadly at statistical modeling – building a statistical model of the process that generated the observed data. Statistical inference is a special case of statistical modeling where the primary purpose of the model (perhaps the only purpose) is to test a specific hypothesis.

In combination with graphical techniques, descriptive statistics form the core methods of *exploratory data analysis* (Tukey, 1977). Exploratory analyses are used to become familiar with a data set and will often throw up specific hypotheses (e.g., potential explanations of what is happening). In contrast, *confirmatory data analysis* is employed to test hypotheses. Sometimes these are derived from scientific theory, but they also often emerge from exploratory analyses. Although this distinction is useful, it is not always clear-cut. In particular, thoughtful use of descriptive statistics and graphical techniques can be a very powerful method for testing hypotheses, while confirmatory analyses sometimes lead to reinterpretation of data (e.g., when checking the quality inferences or predictions).

Descriptive statistics, also called *summary statistics*, are an excellent starting point for most statistical analyses and are a good way to summarize and communicate information about a data set. In some situations, descriptive statistics are sufficient to settle a research question (e.g., on the rare occasions when the sample comprises most or all of the whole population of interest). For example, if you want to know what proportion of babies are male and what proportion female it is probably sufficient to look at descriptive statistics for hospital births (the proportion of males is between 0.48 to 0.49). It can also happen that patterns in the data are strong or clear enough to support inferences using descriptive statistics (e.g., that men tend to be taller than women). However, in my view, the main role for descriptive statistics is to get a feel for a data set. A lot of time and effort can be saved and many mistakes avoided by even a quick exploratory analysis of data. Using appropriate descriptive statistics and graphical methods will often catch basic problems before they cause any serious trouble and will help guide you toward

an appropriate statistical model. Take three of the most elementary descriptive statistics: the sample size (*n*), the minimum (*min*) and the maximum (*max*). These will often reveal problems in coding, transcribing or data entry errors.

Example 1.2 Imagine that we have collected data from 100 students in an introductory statistics class. These students all rate their understanding of statistics after completing the class on a scale from one ('no understanding') to seven ('excellent understanding').

The sample size, minimum and maximum for the ratings appear as follows in computer output:

n	101
min	1
max	77

This sort of output is fairly common for manual entry of data onto a computer. Although there were 100 students the apparent sample size is 101. This is most likely because one of the ratings was entered into the computer twice by accident. The *min* of one is plausible (though disappointing), but the *max* of 77 is a clear error – probably arising from hitting a key twice in computer entry. While it is always good practice to check data entry (even if the data are plausible) these descriptive statistics alert us to serious mistakes.

This may seem like mere common sense – hardly worth mentioning – but trivial errors such as these are often missed (even by experienced researchers). They are also more likely to be missed in a complex analysis – where an unusual outcome may be attributed to all sorts of other causes. There is much more to exploratory analysis than this, but embarking on a statistical analysis without getting the basics right is extremely dangerous.

1.3.2 Types of data

Different contexts provide us with different kinds of data. One of the simplest and most important distinctions is between *discrete* and *continuous* data. Discrete data are restricted in the values that can legitimately occur. For example, binary discrete data can take on only two possible values (usually represented as zero or one). Another common type of discrete data used in research is *frequency* data – often known as *count* data (because it involves counting things). Continuous data can take on intermediate values within a given range, for example, physical measures, such as, time and distance can (in principle) take on any value from zero to infinity. The difference between two such measures can therefore range between minus infinity and infinity.

A widely taught, but controversial, distinction was proposed by Stevens (1946; 1951). He proposed *scales of measurement* that classify data as *nominal* (also known as *categorical*), *ordinal*, *interval* or *ratio*. He argued that these measurement scales are derived from the underlying relationships between the numbers used to represent a data set. Furthermore, he argued that they limit the mathematical operations that are permitted on data of a given type. Nominal data can be represented by numbers, but the relationship between the numbers is arbitrary (e.g., assigning one to represent blue eye colour and two to represent brown eye colour). If data are ordinal, the numbers preserve information about the relative magnitude of what is measured, but not the absolute magnitude (e.g., data about ages are often collected in the form of age groups or

bands with one representing 21–30-year olds, two representing 31–40-year olds and so forth). Interval data preserve continuous, linear relationships between what is measured (e.g., temperature in degrees Centigrade). This means that a given interval between two numbers on the scale (e.g., $5 - 3 = 2$) is equivalent to any other interval of the same magnitude (e.g., $10 - 8$). Ratio scales are interval measurements that have a 'true' zero point (e.g., weight in kilograms or temperature in degrees Kelvin). This means that the number zero represents the point at which the quantity being measured is absent (i.e., nothing is left).

According to this scheme, nominal data is limited to operations, such as counting and ordinal data, to operations, such as placing numbers in rank order, whereas interval data also permits addition and subtraction. Ratio scales permit the full range of arithmetic operations and, as the name suggests, allow meaningful ratios between numbers on the scale to be constructed (e.g., $10/5 = 2$ implies ten is twice as large as five). Ratio scales are probably quite rare for simple measurements, but an interesting observation is that the difference between two numbers on an interval scale is a ratio (because zero represents a 'true' absence of the difference).

There are many critiques (and some defenses) of Stevens' measurement scales (e.g., see Velleman and Wilkinson, 1993, for an overview). Among the more cogent criticisms is the observation that a measurement scale is not a fixed property of data – it also depends on the use to which the data are put. Lord (1953) used the example of football shirt numbers. For many purposes they would be considered nominal data, but it is easy to imagine situations where the numbers convey additional information (perhaps because players derive status from lower or higher numbers, or because they indicate the order in which players joined a team). A major drawback is that the system may also lead people to neglect rather important characteristics of their measurements. Many measurements are bounded in some way (e.g., at zero). Such limits are often much more important for both theoretical and practical purposes when selecting a statistical model or procedure. For example, a statistical model that predicts impossibly low or high values for a measure is problematic (though it may be adequate for some purposes).

Understanding the context of the data that have been sampled and being sensitive to the constraints that context places on a statistical model is important. Classification of data in types (such as those proposed by Stevens) is probably not the best way to go about this. Classification schemes inevitably lose information about the context, so using them in a rigid way to determine what to do is dangerous. Velleman and Wilkinson (1993) go as far as to say, "the single unifying argument against proscribing statistics based on scale type is that it does not work".

An alternative approach – that advocated here – is to consider a range of factors of data that impact on the statistical model you are considering. These factors include whether data are discrete or continuous, but other factors, such as the probability distribution being assumed, the size of the sample and what the model is being used for, are also important. Later chapters will consider several of these factors in greater detail.

1.4 Central tendency

One way to describe data is to reduce it to a single number; a number that is in some way typical or representative of the data as a whole. This corresponds to the everyday notion of an *average*; a notion that encompasses a range of meanings from 'typical', 'most common' to 'mediocre'.

No single way to communicate central tendency will work for all data sets, so it is useful to distinguish between different measures of central tendency. Some common measures, such as the *mode*, *median* and *mean* (and a few less widely known ones) are reviewed below.

As well as being used to describe or summarize samples, many of these measures are also vital in relation to making inferences about the population from which a sample was taken. The mode, median and mean of a sample will nearly always differ from those of the population being sampled. This is an example of the uncertainty that arises through sampling from a population. Even in the ideal situation that every observation in the population has an equal chance of being sampled (e.g., because observations are sampled at random) any sample that does not exhaust all members of the population will almost always differ from it in some way. This leads to the important distinction between a *statistic* and a *parameter* (see Key Concept 1.1).

KEY CONCEPT 1.1

Parameters, statistics and the law of large numbers

A *parameter* can be defined as a property of a population, in contrast to a statistic (which is a property of a sample). Taking a subset of the population makes it unreasonable to conclude that a characteristic of the sample, such as its mean, is the same as that of the population. Instead, the statistic provides a way to estimate a population parameter. It is customary to distinguish statistics from the parameters they estimate by using different (but usually related) symbols. One convention is to use a Greek letter for the population parameter and a Latin letter for the sample statistic. Another convention is to use the same symbol, but differentiate a sample estimate by the 'hat' symbol (^). Thus the population mean is often designated μ (the Greek letter 'mu', pronounced 'myoo') and the sample statistic could be represented by M or $\hat{\mu}$ ('mu-hat'). This is only a convention (and both Latin and Greek letters can have other roles). The mean can also be denoted by a placing a horizontal bar over another symbol. Thus $\bar{x}$ ('x-bar') represents the mean of x.

It is easy to show that statistics are likely to resemble the population parameters they estimate by invoking the notion of the sampling fraction n/N introduced earlier. As sample n approaches N (i.e., the sampling fraction increases) sample statistics tend to resemble population parameters ever more closely. When the sampling fraction is 1 (i.e., $n = N$) a statistic such as the mean is necessarily equal to the parameter being estimated.

It is possible to go further by appealing to the *law of large numbers*. According to this law, a sample average converges on its *expected value* as the sample size n increases.* One way to understand this is to consider sampling without replacement from a finite population of size N. Sampling without replacement means that no data point can be sampled more than once. When sampling with replacement, it is possible to resample the same value at a later stage. As a sample of size n increases, the sample mean, $\hat{\mu}$, computed from n data points is likely to be closer to the population mean μ (computed from N data points). Even though the effective N is infinite $\hat{\mu}$ will be indistinguishable from μ for all practical purposes when n is sufficiently large. A further implication of this law is that parameters that are also averages can be interpreted as the expected value of a statistic in the long run (i.e., repeatedly taking a large number of observations).

* If the statistic is unbiased its expected value is the population parameter.

A parameter is a property of a population whereas a statistic is a property of a sample. The connection between them is that descriptive statistics, for instance a mean, provide estimates of parameters, such as the population mean. The quality of the estimate depends on a range

of factors, including (among others) the nature of the process that generated the sample, the amount of uncertainty in the population and – as already mentioned – the size of the sample.

1.4.1 Mode

The mode is the most common value in a sample. The sample labeled D_1 (from Example 1.1) comprised the following numbers:

12 14 9 11 15 11 D_1

The mode of D_1 is, therefore, 11 (and $n = 6$). One of important features of the mode is that a set of numbers may have a single mode (being *unimodal*) or more than one mode (being *multimodal*). For example, a set of numbers with two modes would be *bimodal*. The mode is often chosen to summarize frequency or count data (particularly for small numbers of unordered categories).

For continuous data the mode can be useful, but it is also common to find samples where the mode is not very informative. For the following nine numbers, the mode is 14:

14 14 18 35 43 51 62 88 91

Although this enables you to predict the most common value in the set, it is atypical of the set as a whole. The mode ignores information about the quantitative relationship between the numbers in the set. This makes it more suitable for categorical data – a situation where the numbers may have no inherent relationship with each other.

The mode – in common with other measures of central tendency – can be used to predict future outcomes. The mode would be the best value to guess if you wanted to predict the *exact* value of a number taken at random from the sample. This tends to work better for discrete outcomes than continuous ones. For instance, if you wanted to know what ice cream flavor people prefer, a random sample of 100 people might reveal that chocolate was the modal choice. Chocolate would be, therefore, the best guess for the favorite flavor in the sample (the guess with the best chance of being correct) and a good estimate of the best guess for the population. For continuous data the situation is slightly different. The mode is still the best guess for the sample, but might be wildly wrong. For instance, if you asked a random sample of 100 people how much they weighed (to the nearest kilogram) the mode might be 105 kg. This would be the best guess for the exact weight of someone in the sample, but might be very far from typical. In addition, it is unlikely to be a good estimate for the population.

Example 1.3 Consider the responses made by a group of ten people to the question: *What colour eyes do you have?* If two people respond 'blue', five respond 'brown' and three respond 'green' the modal eye colour would be brown. Figure 1.1 shows these responses in the form of a bar plot of the frequencies (a standard way of plotting count data).[3] The modal eye colour (brown) is indicated by the tallest bar in Figure 1.1. One advantage of plotting the data in this way is being able to spot the mode or modes immediately (which can be hard to detect in even a short list of numbers). The plot also shows that the sample is unimodal (as the plot has a single peak). Here the mode could be used to predict the most likely eye colour of a random member of the group. If the group were

a random sample from a set of people (e.g., university students), and in the absence of any other information, brown would also be the best guess for eye colour for members of that set (being correct with probability $5/10 = 0.5$). Treating the mode as a best guess at the exact value in the population being sampled can work quite well for discrete data.

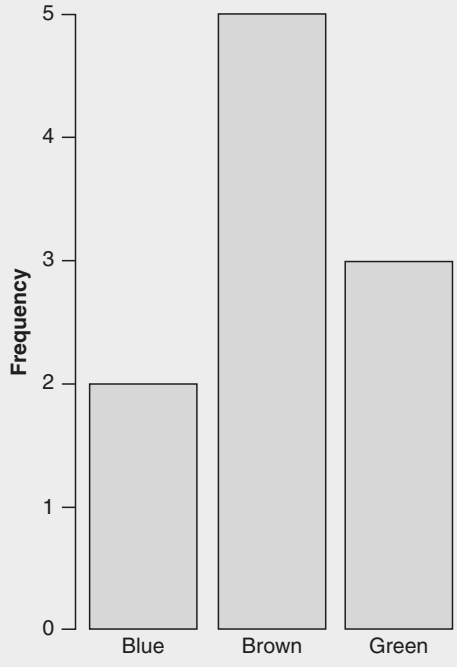

Figure 1.1 Frequency of eye colour for a sample of ten people

1.4.2 Median

The median is one of the most intuitively appealing measures of central tendency. It is the central or middle value in a set of n numbers. If n is odd the median requires little effort to determine. If the numbers are placed in rank order (e.g., lowest to highest) the median is the middle value. So, for the numbers,

$$23 \quad 42 \quad 65 \quad 108 \quad 111$$

the median is 65. If n is even, then the median is – in a strict sense – undefined (except in the unlikely case that the two central numbers take the same value). When n is even the convention is to take the mid-point between the two central values as the median. If the numbers, 64, 11, 7, 10, 4, 22, are placed in rank order, the resulting set is:

$$4 \quad 7 \quad 10 \quad 11 \quad 22 \quad 64$$

The two middle values are ten and 11. The median can therefore be computed as $(10+11)/2=10.5$.

The median has some potentially very valuable properties. Any set of numbers can be described by a single *median* value and the median divides the set roughly in half (exactly so, if n is even and there are no ties). Because the median only uses the central one or two values in its calculation it is not sensitive to extreme scores. For example, the value 64 in the previous example could be replaced with any value greater than or equal to 11 and the median wouldn't change.

The median is generally preferable to the mode if, as with ordinal and continuous data, the relationships within a set of numbers are meaningful. Its insensitivity to non-central values also makes it a good choice if there is reason to doubt the accuracy or authenticity of some of the numbers (e.g., if you think that extreme values may be data entry errors). On the other hand, if someone is interested in all the non-central values, the median is not a good choice. The median ignores potentially vital information about a set of numbers. The median is usually preferred as a description of a typical member of a set of numbers, but is not adequate as a summary of all the numbers in the set.

The insensitivity of the median to extreme values is a particularly attractive feature when sampling from populations with a lot of variability. The median generally provides good estimates of a typical population value in these situations – tending to produce values that are close in absolute distance to the population median (see Box 1.5). One exception, where the median can be rather misleading, is for multimodal distributions (distributions with several modes), where the modes could be far apart and the population median might not be close to any of the modal values.

1.4.3 Arithmetic mean

The arithmetic mean is probably the most widely used measure of central tendency: so widely employed that it is often referred to (without qualification) as the *average* or *mean*. The adjective 'arithmetic' distinguishes it from other forms of mean (some of which will be considered later). An arithmetic mean is calculated by adding up a set of numbers (i.e., taking their sum) and dividing by n. The set of numbers D_1

$$12 \quad 14 \quad 9 \quad 11 \quad 15 \quad 11 \qquad D_1$$

has a sum of 72 and because $n=6$, the arithmetic mean is $72/6=12$. It is common to report the arithmetic mean as M (e.g., $M=12$) when reporting results, but to refer to it as $\hat{\mu}$ or $\bar{x}$ in formulas.[4] The symbol $\hat{\mu}$ emphasizes its role as an estimate of the population mean μ.

One way to present the calculation procedure for the arithmetic mean (and other statistics) is in the form of the equation:

$$\hat{\mu}=\frac{\sum_{i=1}^{n} x_i}{n} \qquad \text{Equation 1.1}$$

This type of formula is ubiquitous in statistics and can be intimidating at first. A brief explanation of how they work is given in Box 1.2.

Box 1.2 Equations involving Σ

Perhaps the simplest way to understand a formula such as that reproduced below is to view it as a set of instructions:

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$$

The Σ symbol (the Greek capital letter sigma) stands for 'sum' and can be interpreted as an instruction to add up the quantities immediately to its right. Underneath sigma is the element of the instruction that indicates where to start adding the numbers (when $i = 1$) and above the sigma it indicates where to stop (when $i = n$, the sample size). The x in the formula refers to the set of numbers in the sample, whereas i is an 'index' value for each data point. So, the first number in the sample D_1, 12, is x_1 and the last number, 11, is x_6. The final element of the instruction involves dividing the resulting sum by n. Putting all this together, the formula is an instruction to add up the set of numbers x_1 to x_n and then to divide the resulting total by n.

Although this type of equation can be off-putting, they are necessary for communicating the exact procedure used in a calculation. In this case the calculation could easily be communicated in words, but for a more complex formula (such as Equation 1.10 below) precise notation of some kind is indispensable.

An important property of the arithmetic mean is that the sum of the distances of each point from the mean (the deviations from the mean) is zero. Thus, a different way of thinking about the mean is that it is the value that balances out these deviations (at least when considering simple arithmetic operations such as addition and subtraction). An important observation is that, unlike the mode or median, the arithmetic mean uses all n numbers in its calculation. Changing any number would therefore always have some impact on the mean (though this impact diminishes as n increases). This has, historically at least, been considered an attractive property for a descriptive statistic because it implies that all the information in the original set of numbers has contributed to the final result.

Example 1.4 The 2009 UK *Annual Survey of Hours and Earnings* reports the median and mean income for a sample of 18,835 employees as £21,320 and £26,470 respectively. For continuous data such as earnings, the mode isn't particularly helpful (perhaps just reflecting the legal minimum wage for a full-time employee). The mean is a bit higher than the median because the data are not evenly distributed either side of the median – the high earners are more spread out than the low earners. This phenomenon is known as *skew* (specifically positive skew – see Key Concept 2.2). So a small number of very high earners pull up the mean relative to the median; someone in the sample might earn £100,000 more than the median, but no one can earn less than zero (£21,320 below median).

Which is the better measure of central tendency here? It depends on what you are trying to measure. The median tells you what a typical member of the sample earns. More people earn close to £21,320 than earn close to £26,470. On the other hand if you want to know how much money the sample as a whole have to spend, the mean is probably a better figure – it better reflects the total earnings in that period.

Faced with data such as these it is tempting to think the median is the most informative measure of central tendency. However, the median can sometimes be very misleading. In 1985 the biologist Stephen Jay Gould wrote the article 'The median isn't the message', in which he described being diagnosed with a rare form of cancer three years earlier. He quickly learned that median mortality for this form of cancer was eight months. Gould (1985) describes his initial 'stunned' reaction, before realizing that while the median might be a reasonable description of a typical patient, he was probably atypical (e.g., being younger and with an early diagnosis). If half of all patients live between zero and eight months after diagnosis, the other half includes patients who survived from eight months upwards. Gould doesn't report the mean survival, but this would have added further information (because it would incorporate those patients still alive many years after diagnosis).

Measures of central tendency can provide reasonable predictions of observations from the same population, but there may well be other information that can be taken into account. The quality of prediction will also depend on the dispersion of observations around the mean. Both incomes and survival times are very variable and so neither the median nor the mean would lead to a particularly accurate prediction on its own. Gould himself lived for another 20 years after diagnosis.

1.4.4 Geometric mean

An important alternative to the arithmetic mean for certain situations is the *geometric* mean. Where the arithmetic mean involves taking the sum of n numbers and dividing by n, the geometric mean involves first calculating the product of n numbers and then taking their n^{th} root.[5] Writing the procedure in equation form gives:

$$\hat{\mu}_{geometric} = \sqrt[n]{x_1 \times x_2 \times \ldots \times x_n} \qquad \text{Equation 1.2}$$

To see how it works, plug the numbers five and 20 into the equation. Their product is 100 and the square root of this product (because $n = 2$) gives a geometric mean of ten:

$$\hat{\mu}_{geometric} = \sqrt[n]{x_1 \times x_2 \times \ldots \times x_n} = \sqrt[2]{5 \times 20} = \sqrt[2]{100} = 10$$

Because Equation 1.2 involves multiplication rather than addition, using Σ would be inappropriate. The equivalent symbol for multiplication is Π (the capital Greek letter *pi*). The geometric mean can therefore be expressed more compactly as:

$$\hat{\mu}_{geometric} = \left(\prod_{i=1}^{n} x_i \right)^{1/n} \qquad \text{Equation 1.3}$$

In what sense is this mean similar to an arithmetic mean? The connection between the arithmetic mean and the geometric mean becomes obvious if you switch to working with logarithms (see Box 1.3). In the examples that follow we will assume that the natural logarithm (*ln*) is used.

Box 1.3 Arithmetic with logarithms

Logarithms are very convenient mathematical functions that provide a link between addition and multiplication (and hence also between subtraction and division). Each logarithm has a *base* that is needed to scale the link between numbers and the logarithms (but is not that important otherwise). In statistics, two popular choices of base are 10 or e (where e is a mathematical constant approximately equal to 2.718282). The logarithm of a number x is defined such that if $base^y = x$, then $\log_{base}(x) = y$. For instance, if you are working with *base* 10, then the logarithm 2 is $10^2 = 100$. Conversely, the logarithm of 100 in base 10 is 2 (i.e., $\log_{10}(100) = 2$).

A major reason for working with logarithms is to simplify mathematical operations using multiplication. This works because addition of logarithms is equivalent to multiplication of the original numbers. Consider the following:

$$\log_{10}(100) = 2$$
$$\log_{10}(1000) = 3$$
$$\log_{10}(100) + \log_{10}(1000) = 5$$

Given that $\log_{10}(100) = 2$ and $\log_{10}(1,000) = 3$, the answer 5 represents $100 \times 1000 = 100,000$ (or 10^5) on the original scale. Adding logarithms of the original numbers gives $2 + 3 = 5$. Although the answer 5 was arrived at by addition using the logarithms, multiplication of the numbers on the original scale gives the same answer. The base 10 logarithm of 5 represents 100,000 (a 1 followed by five zeroes). This is also the answer obtained by multiplying 100 by 1000. This property is true of all addition involving logarithms, hence

$$\log_{base}(a) + \log_{base}(b) = \log_{base}(a + b)$$

is equivalent to

$$base^a \times base^b = base^{(a+b)}$$

Going back to the previous example: $10^2 \times 10^3 = 10^5$.

Subtraction of logarithms is equivalent to division. Thus $\log_{10}(1,000) - \log_{10}(100) = \log_{10}(100/10) = \log_{10}(10) = 1$ on a logarithmic scale (or 10 on the original scale). Less obvious is that logarithms reduce *exponentiation* (raising a base to the power of another number) to multiplication. For example

$$\log_{10}(100,000) = \log_{10}(10^5) = 5 \times \log_{10}(10) = 5$$

The link between exponentiation and multiplication also provides the inverse of the logarithmic function; if $\log_{base}(x) = y$ then $base^y = x$. It follows that the function 10^x is the inverse for $\log_{10}(x)$.

Logarithms to base e are known as *natural logarithms* and usually denoted by the function $\ln(x)$ rather than the clumsier $\log_e(x)$. The inverse is usually denoted by e^x. Most statistical procedures use natural logarithms, but because the choice of base is purely an issue of scaling this is largely a matter of preference (provided the same base is used throughout a set of calculations). This scale

shift involves multiplication by a constant (e.g., by $\ln(10) \approx 2.3026$ to convert from base 10 to natural logarithms). Natural logarithms are usually just as easy to work with as those for any other base when you have real data (because it is unusual to have real data that are neat multiples of 10). For instance:

$$\ln(2) + \ln(3) = \ln(6) = 1.791759 \text{ and}$$
$$e^{1.791759} = 6$$

There are restrictions on the values a base can take (for routine uses of logarithms at least) and bases other than e, 10 or 2 are uncommon. In general, bases are positive numbers greater than one and logarithms can only be computed for real numbers greater than zero. Needing to take the logarithm of zero sometimes makes using logarithms awkward (though there are ways to cope with this problem).

The arithmetic mean of the logarithms of n numbers is given by the equation

$$\frac{\sum_{i=1}^{n} \ln(x_i)}{n} \qquad \text{Equation 1.4}$$

(In this case natural logarithms have been used, but remember that the choice of base is not critical – it just represents a shift of scale.) The statistic expressed by Equation 1.4 is on a logarithmic scale and not easy to interpret in relation to the scale of the original sample. This can be resolved by the transformation e^x (the inverse function for the natural logarithm). The geometric mean is therefore:

$$\hat{\mu}_{geometric} = e^{\left(\frac{1}{n}\sum_{i=1}^{n} \ln(x_i)\right)} \qquad \text{Equation 1.5}$$

Example 1.5 To see how this works in practice, we'll apply the calculation to D_1:

12 14 9 11 15 11 D_1

For these values $M = 12$, and the natural logarithms are:

2.48491 2.63906 2.19722 2.39790 2.70805 2.39790

The arithmetic mean of these values is 2.47084 and $M_{geometric} = e^{2.47084} = 11.83$. Quite a few statistical procedures work with numbers on a logarithmic scale rather than the original scale. In most cases the geometric mean will be much easier to interpret than the arithmetic mean of the numbers on the logarithmic scale. For example, if the original data were earnings per hour in dollars, the value 11.83 is easy to interpret (as \$11.83). The value 2.47084 is not (although it represents the same quantity on a logarithmic scale where e is the base).

1.4.5 Harmonic mean

Where the geometric mean is a generalization of the arithmetic mean using logarithms, the harmonic mean is a generalization using reciprocals (see Box 1.4).

Box 1.4 Reciprocals

A reciprocal is a mathematical function that involves dividing a number into one. The reciprocal of x is therefore $1 \div x$ (usually written as $1/x$). One consequence of this is that when x is a fraction (e.g., 1/2) or a ratio (e.g., 0.25 meters/second) calculating the reciprocal of x just involves 'inverting' the fraction or ratio. For example, 1/2 becomes $2/1 = 2$ and 0.25 meters/second becomes four seconds/meter.

Like logarithms, reciprocals are often used in mathematics to make arithmetic easier. For example, multiplication of the reciprocal of x is equivalent to division by x:

$$5 \times 1/x = 5/x$$

Taking the reciprocal of x is the same as raising x to the power of -1. Hence $1/x = x^{-1}$. The reciprocal function is also its own inverse. So taking the reciprocal of a reciprocal reverses the operation:

$$(1/x)^{-1} = x$$

One drawback of working with reciprocals is that taking the reciprocal of zero is not possible (for standard arithmetic, at least) and, as with logarithms, using the reciprocal function when a set of numbers contains zero can cause problems.

The arithmetic mean of the reciprocals of a set of numbers is

$$\frac{\sum_{i=1}^{n}\left(\frac{1}{x_i}\right)}{n} \qquad \text{Equation 1.6}$$

Taking the reciprocal gives the harmonic mean:

$$\hat{\mu}_{harmonic} = \frac{n}{\sum_{i=1}^{n}\left(\frac{1}{x_i}\right)} \qquad \text{Equation 1.7}$$

Note that Equation 1.7 simply 'flips' the right side of Equation 1.6 – itself a ratio – upside down. As with the geometric mean, symbols for the harmonic mean vary, with both H and $\tilde{x}$ (pronounced 'x-tilda' by analogy to $\bar{x}$ and x-bar) being fairly common.

Example 1.6 Again, let's apply this formula to the sample D_1:

12	14	9	11	15	11	D_1

The reciprocals of the numbers are:

0.083333	0.071429	0.111111	0.090909	0.066667	0.090909

The sum of the reciprocals is .514. The arithmetic mean of these numbers is 12 and $M_{harmonic}$ is 11.67 (because $n = 6$, the harmonic mean equals six divided by .514).

When the numerator is fixed at some total and the denominator of a ratio varies, the harmonic mean is often a sensible choice.[6] If the denominator is fixed, the arithmetic mean is probably more appropriate. Consider the rate at which errors occur on two tests. Test A has an error rate of ten per minute and Test B an error rate of five per minute. If both tests take one minute to complete then the appropriate average is 7.5 (the arithmetic mean). If the tests were of different durations and stopped when a participant made ten errors (i.e., the numerator is fixed), the appropriate average is 6.67 (the harmonic mean).

Why does the appropriate mean change? In the different duration scenario, B takes two minutes relative to one minute for A. The harmonic mean 'weights' the result for the additional length of time that B took. Doing this produces a number that correctly reflects the fact that a total of 20 errors were produced in three minutes ($20/3 = 6.67$).

The harmonic mean is not widely used as a descriptive measure (e.g., perhaps when working with reciprocals rather than data on their original scale). However, it arises from time to time when working with ratios and fractions, such as when averaging rates or ratios within other procedures.

1.4.6 Trimmed mean

A trimmed mean is a measure of central tendency designed to reduce the influence of extreme scores. Consider the following samples:

12	14	9	11	15	11	D_1
12	14	8	11	34	11	D_2

The respective arithmetic means are 12, for D_1, and 15, for D_2. A trimmed mean can be calculated for these samples by discarding the smallest and largest k numbers in each sample. This procedure can be described by the equation

$$\hat{\mu}_{trimmed} = \frac{\sum_{i=k+1}^{n-k} x_{(i)}}{n-2k} \qquad \text{Equation 1.8}$$

The new element to the notation here is $x_{(i)}$. This indicates that the data have been ordered from highest to lowest.

Trimmed means vary according to the extent of trimming. The usual convention is to indicate this by the percentage of data trimmed. It is usual to set k so that between 5% and 20% of each end or 'tail' of the sample is trimmed (where this percentage is $100 \times k/n$). For the above samples, if $k = 1$, the percentage trimmed from each end of the distribution is $1/6$ or roughly 16.7%.[7] The 16.7% trimmed mean, $M_{16.7\%}$, is therefore 12 for both D_1 and D_2. It so happens that the remaining observations of the two samples (11, 11, 12, 14) are identical (though identical trimmed means merely require the untrimmed observations to have the same arithmetic mean).

The trimmed mean forms a natural link between the arithmetic mean and the median. The former is a special case of the trimmed mean when $k = 0$, while the latter is a special case when $n - 2k = 1$ (if n is odd) or $n - 2k = 2$ (if n is even). This makes the trimmed mean a compromise between the mean (that weights all observations equally) and the median (that ignores all non-central values). Trimming can be applied to other statistics (e.g., the geometric or harmonic mean), though this is uncommon in practice. Calculating trimmed means for large data sets can be awkward by hand, but is implemented in most statistics software. As the percentage of trimming approaches 50% in each tail (i.e., 100% in total) the trimmed mean will converge on the sample median.

1.5 Dispersion within a sample

Measures of central tendency such as the median, trimmed mean or arithmetic mean reduce data to a single number. While this can be a very convenient way to summarize a set of numbers, it will fail to capture some essential characteristics of the data. Compare one of the earlier examples in this chapter with a new sample:

12	14	9	11	15	11	D_1
11	22	7	12	15	5	D_3

D_1 and D_3 have identical arithmetic means and medians (12 and 11.5 respectively), but are very different. Both the sample mean and median are a better description of numbers in the sample D_1 than in D_3. The numbers in D_1 fall no more than three units from the mean, whereas D_3 includes one observation that is ten units away from the mean. The numbers in D_3 have greater *dispersion* (i.e., are more spread out) than those in D_1.

Just as a single number can be used to characterize the central tendency of a sample, various options exist to describe the dispersion in a sample. As with measures of central tendency, no single measure is entirely satisfactory for all situations. This section will consider several of the most important measures in relation to a sample of n numbers.

1.5.1 Range

A very simple and intuitive measure of sample dispersion is the *range*: the difference between the *minimum* and *maximum* values of the sample. The range is simple to compute and easy to understand, but it is extremely limited as a measure of dispersion.

From the range alone, it is hard to assess how far a typical data point might be from the mean or median (though halving the range gives an idea of how far the extremes are from the center, provided the sample is fairly symmetrical about the median). A further problem is that the range is determined only by the most extreme values in a sample. Ignoring the dispersion of less extreme numbers implies that the range is particularly vulnerable to aberrant or extreme values – values that researchers will probably not want to dominate the outcome of a statistical procedure.

Example 1.7 Finding the range is easy enough if the numbers are arranged in order. Ordering D_1 and D_2 from highest to lowest gives:

9	11	11	12	14	15	D_1
5	7	11	12	15	22	D_3

The minimum of D_1 is five and its maximum is 15 giving a range of $15 - 9 = 6$.

For D_3 the range is $22 - 5 = 17$. In this case the range does a rather fine job of describing the differences in spread of the two samples. To see why the range is sometimes problematic, compare these results with those for D_4.

5	10	11	12	12	22	D_4

This has the same range as D_3 but most of the numbers are very close to the mean and median. The range is completely insensitive to this clustering of data in the center. This insensitivity to central values is a particular problem in large samples where there are proportionately more central values.

1.5.2 Quartiles, quantiles and the interquartile range (*IQR*)

An alternative to computing the range on the full sample is to compute the range on a trimmed sample. By discarding a proportion of extreme values it is possible to obtain a measure of dispersion that better describes the spread of less extreme, more central values. In principle, this could be carried out for any level of trimming, but it is rare to see anything other than the *interquartile range* (*IQR*) chosen. The *IQR* is defined as the difference between the upper and lower quartile of a set of numbers.

Quartiles are the points on the number line that separate a set of n ordered numbers into subsets of $n/4$ (or as close to $n/4$ as possible). The first (lower) quartile separates the smallest 25% of the numbers from larger numbers. The second (middle) quartile is the median, while the third (upper) quartile separates 25% largest numbers from the smallest. If you are wondering why there are only three quartiles, think about how many cuts you need to make along a length of pipe to divide it into four equal pieces. It should take three cuts. One cut in the middle creates two halves, and then two further cuts are required to divide each of those pieces in half. These cuts are equivalent to the three quartiles. The quartiles are the boundaries used to divide up a set of numbers, they are not the subsets created by the boundaries.

Quartiles are a special case of *quantiles*. Quantiles are the points on the number line used to divide up a set of numbers into q subsets of equal (or as near to equal as possible) size. So for quartiles $q = 4$ and for 'quintiles' $q = 5$. It takes $q - 1$ quantiles to split a set of numbers up in this way (e.g., for quintiles there are $5 - 1 = 4$ boundaries). This makes the size of each subset n/q (or as close to this as you can get). Quartiles are very popular as descriptive statistics. Another common choice is the centile (also called a percentile) where $q = 100$. Centiles therefore describe the percentage of values in the lower portion of a set of numbers (e.g., the 12^{th} centile defines the smallest 12% of the set). It is often convenient to express quartiles as centiles and you will often see the first, second and third quartile referred to as the 25^{th}, 50^{th} and 75^{th} centile. Thus the *IQR* can also be defined as the difference or distance between the 75^{th} and 25^{th} centile.

Example 1.8 As noted above, the samples D_3 and D_4 have identical means, medians and ranges:

5	7	11	12	15	22	D_3
5	10	11	12	12	22	D_4

The *IQR* for D_3 is 6.25 and for D_4 it is 1.75. The *IQR* therefore does a good job of capturing the clustering of the non-extreme values in each sample (relative to the median or mean). Halving the *IQR* also gives a very rough sense of how far the more central values are from the mean or median (provided the sample is fairly symmetrical).

Deriving an *IQR* for small samples such as these (and more generally when n is not neatly divisible by four) is awkward and different methods exist to resolve the problems for such situations (e.g., the figures reported here can be replicated in R or via the QUARTILE() function in Excel, but differ from those provided by SPSS). Software will also sometimes depart from standard conventions when data are very sparse (e.g., for a sample such as D_3). These departures are usually designed to produce plots that are easier to interpret – but it is worth checking exactly what is plotted when n is small. A more detailed discussion of calculation methods can be found in Hyndman and Fan (1996).

While the *IQR* is widely used, it is not common as a stand-alone measure of dispersion. It is usually encountered alongside other descriptive statistics – in particular in graphical summaries of data such as a *box plot*. A box plot is a handy summary of a number of descriptive statistics and can be useful as a quick exploratory tool. Figure 1.2 shows the anatomy of a typical box plot for a sample of simulated data. The dark central line shows the sample median. The *hinges* (the top and bottom of the box) show the upper and lower quartiles respectively. The *whiskers* (the dashed lines extending vertically from the box) show the minimum and maximum values of the sample.

The software that produced the plot extends the whiskers as a multiple of the *IQR* (typically 1.5) from the hinges, provided they do not extend beyond the most extreme values in the data (as would happen in this case). Extreme values that fall beyond the whiskers are often also displayed. One such value occurs in Figure 1.2 and it is labeled here as a 'potential' outlier. Such values can cause problem for a statistical analysis, but are not necessarily unusual or particularly extreme. Working out whether a potential outlier is unusual or extreme is a difficult problem (and dealt with in more detail in Chapter 9). The box plot also indicates the range (this is the distance along the y-axis between the two most extreme features – whether whiskers or individual data points). While box plots can be constructed in many different ways, nearly all will display the

median, *IQR* and range. Regardless of how they are defined, the whiskers tend to give an indication of where the 'bulk' of the data fall, while the *IQR* gives an indication of the clustering around the median.

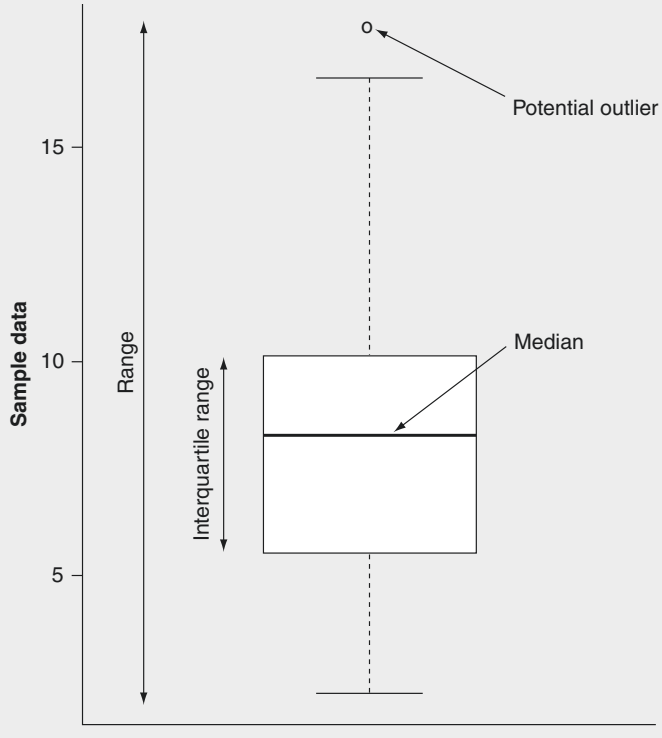

Figure 1.2 The main features of a box plot

1.5.3 Sums of squares

Measures of dispersion such as the range and *IQR* are often reported alongside measures of central tendency that do not use all sample data in their calculation (e.g., measures such as the median, trimmed mean or mode). If you are working with the mean it is natural to report a measure of dispersion that also uses the whole sample.

It might be tempting to start by using the average deviation of each point from the arithmetic mean but, as noted earlier, the sum of these deviations is always zero. The arithmetic mean is the point that balances these deviations (and so they will cancel out when summed). A plausible alternative is to use *absolute deviations* from the mean (i.e., discarding the sign of the deviations). Using absolute deviations is unfortunately not as simple as might first appear (see Box 1.5). Instead, the most widely employed measures of dispersion are based on *squared deviations* from the mean. If the mean is ten and an observation is six, its corresponding deviation would be $6 - 10 = -4$. Its squared deviation would therefore be $(-4)^2 = 16$.

Box 1.5 Advantages and disadvantages of using absolute deviations

Calculating the absolute deviations from the arithmetic mean merely involves subtracting the mean from every number in the sample and discarding the sign of the difference. For D_2 the raw deviations (*residuals*) are:

$$-3 \quad -1 \quad -7 \quad -4 \quad 19 \quad -4$$

The corresponding absolute deviations are therefore:

$$3 \quad 1 \quad 7 \quad 4 \quad 19 \quad 4$$

The arithmetic mean of these absolute deviations – the *mean absolute deviation* (*MAD*) is $38/6 = 6.33$. Using the absolute distance from the mean isn't necessarily the best way to go.

One consideration is that the measure of central tendency that minimizes the absolute deviations is the median (not the mean). For example, the corresponding raw and absolute deviations from the median are

$$0.5 \quad 2.5 \quad -3.5 \quad -0.5 \quad 22.5 \quad -0.5$$

and

$$0.5 \quad 2.5 \quad 3.5 \quad 0.5 \quad 22.5 \quad 0.5$$

Here the total absolute distance from the median is 30 (not 38) and the *MAD* for the median is 5.

This suggests a link between using the arithmetic mean in conjunction with squared deviations (and statistics such as *SS*, variance or *SD*) and using the median with absolute deviations. One reason for preferring squared deviations is therefore the prevailing preference for the mean over the median in statistics. This default use of the mean is not always reasonable. The median is generally a more robust measure (e.g., it is less sensitive to extreme values).

A second consideration is that squared deviations tend to be easier to work with than absolute deviations. This is true both in the sense of deriving the mathematical proofs upon which statistical procedures are based, and in terms of the complexity of the calculations required to implement statistical analyses. Most statistical work still relies on squared deviations from the mean, but owing to advances in computing power and the increasing availability of suitable software other approaches are becoming more popular.

The basic building block for any measure of dispersion using squared deviations from the mean is a *sum of squares* (*SS*). Sums of squares are calculated by squaring each of the deviations from the mean and adding these squared values together. The can be represented as the equation

$$SS = \sum_{i=1}^{n} (x_i - \hat{\mu})^2 \qquad \text{Equation 1.9}$$

where $\hat{\mu}$ is the arithmetic mean of the sample.

Summing squared deviations from the mean is a very convenient way to capture variability within a sample (and many other statistical procedures have this basic calculation at their core). Even so, there are several difficulties inherent in using sums of squares as a descriptive measure. First, summing deviations from the mean ensures that if *n* increases, *SS* increases too (assuming some variability in the sample). This makes comparison between samples with different *n* awkward. Second, sums of squares are scaled in terms of the squared deviations from the mean. Interpreting these squared values in terms of the original (unsquared) units can be tricky.

Note also that any measure using the squared deviations has disadvantages (as well as advantages) relative to using alternatives such as absolute deviations (see Box 1.5). Measuring dispersion in terms of the squared deviations from the mean produces statistics that may be oversensitive to extreme values (although this is not inevitable). This can cause serious problems (e.g., a single very extreme value can sometimes distort the outcome of a study).

Example 1.9 This example will again use the sample D_1:

12 14 9 11 15 11 D_1

For D_1 the deviations from the mean (often termed *residuals* in this context) are obtained by subtracting the mean of 12 from each observation to get:

0 2 −3 −1 3 −1

The corresponding squared deviations are therefore:

0 4 9 1 9 1

The sum of squares is thus $0+4+9+1+9+1=24$.

Now compare this to D_3, a sample with the same mean, but more widely dispersed values. For D_3 the $SS=184$. While for D_4, a sample with the same mean, median and range, $SS=154$. This is consistent with the observation that central values in D_4 tend to be closer to the mean than for D_3. Because sums of squares use the whole data set they are sensitive to the dispersion of both extreme and central values.

1.5.4 Variance

The variance is closely related to sums of squares, but incorporates an adjustment to the *SS* to account for different sample sizes. This is achieved by dividing the sums of the squared deviations by *n*. The equation for the variance can therefore be denoted as:

$$Var = \frac{\sum_{i=1}^{n} (x_i - \hat{\mu})^2}{n} \qquad \text{Equation 1.10}$$

This equation can be interpreted as the arithmetic mean of the sums of the squared deviations from the arithmetic mean. This is an important point: the variance of a sample is itself a form

of arithmetic mean. Thus both the variance and the arithmetic mean are averages and findings such as the law of large numbers apply to both (see Key Concept 1.1).

The variance is a fundamental concept in statistics that, like *SS*, has many further applications. Here the focus is on the variance as a descriptive statistic. Although it deals with the problem of differences in sample size it shares all the other limitations of *SS*. Its main limitation as a descriptive statistic is a consequence of using squared deviations from the mean. Knowing that the mean is ten and that the variance is 16 doesn't make it easy to tell how the data are dispersed around the mean because the number ten (in unsquared units) and the number 16 (in squared units) are on different scales.

Example 1.10 In the previous example, the *SS* of the samples D_1, D_3, and D_4 were reported as 24, 184 and 154 respectively. In each case $n = 6$, so the corresponding variances of the samples can be calculated as $24/6 = 4$, $184/6 = 30.67$, and $154/6 = 25.67$. Because the samples have identical *n*, the variance is only marginally more informative than the *SS*.

To understand the advantages of the variance requires a comparison of samples of different sizes. Let's combine D_3 and D_4 into a single sample with $n = 12$. The variance of the combined sample is *SS*/12 and because the two samples have the same mean there is a shortcut to calculate its overall *SS*. This shortcut is simply to add the *SS* of D_3 and D_4 together. (This won't work if the means differ, because the residuals of the combined and individual samples would no longer be calculated relative to a common value). The new *SS* is $184 + 154 = 338$, and the variance of the combined sample would be $338/12 = 28.17$.

Doubling the sample has little impact on the variance (which necessarily takes a value somewhere between the variances of the two original samples). In contrast, the *SS* of the combined sample is around twice the size of the original *SS* for D_3 and D_4 separately (338 versus 154 or 184). The combined sample and the two subsamples all have similar dispersion and this produces similar variances (25.67, 28.17 or 30.67). The variance is a better statistic than *SS* (or even the range or *IQR*) for comparing otherwise similar samples with different *n* (e.g., different classes in a school).

1.5.5 Standard deviation

The *standard deviation* (*SD*) is a measure of dispersion related to the variance, but scaled to use the same units as the original data. It is the square root of the arithmetic mean of the sum of the squared residuals (where the residuals are deviations of observations from the arithmetic mean). Because the arithmetic mean of the sum of the squared residuals is the variance, the standard deviation is the square root of the variance. This is illustrated by the equation:

$$SD = \sqrt{Var} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \hat{\mu})^2}{n}} \qquad \text{Equation 1.11}$$

As a description of sample dispersion the *SD* has all the advantages of the variance, being a far more sensitive measure of dispersion than measures such as the range or the *IQR*, because it uses all the data (not just extreme or intermediate values) in its calculation. As with the *SS* or variance, it can sometimes be unduly influenced by extreme values (in which case a trimmed

measure such as the *IQR* might be preferred). Taken together, the advantages of using the *SD* often outweigh the disadvantages. For this reason, the *SD* is a popular statistic for summarizing the dispersion of a sample (and it is the standard choice for reporting dispersion alongside the arithmetic mean). Its most important property is that, unlike the variance or *SS*, the *SD* can be directly interpreted as a measure of the degree of clustering of a sample around the mean in a sample. It is therefore an excellent statistic for comparing the dispersion of samples of different sizes on a common scale.

A convenient guideline is to treat the *SD* as an estimate of the distance a typical data point lies from the mean. In a sample with a mean of ten and a variance of 16 the *SD* is four. It is usually reasonable to interpret this as indicating that a typical observation falls roughly four units from the mean.[8]

Example 1.11 In the preceding example the variances of D_1 and D_4 were calculated to be four and 25.67. The *SD*s are $\sqrt{4} = 2$ and $\sqrt{25.67} = 5.07$ respectively. Because the samples are fairly symmetrical and evenly spread around the mean, the *SD* gives a good indication of the distance a typical sample member is from the mean. A typical observation is around two units from the mean of D_1 and around five units from the mean of D_4. If samples are less evenly distributed around the mean the *SD* will still give a rough idea of the average spread of points around the mean (though it might not be the case that any single observation falls around this point). Getting a feel for the distribution of a small sample is easy. In larger samples it is a good idea to plot the data to reveal the overall shape of the distribution. A box plot is one way to do this, but there are many other methods (only one of which will be considered at this point).

A drawback of the box plot (and alternatives such as histograms) is that individual observations are not shown. A simple alternative that does show individual data points is the *stem and leaf plot*. In a stem and leaf plot the numbers are ordered low to high and the first two significant digits (i.e., both digits for numbers in the range −99 to 99) are plotted. Of these digits the first digit is the 'stem' and placed on the left (followed by a vertical line). The second digit is the 'leaf' and placed to the right of the vertical line. Numbers sharing the same stem have their second digit added to the right. Numbers with a larger stem value are added below. So a basic stem and leaf plot of D_3 looks like this:

0 | 5 7
1 | 1 2 5
2 | 2

Here the stem digits are 'tens' and the plot indicates a fairly even and symmetrical spread with no obvious gaps. The *SD* of 5.5 is therefore broadly consistent with typical distance of points from the mean. Contrast this with a stem and leaf plot for D_4:

0 | 5
1 | 0 1 2 2
2 | 2

The *SD* here is 5.1, but most points are much closer to the mean than this (with two points a little further out).

Stem and leaf plots are easy to generate (not requiring sophisticated software) but are not common in published work. Their inclusion here is to show the value of even a simple plot that includes all observations (albeit in an abbreviated form). Such plots can be helpful in deciding what measure of dispersion to report.

1.5.6 Other measures of dispersion

The measures of dispersion described here are among the most popular. Alternative measures of dispersion using absolute deviations from the median or arithmetic mean such as the *MAD* can be constructed (see Box 1.5). It is also possible to calculate measures of dispersion appropriate for the geometric mean, harmonic mean and trimmed mean. For example, the *SD* of the natural logarithms of a sample is very simple to calculate. Taking the exponent of this value (to base e) would rescale the *SD* in terms of the original units and give the *geometric SD*. Likewise a trimmed or harmonic *SD* or variance could be calculated for a sample.

1.6 Description, inference and bias

There are a number of properties that are desirable in estimates of a parameter such as the mean or median. A statistic should, for example, be an *unbiased* and *efficient* estimator of the relevant population parameter.

An efficient estimate has less error (i.e., tends to be close to the population parameter). The degree of error can be assessed in different ways, but the most common criterion is to use the sum of the squared residuals. An efficient estimator assessed using this criterion is therefore one that tends to have a small *SD*. If the estimator is also unbiased, it has zero bias, where bias is defined as the difference between the expected value of the statistic and the true value of the parameter. In the long run (e.g., given sufficient sample size) an unbiased statistic will converge exactly on the parameter it estimates.

The accuracy of an estimate is a combination of its error and bias. Imagine a large number of darts thrown at a target. If the throws are unbiased they will be scattered more-or-less evenly around the center of the target. If they fall consistently slightly right or left (or above or below) the center of the target, this indicates bias. The overall accuracy depends on the sum of the error and bias. An efficient estimator could be very inaccurate if the bias is large (just as a set of throws could be tightly clustered but all land a long way from the target). In the same way, being unbiased doesn't imply accuracy (your throws could fall evenly around the center of the target and yet still fall a long way away from it on average). On the other hand, a known bias can sometimes be corrected, whereas error tends to be quite hard to eliminate. For this reason it is sometimes better to adopt a biased estimator with small error rather than an unbiased but inefficient estimator.

Descriptive statistics such as means, medians and trimmed means are unbiased estimators of central tendency. The expected value of the statistic is the true population parameter (e.g., $\hat{\mu}$ estimates the population mean μ with zero bias). This is not the case, however, for sample statistics used to estimate population dispersion. The expected value of a descriptive measure

of dispersion in a sample (e.g., the range or variance) is an underestimate of the true population value. Where the *law of large numbers* applies, this underestimate declines, as n approaches N. However, for infinitely large populations (i.e., when $N = \infty$) any finite sample will have some bias (though with sufficiently large n this bias will become too small to detect).

Why is the sample dispersion an underestimate of the population dispersion? To understand how this bias arises, consider the case of a finite sample taken without replacement from some population. For example, imagine we have obtained a random sample of the ages (in years) of n people from the population of Nottingham ($N \approx 300,000$). When n is small the sample is unlikely to capture the extremes of the population and will underestimate its dispersion. The lowest age in the population will be zero years and we'll assume that the population maximum is 100 years. A random sample of size, say, $n = 10$ will hardly ever include both a newborn baby and a 100-year-old adult. For this reason the sample *range* will tend to underestimate the population range. As n increases, the probability of sampling the extremes increases (e.g., assuming there is only one 100-year-old in the population the probability of him or her being in the sample increases from $P \approx 1/30,000$ to $P \approx 1/3,000$ as n increases from ten to 100). As n approaches N the sample range is likely to get closer and closer to 100 years (the population range). This argument applies equally to infinite populations (and a mathematical proof is possible).

Other measures of dispersion in a sample will also underestimate the population dispersion (for exactly the same reasons). Measures such as the variance or *SD* require the full range of values in the sample or population for calculation; omitting the extreme values will necessarily reduce the final result. A moment's reflection should suffice to show that it also applies, albeit to a lesser extent, to measures such as the *IQR* or trimmed variance, even though they exclude the minimum and maximum. Excluding the most extreme values just shifts the problem to the next most extreme values (e.g., the first and third quartiles for the *IQR*).

1.6.1 Unbiased estimation of the population variance or *SD*

The aim of collecting data is often to use sample statistics to estimate population parameters. It is therefore undesirable if the descriptive formula for calculating the variance (or other measures of dispersion) provides an underestimate of the population variance.

Fortunately because the degree of bias is known, it can be eliminated. For this reason, a different formula is adopted for estimating the population variance from a sample than for description of the sample itself. Confusingly, both formulas are often labeled as the 'sample variance'. A more sensible designation is to label one formula the *descriptive* formula (which treats the sample as if it were a population) and one as the *inferential* formula (which regards the sample as an estimate of the population): Equation 1.9, Equation 1.10 and Equation 1.11 are all descriptive formulas.

The formulas differ by what is termed a correction factor (which can be derived from the mathematical proof that a sample variance underestimates the population variance) applied to the inferential formula. To go from the uncorrected descriptive formula to the corrected inferential formula requires multiplying the variance by this correction factor. Dividing by the correction factor (or multiplying by its reciprocal) allows you convert the (inferential) population variance estimate back to the (descriptive) sample variance. The correction factor is $n/(n–1)$. As you might expect, it is relatively large only for small n and becomes negligible when n is very large. In practice, the largest possible value of the correction factor is two (when $n = 2$) and approaches one for large samples (e.g., it would be ≈ 1.01 for $n = 100$ and ≈ 1.001 for $n = 1000$).

The population variance is usually represented by the symbol σ^2 (pronounced 'sigma squared' because σ is a lower-case form of the Greek letter sigma). So the descriptive sample *SD* is a biased estimator of σ^2. Knowing the correction factor is $n/(n–1)$, it is possible to construct an unbiased estimate of σ^2. The unbiased estimator of σ^2 can be labeled $\hat{\sigma}^2$ ('sigma-hat squared'). Incorporating the correction factor, the unbiased, inferential formula is:

$$\hat{\sigma}^2 = \frac{\left(\frac{n}{n-1}\right) \times \sum_{i=1}^{n} (X_i - \hat{\mu})^2}{n} = \frac{\sum_{i=1}^{n} (X_i - \hat{\mu})^2}{n-1} \qquad \text{Equation 1.12}$$

Apart from the correction factor, this is identical to Equation 1.10. Many inferential statistical procedures include a similar correction to form estimators of population parameters. However, some measures (e.g., the range or *IQR*) tend to be encountered almost exclusively as descriptive measures (although they remain biased estimates of dispersion in the population). Sums of squares are also used purely as descriptive measures. The *SS* tends to infinity as *n* increases (and is therefore not a sensible estimate of any population parameter). Using *SS* as a descriptive statistic is not necessarily inappropriate, but there is a tendency for people to interpret measures based on *SS* as if they were inferential statistics. So understanding the difference between *SS* and variance – particularly inferential measures of variance – is important.

Given the relationship between the variance and *SD* it is also possible to derive an inferential formula for the population standard deviation σ. The usual sample estimate of σ can be labeled $\hat{\sigma}$ ('sigma-hat'), and is the square root of the unbiased variance estimate:

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \hat{\mu})^2}{n-1}} \qquad \text{Equation 1.13}$$

A curious property of the estimator $\hat{\sigma}$ is that, although derived from an unbiased estimator, it is not itself an unbiased estimate of σ. Equation 1.13 underestimates the magnitude of σ (the underestimate arising, in this case, because the square root function is nonlinear). The underestimate is relatively small and can be safely ignored for most applications.

From this point on, the inferential formulas for the variance and standard deviation will be used by default. However, for the sum of squares (and some closely related statistics) descriptive formulas based on Equation 1.9 will be used.

1.7 R code for Chapter 1

1.7.1 Getting started

Once R is installed (a fairly easy procedure on most desktop computer systems) you will be presented with the *R console* window. Most interaction with R involves typing input into this window and hitting the return key. When R is ready to accept input you will see the following prompt:

```
>
```

Most of the work in R involves assigning things to entities called 'objects' and using functions (themselves types of objects) to manipulate other objects. The most basic object type in R is called a vector and it is a one-dimensional container of information. Vectors can contain all sorts of information, but for the moment we'll just consider vectors as containers of numbers. To assign a number to a vector is very simple. For instance, to assign the number 36 to a vector named `num.vect`[9] just enter (followed by the return key):

```
num.vect <- 36
```

To confirm that the assignment has indeed occurred, just type the name of the vector and hit return. This should bring up the following output:

```
[1] 36
```

The `[1]` in square brackets indicates that the object has a single dimension (and can be ignored for the moment). The <- arrow is an example of an assignment operator (there are others) and works – as you have probably realized – by taking the object on the right and assigning it to the object on the left. An alternative that would have worked just as well is:

```
36 -> num.vect
```

If you are new to R you will be tempted to assign objects left to right (as when writing) but in many cases you'll find it helpful to work right to left. One reason for this is that naming the object tends to be very easy, and it helps to get the easy part out of the way before typing out a complex instruction or formula.

In this way, R allows you to type in data and assign it to an object. It is also possible to read in data from external files (e.g., Excel or SPSS files). To keep things very basic, the examples in this chapter won't require reading in any external data files. On the other hand, the examples do require samples containing more than one data point. How do you get several numbers into an R object? One way is to use the combine function: `c()`. This is one of the most important and useful functions in R. To combine several numbers just use the combine function with the numbers separated by commas:

```
num.vect <- c(36, 49, 64)
```

Now, by entering `num.vect` (typing the object name and hitting return) the R console will generate:

```
[1] 36 49 64
```

The original contents of the vector have now been over-written (and are lost). You now know enough to enter data sets into R. Note that the vector is returned with `[1]` preceding it, indicating that the line starts with the first element of the vector. For large data sets this allows you to find the row with, for example, the 379th observation more easily.

Once you start entering data into R, your workspace (the bit of R that keeps track of objects such as vectors) may start to get untidy. When you quit R you will have the opportunity to save your default workspace (the one that R opens in). You could keep things tidy by not saving your

work, saving workspaces under a different name or by saving work in a text editor.[10] It helps to keep track of your workspace and tidy it up as you go along. To see what objects are in R, you can list them with:

```
ls()
```

The above call requires no arguments. You can remove (i.e., delete) objects with `rm()`, but be sure to use the name of the correct object. To remove `num.vect` enter:

```
rm(num.vect)
```

1.7.2 Arithmetic

All basic arithmetic operators are also built into R, so it is possible to perform a range of calculations either with numbers, objects or both. Standard operators such as + and − behave as expected. For multiplication, R uses the asterisk character `*`; for division, R uses the forward slash character `/`. Try out the following examples:

```
2 + 6
9 - 2
3 * 7
100/4
```

Using R as a calculator pays off when you start integrating calculations with assignment to objects. This makes it possible to store the output from a calculation. One application for this is to update objects:

```
num.vect <- c(36, 49, 64)
num.vect <- num.vect + 4
num.vect
```

```
[1] 40 53 68
```

As well as basic arithmetic operators, R has functions for square roots, logarithms and exponentiation (raising to different powers or orders). These follow the standard order of operations (e.g., associated with acronyms such PEMDAS in the US and BODMAS in the UK). These place operations in parentheses (brackets) first, followed by exponentiation and then division and multiplication. Addition and subtraction bring up the rear. To raise a number (or object) to a power, place the `^` operator after the number, followed by the required power. To square a number you raise it to the power of two (e.g., to square 15 you would enter `15^2` into R). This procedure is quite flexible and allows square, cube and other roots to be found. (To get the nth root of a number you need to raise it to the power of $1/n$.)

One more feature of R is worth introducing here. If a vector contains several numbers you can apply the same operation to every number in the vector simultaneously (if you wish). This is extremely valuable in statistics, because you will often want to do exactly this. We'll illustrate this by taking the square root of several numbers within a vector.

```
num.vect^0.5
[1] 6 7 8
```

Sometimes there are several ways to do the same thing in R (or more generally in mathematics). R has a built-in square root function called `sqrt()`. We could express the preceding operation as `num.vect^.5` or `num.vect^(1/2)` with the same outcome. The preceding example displayed the outcome of the calculation, but was not assigned to an object, so would be lost. To store the results we'd need to use assignment, as below.

```
sqrt.vect <- sqrt(num.vect)
```

1.7.3 Simple functions and measures of central tendency

R has functions for most common descriptive statistics, the exception being the mode. This is usually easy to determine in small samples. For large samples the sample mode is rarely used for continuous data (and different approaches are appropriate).

If the data are discrete (e.g., frequency or count data) then the modal response is the largest count (and obtained as the maximum of the counts). For such data a bar plot or histogram is usually a good idea. The bar plot in Figure 1.1 is produced by the R code below. This creates a vector of the counts for each eye colour and a separate vector of labels for the eye colors. The vector of labels uses text strings enclosed by single or double quotes (the choice is irrelevant – although consistency is important). Strings such as 'Blue' or 'two' can be stored and retrieved by R and are required to label output (e.g., to put titles or legends on figures).[11]

```
eyes <- c(2, 5, 3)
labels <- c('Blue', 'Brown', 'Green')
```

The plot itself uses the `barplot()` function. A basic plot just requires one argument to be defined when the function is 'called'. The arguments are the information supplied to the function. The most basic bar plot just requires a vector of data points to define the heights of the bars.

```
barplot(eyes)
```

This resulting plot has no labels on the *x* or *y*-axis and additional arguments have to be supplied to get a satisfactory plot. Commas must separate all arguments. Depending on the function, the arguments can be defined by order of entry or (if some arguments can be omitted) may need to be named. To tell R that the labels represent the names of the bars the `names.arg` argument is used. To specify the label on the *y*-axis the `ylab` argument is used.

```
barplot(eyes, names.arg = labels, ylab = 'Frequency')
```

This reproduces Figure 1.1 almost exactly. The main difference is the size and shape of the plot. This can also be manipulated via R code, but it is usually easiest just to resize the plot window manually. Manual resizing might sometimes distort the plot, but if so, re-running the command should clean it up. To return to a previous command just use the up arrow key in the R console

window. Repeated use of the up arrow cycles back through previously executed commands. In the preceding call, the text string defining the *y*-axis label is specified within the function (but it could be defined as a separate object such as a vector). Once you are happy with the plot you can save it in one or more common different formats (depending on the platform R is running on).

The flexibility to change the output of a function by specifying different arguments is a considerable benefit, but it can also be frustrating to keep track of the names of all the arguments a function can take. It is worth making your own reference sheet of common functions (or downloading one of the dozens available on the internet). To remember what arguments are available and how to specify them when calling a function, try out the `help()` function. You can get help for any function in the base installation packages (the packages of R functions that are installed by default) with the call `help(function.name)` or the shortcut `?function.name`. The help output may at first appear confusing, but it follows a fairly strict structure across all the main packages and will make more sense as you learn more about R. Try it out with the call `?c` to get help for the combine function. Some functions also have examples that can be accessed via the `example()` function:

```
example(c)
```

To compute the median or mean of a sample the functions `median()` and `mean()` are used. To illustrate this, let's do the calculations for the sample D_1. (Note that R is case sensitive and so the object D1 is different from d1.)

```
D1 <- c(12, 14, 9, 11, 15, 11)
median(D1)
mean(D1)
```

The `mean()` is very versatile and will also calculate a trimmed mean. As a second argument it expects the trimmed proportion in each tail. If you look at the help for this function using `?mean` you will see that the function has a trim argument with a default value of zero. This is designated by the `trim = 0` argument under 'Usage'. Many functions have default values, allowing the function to show quite sophisticated behavior. So a 16.7% trimmed mean for D1 is obtained with the command:

```
mean(D1, .167)
```

Increasing the trim proportion would eventually produce the median (and is guaranteed for `trim = .5`). In this case the trimmed mean doesn't differ from the mean, but for a sample such as D_2 it does matter:

```
D2 <- c(12, 14, 8, 11, 34, 11)
mean(D2, .167)
```

What about the geometric and harmonic mean? Functions to calculate them can be found in user-contributed packages for R, but it is also very easy to calculate them yourself. For the geometric mean, one method is to calculate the logarithms of the sample data (using whatever base you wish). We'll use natural logarithms. The natural logarithm is the default for the R `log()` function. This can be changed by using the `base` argument: the default being `base = exp(1)`. There is also a separate `log10()` function for logarithms to base ten. The inverse of the natural

logarithm is e^x and is provided by the `exp()` function. Thus `exp(1)` is one way to obtain the constant e.

To compute the geometric mean of D_1, first calculate the natural logarithms of the sample. The geometric mean is then obtained by using the inverse of the mean of the natural logarithms:

```
D1.ln <- log(D1)
D1.ln.mean <- mean(D1.ln)
exp(D1.ln.mean)
```

Doing it in three steps is not necessary, and R can roll all the steps into a single command:

```
exp(mean(log(D1)))
```

This instantiates the formula in Equation 1.5. To instantiate the formula in Equation 1.3 it helps to use the `prod()` command for taking the product of a set of numbers.

```
prod(D1)^(1/6)
```

The sample size can also be obtained from the `length()` function – which will count the number of things (in this case numbers) in a vector.

```
prod(D1)^(1/length(D1))
```

The harmonic mean can be obtained by instantiating Equation 1.7. Either of these commands will give the harmonic mean of D_1:

```
1/mean(1/D1)
mean(D1^-1)^-1
```

One final function is worth introducing here. This is the `summary()` function. Summary is a very general function that produces different outputs depending on the type of object in the call. For a vector of numbers it will return the minimum, maximum, mean and the three quartiles (including the median, which is the middle or second quartile).

```
summary(D1)
```

```
Min.  1st Qu.  Median  Mean  3rd Qu.  Max.
9.0   11.0     11.5    12.0  13.5     15.0
```

1.7.4 Measures of dispersion

Although the `range()` function in R returns the minimum and maximum of a sample (often even more informative than the range itself), getting the range is easy using the functions `min()` and `max()` directly:

```
D1.range <- max(D1) - min(D1)
D1.range
```

```
[1] 6
```

The `summary()` function also returns the minimum and maximum for a numeric vector along with the quartiles. Unlike the range, the *IQR* has a dedicated function called `IQR()`. To calculate *IQR* for samples D_3 and D_4:

```
D3 <- c(11, 22, 7, 12, 15, 5)
D4 <- c(10, 22, 11, 12, 12, 5)
IQR(D3)
IQR(D4)
```

R also has a general quantile function that defaults to reporting the minimum, maximum and quartiles (labeled as the 0%, 25%, 50%, 75% and 100% centiles). Note that different computer software may calculate the quartiles in slightly different ways (and may produce different *IQR* values when *n* is small). Hyndman and Fan (1996) describe nine different methods for calculating quantiles – all of which are implemented by `quantile()`. Details are given in the help documentation (e.g., via `?quantile`).

There is no sum of squares function, but R is designed to carry out similar calculations on objects. Calculating the sum of squares for D_1 is therefore not at all hard. The new function required here is the `sum()` function.

```
resids <- D1 - mean(D1)
sq.resids <- resids^2
sum(resids^2)
```

Again it can be combined into a single expression:

```
sum((D1 - mean(D1))^2)
```

To calculate sums of squares for any other numeric vector replace D_1 with the relevant vector name. You can make the procedure more generic by separating the object name from the expression using another vector:

```
vect <- D1
sum((vect-mean(vect))^2)
```

This means we just have to change the line `vect <- D1` to read `vect <- D2` to get the *SS* for D_2. (We could write a function to do this, but for the present it is useful to work through the details of each calculation.)

For the descriptive formulas in Chapter 1, calculating the variance of a sample involves dividing sums of squares by *n*. For a vector this works out as:

```
vect <- D1
sum((vect - mean(vect))^2)/length(vect)
```

Taking the square root of this gives the descriptive *SD* of the sample:

```
(sum((vect - mean(vect))^2)/length(vect))^.5
```

R also has built-in functions for the inferential variance and inferential *SD* of a vector. These are `var()` and `sd()`.

```
var(D1)
sd(D1)
```

1.7.5 Plotting dispersion

This chapter introduced two very basic ways to plot the dispersion of a sample in R (in addition to the bar plot used for frequency data in Figure 1.1). The first of these was the box plot. A box plot can be plotted in a number of different ways. The plot in Figure 1.2 uses the R defaults.

The box is defined by the quartiles (with the median, the second quartile, a line across the middle). The length of the whiskers is defined as either as 1.5 times the *IQR* or as the distance up to the values of the largest or smallest data points (whichever is least extreme). Any points outside the whiskers (potential outliers) would be marked as open circles. A basic box plot using the `boxplot()` function requires only a vector of data points:

```
boxplot(D3)
```

This illustrates the power of R for quick plotting and exploration of data – especially if combined with descriptive data functions such as `summary()`. To make it prettier you can add labels such as the *y*-axis label in Figure 1.2.

```
boxplot(D3, ylab = 'Sample Data')
```

A second plot type considered here is the stem and leaf plot. Again the basic command for this is very easy to run:

```
stem(D3)
stem(D4)
```

One complication is that R default scales the stem of the plot in units of five (rather than units of ten used in Example 1.11). For larger samples the R defaults will usually be very helpful, but to reproduce the plots in this chapter exactly it is possible to tweak the scale argument of the `stem()` function:

```
stem(D3, scale = .5)
stem(D4, scale = .5)
```

1.8 Notes on SPSS syntax for Chapter 1

The examples in this chapter were either worked out by hand or using R (rather than SPSS). The notes below give some pointers for getting broadly equivalent output using SPSS syntax.

1.8.1 Basic descriptive statistics and plots

To obtain common descriptive statistics such as *n*, min, max, mean, *SD* and range use the DESCRIPTIVES command with a /STATISTICS subcommand. It is good practice to separate out the command and subcommand onto different lines like this:

SPSS data file: D1toD4.sav

```
DESCRIPTIVES VARIABLES=D1
 /STATISTICS=MEAN STDDEV MIN MAX.
```

This example assumes you have opened the data file (D1toD4.sav) with each sample specified as a separate variable (and that this is the active data file). To use the syntax you can open a new syntax window and type or paste it in. Highlight the syntax you need from the syntax window, and then go to the <Run> menu and choose <Selection>. You can also access these commands directly from the <Analyze> menu. Choose <Descriptive Statistics> and <Descriptives . . . > to open a dialog box that allows you to generate the syntax you require and either run it by clicking on the 'OK' button or pasting it into the syntax window by clicking on 'Paste'. The latter is an excellent way to explore how SPSS syntax works (and gives your more flexibility than a pure menu-driven approach).

The output from the function includes a box with the requested descriptive statistics for sample D_1. To get the same for D_2, edit the syntax so that the variables statement reads VARIABLES=D1 D2. The DESCRIPTIVES command is rather limited. For a broader set of descriptive statistics including 5% trimmed mean, median and *IQR* use the EXAMINE command.

```
EXAMINE
 VARIABLES=D1
 /PLOT NONE
 /STATISTICS DESCRIPTIVES
```

The EXAMINE command will include plots by default (and the syntax above suppresses that with the /PLOT NONE subcommand. SPSS can produce a wide range of plots, but it is not as versatile as R. A plot similar to the bar plot in Figure 1.1 can be produced with the following syntax:

SPSS data file: eyes.sav

```
GRAPH
 /BAR(SIMPLE)=COUNT BY eye_colour
 /TITLE= 'Frequency of eye colour for a sample of 10 people'.
```

Box plots and stem and leaf plots are also available. A box plot similar to that in Figure 1.2 can be plotted with the following syntax:

SPSS data file: D1toD4.sav

```
EXAMINE VARIABLES=D3
 /PLOT=BOXPLOT
 /STATISTICS=NONE.
```

For a stem and leaf plot (such as that in Example 1.11) try:

```
EXAMINE VARIABLES=D3
 /PLOT STEMLEAF
 /STATISTICS NONE.
```

1.8.2 Other descriptive statistics

SPSS provides a wide range of descriptive statistics and exploratory plots – notably from the EXAMINE and EXPLORE commands. Arithmetic can be performed on variables via the COMPUTE command. For simple arithmetic, however, you may find it easier to do calculations by hand (or to use spreadsheet software such as EXCEL).

The following syntax uses the compute command to calculate the natural logarithms of a variable and DESCRIPTIVES to calculate the mean of the transformed variable. The simplest way to obtain the geometric mean is then to take the exponent e^x of the result (2.478) using a calculator or spreadsheet.

SPSS data file: *C1 sample data.sav*

```
COMPUTE ln_D1 = LN(D1).
EXECUTE.

DESCRIPTIVES
 VARIABLES=ln_D1
 /STATISTICS=MEAN.
```

1.9 Bibliography and further reading

Gould, S. J. (1985) The Median isn't the Message, *Discover*, 6, 40–2.
Tukey, J. W. (1977) *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.
Velleman, P. F. and Wilkinson, L. (1993) Nominal, Ordinal, Interval, and Ratio Typologies are Misleading, *The American Statistician*, 47, 65–72.

2
Probability distributions

Contents

2.1 Chapter overview

This chapter introduces common probability distributions for discrete and continuous data. The main focus is on distribution functions for determining probability or probability density, cumulative probability and quantiles of a distribution. Key characteristics of probability distributions (e.g., skew and kurtosis) and key ideas such as the central limit theorem are reviewed. A central theme is the role of selecting an appropriate probability distribution in building a statistical model.

2.2 Why are probability distributions important in statistics?

The presence of uncertainty is fundamental to statistical inference. If there were no uncertainty when sampling from a population then there would be no need for inferential statistics. In the absence of uncertainty, there will be zero variability and the sample will match the population perfectly.

Example 2.1 Consider the set of numbers {1, 1} as a finite population (of size $N = 2$). As there is no uncertainty in the population there is no role for statistical inference. A sample of size $n = 1$ or $n = 2$ taken from the set {1, 1} will always estimate the population mean (or any other parameter) with perfect accuracy. Because the population does not vary, this would not change even if the population were infinitely large. This could be confirmed by sampling the finite population with replacement (replacing observations after they are sampled and, in effect, sampling from an infinite population). Nor does the probability with which an observation is sampled matter when the population does not vary – the lack of uncertainty makes the relative probability of sampling any particular population value (e.g., the first observation) irrelevant.

Contrast the above situation with the numbers {0, 1} also considered as a population. The probability of sampling a particular value (0 or 1) is now of the utmost importance in estimating a population parameter. For example, if the probability of sampling zero is .25 when sampling with replacement (i.e., $Pr(0) = .25$), then this is effectively sampling from an infinite population where 25% of the population take the value 0 and 75% take the value 1. The distribution of values in this population determines how accurately a sample of size n will estimate the population mean. So even this rather simplistic situation requires us to consider the *probability distribution* defined by the population.

Example 2.1 suggests that without knowing the distribution of the population being sampled it will be difficult to make accurate inferences about a population. An immediate difficulty is how to determine the distribution of values in the population. This may seem like an impossible obstacle to overcome. But it is important to realize that the sample itself contains some information about the population from which it was sampled (and this information increases with n). In addition, at least some further information about the distribution is available in any real study (e.g., about its upper or lower bound). Last, but far from least, you will nearly always be able to get by without knowing the precise population probability distribution. What is required is sufficient information about the distribution to meet the goals of the research (e.g., estimating one or more parameters with a certain degree of accuracy).

For many purposes it is possible to rely on a relatively small set of probability distributions that can capture many of the key characteristics of population. These distributions act as useful building blocks for a statistical model. The distributions focused on here are just a few of those most commonly used in statistical modeling in the human and behavioral sciences. Methods also exist for dealing with situations when data are not easily modeled by assuming one of these distributions (e.g., robust methods). Whatever assumptions are required for a model, however, some understanding of the most important characteristics of the probability distribution being sampled is required (see Box 2.1).

Box 2.1 Characteristics of discrete and continuous probability distributions

One of the most fundamental characteristics of a probability distribution is whether the distribution is *discrete* or *continuous*. A discrete distribution is one where the population (and hence sample) only contains specific values, usually integers, such as {0, 1} or $\{-3, -2, -1, 0, 1, 2, 3\}$. A probability distribution for discrete data involves a mutually exclusive pairing of a probability with each population value that could be sampled. This can be done, in the simplest case, by listing each value with its corresponding probability. If the range of possible population values is large, then this can be done more conveniently by specifying a functional relationship* between the population values and their probability:

$$f(x) = p(X = x) = \text{probability of } X \text{ taking the value } x$$

For discrete data the functional relationship $f(x)$ gives the probably of X taking a particular value and is called a *probability mass function* (*pmf*).

It is common to represent the *pmf* as a plot of probability (from 0 to 1) on the *y*-axis and the values (x) that the distribution can assume on the *x*-axis. A *pmf* for the number of heads obtained when tossing a fair coin ten times is shown in Figure 2.1. In a *pmf* the probabilities of different values (denoted by the height of the lines in Figure 2.1) must sum to 1.

It is possible to obtain the arithmetic mean or expected value of the distribution by multiplying each probability on the *y*-axis by the value on the *x*-axis and summing the results. In this case the expected value is:

$$E(x) = 0.0009765625 \times 0 + 0.0097656250 \times 1 + 0.0439453125 \times 2 +$$
$$0.1171875000 \times 3 + 0.2050781250 \times 4 + 0.2460937500 \times 5 +$$
$$0.2050781250 \times 6 + 0.1171875000 \times 7 + 0.0439453125 \times 8 +$$
$$0.0097656250 \times 9 + 0.0009765625 \times 10 = 5$$

For a discrete distribution the mode is simply the value (or values, if more than one) with the highest probability and therefore the tallest line (e.g., 5 in Figure 2.1). The median is the value that is halfway through (or at the 50th centile) of the probability distribution. This is easiest to show by plotting a *cumulative distribution function* (*cdf*) for the distribution.

A *cdf* is very similar to a *pmf* except that rather than the probability of a value being plotted, the probability of obtaining that value or lower – the cumulative probability of x – is plotted on the *y*-axis. Figure 2.2 shows the *cdf* for the number of heads from ten tosses of a fair coin. In this figure the lengths of the lines indicate the probability of observing x or fewer successes (and hence each line is taller than its height in the *pmf* by an amount equal to the height of the preceding *cdf*

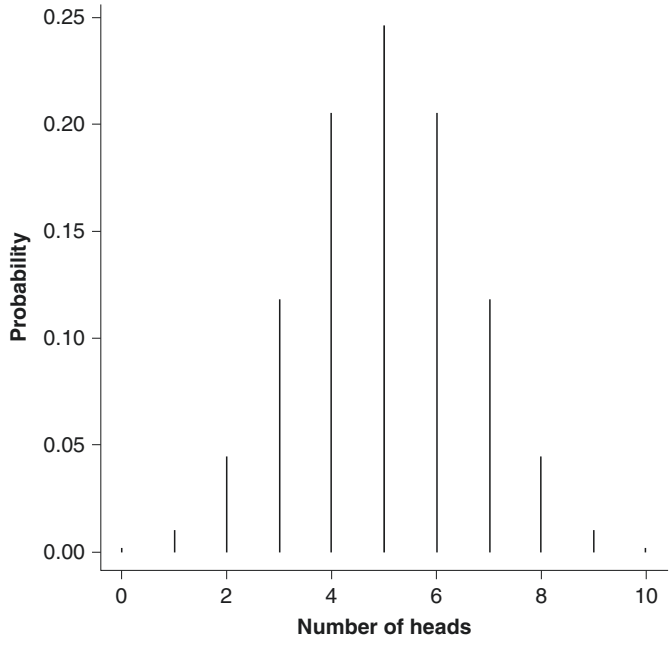

Figure 2.1 Probability mass function for the number of heads observed from 10 tosses of a fair coin

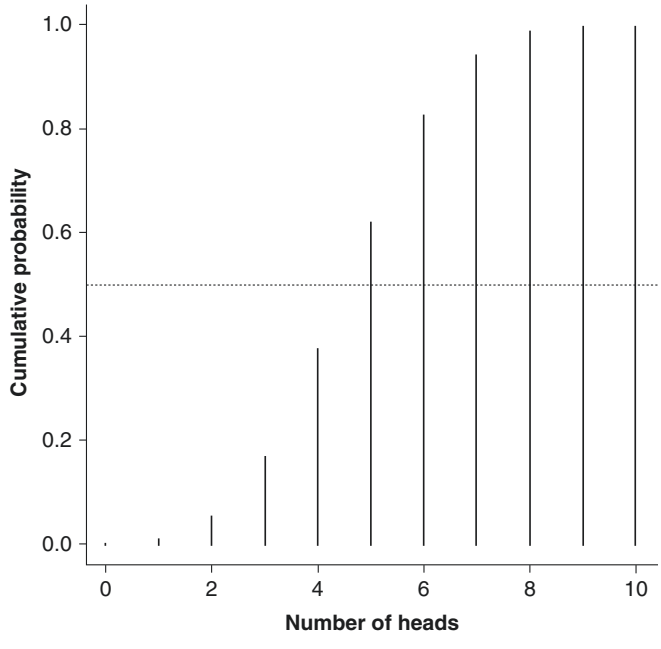

Figure 2.2 Cumulative distribution function for the number of heads observed from 10 tosses of a fair coin

line). The median is the first value for which the probability (or height of the line) includes .5. For the present example this is again 5 (as can be seen by following the dotted line at $P = .5$).

Unlike a discrete distribution, continuous distributions are not restricted to specific values such as integers and can, as a rule, take any value in between the lower and upper bound of the population (assuming it is bounded). A continuous probability distribution therefore has a smooth function rather than the characteristic 'spiky' function of a discrete probability distribution. An implication of this is that, unlike the discrete case, the probability of any particular value is zero. As a consequence, a probability can only be obtained for an interval (a range of values) within a continuous distribution. This makes plotting the probability directly as a *pmf* problematic. The solution is to specify the functional relationship between the population values and their *probability density* rather than the probability *per se*:

$$f(x) = p(X = x) = \text{probability density of } X \text{ at the value } x$$

This relationship for a continuous distribution is known as the *probability density function* (*pdf*). Like a *pmf* a *pdf* is often represented in graphical form, but in this case the *pdf* is a smooth curve representing the density at a given population value, while the probability is represented by the area under the curve (with the total area summing to 1). The probability of observing a value in the interval between the values *a* and *b* is

$$Pr(a \leq X \leq b) = \int_a^b f(x)dx \qquad \text{Equation 2.1}$$

This equation appears more complex than it is. The right-hand side of the equation is the *integral* for the *pdf* between the values *a* and *b*. (Integration is a mathematical procedure for calculating the area under a curve and an *integral* is the calculated area under a section of the curve.) Equation 2.1 thus indicates that calculating the area under the curve of a *pdf* between two values gives the probability of obtaining values in that range.

The population mean, median and mode can also be derived from continuous distribution. The mean is obtained via integration of the product of each value and its probability density:

$$\mu = \int_{-\infty}^{\infty} x\, f(x)\, dx \qquad \text{Equation 2.2}$$

The median is the value that divides the area under the probability density function exactly in half, while the mode is the value with highest peak (or peaks if more than one) in the *pdf*. Note that plotting a *cdf* for a continuous distribution is not a problem (because cumulative probabilities are always defined over intervals from zero to *P*). A *cdf* for a continuous function can also legitimately be plotted as a smooth curve.

It is important to realize that probability distributions also differ in other important ways. In particular, either a discrete or continuous distribution could be *bounded* or *unbounded*. For example the *pmf* in Figure 2.1 is for a distribution bounded at 0 and at 10. Many continuous distributions are also bounded (e.g., heights or weights are bounded at 0). Placing such limits means that values outside these limits are logically impossible. Therefore a continuous distribution such as the time (in seconds) it takes to drink a cup of coffee is bounded at 0 but has no logical upper limit.

Some probability distributions (e.g., that in Figure 2.1) are *symmetrical*, such that the probability function is a mirror image (i.e., identical but reversed) about the median. In a symmetrical distribution the mean and median take the same value. If the distribution is also unimodal (has a single peak and hence mode) then the population median, mean and mode are usually also all equal.** (More generally, in a multimodal *symmetrical* distribution, one of the modes will coincide with the mean only if there are an odd number of modes.) These relationships also hold between sample

distributions and sample statistics (e.g., if the sample is symmetrical around the sample median, this value is also the sample mean).

A vital step in statistical modeling is to match characteristics of the data – more accurately characteristics of the process that generated the data – to a suitable probability distribution (or at least to narrow the choice down to a family or set of related distributions). Particularly important characteristics for this purpose are whether a distribution is discrete or continuous, symmetrical or asymmetrical and whether it is bounded in some way.

* The term $f(x)$ is just a general way to express the output of a function f when the input is the observation x.
** The exception is for discrete distributions with even number population values (see Key Concept 2.2).

2.3 Discrete distributions

2.3.1 The binomial distribution

In introducing probability distributions, the notion of sampling with replacement from a very simple population – a population containing only the values 0 and 1 – was invoked (see Example 2.1). If the probability of sampling the value 1 is fixed at, say, $Pr(1) = .75$ then a single sample of $n = 1$ from such a population is known as a *Bernoulli trial* and its distribution is termed a *Bernoulli distribution*. A Bernoulli trial can be used to model the outcomes of a process that has two mutually exclusive outcomes (e.g., the sex of a newborn baby – in which case the values represent 'male' and 'female').[1] A convenient shorthand is to label 1 as 'success' and 0 as 'failure'. The expected value or mean of a Bernoulli trial is equal to P (defined as the probability of a 'success'). The variance is equal to $P(1 - P)$, where $1 - P$ is the probability of a 'failure'.

What happens if n independent observations are sampled from such a population? The resulting probability distribution is known as the *binomial distribution*. As it happens, just such a situation is shown Figure 2.1 (which shows the probability mass function for a fair coin tossed ten times). Figure 2.1 shows the *pmf* for a binomial distribution with $P = .5$ and $n = 10$ (where 1 represents 'heads').

The binomial distribution is a theoretical distribution that is fundamental to many statistical models. For example, consider a recognition memory experiment in which 20 participants are presented with an item and then offered five options (the correct item and four foils) at test. In this situation, random guessing can be modeled as a binomial distribution with $n = 20$ and $P = .2$.

The *pmf* of a binomial distribution for a variable X is:

$$f(x; n, P) = \binom{n}{x} P^x (1 - P)^{n-x} \qquad \text{Equation 2.3}$$

Again, this formula may appear complex, but if you understand the notation (and some basic probability theory) it is reasonably simple (see Box 2.2). The symbols n and P refer to the number of independent binomial trials and the probability of success respectively, while x is the observed number of 'successes'. Thus $1 - P$ is the probability of a failure and $n - x$ is the observed

number of 'failures'. Overall, the equation indicates that the probability of observing x successes is obtained by multiplying the number of possible combinations of trials with x successes by the probability of a sequence with x successes: $P^x(1-P)^{n-x}$. The number of possible combinations is given by the binomial coefficient described in Box 2.2.

Box 2.2 Probability, combinations and the binomial coefficient

A full introduction to even the basics of probability theory is outside the scope of this chapter, but it is worth introducing (or reviewing) some important points. First, if a and b are independent events (that is, the outcome of one has no influence on the other) then the probability of a occurring followed by b is $Pr(a) \times Pr(b)$. For example, if you toss a fair coin twice the probability of it coming up heads followed by heads again (HH) is

$$Pr(H) \times Pr(H) = Pr(HH) = .5 \times .5 = .5^2 = .25$$

Similarly, the probability of tails followed by tails (TT), or the probability of heads followed by tails (HT) is also .25. Together, these probabilities sum to .75 (not to 1). Why is that? Because a fourth option, tails followed by heads (TH) – also with probability .25 – has been missed out.

To find out how likely any set of outcomes (two heads, two tails or one head and one tail) is, you need to know not just the probability of each particular outcome, but also the number of ways in which it can occur. The number of ways an unordered outcome can occur is known as the number of *combinations* (and is distinct from the four ordered permutations HH, HT, TH and TT). In the above example, two heads can only occur in one way (HH), whereas one head and one tail can occur two ways (HT or TH). While it often helps to think of combinations in terms of sequences, the same principles apply to any set of independent events (e.g., tossing two coins labeled A and B simultaneously; thus HT could indicate that A is heads and B tails, and TH that A is tails and B is heads).

Thus in the *pmf* for the binomial distribution the $P^x(1-P)^{n-x}$ term represents the probability of obtaining a particular combination of x successes in n trials. For two fair coins $n = 2$ and $P = .5$ so the probability of two heads is $.5^2(1-.5)^{(2-2)} = .25(1-.5)^0 = .25(1) = .25$. To get the number of possible combinations there are two main methods. The first method is to list all the possible outcomes and count the relevant ones. Although time-consuming, it is a good way to understand what is going on when n is small. The second method is to use the following formula to find $\binom{n}{x}$:

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} \qquad \text{Equation 2.4}$$

The exclamation mark denotes the factorial function. The factorial of a non-zero integer x (e.g., 4) is obtained by multiplying all integers from 1 up to x together (e.g., $1 \times 2 \times 3 \times 4 = 24$). The factorial of zero is a special case and defined as $0! = 1$. When tossing two coins, there is only one combination of two heads (or two tails):

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

The term $\binom{n}{x}$ is known as the *binomial coefficient* and is a general way to refer to the number of ways of selecting x combinations from a total of n things.

The mean (expected value) of a binomial distribution can be obtained by summing the probability of x successes over all possible outcomes:

$$\mu = \sum_{x=0}^{n} f(x; n, P) = nP \qquad \text{Equation 2.5}$$

Equivalently, given that there are n independent trials, each has P chance of being a success and hence the mean of a binomial distribution is nP. The corresponding variance is $nP(1-P)$.

Although the binomial distribution is unimodal, it is symmetrical only if $P = .5$. In this case, nP is also the median. The median itself always falls between $nP - 1$ and $nP + 1$, while the mode is the largest integer smaller than $(n+1)P$. This carries the implication that the mean, median and mode are also identical if the mean is an integer. When the median and mode differ the mean lies between them (Kaas and Buhrman, 2008). This is necessarily true if the mean is not an integer: it will lie in between the median and the mode (one of which must be $nP + 1$ and the other $nP - 1$).

If a variable X has a binomial distribution this can be written as:

$$X \sim B(n, P)$$

B is shorthand for binomial while n and P refer to the parameters of the distribution (the tilde symbol can be read as 'is distributed as'). Here a particular sense of the term *parameter* is invoked. If you know both n and P, and that the distribution of X is binomial then no other information is needed – between them these parameters completely specify a binomially distributed variable. The estimators of such a set of parameters are sometimes termed *sufficient statistics* (see Box 2.3). Any other parameter of the binomial distribution (e.g., median, mode, range or variance) can be deduced once n and P are known.

Box 2.3 Sufficient statistics

Sufficient statistics are summaries of data that preserve all the information a sample provides about a population parameter. For example, $\hat{P}$ is a sufficient statistic for a Bernoulli distribution, whereas n and $\hat{P}$ are (in combination) sufficient statistics for a binomial distribution. A better-known example is that $\hat{\mu}$ and $\hat{\sigma}$ are sufficient statistics for the normal distribution.

When data are assumed to have been sampled from a particular distribution, sufficient statistics play a particularly crucial role as descriptive statistics. If the assumption is correct then the sufficient statistics offer a complete description of the population distribution. In practice it is unreasonable to assume data are sampled from a perfect binomial or other distribution. Even so, if the distribution is closely approximated by some ideal distribution, sufficient statistics for that distribution are undoubtedly a very powerful way to summarize and communicate what is going on (e.g., potentially allowing readers to check, re-analyze or conduct alternative tests of published data). Strong distributional assumptions are not always adopted, but sufficient statistics (usually in conjunction with other descriptive statistics and checks on a statistical model) are the starting point for most summaries of research. For example, they are the standard for reporting many statistical procedures in psychology (e.g., APA, 2010).

Applications of the binomial distribution often involve not the number of successes x but the proportion of successes x/n. For example, a researcher might be interested in the proportion

of successes from n trials in a memory experiment (e.g., because this might be comparable between experiments with similar stimuli but different n). This follows the same distribution, except that dividing by n produces a mean of P and a variance $P(1-P)$. A sample estimate of the probability of successes is therefore also an estimate of the proportion of successes in the population. I will refer to this sample estimate as $\hat{P}$ ('P-hat') to distinguish it from the population parameter P.

Example 2.2 Consider a rudimentary extra-sensory perception experiment in which a supposed psychic (the 'sender') concentrates on one of three shapes (a circle, square or triangle). On each trial a participant (the 'receiver') is presented with pictures of the three shapes and asked to pick out the one that the sender is concentrating on. This procedure is repeated until $n = 6$ trials are completed. If the receiver is guessing at random then all options are equally probable and the probability of success $P = 1/3$. The assumed population mean is $nP = 6(1/3) = 2$ (as is the median and mean). The variance of this distribution is $nP(1-P) = 2(2/3) = 4/3$ or approximately 1.33. This could easily be re-expressed in terms of proportions of successes rather than number of successes. Dividing by n gives a mean of 1/3 and a variance of $4/18 = 2/9$ or about .22. The corresponding standard deviation (*SD*) could be estimated as .47 (the square root of .22). If four successes are observed then $\hat{P} = 4/6 = .67$. Although the observed proportion is fairly high it is not inconsistent with the assumed distribution of random guesses. Its *SD* of .47 gives a rough idea of how far a typical sample proportion might fall from the true mean.

Given these assumptions about the population it is possible to work out the probability of any particular outcome using Equation 2.3 and Equation 2.4. The probability of four successes turns out to be:

$$\left(\frac{6!}{4!(2)!}\right) \times \left(\frac{1}{3}\right)^4 \times \left(\frac{2}{3}\right)^2 = \frac{720}{48} \times \frac{1}{81} \times \frac{4}{9} = 15 \times \frac{4}{729} \approx .0823$$

Cumulative probabilities can also be calculated using these equations. The probability of four or more successes $Pr(x \geq 4)$ would be approximately 0.0823 plus the probability of five successes (0.0165) and six successes (0.0014) for a total of around 0.10. Although these calculations might seem precise, it is important to remember that they are limited in several ways. First, as noted earlier, the variance is probably quite large. Second, the probability model required an assumption (albeit a fairly reasonable one) about the value of the parameter P. Third, the binomial distribution assumes that the six trials are independent (which is almost certainly not true in this case – as all six responses are made in sequence by the same person). This would be violated if both receiver and sender have a preference for selecting a particular shape or sequence of shapes. Such a problem could be dealt with by generating the targets at random. Thus the design of the study can have an impact on the suitability of the probability model.

2.3.2 The Poisson distribution

The *Poisson distribution* is often selected to model frequency or count data – data that arise from counting the number of occurrences of an outcome within a particular area or time period. A *Poisson process* (something that generates a Poisson distribution) is one in which independent, discrete events occur over time or space at a continuous rate. This means that the number of events observed depends only on the length of the time period or the size of the area sampled.

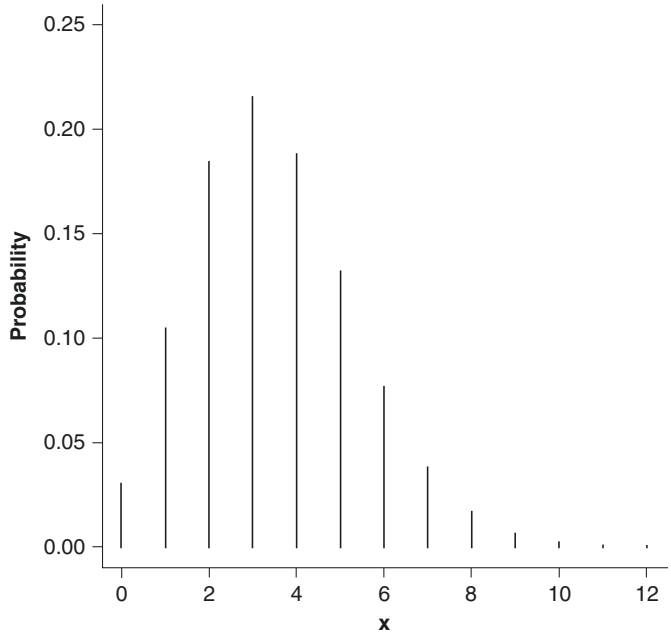

Figure 2.3 Probability mass function for a Poisson distribution with $\lambda = 3.5$

It is important to realize that, although each event is discrete, the time period or area sampled is continuous and could be divided into smaller segments (or pooled into a larger segment). The Poisson distribution is a natural starting point for modeling, say, the number of hits on a website or accidents reported in a one-week period in a particular workplace. The Poisson distribution has a single rate parameter λ (lambda) that determines the number of outcomes observed in a given sample. Figure 2.3 shows a Poisson probability mass function for $\lambda = 3.5$. A Poisson distributed variable, X, can be denoted as:

$$X \sim Pois(\lambda)$$

The probability mass function for the Poisson is:

$$f(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \qquad \text{Equation 2.6}$$

A fundamental characteristic of the Poisson distribution is that the single rate parameter is both the mean and the variance of the distribution. Also, as one would expect from a discrete distribution used to model count data, it is bounded at zero (but has no upper bound). A consequence of this is that the distribution is notably asymmetric when λ is small (the distribution tends to be squashed together toward zero and stretched out toward the upper bound, indicating *positive skew* – see Key Concept 2.2), but becomes more symmetrical as λ increases. Figure 2.4 illustrates this using the *pmf* for the Poisson distribution for $\lambda = 2$ and $\lambda = 5$. Also apparent is how the distribution spreads out as the mean (and hence its variance) also rises.

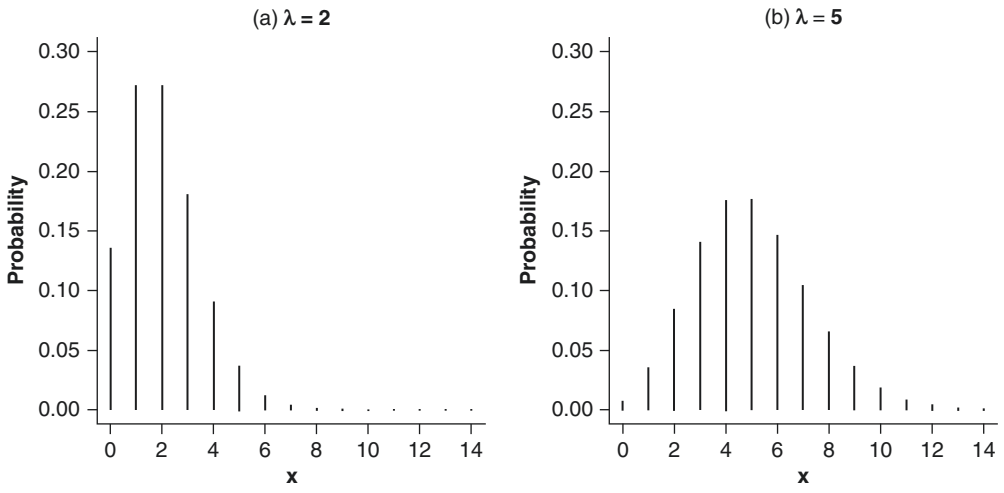

Figure 2.4 Probability mass functions for Poisson distribution with different rates

This relationship between mean and variance in the Poisson distribution offers one indication of its appropriateness in a statistical model. If the mean and variance of a sample differ markedly then a Poisson distribution may not be appropriate (at least not without considering additional factors that may be influencing the variability of the observed counts). Although generally asymmetrical, the Poisson distribution is unimodal. Its mode is the largest integer less than λ, unless λ is an integer (in which case both λ and $\lambda - 1$ are modes[2]). The median is less easy to pin down because of the asymmetry of the distribution. The median will often be close to λ, but there is no simple rule to describe its location (see von Hippel, 2005).

In the preceding section the binomial distribution was introduced as the sum of n independent Bernoulli trials with a fixed probability. What happens if two or more independent Poisson distributions are summed? It turns out that the sum of independent Poisson distributions is also a Poisson distribution with λ equal to the sum of the rates of the constituent distributions. For example, if the frequency of arguments between two couples undergoing therapy is rates of two per week and three per week respectively, the total rate of arguments is five per week. This is useful, because any number of couples (or other measurement units) with different rate parameters might, in principle, be modeled (provided the observations on each unit are independent).

So far, the discussion has focused on the Poisson distribution as the number of independent, discrete events occurring at a given rate. It is also possible to consider the Poisson distribution as an approximation of the binomial distribution appropriate when events are rare (i.e., P is very small). Under these circumstances the binomial distribution approaches the Poisson distribution with $\lambda = nP$ for a fixed value of λ as n becomes very large. Hence you can think of the Poisson distribution as a distribution for rare events if there are very many independent opportunities for the event to occur. At any point in time, a large number of people might potentially carry out some act (e.g., make a fraudulent insurance claim) but each does so with only a small probability. Thus a binomial distribution with $n = 5000$, $P = .001$ is approximated by a Poisson distribution with $\lambda = 5$. The binomial probability of observing exactly two such rare events is 0.08416534, while the Poisson approximation is 0.08422434.

Example 2.3 Consider the case of the number of arguments reported by a couple undergoing therapy. If a couple reported seven arguments in two weeks then $\lambda = 3.5$ per week and the probability of a couple reporting exactly two arguments in a week would be:

$$\frac{3.5^2 e^{-3.5}}{2!} = \frac{12.25 \times .030197}{2} = \frac{.36992}{2} = .18496$$

You'd expect such a couple to report two arguments in a week around 18% of the time. As just noted, this can also be thought of as an approximation to a binomial distribution where arguments occur with fixed probability $\hat{P}$ and an unknown number of trials. Working with a binomial distribution directly would be problematic – partly because it would not be practical to estimate the number of opportunities for arguments in any week.

2.3.3 Other discrete distributions

There are many other common (and many less common) discrete distributions of interest to researchers. Two, in particular, are worth a brief mention for their links to the binomial and Poisson distribution.

The first is the *negative binomial distribution* (sometimes known as the *Pascal distribution*). This is the distribution of the number of independent trials required to achieve a certain number of successes or failures, where P – the probability of a success – is fixed. Thus the negative binomial distribution is a kind of reverse form of the binomial (where the number of outcomes is fixed but n varies). The negative binomial model is also widely employed in a different context, where it acts as a substitute for the Poisson. This application arises because the negative binomial can be set up to mimic the behavior of the Poisson. Because it has an additional parameter it can model count data where the mean is not equal to the variance (see Chapter 17).

The second distribution is the *multinomial distribution*. This can be thought of as a generalization of the binomial distribution to situations with k discrete outcomes each with a fixed, independent probability (P_k) and where $P_1 + P_2 \ldots P_k = 1$. Furthermore, k independent Poisson distributed variables will have a joint distribution that is multinomial (and with parameters determined by the number of Poisson trials and their respective λ). The multinomial distribution is a natural choice when modeling categorical outcomes with more than two options.

2.4 Continuous distributions

2.4.1 The normal distribution

The *normal distribution* (also known as the *Gaussian distribution*) is a symmetrical, unimodal, continuous distribution that plays a key role in many areas of statistics. One way to derive the normal distribution is as the sum of an infinite number of independent, random variables. Thus the normal distribution seems appropriate for situations in which data arise from a process that involves adding together contributions from a large number of independent, random events. For example, the distribution of male heights might reasonably be considered the outcome of

many independent, random genetic or environmental influences and hence (at least approximately) normal. The importance of the distribution in statistics derives primarily from its role in the *central limit theorem* (see Key Concept 2.1).[3] The central limit theorem suggests that under certain, fairly reasonable conditions, many statistics will follow an approximate normal distribution when n is large.

KEY CONCEPT 2.1

Sampling distributions and the central limit theorem

The central limit theorem (CLT) is the justification for many statistical procedures that assume a normal distribution (or at least approximate normality). It is a theorem about the *sampling distribution* of a statistic (a statistic being the sample estimate of a population parameter). The sampling distribution of a statistic is the distribution that is obtained by calculating the same statistic from an infinite number of independent samples of fixed n. It is therefore the hypothetical population distribution of sample statistics for a given sample size. Such a distribution is known as the sampling distribution of the statistic (e.g., in the case of the mean you would refer to it as the *sampling distribution of the mean*). The mean of a sampling distribution is the expected value (i.e., mean) of the original population the samples were drawn from. The *SD* of a sampling distribution is known as its *standard error* (the importance of which will become apparent when statistical inference is introduced). So the *SD* of the sampling distribution of the mean is termed the *standard error of the mean*.

There are a number of reasons why sampling distributions are interesting. First, these distributions determine the probability of observing a particular value of a statistic in any given sample (information essential for statistical inference). Second, the sampling distribution of a statistic can – and often does – differ from the population distribution of the data used to calculate the statistic. Don't assume that a sampling distribution has the same distribution as the population the original data are sampled from.

The central limit theorem states that the sampling distribution of a statistic approaches the normal distribution as n approaches infinity (its *asymptote*). This asymptote is the limit referred to in the theorem. (It is 'central' in the sense that it is fundamental to probability theory and statistics.)

There are restrictions on the generality of this result. The central limit theorem applies to any statistic that is computed by summing or averaging quantities. Thus it holds for variances or means, but not for all descriptive statistics (e.g., the *SD* is the square root of an average and the CLT does not hold for it). It also holds only for distributions with finite mean and finite variance. This might appear to include all possible distributions, but there are distributions (and not necessarily esoteric ones) that do not have finite means or variances. Nevertheless, for many practical applications a finite mean or variance is a reasonable assumption.

What is the practical impact of this? In essence it means that, provided certain fairly plausible assumptions are met and n is sufficiently large, the sampling distribution of a statistic will be approximately normal. In addition, the larger n gets, the better the approximation gets. Armed with the mean and *SD* of the sampling distribution it is therefore possible to use the normal distribution to estimate the probability of observing a particular statistic value or range of values. It should be immediately obvious that this makes the normal distribution incredibly versatile (even if the population that data are sampled from is decidedly non-normal).

Consider data sampled from a binomial distribution. It has finite mean and variance. The number of successes in a sample from a binomial distribution is the sum of n independent Bernoulli trials. It follows that the distribution of successes from a binomial distribution approaches the normal distribution asymptotically. Figure 2.5 shows histograms (frequency bar graphs) of the sampling distribution of successes from a binomial distribution when $n = 10$, 30 or 100 for $P = .35$ and $P = .15$. Each plot is based on only 100,000 samples (but this is enough to show the sampling distribution reasonably clearly). When $n = 100$

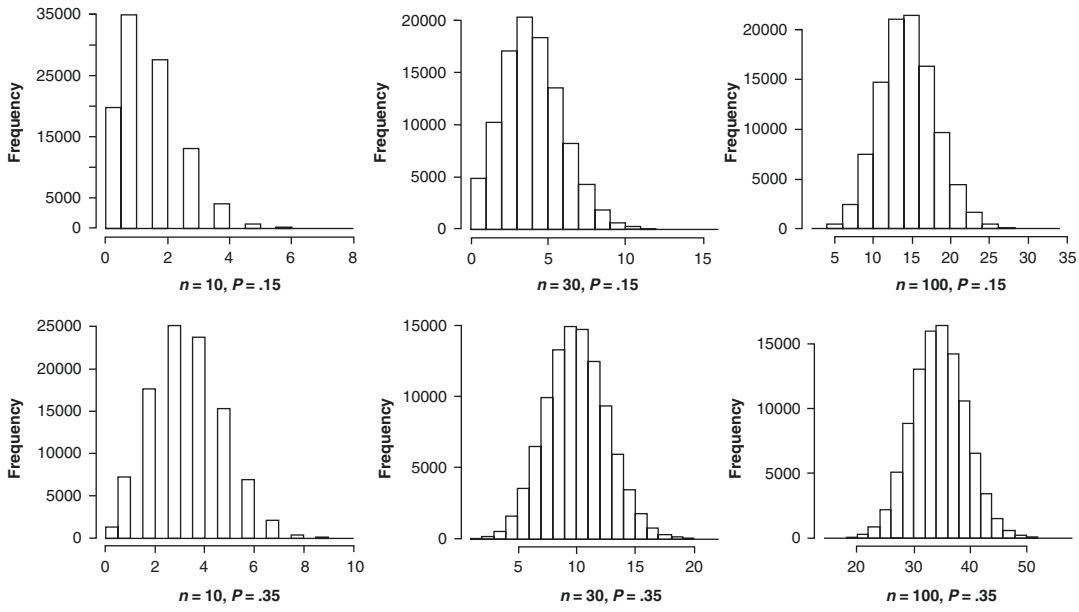

Figure 2.5 Histograms for 100,000 simulated samples from binomial distributions, with $P = .15$ (top row) or $P = .35$ (bottom row) and sample sizes of $n = 10$, $n = 30$ or $n = 100$

the distributions for both $P = .15$ and $P = .35$ look approximately normal, while both distributions look rather asymmetrical when $n = 10$. The main insight to be derived from Figure 2.5, however, is that the rate of convergence of the binomial distribution on the normal depends on P. Remember that the binomial distribution is symmetrical only when $P = .5$ and becomes increasingly asymmetrical as P approaches 0 or 1. Thus, when $P = .35$ the binomial distribution is more symmetrical than when $P = .15$. The similarity to the normal is evident for $P = .35$ even when $n = 30$ (and some skew is evident for $P = .15$ even when $n = 100$).

In general, the closer the original distribution is to the normal in shape, the more swiftly its sampling distribution converges on the normal.

As Figure 2.5 shows, different sampling distributions may well have very different rates of convergence. This poses a problem because it means it is not possible to state with any certainty that a given sample size (e.g., $n = 30$, $n = 100$ or even $n = 1,000,000$) allows the CLT to be invoked. In other words, no matter what n is selected there is no guarantee that the sampling distribution of a statistic will be even approximately normal (without additional conditions being imposed). That said, given some idea of the original distribution of the data (e.g., that it is approximately binomial with $P = .4$ or Poisson with $\lambda = 6$), it is easy to estimate what value of n will provide a reasonable approximation. Figure 2.5 suggests that for $n = 30$ the normal distribution provides a very satisfactory approximation to the binomial if $P = .35$, but might not be adequate for more extreme values of P such as .15 or .85.

Joliffe (1995) provides a neat example of the problem of convergence using the Poisson distribution. Recall that the sum of independent Poisson distributions itself has a Poisson distribution. Let's start by assuming that a distribution has $\lambda = 2$. If this distribution is, let's say, the sum of 100 independent Poisson distributions with $\lambda = 0.02$ it follows that the sum of means from a Poisson distribution is not guaranteed to be approximately normal when $n = 100$ (see Figure 2.4a). Thus Joliffe shows that while it tempting to use the Poisson to illustrate the CLT in action, the scenario works both ways. When λ is large the Poisson distribution is very well approximated by a normal distribution (e.g., Figure 2.4b suggests that the approximation may be acceptable even for $\lambda = 5$). However, we could pick any finite value of n and show

that summing *n* distributions would not be approximately normal (provided λ for the summed distribution was itself small). This state of affairs is not confined to the Poisson. A similar argument can be made for the binomial distribution as a sum of *n* Bernoulli trials for rare events (i.e., when *P* is small).

Although the central limit theorem is a remarkably powerful tool, it is worth reflecting on three common misunderstandings of it. The first misunderstanding is that it applies without restriction. This is not the case – and although widely applicable – there are some statistics and some distributions that are excluded. Second, it is often assumed that a given value of *n* (e.g., $n = 30$ is common) ensures that the theorem can be invoked. Again, this is untrue. Required *n* depends on the distribution of the original data (its shape and the precise parameters involved) and on how close an approximation is desired. The third misconception is the most troubling. The CLT is sometimes interpreted as a statement about the distribution of the original data rather than a statement about the sampling distribution of a statistic. It is possible that this arises because some distributions (e.g., the binomial or Poisson) can themselves be thought of as sampling distributions of statistics. Nevertheless, a moment's reflection should be sufficient to counteract this misunderstanding. If you were to sample a completely flat, uniform distribution (e.g., a process that generates random real numbers between 0 and 1) and calculate the mean, repeating this process a few thousand or million times will produce a reasonable approximation to the sampling distribution (an approximate normal distribution). However, the shape of this sampling distribution will in no way have had any influence on the process that generated those numbers. How could it? Its distribution remains uniform. A related argument could be made for any discrete distribution such as the occurrence of heads or tails when a coin is tossed. Even if the sampling distribution of the number of heads is close to normal (which it will be with large *n*) and well-described by a continuous function, the outcome remains discrete. So someone might predict, but never observe, 4.5 heads from nine tosses of a coin.

The parameters of a normal distribution are its mean μ and its variance σ^2. A normally distributed variable X can therefore be denoted as:

$$X \sim N(\mu, \sigma^2)$$

The probability density function for a normal distribution with $\mu = 100$ and $\sigma = 15$ is shown in Figure 2.6. For convenience, the variability of a normal distribution is often described in terms of σ (sigma) the population *SD* – for the same reasons that the sample *SD* is preferred to the variance as a descriptive statistic. Many psychological scales are deliberately constructed to be approximately normal and Figure 2.6 happens to show the hypothetical population distribution for many common IQ tests. All normal distributions have this characteristic 'bell' shape (though note that a number of other common distributions can also be described as 'bell-shaped'). As the normal distribution is both symmetrical and unimodal, its mean, median and mode are identical in the population.

The probability density function for a normally distributed variable, X is:

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad \text{Equation 2.7}$$

This function looks rather more difficult to understand than it really is – bear in mind that both e and π are constants (and π, the ratio of the diameter of a circle to its circumference, is found in many functions that describe curves). Hays (1973) explains that the 'working' part of the function is $-(x-\mu)^2/2\sigma^2$ (where x appears). For any particular normal distribution $x - \mu$ (the

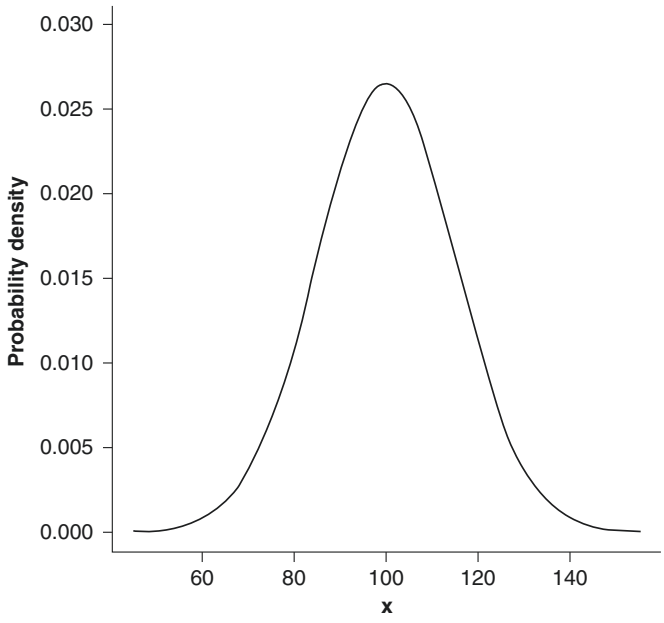

Figure 2.6 Probability density function for a normal distribution, with $\mu = 100$ and $\sigma = 15$

distance from the mean) determines its density. As this quantity is squared in Equation 2.7, the density is symmetrical around the mean. This part of the function appears as an exponent with a negative sign. Larger distances from the mean produce a more negative exponent, so the density is smallest when x is very far from μ and largest when $x = \mu$. Thus the distribution has a single mode. No matter what value x takes, the exponent of a positive number (such as the constant e) always produces an outcome greater than 0, therefore the probability density can never fall below zero (and hence the distribution is unbounded). All this holds provided σ is larger than zero (i.e., provided that there is any variability in X whatsoever). As with any continuous distribution, probabilities are defined by the area under the curve for a given interval (obtained by the integral of the curve between those points). Thus the total area under the curve equals the total probability:

$$Pr(-\infty \leq X \leq \infty) = \int_{-\infty}^{\infty} f(x)dx = 1$$

The spread of points around μ is determined entirely by σ (or equivalently by σ^2). Figure 2.7 shows normal distributions where μ and σ vary. It should be evident that changing the mean merely shifts the distribution right or left along the x-axis, while changing the standard deviation increases or decreases its spread (in the same way that constricting the x-axis would). This constriction is, in effect, just a rescaling of the x-axis: the same visual result is achieved by multiplying all the value labels on the x-axis by two or by halving σ.

An important feature of the normal distribution that is not particularly evident when it is plotted is that the probability density never reaches zero (and one convention is to terminate the left and right ends of the plotted curve in 'mid air' to suggest this). This is because the distribution

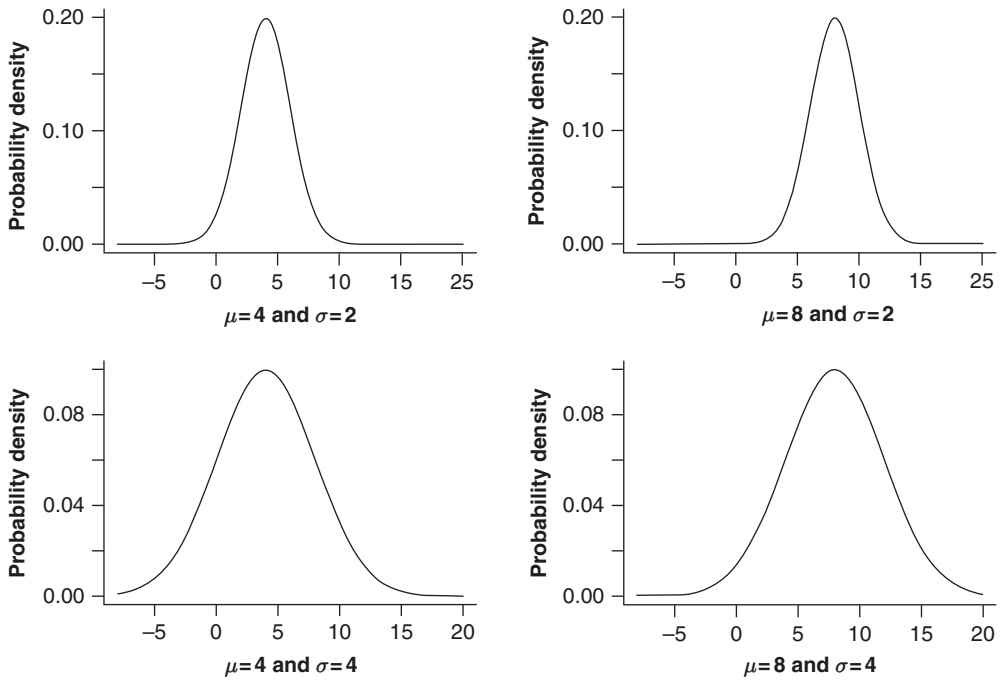

Figure 2.7 Normal distributions with differing parameters

is unbounded; ranging from $-\infty$ to ∞. In principle a value any distance from the mean might be observed (though the probability becomes extremely small as the distance exceeds several multiples of σ). This property may make the normal distribution unsuitable for some purposes (e.g., because the true population values are bounded or because the probability of very extreme values being observed is poorly represented).

If a variable has the distribution

$$z \sim N(0, 1),$$

then it is said to follow the *standard normal distribution*. The *pdf* and *cdf* of the standard normal are depicted in Figure 2.8.

The standard normal distribution (often abbreviated to *z*) is frequently used to simplify working with normal distributions. For the reasons considered above, any normal distribution can be shifted right or left, so that $\mu = 0$ (by subtracting its mean from all values). Likewise you can squash or stretch the distribution, so that $\sigma = 1$, by dividing all values by its *SD*. Aside from simplifying calculation (arithmetic using 0 and 1 being generally quite easy) it also simplifies some mathematical proofs involving the normal distribution. Particularly useful (historically at least) is the fact that the quantiles (and particularly centiles) of the standard normal distribution can be so easily mapped onto any other normal distribution and vice versa. A single set of tables of *z* quantiles, combined with a little arithmetic, can substitute for any normal distribution (and is particularly convenient when working without a computer). For example, the 50th centile (the median) lies at zero and approximately two-thirds of the distribution lie between -1 and 1.

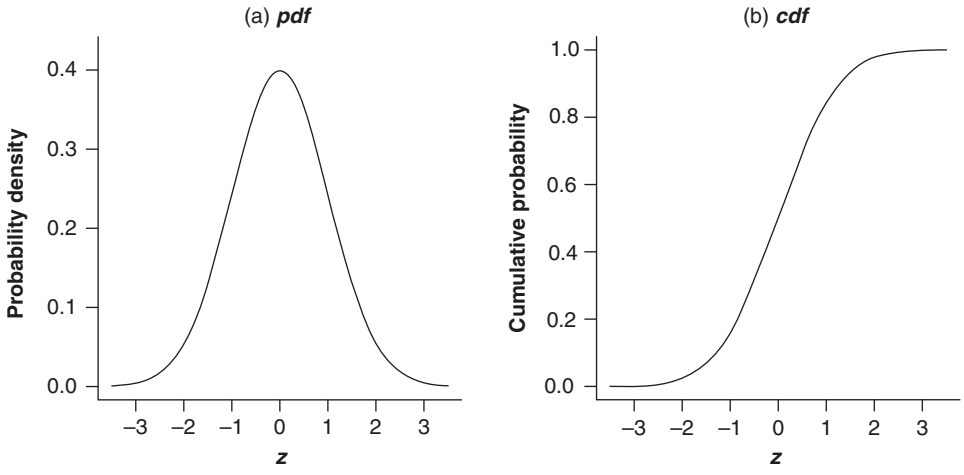

Figure 2.8 Probability density and cumulative distribution functions for the standard normal (z) distribution

For any normal distribution, around 67% of the population lie within $\mu \pm \sigma$, and around 95% of the population lie within $\mu \pm 2\sigma$. As the distribution is symmetrical it follows that about 2.5% lie above $\mu + 2\sigma$ and about 2.5% below $\mu - 2\sigma$. While these integer approximations are handy, more precise values can determined by computer or from tabulated values of z. These values could also be calculated by integrating the area under the standard normal curve bounded by particular values of z. The usual trick here is to use the *cdf* to obtain the cumulative probability for a particular tail value. When using the z distribution, this function is often labeled Φ (the Greek capital letter 'phi') and its inverse as Φ^{-1}. Hence $\Phi(-1.96) = .025$. This gives the cumulative probability of the left tail of z up to and including -1.96. You should be able to check this visually using Figure 2.8b. Because the *pdf* is symmetrical (see Figure 2.8a) you only need to look up values for the lower tail and double it to get the proportion of values falling $+/-1.96\sigma$ from the mean. (Using the left tail is easiest because the convention is to cumulate left-to-right.) Likewise, $\Phi(1.96)$ gives the cumulative probability up to and including $z = 1.96$. This is .975 and implies $1 - .975 = .025$ is the probability for values of z exceeding 1.96. To include 66.67%, 90%, 95% and 99% of the distribution, the correct values of z (to three decimal places) are ± 0.968, ± 1.645, ± 1.960 and ± 2.576.[4]

This property of the normal distribution makes the sample *SD* an especially useful descriptive statistic. Even if a distribution is known not to be normal, there is a theorem (*Tchebycheff's inequality*) that provides limits on the proportion of a distribution that can fall in each tail (see Hays, 1973). For any distribution with finite mean and variance, the probability of obtaining a value $\pm j\sigma$ from the mean of a distribution is always less than or equal to $1/j^2$. For example, the probability of obtaining a value $\pm 2\sigma$ from the mean is no more than .25. This can also be applied to the distribution as a whole. At least 75% of values from any distribution are no further than $\pm 2\sigma$ and at least 93.75% are no further than $\pm 4\sigma$ from the mean. If the distribution is both unimodal and symmetrical, it is possible to narrow the limits a little further using the *Vysochanskij-Petunin inequality*. Here the relevant quantity is $4/9\,(1/j^2)$. This more than halves the tail probability, so that around 89% of observations are no further than $\pm 2\sigma$ from the mean. Thus, even if a variable is not normal, $\hat{\sigma}$ provides at least a rough indication of how unusual an observation is.

Example 2.4 Researchers in the UK have recently argued that differences in teaching ability can have a substantial impact on public examination grades (Slater *et al.*, 2009). One estimate is that the difference between having a teacher in the top 5% or bottom 5% of teaching ability, relative to an average teacher, is around one letter grade (the difference between A and B or B and C) at GCSE (the main public exam for 16-year-olds in England and Wales). If teaching ability is assumed to follow a normal distribution (probably quite a strong assumption), the effect of teaching ability on GCSE grades could be modeled as a normal distribution with $\mu = 0$ and $\sigma \approx 0.3$.

On this basis, what is the expected impact of a teacher on the 25th centile of teaching ability? This can be from $\Phi^{-1}(.25) \approx -0.67$. A teacher at the 25th centile might be expected to reduce grades by about $.67\sigma$ or $0.67 \times 0.3 = 0.2$. This amounts to a fifth of a letter grade lower than they would otherwise obtain. It follows that a teacher on the 75th centile would be expected to increase performance by about a fifth of a grade.

The calculations also work the other way. What proportion of teachers would be expected to increase grades by a whole letter grade? This is approximately $1/0.30 = 3.33\sigma$. As $\Phi(3.33) \approx .9996$ it follows that this proportion is about $1 - .9996 = .0004$. This equates to .04% (or around one teacher in 2500). Note that by requesting the lower (left) tail cumulative probability I would have got directly to $\Phi(-3.33) \approx .0004$.

Although these estimates seem reasonable, they required quite strong assumptions about the distribution. If the distribution is not normal the estimated probabilities – particularly the extreme tail probabilities – could be very inaccurate.

2.4.2 The lognormal distribution

The lognormal distribution is an asymmetric, continuous probability distribution that, as the name implies, is normal under a logarithmic transformation. If a variable X has a normal distribution then e^X has a lognormal distribution. Correspondingly, if Y has a lognormal distribution then $\ln(Y)$ is normal. (Note that the base of the logarithm is irrelevant to this relationship – it merely acts as a scaling factor.) While the normal distribution is *additive* – arising as the sum of an infinite number of independent, random variables, the lognormal distribution is *multiplicative*. The lognormal thus applies when many independent, random variables are multiplied together (i.e., as the distribution of their product rather than their sum). This is reasonable when the effect of many independent and random influences is to induce a proportionate change in something (rather than merely adding or subtracting from it). Just as the sum of independent normal variables is normal, so the product of independent lognormal distributions is lognormal.

It is natural to exploit the link between the lognormal and the normal. For example, the parameters of the lognormal are usually defined as μ and σ^2 where μ and σ^2 are the mean and variance of the logarithms of the population sampled (rather than of the original population). A lognormal variable X can be written as

$$X \sim LogN(\mu, \sigma^2)$$

with the probability density function:

$$f\left(x; \mu, \sigma^2\right) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}} \qquad \text{Equation 2.8}$$

Representing the lognormal in this way (using natural logarithms) means that exponentiation to base e of μ and σ gives e^{μ}, the geometric mean, and e^{σ}, the geometric standard deviation. The variance of X is given by $\left(e^{\sigma^2}-1\right)e^{2\mu+\sigma^2}$, while the arithmetic mean of the original values is given by $e^{\mu+\frac{\sigma^2}{2}}$. The geometric mean of a lognormal distribution, e^{μ}, is also its median, whereas the mode is $e^{\mu-\sigma^2}$. These values are consistent with the asymmetry of the lognormal distribution – the median, geometric mean and mode are smaller than the arithmetic mean (which is weighted toward the long right tail of the distribution). This is shown clearly in Figure 2.9, which depicts the standard lognormal distribution (i.e., with $\mu = 0$ and $\sigma = 1$).

Data from a lognormal distribution are constrained to be greater than zero, but have no upper bound. The distribution is therefore a popular choice of continuous distribution for real world data bounded in this way (e.g., response times). The lognormal distribution is a common example of a positively skewed distribution (see Key Concept 2.2). The degree of skew depends on σ. When σ is small the skew is negligible (approaching symmetry as σ approaches zero). One implication of this is that, as skew decreases, the mean $e^{\mu+\frac{\sigma^2}{2}}$ and geometric mean e^{μ} become more similar.

It is therefore difficult to distinguish samples from normal and lognormal distributions if μ is large relative to σ. Limpert *et al.* (2001) make exactly this argument for much real world data; the normal distribution is often assumed when the lognormal is at least as plausible.

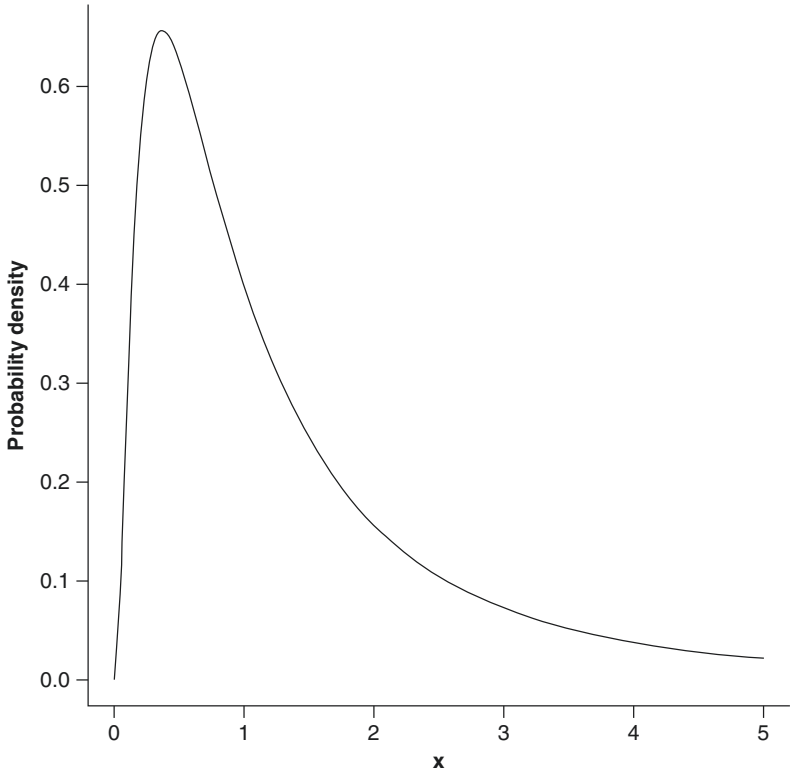

Figure 2.9 Probability density function for the standard lognormal

Example 2.5 Limpert *et al.* (2001, Table 2) report numerous examples of data sets with an approximate lognormal distribution. One, based on historic data from Boag (1949), is for survival after a diagnosis of mouth and throat cancer. Subsequent examples refer to these as the cancer survival data.

Survival data are notable for their positive skew (see Gould, 1985), and it seems reasonable that the lognormal distribution could be used to describe it. The geometric mean (also the median) is approximately 9.6 months with a geometric *SD* of 2.50. This distribution could be modeled as a normal distribution of the natural logarithm of the survival time (in months) or as a lognormal distribution with $\mu = \ln(9.6)$ and $\sigma = \ln(2.50)$. The advantage of using a lognormal distribution is that it retains the original units (survival in months).

Using this information one can predict the survival for patients in the left and right tails of the distribution, or the probability or surviving a certain length of time after diagnosis for a patient picked at random. The probability of surviving two years or more after diagnosis $Pr(x \geq 24)$ is about .16. A patient in the bottom 10% has an estimated survival time of about 3.0 months (6.6 worse than median). The skew of the distribution is obvious when you consider that estimated survival for a patient in the top 10% is about 31.1 months (21.5 months better than median). These summary data are now over 60 years old and survival after diagnosis will probably be much improved (though it may well still be adequately modeled by a lognormal distribution).

KEY CONCEPT 2.2

Skew

An asymmetrical distribution is said to be *skewed*. It is sensible to distinguish between *positively skewed* distributions (weighted to the right of the number line where the larger, more positive numbers are located) and *negatively skewed* distributions (weight toward smaller, more negative numbers on the left). When plotted with data values on the *x*-axis (e.g., for a *pmf* or *pdf*), positively skewed distributions look as if they have been stretched out to the right and negatively skewed stretched out to the left. The terms *right skew* and *left skew* are therefore often used interchangeably with positive and negative skew respectively. However, positive and negative skew are more general terms; if the data values are plotted vertically (on the *y*-axis) the terms right and left skew will be misleading.

Distributions bounded only at their left-most tail (e.g., at or near 0) tend to be positively skewed. This is apparent for both Poisson and lognormal distributions (see Figure 2.3 and Figure 2.9). Distributions bounded only on the right tend to have negative skew. Negative skew is less frequently encountered than positive skew, because many real world processes (e.g., response times, income, number of children) are bounded at zero (or close to zero). However, negatively skewed distributions are not rare. The binomial distribution (bounded on both the left and the right) is a good example. When $P < .5$ it is positively skewed, whereas for $P > .5$ it is negatively skewed. Figure 2.10 shows the *pmf* for the binomial distribution for $n = 20$ when $P = .08$ and $P = .92$.

For a positively skewed distribution the median is typically smaller than the mean. For negative skew it is usually larger than the mean. A common misconception is that this is always true, but there are many situations where this pattern does not hold (von Hippel, 2005). One reason for this misconception is historical. Karl Pearson proposed two simple measures of skew (*skewness*) based on the difference between mean and median. These measures define skewness as positive if the mean is larger than the median (and negative if it is smaller). The Poisson distribution is a particularly useful counter-example (see von Hippel, 2005). For around 30% of all values of λ the median is larger than the mean (even though the Poisson can never have negative skew). For example, if $\lambda = 1.9$ the median is 2. When $\lambda = 2$ the median is also 2, but the distribution is far from symmetrical. If you look back at Figure 2.4a it should be possible to see why.

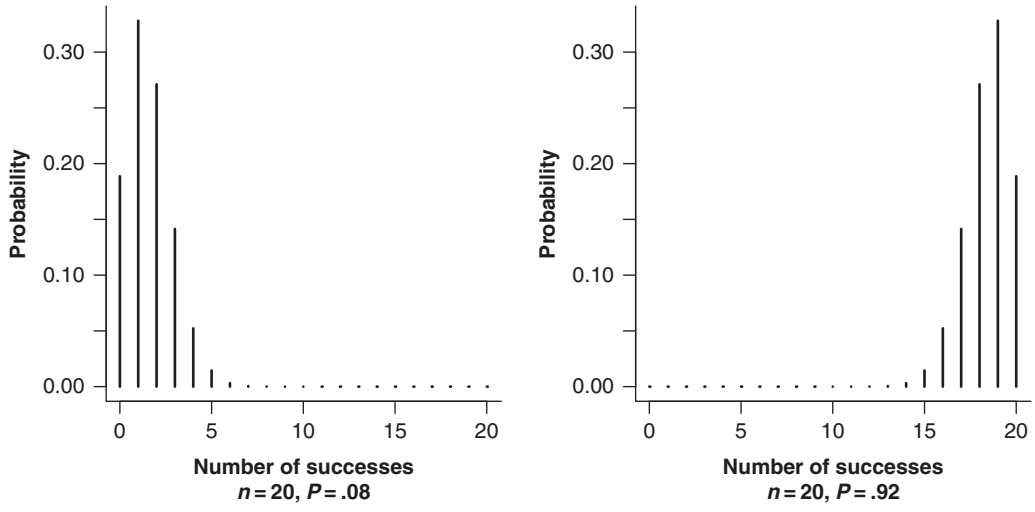

Figure 2.10 An illustration of the asymmetry of the binomial distribution, when $P \neq .5$

First, note that the mode is spread across 1 and 2. When λ falls just below 2, the peak shifts slightly so that the mode equals 1. The median is harder to shift and remains at 2.

Skew is more correctly defined in terms of the 3rd 'moment' (μ_3) of a probability distribution (where the mean is defined as the 1st moment and the variance the 2nd). This indicates that the skew depends on the cubed deviations of values from the mean (while the variance depends on the squared values and the mean on the unsquared, uncubed values). When comparing distributions, skewness is often 'standardized' (i.e., scaled in terms of σ). For example, skewness is typically represented as μ_3/σ^3 (and will be zero if a distribution is symmetrical).

In general, violations of the supposed 'rule' that the median is shifted away from direction of skew (relative to the mean) occur most often for discrete and bimodal (or multimodal) distributions. For continuous distributions such a pattern is never found if a distribution is unimodal, but can sometimes occur when a distribution is not unimodal. In addition, it is important to understand that a sample distribution can (and typically does) differ in shape from the distribution it is drawn from – so a negatively skewed sample could be drawn from a positively skewed population. Furthermore, if measurements are discrete (e.g., because of rounding error or the nature of the measurement tool being used), samples from a continuous distribution might behave somewhat erratically (relative to the continuous distribution they are supposedly sampled from).

2.4.3 The chi-square (χ^2) distribution

A good starting point for understanding the *chi-square* distribution[5] is to consider a squared observation, z^2, drawn at random from the z (i.e., standard normal) distribution. The sampling distribution of z^2 is a chi-square distribution – specifically it is a chi-square with 1 *degree of freedom* (*df*). Figure 2.11 shows the *pdf* for the chi-square distribution with 1 *df*.[6]

If k independent observations were sampled from a z distribution and each observation squared and summed (added together), the distribution would be chi-square with ν degrees

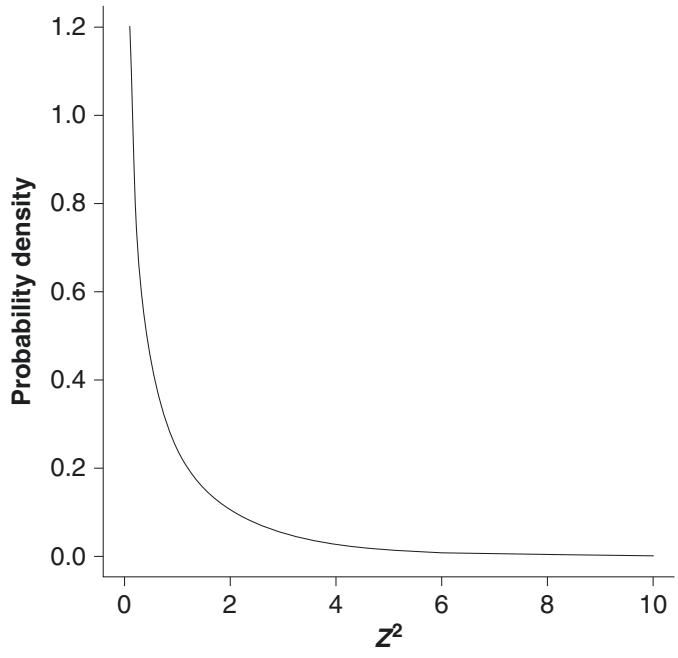

Figure 2.11 Probability density function for a chi-square distribution with 1 *df*

of freedom (where ν is the Greek lower-case letter 'nu'). The distribution has a single parameter ν, and this is always greater than zero. Why might we be interested in this distribution? The fundamental insight here is that z^2 is a special case of *sums of squares* where the data have a variance of 1. Given the intimate link between the calculation of sums of squares and the variance it turns out that the chi-square distribution is useful for modeling variances of samples from normal (or approximately normal) distributions.

If a variable X has a chi-square distribution with ν *df* it can be denoted as:

$$X \sim \chi^2_\nu$$

The mean or expected value of χ^2 is equal to its degrees of freedom ν and its variance is 2ν. Knowledge of the expected value of a chi-square statistic is particularly useful for large ν; it provides a quick way to gauge the fit of a statistical model (which tends to be good if the statistic is similar in value to its *df* and poor when markedly different from its *df*). The mode is $\nu - 2$ if $\nu > 2$ (and zero otherwise). In contrast, the median tends to be in the region of $\nu - 2/3$. Figure 2.12 shows the *pdf* for chi-square when for $\nu = 3$ and $\nu = 10$.

The chi-square distribution is positively skewed and bounded at zero – as should be expected for a distribution derived from sums of squares. Figure 2.12 hints that as the *df* rise, the distribution will become more symmetrical (and, in accordance with the CLT, it will ultimately converge on the normal distribution). This is indeed the case – and follows from the fact that chi-square is itself a form of sampling distribution (for the sums of squares of independent z).

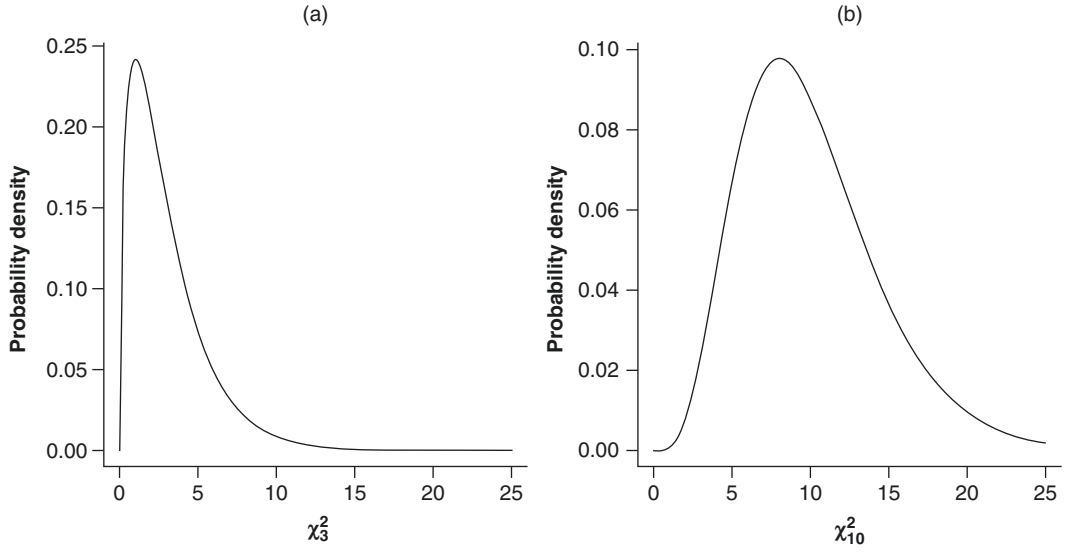

Figure 2.12 Probability density functions for chi-square distributions, with (a) $\nu = 3$, and (b) $\nu = 10$ *df*

The link between sums of squares and χ^2 is not restricted to the standard normal. The sums of squares of independent, random normal variables with variance σ^2 have a chi-square distribution scaled by σ^2:

$$\sum_{i=1}^{n} (x_i - \mu)^2 \quad \sim \quad \sigma^2 \chi^2_{n-1} \qquad \text{Equation 2.9}$$

An oddity of the chi-square distribution is that, although it can be derived as the distribution of sums of squares of discrete observations, the *df* are not restricted to integer values (though $\nu \geq 1$). Some statistical procedures make use of fractional *df* and it is helpful to realize that this is not necessarily an error.

Strictly speaking, the preceding discussion applies to the *central chi-square distribution*, and for some situations a *non-central chi-square distribution* is appropriate. This point is also relevant for the final two continuous distributions discussed in this chapter – the central *t* and central *F*. The distinction between central and non-central distributions is dealt with later, in relation to statistical power (see Chapter 8).

2.4.4 The *t* distribution

The *t distribution* (also called *Student's t*) is a sampling distribution for means from a normal distribution. As has already been established, the sampling distribution of means from a normal population is itself normal. However, estimating this distribution when the population standard deviation σ is unknown presents a practical difficulty. William Gossett (publishing under the

pseudonym 'Student') addressed this problem by showing that if σ is estimated from the unbiased variance estimate $\hat{\sigma}^2$, the resulting 'standardized' sample mean has what is now known as a t distribution with df equal to $\nu = n - 1$ (Student, 1908). A variable, X, with a t distribution can be denoted as:

$$X \sim t(\nu)$$

The t distribution is closely linked to both z and the χ^2 and it is the probability distribution of the ratio:

$$\frac{Z}{\sqrt{V/\nu}} \qquad \text{Equation 2.10}$$

The numerator Z is a variable with a standard normal distribution. The denominator is a chi-squared variable V with ν df. In addition, Z and V are assumed to be independent.

A common misconception is that the t distribution applies to the sampling distribution itself (whereas it applies only to this standardized form). This is, however, the form that is most likely to arise in practice, because σ^2 is rarely known for real data sets. Figure 2.13 shows the t distribution relative to the standard normal when $\nu = 1$ and $\nu = 29$ (corresponding to a single sample with $n = 2$ and $n = 30$).

Like the normal distribution t, is always symmetrical, but it tends to have a relatively narrower peak and 'fatter' or 'heavier' tails than z (at least when n is small). This characteristic is known as *leptokurtosis* (see Key Concept 2.3). Notably, the t distribution converges rapidly on z and is only barely distinguishable from z in Figure 2.13b when $n = 30$. Even so, the difference between z and t can be substantial for very small samples.

Being related to z (and hence unimodal and symmetrical), the median and mode of t are both equal to zero (and the mean is also zero if $\nu > 1$). If t has at least 2 df it has a variance of $\frac{\nu}{\nu-2}$. When t has only 1 df its mean and variance are undefined and it coincides with the standard

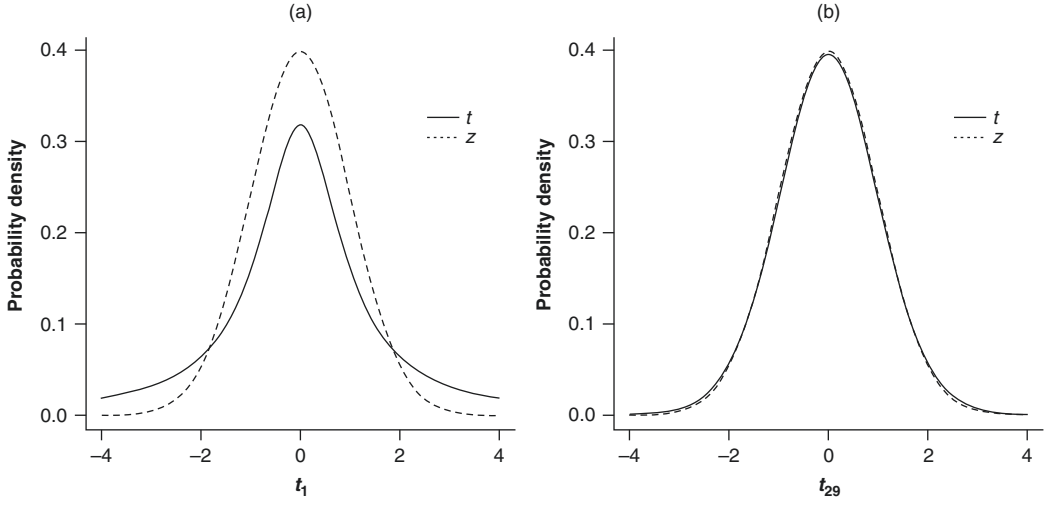

Figure 2.13 Probability density of t_1 and t_{29} relative to z

Cauchy distribution (a distribution which is the ratio of two standard normal distributions). This is an important reminder that the distribution of something as 'basic' as a *t* distribution with 1 *df* can fall outside the scope of the CLT. The CLT requires that the mean and variance are finite (see Key Concept 2.1).[7]

KEY CONCEPT 2.3

Kurtosis

Kurtosis is a term describing the relative proportions or 'weight' of a probability distribution in its middle versus its tails. For unimodal, symmetrical distributions (such as the normal or *t* distributions) it therefore refers to the relative weight of the distribution in its peak or its tails. A distribution with flat peak and relatively 'thin' or 'light' tails indicates *platykurtosis* (i.e., it is *platykurtotic*). A distribution with 'pointy' peak and relatively 'fat' or 'heavy' tails indicates *leptokurtosis* (i.e., it is *leptokurtotic*).

Kurtosis is widely misunderstood and is often incorrectly described solely in terms of how 'pointy' or 'peaked' a distribution is (see DeCarlo, 1997). Think about the normal distribution – it has the same degree of kurtosis regardless what values μ and σ take, yet it can be made arbitrarily more or less 'pointy' simply by decreasing or increasing σ. So adjusting the 'pointiness', either by altering σ or by adjusting the aspect ratio (the ratio of the scales of the *x*-axis and *y*-axis) of a plot, has no impact on kurtosis. Kurtosis must involve a *relative* shift of probability density (or mass) to or from the middle to the tails such that the variance is unchanged. For this reason the weight of density falling on the 'shoulders' of a unimodal distribution (the bits either side of the peak) tends to be particularly important.

Kurtosis is defined formally in terms of the 4$^{\text{th}}$ moment around the mean (μ_4). Although a normal distribution has positive standardized kurtosis (with $\mu_4/\sigma^4 = 3$), it is common to represent kurtosis as *excess kurtosis* relative to the normal distribution (which is defined as having zero excess kurtosis). For example, the *t* distribution with 5 *df* has excess kurtosis of 6 (relative to 0 excess kurtosis for the normal). The kurtosis of a distribution is a somewhat neglected topic in statistics (relative to skew), but the relative weight of data in the middle and tails of a distribution can be vital for statistical inference. Distributions with heavy tails (i.e., leptokurtosis) are particularly awkward to work with.

2.4.5 The *F* distribution

The *F distribution* (sometimes termed the Fisher or Fisher-Snedecor distribution) is a probability distribution for the ratio of variances of independent, random samples from populations with a normal distribution. The distribution is appropriate if the sample variances are unbiased estimates and provided the population variances (but not necessarily the means) are equal. If they are not equal, the *non-central F distribution* is appropriate. If a variable, *X*, has an *F* distribution it can be written as:

$$X \sim F(\nu_1, \nu_2)$$

The parameters ν_1 and ν_2 are the number of observations in each sample minus 1 (i.e., the denominator used in calculating an unbiased variance estimate $\hat{\sigma}^2$). Thus ν_1 and ν_2 are the *df* of two independent χ^2 variables. This is because *F* can also be defined as the ratio of two chi-square distributions divided by their respective *df*. Equation 2.9 indicates that (assuming

a normal distribution) independent sums of squares have the distribution: $\sigma^2\chi^2_{n-1}$. This implies that an unbiased sample variance has the distribution:

$$\hat{\sigma}^2 \sim \frac{\sigma^2\chi^2_{n-1}}{n-1} \qquad \text{Equation 2.11}$$

If it is assumed that the populations the two samples are drawn from have equal variance (as is the case for the *central F* distribution) then the unknown σ^2 will cancel out and the ratio of the two sample variances has the distribution

$$F = \frac{\chi^2_{\nu_1}/\nu_1}{\chi^2_{\nu_2}/\nu_2} \qquad \text{Equation 2.12}$$

where ν_1 and ν_2 are $(n_1 - 1)$ and $(n_2 - 2)$ respectively.

The shape of *F* distribution (being a ratio of two other distributions) is difficult to characterize. Figure 2.14 shows the *F* distribution with values of ν_1 ranging from 1 to 10 and with ν_2 set at

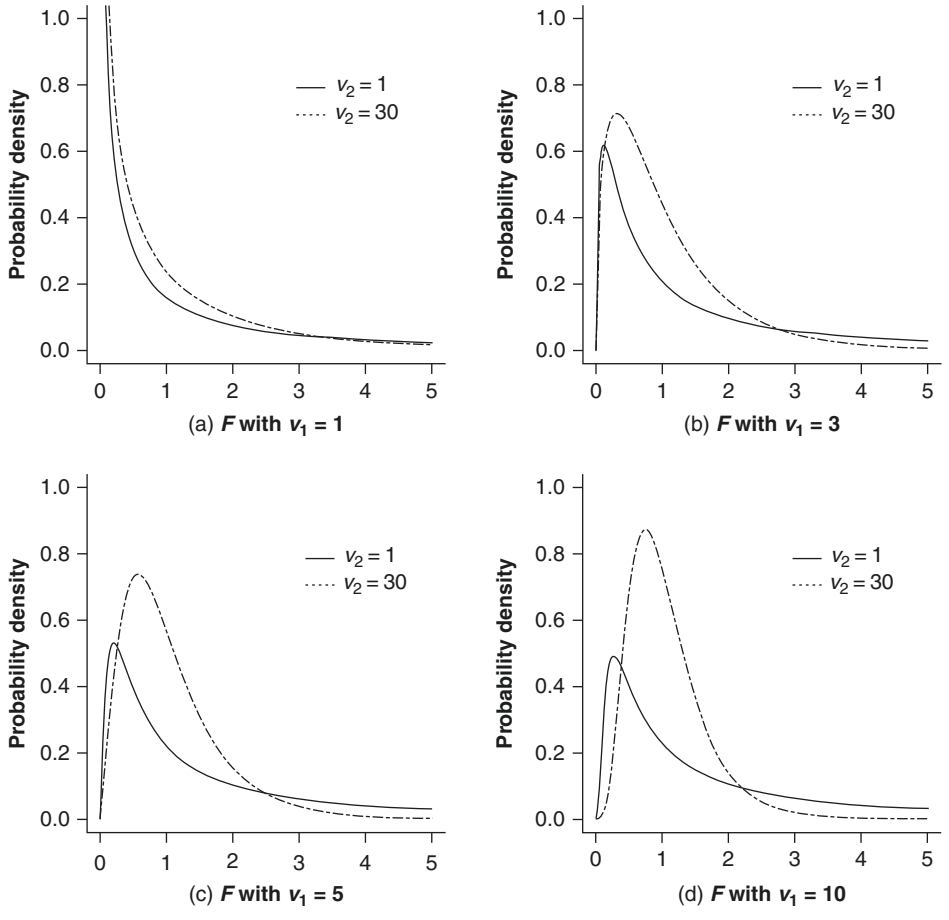

Figure 2.14 Probability density of *F*, with denominator *df* of $\nu_2 = 1$ or $\nu_2 = 30$ with numerator *df* equal to (a) $\nu_1 = 1$, (b) $\nu_1 = 3$, (c) $\nu_1 = 5$, or (d) $\nu_1 = 10$

either 1 or 30. This gives a rough idea of how the shape of the distribution can vary. As a ratio of two quantities that must be greater than zero, F is also bounded at zero (but can take any positive value greater than zero). Provided both that ν_1 is smaller than ν_2 and that ν_2 is larger than 2 (which is commonly the case for real applications) the distribution is positively skewed and unimodal. For these values of ν_2 the mean of the F distribution is $\nu_2/(\nu_2 - 2)$. This, you may recall, is also the variance of a t distribution when $\nu > 1$. As ν_2 approaches infinity the value of F tends toward 1.

F and t are intimately related (through the chi-square distribution). Using Equation 2.10 it is easy to show that

$$t^2 = \left(\frac{Z}{\sqrt{V/\nu}}\right)^2 = \frac{Z^2}{V/\nu}$$

The square of a standard normal variable such as Z has, by definition, a chi-square distribution with 1 *df* (i.e., $Z^2 \sim \chi^2_1$). V is chi-square with ν *df*. Given Equation 2.12 it follows that

$$t^2_\nu = \frac{\chi^2_1/1}{\chi^2_\nu/\nu} = F_{1,\nu} \qquad \text{Equation 2.13}$$

In other words t^2 is distributed as F with numerator *df* (ν_1) equal to 1 and denominator *df* (ν_2) equal to ν. It is therefore not surprising that the shape of F be similar to a χ^2_1 distribution when $\nu_1 = 1$ (as can be seen by comparing Figure 2.11 and Figure 2.14a). It is necessarily also true that $\sqrt{F_{1,\nu}} = t_\nu$.

Example 2.6 Distribution functions for the χ^2, t and F distribution can be employed in the same way as for the normal or standard normal. As they are continuous distributions the *pdf* tends to be preferred for plotting, but the *cdf* and its inverse are used more heavily in calculations. As with the Φ function and z, the *cdf* gives cumulative probabilities from the lower, left tail of the distribution up to a desired value, while its inverse produces quantiles of the distribution for a given cumulative probability (and the inverse of a *cdf* is therefore a quantile function).

Of the three distributions, t (being symmetrical) is the easiest to work with. Just as for z, either tail can be interrogated to get a particular tail probability. If ν (the *df*) is very large the *cdf* for t will return values identical to z for all practical purposes. Thus $Pr(t_{999} \leq 1.96)$, the cumulative probability of t up to and including 1.96 is .975 (rounding these and subsequent values to 3 decimal places). Likewise, $t_{999,\ .025}$ (the quantile of t with cumulative probability .025 and 999 *df*) is approximately -1.96. Discrepancies between the z and t distribution are apparent for low *df*. For a single sample with $n = 10$, $Pr(t_9 \leq 1.96) = .960$ and $t_{9,\ .025} \approx -2.26$.

The *cdf* and its inverse (the quantile function) can be applied to obtain the tail probability for any quantile, or the quantile (often expressed as a percentage) for any tail probability. For instance, you might need to find the tail probability for an observed t of 2.14 with 29 *df*. Using the symmetry of the distribution this can be obtained directly by looking for the cumulative probability for $t_{29} = -2.14$ which is .020. (The cumulative probability for $t_{29} = 2.14$ is $1 - .020 = .98$.) Alternatively, you may want to find the upper limit of t that covers 90% of the distribution when $df = 29$. To do this you'd want to know the value of t that excludes the most extreme 5% at either end. These values are $t_{29,.05} \approx -1.70$ and $t_{29,\ .95} \approx 1.70$ respectively. So with $df = 29$, 90% of the distribution falls in the range $-1.70 \leq t \leq 1.70$.

Working with χ^2 or F is different. For these distributions interest will almost always focus on one tail of the distribution. Because the distributions are related to the square of z or t, extreme values