Statistics 553: Asymptotic Tools

David R. Hunter

Fall 2011
Penn State University

Contents

Preface

These notes are designed to accompany STAT 553, a graduate-level course in large-sample theory at Penn State intended for students who may not have had any exposure to measure-theoretic probability. While many excellent large-sample theory textbooks already exist, the majority (though not all) of them reflect a traditional view in graduate-level statistics education that students should learn measure-theoretic probability before large-sample theory. The philosophy of these notes is that these priorities are backwards, and that in fact statisticians have more to gain from an understanding of large-sample theory than of measure theory. The intended audience will have had a year-long sequence in mathematical statistics, along with the usual calculus and linear algebra prerequisites that usually accompany such a course, but no measure theory.

Many exercises require students to do some computing, based on the notion that computing skills should be emphasized in *all* statistics courses whenever possible, provided that the computing enhances the understanding of the subject matter. The study of large-sample theory lends itself very well to computing, since frequently the theoretical large-sample results we prove do not give any indication of how well asymptotic approximations work for finite samples. Thus, simulation for the purpose of checking the quality of asymptotic approximations for small samples is very important in understanding the limitations of the results being learned. Of course, all computing activities will force students to choose a particular computing environment. Occasionally, hints are offered in the notes using R (http://www.r-project.org), though these exercises can be completed using other packages or languages, provided that they possess the necessary statistical and graphical capabilities.

Credit where credit is due: These notes originally evolved as an accompaniment to the book *Elements of Large-Sample Theory* by the late Erich Lehmann; the strong influence of that great book, which shares the philosophy of these notes regarding the mathematical level at which an introductory large-sample theory course should be taught, is still very much evident here. I am fortunate to have had the chance to correspond with Professor Lehmann several times about his book, as my students and I provided lists of typographical errors that we had spotted. He was extremely gracious and I treasure the letters that he sent me, written out longhand and sent through the mail even though we were already well into the

era of electronic communication.

I have also drawn on many other sources for ideas or for exercises. Among these are the fantastic and concise *A Course in Large Sample Theory* by Thomas Ferguson, the comprehensive and beautifully written *Asymptotic Statistics* by A. W. van der Vaart, and the classic probability textbooks *Probability and Measure* by Patrick Billingsley and *An Introduction to Probability Theory and Its Applications, Volumes 1 and 2* by William Feller. Arkady Tempelman at Penn State helped with some of the Strong-Law material in Chapter 3, and it was Tom Hettmansperger who originally convinced me to design this course at Penn State back in 2000 when I was a new assistant professor. My goal in doing so was to teach a course that I wished I had had as a graduate student, and I hope that these notes help to achieve that goal.

Chapter 1

Mathematical and Statistical Preliminaries

We assume that many readers are familiar with much of the material presented in this chapter. However, we do not view this material as superfluous, and we feature it prominently as the first chapter of these notes for several reasons. First, some of these topics may have been learned long ago by readers, and a review of this chapter may remind them of knowledge they have forgotten. Second, including these preliminary topics as a separate chapter makes the notes more self-contained than if the topics were omitted: We do not have to refer readers to "a standard calculus textbook" or "a standard mathematical statistics textbook" whenever an advanced result relies on this preliminary material. Third, some of the topics here are likely to be new to some readers, particularly readers who have not taken a course in real analysis.

Fourth, and perhaps most importantly, we wish to set the stage in this chapter for a mathematically rigorous treatment of large-sample theory. By "mathematically rigorous," we do not mean "difficult" or "advanced"; rather, we mean logically sound, relying on arguments in which assumptions and definitions are unambiguously stated and assertions must be provable from these assumptions and definitions. Thus, even well-prepared readers who know the material in this chapter often benefit from reading it and attempting the exercises, particularly if they are new to rigorous mathematics and proof-writing. We strongly caution against the alluring idea of saving time by skipping this chapter when teaching a course, telling students "you can always refer to Chapter 1 when you need to"; we have learned the hard way that this is a dangerous approach that can waste more time in the long run than it saves!

1.1 Limits and Continuity

Fundamental to the study of large-sample theory is the idea of the limit of a sequence. Much of these notes will be devoted to sequences of random variables; however, we begin here by focusing on sequences of real numbers. Technically, a sequence of real numbers is a function from the natural numbers $\{1, 2, 3, \ldots\}$ into the real numbers $\mathbb{R}$; yet we always write $a_1, a_2, \ldots$ instead of the more traditional function notation $a(1), a(2), \ldots$.

We begin by definiting a limit of a sequence of real numbers. This is a concept that will be intuitively clear to readers familiar with calculus. For example, the fact that the sequence $a_1 = 1.3, a_2 = 1.33, a_3 = 1.333, \ldots$ has a limit equal to $4/3$ is unsurprising. Yet there are some subtleties that arise with limits, and for this reason and also to set the stage for a rigorous treatment of the topic, we provide two separate definitions. It is important to remember that even these two definitions do not cover all possible sequences; that is, not every sequence has a well-defined limit.

Definition 1.1 A sequence of real numbers $a_1, a_2, \ldots$ has limit equal to the real number a if for every $\epsilon > 0$, there exists N such that

$$|a_n - a| < \epsilon \text{ for all } n > N.$$

In this case, we write $a_n \to a$ as $n \to \infty$ or $\lim_{n\to\infty} a_n = a$ and we could say that "a_n converges to a".

Definition 1.2 A sequence of real numbers $a_1, a_2, \ldots$ has limit ∞ if for every real number M, there exists N such that

$$a_n > M \text{ for all } n > N.$$

In this case, we write $a_n \to \infty$ as $n \to \infty$ or $\lim_{n\to\infty} a_n = \infty$ and we could say that "a_n diverges to ∞". Similarly, $a_n \to -\infty$ as $n \to \infty$ if for all M, there exists N such that $a_n < M$ for all $n > N$.

Implicit in the language of Definition 1.1 is that N may depend on ϵ. Similarly, N may depend on M (in fact, it *must* depend on M) in Definition 1.2.

The symbols $+\infty$ and $-\infty$ are not considered real numbers; otherwise, Definition 1.1 would be invalid for $a = \infty$ and Definition 1.2 would never be valid since M could be taken to be ∞. Throughout these notes, we will assume that symbols such as a_n and a denote real numbers unless stated otherwise; if situations such as $a = \pm\infty$ are allowed, we will state this fact explicitly.

A crucial fact regarding sequences and limits is that not every sequence has a limit, even when "has a limit" includes the possibilities $\pm\infty$. (However, see Exercise 1.4, which asserts

that every *nondecreasing* sequence has a limit.) A simple example of a sequence without a limit is given in Example 1.3. A common mistake made by students is to "take the limit of both sides" of an equation $a_n = b_n$ or an inequality $a_n \leq b_n$. This is a meaningless operation unless it has been established that such limits exist. On the other hand, an operation that *is* valid is to take the limit superior or limit inferior of both sides, concepts that will be defined in Section 1.1.1. One final word of warning, though: When taking the limit superior of a strict inequality, $<$ or $>$ must be replaced by $\leq$ or $\geq$; see the discussion following Lemma 1.10.

Example 1.3 Define

$$a_n = \log n; \qquad b_n = 1 + (-1)^n/n; \qquad c_n = 1 + (-1)^n/n^2; \qquad d_n = (-1)^n.$$

Then $a_n \to \infty$, $b_n \to 1$, and $c_n \to 1$; but the sequence $d_1, d_2, \ldots$ does not have a limit. (We do not always write "as $n \to \infty$" when this is clear from the context.) Let us prove one of these limit statements, say, $b_n \to 1$. By Definition 1.1, given an arbitrary $\epsilon > 0$, we must prove that there exists some N such that $|b_n - 1| < \epsilon$ whenever $n > N$. Since $|b_n - 1| = 1/n$, we may simply take $N = 1/\epsilon$: With this choice, whenever $n > N$, we have $|b_n - 1| = 1/n < 1/N = \epsilon$, which completes the proof.

We always assume that $\log n$ denotes the natural logarithm, or logarithm base e, of n. This is fairly standard in statistics, though in some other disciplines it is more common to use $\log n$ to denote the logarithm base 10, writing $\ln n$ instead of the natural logarithm. Since the natural logarithm and the logarithm base 10 differ only by a constant ratio—namely, $\log_e n = 2.3026 \log_{10} n$—the difference is often not particularly important. (However, see Exercise 1.27.)

Finally, note that although $\lim_n b_n = \lim_n c_n$ in Example 1.3, there is evidently something different about the manner in which these two sequences approach this limit. This difference will prove important when we study rates of convergence beginning in Section 1.3.

Example 1.4 A very important example of a limit of a sequence is

$$\lim_{n\to\infty} \left(1 + \frac{c}{n}\right)^n = \exp(c)$$

for any real number c. This result is proved in Example 1.20 using l'Hôpital's rule (Theorem 1.19).

Two or more sequences may be added, multiplied, or divided, and the results follow intuitively pleasing rules: The sum (or product) of limits equals the limit of the sums (or products); and as long as division by zero does not occur, the ratio of limits equals the limit

of the ratios. These rules are stated formally as Theorem 1.5, whose complete proof is the subject of Exercise 1.1. To prove only the "limit of sums equals sum of limits" part of the theorem, if we are given $a_n \to a$ and $b_n \to b$ then we need to show that for a given $\epsilon > 0$, there exists N such that for all $n > N$, $|a_n + b_n - (a + b)| < \epsilon$. But the triangle inequality gives

$$|a_n + b_n - (a + b)| \leq |a_n - a| + |b_n - b|, \tag{1.1}$$

and furthermore we know that there must be N_1 and N_2 such that $|a_n - a| < \epsilon/2$ for $n > N_1$ and $|b_n - b| < \epsilon/2$ for $n > N_2$ (since $\epsilon/2$ is, after all, a positive constant and we know $a_n \to a$ and $b_n \to b$). Therefore, we may take $N = \max\{N_1, N_2\}$ and conclude by inequality (1.1) that for all $n > N$,

$$|a_n + b_n - (a + b)| < \frac{\epsilon}{2} + \frac{\epsilon}{2},$$

which proves that $a_n + b_n \to a + b$.

Theorem 1.5 Suppose $a_n \to a$ and $b_n \to b$ as $n \to \infty$. Then $a_n + b_n \to a + b$ and $a_n b_n \to ab$; furthermore, if $b \neq 0$ then $a_n/b_n \to a/b$.

A similar result states that continuous transformations preserve limits; see Theorem 1.16. Theorem 1.5 may be extended by replacing a and/or b by $\pm\infty$, and the results remain true as long as they do not involve the indeterminate forms $\infty - \infty$, $\pm\infty \times 0$, or $\pm\infty/\infty$.

1.1.1 Limit Superior and Limit Inferior

The limit superior and limit inferior of a sequence, unlike the limit itself, are defined for *any* sequence of real numbers. Before considering these important quantities, we must first define supremum and infimum, which are generalizations of the ideas of maximum and minumum. That is, for a set of real numbers that has a minimum, or smallest element, the infimum is equal to this minimum; and similarly for the maximum and supremum. For instance, any finite set contains both a minimum and a maximum. ("Finite" is not the same as "bounded"; the former means having finitely many elements and the latter means contained in an interval neither of whose endpoints are $\pm\infty$.) However, not all sets of real numbers contain a minimum (or maximum) value. As a simple example, take the open interval $(0, 1)$. Since neither 0 nor 1 is contained in this interval, there is no single element of this interval that is smaller (or larger) than all other elements. Yet clearly 0 and 1 are in some sense important in bounding this interval below and above. It turns out that 0 and 1 are the infimum and supremum, respectively, of $(0, 1)$.

An upper bound of a set S of real numbers is (as the name suggests) any value m such that $s \leq m$ for all $s \in S$. A *least upper bound* is an upper bound with the property that no smaller

upper bound exists; that is, m is a least upper bound if m is an upper bound such that for any $\epsilon > 0$, there exists $s \in S$ such that $s > m - \epsilon$. A similar definition applies to *greatest lower bound*. A useful fact about the real numbers—a consequence of the completeness of the real numbers which we do not prove here—is that every set that has an upper (or lower) bound has a least upper (or greatest lower) bound.

Definition 1.6 For any set of real numbers, say S, the supremum $\sup S$ is defined to be the least upper bound of S (or $+\infty$ if no upper bound exists). The infimum $\inf S$ is defined to be the greatest lower bound of S (or $-\infty$ if no lower bound exists).

Example 1.7 Let $S = \{a_1, a_2, a_3, \ldots\}$, where $a_n = 1/n$. Then $\inf S$, which may also be denoted $\inf_n a_n$, equals 0 even though $0 \notin S$. But $\sup_n a_n = 1$, which is contained in S. In this example, $\max S = 1$ but $\min S$ is undefined.

If we denote by $\sup_{k \geq n} a_k$ the supremum of $\{a_n, a_{n+1}, \ldots\}$, then we see that this supremum is taken over a smaller and smaller set as n increases. Therefore, $\sup_{k \geq n} a_k$ is a nonincreasing sequence in n, which implies that it has a limit as $n \to \infty$ (see Exercise 1.4). Similarly, $\inf_{k \geq n} a_k$ is a nondecreasing sequence, which implies that it has a limit.

Definition 1.8 The limit superior of a sequence $a_1, a_2, \ldots$, denoted $\limsup_n a_n$ or sometimes $\overline{\lim}_n a_n$, is the limit of the nonincreasing sequence

$$\sup_{k \geq 1} a_k, \quad \sup_{k \geq 2} a_k, \quad \ldots \quad .$$

The limit inferior, denoted $\liminf_n a_n$ or sometimes $\underline{\lim}_n a_n$, is the limit of the nondecreasing sequence

$$\inf_{k \geq 1} a_k, \inf_{k \geq 2} a_k, \ldots .$$

Intuitively, the limit superior and limit inferior may be understood as follows: If we define a limit point of a sequence to be any number which is the limit of some subsequence, then lim inf and lim sup are the smallest and largest limit points, respectively (more precisely, they are the infimum and supremum, respectively, of the set of limit points).

Example 1.9 In Example 1.3, the sequence $d_n = (-1)^n$ does not have a limit. However, since $\sup_{k \geq n} d_k = 1$ and $\inf_{k \leq n} d_k = -1$ for all n, it follows that

$$\limsup_n d_n = 1 \quad \text{and} \quad \liminf_n d_n = -1.$$

In this example, the set of limit points of the sequence $d_1, d_2, \ldots$ is simply $\{-1, 1\}$.

Here are some useful facts regarding limits superior and inferior:

Lemma 1.10 Let $a_1, a_2, \ldots$ and $b_1, b_2, \ldots$ be arbitrary sequences of real numbers.

- $\limsup_n a_n$ and $\liminf_n a_n$ always exist, unlike $\lim_n a_n$.
- $\liminf_n a_n \leq \limsup_n a_n$
- $\lim_n a_n$ exists if and only if $\liminf_n a_n = \limsup_n a_n$, in which case

$$\lim_n a_n = \liminf_n a_n = \limsup_n a_n.$$

- Both lim sup and lim inf preserve nonstrict inequalities; that is, if $a_n \leq b_n$ for all n, then $\limsup_n a_n \leq \limsup_n b_n$ and $\liminf_n a_n \leq \liminf_n b_n$.
- $\limsup_n(-a_n) = -\liminf_n a_n$.

The next-to-last claim in Lemma 1.10 is no longer true if "nonstrict inequalities" is replaced by "strict inequalities". For instance, $1/(n+1) < 1/n$ is true for all positive n, but the limit superior of each side equals zero. Thus, it is *not* true that

$$\limsup_n \frac{1}{n+1} < \limsup_n \frac{1}{n}.$$

We must replace $<$ by $\leq$ (or $>$ by $\geq$) when taking the limit superior or limit inferior of both sides of an inequality.

1.1.2 Continuity

Although Definitions 1.1 and 1.2 concern limits, they apply only to sequences of real numbers. Recall that a sequence is a real-valued function of the *natural* numbers. We shall also require the concept of a limit of a real-valued function of a *real* variable. To this end, we make the following definition.

Definition 1.11 For a real-valued function $f(x)$ defined for all points in a neighborhood of x_0 except possibly x_0 itself, we call the real number a the limit of $f(x)$ as x goes to x_0, written

$$\lim_{x \to x_0} f(x) = a,$$

if for each $\epsilon > 0$ there is a $\delta > 0$ such that $|f(x) - a| < \epsilon$ whenever $0 < |x - x_0| < \delta$.

First, note that Definition 1.11 is sensible only if both x_0 and a are finite (but see Definition 1.13 for the case in which one or both of them is $\pm\infty$). Furthermore, it is very important to remember that $0 < |x - x_0| < \delta$ may *not* be replaced by $|x - x_0| < \delta$: The latter would

imply something specific about the value of $f(x_0)$ itself, whereas the correct definition does not even require that this value be defined. In fact, by merely replacing $0 < |x - x_0| < \delta$ by $|x - x_0| < \delta$ (and insisting that $f(x_0)$ be defined), we could take Definition 1.11 to be the definition of *continuity* of $f(x)$ at the point x_0 (see Definition 1.14 for an equivalent formuation).

Implicit in Definition 1.11 is the fact that a is the limiting value of $f(x)$ no matter whether x approaches x_0 from above or below; thus, $f(x)$ has a two-sided limit at x_0. We may also consider one-sided limits:

Definition 1.12 The value a is called the right-handed limit of $f(x)$ as x goes to x_0, written

$$\lim_{x \to x_0+} f(x) = a \quad \text{or} \quad f(x_0+) = a,$$

if for each $\epsilon > 0$ there is a $\delta > 0$ such that $|f(x) - a| < \epsilon$ whenever $0 < x - x_0 < \delta$.

The left-handed limit, $\lim_{x \to x_0-} f(x)$ or $f(x_0-)$, is defined analagously: $f(x_0-) = a$ if for each $\epsilon > 0$ there is a $\delta > 0$ such that $|f(x) - a| < \epsilon$ whenever $-\delta < x - x_0 < 0$.

The preceding definitions imply that

$$\lim_{x \to x_0} f(x) = a \quad \text{if and only if} \quad f(x_0+) = f(x_0-) = a; \tag{1.2}$$

in other words, the (two-sided) limit exists if and only if both one-sided limits exist and they coincide. Before using the concept of a limit to define continuity, we conclude the discussion of limits by addressing the possibilities that $f(x)$ has a limit as $x \to \pm\infty$ or that $f(x)$ tends to $\pm\infty$:

Definition 1.13 Definition 1.11 may be expanded to allow x_0 or a to be infinite:

(a) We write $\lim_{x \to \infty} f(x) = a$ if for every $\epsilon > 0$, there exists N such that $|f(x) - a| < \epsilon$ for all $x > N$.

(b) We write $\lim_{x \to x_0} f(x) = \infty$ if for every M, there exists $\delta > 0$ such that $f(x) > M$ whenever $0 < |x - x_0| < \delta$.

(c) We write $\lim_{x \to \infty} f(x) = \infty$ if for every M, there exists N such that $f(x) > M$ for all $x > N$.

Definitions involving $-\infty$ are analogous, as are definitions of $f(x_0+) = \pm\infty$ and $f(x_0-) = \pm\infty$.

As mentioned above, the value of $f(x_0)$ in Definitions 1.11 and 1.12 is completely irrelevant; in fact, $f(x_0)$ might not even be defined. In the special case that $f(x_0)$ is defined and equal to a, then we say that $f(x)$ is *continuous* (or right- or left-continuous) at x_0, as summarized by Definition 1.14 below. Intuitively, $f(x)$ is continuous at x_0 if it is possible to draw the graph of $f(x)$ through the point $[x_0, f(x_0)]$ without lifting the pencil from the page.

Definition 1.14 If $f(x)$ is a real-valued function and x_0 is a real number, then

- we say $f(x)$ is continuous at x_0 if $\lim_{x \to x_0} f(x) = f(x_0)$;
- we say $f(x)$ is right-continuous at x_0 if $\lim_{x \to x_0+} f(x) = f(x_0)$;
- we say $f(x)$ is left-continuous at x_0 if $\lim_{x \to x_0-} f(x) = f(x_0)$.

Finally, even though continuity is inherently a *local* property of a function (since Definition 1.14 applies only to the particular point x_0), we often speak *globally* of "a continuous function," by which we mean a function that is continuous at every point in its domain.

Statement (1.2) implies that every (globally) continuous function is right-continuous. However, the converse is not true, and in statistics the canonical example of a function that is right-continuous but not continuous is the cumulative distribution function for a discrete random variable.

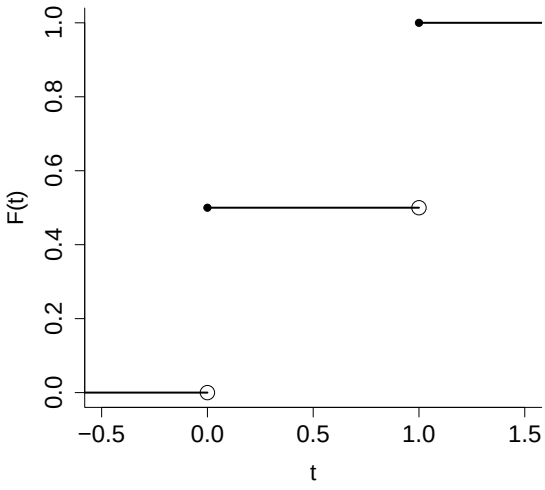

Figure 1.1: *The cumulative distribution function for a Bernoulli* (1/2) *random variable is discontinuous at the points* $t = 0$ *and* $t = 1$*, but it is everywhere right-continuous.*

Example 1.15 Let X be a Bernoulli $(1/2)$ random variable, so that the events $X = 0$ and $X = 1$ each occur with probability $1/2$. Then the distribution function $F(t) = P(X \leq t)$ is right-continuous but it is not continuous because it has "jumps" at $t = 0$ and $t = 1$ (see Figure 1.1). Using one-sided limit notation of Definition 1.12, we may write

$$0 = F(0-) \neq F(0+) = 1/2 \quad \text{and} \quad 1/2 = F(1-) \neq F(1+) = 1.$$

Although $F(t)$ is not (globally) continuous, it is continuous at every point in the set $\mathbb{R} \setminus \{0, 1\}$ that does not include the points 0 and 1.

We conclude with a simple yet important result relating continuity to the notion of the limit of a sequence. Intuitively, this result states that continuous functions preserve limits of sequences.

Theorem 1.16 If a is a real number such that $a_n \to a$ as $n \to \infty$ and the real-valued function $f(x)$ is continuous at the point a, then $f(a_n) \to f(a)$.

Proof: We need to show that for any $\epsilon > 0$, there exists N such that $|f(a_n) - f(a)| < \epsilon$ for all $n > N$. To this end, let $\epsilon > 0$ be a fixed arbitrary constant. From the definition of continuity, we know that there exists some $\delta > 0$ such that $|f(x) - f(a)| < \epsilon$ for all x such that $|x - a| < \delta$. Since we are told $a_n \to a$ and since $\delta > 0$, there must by definition be some N such that $|a_n - a| < \delta$ for all $n > N$. We conclude that for all n greater than this particular N, $|f(a_n) - f(a)| < \epsilon$. Since ϵ was arbitrary, the proof is finished. ■

Exercises for Section 1.1

Exercise 1.1 Assume that $a_n \to a$ and $b_n \to b$, where a and b are real numbers.

(a) Prove that $a_n b_n \to ab$

Hint: Show that $|a_n b_n - ab| \leq |(a_n - a)(b_n - b)| + |a(b_n - b)| + |b(a_n - a)|$ using the triangle inequality.

(b) Prove that if $b \neq 0$, $a_n / b_n \to a/b$.

Exercise 1.2 For a fixed real number c, define $a_n(c) = (1 + c/n)^n$. Then Equation (1.9) states that $a_n(c) \to \exp(c)$. A different sequence with the same limit is obtained from the power series expansion of $\exp(c)$:

$$b_n(c) = \sum_{i=0}^{n-1} \frac{c^i}{i!}$$

For each of the values $c \in \{-10, -1, 0.2, 1, 5\}$, find the smallest value of n such that $|a_n(c) - \exp(c)|/\exp(c) < .01$. Now replace $a_n(c)$ by $b_n(c)$ and repeat. Comment on any general differences you observe between the two sequences.

Exercise 1.3 **(a)** Suppose that $a_k \to c$ as $k \to \infty$ for a sequence of real numbers $a_1, a_2, \ldots$. Prove that this implies convergence in the sense of Cesáro, which means that

$$\frac{1}{n}\sum_{k=1}^{n} a_k \to c \qquad \text{as } n \to \infty. \tag{1.3}$$

In this case, c may be real or it may be $\pm\infty$.

Hint: If c is real, consider the definition of $a_k \to c$: There exists N such that $|a_k - c| < \epsilon$ for all $k > N$. Consider what happens when the sum in expression (1.3) is broken into two sums, one for $k \le N$ and one for $k > N$. The case $c = \pm\infty$ follows a similar line of reasoning.

(b) Is the converse true? In other words, does (1.3) imply $a_k \to c$?

Exercise 1.4 Prove that if $a_1, a_2, \ldots$ is a nondecreasing (or nonincreasing) sequence, then $\lim_n a_n$ exists and is equal to $\sup_n a_n$ (or $\inf_n a_n$). We allow the possibility $\sup_n a_n = \infty$ (or $\inf_n a_n = -\infty$) here.

Hint: For the case in which $\sup_n a_n$ is finite, use the fact that the least upper bound M of a set S is defined by the fact that $s \le M$ for all $s \in S$, but for any $\epsilon > 0$ there exists $s \in S$ such that $s > M - \epsilon$.

Exercise 1.5 Let $a_n = \sin n$ for $n = 1, 2, \ldots$.

(a) What is $\sup_n a_n$? Does $\max_n a_n$ exist?

(b) What is the set of limit points of $\{a_1, a_2, \ldots\}$? What are $\limsup_n a_n$ and $\liminf_n a_n$? (Recall that a limit point is any point that is the limit of a subsequence $a_{k_1}, a_{k_2}, \ldots$, where $k_1 < k_2 < \cdots$.)

(c) As usual in mathematics, we assume above that angles are measured in radians. How do the answers to (a) and (b) change if we use degrees instead (i.e., $a_n = \sin n^\circ$)?

Exercise 1.6 Prove Lemma 1.10.

Exercise 1.7 For $x \notin \{0, 1, 2\}$, define

$$f(x) = \frac{|x^3 - x|}{x(x-1)(x-2)}.$$

(a) Graph $f(x)$. Experiment with various ranges on the axes until you attain a visually pleasing and informative plot that gives a sense of the overall behavior of the function.

(b) For each of $x_0 \in \{-1, 0, 1, 2\}$, answer these questions: Is $f(x)$ continuous at x_0, and if not, could $f(x_0)$ be defined so as to make the answer yes? What are the right- and left-hand limits of $f(x)$ at x_0? Does it have a limit at x_0? Finally, what are $\lim_{x\to\infty} f(x)$ and $\lim_{x\to-\infty} f(x)$?

Exercise 1.8 Define $F(t)$ as in Example 1.15 (and as pictured in Figure 1.1). This function is not continuous, so Theorem 1.16 does not apply. That is, $a_n \to a$ does not imply that $F(a_n) \to F(a)$.

(a) Give an example of a sequence $\{a_n\}$ and a real number a such that $a_n \to a$ but $\limsup_n F(a_n) \neq F(a)$.

(b) Change your answer to part (a) so that $a_n \to a$ and $\limsup_n F(a_n) = F(a)$, but $\lim_n F(a_n)$ does not exist.

(c) Explain why it is not possible to change your answer so that $a_n \to a$ and $\liminf_n F(a_n) = F(a)$, but $\lim_n F(a_n)$ does not exist.

1.2 Differentiability and Taylor's Theorem

Differential calculus plays a fundamental role in much asymptotic theory. In this section we review simple derivatives and one form of Taylor's well-known theorem. Approximations to functions based on Taylor's Theorem, often called Taylor expansions, are ubiquitous in large-sample theory.

We assume that readers are familiar with the definition of a derivative of a real-valued function $f(x)$:

Definition 1.17 If $f(x)$ is continuous in a neighborhood of x_0 and

$$\lim_{x\to x_0} \frac{f(x) - f(x_0)}{x - x_0} \tag{1.4}$$

exists, then $f(x)$ is said to be differentiable at x_0 and the limit (1.4) is called the derivative of $f(x)$ at x_0 and is denoted by $f'(x_0)$ or $f^{(1)}(x_0)$.

We use the standard notation for second- and higher-order derivatives. Thus, if $f'(x)$ is itself differentiable at x_0, we express its derivative as $f''(x_0)$ or $f^{(2)}(x_0)$. In general, if the kth derivative $f^{(k)}(x)$ is differentiable at x_0, then we denote this derivative by $f^{(k+1)}(x_0)$. We

also write $(d^k/dx^k)f(x)$ (omitting the k when $k = 1$) to denote the function $f^{(k)}(x)$, and to denote the evaluation of this function at a specific point (say x_0), we may use the following notation, which is equivalent to $f^{(k)}(x_0)$:

$$\frac{d^k}{dx^k} f(x) \bigg|_{x=x_0}$$

In large-sample theory, differential calculus is most commonly applied in the construction of Taylor expansions. There are several different versions of Taylor's Theorem, distinguished from one another by the way in which the remainder term is expressed. The first form we present here (Theorem 1.18), which is proved in Exercise 1.11, does not state an explicit form for the remainder term. This gives it the advantage that it does not require that the function have an extra derivative. For instance, a second-order Taylor expansion requires only two derivatives using this version of Taylor's Theorem (and the second derivative need only exist at a single point), whereas other forms of Taylor's Theorem require the existence of a third derivative over an entire interval. The disadvantage of this form of Taylor's Theorem is that we do not get any sense of what the remainder term is, only that it goes to zero; however, for many applications in these notes, this form of Taylor's Theorem will suffice.

Theorem 1.18 If $f(x)$ has d derivatives at a, then

$$f(x) \;=\; f(a) + (x-a)f'(a) + \cdots + \frac{(x-a)^d}{d!} f^{(d)}(a) + r_d(x,a), \qquad (1.5)$$

where $r_d(x,a)/(x-a)^d \to 0$ as $x \to a$.

In some cases, we will find it helpful to have an explicit form of $r_d(x,a)$. This is possible under stronger assumptions, namely, that $f(x)$ has $d+1$ derivatives on the closed interval from x to a. In this case, we may write

$$r_d(x,a) = \int_a^x \frac{(x-t)^d}{d!} f^{(d+1)}(t)\, dt \qquad (1.6)$$

in equation (1.5). Equation (1.6) is often called the Lagrange form of the remainder. By the Mean Value Theorem of calculus, there exists x^* somewhere in the closed interval from x to a such that

$$r_d(x,a) = \frac{(x-a)^{d+1}}{(d+1)!} f^{(d+1)}(x^*). \qquad (1.7)$$

Expression (1.7), since it follows immediately from Equation (1.6), is also referred to as the Lagrange form of the remainder.

To conclude this section, we state the well-known calculus result known as l'Hôpital's Rule. This useful Theorem provides an elegant way to prove Theorem 1.18, among other things.

Theorem 1.19 *l'Hôpital's Rule:* For a real number c, suppose that $f(x)$ and $g(x)$ are differentiable for all points in a neighborhood containing c except possibly c itself. If $\lim_{x\to c} f(x) = 0$ and $\lim_{x\to c} g(x) = 0$, then

$$\lim_{x\to c} \frac{f(x)}{g(x)} = \lim_{x\to c} \frac{f'(x)}{g'(x)}, \tag{1.8}$$

provided the right-hand limit exists. Similarly, if $\lim_{x\to c} f(x) = \infty$ and $\lim_{x\to c} g(x) = \infty$, then Equation (1.8) also holds. Finally, the theorem also applies if $c = \pm\infty$, in which case a "neighborhood containing c" refers to an interval (a, ∞) or $(-\infty, a)$.

Example 1.20 Example 1.4 states that

$$\lim_{n\to\infty} \left(1 + \frac{c}{n}\right)^n = \exp(c) \tag{1.9}$$

for any real number c. Let us prove this fact using l'Hôpital's Rule. Care is necessary in this proof, since l'Hôpital's Rule applies to limits of differentiable functions, whereas the left side of Equation (1.9) is a function of an *integer-valued* n.

Taking logarithms in Equation (1.9), we shall first establish that $n\log(1+c/n) \to c$ as $n \to \infty$. Define $f(x) = \log(1 + cx)$ and $g(x) = x$. The strategy is to treat n as $1/x$, so we will see what happens to $f(x)/g(x)$ as $x \to 0$. By l'Hôpital's Rule, we obtain

$$\lim_{x\to 0} \frac{\log(1 + cx)}{x} = \lim_{x\to 0} \frac{c/(1 + cx)}{1} = c.$$

Since this limit must be valid no matter how x approaches 0, in particular we may conclude that if we define $x_n = 1/n$ for $n = 1, 2, \ldots$, then

$$\lim_{n\to\infty} \frac{\log(1 + cx_n)}{x_n} = \lim_{n\to\infty} n\log\left(1 + \frac{c}{n}\right) = c, \tag{1.10}$$

which was to be proved. Now we use the fact that the exponential function $h(t) = \exp t$ is a continuous function, so Equation (1.9) follows from Theorem 1.16 once we apply the exponential function to Equation (1.10).

Exercises for Section 1.2

Exercise 1.9 The well-known derivative of the polynomial function $f(x) = x^n$ for a positive integer n is given by nx^{n-1}. Prove this fact directly using Definition 1.17.

Exercise 1.10 For $f(x)$ continuous in a neighborhood of x_0, consider

$$\lim_{x \to x_0} \frac{f(x) - f(2x_0 - x)}{2(x - x_0)}. \tag{1.11}$$

(a) Prove or give a counterexample: When $f'(x_0)$ exists, limit (1.11) also exists and it is equal to $f'(x_0)$.

(b) Prove or give a counterexample: When limit (1.11) exists, it equals $f'(x_0)$, which also exists.

Exercise 1.11 Prove Theorem 1.18.

Hint: Let $P_d(x)$ denote the Taylor polynomial such that

$$r_d(x, a) = f(x) - P_d(x).$$

Then use l'Hôpital's rule, Theorem 1.19, $d - 1$ times. (You can do this because the existence of $f^{(d)}(a)$ implies that all lower-order derivatives exist on an interval containing a.) You cannot use l'Hôpital's rule d times, but you won't need to if you use Definition 1.17.

Exercise 1.12 Let $f(t) = \log t$. Taking $a = 1$ and $x = a + h$, find the explicit remainder term $r_d(x, a)$ in Equation (1.5) for all values of $d \in \{2, 3\}$ and $h \in \{0.1, 0.01, 0.001\}$. Give your results in a table. How does $r_d(x, a)$ appear to vary with d? How does $r_d(a + h, a)$ appear to vary with h?

Exercise 1.13 The idea for Exercise 1.10 is based on a numerical trick for accurately approximating the derivative of a function that can be evaluated directly but for which no formula for the derivative is known.

(a) First, construct a "first-order" approximation to a derivative. Definition 1.17 with $d = 1$ suggests that we may choose a small h and obtain

$$f'(a) \approx \frac{f(a + h) - f(a)}{h}. \tag{1.12}$$

For $f(x) = \log x$ and $a = 2$, calculate the approximation to $f'(a)$ in Equation (1.12) using $h \in \{0.5, 0.05, 0.005\}$. How does the difference between the true value (which you happen to know in this case) and the approximation appear to vary as a function of h?

(b) Next, expand both $f(a + h)$ and $f(a - h)$ using Taylor's theorem with $d = 2$. Subtract one expansion from the other and solve for $f'(a)$. Ignore the

remainder terms and you have a "second-order" approximation. (Compare this approximation with Exercise 1.10, substituting x_0 and $x - x_0$ for a and h.) Repeat the computations of part (a). Now how does the error appear to vary as a function of h?

(c) Finally, construct a "fourth-order" approximation. Perform Taylor expansions of $f(x + 2h)$, $f(x + h)$, $f(x - h)$, and $f(x - 2h)$ with $d = 4$. Ignore the remainder terms, then find constants C_1 and C_2 such that the second, third, and fourth derivatives all disappear and you obtain

$$f'(a) \approx \frac{C_1 [f(a+h) - f(a-h)] + C_2 [f(a+2h) - f(a-2h)]}{h}. \tag{1.13}$$

Repeat the computations of parts (a) and (b) using the approximation in Equation (1.13).

Exercise 1.14 The gamma function $\Gamma(x)$ is defined for positive real x as

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t}\, dt \tag{1.14}$$

[in fact, equation (1.14) is also valid for complex x with positive real part]. The gamma function may be viewed as a continuous version of the factorial function in the sense that $\Gamma(n) = (n-1)!$ for all positive integers n. The gamma function satisfies the identity

$$\Gamma(x+1) = x\Gamma(x) \tag{1.15}$$

even for noninteger positive values of x. Since $\Gamma(x)$ grows very quickly as x increases, it is often convenient in numerical calculations to deal with the logarithm of the gamma function, which we term the log-gamma function. The *digamma function* $\Psi(x)$ is defined to be the derivative of the log-gamma function; this function often arises in statistical calculations involving certain distributions that use the gamma function.

(a) Apply the result of Exercise 1.13(b) using $h = 1$ to demonstrate how to obtain the approximation

$$\Psi(x) \approx \frac{1}{2} \log [x(x-1)] \tag{1.16}$$

for $x > 2$.

Hint: Use Identity (1.15).

(b) Test Approximation (1.16) numerically for all x in the interval $(2, 100)$ by plotting the ratio of the approximation to the true $\Psi(x)$. What do you notice about the quality of the approximation? If you are using R or Splus, then `digamma(x)` gives the value of $\Psi(x)$.

Exercise 1.15 The second derivative of the log-gamma function is called the trigamma function:

$$\Psi'(x) = \frac{d^2}{dx^2} \log \Gamma(x). \tag{1.17}$$

Like the digamma function, it often arises in statistical calculations; for example, see Exercise 1.35.

(a) Using the method of Exercise 1.13(c) with $h = 1$ [that is, expanding $f(x+2h)$, $f(x+h)$, $f(x-h)$, and $f(x-2h)$ and then finding a linear combination that makes all but the *second* derivative of the log-gamma function disappear], show how to derive the following approximation to $\Psi'(x)$ for $x > 2$:

$$\Psi'(x) \approx \frac{1}{12} \log \left[\left(\frac{x}{x-1} \right)^{15} \left(\frac{x-2}{x+1} \right) \right]. \tag{1.18}$$

(b) Test Approximation (1.18) numerically as in Exercise 1.14(b). In R or Splus, `trigamma(x)` gives the value of $\Psi'(x)$.

1.3 Order Notation

As we saw in Example 1.3, the limiting behavior of a sequence is not fully characterized by the value of its limit alone, if the limit exists. In that example, both $1 + (-1)^n/n$ and $1 + (-1)^n/n^2$ converge to the same limit, but they approach this limit at different rates. In this section we consider not only the value of the limit, but the rate at which that limit is approached. In so doing, we present some convenient notation for comparing the limiting behavior of different sequences.

Definition 1.21 We say that the sequence of real numbers $a_1, a_2, \ldots$ is asymptotically equivalent to the sequence $b_1, b_2, \ldots$, written $a_n \sim b_n$, if $(a_n/b_n) \to 1$ as $n \to \infty$.

Equivalently, $a_n \sim b_n$ if and only if

$$\left| \frac{a_n - b_n}{a_n} \right| \to 0.$$

The expression $|(a_n - b_n)/a_n|$ above is called the relative error in approximating a_n by b_n.

The definition of asymptotic equivalence does *not* say that

$$\frac{\lim a_n}{\lim b_n} = 1;$$

the above fraction might equal $0/0$ or ∞/∞, or the limits might not even exist! (See Exercise 1.17.)

Example 1.22 A well-known asymptotic equivalence is Stirling's formula, which states

$$n! \sim \sqrt{2\pi} n^{n+.5} \exp(-n). \tag{1.19}$$

There are multiple ways to prove Stirling's formula. We outline one proof, based on the Poisson distribution, in Exercise 4.5.

Example 1.23 For any $k > -1$,

$$\sum_{i=1}^{n} i^k \sim \frac{n^{k+1}}{k+1}. \tag{1.20}$$

This is proved in Exercise 1.19. But what about the case $k = -1$? Let us prove that

$$\sum_{i=1}^{n} \frac{1}{i} \sim \log n. \tag{1.21}$$

Proof: Since $1/x$ is a strictly decreasing function of x, we conclude that

$$\int_i^{i+1} \frac{1}{x}\, dx < \frac{1}{i} < \int_{i-1}^{i} \frac{1}{x}\, dx$$

for $i = 2, 3, 4, \ldots$. Summing on i (and using $1/i = 1$ for $i = 1$) gives

$$1 + \int_2^{n+1} \frac{1}{x}\, dx < \sum_{i=1}^{n} \frac{1}{i} < 1 + \int_1^{n} \frac{1}{x}\, dx.$$

Evaluating the integrals and dividing through by $\log n$ gives

$$\frac{1 + \log(n+1) - \log 2}{\log n} < \frac{\sum_{i=1}^{n} \frac{1}{i}}{\log n} < \frac{1}{\log n} + 1.$$

The left and right sides of this expression have limits, both equal to 1 (do you see why?). A standard trick is therefore to take the limit inferior of the left inequality

and combine this with the limit superior of the right inequality (remember to change $<$ to $\leq$ when doing this; see the discussion following Lemma 1.10) to obtain

$$1 \leq \liminf_n \frac{\sum_{i=1}^n \frac{1}{i}}{\log n} \leq \limsup_n \frac{\sum_{i=1}^n \frac{1}{i}}{\log n} \leq 1.$$

This implies that the limit inferior and limit superior are in fact the same, so the limit exists and is equal to 1. This is what we wished to show. ∎

The next notation we introduce expresses the idea that one sequence is asymptotically negligible compared to another sequence.

Definition 1.24 We write $a_n = o(b_n)$ ("a_n is little-o of b_n") as $n \to \infty$ if $a_n/b_n \to 0$ as $n \to \infty$.

Among other advantages, the o-notation makes it possible to focus on the most important terms of a sequence while ignoring the terms that are comparatively negligible.

Example 1.25 According to Definition 1.24, we may write

$$\frac{1}{n} - \frac{2}{n^2} + \frac{4}{n^3} = \frac{1}{n} + o\left(\frac{1}{n}\right) \qquad \text{as } n \to \infty.$$

This makes it clear at a glance how fast the sequence on the left tends to zero, since all terms other than the dominant term are lumped together as $o(1/n)$.

Some of the exercises in this section require proving that one sequence is little-o of another sequence. Sometimes, l'Hôpital's rule may be helpful; yet as in Example 1.20, care must be exercised because l'Hôpital's rule applies to functions of real numbers whereas a sequence is a function of the positive integers.

Example 1.26 Let us prove that $\log\log n = o(\log n)$. The function $(\log\log x)/\log x$, defined for $x > 1$, agrees with $(\log\log n)/\log n$ on the positive integers; thus, since l'Hôpital's rule implies

$$\lim_{x\to\infty} \frac{\log\log x}{\log x} = \lim_{x\to\infty} \frac{\frac{1}{x\log x}}{\frac{1}{x}} = \lim_{x\to\infty} \frac{1}{\log x} = 0,$$

we conclude that $(\log\log n)/\log n$ must also tend to 0 as n tends to ∞ as an integer.

Often, however, one may simply prove $a_n = o(b_n)$ without resorting to l'Hôpital's rule, as in the next example.

Example 1.27 Prove that

$$n = o\left(\sum_{i=1}^{n} \sqrt{i}\right). \tag{1.22}$$

Proof: Letting $\lfloor n/2 \rfloor$ denote the largest integer less than or equal to $n/2$,

$$\sum_{i=1}^{n} \sqrt{i} \;\geq\; \sum_{i=\lfloor n/2 \rfloor}^{n} \sqrt{i} \;\geq\; \frac{n}{2}\sqrt{\left\lfloor \frac{n}{2} \right\rfloor}.$$

Since $n = o(n\sqrt{n})$, the desired result follows. ∎

Equation (1.22) could have been proved using the result of Example 1.23, in which Equation (1.20) with $k = 1/2$ implies that

$$\sum_{i=1}^{n} \sqrt{i} \sim \frac{2n^{3/2}}{3}. \tag{1.23}$$

However, we urge extreme caution when using asymptotic equivalences like Expression (1.23). It is tempting to believe that expressions that are asymptotically equivalent may be substituted for one another under any circumstances, and this is *not* true! In this particular example, we may write

$$\frac{n}{\sum_{i=1}^{n} \sqrt{i}} = \left(\frac{3}{2\sqrt{n}}\right)\left(\frac{2n^{3/2}}{3\sum_{i=1}^{n} \sqrt{i}}\right),$$

and because we know that the second fraction in parentheses tends to 1 by Expression (1.23) and the first fraction in parentheses tends to 0, we conclude that the product of the two converges to 0 and Equation (1.22) is proved.

We define one additional order notation, the capital O.

Definition 1.28 We write $a_n = O(b_n)$ ("a_n is big-o of b_n") as $n \to \infty$ if there exist $M > 0$ and $N > 0$ such that $|a_n/b_n| < M$ for all $n > N$.

In particular, $a_n = o(b_n)$ implies $a_n = O(b_n)$. In a vague sense, o and O relate to sequences as $<$ and $\leq$ relate to real numbers. However, this analogy is not perfect: For example, note that it is not always true that either $a_n = O(b_n)$ or $b_n = O(a_n)$.

Although the notation above is very precisely defined, unfortunately this is not the case with the *language* used to describe the notation. In particular, "a_n is of order b_n" is ambiguous; it may mean simply that $a_n = O(b_n)$, or it may mean something more precise: Some authors define $a_n \asymp b_n$ or $a_n = \Theta(b_n)$ to mean that $|a_n|$ remains bounded between $m|b_n|$ and $M|b_n|$

for large enough n for some constants $0 < m < M$. Although the language can be imprecise, it is usually clear from context what the speaker's intent is.

This latter case, where $a_n = O(b_n)$ but $a_n \neq o(b_n)$, is one in which the ratio $|a_n/b_n|$ remains bounded and also bounded away from zero: There exist positive constants m and M, and an integer N, such that

$$m < \left|\frac{a_n}{b_n}\right| < M \quad \text{for all } n > N. \tag{1.24}$$

Some books introduce a special symbol for (1.24), such as $a_n \asymp b_n$ or $a_n = \Theta(b_n)$.

Do not forget that the use of o, O, or $\sim$ always implies that there is some sort of limit being taken. Often, an expression involves n, in which case we usually assume n tends to ∞ even if this is not stated; however, sometimes things are not so clear, so it helps to be explicit:

Example 1.29 According to Definition 1.24, a sequence that is $o(1)$ tends to zero. Therefore, Equation (1.5) of Taylor's Theorem may be rewritten

$$f(x) \;=\; f(a) + (x-a)f'(a) + \cdots + \frac{(x-a)^d}{d!}\left\{f^{(d)}(a) + o(1)\right\} \quad \text{as } x \to a.$$

It is important to write "as $x \to a$" in this case.

It is often tempting, when faced with an equation such as $a_n = o(b_n)$, to attempt to apply a function $f(x)$ to each side and claim that $f(a_n) = o[f(b_n)]$. Unfortunately, however, this is not true in general and it is not hard to find a counterexample [see Exercise 1.18(d)]. There are certain circumstances in which it *is* possible to claim that $f(a_n) = o[f(b_n)]$, and one such circumstance is particularly helpful. It involves a convex function $f(x)$, defined as follows:

Definition 1.30 We say that a function $f(x)$ is convex if for all x, y and any $\alpha \in [0,1]$, we have

$$f[\alpha x + (1-\alpha)y] \leq \alpha f(x) + (1-\alpha)f(y). \tag{1.25}$$

If $f(x)$ is everywhere differentiable and $f''(x) > 0$ for all x, then $f(x)$ is convex (this is proven in Exercise 1.24). For instance, the function $f(x) = \exp(x)$ is convex because its second derivative is always positive.

We now see a general case in which it may be shown that $f(a_n) = o[f(b_n)]$.

Theorem 1.31 Suppose that $a_1, a_2, \ldots$ and $b_1, b_2, \ldots$ are sequences of real numbers such that $a_n \to \infty$, $b_n \to \infty$, and $a_n = o(b_n)$; and $f(x)$ is a convex function such that $f(x) \to \infty$ as $x \to \infty$. Then $f(a_n) = o[f(b_n)]$.

The proof of Theorem 1.31 is the subject of Exercise 1.25.

There are certain rates of growth toward ∞ that are so common that they have names, such as logarithmic, polynomial, and exponential growth. If α, β, and γ are arbitrary positive constants, then the sequences $(\log n)^\alpha$, n^β, and $(1+\gamma)^n$ exhibit logarithmic, polynomial, and exponential growth, respectively. Furthermore, we always have

$$(\log n)^\alpha = o(n^\beta) \quad \text{and} \quad n^\beta = o([1+\gamma]^n). \tag{1.26}$$

Thus, in the sense of Definition 1.24, logarithmic growth is always slower than polynomial growth and polynomial growth is always slower than exponential growth.

To prove Statement (1.26), first note that $\log \log n = o(\log n)$, as shown in Example 1.26. Therefore, $\alpha \log \log n = o(\beta \log n)$ for arbitrary positive constants α and β. Since $\exp(x)$ is a convex function, Theorem 1.31 gives

$$(\log n)^\alpha = o(n^\beta). \tag{1.27}$$

As a special case of Equation (1.27), we obtain $\log n = o(n)$, which immediately gives $\beta \log n = o[n \log(1+\gamma)]$ for arbitrary positive constants β and γ. Exponentiating once again and using Theorem 1.31 yields

$$n^\beta = o[(1+\gamma)^n].$$

Exercises for Section 1.3

Exercise 1.16 Prove that $a_n \sim b_n$ if and only if $|(a_n - b_n)/a_n| \to 0$.

Exercise 1.17 For each of the following statements, prove the statement or provide a counterexample that disproves it.

(a) If $a_n \sim b_n$, then $\lim_n a_n / \lim_n b_n = 1$.

(b) If $\lim_n a_n / \lim_n b_n$ is well-defined and equal to 1, then $a_n \sim b_n$.

(c) If neither $\lim_n a_n$ nor $\lim_n b_n$ exists, then $a_n \sim b_n$ is impossible.

Exercise 1.18 Suppose that $a_n \sim b_n$ and $c_n \sim d_n$.

(a) Prove that $a_n c_n \sim b_n d_n$.

(b) Show by counterexample that it is not generally true that $a_n + c_n \sim b_n + d_n$.

(c) Prove that $|a_n| + |c_n| \sim |b_n| + |d_n|$.

(d) Show by counterexample that it is not generally true that $f(a_n) \sim f(b_n)$ for a continuous function $f(x)$.

Exercise 1.19 Prove the asymptotic relationship in Example 1.23.

Hint: One way to proceed is to prove that the sum lies between two simple-to-evaluate integrals that are themselves asymptotically equivalent. Consult the proof of Expression (1.21) as a model.

Exercise 1.20 According to the result of Exercise 1.16, the limit (1.21) implies that the *relative* difference between $\sum_{i=1}^{n}(1/i)$ and $\log n$ goes to zero. But this does not imply that the difference itself goes to zero (in general, the difference may not even have any limit at all). In this particular case, the difference converges to a constant called Euler's constant that is sometimes used to define the complex-valued gamma function.

Evaluate $\sum_{i=1}^{n}(1/i) - \log n$ for various large values of n (say, $n \in \{100, 1000, 10000\}$) to approximate the Euler constant.

Exercise 1.21 Let $X_1, \ldots, X_n$ be a simple random sample from an exponential distribution with density $f(x) = \theta \exp(-\theta x)$ and consider the estimator $\delta_n(X) = \sum_{i=1}^{n} X_i/(n+2)$ of $g(\theta) = 1/\theta$. Show that for some constants c_1 and c_2 depending on θ,

$$\text{bias of } \delta_n \sim c_1 \,(\text{variance of } \delta_n) \sim \frac{c_2}{n}$$

as $n \to \infty$. The bias of δ_n equals its expectation minus $(1/\theta)$.

Exercise 1.22 Let $X_1, \ldots, X_n$ be independent with identical density functions $f(x) = \theta x^{\theta-1} I\{0 < x < 1\}$.

(a) Let δ_n be the posterior mean of θ, assuming a standard exponential prior for θ (i.e., $p(\theta) = e^{-\theta} I\{\theta > 0\}$). Compute δ_n.

Hints: The posterior distribution of θ is gamma. If Y is a gamma random variable, then $f(y) \propto y^{\alpha-1} e^{-y\beta}$ and the mean of Y is α/β. To determine α and β for the posterior distribution of θ, simply multiply the prior density times the likelihood function to get an expression equal to the posterior density up to a normalizing constant that is irrelevant in determining α and β.

(b) For each $n \in \{10, 50, 100, 500\}$, simulate 1000 different samples of size n from the given distribution with $\theta = 2$. Use these to calculate the value of δ_n 1000 times for each n. Make a table in which you report, for each n, your estimate of the bias (the sample mean of $\delta_n - 2$) and the variance (the sample variance of δ_n). Try to estimate the asymptotic order of the bias and the variance of this estimator by finding "nice" positive exponents a and b such that $n^a |\text{bias}_n|$

and $n^b \text{variance}_n$ are roughly constant. ("Nice" here may be interpreted to mean integers or half-integers.)

Hints: To generate a sample from the given distribution, use the fact that if $U_1, U_2, \ldots$ is a sample from a uniform $(0, 1)$ density and the continuous distribution function $F(x)$ may be inverted explicitly, then letting $X_i = F^{-1}(U_i)$ results in $X_1, X_2, \ldots$ being a simple random sample from $F(x)$. When using Splus or R, a sample from uniform $(0, 1)$ of size, say, 50 may be obtained by typing `runif(50)`.

Calculating δ_n involves taking the sum of logarithms. Mathematically, this is the same as the logarithm of the product. However, mathematically equivalent expressions are not necessarily computationally equivalent! For a large sample, multiplying all the values could result in overflow or underflow, so the logarithm of the product won't always work. Adding the logarithms is safer even though it requires more computation due to the fact that many logarithms are required instead of just one.

Exercise 1.23 Let $X_1, X_2, \ldots$ be defined as in Exercise 1.22.

(a) Derive a formula for the maximum likelihood estimator of θ for a sample of size n. Call it $\hat{\theta}_n$.

(b) Follow the directions for Exercise 1.22(b) using $\hat{\theta}_n$ instead of δ_n.

Exercise 1.24 Prove that if $f(x)$ is everywhere twice differentiable and $f''(x) \geq 0$ for all x, then $f(x)$ is convex.

Hint: Expand both $\alpha f(x)$ and $(1 - \alpha)f(y)$ using Taylor's theorem 1.18 with $d = 1$, then add. Use the mean value theorem version of the Lagrange remainder (1.7).

Exercise 1.25 Prove Theorem 1.31.

Hint: Let c be an arbitrary constant for which $f(c)$ is defined. Then in inequality (1.25), take $x = b_n$, $y = c$, and $\alpha = (a_n - c)/(b_n - c)$. Be sure your proof uses all of the hypotheses of the theorem; as Exercise 1.26 shows, all of the hypotheses are necessary.

Exercise 1.26 Create counterexamples to the result in Theorem 1.31 if the hypotheses of the theorem are weakened as follows:

(a) Find a_n, b_n, and convex $f(x)$ with $\lim_{x\to\infty} f(x) = \infty$ such that $a_n = o(b_n)$ but $f(a_n) \neq o[f(b_n)]$.

(b) Find a_n, b_n, and convex $f(x)$ such that $a_n \to \infty$, $b_n \to \infty$, and $a_n = o(b_n)$

but $f(a_n) \neq o[f(b_n)]$.

(c) Find a_n, b_n, and $f(x)$ with $\lim_{x\to\infty} f(x) = \infty$ such that $a_n \to \infty$, $b_n \to \infty$, and $a_n = o(b_n)$ but $f(a_n) \neq o[f(b_n)]$.

Exercise 1.27 Recall that $\log n$ always denotes the natural logarithm of n. Assuming that $\log n$ means $\log_{10} n$ will change some of the answers in this exercise!

(a) The following 5 sequences have the property that each tends to ∞ as $n \to \infty$, and for any pair of sequences, one is little-o of the other. List them in order of rate of increase from slowest to fastest. In other words, give an ordering such that first sequence = o(second sequence), second sequence = o(third sequence), etc.

$$n \qquad \sqrt{\log n!} \qquad \sum_{i=1}^{n} \sqrt[3]{i} \qquad 2^{\log n} \qquad (\log n)^{\log\log n}$$

Prove the 4 order relationships that result from your list.

Hint: Here and in part (b), using a computer to evaluate some of the sequences for large values of n can be helpful in suggesting the correct ordering. However, note that this procedure does not constitute a proof!

(b) Follow the directions of part (a) for the following 13 sequences.

$$\log(n!) \qquad n^2 \qquad n^n \qquad 3^n$$

$$\log(\log n) \qquad n \qquad \log n \qquad 2^{3\log n} \qquad n^{n/2}$$

$$n! \qquad 2^{2^n} \qquad n^{\log n} \qquad (\log n)^n$$

Proving the 12 order relationships is challenging but not quite as tedious as it sounds; some of the proofs will be very short.

1.4 Multivariate Extensions

We now consider vectors in $\mathbb{R}^k$, $k > 1$. We denote vectors by bold face and their components by regular type with subscripts; thus, $\mathbf{a}$ is equivalent to $(a_1, \ldots, a_k)$. For sequences of vectors, we use bold face with subscripts, as in $\mathbf{a}_1, \mathbf{a}_2, \ldots$. This notation has a drawback: Since subscripts denote both component numbers and sequence numbers, it is awkward to denote specific components of specific elements in the sequence. When necessary, we will denote the jth component of the ith vector by a_{ij}. In other words, $\mathbf{a}_i = (a_{i1}, \ldots, a_{ik})^\top$ for $i = 1, 2, \ldots$. We follow the convention that vectors are to be considered as columns instead of rows unless stated otherwise, and the transpose of a matrix or vector is denoted by a superscripted $\top$.

The extension to the multivariate case from the univariate case is often so trivial that it is reasonable to ask why we consider the cases separately at all. There are two main reasons. The first is pedagogical: We feel that any disadvantage due to repeated or overlapping material is outweighed by the fact that concepts are often intuitively easier to grasp in $\mathbb{R}$ than in $\mathbb{R}^k$. Furthermore, generalizing from $\mathbb{R}$ to $\mathbb{R}^k$ is often instructive in and of itself, as in the case of the multivariate concept of differentiability. The second reason is mathematical: Some one-dimensional results, like Taylor's Theorem 1.18 for $d > 2$, need not (or cannot, in some cases) be extended to multiple dimensions in these notes. In later chapters in these notes, we will treat univariate and multivariate topics together sometimes and separately sometimes, and we will maintain the bold-face notation for vectors throughout.

To define a limit of a sequence of vectors, we must first define a norm on $\mathbb{R}^k$. We are interested primarily in whether the norm of a vector goes to zero, a concept for which any norm will suffice, so we may as well take the Euclidean norm:

$$\|\mathbf{a}\| \stackrel{\text{def}}{=} \sqrt{\sum_{i=1}^{k} a_i^2} = \sqrt{\mathbf{a}^\top \mathbf{a}}.$$

We may now write down the analogue of Definition 1.1.

Definition 1.32 The sequence $\mathbf{a}_1, \mathbf{a}_2, \ldots$ is said to have limit $\mathbf{c} \in \mathbb{R}^k$, written $\mathbf{a}_n \to \mathbf{c}$ as $n \to \infty$ or $\lim_{n\to\infty} \mathbf{a}_n = \mathbf{c}$, if $\|\mathbf{a}_n - \mathbf{c}\| \to 0$ as $n \to \infty$. That is, $\mathbf{a}_n \to \mathbf{c}$ means that for any $\epsilon > 0$ there exists N such that $\|\mathbf{a}_n - \mathbf{c}\| < \epsilon$ for all $n > N$.

It is sometimes possible to define multivariate concepts by using the univariate definition on each of the components of the vector. For instance, the following lemma gives an alternative way to define $\mathbf{a}_n \to \mathbf{c}$:

Lemma 1.33 $\mathbf{a}_n \to \mathbf{c}$ if and only if $a_{nj} \to c_j$ for all $1 \le j \le k$.

Proof: Since

$$\|\mathbf{a}_n - \mathbf{c}\| = \sqrt{(a_{n1} - c_1)^2 + \cdots + (a_{nk} - c_k)^2},$$

the "if" part follows from repeated use of Theorem 1.5 (which says that the limit of a sum is the sum of the limits and the limit of a product is the product of the limits) and Theorem 1.16 (which says that continuous functions preserve limits). The "only if" part follows because $|a_{nj} - c_j| \le \|\mathbf{a}_n - \mathbf{c}\|$ for each j. ∎

There is no multivariate analogue of Definition 1.2; it is nonsensical to write $\mathbf{a}_n \to \infty$. However, since $\|\mathbf{a}_n\|$ is a real number, writing $\|\mathbf{a}_n\| \to \infty$ is permissible. If we write $\lim_{\|\mathbf{x}\|\to\infty} f(\mathbf{x}) = c$ for a real-valued function $f(\mathbf{x})$, then it must be true that $f(\mathbf{x})$ tends to the same limit c no matter what path $\mathbf{x}$ takes as $\|\mathbf{x}\| \to \infty$.

Suppose that the function $\mathbf{f}(\mathbf{x})$ maps vectors in some open subset U of $\mathbb{R}^k$ to vectors in $\mathbb{R}^\ell$, a property denoted by $\mathbf{f} : U \to \mathbb{R}^\ell$. In order to define continuity, we first extend Definition 1.11 to the multivariate case:

Definition 1.34 For a function $\mathbf{f} : U \to \mathbb{R}^\ell$, where U is open in $\mathbb{R}^k$, we write $\lim_{\mathbf{x}\to\mathbf{a}} \mathbf{f}(\mathbf{x}) = \mathbf{c}$ for some $\mathbf{a} \in U$ and $\mathbf{c} \in \mathbb{R}^\ell$ if for every $\epsilon > 0$ there exists a $\delta > 0$ such that $\|\mathbf{f}(\mathbf{x}) - \mathbf{c}\| < \epsilon$ whenever $\mathbf{x} \in U$ and $0 < \|\mathbf{x} - \mathbf{a}\| < \delta$.

In Definition 1.34, $\|\mathbf{f}(\mathbf{x}) - \mathbf{c}\|$ refers to the norm on $\mathbb{R}^\ell$, while $\|\mathbf{x} - \mathbf{a}\|$ refers to the norm on $\mathbb{R}^k$.

Definition 1.35 A function $\mathbf{f} : U \to \mathbb{R}^\ell$ is continuous at $\mathbf{a} \in U \subset \mathbb{R}^k$ if $\lim_{\mathbf{x}\to\mathbf{a}} \mathbf{f}(\mathbf{x}) = \mathbf{f}(\mathbf{a})$.

Since there is no harm in letting $k = 1$ or $\ell = 1$ or both, Definitions 1.34 and 1.35 include Definitions 1.11 and 1.14(a), respectively, as special cases.

The extension of differentiation from the univariate to the multivariate setting is not quite as straightforward as the extension of continuity. Part of the difficulty lies merely in notation, but we will also rely on a qualitatively different definition of the derivative in the multivariate setting. Recall that in the univariate case, Taylor's Theorem 1.18 implies that the derivative $f'(x)$ of a function $f(x)$ satisfies

$$\frac{f(x+h) - f(x) - hf'(x)}{h} \to 0 \quad \text{as } h \to 0. \tag{1.28}$$

It turns out that Equation (1.28) could have been taken as the *definition* of the derivative $f'(x)$. To do so would have required just a bit of extra work to prove that Equation (1.28) uniquely defines $f'(x)$, but this is precisely how we shall now extend differentiation to the multivariate case:

Definition 1.36 Suppose that $\mathbf{f} : U \to \mathbb{R}^\ell$, where $U \subset \mathbb{R}^k$ is open. For a point $\mathbf{a} \in U$, suppose there exists an $\ell \times k$ matrix $J_\mathbf{f}(\mathbf{a})$, depending on $\mathbf{a}$ but not on $\mathbf{h}$, such that

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\mathbf{f}(\mathbf{a}+\mathbf{h}) - \mathbf{f}(\mathbf{a}) - J_\mathbf{f}(\mathbf{a})\mathbf{h}}{\|\mathbf{h}\|} = \mathbf{0}. \tag{1.29}$$

Then $J_\mathbf{f}(\mathbf{a})$ is unique and we call $J_\mathbf{f}(\mathbf{a})$ the Jacobian matrix of $\mathbf{f}(\mathbf{x})$ at $\mathbf{a}$. We say that $\mathbf{f}(\mathbf{x})$ is differentiable at the point $\mathbf{a}$, and $J_\mathbf{f}(\mathbf{x})$ may be called the derivative of $\mathbf{f}(\mathbf{x})$.

The assertion in Definition 1.36 that $J_\mathbf{f}(\mathbf{a})$ is unique may be proved as follows: Suppose that $J^{(1)}_\mathbf{f}(\mathbf{a})$ and $J^{(2)}_\mathbf{f}(\mathbf{a})$ are two versions of the Jacobian matrix. Then Equation (1.29) implies

that

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\left(J_{\mathbf{f}}^{(1)}(\mathbf{a}) - J_{\mathbf{f}}^{(2)}(\mathbf{a})\right)\mathbf{h}}{\|\mathbf{h}\|} = \mathbf{0};$$

but $\mathbf{h}/\|\mathbf{h}\|$ is an arbitrary unit vector, which means that $\left(J_{\mathbf{f}}^{(1)}(\mathbf{a}) - J_{\mathbf{f}}^{(2)}(\mathbf{a})\right)$ must be the zero matrix, proving the assertion. ■

Although Definition 1.36, sometimes called the Fréchet derivative, is straightforward and quite common throughout the calculus literature, there is unfortunately not a universally accepted notation for multivariate derivatives. Various authors use notation such as $\mathbf{f}'(\mathbf{x})$, $\dot{\mathbf{f}}(\mathbf{x})$, $D\mathbf{f}(\mathbf{x})$, or $\nabla\mathbf{f}(\mathbf{x})$ to denote the Jacobian matrix or its transpose, depending on the situation. In these notes, we adopt perhaps the most widespread of these notations, letting $\nabla\mathbf{f}(\mathbf{x})$ denote the *transpose* of the Jacobian matrix $J_{\mathbf{f}}(\mathbf{x})$. We often refer to $\nabla\mathbf{f}$ as the *gradient* of $\mathbf{f}$.

When the Jacobian matrix exists, it is equal to the matrix of partial derivatives, which are defined as follows:

Definition 1.37 Let $g(\mathbf{x})$ be a real-valued function defined on a neighborhood of $\mathbf{a}$ in $\mathbb{R}^k$. For $1 \le i \le k$, let $\mathbf{e}_i$ denote the ith standard basis vector in $\mathbb{R}^k$, consisting of a one in the ith component and zeros elsewhere. We define the ith partial derivative of $g(\mathbf{x})$ at $\mathbf{a}$ to be

$$\left.\frac{\partial g(\mathbf{x})}{\partial x_i}\right|_{\mathbf{x}=\mathbf{a}} \stackrel{\text{def}}{=} \lim_{h\to 0} \frac{g(\mathbf{a}+h\mathbf{e}_i) - g(\mathbf{a})}{h},$$

if this limit exists.

Now we are ready to state that the Jacobian matrix is the matrix of partial derivatives.

Theorem 1.38 Suppose $\mathbf{f}(\mathbf{x})$ is differentiable at $\mathbf{a}$ in the sense of Definition 1.36. Define the gradient matrix $\nabla\mathbf{f}(\mathbf{a})$ to be the transpose of the Jacobian matrix $J_{\mathbf{f}}(\mathbf{a})$. Then

$$\nabla\mathbf{f}(\mathbf{a}) = \left.\begin{pmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_\ell(\mathbf{x})}{\partial x_1} \\ \vdots & & \vdots \\ \frac{\partial f_1(\mathbf{x})}{\partial x_k} & \cdots & \frac{\partial f_\ell(\mathbf{x})}{\partial x_k} \end{pmatrix}\right|_{\mathbf{x}=\mathbf{a}.} \tag{1.30}$$

The converse of Theorem 1.38 is not true, in the sense that the existence of partial derivatives of a function does not guarantee the differentiability of that function (see Exercise 1.31).

When $\mathbf{f}$ maps k-vectors to ℓ-vectors, $\nabla\mathbf{f}(\mathbf{x})$ is a $k \times \ell$ matrix, a fact that is important to memorize; it is often very helpful to remember the dimensions of the gradient matrix when

trying to recall the form of various multivariate results. To try to simplify the admittedly confusing notational situation resulting from the introduction of both a Jacobian matrix and a gradient, we will use only the gradient notation $\nabla \mathbf{f}(\mathbf{x})$, defined in Equation (1.30), throughout these notes.

By Definition 1.36, the gradient matrix satisfies the first-order Taylor formula

$$\mathbf{f}(\mathbf{x}) = \mathbf{f}(\mathbf{a}) + \nabla \mathbf{f}(\mathbf{a})^\top (\mathbf{x} - \mathbf{a}) + \mathbf{r}(\mathbf{x}, \mathbf{a}), \tag{1.31}$$

where $\mathbf{r}(\mathbf{x}, \mathbf{a})/\|\mathbf{x} - \mathbf{a}\| \to \mathbf{0}$ as $\mathbf{x} \to \mathbf{a}$.

Now that we have generalized Taylor's Theorem 1.18 for the linear case $d = 1$, it is worthwhile to ask whether a similar generalization is necessary for larger d. The answer is no, except for one particular case: We will require a second-order Taylor expansion (that is, $d = 2$) when $f(\mathbf{x})$ is real-valued but its argument $\mathbf{x}$ is a vector. To this end, suppose that $U \subset \mathbb{R}^k$ is open and that $f(\mathbf{x})$ maps U into $\mathbb{R}$. Then according to Equation (1.30), $\nabla f(\mathbf{x})$ is a $k \times 1$ vector of partial derivatives, which means that $\nabla f(\mathbf{x})$ maps k-vectors to k-vectors. If we differentiate once more and evaluate the result at $\mathbf{a}$, denoting the result by $\nabla^2 f(\mathbf{a})$, then Equation (1.30) with $\partial/\partial x_i f(\mathbf{x})$ substituted for $f_i(\mathbf{x})$ gives

$$\nabla^2 f(\mathbf{a}) = \left(\begin{array}{ccc} \frac{\partial^2 f(\mathbf{x})}{\partial x_1^2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_k} \\ \vdots & & \vdots \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_k \partial x_1} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_k^2} \end{array} \right) \Bigg|_{\mathbf{x}=\mathbf{a}}. \tag{1.32}$$

Definition 1.39 The $k \times k$ matrix on the right hand side of Equation (1.32), when it exists, is called the *Hessian* matrix of the function $f(\mathbf{x})$ at $\mathbf{a}$.

Twice differentiability guarantees the existence (by two applications of Theorem 1.38) and symmetry (by Theorem 1.40 below) of the Hessian matrix. The Hessian may exist for a function that is not twice differentiable, as seen in Exercise 1.33, but this mathematical curiosity will not concern us elsewhere in these notes.

We state the final theorem of this section, which extends second-order Taylor expansions to a particular multivariate case, without proof, but the interested reader may consult Magnus and Neudecker (1999) for an encyclopedic treatment of this and many other topics involving differentiation.

Theorem 1.40 Suppose that the real-valued function $f(\mathbf{x})$ is twice differentiable at some point $\mathbf{a} \in \mathbb{R}^k$. Then $\nabla^2 f(\mathbf{a})$ is a symmetric matrix, and

$$f(\mathbf{x}) = f(\mathbf{a}) + \nabla f(\mathbf{a})^\top (\mathbf{x} - \mathbf{a}) + \frac{1}{2}(\mathbf{x} - \mathbf{a})^\top \nabla^2 f(\mathbf{a})(\mathbf{x} - \mathbf{a}) + r_2(\mathbf{x}, \mathbf{a}),$$

where $r_2(\mathbf{x}, \mathbf{a})/\|\mathbf{x} - \mathbf{a}\|^2 \to 0$ as $\mathbf{x} \to \mathbf{a}$.

Exercises for Section 1.4

Exercise 1.28 **(a)** Suppose that $f(\mathbf{x})$ is continuous at $\mathbf{0}$. Prove that $f(t\mathbf{e}_i)$ is continuous as a function of t at $t = 0$ for each i, where $\mathbf{e}_i$ is the ith standard basis vector.

(b) Prove that the converse of (a) is not true by inventing a function $f(\mathbf{x})$ that is not continuous at $\mathbf{0}$ but such that $f(t\mathbf{e}_i)$ is continuous as a function of t at $t = 0$ for each i.

Exercise 1.29 Suppose that $a_{nj} \to c_j$ as $n \to \infty$ for $j = 1, \ldots, k$. Prove that if $f : \mathbb{R}^k \to \mathbb{R}$ is continuous at the point $\mathbf{c}$, then $f(\mathbf{a}_n) \to f(\mathbf{c})$. This proves every part of Exercise 1.1. (The hard work of an exercise like 1.1(b) is in showing that multiplication is continuous).

Exercise 1.30 Prove Theorem 1.38.

Hint: Starting with Equation (1.29), take $\mathbf{x} = \mathbf{a} + t\mathbf{e}_i$ and let $t \to 0$, where $\mathbf{e}_i$ is defined in Definition 1.37.

Exercise 1.31 Prove that the converse of Theorem 1.38 is not true by finding a function that is not differentiable at some point but whose partial derivatives at that point all exist.

Exercise 1.32 Suppose that $X_1, \ldots, X_n$ comprises a sample of independent and identically distributed normal random variables with density

$$f(x_i; \mu, \sigma^2) = \frac{\exp\{-\frac{1}{2\sigma^2}(x_i - \mu)^2\}}{\sqrt{2\pi\sigma^2}}.$$

Let $\ell(\mu, \sigma^2)$ denote the loglikelihood function; i.e., $\ell(\mu, \sigma^2)$ is the logarithm of the joint density $\prod_i f(X_i; \mu, \sigma^2)$, viewed as a function of the parameters μ and σ^2.

The *score vector* is defined to be the gradient of the loglikelihood. Find the score vector for this example.

Hint: The score vector is a vector with two components and it is a function of $X_1, \ldots, X_n$, μ, and σ^2. Setting the score vector equal to zero and solving for μ and σ^2 gives the well-known maximum likelihood estimators of μ and σ^2, namely $\overline{X}$ and $\frac{1}{n}\sum_i (X_i - \overline{X})^2$.

Exercise 1.33 Define

$$f(x, y) = \begin{cases} 0 & \text{if } x = y = 0; \\ \frac{x^3 y - x y^3}{x^2 + y^2} & \text{otherwise.} \end{cases}$$

Use Theorem 1.40 to demonstrate that $f(x, y)$ is not twice differentiable at $(0, 0)$ by showing that $\nabla^2 f(0, 0)$, which does exist, is not symmetric.

Exercise 1.34 **(a)** Find the Hessian matrix of the loglikelihood function defined in Exercise 1.32.

(b) Suppose that $n = 10$ and that we observe this sample:

```
2.946  0.975  1.333  4.484  1.711
2.627 -0.628  2.476  2.599  2.143
```

Evaluate the Hessian matrix at the maximum likelihood estimator $(\hat{\mu}, \hat{\sigma}^2)$. (A formula for the MLE is given in the hint to Exercise 1.32).

(c) As we shall see in Chapter 7, the negative inverse of the Hessian matrix is a reasonable large-sample estimator of the covariance matrix of the MLE (though with only $n = 10$, it is not clear how good this estimator would be in this example!). Invert your answer from part (b), then put a negative sign in front and use the answer to give approximate standard errors (the square roots of the diagonal entries) for $\hat{\mu}$ and $\hat{\sigma}^2$.

Exercise 1.35 Suppose $X_1, \ldots, X_n$ is a sample of independent and identically distributed random variables from a Beta(α, β) distribution, for which the density function is

$$f(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1 - x)^{\beta-1} \qquad \text{for } 0 < x < 1,$$

where α and β are assumed to be positive parameters.

(a) Calculate the score vector (the gradient of the loglikelihood) and the Hessian of the loglikelihood. Recall the definitions of the digamma and trigamma functions in Exercises (1.14) and (1.15).

Exercise 1.36 The gamma distribution with shape parameter $\alpha > 0$ and rate parameter $\beta > 0$ has density function

$$f(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \qquad \text{for } x > 0.$$

(a) Calculate the score vector for an independent and identically distributed gamma(α, β) sample of size n.

(b) Using the approximation to the digamma function $\Psi(x)$ given in Equation (1.16), find a closed-form approximation to the maximum likelihood estimator

(obtained by setting the score vector equal to zero and solving for α and β). Simulate 1000 samples of size $n = 50$ from gamma$(5, 1)$ and calculate this approximation for each. Give histograms of these estimators. Can you characterize their performance?

The approximation of $\Psi(x)$ in Equation (1.16) can be extremely poor for $x < 2$, so the method above is *not* a reliable general-purpose estimation procedure.

1.5 Expectation and Inequalities

While random variables have made only occasional appearances in these notes before now, they will be featured prominently from now on. We do not wish to make the definition of a random variable rigorous here—to do so requires measure theory—but we assume that the reader is familiar with the basic idea: A random variable is a function from a sample space Ω into $\mathbb{R}$. (We often refer to "random vectors" rather than "random variables" if the range space is $\mathbb{R}^k$ rather than $\mathbb{R}$.)

For any random variable X, we denote the expected value of X, if this value exists, by $\mathrm{E}\, X$. We assume that the reader is already familiar with expected values for commonly-encountered random variables, so we do not attempt here to define the expectation operator E rigorously. In particular, we avoid writing explicit formulas for $\mathrm{E}\, X$ (e.g., sums if X is discrete or integrals if X is continuous) except when necessary. Much of the theory in these notes may be developed using only the $\mathrm{E}\, X$ notation; exceptions include cases in which we wish to evaluate particular expectations and cases in which we must deal with density functions (such as the topic of maximum likelihood estimation). For students who have not been exposed to any sort of a rigorous treatment of random variables and expectation, we hope that the many applications of this theory presented here will pique your curiosity and encourage you to delve further into the technical details of random variables, expectations, and conditional expectations. Nearly any advanced probability textbook will develop these details. For a quick, introductory-level exposure to these intricacies, we recommend the first chapter of Lange (2003).

Not all random variables have expectations, even if we allow the possibilities $\mathrm{E}\, X = \pm\infty$: Let $X^+ = \max\{X, 0\}$ and $X^- = \max\{-X, 0\}$ denote the positive and negative parts of X, so that $X = X^+ - X^-$. Now both $\mathrm{E}\, X^+$ and $\mathrm{E}\, X^-$ are always well-defined if we allow ∞ as a possibility, but if both X^+ and X^- have infinite expectation, then there is no sensible way to define $\mathrm{E}\, X$. It is easy to find examples of random variables X for which $\mathrm{E}\, X$ is undefined. Perhaps the best-known example is a Cauchy random variable (whose density function is given in Exercise 7.3), but we may construct other examples by taking any two independent nonnegative random variables Y_1 and Y_2 with infinite expectation—e.g., let Y_i take the value

2^n with probability 2^{-n} for all positive integers n—and simply defining $X = Y_1 - Y_2$.

The expectation operator has several often-used properties, listed here as axioms because we will not derive them from first principles. We assume below that X and Y are defined on the same sample space Ω and $\mathrm{E}\ X$ and $\mathrm{E}\ Y$ are well-defined.

1. *Linearity:* For any real numbers a and b, $\mathrm{E}\,(aX + bY) = a\,\mathrm{E}\,(X) + b\,\mathrm{E}\,(Y)$ (and if $a\,\mathrm{E}\,(X) + b\,\mathrm{E}\,(Y)$ is undefined, then so is $\mathrm{E}\,(aX + bY)$).
2. *Monotonicity:* If $X(\omega) \leq Y(\omega)$ for all $\omega \in \Omega$, then $\mathrm{E}\ X \leq \mathrm{E}\ Y$.
3. *Conditioning:* If $\mathrm{E}\,(X|Y)$ denotes the conditional expectation of X given Y (which, as a function of Y, is itself a random variable), then $\mathrm{E}\ X = \mathrm{E}\ \{\mathrm{E}\,(X|Y)\}$.

As a special case of the conditioning property, note that if X and Y are independent, then $\mathrm{E}\,(X|Y) = \mathrm{E}\ X$, which gives the well-known identity

$$\mathrm{E}\ XY = \mathrm{E}\ \{\mathrm{E}\,(XY|Y)\} = \mathrm{E}\ \{Y\,\mathrm{E}\,(X|Y)\} = \mathrm{E}\ \{Y\,\mathrm{E}\ X\} = \mathrm{E}\ X\,\mathrm{E}\ Y,$$

where we have used the fact that $\mathrm{E}\,(XY|Y) = Y\,\mathrm{E}\,(X|Y)$, which is always true because conditioning on Y is like holding it constant.

The variance and covariance operators are defined as usual, namely,

$$\mathrm{Cov}\,(X, Y) \stackrel{\mathrm{def}}{=} \mathrm{E}\ XY - (\mathrm{E}\ X)(\mathrm{E}\ Y)$$

and $\mathrm{Var}\,(X) \stackrel{\mathrm{def}}{=} \mathrm{Cov}\,(X, X)$. The linearity property above extends to random vectors: For scalars a and b we have $\mathrm{E}\,(a\mathbf{X} + b\mathbf{Y}) = a\,\mathrm{E}\,(\mathbf{X}) + b\,\mathrm{E}\,(\mathbf{Y})$, and for matrices P and Q with dimensions such that $P\mathbf{X} + Q\mathbf{Y}$ is well-defined, $\mathrm{E}\,(P\mathbf{X} + Q\mathbf{Y}) = P\,\mathrm{E}\,(\mathbf{X}) + Q\,\mathrm{E}\,(\mathbf{Y})$. The covariance between two random vectors is

$$\mathrm{Cov}\,(\mathbf{X}, \mathbf{Y}) \stackrel{\mathrm{def}}{=} \mathrm{E}\ \mathbf{X}\mathbf{Y}^\top - (\mathrm{E}\ \mathbf{X})(\mathrm{E}\ \mathbf{Y})^\top,$$

and the variance matrix of a random vector (sometimes referred to as the covariance matrix) is $\mathrm{Var}\,(\mathbf{X}) \stackrel{\mathrm{def}}{=} \mathrm{Cov}\,(\mathbf{X}, \mathbf{X})$. Among other things, these properties imply that

$$\mathrm{Var}\,(P\mathbf{X}) = P\,\mathrm{Var}\,(\mathbf{X})P^\top \tag{1.33}$$

for any constant matrix P with as many columns as $\mathbf{X}$ has rows.

Example 1.41 As a first application of the monotonicity of the expectation operator, we derive a useful inequality called Chebyshev's inequality. For any positive constants a and r and any random variable X, observe that

$$|X|^r \geq |X|^r I\{|X| \geq a\} \geq a^r I\{|X| \geq a\},$$

where throughout these notes, $I\{\cdot\}$ denotes the indicator function

$$I\{\textit{expression}\} \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } \textit{expression} \text{ is true} \\ 0 & \text{if } \textit{expression} \text{ is not true.} \end{cases} \tag{1.34}$$

Since $\mathrm{E}\ I\{|X| \geq a\} = P(|X| \geq a)$, the monotonicity of the expectation operator implies

$$P(|X| \geq a) \leq \frac{\mathrm{E}\ |X|^r}{a^r}. \tag{1.35}$$

Inequality (1.35) is sometimes called Markov's inequality. In the special case that $X = Y - \mathrm{E}\ Y$ and $r = 2$, we obtain Chebyshev's inequality: For any $a > 0$ and any random Y,

$$P(|Y - \mathrm{E}\ Y| \geq a) \leq \frac{\mathrm{Var}\ Y}{a^2}. \tag{1.36}$$

Example 1.42 We now derive another inequality, Jensen's, that takes advantage of linearity as well as monotonicity. Jensen's inequality states that

$$f\left(\mathrm{E}\ X\right) \leq \mathrm{E}\ f(X) \tag{1.37}$$

for any convex function $f(x)$ and any random variable X. Definition 1.30 tells precisely what a convex function is, but the intuition is simple: Any line segment connecting two points on the graph of a convex function must never go below the graph (valley-shaped graphs are convex; hill-shaped graphs are not). To prove inequality 1.37, we require another property of any convex function, called the supporting hyperplane property. This property, whose proof is the subject of Exercise 1.38, essentially guarantees that for any point on the graph of a convex function, it is possible to construct a hyperplane through that point that puts the entire graph on one side of that hyperplane.

In the context of inequality (1.37), the supporting hyperplane property guarantees that there exists a line $g(x) = ax + b$ through the point $[\mathrm{E}\ X, f(\mathrm{E}\ X)]$ such that $g(x) \leq f(x)$ for all x (see Figure 1.2). By monotonicity, we know that $\mathrm{E}\ g(X) \leq \mathrm{E}\ f(X)$. We now invoke the linearity of the expectation operator to conclude that

$$\mathrm{E}\ g(X) = g(\mathrm{E}\ X) = f(\mathrm{E}\ X),$$

which proves inequality (1.37).

Exercises for Section 1.5

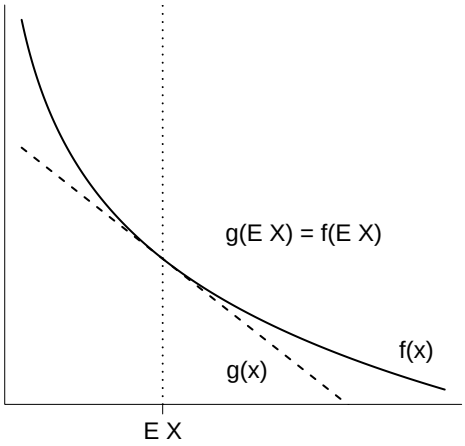

Figure 1.2: *The solid curve is a convex function $f(x)$ and the dotted line is a supporting hyperplane $g(x)$, tangent at $x = \mathrm{E}\ X$. This figure shows how to prove Jensen's inequality.*

Exercise 1.37 Show by example that equality can hold in inequality 1.36.

Exercise 1.38 Let $f(x)$ be a convex function on some interval, and let x_0 be any point on the interior of that interval.

(a) Prove that

$$\lim_{x \to x_0+} \frac{f(x) - f(x_0)}{x - x_0} \tag{1.38}$$

exists and is finite; that is, a one-sided derivative exists at x_0.

Hint: Using Definition 1.30, show that the fraction in expression (1.38) is non-increasing and bounded below as x decreases to x_0.

(b) Prove that there exists a linear function $g(x) = ax+b$ such that $g(x_0) = f(x_0)$ and $g(x) \leq f(x)$ for all x in the interval. This fact is the *supporting hyperplane* property in the case of a convex function taking a real argument.

Hint: Let $f'(x_0+)$ denote the one-sided derivative of part (a). Consider the line $f(x_0) + f'(x_0+)(x - x_0)$.

Exercise 1.39 Prove Hölder's inequality: For random variables X and Y and positive p and q such that $p + q = 1$,

$$\mathrm{E}\ |XY| \leq \left(\mathrm{E}\ |X|^{1/p}\right)^p \left(\mathrm{E}\ |Y|^{1/q}\right)^q. \tag{1.39}$$

(If $p = q = 1/2$, inequality (1.39) is also called the Cauchy-Schwartz inequality.)

Hint: Use the convexity of $\exp(x)$ to prove that $|abXY| \leq p|aX|^{1/p} + q|bY|^{1/q}$ whenever $aX \neq 0$ and $bY \neq 0$ (the same inequality is also true if $aX = 0$ or $bY = 0$). Take expectations, then find values for the scalars a and b that give the desired result when the right side of inequality (1.39) is nonzero.

Exercise 1.40 Use Hölder's Inequality (1.39) to prove that if $\alpha > 1$, then

$$(\mathrm{E}\ |X|)^{\alpha} \leq \mathrm{E}\ |X|^{\alpha}.$$

Hint: Take Y to be a constant in Inequality (1.39).

Exercise 1.41 Kolmogorov's inequality is a strengthening of Chebyshev's inequality for a sum of independent random variables: If $X_1, \ldots, X_n$ are independent random variables, define

$$S_k = \sum_{i=1}^{k} (X_i - \mathrm{E}\ X_i)$$

to be the centered kth partial sum for $1 \leq k \leq n$. Then for $a > 0$, Kolmogorov's inequality states that

$$P\left(\max_{1 \leq k \leq n} |S_k| \geq a\right) \leq \frac{\mathrm{Var}\ S_n}{a^2}. \tag{1.40}$$

(a) Let A_k denote the event that $|S_i| \geq a$ for the first time when $i = k$; that is, that $|S_k| \geq a$ and $|S_j| < a$ for all $j < k$. Prove that

$$a^2 P\left(\max_{1 \leq k \leq n} |S_k| \geq a\right) \leq \sum_{i=1}^{n} \mathrm{E}\ \left[I\{A_k\} S_k^2\right].$$

Hint: Argue that

$$\sum_{i=1}^{n} \mathrm{E}\ I\{A_i\} = P\left(\max_{1 \leq k \leq n} |S_k| \geq a\right)$$

and $\mathrm{E}\ [I\{A_k\} S_k^2] \geq a^2\, \mathrm{E}\ I\{A_k\}$.

(b) Prove that

$$\mathrm{E}\ S_n^2 \geq \sum_{k=1}^{n} \mathrm{E}\ \left[I\{A_k\}\{S_k^2 + 2S_k(S_n - S_k)\}\right].$$

Hint: Use the fact that the A_k are nonoverlapping, which implies that $1 \geq I(A_1) + \cdots + I(A_n)$. Also use $S_n^2 = S_k^2 + 2S_k(S_n - S_k) + (S_n - S_k)^2$.

(c) Using parts (a) and (b), prove inequality (1.40).

Hint: By independence,

$$\mathrm{E}\,[I\{A_k\}S_k(S_n - S_k)] = \mathrm{E}\,[I\{A_k\}S_k]\,\mathrm{E}\,(S_n - S_k).$$

What is $\mathrm{E}\,(S_n - S_k)$?

Exercise 1.42 Try a simple numerical example to check how much sharper Kolmogorov's inequality (1.40) is than Chebyshev's inequality (1.36).

(a) Take $n = 8$ and assume that $X_1, \ldots, X_n$ are independent normal random variables with $\mathrm{E}\,X_i = 0$ and $\mathrm{Var}\,X_i = 9 - i$. Take $a = 12$. Calculate the exact values on both sides of Chebyshev's inequality (1.36).

(b) Simulate 10^4 realizations of the situation described in part (a). For each, record the maximum value attained by $|S_k|$ for $k = 1, \ldots, 8$. Approximate the probability on the left hand side of Kolmogorov's inequality (1.40). Describe what you find when you compare parts (a) and (b). How does a histogram of the maxima found in part (b) compare with the distribution of $|S_n|$?

Exercise 1.43 The complex plane $\mathbb{C}$ consists of all points $x + \mathrm{i}y$, where x and y are real numbers and $\mathrm{i} = \sqrt{-1}$. The elegant result known as Euler's formula relates the points on the unit circle to the complex exponential function:

$$\exp\{\mathrm{i}t\} = \cos t + \mathrm{i}\sin t \qquad \text{for all } t \in \mathbb{R}. \tag{1.41}$$

Because $e^{\mathrm{i}t}$ is on the unit circle for all real-valued t, the norm (also known as the modulus) of $e^{\mathrm{i}t}$, denoted $|e^{\mathrm{i}t}|$, equals 1. This fact leads to the following generalization of the triangle inequality: For any real-valued function $g(x)$ and any real number t,

$$\left| \int_0^t g(x)e^{\mathrm{i}x}\,dx \right| \leq \left| \int_0^t \left|g(x)e^{\mathrm{i}x}\right|\,dx \right| = \left| \int_0^t |g(x)|\,dx \right|. \tag{1.42}$$

The inequalities below in parts (a) through (d) involving $\exp\{\mathrm{i}t\}$ will be used in Chapter 4. Assume t is a real number, then use Equations (1.6) and (1.41), together with Inequality (1.42), to prove them. [Since we only claim Equation (1.6) to be valid for real-valued functions of real variables, it is necessary here to use Euler's formula to separate $e^{\mathrm{i}t}$ into its real and imaginary parts, namely $\cos t$ and $\sin t$, then Taylor-expand them separately before reassembling the parts using Euler's formula again.]

(a) In Equation (1.6), use $a = 0$ and $d = 0$ on both $\cos t$ and $\sin t$ to show that for any $t \in \mathbb{R}$,

$$|\exp\{\mathrm{i}t\} - 1| \leq |t|.$$

(b) Proceed as above but with $d = 1$ to show that

$$|\exp\{\mathrm{i}t\} - 1 - \mathrm{i}t| \leq t^2/2.$$

(c) Proceed as above but with $d = 2$ to show that

$$\left|\exp\{\mathrm{i}t\} - 1 - \mathrm{i}t + \frac{1}{2}t^2\right| \leq |t|^3/6.$$

(d) Proceed as above but using $d = 1$ for $\sin t$, then $d = 2$ together with integration by parts for $\cos t$, to show that

$$\left|\exp\{\mathrm{i}t\} - 1 - \mathrm{i}t + \frac{1}{2}t^2\right| \leq t^2.$$

Exercise 1.44 Refer to Exercise 1.43. Graph the functions $\left|\exp\{\mathrm{i}t\} - 1 - \mathrm{i}t + \frac{1}{2}t^2\right|$, $|t|^3/6$, and t^2 for t in the interval $[-10, 10]$. Graph the three curves on the same set of axes, using different plotting styles so they are distinguishable from one another. As a check, verify that the inequalities in Exercises 1.43(c) and (d) appear to be satisfied.

Hint: The modulus $|z|$ of a complex number $z = x + \mathrm{i}y$ equals $\sqrt{x^2 + y^2}$. Refer to Equation (1.41) to deal with the expression $\exp\{\mathrm{i}t\}$.

Exercise 1.45 For any nonnegative random variable Y with finite expectation, prove that

$$\sum_{i=1}^{\infty} P(Y \geq i) \leq \mathrm{E}\ Y. \tag{1.43}$$

Hint: First, prove that equality holds if Y is supported on the nonnegative integers. Then note for a general Y that $\mathrm{E}\lfloor Y \rfloor \leq \mathrm{E}\ Y$, where $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x.

Though we will not do so here, it is possible to prove a statement stronger than inequality (1.43) for nonnegative random variables, namely,

$$\int_0^{\infty} P(Y \geq t)\, dt = \mathrm{E}\ Y.$$

(This equation remains true if $\mathrm{E}\, Y = \infty$.) To sketch a proof, note that if we can prove $\int \mathrm{E}\, f(Y,t)\, dt = \mathrm{E} \int f(Y,t)\, dt$, the result follows immediately by taking $f(Y,t) = I\{Y \geq t\}$.

Chapter 2

Weak Convergence

Chapter 1 discussed limits of sequences of constants, either scalar-valued or vector-valued. Chapters 2 and 3 extend this notion by defining what it means for a sequence of *random variables* to have a limit. As it turns out, there is more than one sensible way to do this.

Chapters 2 and 4 (and, to a lesser extent, Chapter 3) lay the theoretical groundwork for nearly all of the statistical topics that will follow. While the material in Chapter 2 is essential, readers may wish to skip Chapter 3 on a first reading. As is common throughout the book, some of the proofs here have been relegated to the exercises.

2.1 Modes of Convergence

Whereas the limit of a sequence of real numbers is unequivocally expressed by Definition 1.32, in the case of random variables there are several ways to define the convergence of a sequence. This section discusses three such definitions, or modes, of convergence; Section 3.1 presents a fourth. Because it is often easier to understand these concepts in the univariate case than the multivariate case, we only consider univariate random vectors here, deferring the analogous multivariate topics to Section 2.3.

2.1.1 Convergence in Probability

What does it mean for the sequence $X_1, X_2, \ldots$ of random variables to converge to, say, the random variable X? Under what circumstances should one write $X_n \to X$? We begin by considering a definition of convergence that requires that X_n and X be defined on the same sample space. For this form of convergence, called convergence in probability, the

absolute difference $|X_n - X|$, itself a random variable, should be arbitrarily close to zero with probability arbitrarily close to one. More precisely, we make the following definition.

Definition 2.1 Let $\{X_n\}_{n\geq 1}$ and X be defined on the same probability space. We say that X_n converges in probability to X, written $X_n \xrightarrow{P} X$, if for any $\epsilon > 0$,

$$P(|X_n - X| < \epsilon) \to 1 \text{ as } n \to \infty. \tag{2.1}$$

It is very common that the X in Definition 2.1 is a constant, say $X \equiv c$. In such cases, we simply write $X_n \xrightarrow{P} c$. When we replace X by c in Definition 2.1, we do not need to concern ourselves with the question of whether X is defined on the same sample space as X_n because any constant may be defined as a random variable on any sample space. In the most common statistical usage of convergence to a constant c, we take c to be some parameter θ and X_n to be an estimator of θ:

Definition 2.2 If $X_n \xrightarrow{P} \theta$, X_n is said to be *consistent* (or weakly consistent) for θ.

As the name suggests, weak consistency is weaker than (i.e., implied by) a condition called "strong consistency," which will be defined in Chapter 3. "Consistency," used without the word "strong" or "weak," generally refers to weak consistency. Throughout this book, we shall refer repeatedly to (weakly) consistent estimators, whereas strong consistency plays a comparatively small role.

Example 2.3 Suppose that $X_1, X_2, \ldots$ are independent and identically distributed (i.i.d.) uniform $(0, \theta)$ random variables, where θ is an unknown positive constant. For $n \geq 1$, let $X_{(n)}$ be defined as the largest value among X_1 through X_n: That is, $X_{(n)} \stackrel{\text{def}}{=} \max_{1\leq i\leq n} X_i$. Then we may show that $X_{(n)}$ is a consistent estimator of θ as follows:

By Definition 2.1, we wish to show that for an arbitrary $\epsilon > 0$, $P(|X_{(n)} - \theta| < \epsilon) \to 1$ as $n \to \infty$. In this particular case, we can evalulate $P(|X_{(n)} - \theta| < \epsilon)$ directly by noting that $X_{(n)}$ cannot possibly be larger than θ, so that

$$P(|X_{(n)} - \theta| < \epsilon) = P(X_{(n)} > \theta - \epsilon) = 1 - P(X_{(n)} \leq \theta - \epsilon).$$

The maximum $X_{(n)}$ is less than some constant if and only if each of the random variables $X_1, \ldots, X_n$ is less than that constant. Therefore, since the X_i are i.i.d.,

$$P(X_{(n)} \leq \theta - \epsilon) = [P(X_1 \leq \theta - \epsilon)]^n = \begin{cases} [1 - (\epsilon/\theta)]^n & \text{if } 0 < \epsilon < \theta \\ 0 & \text{if } \epsilon \geq \theta. \end{cases}$$

Since $1 - (\epsilon/\theta)$ is strictly less than 1, we conclude that no matter what positive value ϵ takes, $P(X_n \leq \theta - \epsilon) \to 0$ as desired.

2.1.2 Probabilistic Order Notation

There are probabilistic analogues of the o and O notations of Section 1.3 that apply to random variable sequences instead of real number sequences.

Definition 2.4 We write $X_n = o_P(Y_n)$ if $X_n/Y_n \xrightarrow{P} 0$.

In particular, $o_P(1)$ is shorthand notation for a sequence of random variables that converges to zero in probability, as illustrated in Equation (2.2) below.

Definition 2.5 We write $X_n = O_P(Y_n)$ if for every $\epsilon > 0$, there exist M and N such that

$$P\left(\left|\frac{X_n}{Y_n}\right| < M\right) > 1 - \epsilon \text{ for all } n > N.$$

As a special case of Definition 2.5, we refer to any $O_P(1)$ sequence as a *bounded in probability* sequence:

Definition 2.6 We say that $X_1, X_2, \ldots$ is *bounded in probability* if $X_n = O_P(1)$, i.e., if for every $\epsilon > 0$, there exist M and N such that $P(|X_n| < M) > 1-\epsilon$ for $n > N$.

Definition 2.6 is primarily useful because of the properties of bounded in probability sequences established in Exercise 2.2.

Example 2.7 In Example 2.3, we showed that if $X_1, X_2, \ldots$ are independent and identically distributed uniform $(0, \theta)$ random variables, then

$$\max_{1 \le i \le n} X_i \xrightarrow{P} \theta \qquad \text{as } n \to \infty.$$

Equivalently, we may say that

$$\max_{1 \le i \le n} X_i = \theta + o_P(1) \qquad \text{as } n \to \infty. \tag{2.2}$$

It is also technically correct to write

$$\max_{1 \le i \le n} X_i = \theta + O_P(1) \qquad \text{as } n \to \infty, \tag{2.3}$$

though Statement (2.3) is less informative than Statement (2.2). On the other hand, we will see in Example 6.1 that Statement (2.3) may be sharpened considerably—and made more informative than Statement (2.2)—by writing

$$\max_{1 \le i \le n} X_i = \theta + O_P\left(\frac{1}{n}\right) \qquad \text{as } n \to \infty.$$

Using the o_P notation defined above, it is possible to rewrite Taylor's theorem 1.18 in a form involving random variables. This theorem will prove to be useful in later chapters; for instance, it is used to prove the result known as the delta method in Section 5.1.1.

Theorem 2.8 Suppose that $X_n \xrightarrow{P} \theta_0$ for a sequence of random variables $X_1, X_2, \ldots$ and a constant θ_0. Furthermore, suppose that $f(x)$ has d derivatives at the point θ_0. Then there is a random variable Y_n such that

$$f(X_n) = f(\theta_0) + (X_n - \theta_0)f'(\theta_0) + \cdots + \frac{(X_n - \theta_0)^d}{d!}\left\{f^{(d)}(\theta_0) + Y_n\right\} \quad (2.4)$$

and $Y_n = o_P(1)$ as $n \to \infty$.

The proof of Theorem 2.8 is a useful example of an "epsilon-delta" proof (named for the ϵ and δ in Definition 1.11).

Proof: Let

$$Y_n = \begin{cases} \frac{d!}{(X_n-\theta_0)^d}\left[f(X_n) - f(\theta_0) - (X_n - \theta_0)f'(\theta_0) - \cdots - \frac{(X_n-\theta_0)^{d-1}}{(d-1)!}\right] & \text{if } X_n \neq \theta_0 \\ 0 & \text{if } X_n = \theta_0. \end{cases}$$

Then Equation (2.4) is trivially satisfied. We will show that $Y_n = o_P(1)$, which means $Y_n \xrightarrow{P} 0$, by demonstrating that for an arbitrary $\epsilon > 0$, there exists N such that $P(|Y_n| < \epsilon) > 1 - \epsilon$ for all $n > N$. By Taylor's Theorem 1.18, there exists some $\delta > 0$ such that $|X_n - \theta_0| < \delta$ implies $|Y_n| < \epsilon$ (that is, the event $\{\omega : |X_n(\omega) - \theta_0| < \delta\}$ is contained in the event $\{\omega : |Y_n(\omega)| < \epsilon\}$). Furthermore, because $X_n \xrightarrow{P} \theta_0$, we know that there exists some N such that $P(|X_n - \theta_0| < \delta) > 1 - \epsilon$ for all $n > N$. Putting these facts together, we conclude that for all $n > N$,

$$P(|Y_n| < \epsilon) \geq P(|X_n - \theta_0| < \delta) > 1 - \epsilon,$$

which proves the result. ■

In later chapters, we will generally write simply

$$f(X_n) = f(\theta_0) + (X_n - \theta_0)f'(\theta_0) + \cdots + \frac{(X_n - \theta_0)^d}{d!}\left\{f^{(d)}(\theta_0) + o_P(1)\right\} \quad \text{as } n \to \infty \quad (2.5)$$

when referring to the result of Theorem 2.8. A technical quibble with Expression (2.5) is that it suggests that *any* random variable Y_n satisfying (2.4) must also be $o_P(1)$. This is not quite true: Since Y_n may be defined arbitrarily in the event that $X_n = \theta_0$ and still satisfy (2.4), if

$$P(X_n = \theta_0) > c \qquad \text{for all } n$$

for some positive constant c, then $Y_n \neq o_P(1)$ may still satisfy (2.4). However, as long as one remembers what Theorem 2.8 says, there is little danger in using Expression (2.5).

2.1.3 Convergence in Distribution

As the name suggests, convergence in distribution (also known as convergence in law) has to do with convergence of the distribution functions (or "laws") of random variables. Given a random variable X, the distribution function of X is the function

$$F(x) = P(X \leq x). \tag{2.6}$$

Any distribution function $F(x)$ is nondecreasing and right-continuous, and it has limits $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$. Conversely, any function $F(x)$ with these properties is a distribution function for some random variable.

It is not enough to define convergence in distribution as simple pointwise convergence of a sequence of distribution functions; there are technical reasons that such a simplistic definition fails to capture any useful concept of convergence of random variables. These reasons are illustrated by the following two examples.

Example 2.9 Let X_n be normally distributed with mean 0 and variance n. Then the distribution function of X_n is $F_n(x) = \Phi(x/\sqrt{n})$, where $\Phi(z)$ denotes the standard normal distribution function. Because $\Phi(0) = 1/2$, we see that for any fixed x, $F_n(x) \to 1/2$ as $n \to \infty$. But the function that is constant at $1/2$ is not a distribution function. This example shows that not all convergent sequences of distribution functions have limits that are distribution functions.

Example 2.10 By any sensible definition of convergence, $1/n$ should converge to 0 as $n \to \infty$. But consider the distribution functions $F_n(x) = I\{x \geq 1/n\}$ and $F(x) = I\{x \geq 0\}$ corresponding to the constant random variables $1/n$ and 0. We do *not* have pointwise convergence of $F_n(x)$ to $F(x)$, since $F_n(0) = 0$ for all n but $F(0) = 1$. However, $F_n(x) \to F(x)$ is true for all $x \neq 0$. Not coincidentally, the point $x = 0$ where convergence of $F_n(x)$ to $F(x)$ fails is the only point at which the function $F(x)$ is not continuous.

To write a sensible definition of convergence in distribution, Example 2.9 demonstrates that we should require that the limit of distribution functions be a distribution function itself, say $F(x)$, while Example 2.10 suggests that we should exclude points where $F(x)$ is not continuous. We therefore arrive at the following definition:

Definition 2.11 Suppose that X has distribution function $F(x)$ and that X_n has distribution function $F_n(x)$ for each n. Then we say X_n converges in distribution to X, written $X_n \xrightarrow{d} X$, if $F_n(x) \to F(x)$ as $n \to \infty$ for all x at which $F(x)$ is continuous. Convergence in distribution is sometimes called convergence in law and written $X_n \xrightarrow{\mathcal{L}} X$.

The notation of Definition 2.11 may be stretched a bit; sometimes the expressions on either side of the $\xrightarrow{d}$ symbol may be distribution functions or other notations indicating certain distributions, rather than actual random variables as in the definition. The meaning is always clear even if the notation is not consistent.

However, one common mistake should be avoided at all costs: If $\xrightarrow{d}$ (or $\xrightarrow{P}$ or any other "limit arrow") indicates that $n \to \infty$, then n *must never appear on the right side of the arrow.* See Expression (2.8) in Example 2.12 for an example of how this rule is sometimes violated.

Example 2.12 *The Central Limit Theorem for i.i.d. sequences:* Let $X_1, \ldots, X_n$ be independent and identically distributed (i.i.d.) with mean μ and finite variance σ^2. Then by a result that will be covered in Chapter 4 (but which is perhaps already known to the reader),

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right) \xrightarrow{d} N(0, \sigma^2), \tag{2.7}$$

where $N(0, \sigma^2)$ denotes a normal distribution with mean 0 and variance σ^2. [$N(0, \sigma^2)$ is not actually a random variable; this is an example of "stretching the $\xrightarrow{d}$ notation" referred to above.]

Because Equation 2.7 may be interpreted as saying that the sample mean $\overline{X}_n$ has approximately a $N(\mu, \sigma^2/n)$ distribution, it may seem tempting to "rewrite" Equation (2.7) as

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{d} N\left(\mu, \frac{\sigma^2}{n}\right). \tag{2.8}$$

Resist the temptation to do this! As pointed out above, n should never appear on the right side of a limit arrow (as long as that limit arrow expresses the idea that n is tending to ∞).

By the result of Exercise 2.2, the limit statement (2.7) implies that the left side of that statement is $O_P(1)$. We may therefore write (after dividing through by $\sqrt{n}$ and adding μ)

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \mu + O_P\left(\frac{1}{\sqrt{n}}\right) \qquad \text{as } n \to \infty. \tag{2.9}$$

Unlike Expression (2.8), Equation (2.9) is perfectly legal; and although it is less specific than Expression (2.7), it expresses at a glance the $\sqrt{n}$-rate of convergence of the sample mean to μ.

Unlike $X_n \xrightarrow{P} X$, the expression $X_n \xrightarrow{d} X$ does not require $X_n - X$ to be a random variable; in fact, $X_n \xrightarrow{d} X$ is possible even if X_n and X are not defined on the same sample space. Even if X_n and X do have a joint distribution, it is easy to construct an example in which $X_n \xrightarrow{d} X$ but X_n does not converge to X in probability: Take Z_1 and Z_2 to be independent and identically distributed standard normal random variables, then let $X_n = Z_1$ for all n and $X = Z_2$. Since X_n and X have *exactly* the same distribution by construction, $X_n \xrightarrow{d} X$ in this case. However, since $X_n - X$ is a $N(0, 2)$ random variable for all n, we do not have $X_n \xrightarrow{P} X$.

We conclude that $X_n \xrightarrow{d} X$ cannot possibly imply $X_n \xrightarrow{P} X$ (but see Theorem 2.14 for a special case in which it does). However, the implication in the other direction is always true:

Theorem 2.13 If $X_n \xrightarrow{P} X$, then $X_n \xrightarrow{d} X$.

Proof: Let $F_n(x)$ and $F(x)$ denote the distribution functions of X_n and X, respectively. Assume that $X_n \xrightarrow{P} X$. We need to show that $F_n(t) \to F(t)$, where t is any point of continuity of $F(x)$.

Choose any $\epsilon > 0$. Whenever $X_n \le t$, it must be true that either $X \le t + \epsilon$ or $|X_n - X| > \epsilon$. This implies that

$$F_n(t) \le F(t + \epsilon) + P(|X_n - X| > \epsilon).$$

Similarly, whenever $X \le t - \epsilon$, either $X_n \le t$ or $|X_n - X| > \epsilon$, implying

$$F(t - \epsilon) \le F_n(t) + P(|X_n - X| > \epsilon).$$

We conclude that for arbitrary n and $\epsilon > 0$,

$$F(t - \epsilon) - P(|X_n - X| > \epsilon) \le F_n(t) \le F(t + \epsilon) + P(|X_n - X| > \epsilon). \tag{2.10}$$

Taking both the $\liminf_n$ and the $\limsup_n$ of the above inequalities, we conclude [since $X_n \xrightarrow{P} X$ implies $P(|X_n - X| > \epsilon) \to 0$] that

$$F(t - \epsilon) \le \liminf_n F_n(t) \le \limsup_n F_n(t) \le F(t + \epsilon)$$

for all ϵ. Since t is a continuity point of $F(x)$, letting $\epsilon \to 0$ implies

$$F(t) = \liminf_n F_n(t) = \limsup_n F_n(t),$$

so we conclude $F_n(t) \to F(t)$ and the theorem is proved. ■

We remarked earlier that $X_n \xrightarrow{d} X$ could not possibly imply $X_n \xrightarrow{P} X$ because the latter expression requires that X_n and X be defined on the same sample space for every n. However, a constant c may be considered to be a random variable defined on any sample space; thus, it is reasonable to ask whether $X_n \xrightarrow{d} c$ implies $X_n \xrightarrow{P} c$. The answer is yes:

Theorem 2.14 $X_n \xrightarrow{d} c$ if and only if $X_n \xrightarrow{P} c$.

Proof: We only need to prove that $X_n \xrightarrow{d} c$ implies $X_n \xrightarrow{P} c$, since the other direction is a special case of Theorem 2.13. If $F(x)$ is the distribution function $I\{x \geq c\}$ of the constant random variable c, then $c + \epsilon$ and $c - \epsilon$ are points of continuity of $F(x)$ for any $\epsilon > 0$. Therefore, $X_n \xrightarrow{d} c$ implies that $F_n(c - \epsilon) \to F(c - \epsilon) = 0$ and $F_n(c + \epsilon) \to F(c + \epsilon) = 1$ as $n \to \infty$. We conclude that

$$P(-\epsilon < X_n - c \leq \epsilon) = F_n(c + \epsilon) - F_n(c - \epsilon) \to 1,$$

which means $X_n \xrightarrow{P} c$. ■

When we speak of convergence of random variables to a constant in this book, most commonly we refer to convergence in probability, which (according to Theorem 2.14) is equivalent to convergence in distribution. On the other hand, when we speak of convergence to a random variable, we nearly always refer to convergence in distribution. Therefore, in a sense, Theorem 2.14 makes convergence in distribution the most important form of convergence in this book. This type of convergence is often called "weak convergence".

2.1.4 Convergence in Mean

The third and final mode of convergence in this chapter is useful primarily because it is sometimes easy to verify and thus gives a quick way to prove convergence in probability, as Theorem 2.17 below implies.

Definition 2.15 Let a be a positive constant. We say that X_n converges in ath mean to X, written $X_n \xrightarrow{a} X$, if

$$\mathrm{E}\ |X_n - X|^a \to 0 \text{ as } n \to \infty. \tag{2.11}$$

Two specific cases of Definition 2.15 deserve special mention. When $a = 1$, we normally omit mention of the a and simply refer to the condition $\mathrm{E}\ |X_n - X| \to 0$ as *convergence in mean*. Convergence in mean is *not* equivalent to $\mathrm{E}\ X_n \to \mathrm{E}\ X$: For one thing, $\mathrm{E}\ X_n \to \mathrm{E}\ X$ is possible without any regard to the joint distribution of X_n and X, whereas $\mathrm{E}\ |X_n - X| \to 0$ clearly requires that $X_n - X$ be a well-defined random variable.

Even more important than $a = 1$ is the special case $a = 2$:

Definition 2.16 We say that X_n converges in quadratic mean to X, written $X_n \xrightarrow{\mathrm{qm}} X$, if

$$\mathrm{E}\ |X_n - X|^2 \to 0 \text{ as } n \to \infty.$$

Convergence in quadratic mean is important for two reasons. First, it is often quite easy to check; in Exercise 2.6, you are asked to prove that $X_n \xrightarrow{\text{qm}} c$ if and only if $\text{E } X_n \to c$ and $\text{Var } X_n \to 0$ for some constant c. Second, quadratic mean convergence (indeed, ath mean convergence for *any* $a > 0$) is stronger than convergence in probability, which means that weak consistency of an estimator may be established by checking that it converges in quadratic mean. This latter property is a corollary of the following result:

Theorem 2.17 **(a)** For a constant c, $X_n \xrightarrow{\text{qm}} c$ if and only if $\text{E } X_n \to c$ and $\text{Var } X_n \to 0$.

(b) For fixed $a > 0$, $X_n \xrightarrow{a} X$ implies $X_n \xrightarrow{P} X$.

Proof: Part (a) is the subject of Exercise 2.6. Part (b) relies on Markov's inequality (1.35), which states that

$$P(|X_n - X| \geq \epsilon) \leq \frac{1}{\epsilon^a} \text{E } |X_n - X|^a \tag{2.12}$$

for an arbitrary fixed $\epsilon > 0$. If $X_n \xrightarrow{a} X$, then by definition the right hand side of inequality (2.12) goes to zero as $n \to \infty$, so the left side also goes to zero and we conclude that $X_n \xrightarrow{P} X$ by definition. ■

Example 2.18 Any unbiased estimator is consistent if its variance goes to zero. This fact follows directly from Theorem 2.17(a) and (b). As an example, consider a sequence of independent and identically distributed random variables $X_1, X_2, \ldots$ with mean μ and finite variance σ^2. The sample mean

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

has mean μ and variance σ^2/n. Therefore, $\overline{X}_n$ is unbiased and its variance goes to zero, so we conclude that it is consistent; i.e., $\overline{X}_n \xrightarrow{P} \mu$. This fact is the Weak Law of Large Numbers (see Theorem 2.19) for the case of random variables with finite variance.

Exercises for Section 2.1

Exercise 2.1 For each of the three cases below, prove that $X_n \xrightarrow{P} 1$:

(a) $X_n = 1 + nY_n$, where Y_n is a Bernoulli random variable with mean $1/n$.

(b) $X_n = Y_n / \log n$, where Y_n is a Poisson random variable with mean $\sum_{i=1}^{n} (1/i)$.

(c) $X_n = \frac{1}{n} \sum_{i=1}^{n} Y_i^2$, where the Y_i are independent standard normal random variables.

Exercise 2.2 This exercise deals with bounded in probability sequences; see Definition 2.6.

(a) Prove that if $X_n \xrightarrow{d} X$ for some random variable X, then X_n is bounded in probability.

Hint: You may use the fact that any interval of real numbers must contain a point of continuity of $F(x)$. Also, recall that $F(x) \to 1$ as $x \to \infty$.

(b) Prove that if X_n is bounded in probability and $Y_n \xrightarrow{P} 0$, then $X_n Y_n \xrightarrow{P} 0$.

Hint: For fixed $\epsilon > 0$, argue that there must be M and N such that $P(|X_n| < M) > 1 - \epsilon/2$ and $P(|Y_n| < \epsilon/M) > 1 - \epsilon/2$ for all $n > N$. What is then the smallest possible value of $P(|X_n| < M \text{ and } |Y_n| < \epsilon/M)$? Use this result to prove $X_n Y_n \xrightarrow{P} 0$.

Exercise 2.3 *The Poisson approximation to the binomial:*

(a) Suppose that X_n is a binomial random variable with n trials, where the probability of success on each trial is λ/n. Let X be a Poisson random variable with the same mean as X_n, namely λ. Prove that $X_n \xrightarrow{d} X$.

Hint: Argue that it suffices to show that $P(X_n = k) \to P(X = k)$ for all nonnegative integers k. Then use Stirling's formula (1.19).

(b) Part (a) can be useful in approximating binomial probabilities in cases where the number of trials is large but the success probability is small: Simply consider a Poisson random variable with the same mean as the binomial variable. Assume that X_n is a binomial random variable with parameters n and $2/n$. Create a plot on which you plot $P(X_{10} = k)$ for $k = 0, \ldots, 10$. On the same set of axes, plot the same probabilities for X_{20}, X_{50}, and the Poisson variable we'll denote by X_∞. Try looking at the same plot but with the probabilities transformed using the logit (log-odds) transformation $\text{logit}(t) = \log(t) - \log(1 - t)$. Which plot makes it easier to characterize the trend you observe?

Exercise 2.4 Suppose that $X_1, \ldots, X_n$ are independent and identically distributed Uniform$(0, 1)$ random variables. For a real number t, let

$$G_n(t) = \sum_{i=1}^{n} I\{X_i \leq t\}.$$

(a) What is the distribution of $G_n(t)$ if $0 < t < 1$?

(b) Suppose $c > 0$. Find the distribution of a random variable X such that

$G_n(c/n) \xrightarrow{d} X$. Justify your answer.

(c) How does your answer to part (b) change if $X_1, \ldots, X_n$ are from a standard exponential distribution instead of a uniform distribution? The standard exponential distribution function is $F(t) = 1 - e^{-t}$.

Exercise 2.5 For each of the three examples in Exercise 2.1, does $X_n \xrightarrow{\text{qm}} 1$? Justify your answers.

Exercise 2.6 Prove Theorem 2.17(a).

Exercise 2.7 The converse of Theorem 2.17(b) is not true. Construct a counterexample in which $X_n \xrightarrow{P} 0$ but $\text{E } X_n = 1$ for all n (by Theorem 2.17, if $\text{E } X_n = 1$, then X_n cannot converge in quadratic mean to 0).

Hint: The mean of a random variable may be strongly influenced by a large value that occurs with small probability (and if this probability goes to zero, then the mean can be influenced in this way without destroying convergence in probability).

Exercise 2.8 Prove or disprove this statement: If there exists M such that $P(|X_n| < M) = 1$ for all n, then $X_n \xrightarrow{P} c$ implies $X_n \xrightarrow{\text{qm}} c$.

Exercise 2.9 **(a)** Prove that if $0 < a < b$, then convergence in bth mean is stronger than convergence in ath mean; i.e., $X_n \xrightarrow{b} X$ implies $X_n \xrightarrow{a} X$.

Hint: Use Exercise 1.40 with $\alpha = b/a$.

(b) Prove by counterexample that the conclusion of part (a) is not true in general if $0 < b < a$.

2.2 Consistent Estimates of the Mean

For a sequence of random vectors $X_1, X_2, \ldots$, we denote the nth sample mean by

$$\overline{X}_n \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} X_i.$$

We begin with a formal statement of the weak law of large numbers for an independent and identically distributed sequence. Later, we discuss some cases in which the sequence of random vectors is not independent and identically distributed.

2.2.1 The Weak Law of Large Numbers

Theorem 2.19 *Weak Law of Large Numbers (univariate version):* Suppose that $X_1, X_2, \ldots$ are independent and identically distributed and have finite mean μ. Then $\overline{X}_n \xrightarrow{P} \mu$.

The proof of Theorem 2.19 in its full generality is beyond the scope of this chapter, though it may be proved using the tools in Section 4.1. However, by tightening the assumptions a bit, a proof can be made simple. For example, if the X_i are assumed to have finite variance (not a terribly restrictive assumption), the weak law may be proved in a single line: Chebyshev's inequality (1.36) implies that

$$P\left(|\overline{X}_n - \mu| \geq \epsilon\right) \leq \frac{\text{Var } \overline{X}_n}{\epsilon^2} = \frac{\text{Var } X_1}{n\epsilon^2} \to 0,$$

so $\overline{X}_n \xrightarrow{P} \mu$ follows by definition. (This is an alternative proof of the same result in Example 2.18.)

Example 2.20 If $X \sim \text{binomial}(n, p)$, then $X/n \xrightarrow{P} p$. Although we could prove this fact directly using the definition of convergence in probability, it follows immediately from the Weak Law of Large Numbers due to the fact that X/n is the sample mean of n independent and identically distributed Bernoulli random variables, each with mean p.

Example 2.21 Suppose that $X_1, X_2, \ldots$ are independent and identically distributed with mean μ and finite variance σ^2. Then the estimator

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

is consistent for σ^2 because of the Weak Law of Large Numbers: The $(X_i - \mu)^2$ are independent and identically distributed and they have mean σ^2.

Ordinarily, of course, we do not know μ, so we replace μ by $\overline{X}_n$ (and often replace $\frac{1}{n}$ by $\frac{1}{n-1}$) to obtain the sample variance. We need a bit more theory to establish the consistency of the sample variance, but we will revisit this topic later.

2.2.2 Independent but not Identically Distributed Variables

Let us now generalize the conditions of the previous section: Suppose that $X_1, X_2, \ldots$ are independent but not necessarily identically distributed but have at least the same mean, so

that E $X_i = \mu$ and Var $X_i = \sigma_i^2$. When is $\overline{X}_n$ consistent for μ? Since $\overline{X}_n$ is unbiased, the result of Example 2.18 tells us that we may conclude that $\overline{X}_n$ is consistent as long as the variance of $\overline{X}_n$ tends to 0 as $n \to \infty$. (However, $\overline{X}_n \xrightarrow{P} \mu$ does not imply Var $\overline{X}_n \to 0$; see Exercise 2.10.) Since

$$\text{Var } \overline{X}_n = \frac{1}{n^2} \sum_{i=1}^{n} \sigma_i^2, \tag{2.13}$$

we conclude that $\overline{X}_n \xrightarrow{P} \mu$ if $\sum_{i=1}^n \sigma_i^2 = o(n^2)$.

What about alternatives to the sample mean? Suppose we restrict attention to *weighted* mean estimators of the form

$$\hat{\mu}_n = \frac{\sum_{i=1}^n c_i X_i}{\sum_{i=1}^n c_i}$$

for some sequence of positive constants $c_1, c_2, \ldots$. The $\hat{\mu}_n$ estimator above is unbiased, so we consider whether its variance tends to zero. By independence, we may write

$$\text{Var } \hat{\mu}_n = \frac{\sum_{i=1}^n c_i^2 \sigma_i^2}{\left(\sum_{i=1}^n c_i\right)^2}.$$

How may we obtain the smallest possible variance for $\hat{\mu}_n$? To find the answer, we may set $\gamma_i = c_i / \sum_{i=1}^n c_i$ and finding partial derivatives of Var $\hat{\mu}_n$ with respect to $\gamma_1, \ldots, \gamma_{n-1}$ (after making the substitution $\gamma_n = 1 - \gamma_1 - \cdots - \gamma_{n-1}$). Setting these partial derivatives equal to zero gives the equations

$$\gamma_i \sigma_i^2 = \gamma_n \sigma_n^2 \qquad \text{for } 1 \le i \le n-1.$$

After checking to ensure that the solution is indeed a minimizer of the variance, we conclude that Var $\hat{\mu}_n$ is minimized when each c_i is proportional to $1/\sigma_i^2$. Thus, the variance is minimized by

$$\delta_n = \frac{\sum_{i=1}^n X_i/\sigma_i^2}{\sum_{j=1}^n 1/\sigma_j^2}, \tag{2.14}$$

which attains the variance

$$\text{Var } \delta_n = \frac{1}{\sum_{j=1}^n 1/\sigma_j^2}. \tag{2.15}$$

An interesting fact about n Var $\overline{X}_n$ and n Var δ_n is that they are, respectively, the arithmetic and harmonic means of $\sigma_1^2, \ldots, \sigma_n^2$. In other words, our conclusion that Var $\delta_n \le$ Var $\overline{X}_n$,

with equality only when the σ_i^2 are all equal, is simply a restatement of a well-known mathematical inequality relating the harmonic and arithmetic means! See Exercise 2.11 for a particularly compelling demonstration of the discrepancy between these two means.

Suppose for a sequence of independent random variables that instead of estimating their mean μ, we wish to estimate their *conditional* mean, given a covariate. This is exactly the case in regression:

Example 2.22 In the case of simple linear regression, let

$$Y_i = \beta_0 + \beta_1 z_i + \epsilon_i,$$

where we assume the z_i are known covariates and the ϵ_i are independent and identically distributed with mean 0 and finite variance σ^2. (This implies that the Y_i are independent but not identically distributed.) If we define

$$w_i^{(n)} = \frac{z_i - \overline{z}_n}{\sum_{j=1}^n (z_j - \overline{z}_n)^2} \quad \text{and} \quad v_i^{(n)} = \frac{1}{n} - \overline{z}_n w_i^{(n)},$$

then the least squares estimators of β_0 and β_1 are

$$\hat{\beta}_{0n} = \sum_{i=1}^n v_i^{(n)} Y_i \quad \text{and} \quad \hat{\beta}_{1n} = \sum_{i=1}^n w_i^{(n)} Y_i, \tag{2.16}$$

respectively. One may prove, as in Exercise 2.14(a), that $\hat{\beta}_{0n}$ and $\hat{\beta}_{1n}$ are unbiased estimators of β_0 and β_1, so Example 2.18 tells us that they are consistent as long as their variances tend to zero as $n \to \infty$. It is therefore possible to show, as in Exercise 2.14(b), that $\hat{\beta}_{0n}$ is consistent if

$$\frac{\overline{z}_n^2}{\sum_{j=1}^n (z_j - \overline{z}_n)^2} \to 0 \tag{2.17}$$

and $\hat{\beta}_{1n}$ is consistent if

$$\frac{1}{\sum_{j=1}^n (z_j - \overline{z}_n)^2} \to 0. \tag{2.18}$$

2.2.3 Identically Distributed but not Independent Variables

Suppose that $X_1, X_2, \ldots$ have the same mean, say μ, but that they are not necessarily independent. Since the unbiasedness of $\overline{X}_n$ does not rely on independence, we still have $\mathrm{E}\, \overline{X}_n = \mu$, so $\overline{X}_n$ is consistent if its variance tends to zero. A direct calculation gives

$$\mathrm{Var}\, \overline{X}_n = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \mathrm{Cov}\,(X_i, X_j). \tag{2.19}$$

Suppose we also assume that the X_i are identically distributed. In the presence of independence (the first "i" in "i.i.d"), the "i.d." assumption is sufficient to specify the joint distribution of the X_i. However, in the case of dependent random variables, assuming they are identically distributed (with, say, variance σ^2) allows only a small simplification of Equation (2.19):

$$\text{Var } \overline{X}_n = \frac{\sigma^2}{n} + \frac{2}{n^2} \sum \sum_{i<j} \text{Cov}\,(X_i, X_j).$$

In order to deal with the $\binom{n}{2}$ covariances above, all of which could in principle be distinct, it would help to make some additional assumptions beyond "identically distributed". One possibility is to assume that the $X_1, X_2, \ldots$ are *exchangeable*:

Definition 2.23 Let π denote an arbitrary permutation on n elements (that is, a function that maps $\{1, \ldots, n\}$ onto itself). The finite sequence $X_1, \ldots, X_n$ is said to be *exchangeable* if the joint distribution of the permuted random vector $(X_{\pi(1)}, \ldots, X_{\pi(n)})$ is the same no matter which π is chosen. The infinite sequence $X_1, X_2, \ldots$ is said to be exchangeable if any finite subsequence is exchangeable.

Under exchangeability, the covariance between X_i and X_j is always the same, say $\text{Cov}\,(X_1, X_2)$, when $i \neq j$. Therefore, Equation (2.19) reduces to

$$\text{Var } \overline{X}_n = \frac{\sigma^2}{n} + \frac{(n-1)\,\text{Cov}\,(X_1, X_2)}{n},$$

and we conclude that exchangeability implies $\text{Var } \overline{X}_n \to \text{Cov}\,(X_1, X_2)$ as $n \to \infty$. Since this is a nonzero limit unless the X_i are pairwise uncorrelated, exchangeability appears to be too stringent a condition to place on the X_i in the context of searching for consistent estimators of μ.

Thus, we turn to a weaker concept than exchangeability:

Definition 2.24 The sequence $X_1, X_2, \ldots$ is said to be *stationary* if, for a fixed $k \geq 0$, the joint distribution of $(X_i, \ldots, X_{i+k})$ is the same no matter what positive value of i is chosen.

We see that i.i.d. implies exchangeability, which implies stationarity, which implies identically distributed. To obtain an interesting simplification of Equation (2.19), it turns out that stationarity is just about the right level in this hierarchy.

Under stationarity, $\text{Cov}\,(X_i, X_j)$ depends only on the "gap" $j - i$. For example, stationarity implies $\text{Cov}\,(X_1, X_4) = \text{Cov}\,(X_2, X_5) = \text{Cov}\,(X_5, X_8) = \cdots$. Therefore, Equation (2.19)

becomes

$$\text{Var } \overline{X}_n \;=\; \frac{\sigma^2}{n} + \frac{2}{n^2}\sum_{k=1}^{n-1}(n-k)\,\text{Cov}\,(X_1, X_{1+k}). \tag{2.20}$$

Lemma 2.25 Expression (2.20) tends to 0 as $n \to \infty$ if $\sigma^2 < \infty$ and $\text{Cov}\,(X_1, X_{1+k}) \to 0$ as $k \to \infty$.

Proof: It is immediate that $\sigma^2/n \to 0$ if $\sigma^2 < \infty$. Assuming that $\text{Cov}\,(X_1, X_{1+k}) \to 0$, select $\epsilon > 0$ and note that if N is chosen so that $|\,\text{Cov}\,(X_1, X_{1+k})| < \epsilon/2$ for all $k > N$, we have

$$\left|\frac{2}{n^2}\sum_{k=1}^{n-1}(n-k)\,\text{Cov}\,(X_1, X_{1+k})\right| \le \frac{2}{n}\sum_{k=1}^{N}|\text{Cov}\,(X_1, X_{1+k})| + \frac{2}{n}\sum_{k=N+1}^{n-1}|\text{Cov}\,(X_1, X_{1+k})|\,.$$

The second term on the right is strictly less than $\epsilon/2$, and the first term is a constant divided by n, which may be made smaller than $\epsilon/2$ by choosing n large enough. (This lemma is also a corollary of the result stated in Exercise 1.3.) ■

Because of Lemma 2.25, it is sensible to impose conditions guaranteeing that $\text{Cov}\,(X_1, X_{1+k})$ tends to zero as $k \to \infty$. For instance, we might consider sequences for which $\text{Cov}\,(X_1, X_{1+k})$ is exactly equal to zero for all k larger than some cutoff value, say m. This is the idea of m-dependence:

Definition 2.26 For a fixed nonnegative integer m, the sequence $X_1, X_2, \ldots$ is called m-dependent if the random vectors $(X_1, \ldots, X_i)$ and $(X_j, X_{j+1}, \ldots)$ are independent whenever $j - i > m$.

Any stationary m-dependent sequence trivially satisfies $\text{Cov}\,(X_1, X_{1+k}) \to 0$ as $k \to \infty$, so by Lemma 2.25, $\overline{X}_n$ is consistent for any stationary m-dependent sequence with finite variance. As a special case of m-dependence, any independent sequence is 0-dependent.

Exercises for Section 2.2

Exercise 2.10 The goal of this Exercise is to construct an example of an independent sequence $X_1, X_2, \ldots$ with $\text{E } X_i = \mu$ such that $\overline{X}_n \xrightarrow{P} \mu$ but $\text{Var } \overline{X}_n$ does not converge to 0. There are numerous ways we could proceed, but let us suppose that for some positive constants c_i and p_i, $X_i = c_i Y_i(2Z_i - 1)$, where Y_i and Z_i are independent Bernoulli random variables with $\text{E } Y_i = p_i$ and $\text{E } Z_i = 1/2$.

(a) Verify that $\text{E } X_i = 0$ and find $\text{Var } \overline{X}_n$.

(b) Show that $\overline{X}_n \xrightarrow{P} 0$ if

$$\frac{1}{n}\sum_{i=1}^{n} c_i p_i \to 0. \tag{2.21}$$

Hint: Use the triangle inequality to show that if Condition (2.21) is true, then $\overline{X}_n$ converges in mean to 0 (see Definition 2.15).

(c) Now specify c_i and p_i so that Var $\overline{X}_n$ does not converge to 0 but Condition (2.21) holds. Remember that p_i must be less than or equal to 1 because it is the mean of a Bernoulli random variable.

Exercise 2.11 Suppose that $X_1, X_2, \ldots$ are independent with mean zero and Var $X_i = (i+1)\log(i+1)$. Let δ_n be the minimum variance linear estimator defined in Equation (2.14) and let $\overline{X}_n$ denote the sample mean. Find the relative efficiency of δ_n with respect to $\overline{X}_n$ (defined as Var $\overline{X}_n$/ Var δ_n) for $n = 10^k$, $k = 1, \ldots, 6$. What seems to be happening? Find, with proof, the limits of Var $\overline{X}_n$ and Var δ_n as $n \to \infty$ to try to verify your conjecture.

Exercise 2.12 Suppose $X_1, X_2, \ldots$ are independent and identically distributed with mean μ and finite variance σ^2. Let $Y_i = \overline{X}_i = (\sum_{j=1}^{i} X_j)/i$.

(a) Prove that $\overline{Y}_n = (\sum_{i=1}^{n} Y_i)/n$ is a consistent estimator of μ.

(b) Compute the relative efficiency $e_{\overline{Y}_n, \overline{X}_n}$ of $\overline{Y}_n$ to $\overline{X}_n$, defined as Var $(\overline{X}_n)$/ Var $(\overline{Y}_n)$, for $n \in \{5, 10, 20, 50, 100, \infty\}$ and report the results in a table. For $n = \infty$, give the limit (with proof) of the efficiency.

Exercise 2.13 Let $Y_1, Y_2, \ldots$ be independent and identically distributed with mean μ and variance $\sigma^2 < \infty$. Let

$$X_1 = Y_1, \quad X_2 = \frac{Y_2 + Y_3}{2}, \quad X_3 = \frac{Y_4 + Y_5 + Y_6}{3}, \quad \text{etc.}$$

Define δ_n as in Equation (2.14).

(a) Show that δ_n and $\overline{X}_n$ are both consistent estimators of μ.

(b) Calculate the relative efficiency $e_{\overline{X}_n, \delta_n}$ of $\overline{X}_n$ to δ_n, defined as Var (δ_n)/ Var $(\overline{X}_n)$, for $n = 5, 10, 20, 50, 100$, and ∞ and report the results in a table. For $n = \infty$, give the limit (with proof) of the efficiency.

(c) Using Example 1.23, give a simple expression asymptotically equivalent to $e_{\overline{X}_n, \delta_n}$. Report its values in your table for comparison. How good is the approximation for small n?

Exercise 2.14 Consider the case of simple linear regression in Example 2.22.

(a) Prove that the least squares regression estimators defined in equation (2.16) are unbiased. In other words, show that $\mathrm{E}\ \hat{\beta}_{0n} = \beta_0$ and $\mathrm{E}\ \hat{\beta}_{1n} = \beta_1$.

Hint: Prove and use the facts that $\sum_{i=1}^{n} w_i^{(n)} = 0$ and $\sum_{i=1}^{n} w_i^{(n)} z_i = 1$.

(b) Prove consistency of $\hat{\beta}_{0n}$ and $\hat{\beta}_{1n}$ under conditions (2.17) and (2.18), respectively.

2.3 Convergence of Transformed Sequences

Many statistical estimators of interest may be written as functions of simpler statistics whose convergence properties are known. Therefore, results that describe the behavior of transformed sequences have central importance for the study of statistical large-sample theory. We begin with some results about continuous transformations of univariate random variable sequences. Yet the important result near the end of this section, called Slutsky's theorem, is intrinsically *multivariate* in nature. For this reason, after presenting a few results on continuous transformations, we will extend these and other univariate concepts from earlier in this chapter to the k-dimensional setting for $k > 1$.

2.3.1 Continuous Transformations: The Univariate Case

Just as they do for sequences of real numbers, continuous functions preserve convergence of sequences of random variables. We state this result formally for both convergence in probability and convergence in distribution.

Theorem 2.27 Suppose that $f(x)$ is a continuous function.

(a) If $X_n \xrightarrow{P} X$, then $f(X_n) \xrightarrow{P} f(X)$.

(b) If $X_n \xrightarrow{d} X$, then $f(X_n) \xrightarrow{d} f(X)$.

Theorem 2.27 is the random-variable analogue of Theorem 1.16, which is proved using a straightforward ϵ-δ argument. It is therefore surprising that proving Theorem 2.27 is quite difficult. Indeed, each of its two statements relies on an additional theorem for its proof. For statement (a), this additional theorem (Theorem 3.10) involves almost sure convergence, a mode of convergence not defined until Chapter 3. Statement (b) about convergence in distribution, on the other hand, follows from a powerful characterization of convergence in distribution (its proof is the subject of Exercises 2.15 and 2.16).

Theorem 2.28 $X_n \xrightarrow{d} X$ if and only if E $g(X_n) \to$ E $g(X)$ for all bounded and continuous real-valued functions $g(x)$.

The forward half of Theorem 2.28, the fact that $X_n \xrightarrow{d} X$ implies E $g(X_n) \to$ E $g(X)$ if $g(x)$ is bounded and continuous, is called the Helly-Bray theorem. Taken as a whole, the theorem establishes a condition equivalent to convergence in distribution, and in fact this condition is sometimes used as the *definition* of $X_n \xrightarrow{d} X$. It is very important to remember that Theorem 2.28 does not say that $X_n \xrightarrow{d} X$ implies E $X_n \to$ E X (because the function $g(t) = t$, while certainly continuous, is not bounded). In fact, the special conditions under which $X_n \xrightarrow{d} X$ implies E $X_n \to$ E X are the subject of Section 3.3.

Using Theorem 2.28, Theorem 2.27(b) follows quickly.

Proof of Theorem 2.27(b) Let $g(x)$ be any bounded and continuous function. By Theorem 2.28, it suffices to show that E $g[f(X_n)] \to$ E $g[f(X)]$. Since $f(x)$ is continuous, the composition $x \mapsto g[f(x)]$ is bounded and continuous. Therefore, another use of Theorem 2.28 proves that E $g[f(X_n)] \to$ E $g[f(X)]$ as desired. ■

2.3.2 Multivariate Extensions

We now extend our notions of random-vector convergence to the multivariate case. Several earlier results from this chapter, such as the weak law of large numbers and the results on continuous functions, generalize immediately to this case. Note the use of bold type to signify random vectors: Whereas X_n and X denote univariate random variables, the possibly-multidimensional analogues are $\mathbf{X}_n$ and $\mathbf{X}$.

A k-dimensional random vector is a function $\mathbf{X}(\omega)$, usually abbreviated $\mathbf{X}$, that maps a probability space Ω into k-dimensional Euclidean space $\mathbb{R}^k$. As we remarked in Section 1.5, it is not possible to develop a coherent theory if we consider *all possible* functions $\mathbf{X}(\omega)$ to be random vectors; therefore, strictly speaking we must restrict attention only to *measurable* functions $\mathbf{X}(\omega)$. Yet a reasonable treatment of measurability is beyond the scope of this book, and instead of delving into technicalities we rest assured that basically every interesting function $\mathbf{X}(\omega)$ is a legitimate random variable. (Indeed, it is a fairly challenging mathematical exercise to construct a nonmeasurable function.)

The multivariate definitions of convergence in probability and convergence in ath mean are both based on the sequence $\|\mathbf{X}_n - \mathbf{X}\|$, which is a *univariate* sequence, so they are straightforward and require no additional development:

Definition 2.29 $\mathbf{X}_n$ converges in probability to $\mathbf{X}$ (written $\mathbf{X}_n \xrightarrow{P} \mathbf{X}$) if for any $\epsilon > 0$,

$$P\left(\|\mathbf{X}_n - \mathbf{X}\| < \epsilon\right) \to 1 \text{ as } n \to \infty.$$

Definition 2.30 For $a > 0$, $\mathbf{X}_n$ converges in ath mean to $\mathbf{X}$ (written $\mathbf{X}_n \xrightarrow{a} \mathbf{X}$) if

$$\mathrm{E}\,\|\mathbf{X}_n - \mathbf{X}\|^a \to 0 \text{ as } n \to \infty.$$

Convergence in quadratic mean, written $\mathbf{X}_n \xrightarrow{\text{qm}} \mathbf{X}$, is the special case of Definition 2.30 when $a = 2$. Since $\|\mathbf{c}\|^2 = \mathbf{c}^\top\mathbf{c}$, $\mathbf{X}_n$ converges in quadratic mean to $\mathbf{X}$ if

$$\mathrm{E}\left[(\mathbf{X}_n - \mathbf{X})^\top(\mathbf{X}_n - \mathbf{X})\right] \to 0 \text{ as } n \to \infty. \tag{2.22}$$

Because Definitions 2.29 and 2.30 rely only on the *univariate* random variables $\|\mathbf{X}_n - \mathbf{X}\|$, Theorem 2.17 immediately implies that (a) $\mathbf{X}_n \xrightarrow{\text{qm}} \mathbf{c}$ if and only if $\mathrm{E}\,\|\mathbf{X}_n - \mathbf{c}\| \to 0$ and $\mathrm{Var}\,\|\mathbf{X}_n - \mathbf{c}\| \to 0$; and (b) if $\mathbf{X}_n \xrightarrow{a} \mathbf{X}$, then $\mathbf{X}_n \xrightarrow{P} \mathbf{X}$. However, fact (a) is less useful than in the univariate setting; consider the comments following the proof of Theorem 2.31 below.

As an immediate application of the above results, we may extend the weak law of large numbers to the multivariate case. For a sequence $\mathbf{X}_1, \mathbf{X}_2, \ldots$, define the nth sample mean to be

$$\overline{\mathbf{X}}_n \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i.$$

Theorem 2.31 *The Weak Law of Large Numbers:* Suppose that $\mathbf{X}_1, \mathbf{X}_2, \ldots$ are independent and identically distributed and have finite mean $\boldsymbol{\mu}$. Then $\overline{\mathbf{X}}_n \xrightarrow{P} \boldsymbol{\mu}$.

Partial Proof: We do not prove this theorem in full generality until Section 4.1. However, in the special (and very common) case in which the $k \times k$ covariance matrix $\Sigma = \mathrm{Var}\,\mathbf{X_1}$ has only finite entries,

$$\begin{aligned}\mathrm{E}\,(\overline{\mathbf{X}}_n - \boldsymbol{\mu})^\top(\overline{\mathbf{X}}_n - \boldsymbol{\mu}) &= \frac{1}{n^2}\,\mathrm{E}\left[\sum_{i=1}^{n}(\mathbf{X}_i - \boldsymbol{\mu})\right]^\top\left[\sum_{i=1}^{n}(\mathbf{X}_i - \boldsymbol{\mu})\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{E}\,(\mathbf{X}_i - \boldsymbol{\mu})^\top(\mathbf{X}_i - \boldsymbol{\mu}) = \frac{1}{n}\mathrm{Tr}(\Sigma) \to 0\end{aligned}$$

as $n \to \infty$. Therefore, $\overline{\mathbf{X}}_n \xrightarrow{\text{qm}} \boldsymbol{\mu}$ by definition, which implies that $\overline{\mathbf{X}}_n \xrightarrow{P} \boldsymbol{\mu}$. ■

It is instructive to compare the proof outlined in the univariate Example 2.18 to the multivariate proof above because the former method may not be adapted to the latter situation. It is still true that any unbiased estimator whose covariance matrix converges to the zero matrix is consistent [by Equation (2.22) with $\mathbf{X}$ replaced by $\mathrm{E}\,\mathbf{X}_n$], but an argument for this cannot easily be based on Theorem 2.17(a): The fact that an estimator like $\overline{\mathbf{X}}_n$ is unbiased for $\boldsymbol{\mu}$ does *not* immediately imply that $\mathrm{E}\,\left\|\overline{\mathbf{X}}_n - \boldsymbol{\mu}\right\| = 0$.

To extend convergence in distribution to random vectors, we need the multivariate analogue of Equation (2.6), the distribution function. To this end, let

$$F(\mathbf{x}) \stackrel{\text{def}}{=} P(\mathbf{X} \leq \mathbf{x}),$$

where $\mathbf{X}$ is a random vector in $\mathbb{R}^d$ and $\mathbf{X} \leq \mathbf{x}$ means that $X_i \leq x_i$ for all $1 \leq i \leq d$.

Definition 2.32 $\mathbf{X}_n$ converges in distribution to $\mathbf{X}$ (written $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$) if for any point $\mathbf{c}$ at which $F(\mathbf{x})$ is continuous,

$$F_n(\mathbf{c}) \to F(\mathbf{c}) \text{ as } n \to \infty.$$

There is one subtle way in which the multivariate situation is not quite the same as the univariate situation for distribution functions. In the univariate case, it is very easy to characterize the points of continuity of $F(x)$: The distribution function of the univariate random variable X is continuous at x if and only if $P(X = x) = 0$. However, this simple characterization no longer holds true for random vectors; a point $\mathbf{x}$ may be a point of discontinuity yet still satisfy $P(\mathbf{X} = \mathbf{x}) = 0$. The task in Exercise 2.18 is to produce an example of this phenomenon.

We may now extend Theorems 2.13, 2.14, and 2.27 to the multivariate case. The proofs of these results do not differ substantially from their univariate counterparts; all necessary modifications are straightforward. For instance, in the proofs of Theorems 2.13 and 2.14, the scalar ϵ should be replaced by the vector $\boldsymbol{\epsilon} = \epsilon\mathbf{1}$, each of whose entries equals ϵ; with this change, modified statements such as "whenever $\mathbf{X}_n \leq \mathbf{t}$, it must be true that either $\mathbf{X} \leq \mathbf{t} + \boldsymbol{\epsilon}$ or $\|\mathbf{X}_n - \mathbf{X}\| > \epsilon$" remain true.

Theorem 2.33 $\mathbf{X}_n \xrightarrow{P} \mathbf{X}$ implies $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$. Furthermore, if $\mathbf{c}$ is a constant, then $\mathbf{X}_n \xrightarrow{P} \mathbf{c}$ if and only if $\mathbf{X}_n \xrightarrow{d} \mathbf{c}$.

Theorem 2.34 Suppose that $\mathbf{f} : S \to \mathbb{R}^\ell$ is a continuous function defined on some subset $S \subset \mathbb{R}^k$, $\mathbf{X}_n$ is a k-component random vector, and $P(\mathbf{X} \in S) = 1$.

(a) If $\mathbf{X}_n \xrightarrow{P} \mathbf{X}$, then $\mathbf{f}(\mathbf{X}_n) \xrightarrow{P} \mathbf{f}(\mathbf{X})$.

(b) If $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$, then $\mathbf{f}(\mathbf{X}_n) \xrightarrow{d} \mathbf{f}(\mathbf{X})$.

The proof of Theorem 2.34 is basically the same as in the univariate case. For proving part (b), we use the multivariate version of Theorem 2.28:

Theorem 2.35 $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$ if and only if $\mathrm{E}\, g(\mathbf{X}_n) \to \mathrm{E}\, g(\mathbf{X})$ for all bounded and continuous real-valued functions $g : \mathbb{R}^k \to \mathbb{R}$.

Proving Theorem 2.35 involves a few complications related to the use of multivariate distribution functions, but the essential idea is the same as in the univariate case (the univariate proof is the subject of Exercises 2.15 and 2.16).

2.3.3 Slutsky's Theorem

As we have seen in the preceding few pages, many univariate definitions and results concerning convergence of sequences of random vectors are basically the same as in the univariate case. Here, however, we consider a result that has no one-dimensional analogue.

At issue is the question of when we may "stack" random variables to make random vectors while preserving convergence. It is here that we encounter perhaps the biggest surprise of this section: Convergence in distribution is not preserved by "stacking".

To understand what we mean by "stacking" preserving convergence, consider the case of convergence in probability. By definition, it is straightforward to show that

$$X_n \xrightarrow{P} X \text{ and } Y_n \xrightarrow{P} Y \text{ together imply that } \begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{P} \begin{pmatrix} X \\ Y \end{pmatrix}. \tag{2.23}$$

Thus, two convergent-in-probability sequences X_n and Y_n may be stacked, one on top of the other, to make a vector, and this vector must still converge in probability to the vector of stacked limits.

Example 2.36 Statement (2.23) gives a way to show that the multivariate Weak Law of Large Numbers (Theorem 2.31) follows immediately from the univariate version (Theorem 2.19).

Example 2.37 If $X_1, X_2, \ldots$ are independent and identically distributed with mean μ and positive variance σ^2, we often take as an estimator of σ^2 the so-called sample variance

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}_n \right)^2 .$$

We may use Statement (2.23) and Theorem 2.34(a), together with the univariate Weak Law of Large Numbers (Theorem 2.19), to prove that s_n^2 is a consistent estimator of σ^2.

To accomplish this, we first rewrite s_n^2 as

$$s_n^2 = \frac{n}{n-1} \left[\frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2 - (\overline{X}_n - \mu)^2 \right].$$

The WLLN tells us that $\overline{X}_n - \mu \xrightarrow{P} 0$, from which we deduce that $(\overline{X}_n - \mu)^2 \xrightarrow{P} 0$ by Theorem 2.27(a). Since $\mathrm{E}\,(X_i - \mu)^2 = \sigma^2$, the Weak Law also tells us that $(1/n)\sum_i (X_i - \mu)^2 \xrightarrow{P} \sigma^2$. Combining these two facts by "stacking" as in (2.23) yields

$$\begin{pmatrix} \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2 \\ (\overline{X}_n - \mu)^2 \end{pmatrix} \xrightarrow{P} \begin{pmatrix} \sigma^2 \\ 0 \end{pmatrix}.$$

We now apply the continuous function $f(a, b) = a - b$ to both sides of the above result, as allowed by Theorem 2.34(a), to conclude that

$$\frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2 - (\overline{X}_n - \mu)^2 \xrightarrow{P} \sigma^2 - 0. \tag{2.24}$$

Finally, we may perform a similar stacking operation using Equation (2.24) together with the fact that $n/(n-1) \xrightarrow{P} 1$, whereupon multiplication yields the final conclusion that $s_n^2 \xrightarrow{P} \sigma^2$.

If it seems as though we spent too much time in the above proof worrying about "obvious" steps such as stacking followed by addition or multiplication, we did so in order to make a point: When we replace $\xrightarrow{P}$ by $\xrightarrow{d}$, the "obvious" is no longer correct.

The converse of (2.23) is true by Theorem 2.27 because the function $f(x, y) = x$ is a continuous function from $\mathbb{R}^2$ to $\mathbb{R}$. By induction, we can therefore stack or unstack arbitrarily many random variables or vectors without distrurbing convergence in probability. Combining this fact with Theorem 2.27 yields a useful result; see Exercise 2.22. By Definition 2.30, Statement 2.23 remains true if we replace $\xrightarrow{P}$ by $\xrightarrow{a}$ throughout the statement for some $a > 0$.

However, Statement 2.23 is *not* true if $\xrightarrow{P}$ is replaced by $\xrightarrow{d}$. Consider the following simple counterexample.

Example 2.38 Take X_n and Y_n to be independent standard normal random variables for all n. These distributions do not depend on n at all, and it is correct to write $X_n \xrightarrow{d} Z$ and $Y_n \xrightarrow{d} Z$, where $Z \sim N(0, 1)$. But it is certainly not true that

$$\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} Z \\ Z \end{pmatrix}, \tag{2.25}$$

since the distribution on the left is bivariate normal with correlation 0, while the distribution on the right is the (degenerate) bivariate normal with correlation 1.

Expression (2.25) is untrue precisely because the marginal distributions do not in general uniquely determine the joint distribution. However, there are certain special cases in which the marginals do determine the joint distribution. For instance, if random variables are independent, then their marginal distributions uniquely determine their joint distribution. Indeed, we can say that if $X_n \to X$ and $Y_n \to Y$, where X_n is independent of Y_n and X is independent of Y, then statement (2.23) remains true when $\xrightarrow{P}$ is replaced by $\xrightarrow{d}$ (see Exercise 2.23). As a special case, the constant c, when viewed as a random variable, is automatically independent of any other random variable. Since $Y_n \xrightarrow{d} c$ is equivalent to $Y_n \xrightarrow{P} c$ by Theorem 2.14, it must be true that

$$X_n \xrightarrow{d} X \text{ and } Y_n \xrightarrow{P} c \text{ implies that } \begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} X \\ c \end{pmatrix} \tag{2.26}$$

if X_n is independent of Y_n for every n. The content of a powerful theorem called Slutsky's Theorem is that statement (2.26) remains true even if the X_n and Y_n are not independent. Although the preceding discussion involves stacking only random (univariate) variables, we present Slutsky's theorem in a more general version involving random vectors.

Theorem 2.39 *Slutsky's Theorem:* For random vectors $\mathbf{X}_n$, $\mathbf{Y}_n$, and $\mathbf{X}$ and a constant $\mathbf{c}$, if $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$ and $\mathbf{Y}_n \xrightarrow{P} \mathbf{c}$ as $n \to \infty$, then

$$\begin{pmatrix} \mathbf{X}_n \\ \mathbf{Y}_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \mathbf{X} \\ \mathbf{c} \end{pmatrix}.$$

A proof of Theorem 2.39 is outlined in Exercise 2.24.

Putting several of the preceding results together yields the following corollary.

Corollary 2.40 If $\mathbf{X}$ is a k-vector such that $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$, and $Y_{nj} \xrightarrow{P} c_j$ for $1 \le j \le m$, then

$$\mathbf{f}\begin{pmatrix} \mathbf{X}_n \\ \mathbf{Y}_n \end{pmatrix} \xrightarrow{d} \mathbf{f}\begin{pmatrix} \mathbf{X} \\ \mathbf{c} \end{pmatrix}$$

for any continuous function $\mathbf{f} : S \subset \mathbb{R}^{k+m} \to \mathbb{R}^{\ell}$.

It is very common practice in statistics to use Corollary 2.40 to obtain a result, then state that the result follows "by Slutsky's Theorem". In fact, there is not a unanimously held view in the statistical literature about what precisely "Slutsky's Theorem" refers to; some consider the Corollary itself, or particular cases of the Corollary, to be Slutsky's Theorem. These minor differences are unimportant; the common feature of all references to "Slutsky's Theorem" is some combination of one sequence that converges in distribution with one or more sequences that converge in probability to constants.

Example 2.41 *Asymptotic normality of the t-statistic:* Let $X_1, \ldots, X_n$ be independent and identically distributed with mean μ and finite positive variance σ^2. By Example 2.12, $\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$. Suppose that $\hat{\sigma}_n^2$ is any consistent estimator of σ^2; that is, $\hat{\sigma}_n^2 \xrightarrow{P} \sigma^2$. (For instance, we might take $\hat{\sigma}_n^2$ to be the usual unbiased sample variance estimator s_n^2 of Example 2.37, whose asymptotic properties will be studied later.) If Z denotes a standard normal random variable, Theorem 2.39 implies

$$\begin{pmatrix} \sqrt{n}(\overline{X}_n - \mu) \\ \hat{\sigma}_n^2 \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \sigma Z \\ \sigma^2 \end{pmatrix}. \tag{2.27}$$

Therefore, since $f(a, b) = a/b$ is a continuous function for $b > 0$ (and σ^2 is assumed positive),

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sqrt{\hat{\sigma}_n^2}} \xrightarrow{d} Z. \tag{2.28}$$

It is common practice to skip step (2.27), attributing equation (2.28) directly to "Slutsky's Theorem".

Exercises for Section 2.3

Exercise 2.15 Here we prove half (the "only if" part) of Theorem 2.28: If $X_n \xrightarrow{d} X$ and $g(x)$ is a bounded, continuous function on $\mathbb{R}$, then $\mathrm{E}\ g(X_n) \to \mathrm{E}\ g(X)$. (This half of Theorem 2.28 is sometimes called the univariate Helly-Bray Theorem.)

Let $F_n(x)$ and $F(x)$ denote the distribution functions of X_n and X, as usual. For $\epsilon > 0$, take $b < c$ to be constant real numbers such that $F(b) < \epsilon$ and $F(c) > 1 - \epsilon$. First, we note that since $g(x)$ is continuous, it must be *uniformly continuous* on $[b, c]$: That is, for any $\epsilon > 0$ there exists $\delta > 0$ such that $|g(x) - g(y)| < \epsilon$ whenever $|x - y| < \delta$. This fact, along with the boundedness of $g(x)$, ensures that there exists a finite set of real numbers $b = t_0 < t_1 < \cdots < t_m = c$ such that:

- Each t_i is a continuity point of $F(x)$.
- $F(t_0) < \epsilon$ and $F(t_m) > 1 - \epsilon$.
- For $1 \leq i \leq m$, $|g(x) - g(t_i)| < \epsilon$ for all $x \in [t_{i-1}, t_i]$.

(a) As in Figure 2.1, define

$$h(x) = \begin{cases} g(t_i) & \text{if } t_{i-1} < x \leq t_i \text{ for some } 1 \leq i \leq m. \\ 0 & \text{otherwise.} \end{cases}$$

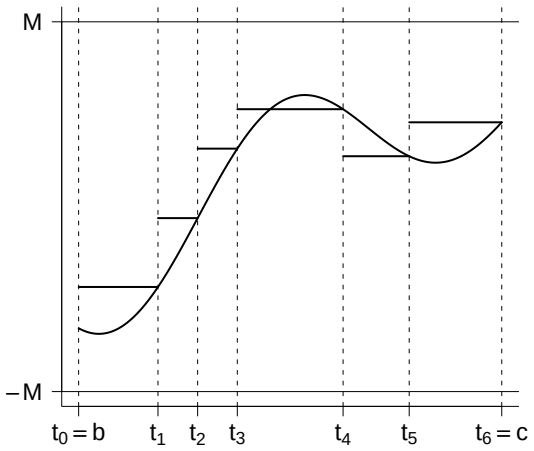

Figure 2.1: *The solid curve is the function $g(x)$, assumed to be bounded by $-M$ and M for all x, and the horizontal segments are the function $h(x)$. The points $t_0, \ldots, t_6$ are chosen so that $|g(x) - h(x)|$ is always less than ϵ. The t_i are continuity points of $F(x)$, and both $F(t_0)$ and $1 - F(t_6)$ are less than ϵ.*

Prove that there exists N such that $|\,\mathrm{E}\, h(X_n) - \mathrm{E}\, h(X)| < \epsilon$ for all $n > N$.

Hint: Use the fact that for any random variable Y,

$$\mathrm{E}\, h(Y) = \sum_{i=1}^{m} g(t_i) P(t_{i-1} < Y \leq t_i).$$

Also, please note that we may not write $\mathrm{E}\, h(X_n) - \mathrm{E}\, h(X)$ as $\mathrm{E}\,[h(X_n) - h(X)]$ because it is not necessarily the case that X_n and X are defined on the same sample space.

(b) Prove that $\mathrm{E}\, g(X_n) \to \mathrm{E}\, g(X)$.

Hint: Use the fact that

$$\begin{aligned} |\,\mathrm{E}\, g(X_n) - \mathrm{E}\, g(X)| \;\leq\; & |\,\mathrm{E}\, g(X_n) - \mathrm{E}\, h(X_n)| + |\,\mathrm{E}\, h(X_n) - \mathrm{E}\, h(X)| \\ & + |\,\mathrm{E}\, h(X) - \mathrm{E}\, g(X)|. \end{aligned}$$

Exercise 2.16 Prove the other half (the "if" part) of Theorem 2.28, which states that if $\mathrm{E}\, g(X_n) \to \mathrm{E}\, g(X)$ for all bounded, continuous functions $g : \mathbb{R} \to \mathbb{R}$, then $X_n \xrightarrow{d} X$.

(a) Let t be any continuity point of $F(x)$. Let $\epsilon > 0$ be arbitrary. Show that there exists $\delta > 0$ such that $F(t-\delta) > F(t) - \epsilon$ and $F(t+\delta) < F(t) + \epsilon$.

(b) Show how to define continuous functions $g_1 : \mathbb{R} \to [0,1]$ and $g_2 : \mathbb{R} \to [0,1]$ such that for all $x \leq t$, $g_1(x) = g_2(x-\delta) = 1$ and for all $x > t$, $g_1(x+\delta) = g_2(x) = 0$. Use these functions to bound the difference between $F_n(t)$ and $F(t)$ in such a way that this difference must tend to 0.

Exercise 2.17 To illustrate a situation that can arise in the multivariate setting that cannot arise in the univariate setting, construct an example of a sequence (X_n, Y_n), a joint distribution (X, Y), and a connected subset $S \in R^2$ such that

(i) $(X_n, Y_n) \xrightarrow{d} (X, Y)$;

(ii) every point of R^2 is a continuity point of the distribution function of (X, Y);

(iii) $P[(X_n, Y_n) \in S]$ does not converge to $P[(X, Y) \in S]$.

Hint: Condition (ii) may be satisfied even if the distribution of (X, Y) is concentrated on a line.

Exercise 2.18 If X is a univariate random variable with distribution function $F(x)$, then $F(x)$ is continuous at c if and only if $P(X = c) = 0$. Prove by counterexample that this is not true if variables X and c are replaced by vectors $\mathbf{X}$ and $\mathbf{c}$.

Exercise 2.19 Suppose that (X, Y) is a bivariate normal vector such that both X and Y are marginally standard normal and $\text{Corr}\,(X, Y) = \rho$. Construct a computer program that simulates the distribution function $F_\rho(x, y)$ of the joint distribution of X and Y. For a given (x, y), the program should generate at least 50,000 random realizations from the distribution of (X, Y), then report the proportion for which $(X, Y) \leq (x, y)$. (If you wish, you can also report a confidence interval for the true value.) Use your function to approximate $F_{.5}(1, 1)$, $F_{.25}(-1, -1)$, and $F_{.75}(0, 0)$. As a check of your program, you can try it on $F_0(x, y)$, whose true values are not hard to calculate directly for an arbitrary x and y assuming your software has the ability to evaluate the standard normal distribution function.

Hint: To generate a bivariate normal random vector (X, Y) with covariance matrix $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, start with independent standard normal U and V, then take $X = U$ and $Y = \rho U + \sqrt{1-\rho^2}V$.

Exercise 2.20 Adapt the method of proof in Exercise 2.15 to the multivariate case, proving half of Theorem 2.35: If $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$, then $\text{E}\; g(\mathbf{X}_n) \to \text{E}\; g(\mathbf{X})$ for any

bounded, continuous $g : \mathbb{R}^k \to \mathbb{R}$.

Hint: Instead of intervals $(a_{i-1}, a_i]$ as in Exercise 2.15, use small regions $\{\mathbf{x} : a_{i,j-1} < x_i \leq a_{i,j} \text{ for all } i\}$ of $\mathbb{R}^k$. Make sure these regions are chosen so that their boundaries contain only continuity points of $F(\mathbf{x})$.

Exercise 2.21 Construct a counterexample to show that Slutsky's Theorem 2.39 may not be strengthened by changing $Y_n \xrightarrow{P} c$ to $Y_n \xrightarrow{P} Y$.

Exercise 2.22 **(a)** Prove that if $f : \mathbb{R}^k \to \mathbb{R}^\ell$ is continuous and $X_{nj} \xrightarrow{P} X_j$ for all $1 \leq j \leq k$, then $f(\mathbf{X}_n) \xrightarrow{P} f(\mathbf{X})$.

(b) Taking $f(a, b) = a + b$ for simplicity, construct an example demonstrating that part (a) is not true if $\xrightarrow{P}$ is replaced by $\xrightarrow{d}$.

Exercise 2.23 Prove that if X_n is independent of Y_n for all n and X is independent of Y, then

$$X_n \xrightarrow{d} X \text{ and } Y_n \xrightarrow{d} Y \text{ implies that } \begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} X \\ Y \end{pmatrix}.$$

Hint: Be careful to deal with points of discontinuity: If X_n and Y_n are independent, what characterizes a point of discontinuity of the joint distribution?

Exercise 2.24 Prove Slutsky's Theorem, Theorem 2.39, using the following approach:

(a) Prove the following lemma:

Lemma 2.42 Let $\mathbf{V}_n$ and $\mathbf{W}_n$ be k-dimensional random vectors on the same sample space.

$$\text{If } \mathbf{V}_n \xrightarrow{d} \mathbf{V} \text{ and } \mathbf{W}_n \xrightarrow{P} \mathbf{0}, \text{ then } \mathbf{V}_n + \mathbf{W}_n \xrightarrow{d} \mathbf{V}.$$

Hint: For $\epsilon > 0$, let $\boldsymbol{\epsilon}$ denote the k-vector all of whose entries are ϵ. Take $\mathbf{a} \in \mathbb{R}^k$ to be a continuity point of $F_{\mathbf{V}}(\mathbf{v})$. Now argue that $\mathbf{a}$, since it is a point of continuity, must be contained in a neighborhood consisting only of points of continuity; therefore, ϵ may be taken small enough so that $\mathbf{a} - \boldsymbol{\epsilon}$ and $\mathbf{a} + \boldsymbol{\epsilon}$ are also points of continuity. Prove that

$$\begin{aligned} P(\mathbf{V}_n \leq \mathbf{a} - \boldsymbol{\epsilon}) - P(\|\mathbf{W}_n\| \geq \epsilon) &\leq P(\mathbf{V}_n + \mathbf{W}_n \leq \mathbf{a}) \\ &\leq P(\mathbf{V}_n \leq \mathbf{a} + \boldsymbol{\epsilon}) + P(\|\mathbf{W}_n\| \geq \epsilon). \end{aligned}$$

Next, take $\limsup_n$ and $\liminf_n$. Finally, let $\epsilon \to 0$.

(b) Show how to prove Theorem 2.39 using Lemma 2.42.

Hint: Consider the random vectors

$$\mathbf{V}_n = \begin{pmatrix} \mathbf{X}_n \\ \mathbf{c} \end{pmatrix} \quad \text{and} \quad \mathbf{W}_n = \begin{pmatrix} \mathbf{0} \\ \mathbf{Y}_n - \mathbf{c} \end{pmatrix}.$$

Chapter 3

Strong convergence

There are multiple ways to define the convergence of a sequence of random variables. Chapter 2 introduced convergence in probability, convergence in distribution, and convergence in quadratic mean. We now consider a fourth mode of convergence, almost sure convergence or convergence with probability one. We will see that almost sure convergence implies both convergence in probability and convergence in distribution, which is why we sometimes use the term "strong" for almost sure convergence and "weak" for the other two.

The terms "weak" and "strong" do not indicate anything about their importance; indeed, the "weak" modes of convergence are used much more frequently in asymptotic statistics than the strong mode. Because weak convergence dominates the remainder of this book beginning with Chapter 4, a reader may safely skip much of the material in the current chapter if time is limited; however, the quantile function and the Dominated Convergence Theorem of Section 3.3 are used elsewhere, and at least these topics should be reviewed before moving on. Due to the technical nature of the material of this chapter, the exercises are almost exclusively devoted to proofs.

3.1 Strong Consistency Defined

A random variable like X_n or X is a function on a sample space, say Ω. Suppose that we fix a particular element of that space, say ω_0, so we obtain the real numbers $X_n(\omega_0)$ and $X(\omega_0)$. If $X_n(\omega_0) \to X(\omega_0)$ as $n \to \infty$ in the sense of Definition 1.1, then ω_0 is contained in the event

$$S = \{\omega \in \Omega : X_n(\omega) \to X(\omega)\}. \tag{3.1}$$

If the probability of S—that is, E $I\{X_n \to X\}$—equals 1, then we say that X_n converges almost surely to X:

Definition 3.1 Suppose X and $X_1, X_2, \ldots$ are random variables defined on the same sample space Ω (and as usual P denotes the associated probability measure). If

$$P\left(\{\omega \in \Omega : X_n(\omega) \to X(\omega)\}\right) = 1,$$

then X_n is said to converge almost surely (or with probability one) to X, denoted $X_n \overset{\text{a.s.}}{\to} X$ or $X_n \to X$ a.s. or $X_n \to X$ w.p. 1.

In other words, convergence with probability one means exactly what it sounds like: The probability that X_n converges to X equals one. Later, in Theorem 3.3, we will formulate an equivalent definition of almost sure convergence that makes it much easier to see why it is such a strong form of convergence of random variables. Yet the intuitive simplicity of Definition 3.1 makes it the standard definition.

As in the case of convergence in probability, we may replace the limiting random variable X by any constant c, in which case we write $X_n \overset{\text{a.s.}}{\to} c$. In the most common statistical usage of convergence to a constant, the random variable X_n is some estimator of a particular parameter, say θ:

Definition 3.2 If $X_n \overset{\text{a.s.}}{\to} \theta$, X_n is said to be **strongly consistent** for θ.

As the names suggest, strong consistency implies consistency (also known as weak consistency), a fact we now explore in more depth.

3.1.1 Strong Consistency versus Consistency

As before, suppose that X and $X_1, X_2, \ldots$ are random variables defined on the same sample space, Ω. For given n and $\epsilon > 0$, define the events

$$A_n = \{\omega \in \Omega : |X_k(\omega) - X(\omega)| < \epsilon \text{ for all } k \geq n\} \tag{3.2}$$

and

$$B_n = \{\omega \in \Omega : |X_n(\omega) - X(\omega)| < \epsilon\}. \tag{3.3}$$

First, note that A_n must be contained in B_n and that both A_n and B_n imply that X_n is close to X as long as ϵ is small. Therefore, both $P(A_n) \to 1$ and $P(B_n) \to 1$ seem like reasonable ways to define the convergence of X_n to X. Indeed, as we have already seen in Definition 2.1, convergence in probability means precisely that $P(B_n) \to 1$ for any $\epsilon > 0$.

Yet what about the sets A_n? One fact is immediate: Since $A_n \subset B_n$, we must have $P(A_n) \leq P(B_n)$. Therefore, $P(A_n) \to 1$ implies $P(B_n) \to 1$. In other words, if we were to

take $P(A_n) \to 1$ for all $\epsilon > 0$ to be the definition of a new form of convergence of random sequences, then this form of convergence would be stronger than (i.e., it would imply) convergence in probability. By now, the reader may already have guessed that this new form of convergence is actually equivalent to almost sure convergence:

Theorem 3.3 With A_n defined as in Equation (3.2), $P(A_n) \to 1$ for any $\epsilon > 0$ if and only if $X_n \overset{\text{a.s.}}{\to} X$.

Proving Theorem 3.3 is the subject of Exercise 3.1. The following corollary now follows from the preceding discussion:

Corollary 3.4 If $X_n \overset{\text{a.s.}}{\to} X$, then $X_n \overset{P}{\to} X$.

The converse of Corollary 3.4 is not true, as the following example illustrates.

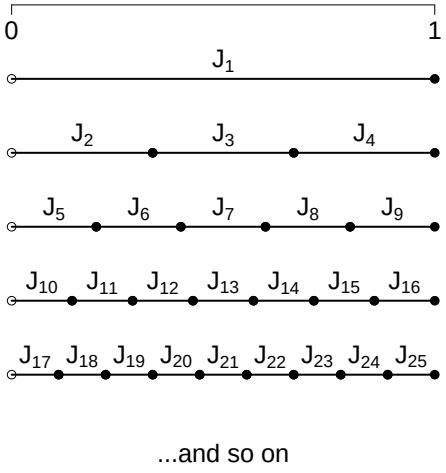

Figure 3.1: *Example 3.5, in which* $P(J_n) \to 0$ *as* $n \to \infty$*, which means that* $I\{J_n\} \overset{P}{\to} 0$*. However, the intervals* J_n *repeatedly cover the entire interval* $(0, 1]$*, so the subset of* $(0, 1]$ *on which* $I\{J_n\}$ *converges to* 0 *is empty!*

Example 3.5 Take Ω to be the half-open interval $(0, 1]$, and for any interval $J \subset \Omega$, say $J = (a, b]$, take $P(J) = b - a$ to be the length of that interval. Define a sequence of intervals $J_1, J_2, \ldots$ as follows (see Figure 3.1):

$$\begin{aligned} J_1 &= (0, 1] \\ J_2 \text{ through } J_4 &= \left(0, \tfrac{1}{3}\right], \left(\tfrac{1}{3}, \tfrac{2}{3}\right], \left(\tfrac{2}{3}, 1\right] \end{aligned}$$

$$\begin{aligned} J_5 \text{ through } J_9 &= \left(0,\tfrac{1}{5}\right], \left(\tfrac{1}{5},\tfrac{2}{5}\right], \left(\tfrac{2}{5},\tfrac{3}{5}\right], \left(\tfrac{3}{5},\tfrac{4}{5}\right], \left(\tfrac{4}{5},1\right] \\ &\vdots \\ J_{m^2+1} \text{ through } J_{(m+1)^2} &= \left(0, \frac{1}{2m+1}\right], \ldots, \left(\frac{2m}{2m+1}, 1\right] \\ &\vdots \end{aligned}$$

Note in particular that $P(J_n) = 1/(2m+1)$, where $m = \lfloor \sqrt{n-1} \rfloor$ is the largest integer not greater than $\sqrt{n-1}$. Now, define $X_n = I\{J_n\}$ and take $0 < \epsilon < 1$. Then $P(|X_n - 0| < \epsilon)$ is the same as $1 - P(J_n)$. Since $P(J_n) \to 0$, we conclude $X_n \xrightarrow{P} 0$ by definition.

However, it is *not* true that $X_n \xrightarrow{\text{a.s.}} 0$. Since every $\omega \in \Omega$ is contained in infinitely many J_n, the set A_n defined in Equation (3.2) is *empty* for all n. Alternatively, consider the set $S = \{\omega : X_n(\omega) \to 0\}$. For any ω, $X_n(\omega)$ has no limit because $X_n(\omega) = 1$ and $X_n(\omega) = 0$ both occur for infinitely many n. Thus S is empty. This is not convergence with probability one; it is convergence with probability zero!

3.1.2 Multivariate Extensions

We may extend Definition 3.1 to the multivariate case in a completely straightforward way:

Definition 3.6 $\mathbf{X}_n$ is said to converge almost surely (or with probability one) to $\mathbf{X}$ ($\mathbf{X}_n \xrightarrow{\text{a.s.}} \mathbf{X}$) if

$$P\left(\mathbf{X}_n \to \mathbf{X} \text{ as } n \to \infty\right) = 1.$$

Alternatively, since the proof of Theorem 3.3 applies to random vectors as well as random variables, we say $\mathbf{X}_n \xrightarrow{\text{a.s.}} \mathbf{X}$ if for any $\epsilon > 0$,

$$P\left(\|\mathbf{X}_k - \mathbf{X}\| < \epsilon \text{ for all } k \geq n\right) \to 1 \text{ as } n \to \infty. \tag{3.4}$$

We saw in Theorems 2.27 and 2.34 that continuous functions preserve both convergence in probability and convergence in distribution. Yet these facts were quite difficult to prove. Fortunately, the analogous result for almost sure convergence follows immediately from the results of Chapter 1. Similarly, unlike with convergence in distribution, there is no problem "stacking" random sequences into vectors while preserving almost sure convergence. The following theorem is really just a corollary of earlier results (specifically, Theorem 1.16 and Lemma 1.33).

Theorem 3.7 **(a)** Suppose that $\mathbf{f} : S \to \mathbb{R}^\ell$ is a continuous function defined on some subset $S \subset \mathbb{R}^k$, $\mathbf{X}_n$ is a k-component random vector, and the range of $\mathbf{X}$ and of each X_n is contained in S with probability 1. If $\mathbf{X}_n \xrightarrow{\text{a.s.}} \mathbf{X}$, then $\mathbf{f}(\mathbf{X}_n) \xrightarrow{\text{a.s.}} \mathbf{f}(\mathbf{X})$.

(b) $\mathbf{X}_n \xrightarrow{\text{a.s.}} \mathbf{X}$ if and only if $X_{nj} \xrightarrow{\text{a.s.}} X_j$ for all j.

We conclude this section with a simple diagram summarizing the implications among the modes of convergence defined so far. In the diagram, a double arrow like $\Rightarrow$ means "implies". Note that the picture changes slightly when convergence is to a constant $\mathbf{c}$ rather than a random vector $\mathbf{X}$.

$$\begin{array}{ccccc} & & \mathbf{X}_n \xrightarrow{\text{qm}} \mathbf{X} & & \\ & & \Downarrow & & \\ \mathbf{X}_n \xrightarrow{\text{a.s.}} \mathbf{X} & \Rightarrow & \mathbf{X}_n \xrightarrow{P} \mathbf{X} & \Rightarrow & \mathbf{X}_n \xrightarrow{d} \mathbf{X} \end{array} \qquad \begin{array}{ccccccc} & & \mathbf{X}_n \xrightarrow{\text{qm}} \mathbf{c} & & & & \\ & & \Downarrow & & & & \\ \mathbf{X}_n \xrightarrow{\text{a.s.}} \mathbf{c} & \Rightarrow & \mathbf{X}_n \xrightarrow{P} \mathbf{c} & \Leftrightarrow & \mathbf{X}_n \xrightarrow{d} \mathbf{c} \end{array}$$

Exercises for Section 3.1

Exercise 3.1 Let S be the set defined in equation (3.1), so $X_n \xrightarrow{\text{a.s.}} X$ is equivalent to $P(S) = 1$ by definition.

(a) Let A_n be defined as in Equation (3.2). Prove that

$$\omega_0 \in \cup_{n=1}^{\infty} A_n \text{ for all } \epsilon > 0$$

if and only if $\omega_0 \in S$.

Hint: Use Definition 1.1.

(b) Prove Theorem 3.3.

Hint: Note that the sets A_n are increasing in n, so that by the lower continuity of any probability measure (which you may assume without proof), $\lim_n P(A_n)$ exists and is equal to $P(\cup_{n=1}^{\infty} A_n)$.

Exercise 3.2 The diagram at the end of this section suggests that neither $X_n \xrightarrow{\text{a.s.}} X$ nor $X_n \xrightarrow{\text{qm}} X$ implies the other. Construct two counterexamples, one to show that $X_n \xrightarrow{\text{a.s.}} X$ does not imply $X_n \xrightarrow{\text{qm}} X$ and the other to show that $X_n \xrightarrow{\text{qm}} X$ does not imply $X_n \xrightarrow{\text{a.s.}} X$.

3.2 The Strong Law of Large Numbers

Some of the results in this section are presented for univariate random variables and some are presented for random vectors. Take note of the use of bold print to denote vectors. Nearly

all of the technical proofs are posed as exercises (with hints, of course).

Theorem 3.8 *Strong Law of Large Numbers:* Suppose that $\mathbf{X}_1, \mathbf{X}_2, \ldots$ are independent and identically distributed and have finite mean $\boldsymbol{\mu}$. Then $\overline{\mathbf{X}}_n \xrightarrow{\text{a.s.}} \boldsymbol{\mu}$.

It is possible to use fairly simple arguments to prove a version of the Strong Law under more restrictive assumptions than those given above. See Exercise 3.4 for details of a proof of the univariate Strong Law under the additional assumption that $\mathrm{E}\, X_n^4 < \infty$. To aid the proof of the Strong Law, with or without such an additional assumption, we first establish a useful lemma.

Lemma 3.9 If $\sum_{k=1}^{\infty} P(\|\mathbf{X}_k - \mathbf{X}\| > \epsilon) < \infty$ for any $\epsilon > 0$, then $\mathbf{X}_n \xrightarrow{\text{a.s.}} \mathbf{X}$.

Proof: The proof relies on the *countable subadditivity* of any probability measure, an axiom stating that for any sequence $C_1, C_2, \ldots$ of events,

$$P\left(\bigcup_{k=1}^{\infty} C_k\right) \leq \sum_{k=1}^{\infty} P(C_k). \tag{3.5}$$

To prove the lemma using (3.4), we must demonstrate that $P\left(\|\mathbf{X}_k - \mathbf{X}\| \leq \epsilon \text{ for all } k \geq n\right) \to 1$ as $n \to \infty$, which (taking complements) is equivalent to $P\left(\|\mathbf{X}_k - \mathbf{X}\| > \epsilon \text{ for some } k \geq n\right) \to 0$. Letting C_k denote the event that $\|\mathbf{X}_k - \mathbf{X}\| > \epsilon$, countable subadditivity implies

$$P\left(C_k \text{ for some } k \geq n\right) = P\left(\bigcup_{k=n}^{\infty} C_k\right) \leq \sum_{k=n}^{\infty} P\left(C_k\right),$$

and the right hand side tends to 0 as $n \to \infty$ because it is the tail of a convergent series. ■

Lemma 3.9 is nearly the same as a famous result called the First Borel-Cantelli Lemma, or sometimes simply the Borel-Cantelli Lemma; see Exercise 3.3. Lemma 3.9 is extremely useful for establishing almost sure convergence of sequences. As an illustration of the type of result this lemma helps to prove, consider the following theorem (see Exercise 3.8 for hints on how to prove it).

Theorem 3.10 $\mathbf{X}_n \xrightarrow{P} \mathbf{X}$ if and only if each subsequence $\mathbf{X}_{n_1}, \mathbf{X}_{n_2}, \ldots$ contains a further subsequence that converges almost surely to $\mathbf{X}$.

Using Theorem 3.10, it is now—finally—possible to prove that continuous transformations preserve convergence in probability. This fact was stated in Theorem 2.27(a) (for the univariate case) and Theorem 2.34(a) (for the multivariate case). It suffices to complete the proof for the multivariate case.

Proof of Theorem 2.34(a): If $\mathbf{X}_n \xrightarrow{P} \mathbf{X}$ and $f(\mathbf{x})$ is continuous, it suffices to prove that each subsequence $f(\mathbf{X}_{n_1}), f(\mathbf{X}_{n_2}), \ldots$ contains a further subsequence that converges

almost surely to $f(\mathbf{X})$. But we know that $\mathbf{X}_{n_1}, \mathbf{X}_{n_2}, \ldots$ must contain a further subsequence, say $\mathbf{X}_{n_{i(1)}}, \mathbf{X}_{n_{i(2)}}, \ldots$, such that $\mathbf{X}_{n_{i(j)}} \stackrel{\text{a.s.}}{\to} \mathbf{X}$ and therefore $f(\mathbf{X}_{n_{i(j)}}) \stackrel{\text{a.s.}}{\to} f(\mathbf{X})$ as $j \to \infty$ by Theorem 3.7(a). This proves the theorem! ■

To conclude this section, we provide a proof of the Strong Law (Theorem 3.8). The approach we give here is based on a powerful theorem of Kolmogorov:

Theorem 3.11 *Kolmogorov's Strong Law of Large Numbers:* Suppose that $X_1, X_2, \ldots$ are independent with mean μ and

$$\sum_{i=1}^{\infty} \frac{\text{Var } X_i}{i^2} < \infty.$$

Then $\overline{X}_n \stackrel{\text{a.s.}}{\to} \mu$.

Note that there is no reason the X_i in Theorem 3.11 must have the same means: If E $X_i = \mu_i$, then the conclusion of the theorem becomes $(1/n)\sum_i (X_i - \mu_i) \stackrel{\text{a.s.}}{\to} 0$. Theorem 3.11 may be proved using Kolmogorov's inequality from Exercise 1.41; this proof is the focus of Exercise 3.7.

The key to completing the Strong Law of Large Numbers for an independent and identically distributed sequence using Theorem 3.11 is to introduce truncated versions of $X_1, X_2, \ldots$ as in the following lemmas, which are proved in Exercises 3.5 and 3.6.

Lemma 3.12 Suppose that $X_1, X_2, \ldots$ are independent and identically distributed and have finite mean μ. Define $X_i^* = X_i I\{|X_i| \le i\}$. Then

$$\sum_{i=1}^{\infty} \frac{\text{Var } X_i^*}{i^2} < \infty. \tag{3.6}$$

Lemma 3.13 Under the assumptions of Lemma 3.12, let $\overline{X}_n^* = (1/n)\sum_{i=1}^n X_i^*$. Then $\overline{X}_n - \overline{X}_n^* \stackrel{\text{a.s.}}{\to} 0$.

Finally, it is possible to put the preceding results together to prove the Strong Law of Large Numbers:

Proof of Theorem 3.8: Let $X_1, X_2, \ldots$ be independent and identically distributed with finite mean μ, and let $X_i^* = X_i I\{|X_i| \le i\}$. Then Lemma 3.12 and Theorem 3.11 together imply that $\overline{X}_n^* \stackrel{\text{a.s.}}{\to} \mu$. From Lemma 3.13, we obtain $\overline{X}_n - \overline{X}_n^* \stackrel{\text{a.s.}}{\to} 0$. Adding these two limit statements (which is legal because of Theorem 3.7), we obtain

$$\overline{X}_n = \overline{X}_n^* + (\overline{X}_n - \overline{X}_n^*) \stackrel{\text{a.s.}}{\to} \mu,$$

which establishes the Strong Law for the univariate case. Since "stacking" sequences presents no problems for almost sure convergence [Theorem 3.7(b)], the multivariate version follows immediately. ■

Exercises for Section 3.2

Exercise 3.3 Let $B_1, B_2, \ldots$ denote a sequence of events. Let B_n i.o., which stands for B_n infinitely often, denote the set

$$B_n \text{ i.o.} \stackrel{\text{def}}{=} \{\omega \in \Omega : \text{ for every } n, \text{ there exists } k \geq n \text{ such that } \omega \in B_k\}.$$

Prove the *First Borel-Cantelli Lemma*, which states that if $\sum_{n=1}^{\infty} P(B_n) < \infty$, then $P(B_n \text{ i.o.}) = 0$.

Hint: Argue that

$$B_n \text{ i.o.} = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} B_k,$$

then adapt the proof of Lemma 3.9.

Exercise 3.4 Use the steps below to prove a version of the Strong Law of Large Numbers for the special case in which the random variables $X_1, X_2, \ldots$ have a finite fourth moment, $\text{E } X_1^4 < \infty$.

(a) Assume without loss of generality that $\text{E } X_1 = 0$. Expand $\text{E}\,(X_1 + \ldots + X_n)^4$ and then count the nonzero terms.

Hint: The only nonzero terms are of the form $\text{E } X_i^4$ or $(\text{E } X_i^2)^2$.

(b) Use Markov's inequality (1.35) with $r = 4$ to put an upper bound on

$$P\left(\left|\overline{X}_n\right| > \epsilon\right)$$

involving $\text{E}\,(X_1 + \ldots + X_n)^4$.

(c) Combine parts (a) and (b) with Lemma 3.9 to show that $\overline{X}_n \stackrel{\text{a.s.}}{\to} 0$.

Hint: Use the fact that $\sum_{n=1}^{\infty} n^{-2} < \infty$.

Exercise 3.5 Lemmas 3.12 and 3.13 make two assertions about the random variables $X_i^* = X_i I\{|X_i| \leq i\}$, where $X_1, X_2, \ldots$ are independent and identically distributed with finite mean μ.

(a) Prove that for an arbitrary real number c,

$$c^2 \sum_{i=1}^{\infty} \frac{1}{i^2} I\{|c| \leq i\} \leq 1 + |c|.$$

Hint: Bound the sum on the left hand side by an easy-to-evaluate integral.

(b) Prove Lemma 3.12, which states that

$$\sum_{i=1}^{\infty} \frac{\text{Var } X_i^*}{i^2} < \infty.$$

Hint: Use the fact that

$$\text{Var } X_i^* \le \text{E}\,(X_i^*)^2 = \text{E } X_1^2 I\{|X_1| \le i\}$$

together with part (a) and the fact that $\text{E } |X_1| < \infty$.

Exercise 3.6 Assume the conditions of Exercise 3.5.

(a) Prove that $X_n - X_n^* \xrightarrow{\text{a.s.}} 0$.

Hint: Note that X_n and X_n^* do not have bars here. Use Exercise 1.45 together with Lemma 3.9.

(b) Prove Lemma 3.13, which states that

$$\overline{X}_n - \overline{X}_n^* \xrightarrow{\text{a.s.}} 0.$$

Hint: Use Exercise 1.3.

Exercise 3.7 Prove Theorem 3.11. Use the following steps:

(a) For $k = 1, 2, \ldots$, define

$$Y_k = \max_{2^{k-1} \le n < 2^k} |\overline{X}_n - \mu|.$$

Use the Kolmogorov inequality from Exercise 1.41 to show that

$$P(Y_k \ge \epsilon) \le \frac{4 \sum_{i=1}^{2^k} \text{Var } X_i}{4^k \epsilon^2}.$$

(b) Use Lemma 3.9 to show that $Y_k \xrightarrow{\text{a.s.}} 0$, then argue that this proves $\overline{X}_n \xrightarrow{\text{a.s.}} \mu$.

Hint: Letting $\lceil \log_2 i \rceil$ denote the smallest integer greater than or equal to $\log_2 i$ (the base-2 logarithm of i), verify and use the fact that

$$\sum_{k=\lceil \log_2 i \rceil}^{\infty} \frac{1}{4^k} \le \frac{4}{3i^2}.$$

Exercise 3.8 Prove Theorem 3.10:

(a) To simplify notation, let $\mathbf{Y}_1 = \mathbf{X}_{n_1}, \mathbf{Y}_2 = \mathbf{X}_{n_2}, \ldots$ denote an arbitrary subseqence of $\mathbf{X}_1, \mathbf{X}_2, \ldots$.

Prove the "only if" part of Theorem 3.10, which states that if $\mathbf{X}_n \xrightarrow{P} \mathbf{X}$, then there exists a subsequence $\mathbf{Y}_{m_1}, \mathbf{Y}_{m_2}, \ldots$ such that $\mathbf{Y}_{m_j} \xrightarrow{\text{a.s.}} \mathbf{X}$ as $j \to \infty$.

Hint: Show that there exist $m_1, m_2, \ldots$ such that

$$P(\|\mathbf{Y}_{m_j} - \mathbf{X}\| > \epsilon) < \frac{1}{2^j},$$

then use Lemma 3.9.

(b) Now prove the "if" part of the theorem by arguing that if $\mathbf{X}_n$ does not converge in probability to $\mathbf{X}$, there exists a subsequence $\mathbf{Y}_1 = \mathbf{X}_{n_1}, \mathbf{Y}_2 = \mathbf{X}_{n_2}, \ldots$ and $\epsilon > 0$ such that

$$P(\|\mathbf{Y}_k - \mathbf{X}\| > \epsilon) > \epsilon$$

for all k. Then use Corollary 3.4 to argue that $\mathbf{Y}_1, \mathbf{Y}_2, \ldots$ does not have a subsequence that converges almost surely.

3.3 The Dominated Convergence Theorem

In this section, the key question is this: When does $X_n \xrightarrow{d} X$ imply $\mathrm{E}\ X_n \to \mathrm{E}\ X$? The answer to this question impacts numerous results in statistical large-sample theory. Yet because the question involves only convergence in distribution, it may seem odd that it is being asked here, in the chapter on almost sure convergence. We will see that one of the most useful conditions under which $\mathrm{E}\ X_n \to \mathrm{E}\ X$ follows from $X_n \xrightarrow{d} X$, the Dominated Convergence Theorem, is proved using almost sure convergence.

3.3.1 Moments Do Not Always Converge

It is easy to construct cases in which $X_n \xrightarrow{d} X$ does *not* imply $\mathrm{E}\ X_n \to \mathrm{E}\ X$, and perhaps the easiest way to construct such examples is to recall that if X is a constant, then $\xrightarrow{d} X$ and $\xrightarrow{P} X$ are equivalent: For $X_n \xrightarrow{P} c$ means only that X_n is close to c *with probability approaching one*. What happens to X_n when it is not close to c can have an arbitrarily extreme influence on the mean of X_n. In other words, $X_n \xrightarrow{P} c$ certainly does not imply that $\mathrm{E}\ X_n \to c$, as the next two examples show.

Example 3.14 Suppose that U is a standard uniform random variable. Define

$$X_n = nI\{U < 1/n\} = \begin{cases} n & \text{with probability } 1/n \\ 0 & \text{with probability } 1 - 1/n. \end{cases} \tag{3.7}$$

We see immediately that $X_n \xrightarrow{P} 0$ as $n \to \infty$, but the mean of X_n, which is 1 for all n, is certainly not converging to zero!. Furthermore, we may generalize this example by defining

$$X_n = c_n I\{U < p_n\} = \begin{cases} c_n & \text{with probability } p_n \\ 0 & \text{with probability } 1 - p_n. \end{cases}$$

In this case, a sufficient condition for $X_n \xrightarrow{P} 0$ is $p_n \to 0$. But the mean $\mathrm{E}\ X_n = c_n p_n$ may be specified arbitrarily by an appropriate choice of c_n, no matter what nonzero value p_n takes.

Example 3.15 Let X_n be a contaminated standard normal distribution with mixture distribution function

$$F_n(x) = \left(1 - \frac{1}{n}\right)\Phi(x) + \frac{1}{n}G_n(x), \tag{3.8}$$

where $\Phi(x)$ denotes the standard normal distribution function. No matter how the distribution functions G_n are defined, $X_n \xrightarrow{d} \Phi$. However, letting μ_n denote the mean of G_n, $\mathrm{E}\ X_n = \mu_n/n$ may be set arbitrarily by an appropriate choice of G_n.

Consider Example 3.14, specifically Equation (3.7), once again. Note that each X_n in that example can take only two values, 0 or n. In particular, each X_n is bounded, as is any random variable with finite support. Yet taken as a sequence, the X_n are not *uniformly* bounded—that is, there is no single constant that bounds all of the X_n simultaneously.

On the other hand, suppose that a sequence $X_1, X_2, \ldots$ *does* have some uniform bound, say, M such that $|X_n| \leq M$ for all n. Then define the following function:

$$g(t) = \begin{cases} M & \text{if } t > M \\ t & \text{if } |t| \leq M \\ -M & \text{if } t < -M. \end{cases}$$

Note that $g(t)$ is bounded and continuous. Therefore, $X_n \xrightarrow{d} X$ implies that $\mathrm{E}\ g(X_n) \to \mathrm{E}\ g(X)$ by the Helly-Bray Theorem (see Theorem 2.28). Because of the uniform bound on the X_n, we know that $g(X_n)$ is always equal to X_n. We conclude that $\mathrm{E}\ X_n \to \mathrm{E}\ g(X)$, and furthermore it is not difficult to show that $|X|$ must be bounded by M with probability 1, so $\mathrm{E}\ g(X) = \mathrm{E}\ X$. We may summarize this argument by the following Corollary of Theorem 2.28:

Corollary 3.16 If $X_1, X_2, \ldots$ are uniformly bounded (i.e., there exists M such that $|X_n| < M$ for all n) and $X_n \xrightarrow{d} X$, then $\mathrm{E}\ X_n \to \mathrm{E}\ X$.

Corollary 3.16 gives a vague sense of the type of condition—a uniform bound on the X_n—sufficient to ensure that $X_n \xrightarrow{d} X$ implies $\mathrm{E}\ X_n \to \mathrm{E}\ X$. However, the corollary is not very broadly applicable, since many common random variables are not bounded. In Example 3.15, for instance, the X_n could not possibly have a uniform bound (no matter how the G_n are defined) because the standard normal component of each random vector is supported on all of $\mathbb{R}$ and is therefore not bounded. It is thus desirable to generalize the idea of a "uniform bound" of a sequence. There are multiple ways to do this, but probably the best-known generalization is the Dominated Convergence Theorem introduced later in this section. To prove this important theorem, we must first introduce quantile functions and another theorem called the Skorohod Representation Theorem.

3.3.2 Quantile Functions and the Skorohod Representation Theorem

Roughly speaking, the q quantile of a variable X, for some $q \in (0, 1)$, is a value ξ_q such that $P(X \leq \xi_q) = q$. Therefore, if $F(x)$ denotes the distribution function of X, we ought to define $\xi_q = F^{-1}(q)$. However, not all distribution functions $F(x)$ have well-defined inverse functions $F^{-1}(q)$. To understand why not, consider Example 3.17.

Example 3.17 Suppose that $U \sim \text{Uniform}(0, 1)$ and $V \sim \text{Binomial}(2, 0.5)$ are independent random variables. Let $X = V/4 + UV^2/8$. The properties of X are most easily understood by noticing that X is either a constant 0, uniform on $(1/4, 3/8)$, or uniform on $(1/2, 1)$, conditional on $V = 0$, $V = 1$, or $V = 2$, respectively. The distribution function of X, $F(x)$, is shown in Figure 3.2.

There are two problems that can arise when trying to define $F^{-1}(q)$ for an arbitrary $q \in (0, 1)$ and an arbitrary distribution function $F(x)$, and the current example suffers from both: First, in the range $q \in (0, 1/4)$, there is no x for which $F(x)$ equals q because $F(x)$ jumps from 0 to 1/4 at $x = 0$. Second, for $q = 1/4$ or $q = 3/4$, there is not a *unique* x for which $F(x)$ equals q because $F(x)$ is flat (constant) at 1/4 and again at 3/4 for whole intervals of x values.

From Example 3.17, we see that a meaningful general inverse of a distribution function must deal both with "jumps" and "flat spots". The following definition does this.

Definition 3.18 If $F(x)$ is a distribution function, then we define the *quantile func-*

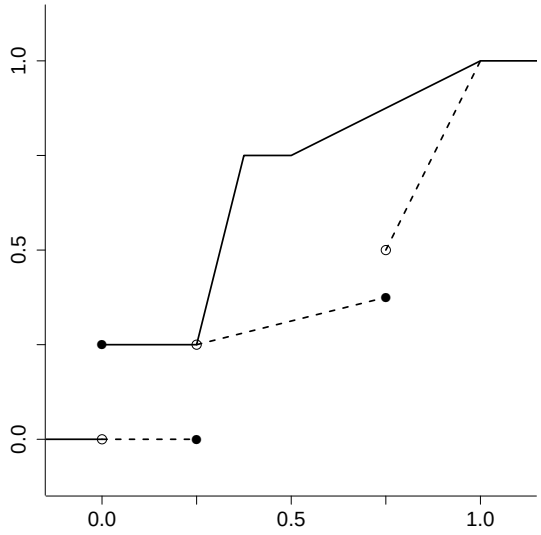

Figure 3.2: *The solid line is a distribution function $F(x)$ that does not have a uniquely defined inverse $F^{-1}(q)$ for all $q \in (0,1)$. However, the dotted quantile function $F^{-}(q)$ of Definition 3.18 is well-defined. Note that $F^{-}(q)$ is the reflection of $F(x)$ over the line $q = x$, so "jumps" in $F(x)$ correspond to "flat spots" in $F^{-}(q)$ and vice versa.*

tion $F^{-} : (0,1) \to \mathbb{R}$ by

$$F^{-}(q) \stackrel{\text{def}}{=} \inf\{x \in \mathbb{R} : q \leq F(x)\}. \tag{3.9}$$

With the quantile function thus defined, we may prove a useful lemma:

Lemma 3.19 $q \leq F(x)$ if and only if $F^{-}(q) \leq x$.

Proof: Using the facts that $F^{-}(\cdot)$ is nondecreasing and $F^{-}[F(x)] \leq x$ by definition,

$$q \leq F(x) \quad \text{implies} \quad F^{-}(q) \leq F^{-}[F(x)] \leq x,$$

which proves the "only if" statement. Conversely, assume that $F^{-}(q) \leq x$. Since $F(\cdot)$ is nondecreasing, we may apply it to both sides of the inequality to obtain

$$F[F^{-}(q)] \leq F(x).$$

Thus, $q \leq F(x)$ follows if we can prove that $q \leq F[F^{-}(q)]$. To this end, consider the set $\{x \in \mathbb{R} : q \leq F(x)\}$ in Equation (3.9). Because any distribution function is right-continuous, this set always contains its infimum, which is $F^{-}(q)$ by definition. This proves that $q \leq F[F^{-}(q)]$. ■

Corollary 3.20 If X is a random variable with distribution function $F(x)$ and U is a uniform$(0,1)$ random variable, then X and $F^-(U)$ have the same distribution.

Proof: For any $x \in \mathbb{R}$, Lemma 3.19 implies that

$$P[F^-(U) \leq x] = P[U \leq F(x)] = F(x).$$

■

Assume now that $X_1, X_2, \ldots$ is a sequence of random variables converging in distribution to X, and let $F_n(x)$ and $F(x)$ denote the distribution functions of X_n and X, respectively. We will show how to construct a new sequence $Y_1, Y_2, \ldots$ such that $Y_n \stackrel{\text{a.s.}}{\to} Y$ and $Y_n \sim F_n(\cdot)$ and $Y \sim F(\cdot)$.

Take $\Omega = (0,1)$ to be a sample space and adopt the probability measure that assigns to each interval subset $(a,b) \subset \Omega$ its length $(b-a)$. [There is a unique probability measure on $(0,1)$ with this property, a fact we do not prove here.] Then for every $\omega \in \Omega$, define

$$Y_n(\omega) \stackrel{\text{def}}{=} F_n^-(\omega) \quad \text{and} \quad Y(\omega) \stackrel{\text{def}}{=} F^-(\omega). \tag{3.10}$$

Note that the random variable defined by $U(\omega) = \omega$ is a uniform$(0,1)$ random variable, so Corollary 3.20 demonstrates that $Y_n \sim F_n(\cdot)$ and $Y \sim F(\cdot)$. It remains to prove that $Y_n \stackrel{\text{a.s.}}{\to} Y$, but once this is proven we will have established the following theorem:

Theorem 3.21 *Skorohod Representation Theorem:* Assume $F, F_1, F_2, \ldots$ are distribution functions and $F_n \stackrel{d}{\to} F$. Then there exist random variables $Y, Y_1, Y_2, \ldots$ such that

1. $P(Y_n \leq t) = F_n(t)$ for all n and $P(Y \leq t) = F(t)$;
2. $Y_n \stackrel{\text{a.s.}}{\to} Y$.

A completion of the proof of Theorem 3.21 is the subject of Exercise 3.10.

Having thus established the Skorohod Represenation Theorem, we now introduce the Dominated Convergence Theorem.

Theorem 3.22 *Dominated Convergence Theorem:* If for a nonnegative random variable Z, $|X_n| \leq Z$ for all n and $\mathrm{E}\ Z < \infty$, then $X_n \stackrel{d}{\to} X$ implies that $\mathrm{E}\ X_n \to \mathrm{E}\ X$.

Proof: Use the Skorohod Representation Theorem to construct a sequence Y_n converging to Y almost surely such that $Y_n \stackrel{d}{=} X_n$ for all n and $Y \stackrel{d}{=} X$. Furthermore, construct a nonnegative random variable Z^* on the same sample space satisfying $Z^* \stackrel{d}{=} Z$ and $|Y_n| \leq Z^*$; this is possible by defining $Z^* = \sup_n |Y_n| + W$, where W is constructed to have the distribution

of $Z - \sup_n |X_n|$ using the idea of expression (3.10). In other words, we now have $Y_n \xrightarrow{\text{a.s.}} Y$, $|Y_n| \leq Z^*$, and $\text{E } Z^* < \infty$. Since $\text{E } X_n = \text{E } Y_n$ for all n and $\text{E } X = \text{E } Y$, it suffices to prove now that $\text{E } Y_n \to \text{E } Y$.

Fatou's Lemma (see Exercise 3.11) states that

$$\text{E} \liminf_n |Y_n| \leq \liminf_n \text{E } |Y_n|. \tag{3.11}$$

A second application of Fatou's Lemma to the nonnegative random variables $Z^* - |Y_n|$ implies

$$\text{E } Z^* - \text{E} \limsup_n |Y_n| \leq \text{E } Z^* - \limsup_n \text{E } |Y_n|.$$

Because $\text{E } Z^* < \infty$, subtracting $\text{E } Z^*$ preserves the inequality, so we obtain

$$\limsup_n \text{E } |Y_n| \leq \text{E} \limsup_n |Y_n|. \tag{3.12}$$

Together, inequalities (3.11) and (3.12) imply

$$\text{E} \liminf_n |Y_n| \leq \liminf_n \text{E } |Y_n| \leq \limsup_n \text{E } |Y_n| \leq \text{E} \limsup_n |Y_n|.$$

Since $Y_n \xrightarrow{\text{a.s.}} Y$, both $\liminf_n |Y_n|$ and $\limsup_n |Y_n|$ are equal to $|Y|$ with probability one, so we conclude that $\lim \text{E } |Y_n|$ exists and is equal to $\text{E } |Y|$. ■

The Dominated Convergence Theorem essentially tells us when it is possible to interchange the operations of limit and expectation, that is, when the limit of the expectations (of the X_n) equals the expectation of their limit.

Exercises for Section 3.3

Exercise 3.9 Prove that any nondecreasing function must have countably many points of discontinuity. (This fact is used in proving the Skorohod Representation Theorem.)

Hint: Use the fact that the set of rational numbers is a countably infinite set and that any real interval must contain a rational number.

Exercise 3.10 To complete the proof of Theorem 3.21, it only remains to show that $Y_n \xrightarrow{\text{a.s.}} Y$, where Y_n and Y are defined as in Equation (3.10).

(a) Let $\delta > 0$ and $\omega \in (0, 1)$ be arbitrary. Show that there exists N_1 such that

$$Y(\omega) - \delta < Y_n(\omega) \tag{3.13}$$

for all $n > N_1$.

Hint: There exists a point of continuity of $F(x)$, say x_0, such that

$$Y(\omega) - \delta < x_0 < Y(\omega).$$

Use the fact that $F_n(x_0) \to F(x_0)$ together with Lemma 3.19 to show how this fact leads to the desired conclusion.

(b) Take δ and ω as in part (a) and let $\epsilon > 0$ be arbitrary. Show that there exists N_2 such that

$$Y_n(\omega) < Y(\omega + \epsilon) + \delta \tag{3.14}$$

for all $n > N_2$.

Hint: There exists a point of continuity of $F(x)$, say x_1, such that

$$Y(\omega + \epsilon) < x_1 < Y(\omega + \epsilon) + \delta.$$

Use the fact that $F_n(x_1) \to F(x_1)$ together with Lemma 3.19 to show how this fact leads to the desired conclusion.

(c) Suppose that $\omega \in (0, 1)$ is a continuity point of $F(x)$. Prove that $Y_n(\omega) \to Y(\omega)$.

Hint: Take $\liminf_n$ in Inequality (3.13) and $\limsup_n$ in Inequality (3.14). Put these inequalities together, then let $\delta \to 0$. Finally, let $\epsilon \to 0$.

(d) Use Exercise 3.9 to prove that $Y_n \xrightarrow{\text{a.s.}} Y$.

Hint: Use countable subadditivity, Inequality (3.5), to show that the set of discontinuity points of $F(x)$ has probability zero.

Exercise 3.11 Prove Fatou's lemma:

$$\mathrm{E} \liminf_n |X_n| \leq \liminf_n \mathrm{E}\ |X_n|. \tag{3.15}$$

Hint: Argue that $\mathrm{E}\ |X_n| \geq \mathrm{E} \inf_{k \geq n} |X_k|$, then take the limit inferior of each side. Use (without proof) the monotone convergence property of the expectation operator: If

$$0 \leq X_1(\omega) \leq X_2(\omega) \leq \cdots \quad \text{and} \quad X_n(\omega) \to X(\omega) \quad \text{for all } \omega \in \Omega,$$

then $\mathrm{E}\ X_n \to \mathrm{E}\ X$.

Exercise 3.12 If $Y_n \xrightarrow{d} Y$, a sufficient condition for $\mathrm{E}\, Y_n \to \mathrm{E}\, Y$ is the **uniform integrability** of the Y_n.

Definition 3.23 The sequence $Y_1, Y_2, \ldots$ of random variables is said to be **uniformly integrable** if

$$\sup_n E\left(|Y_n|\, I\{|Y_n| \geq \alpha\}\right) \to 0 \quad \text{as } \alpha \to \infty.$$

Use the following steps to prove that if $Y_n \xrightarrow{d} Y$ and the Y_n are uniformly integrable, then $\mathrm{E}\, Y_n \to \mathrm{E}\, Y$.

(a) Prove that if $A_1, A_2, \ldots$ and $B_1, B_2, \ldots$ are both uniformly integrable sequences defined on the same probability space, then $A_1 + B_1, A_2 + B_2, \ldots$ is a uniformly integrable sequence.

Hint: First prove that

$$|a+b|I\{|a+b| \geq \alpha\} \leq 2|a|I\{|a| \geq \alpha/2\} + 2|b|I\{|b| \geq \alpha/2\}.$$

(b) Define Z_n and Z such that Z_n has the same distribution as Y_n, Z has the same distribution as Y, and $Z_n \xrightarrow{\text{a.s.}} Z$. (We know that such random variables exist because of the Skorohod Representation Theorem.) Show that if $X_n = |Z_n - Z|$, then $X_1, X_2, \ldots$ is a uniformly integrable sequence.

Hint: Use Fatou's Lemma (Exercise 3.11) to show that $\mathrm{E}\, |Z| < \infty$, i.e., Z is integrable. Then use part (a).

(c) By part (b), the desired result now follows from the following result, which you are asked to prove: If $X_1, X_2, \ldots$ is a uniformly integrable sequence with $X_n \xrightarrow{\text{a.s.}} 0$, then $\mathrm{E}\, X_n \to 0$.

Hint: Use the Dominated Convergence Theorem 3.22 and the fact that

$$\mathrm{E}\, X_n = \mathrm{E}\, X_n I\{|X_n| \geq \alpha\} + \mathrm{E}\, X_n I\{|X_n| < \alpha\}.$$

Exercise 3.13 Prove that if there exists $\epsilon > 0$ such that $\sup_n \mathrm{E}\, |Y_n|^{1+\epsilon} < \infty$, then $Y_1, Y_2, \ldots$ is a uniformly integrable sequence.

Hint: First prove that

$$|Y_n|\, I\{|Y_n| \geq \alpha\} \leq \frac{1}{\alpha^\epsilon}|Y_n|^{1+\epsilon}.$$

Exercise 3.14 Prove that if there exists a random variable Z such that $\mathrm{E}\,|Z| = \mu < \infty$ and $P(|Y_n| \geq t) \leq P(|Z| \geq t)$ for all n and for all $t > 0$, then $Y_1, Y_2, \ldots$ is a uniformly integrable sequence. You may use the fact (without proof) that for a nonnegative X,

$$\mathrm{E}\,(X) = \int_0^\infty P(X \geq t)\, dt.$$

Hints: Consider the random variables $|Y_n| I\{|Y_n| \geq t\}$ and $|Z| I\{|Z| \geq t\}$. In addition, use the fact that

$$\mathrm{E}\,|Z| = \sum_{i=1}^{\infty} \mathrm{E}\,\left(|Z| I\{i-1 \leq |Z| < i\}\right)$$

to argue that $\mathrm{E}\,\left(|Z| I\{|Z| < \alpha\}\right) \to \mathrm{E}\,|Z|$ as $\alpha \to \infty$.

Chapter 4

Central Limit Theorems

The main result of this chapter, in Section 4.2, is the Lindeberg-Feller Central Limit Theorem, from which we obtain the result most commonly known as "The Central Limit Theorem" as a corollary. As in Chapter 3, we mix univariate and multivariate results here. As a general summary, much of Section 4.1 is multivariate and most of the remainder of the chapter is univariate. The interplay between univariate and multivariate results is exemplified by the Central Limit Theorem itself, Theorem 4.9, which is stated for the multivariate case but whose proof is a simple combination of the analagous univariate result with Theorem 4.12, the Cramér-Wold theorem.

Before we discuss central limit theorems, we include one section of background material for the sake of completeness. Section 4.1 introduces the powerful Continuity Theorem, Theorem 4.3, which is the basis for proofs of various important results including the Lindeberg-Feller Theorem. This section also defines multivariate normal distributions.

4.1 Characteristic Functions and Normal Distributions

While it may seem odd to group two such different-sounding topics into the same section, there are actually many points of overlap between characteristic function theory and the multivariate normal distribution. Characteristic functions are essential for proving the Central Limit Theorems of this chapter, which are fundamentally statements about normal distributions. Furthermore, the simplest way to define normal distributions is by using their characteristic functions. The standard univariate method of defining a normal distribution by writing its density does not work here (at least not in a simple way), since not all normal distributions have densities in the usual sense. We even provide a proof of an important result—that characteristic functions determine their distributions uniquely—that uses nor-

mal distributions in an essential way. Thus, the study of characteristic functions and the study of normal distributions are so closely related in statistical large-sample theory that it is perfectly natural for us to introduce them together.

4.1.1 The Continuity Theorem

Definition 4.1 For a random vector $\mathbf{X}$, we define the characteristic function $\phi_{\mathbf{X}} : \mathbb{R}^k \to \mathbb{C}$ by

$$\phi_{\mathbf{X}}(\mathbf{t}) \;=\; \mathrm{E}\,\exp(\mathrm{i}\mathbf{t}^\top\mathbf{X}) \;=\; \mathrm{E}\,\cos(\mathbf{t}^\top\mathbf{X}) + \mathrm{i}\,\mathrm{E}\,\sin(\mathbf{t}^\top\mathbf{X}),$$

where $\mathrm{i}^2 = -1$ and $\mathbb{C}$ denotes the complex numbers.

The characteristic function, which is defined on all of $\mathbb{R}^k$ for *any* $\mathbf{X}$ (unlike the moment generating function, which requires finite moments), has some basic properties. For instance, $\phi_{\mathbf{X}}(\mathbf{t})$ is always a continuous function with $\phi_{\mathbf{X}}(\mathbf{0}) = 1$ and $|\phi_{\mathbf{X}}(\mathbf{t})| \le 1$. Also, inspection of Definition 4.1 reveals that for any constant vector $\mathbf{a}$ and scalar b,

$$\phi_{\mathbf{X}+\mathbf{a}}(\mathbf{t}) = \exp(\mathrm{i}\mathbf{t}^\top\mathbf{a})\phi_{\mathbf{X}}(\mathbf{t}) \;\text{ and }\; \phi_{b\mathbf{X}}(\mathbf{t}) = \phi_{\mathbf{X}}(b\mathbf{t}). \tag{4.1}$$

Also, if $\mathbf{X}$ and $\mathbf{Y}$ are independent,

$$\phi_{\mathbf{X}+\mathbf{Y}}(\mathbf{t}) = \phi_{\mathbf{X}}(\mathbf{t})\phi_{\mathbf{Y}}(\mathbf{t}). \tag{4.2}$$

One of the main reasons that characteristic functions are so useful is the fact that they uniquely determine the distributions from which they are derived. This fact is so important that we state it as a theorem:

Theorem 4.2 The random vectors $\mathbf{X}_1$ and $\mathbf{X}_2$ have the same distribution if and only if $\phi_{\mathbf{X}_1}(\mathbf{t}) = \phi_{\mathbf{X}_2}(\mathbf{t})$ for all $\mathbf{t}$.

Now suppose that $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$, which implies $\mathbf{t}^\top\mathbf{X}_n \xrightarrow{d} \mathbf{t}^\top\mathbf{X}$. Since both $\sin x$ and $\cos x$ are bounded continuous functions, Theorem 2.28 implies that $\phi_{\mathbf{X}_n}(\mathbf{t}) \to \phi_{\mathbf{X}}(\mathbf{t})$. The converse, which is much harder to prove, is also true:

Theorem 4.3 *Continuity Theorem:* $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$ if and only if $\phi_{\mathbf{X}_n}(\mathbf{t}) \to \phi_{\mathbf{X}}(\mathbf{t})$ for all $\mathbf{t}$.

Here is a partial proof that $\phi_{\mathbf{X}_n}(\mathbf{t}) \to \phi_{\mathbf{X}}(\mathbf{t})$ implies $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$. First, we note that the distribution functions F_n must contain a convergent subsequence, say $F_{n_k} \to G$ as $k \to \infty$, where $G : \mathbb{R} \to [0, 1]$ must be a nondecreasing function but G is not necessarily a true distribution function (and, of course, convergence is guaranteed only at continuity points of G). It is possible to define the characteristic function of G—though we will not prove this

assertion—and it must follow that $\phi_{F_{n_k}}(t) \to \phi_G(t)$. But this implies that $\phi_G(t) = \phi_{\mathbf{X}}(t)$ because it was assumed that $\phi_{\mathbf{X}_n}(t) \to \phi_{\mathbf{X}}(t)$. By Theorem 4.2, G must be the distribution function of $\mathbf{X}$. Therefore, every convergent subsequence of $\{\mathbf{X}_n\}$ converges to $\mathbf{X}$, which gives the result.

Theorem 4.3 is an extremely useful tool for proving facts about convergence in distribution. Foremost among these will be the Lindeberg-Feller Theorem in Section 4.2, but other results follow as well. For example, a quick proof of the Cramér-Wold Theorem, Theorem 4.12, is possible (see Exercise 4.3).

4.1.2 Moments

One of the facts that allows us to prove results about distributions using results about characteristic functions is the relationship between the moments of a distribution and the derivatives of a characteristic function. We emphasize here that *all* random variables have well-defined characteristic functions, even if they do not have any moments. What we will see is that existence of moments is related to differentiability of the characteristic function.

We derive $\partial\phi_{\mathbf{X}}(\mathbf{t})/\partial t_j$ directly by considering the limit, if it exists, of

$$\frac{\phi_{\mathbf{X}}(\mathbf{t} + h\mathbf{e}_j) - \phi_{\mathbf{X}}(\mathbf{t})}{h} = \mathrm{E}\left[\exp\{i\mathbf{t}^\top\mathbf{X}\}\left(\frac{\exp\{ihX_j\} - 1}{h}\right)\right]$$

as $h \to 0$, where $\mathbf{e}_j$ denotes the jth unit vector with 1 in the jth component and 0 elsewhere. Note that

$$\left|\exp\{i\mathbf{t}^\top\mathbf{X}\}\left(\frac{\exp\{ihX_j\} - 1}{h}\right)\right| = \left|\int_0^{X_j} \exp\{iht\}\,dt\right| \le |X_j|,$$

so if $\mathrm{E}\,|X_j| < \infty$ then the dominated convergence theorem, Theorem 3.22, implies that

$$\frac{\partial}{\partial t_j}\phi_{\mathbf{X}}(\mathbf{t}) = \mathrm{E}\lim_{h\to 0}\left[\exp\{i\mathbf{t}^\top\mathbf{X}\}\left(\frac{\exp\{ihX_j\} - 1}{h}\right)\right] = i\,\mathrm{E}\left[X_j \exp\{i\mathbf{t}^\top\mathbf{X}\}\right].$$

We conclude that

Lemma 4.4 If $\mathrm{E}\,\|\mathbf{X}\| < \infty$, then $\nabla\phi_{\mathbf{X}}(\mathbf{0}) = i\,\mathrm{E}\,\mathbf{X}$.

A similar argument gives

Lemma 4.5 If $\mathrm{E}\,\mathbf{X}^\top\mathbf{X} < \infty$, then $\nabla^2\phi_{\mathbf{X}}(\mathbf{0}) = -\,\mathrm{E}\,\mathbf{X}\mathbf{X}^\top$.

It is possible to relate higher-order moments of X to higher-order derivatives of $\phi_{\mathbf{X}}(\mathbf{t})$ using the same logic, but for our purposes, only Lemmas 4.4 and 4.5 are needed.

4.1.3 The Multivariate Normal Distribution

It is easy to define a univariate normal distribution. If μ and σ^2 are the mean and variance, respectively, then if $\sigma^2 > 0$ the corresponding normal distribution is by definition the distribution whose density is the well-known function

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

If $\sigma^2 = 0$, on the other hand, we simply take the corresponding normal distribution to be the constant μ. However, it is not quite so easy to define a multivariate normal distribution. This is due to the fact that not all nonconstant multivariate normal distributions have densities on $\mathbb{R}^k$ in the usual sense. It turns out to be much simpler to define multivariate normal distributions using their characteristic functions:

Definition 4.6 Let Σ be any symmetric, nonnegative definite, $k \times k$ matrix and let $\boldsymbol{\mu}$ be any vector in $\mathbb{R}^k$. Then the normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ is defined to be the distribution with characteristic function

$$\phi_{\mathbf{X}}(\mathbf{t}) = \exp\left(\mathrm{i}\mathbf{t}^\top \boldsymbol{\mu} - \frac{\mathbf{t}^\top \Sigma \mathbf{t}}{2}\right). \tag{4.3}$$

Definition 4.6 has a couple of small flaws. First, because it does not stipulate $k \neq 1$, it offers a definition of *univariate* normality that might compete with the already-established definition. However, Exercise 4.1(a) verifies that the two definitions coincide. Second, Definition 4.6 asserts without proof that equation (4.3) actually defines a legitimate characteristic function. How do we know that a distribution with this characteristic function really exists for all possible Σ and $\boldsymbol{\mu}$? There are at least two ways to mend this flaw. One way is to establish sufficient conditions for a particular function to be a legitimate characteristic function, then prove that the function in Equation (4.3) satisfies them. This is possible, but it would take us too far from the aim of this section, which is to establish just enough background to aid the study of statistical large-sample theory. Another method is to construct a random variable whose characteristic function coincides with equation (4.3); yet to do this requires that we delve into some linear algebra. Since this linear algebra will prove useful later, this is the approach we now take.

Before constructing a multivariate normal random vector in full generality, we first consider the case in which Σ is diagonal, say $\Sigma = D = \mathrm{diag}(d_1, \ldots, d_k)$. The stipulation in Definition 4.6 that Σ be nonnegative definite means in this special case that $d_i \geq 0$ for all i. Now take $X_1, \ldots, X_k$ to be independent, univariate normal random variables with zero means and Var $X_i = d_i$. We assert without proof—the assertion will be proven later—that $\mathbf{X} = (X_1, \ldots, X_k)$ is then a multivariate normal random vector, according to Definition 4.6, with mean $\mathbf{0}$ and covariance matrix D.

To define a multivariate normal random vector with a general (non-diagonal) covariance matrix Σ, we make use of the fact that any symmetric matrix may be diagonalized by an orthogonal matrix. We first define orthogonal, then state the diagonalizability result as a lemma that will not be proven here.

Definition 4.7 A square matrix Q is orthogonal if Q^{-1} exists and is equal to $Q^\top$.

Lemma 4.8 If A is a symmetric $k \times k$ matrix, then there exists an orthogonal matrix Q such that $QAQ^\top$ is diagonal.

Note that the diagonal elements of the matrix $QAQ^\top$ in the matrix above must be the eigenvalues of A. This follows since if λ is a diagonal element of $QAQ^\top$, then it is an eigenvalue of $QAQ^\top$. Hence, there exists a vector x such that $QAQ^\top x = \lambda x$, which implies that $A(Q^\top x) = \lambda(Q^\top x)$ and so λ is an eigenvalue of A.

Taking Σ and $\boldsymbol{\mu}$ as in Definition 4.6, Lemma 4.8 implies that there exists an orgthogonal matrix Q such that $Q\Sigma Q^\top$ is diagonal. Since we know that every diagonal entry in $Q\Sigma Q^\top$ is nonnegative, we may define $\mathbf{Y} = (Y_1, \ldots, Y_k)$, where $Y_1, \ldots, Y_k$ are independent normal random vectors with mean zero and Var Y_i equal to the ith diagonal entry of $Q\Sigma Q^\top$. Then the random vector

$$\mathbf{X} = \boldsymbol{\mu} + Q^\top \mathbf{Y} \tag{4.4}$$

has the characteristic function in equation (4.3), a fact whose proof is the subject of Exercise 4.1. Thus, Equation (4.3) of Definition 4.6 always gives the characteristic function of an actual distribution. We denote this multivariate normal distribution by $N_k(\boldsymbol{\mu}, \Sigma)$, or simply $N(\mu, \sigma^2)$ if $k = 1$.

To conclude this section, we point out that in case Σ *is* invertible, then $N_k(\boldsymbol{\mu}, \Sigma)$ has a density in the usual sense on $\mathbb{R}^k$:

$$f(\mathbf{x}) = \frac{1}{\sqrt{2^k \pi^k |\Sigma|}} \exp\left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}, \tag{4.5}$$

where $|\Sigma|$ denotes the determinant of Σ. However, this density will be of little value in the large-sample topics to follow.

4.1.4 Asymptotic Normality

Now that $N_k(\boldsymbol{\mu}, \Sigma)$ is defined, we may use it to state one of the most useful theorems in all of statistical large-sample theory, the Central Limit Theorem for independent and identically distributed (iid) sequences of random vectors. We defer the proof of this theorem to the next section, where we establish a much more general result called the Lindeberg-Feller Central Limit Theorem.

Theorem 4.9 *Central Limit Theorem for independent and identically distributed multivariate sequences:* If $\mathbf{X}_1, \mathbf{X}_2, \ldots$ are independent and identically distributed with mean $\boldsymbol{\mu} \in R^k$ and covariance Σ, where Σ has finite entries, then

$$\sqrt{n}(\overline{\mathbf{X}}_n - \boldsymbol{\mu}) \xrightarrow{d} N_k(\mathbf{0}, \Sigma).$$

Although we refer to several different theorems in this chapter as central limit theorems of one sort or another, we also employ the standard statistical usage in which the phrase "The Central Limit Theorem," with no modifier, refers to Theorem 4.9 or its univariate analogue.

Before exhibiting some examples that apply Theorem 4.9, we discuss what is generally meant by the phrase "asymptotic distribution". Suppose we are given a sequence $X_1, X_2, \ldots$ of random variables and asked to determine the asymptotic distribution of this sequence. This might mean to find X such that $X_n \xrightarrow{d} X$. However, depending on the context, this might not be the case; for example, if $X_n \xrightarrow{d} c$ for a constant c, then we mean something else by "asymptotic distribution".

In general, the "asymptotic distribution of X_n" means a *nonconstant* random variable X, along with real-number sequences $\{a_n\}$ and $\{b_n\}$, such that $a_n(X_n - b_n) \xrightarrow{d} X$. In this case, the distribution of X might be referred to as the asymptotic or limiting distribution of either X_n or of $a_n(X_n - b_n)$, depending on the context.

Example 4.10 Suppose that X_n is the sum of n independent Bernoulli(p) random variables, so that $X_n \sim$ binomial(n, p). Even though we know that $X_n/n \xrightarrow{P} p$ by the weak law of large numbers, this is not generally what we mean by the asymptotic distribution of X_n/n. Instead, the asymptotic distribution of X_n/n is expressed by

$$\sqrt{n}\left(\frac{X_n}{n} - p\right) \xrightarrow{d} N\{0, p(1-p)\},$$

which follows from the Central Limit Theorem because a Bernoulli(p) random variable has mean p and variance $p(1-p)$.

Example 4.11 *Asymptotic distribution of sample variance:* Suppose that $X_1, X_2, \ldots$ are independent and identically distributed with $\mathrm{E}(X_i) = \mu$, $\mathrm{Var}(X_i) = \sigma^2$, and $\mathrm{Var}\{(X_i - \mu)^2\} = \tau^2 < \infty$. Define

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2. \tag{4.6}$$

We wish to determine the asymptotic distribution of S_n^2.

Since the distribution of $X_i - \overline{X}_n$ does not change if we replace each X_i by $X_i - \mu$, we may assume without loss of generality that $\mu = 0$. By the Central Limit Theorem, we know that

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \sigma^2\right) \xrightarrow{d} N(0, \tau^2).$$

Furthermore, the Central Limit Theorem and the Weak Law imply $\sqrt{n}(\overline{X}_n) \xrightarrow{d} N(0, \sigma^2)$ and $\overline{X}_n \xrightarrow{P} 0$, respectively, so Slutsky's theorem implies $\sqrt{n}\left(\overline{X}_n^2\right) \xrightarrow{P} 0$. Therefore, since

$$\sqrt{n}(S_n^2 - \sigma^2) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \sigma^2\right) + \sqrt{n}\left(\overline{X}_n^2\right),$$

Slutsky's theorem implies that $\sqrt{n}\,(S_n^2 - \sigma^2) \xrightarrow{d} N(0, \tau^2)$, which is the desired result.

Note that the definition of S_n^2 in Equation 4.6 is not the usual unbiased sample variance, which uses the denominator $n - 1$ instead of n. However, since

$$\sqrt{n}\left(\frac{n}{n-1}S_n^2 - \sigma^2\right) = \sqrt{n}(S_n^2 - \sigma^2) + \frac{\sqrt{n}}{n-1}S_n^2$$

and $\sqrt{n}/(n-1) \to 0$, we see that the simpler choice of n does not change the asymptotic distribution at all.

4.1.5 The Cramér-Wold Theorem

Suppose that $\mathbf{X}_1, \mathbf{X}_2, \ldots$ is a sequence of random k-vectors. By Theorem 2.34, we see immediately that

$$\mathbf{X}_n \xrightarrow{d} \mathbf{X} \text{ implies } \mathbf{a}^\top \mathbf{X}_n \xrightarrow{d} \mathbf{a}^\top \mathbf{X} \text{ for any } \mathbf{a} \in \mathbb{R}^k. \tag{4.7}$$

This is because multiplication by a constant vector $\mathbf{a}^T$ is a continuous transformation from $\mathbb{R}^k$ to $\mathbb{R}$. It is not clear, however, whether the converse of statement (4.7) is true. Such a converse would be useful because it would give a means for proving multivariate convergence in distribution using only univariate methods. As the counterexample in Example 2.38 shows, multivariate convergence in distribution does *not* follow from the mere fact that each of the components converges in distribution. Yet the converse of statement (4.7) is much stronger than the statement that each component converges in distribution; could it be true that requiring *all* linear combinations to converge in distribution is strong enough to guarantee multivariate convergence? The answer is yes:

Theorem 4.12 *Cramér-Wold Theorem:* $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$ if and only if $\mathbf{a}^\top \mathbf{X}_n \xrightarrow{d} \mathbf{a}^\top \mathbf{X}$ for all $\mathbf{a} \in \mathbb{R}^k$.

Using the machinery of characteristic functions, to be presented in Section 4.1, the proof of the Cramér-Wold Theorem is immediate; see Exercise 4.3. This theorem in turn provides a straightforward method for proving cerain multivariate theorems using univariate results. For instance, once we establish the univariate Central Limit Theorem (Theorem 4.19), we will show how to use the Cramér-Wold Theorem to prove the multivariate CLT, Theorem 4.9.

Exercises for Section 4.1

Exercise 4.1 **(a)** Prove that if $Y \sim N(0, \sigma^2)$ with $\sigma^2 > 0$, then $\phi_Y(t) = \exp\left(-\frac{1}{2}t^2\sigma^2\right)$. Argue that this demonstrates that Definition 4.6 is valid in the case $k = 1$.

Hint: Verify and solve the differential equation $\phi_Y'(t) = -t\sigma^2\phi_Y(t)$. Use integration by parts.

(b) Using part (a), prove that if $\mathbf{X}$ is defined as in Equation (4.4), then $\phi_{\mathbf{X}}(\mathbf{t}) = \exp\left(i\mathbf{t}^\top\boldsymbol{\mu} - \frac{1}{2}\mathbf{t}^\top\Sigma\mathbf{t}\right)$.

Exercise 4.2 We will prove Theorem 4.2, which states that chacteristic functions uniquely determine their distributions.

(a) First, prove the *Parseval relation* for random $\mathbf{X}$ and $\mathbf{Y}$:

$$\mathrm{E}\left[\exp(-\mathrm{i}\mathbf{a}^\top\mathbf{Y})\phi_{\mathbf{X}}(\mathbf{Y})\right] = \mathrm{E}\ \phi_{\mathbf{Y}}(\mathbf{X} - \mathbf{a}).$$

Hint: Use conditioning to evaluate $\mathrm{E}\exp\{\mathrm{i}(\mathbf{X} - \mathbf{a})^\top\mathbf{Y}\}$.

(b) Suppose that $\mathbf{Y} = (Y_1, \ldots, Y_k)$, where $Y_1, \ldots, Y_k$ are independent and identically distributed normal random variables with mean 0 and variance σ^2. That is, $\mathbf{Y}$ has density

$$f_{\mathbf{Y}}(\mathbf{y}) = (\sqrt{2\pi\sigma^2})^{-k}\exp(-\mathbf{y}^\top\mathbf{y}/2\sigma^2).$$

Show that $\mathbf{X} + \mathbf{Y}$ has density

$$f_{\mathbf{X}+\mathbf{Y}}(\mathbf{s}) = \mathrm{E}\ f_{\mathbf{Y}}(\mathbf{s} - \mathbf{X}).$$

(c) Use the result of Exercise 4.1 along with part (b) to show that

$$f_{\mathbf{X}+\mathbf{Y}}(\mathbf{s}) = (\sqrt{2\pi\sigma^2})^{-k}\,\mathrm{E}\ \phi_{\mathbf{Y}}\left(\frac{\mathbf{X}}{\sigma^2} - \frac{s}{\sigma^2}\right).$$

Argue that this fact proves $\phi_{\mathbf{X}}(\mathbf{t})$ uniquely determines the distribution of $\mathbf{X}$.

Hint: Use parts (a) and (b) to show that the distribution of $\mathbf{X}+\mathbf{Y}$ depends on $\mathbf{X}$ only through $\phi_{\mathbf{X}}$. Then note that $\mathbf{X}+\mathbf{Y} \xrightarrow{d} \mathbf{X}$ as $\sigma^2 \to 0$.

Exercise 4.3 Use the Continuity Theorem to prove the Cramér-Wold Theorem, Theorem 4.12.

Hint: $\mathbf{a}^\top \mathbf{X}_n \xrightarrow{d} \mathbf{a}^\top \mathbf{X}$ implies that $\phi_{\mathbf{a}^\top \mathbf{X}_n}(1) \to \phi_{\mathbf{a}^\top \mathbf{X}}(1)$.

Exercise 4.4 Suppose $\mathbf{X} \sim N_k(\boldsymbol{\mu}, \Sigma)$, where Σ is invertible. Prove that

$$(\mathbf{X}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{X}-\boldsymbol{\mu}) \sim \chi^2_k.$$

Hint: If Q diagonalizes Σ, say $Q\Sigma Q^\top = \Lambda$, let $\Lambda^{1/2}$ be the diagonal, nonnegative matrix satisfying $\Lambda^{1/2}\Lambda^{1/2} = \Lambda$ and consider $\mathbf{Y}^\top\mathbf{Y}$, where $\mathbf{Y} = (\Lambda^{1/2})^{-1}Q(\mathbf{X}-\boldsymbol{\mu})$.

Exercise 4.5 Let $X_1, X_2, \ldots$ be independent Poisson random variables with mean $\lambda = 1$. Define $Y_n = \sqrt{n}(\overline{X}_n - 1)$.

(a) Find $\mathrm{E}\,(Y_n^+)$, where $Y_n^+ = Y_n I\{Y_n > 0\}$.

(b) Find, with proof, the limit of $\mathrm{E}\,(Y_n^+)$ and prove Stirling's formula

$$n! \;\sim\; \sqrt{2\pi}\, n^{n+1/2} e^{-n}.$$

Hint: Use the result of Exericse 3.12.

Exercise 4.6 Use the Continuity Theorem to prove Theorem 2.19, the univariate Weak Law of Large Numbers.

Hint: Use a Taylor expansion (1.5) with $d = 2$ for both the real and imaginary parts of the characteristic function of $\overline{X}_n$.

Exercise 4.7 Use the Cramér-Wold Theorem along with the univariate Central Limit Theorem (from Example 2.12) to prove Theorem 4.9.

4.2 The Lindeberg-Feller Central Limit Theorem

The Lindeberg-Feller Central Limit Theorem states in part that sums of independent random variables, properly standardized, converge in distribution to standard normal as long as a certain condition, called the Lindeberg Condition, is satisfied. Since these random variables do not have to be identically distributed, this result generalizes the Central Limit Theorem for independent and identically distributed sequences.

4.2.1 The Lindeberg and Lyapunov Conditions

Suppose that $X_1, X_2, \ldots$ are independent random variables such that $\mathrm{E}\, X_n = \mu_n$ and $\mathrm{Var}\, X_n = \sigma_n^2 < \infty$. Define

$$\begin{aligned} Y_n &= X_n - \mu_n, \\ T_n &= \textstyle\sum_{i=1}^n Y_i, \\ s_n^2 &= \mathrm{Var}\, T_n = \textstyle\sum_{i=1}^n \sigma_i^2. \end{aligned}$$

Instead of defining Y_n to be the centered version of X_n, we could have simply taken μ_n to be zero without loss of generality. However, when these results are used in practice, it is easy to forget the centering step, so we prefer to make it explicit here.

Note that T_n/s_n has mean zero and variance 1. We wish to give sufficient conditions that ensure $T_n/s_n \xrightarrow{d} N(0,1)$. We give here two separate conditions, one called the Lindeberg condition and the other called the Lyapunov condition. The *Lindeberg Condition* for sequences states that

$$\text{for every } \epsilon > 0,\ \frac{1}{s_n^2}\sum_{i=1}^n \mathrm{E}\left(Y_i^2 I\left\{|Y_i| \geq \epsilon s_n\right\}\right) \to 0 \text{ as } n \to \infty; \tag{4.8}$$

the *Lyapunov Condition* for sequences states that

$$\text{there exists } \delta > 0 \text{ such that } \frac{1}{s_n^{2+\delta}}\sum_{i=1}^n \mathrm{E}\left(|Y_i|^{2+\delta}\right) \to 0 \text{ as } n \to \infty. \tag{4.9}$$

We shall see later (in Theorem 4.16, the Lindeberg-Feller Theorem) that Condition (4.8) implies $T_n/s_n \to N(0,1)$. For now, we show only that Condition (4.9)—the Lyapunov Condition—is stronger than Condition (4.8). Thus, the Lyapunov Condition also implies $T_n/s_n \to N(0,1)$:

Theorem 4.13 The Lyapunov Condition (4.9) implies the Lindeberg Condition (4.8).

Proof: Assume that the Lyapunov Condition is satisfied and fix $\epsilon > 0$. Since $|Y_i| \geq \epsilon s_n$ implies $|Y_i/\epsilon s_n|^\delta \geq 1$, we obtain

$$\begin{aligned} \frac{1}{s_n^2}\sum_{i=1}^n \mathrm{E}\left(Y_i^2 I\left\{|Y_i| \geq \epsilon s_n\right\}\right) &\leq \frac{1}{\epsilon^\delta s_n^{2+\delta}}\sum_{i=1}^n \mathrm{E}\left(|Y_i|^{2+\delta} I\left\{|Y_i| \geq \epsilon s_n\right\}\right) \\ &\leq \frac{1}{\epsilon^\delta s_n^{2+\delta}}\sum_{i=1}^n \mathrm{E}\left(|Y_i|^{2+\delta}\right). \end{aligned}$$

Since the right hand side tends to 0, the Lindeberg Condition is satisfied. ■

Example 4.14 Suppose that we perform a series of independent Bernoulli trials with possibly different success probabilities. Under what conditions will the proportion of successes, properly standardized, tend to a normal distribution?

Let $X_n \sim \text{Bernoulli}(p_n)$, so that $Y_n = X_n - p_n$ and $\sigma_n^2 = p_n(1-p_n)$. As we shall see later (Theorem 4.16), either the Lindeberg Condition (4.8) or the Lyapunov Condition (4.9) will imply that $\sum_{i=1}^n Y_i/s_n \xrightarrow{d} N(0,1)$.

Let us check the Lyapunov Condition for, say, $\delta = 1$. First, verify that

$$\text{E}\,|Y_n|^3 = p_n(1-p_n)^3 + (1-p_n)p_n^3 = \sigma_n^2[(1-p_n)^2 - p_n^2] \leq \sigma_n^2.$$

Using this upper bound on $\text{E}\,|Y_n|^3$, we obtain $\sum_{i=1}^n \text{E}\,|Y_i|^3 \leq s_n^2$. Therefore, the Lyapunov condition is satisfied whenever $s_n^2/s_n^3 \to 0$, which implies $s_n \to \infty$. We conclude that the proportion of successes tends to a normal distribution whenever

$$s_n^2 = \sum_{i=1}^n p_n(1-p_n) \to \infty,$$

which will be true as long as $p_n(1-p_n)$ does not tend to 0 too fast.

4.2.2 Independent and Identically Distributed Variables

We now set the stage for proving a central limit theorem for independent and identically distributed random variables by showing that the Lindeberg Condition is satisfied by such a sequence as long as the common variance is finite.

Example 4.15 Suppose that $X_1, X_2, \ldots$ are independent and identically distributed with $\text{E}\,(X_i) = \mu$ and $\text{Var}\,(X_i) = \sigma^2 < \infty$. The case $\sigma^2 = 0$ is uninteresting, so we assume $\sigma^2 > 0$.

Let $Y_i = X_i - \mu$ and $s_n^2 = \text{Var} \sum_{i=1}^n Y_i = n\sigma^2$. Fix $\epsilon > 0$. The Lindeberg Condition states that

$$\frac{1}{n\sigma^2} \sum_{i=1}^n \text{E}\,\left(Y_i^2 I\{|Y_i| \geq \epsilon\sigma\sqrt{n}\}\right) \to 0 \quad \text{as } n \to \infty. \tag{4.10}$$

Since the Y_i are identically distributed, the left hand side of expression (4.10) simplifies to

$$\frac{1}{\sigma^2}\,\text{E}\,\left(Y_1^2 I\{|Y_1| \geq \epsilon\sigma\sqrt{n}\}\right). \tag{4.11}$$

To simplify notation, let Z_n denote the random variable $Y_1^2 I\{|Y_1| \geq \epsilon\sigma\sqrt{n}\}$. Thus, we wish to prove that E $Z_n \to 0$. Note that Z_n is nonzero if and only if $|Y_1| \geq \epsilon\sigma\sqrt{n}$. Since this event has probability tending to zero as $n \to \infty$, we conclude that $Z_n \xrightarrow{P} 0$ by the definition of convergence in probability. We can also see that $|Z_n| \leq Y_1^2$, and we know that E $Y_1^2 < \infty$. Therefore, we may apply the Dominated Convergence Theorem, Theorem 3.22, to conclude that E $Z_n \to 0$. This demonstrates that the Lindeberg Condition is satisfied.

The preceding argument, involving the Dominated Convergence Theorem, is quite common in proofs that the Lindeberg Condition is satisfied. Any beginning student is well-advised to study this argument carefully.

Note that the assumptions of Example 4.15 are not strong enough to ensure that the Lyapunov Condition (4.9) is satisfied. This is because there are some random variables that have finite variances but no finite $2 + \delta$ moment for any $\delta > 0$. Construction of such an example is the subject of Exercise 4.10. However, such examples are admittedly somewhat pathological, and if one is willing to assume that $X_1, X_2, \ldots$ are independent and identically distributed with E $|X_1|^{2+\delta} = \gamma < \infty$ for some $\delta > 0$, then the Lyapunov Condition is much easier to check than the Lindeberg Condition. Indeed, because $s_n = \sigma\sqrt{n}$, the Lyapunov Condition reduces to

$$\frac{n\gamma}{(n\sigma^2)^{1+\delta/2}} = \frac{\gamma}{n^{\delta/2}\sigma^{2+\delta}} \to 0,$$

which follows immediately.

4.2.3 Triangular Arrays

It is sometimes the case that $X_1, \ldots, X_n$ are independent random variables—possibly even identically distributed—but their distributions depend on n. Take the simple case of the binomial(n, p_n) distribution as an example, where the probability p_n of success on any trial changes as n increases. What can we say about the asymptotic distribution in such a case? It seems that what we need is some way of dealing with a sequence of sequences, say, $X_{n1}, \ldots, X_{nn}$ for $n \geq 1$. This is exactly the idea of a triangular array of random variables.

Generalizing the concept of "sequence of independent random variables," a triangular array or random variables may be visualized as follows:

$$\begin{array}{llll} X_{11} & & & \leftarrow \text{independent} \\ X_{21} & X_{22} & & \leftarrow \text{independent} \\ X_{31} & X_{32} & X_{33} & \leftarrow \text{independent} \\ & \vdots & & \end{array}$$

Thus, we assume that for each n, $X_{n1}, \ldots, X_{nn}$ are independent. Carrying over the notation from before, we assume E $X_{ni} = \mu_{ni}$ and Var $X_{ni} = \sigma_{ni}^2 < \infty$. Let

$$\begin{aligned} Y_{ni} &= X_{ni} - \mu_{ni}, \\ T_n &= \textstyle\sum_{i=1}^n Y_{ni}, \\ s_n^2 &= \text{Var } T_n = \textstyle\sum_{i=1}^n \sigma_{ni}^2. \end{aligned}$$

As before, T_n/s_n has mean 0 and variance 1; our goal is to give conditions under which

$$\frac{T_n}{s_n} \xrightarrow{d} N(0,1). \tag{4.12}$$

Such conditions are given in the Lindeberg-Feller Central Limit Theorem. The key to this theorem is the *Lindeberg condition* for triangular arrays:

$$\text{For every } \epsilon > 0, \ \frac{1}{s_n^2} \sum_{i=1}^n \text{E}\left(Y_{ni}^2 I\left\{|Y_{ni}| \geq \epsilon s_n\right\}\right) \to 0 \text{ as } n \to \infty. \tag{4.13}$$

Before stating the Lindeberg-Feller theorem, we need a technical condition that says essentially that the contribution of each X_{ni} to s_n^2 should be negligible:

$$\frac{1}{s_n^2} \max_{i \leq n} \sigma_{ni}^2 \to 0 \text{ as } n \to \infty. \tag{4.14}$$

Now that Conditions (4.12), (4.13), and (4.14) have been written, the main result may be stated in a single line:

Theorem 4.16 *Lindeberg-Feller Central Limit Theorem:* Condition (4.13) holds if and only if Conditions (4.12) and (4.14) hold.

A proof of the Lindeberg-Feller Theorem is the subject of Exercises 4.8 and 4.9. In most practical applications of this theorem, the Lindeberg Condition (4.13) is used to establish asymptotic normality (4.12); the remainder of the theorem's content is less useful.

Example 4.17 As an extension of Example 4.14, suppose $X_n \sim \text{binomial}(n, p_n)$. The calculations here are not substantially different from those in Example 4.14, so we use the Lindeberg Condition here for the purpose of illustration. We claim that

$$\frac{X_n - np_n}{\sqrt{np_n(1-p_n)}} \xrightarrow{d} N(0,1) \tag{4.15}$$

whenever $np_n(1-p_n) \to \infty$ as $n \to \infty$. In order to use Theorem 4.16 to prove this result, let $Y_{n1}, \ldots, Y_{nn}$ be independent and identically distributed with

$$P(Y_{ni} = 1 - p_n) = 1 - P(Y_{ni} = -p_n) = p_n.$$

Then with $X_n = np_n + \sum_{i=1}^n Y_{ni}$, we obtain $X_n \sim$ binomial(n, p_n) as specified. Furthermore, $\mathrm{E}\, Y_{ni} = 0$ and $\mathrm{Var}\, Y_{ni} = p_n(1-p_n)$, so the Lindeberg condition says that for any $\epsilon > 0$,

$$\frac{1}{np_n(1-p_n)} \sum_{i=1}^n \mathrm{E}\, \left(Y_{ni}^2 I\left\{ |Y_{ni}| \geq \epsilon \sqrt{np_n(1-p_n)} \right\} \right) \to 0. \tag{4.16}$$

Since $|Y_{ni}| \leq 1$, the left hand side of expression (4.16) will be identically zero whenever $\epsilon\sqrt{np_n(1-p_n)} > 1$. Thus, a sufficient condition for (4.15) to hold is that $np_n(1-p_n) \to \infty$. One may show that this is also a necessary condition (this is Exercise 4.11).

Note that any independent sequence $X_1, X_2, \ldots$ may be considered a triangular array by simply taking $X_{n1} = X_1$ for all $n \geq 1$, $X_{n2} = X_2$ for all $n \geq 2$, and so on. Therefore, Theorem 4.16 applies equally to the Lindeberg Condition (4.8) for sequences. Furthermore, the proof of Theorem 4.13 is unchanged if the sequence Y_i is replaced by the array Y_{ni}. Therefore, we obtain an alternative means for checking asymptotic normality:

Corollary 4.18 Asymptotic normality (4.12) follows if the triangular array above satisfies the *Lyapunov Condition* for triangular arrays:

$$\text{there exists } \delta > 0 \text{ such that } \frac{1}{s_n^{2+\delta}} \sum_{i=1}^n \mathrm{E}\, \left(|Y_{ni}|^{2+\delta} \right) \to 0 \text{ as } n \to \infty. \tag{4.17}$$

Combining Theorem 4.16 with Example 4.15, in which the Lindeberg condition is verified for a sequence of independent and identically distributed variables with finite positive variance, gives the result commonly referred to simply as "The Central Limit Theorem":

Theorem 4.19 *Univariate Central Limit Theorem for iid sequences:* Suppose that $X_1, X_2, \ldots$ are independent and identically distributed with $\mathrm{E}\,(X_i) = \mu$ and $\mathrm{Var}\,(X_i) = \sigma^2 < \infty$. Then

$$\sqrt{n}\left(\overline{X}_n - \mu \right) \xrightarrow{d} N(0, \sigma^2). \tag{4.18}$$

The case $\sigma^2 = 0$ is not covered by Example 4.15, but in this case limit (4.18) holds automatically.

We conclude this section by generalizing Theorem 4.19 to the multivariate case, Theorem 4.9. The proof is straightforward using theorem 4.19 along with the Cramér-Wold theorem, theorem 4.12. Recall that the Cramér-Wold theorem allows us to establish multivariate convergence in distribution by proving univariate convergence in distribution for arbitrary linear combinations of the vector components.

Proof of Theorem 4.9: Let $\mathbf{X} \sim N_k(\mathbf{0}, \Sigma)$ and take any vector $\mathbf{a} \in \mathbb{R}^k$. We wish to show that

$$\mathbf{a}^\top \left[\sqrt{n}\left(\overline{\mathbf{X}}_n - \boldsymbol{\mu}\right)\right] \xrightarrow{d} \mathbf{a}^\top \mathbf{X}.$$

But this follows immediately from the univariate Central Limit Theorem, since $\mathbf{a}^\top(\mathbf{X}_1 - \boldsymbol{\mu}), \mathbf{a}^\top(\mathbf{X}_2 - \boldsymbol{\mu}), \ldots$ are independent and identically distributed with mean 0 and variance $\mathbf{a}^\top \Sigma \mathbf{a}$. ■

We will see many, many applications of the univariate and multivariate Central Limit Theorems in the chapters that follow.

Exercises for Section 4.2

Exercise 4.8 Prove that (4.13) implies both (4.12) and (4.14) (the "forward half" of the Lindeberg-Feller Theorem). Use the following steps:

(a) Prove that for any complex numbers $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ with $|a_i| \leq 1$ and $|b_i| \leq 1$,

$$|a_1 \cdots a_n - b_1 \cdots b_n| \leq \sum_{i=1}^{n} |a_i - b_i| \,. \tag{4.19}$$

Hint: First prove the identity when $n = 2$, which is the key step. Then use mathematical induction.

(b) Prove that

$$\left|\phi_{Y_{ni}}\left(\frac{t}{s_n}\right) - \left(1 - \frac{t^2\sigma_{ni}^2}{2s_n^2}\right)\right| \leq \frac{\epsilon|t|^3\sigma_{ni}^2}{s_n^2} + \frac{t^2}{s_n^2}\,\mathrm{E}\left(Y_{ni}^2 I\{|Y_{ni}| \geq \epsilon s_n\}\right). \tag{4.20}$$

Hint: Use the results of Exercise 1.43, parts (c) and (d), to argue that for any Y,

$$\left|\exp\left\{\frac{\mathrm{i}tY}{s_n}\right\} - \left(1 + \frac{\mathrm{i}tY}{s_n} - \frac{t^2Y^2}{2s_n^2}\right)\right| \leq \left|\frac{tY}{s_n}\right|^3 I\left\{\left|\frac{Y}{s_n}\right| < \epsilon\right\} + \left(\frac{tY}{s_n}\right)^2 I\{|Y| \geq \epsilon s_n\}.$$

(c) Prove that (4.13) implies (4.14).

Hint: For any i, show that

$$\frac{\sigma_{ni}^2}{s_n^2} < \epsilon^2 + \frac{\mathrm{E}\left(Y_{ni}^2 I\{|Y_{ni}| \geq \epsilon s_n\}\right)}{s_n^2}.$$

(d) Use parts (a) and (b) to prove that, for n large enough so that $t^2 \max_i \sigma_{ni}^2/s_n^2 \leq 1$,

$$\left| \prod_{i=1}^{n} \phi_{Y_{ni}}\left(\frac{t}{s_n}\right) - \prod_{i=1}^{n}\left(1 - \frac{t^2\sigma_{ni}^2}{2s_n^2}\right)\right| \leq \epsilon|t|^3 + \frac{t^2}{s_n^2}\sum_{i=1}^{n} \mathrm{E}\left(Y_{ni}^2 I\left\{|Y_{ni}| \geq \epsilon s_n\right\}\right).$$

(e) Use part (a) to prove that

$$\left| \prod_{i=1}^{n}\left(1 - \frac{t^2\sigma_{ni}^2}{2s_n^2}\right) - \prod_{i=1}^{n} \exp\left(-\frac{t^2\sigma_{ni}^2}{2s_n^2}\right)\right| \leq \frac{t^4}{4s_n^4}\sum_{i=1}^{n}\sigma_{ni}^4 \leq \frac{t^4}{4s_n^2}\max_{1\leq i\leq n}\sigma_{ni}^2.$$

Hint: Prove that for $x \leq 0$, $|1 + x - \exp(x)| \leq x^2$.

(f) Now put it all together. Show that

$$\left| \prod_{i=1}^{n} \phi_{Y_{ni}}\left(\frac{t}{s_n}\right) - \prod_{i=1}^{n} \exp\left(-\frac{t^2\sigma_{ni}^2}{2s_n^2}\right)\right| \to 0,$$

proving (4.12).

Exercise 4.9 In this problem, we prove the converse of Exercise 4.8, which is the part of the Lindeberg-Feller Theorem due to Feller: Under the assumptions of the Exercise 4.8, the variance condition (4.14) and the asymptotic normality (4.12) together imply the Lindeberg condition (4.13).

(a) Define

$$\alpha_{ni} = \phi_{Y_{ni}}(t/s_n) - 1.$$

Prove that

$$\max_{1\leq i\leq n}|\alpha_{ni}| \leq 2\max_{1\leq i\leq n} P\left(|Y_{ni}| \geq \epsilon s_n\right) + 2\epsilon|t|$$

and thus

$$\max_{1\leq i\leq n}|\alpha_{ni}| \to 0 \quad \text{as } n \to \infty.$$

Hint: Use the result of Exercise 1.43(a) to show that $|\exp\{\mathrm{i}t\} - 1| \leq 2\min\{1, |t|\}$ for $t \in \mathbb{R}$. Then use Chebyshev's inequality along with condition (4.14).

(b) Use part (a) to prove that

$$\sum_{i=1}^{n} |\alpha_{ni}|^2 \to 0$$

as $n \to \infty$.

Hint: Use the result of Exercise 1.43(b) to show that $|\alpha_{ni}| \leq t^2\sigma_{ni}^2/s_n^2$. Then write $|\alpha_{ni}|^2 \leq |\alpha_{ni}| \max_i |\alpha_{ni}|$.

(c) Prove that for n large enough so that $\max_i |\alpha_{ni}| \leq 1/2$,

$$\left| \prod_{i=1}^{n} \exp(\alpha_{ni}) - \prod_{i=1}^{n} (1 + \alpha_{ni}) \right| \leq \sum_{i=1}^{n} |\alpha_{ni}|^2.$$

Hint: Use the fact that $|\exp(z-1)| = \exp(\text{Re } z - 1) \leq 1$ for $|z| \leq 1$ to argue that Inequality (4.19) applies. Also use the fact that $|\exp(z) - 1 - z| \leq |z|^2$ for $|z| \leq 1/2$.

(d) From part (c) and condition (4.12), conclude that

$$\sum_{i=1}^{n} \text{Re}\,(\alpha_{ni}) \to -\frac{1}{2}t^2.$$

(e) Show that

$$0 \leq \sum_{i=1}^{n} \text{E}\left(\cos\frac{tY_{ni}}{s_n} - 1 + \frac{t^2 Y_{ni}^2}{2s_n^2} \right) \to 0.$$

(f) For arbitrary $\epsilon > 0$, choose t large enough so that $t^2/2 > 2/\epsilon^2$. Show that

$$\left(\frac{t^2}{2} - \frac{2}{\epsilon^2} \right) \frac{1}{s_n^2} \sum_{i=1}^{n} \text{E}\left(Y_{ni}^2 I\{|Y_{ni}| \geq \epsilon s_n\} \right) \leq \sum_{i=1}^{n} \text{E}\left(\cos\frac{tY_{ni}}{s_n} - 1 + \frac{t^2 Y_{ni}^2}{2s_n^2} \right),$$

which completes the proof.

Hint: Bound the expression in part (e) below by using the fact that -1 is a lower bound for $\cos x$. Also note that $|Y_{ni}| \geq \epsilon s_n$ implies $-2 \geq -2Y_{ni}^2/(\epsilon^2 s_n^2)$.

Exercise 4.10 Give an example of an independent and identically distributed sequence to which the Central Limit Theorem 4.19 applies but for which the Lyapunov condition is not satisfied.

Exercise 4.11 In Example 4.17, we show that $np_n(1-p_n) \to \infty$ is a sufficient condition for (4.15) to hold. Prove that it is also a necessary condition. You may assume that $p_n(1-p_n)$ is always nonzero.

Hint: Use the Lindeberg-Feller Theorem.

Exercise 4.12 **(a)** Suppose that $X_1, X_2, \ldots$ are independent and identically distributed with $\mathrm{E}\ X_i = \mu$ and $0 < \mathrm{Var}\ X_i = \sigma^2 < \infty$. Let $a_{n1}, \ldots, a_{nn}$ be constants satisfying

$$\frac{\max_{i \le n} a_{ni}^2}{\sum_{j=1}^n a_{nj}^2} \to 0 \quad \text{as } n \to \infty.$$

Let $T_n = \sum_{i=1}^n a_{ni} X_i$, and prove that $(T_n - \mathrm{E}\ T_n)/\sqrt{\mathrm{Var}\ T_n} \xrightarrow{d} N(0,1)$.

(b) Reconsider Example 2.22, the simple linear regression case in which

$$\hat{\beta}_{0n} = \sum_{i=1}^n v_i^{(n)} Y_i \quad \text{and} \quad \hat{\beta}_{1n} = \sum_{i=1}^n w_i^{(n)} Y_i,$$

where

$$w_i^{(n)} = \frac{z_i - \overline{z}_n}{\sum_{j=1}^n (z_j - \overline{z}_n)^2} \quad \text{and} \quad v_i^{(n)} = \frac{1}{n} - \overline{z}_n w_i^{(n)}$$

for constants $z_1, z_2, \ldots$. Using part (a), state and prove sufficient conditions on the constants z_i that ensure the asymptotic normality of $\sqrt{n}(\hat{\beta}_{0n} - \beta_0)$ and $\sqrt{n}(\hat{\beta}_{1n} - \beta_1)$. You may assume the results of Example 2.22, where it was shown that $\mathrm{E}\ \hat{\beta}_{0n} = \beta_0$ and $\mathrm{E}\ \hat{\beta}_{1n} = \beta_1$.

Exercise 4.13 Give an example (with proof) of a sequence of independent random variables $Z_1, Z_2, \ldots$ with $\mathrm{E}\,(Z_i) = 0$, $\mathrm{Var}\,(Z_i) = 1$ such that $\sqrt{n}(\overline{Z}_n)$ does not converge in distribution to $N(0,1)$.

Exercise 4.14 Let $(a_1, \ldots, a_n)$ be a random permutation of the integers $1, \ldots, n$. If $a_j < a_i$ for some $i < j$, then the pair (i, j) is said to form an inversion. Let X_n be the total number of inversions:

$$X_n \;=\; \sum_{j=2}^n \sum_{i=1}^{j-1} I\{a_j < a_i\}.$$

For example, if $n = 3$ and we consider the permutation $(3, 1, 2)$, there are 2 inversions since $1 = a_2 < a_1 = 3$ and $2 = a_3 < a_1 = 3$. This problem asks you to find the asymptotic distribution of X_n.

(a) Define $Y_1 = 0$ and for $j > 1$, let

$$Y_j = \sum_{i=1}^{j-1} I\{a_j < a_i\}$$

be the number of a_i greater than a_j to the left of a_j. Then the Y_j are independent (you don't have to show this; you may wish to think about why, though). Find $\mathrm{E}(Y_j)$ and Var Y_j.

(b) Use $X_n = Y_1 + Y_2 + \cdots + Y_n$ to prove that

$$\frac{3}{2}\sqrt{n}\left(\frac{4X_n}{n^2} - 1\right) \xrightarrow{d} N(0, 1).$$

(c) For $n = 10$, evaluate the distribution of inversions as follows. First, simulate 1000 permutations on $\{1, 2, \ldots, 10\}$ and for each permutation, count the number of inversions. Plot a histogram of these 1000 numbers. Use the results of the simulation to estimate $P(X_{10} \leq 24)$. Second, estimate $P(X_{10} \leq 24)$ using a normal approximation. Can you find the exact integer c such that $10!P(X_{10} \leq 24) = c$?

Exercise 4.15 Suppose that X_1, X_2, X_3 is a sample of size 3 from a beta $(2, 1)$ distribution.

(a) Find $P(X_1 + X_2 + X_3 \leq 1)$ exactly.

(b) Find $P(X_1 + X_2 + X_3 \leq 1)$ using a normal approximation derived from the central limit theorem.

(c) Let $Z = I\{X_1 + X_2 + X_3 \leq 1\}$. Approximate $\mathrm{E}\ Z = P(X_1 + X_2 + X_3 \leq 1)$ by $\overline{Z} = \sum_{i=1}^{1000} Z_i/1000$, where $Z_i = I\{X_{i1} + X_{i2} + X_{i3} \leq 1\}$ and the X_{ij} are independent beta $(2, 1)$ random variables. In addition to $\overline{Z}$, report Var Z for your sample. (To think about: What is the theoretical value of Var Z?)

(d) Approximate $P(X_1 + X_2 + X_3 \leq \frac{3}{2})$ using the normal approximation and the simulation approach. (Don't compute the exact value, which is more difficult to than in part (a); do you see why?)

Exercise 4.16 Lindeberg and Lyapunov impose conditions on moments so that asymptotic normality occurs. However, it is possible to have asymptotic normality even

if there are no moments at all. Let X_n assume the values $+1$ and -1 with probability $(1 - 2^{-n})/2$ each and the value 2^k with probability 2^{-k} for $k > n$.

(a) Show that $\mathrm{E}\,(X_n^j) = \infty$ for all positive integers j and n.

(b) Show that $\sqrt{n}\left(\overline{X}_n\right) \xrightarrow{d} N(0,1)$.

Exercise 4.17 Assume that elements ("coupons") are drawn from a population of size n, randomly and with replacement, until the number of distinct elements that have been sampled is r_n, where $1 \leq r_n \leq n$. Let S_n be the drawing on which this first happens. Suppose that $r_n/n \to \rho$, where $0 < \rho < 1$.

(a) Suppose $k-1$ distinct coupons have thus far entered the sample. Let X_{nk} be the waiting time until the next distinct one appears, so that

$$S_n = \sum_{k=1}^{r_n} X_{nk}.$$

Find the expectation and variance of X_{nk}.

(b) Let $m_n = \mathrm{E}\,(S_n)$ and $\tau_n^2 = \mathrm{Var}\,(S_n)$. Show that

$$\frac{S_n - m_n}{\tau_n} \xrightarrow{d} N(0,1).$$

Tip: One approach is to apply Lyapunov's condition with $\delta = 2$. This involves demonstrating an asymptotic expression for τ_n^2 and a bound on $\mathrm{E}\,[X_{nk} - E(X_{nk})]^4$. There are several ways to go about this.

Exercise 4.18 Suppose that $X_1, X_2, \ldots$ are independent binomial$(2, p)$ random variables. Define $Y_i = I\{X_i = 0\}$.

(a) Find $\mathbf{a}$ such that the joint asymptotic distribution of

$$\sqrt{n}\left[\begin{pmatrix}\overline{X}_n \\ \overline{Y}_n\end{pmatrix} - \mathbf{a}\right]$$

is nontrivial, and find this joint asymptotic distribution.

(b) Using the Cramér-Wold Theorem, Theorem 4.12, find the asymptotic distribution of $\sqrt{n}(\overline{X}_n + \overline{Y}_n - 1 - p^2)$.

4.3 Stationary m-Dependent Sequences

Here we consider sequences that are identically distributed but not independent. In fact, we make a stronger assumption than identically distributed; namely, we assume that $X_1, X_2, \ldots$ is a stationary sequence. (Stationary is defined in Definition 2.24.) Denote E X_i by μ and let $\sigma^2 = \text{Var } X_i$.

We seek sufficient conditions for the asymptotic normality of $\sqrt{n}(\overline{X}_n - \mu)$. The variance of $\overline{X}_n$ for a stationary sequence is given by Equation (2.20). Letting $\gamma_k = \text{Cov}\,(X_1, X_{1+k})$, we conclude that

$$\text{Var } \left\{\sqrt{n}(\overline{X}_n - \mu)\right\} = \sigma^2 + \frac{2}{n}\sum_{k=1}^{n-1}(n-k)\gamma_k. \tag{4.21}$$

Suppose that

$$\frac{2}{n}\sum_{k=1}^{n-1}(n-k)\gamma_k \to \gamma \tag{4.22}$$

as $n \to \infty$. Then based on Equation (4.21), it seems reasonable to ask whether

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2 + \gamma).$$

The answer, in many cases, is yes. This section explores one such case.

Recall from Definition 2.26 that $X_1, X_2, \ldots$ is m-dependent for some $m \geq 0$ if the vector $(X_1, \ldots, X_i)$ is independent of $(X_{i+j}, X_{i+j+1}, \ldots)$ whenever $j > m$. Therefore, for an m-dependent sequence we have $\gamma_k = 0$ for all $k > m$, so limit (4.22) becomes

$$\frac{2}{n}\sum_{k=1}^{n-1}(n-k)\gamma_k \to 2\sum_{k=1}^{m}\gamma_k.$$

For a stationary m-dependent sequence, the following theorem asserts the asymptotic normality of $\overline{X}_n$ as long as the X_i are bounded:

Theorem 4.20 If for some $m \geq 0$, $X_1, X_2, \ldots$ is a stationary m-dependent sequence of bounded random variables with E $X_i = \mu$ and Var $X_i = \sigma^2$, then

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} N\left(0, \sigma^2 + 2\sum_{k=1}^{m}\text{Cov}\,[X_1, X_{1+k}]\right).$$

The assumption in Theorem 4.20 that the X_i are bounded is not necessary, as long as $\sigma^2 < \infty$. However, the proof of the theorem is quite tricky without the boundedness assumption, and the theorem is strong enough for our purposes as it stands. See, for instance, Ferguson (1996) for a complete proof. The theorem may be proved using the following strategy: For some integer k_n, define random variables $V_1, V_2, \ldots$ and $W_1, W_2, \ldots$ as follows:

$$\begin{array}{ll} V_1 = X_1 + \cdots + X_{k_n}, & W_1 = X_{k_n+1} + \cdots + X_{k_n+m}, \\ V_2 = X_{k_n+m+1} + \cdots + X_{2k_n+m}, & W_2 = X_{2k_n+m+1} + \cdots + X_{2k_n+2m}, \\ \qquad\qquad \vdots & \end{array} \tag{4.23}$$

In other words, each V_i is the sum of k_n of the X_i and each W_i is the sum of m of the X_i. Because the sequence of X_i is m-dependent, we conclude that the V_i are independent. For this reason, we may apply the Lindeberg-Feller theorem to the V_i. If k_n is defined carefully, then the contribution of the W_i may be shown to be negligible. This strategy is implemented in Exercise 4.19, where a proof of Theorem 4.20 is outlined.

Example 4.21 *Runs of successes:* Suppose $X_1, X_2, \ldots$ are independent Bernoulli(p) variables. Let T_n denote the number of runs of successes in $X_1, \ldots, X_n$, where a run of successes is defined as a sequence of consecutive X_i, all of which equal 1, that is both preceded and followed by zeros (unless the run begins with X_1 or ends with X_n). What is the asymptotic distribution of T_n?

We note that

$$\begin{aligned} T_n &= \sum_{i=1}^{n} I\{\text{run starts at } i\text{th position}\} \\ &= X_1 + \sum_{i=2}^{n} X_i(1 - X_{i-1}), \end{aligned}$$

since a run starts at the ith position for $i > 1$ if and only if $X_i = 1$ and $X_{i-1} = 0$.

Letting $Y_i = X_{i+1}(1 - X_i)$, we see immediately that $Y_1, Y_2, \ldots$ is a stationary 1-dependent sequence with $\mathrm{E}\, Y_i = p(1-p)$, so that by Theorem 4.20, $\sqrt{n}\{\overline{Y}_n - p(1-p)\} \xrightarrow{d} N(0, \tau^2)$, where

$$\begin{aligned} \tau^2 &= \mathrm{Var}\, Y_1 + 2\,\mathrm{Cov}\,(Y_1, Y_2) \\ &= \mathrm{E}\, Y_1^2 - (\mathrm{E}\, Y_1)^2 + 2\,\mathrm{E}\, Y_1 Y_2 - 2(\mathrm{E}\, Y_1)^2 \\ &= \mathrm{E}\, Y_1 - 3(\mathrm{E}\, Y_1)^2 \;=\; p(1-p) - 3p^2(1-p)^2. \end{aligned}$$

Since

$$\frac{T_n - np(1-p)}{\sqrt{n}} = \sqrt{n}\{\overline{Y}_n - p(1-p)\} + \frac{X_1 - Y_n}{\sqrt{n}},$$

we conclude that

$$\frac{T_n - np(1-p)}{\sqrt{n}} \xrightarrow{d} N(0, \tau^2).$$

Exercises for Section 4.3

Exercise 4.19 We wish to prove theorem 4.20. Suppose $X_1, X_2, \ldots$ is a stationary m-dependent sequence of bounded random variables such that $\text{Var } X_i = \sigma^2$. Without loss of generality, assume $\text{E } X_i = 0$. We wish to prove that $\sqrt{n}(\overline{X}_n) \xrightarrow{d} N(0, \tau^2)$, where

$$\tau^2 = \sigma^2 + 2\sum_{k=1}^{m} \text{Cov}\,(X_1, X_{1+k}).$$

For all n, define $k_n = \lfloor n^{1/4} \rfloor$ and $\ell_n = \lfloor n/(k_n + m) \rfloor$ and $t_n = \ell_n(k_n + m)$. Define $V_1, \ldots, V_{\ell_n}$ and $W_1, \ldots, W_{\ell_n}$ as in Equation (4.23). Then

$$\sqrt{n}(\overline{X}_n) = \frac{1}{\sqrt{n}} \sum_{i=1}^{\ell_n} V_i + \frac{1}{\sqrt{n}} \sum_{i=1}^{\ell_n} W_i + \frac{1}{\sqrt{n}} \sum_{i=t_n+1}^{n} X_i.$$

(a) Prove that

$$\frac{1}{\sqrt{n}} \sum_{i=t_n+1}^{n} X_i \xrightarrow{P} 0. \tag{4.24}$$

Hint: Bound the left hand side of expression (4.24) using Markov's inequality (1.35) with $r = 1$. What is the greatest possible number of summands?

(b) Prove that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{\ell_n} W_i \xrightarrow{P} 0.$$

Hint: For $k_n > m$, the W_i are independent and identically distributed with distributions that do not depend on n. Use the central limit theorem on $(1/\sqrt{\ell_n}) \sum_{i=1}^{\ell_n} W_i$.

(c) Prove that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{\ell_n} V_i \xrightarrow{d} N(0, \tau^2),$$

then use Slutsky's theorem to prove theorem 4.20.

Hint: Use the Lindeberg-Feller theorem.

Exercise 4.20 Suppose $X_0, X_1, \ldots$ is an independent sequence of Bernoulli trials with success probability p. Suppose X_i is the indicator of your team's success on rally i in a volleyball game. Your team scores a point each time it has a success that follows another success. Let $S_n = \sum_{i=1}^{n} X_{i-1} X_i$ denote the number of points your team scores by time n.

(a) Find the asymptotic distribution of S_n.

(b) Simulate a sequence $X_0, X_1, \ldots, X_{1000}$ as above and calculate S_{1000} for $p = .4$. Repeat this process 100 times, then graph the empirical distribution of S_{1000} obtained from simulation on the same axes as the theoretical asymptotic distribution from (a). Comment on your results.

Exercise 4.21 Let $X_0, X_1, \ldots$ be independent and identically distributed random variables from a continuous distribution $F(x)$. Define $Y_i = I\{X_i < X_{i-1} \text{ and } X_i < X_{i+1}\}$. Thus, Y_i is the indicator that X_i is a relative minimum. Let $S_n = \sum_{i=1}^{n} Y_i$.

(a) Find the asymptotic distribution of S_n.

(b) Let $n = 5000$. For a sample $X_0, \ldots, X_{5001}$ of size 5002 from the uniform $(0, 1)$ random number generator in R, compute an approximate two-sided p-value based on the observed value of S_n and the answer to part (a). The null hypothesis is that the sequence of "random" numbers generated is independent and identically distributed. (Naturally, the "random" numbers are not random at all, but are generated by a deterministic formula that is supposed to mimic randomness.)

4.4 Univariate extensions

This section discusses two different extensions of Theorem 4.19, the univariate Central Limit Theorem. As in the statement of that theorem, we assume here that $X_1, X_2, \ldots$ are independent and identically distributed with $\mathrm{E}(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$.

4.4.1 The Berry-Esseen theorem

Let us define $Y_i = (X_i - \mu)/\sigma$ and $S_n = \sqrt{n}\overline{Y}_n$. Furthermore, let $G_n(s)$ denote the cumulative distribution function of S_n, i.e., $G_n(s) = P(S_n \leq s)$. Then the Central Limit Theorem tells us that for any real number s, $G_n(s) \to \Phi(s)$ as $n \to \infty$, where as usual we let Φ denote the cumulative distribution function of the standard normal distribution. Since $\Phi(s)$ is bounded and continuous, we know that this convergence is uniform, which is to say that

$$\sup_{s \in \mathbb{R}} |G_n(s) - \Phi(s)| \to 0 \quad \text{as } n \to \infty.$$

Of course, this limit result says nothing about how close the left and right sides must be for any fixed (finite) n. However, theorems discovered independently by Andrew Berry and Carl-Gustav Esseen in the early 1940s do just this (each of these mathematicians actually proved a result that is slightly more general than the one that typically bears their names). The so-called Berry-Esseen Theorem is as follows:

Theorem 4.22 There exists a constant c such that if $Y_1, Y_2, \ldots$ are independent and identically distributed random variables with mean 0 and variance 1, then

$$\sup_{s \in \mathbb{R}} |G_n(s) - \Phi(s)| \leq \frac{c\,\mathrm{E}\,|Y_1^3|}{\sqrt{n}}$$

for all n, where $G_n(s)$ is the cumulative distribution function of $\sqrt{n}\overline{Y}_n$.

Notice that the inequality is vacuously true whenever $\mathrm{E}\,|Y_1^3|$ is infinite. In terms of the original sequence $X_1, X_2, \ldots$, the theorem is therefore sometimes stated by saying that when $\lambda = \mathrm{E}\,|X_1^3| < \infty$,

$$\sup_{s \in \mathbb{R}} |G_n(s) - \Phi(s)| \leq \frac{c\lambda}{\sigma^3\sqrt{n}}.$$

We will not give a proof of Theorem 4.22 here, though the interested reader might wish to consult papers by Ho and Chen (1978, *Annals of Probability*, pp. 231–249) and Stein (1972, *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2*, pp. 583–602). The former authors give a proof of Theorem 4.22 based on the Stein paper that gives the value $c = 6.5$. However, they do not prove that 6.5 is the smallest possible value of c, and in fact an interesting aspect of the Berry-Esseen Theorem is that the smallest possible value is not known. Currently, one author (Irina Shevtsova, arXiv:1111.6554v1) has shown that the inequality is valid for $c = 0.4748$. Furthermore, Esseen himself proved that c cannot be less than 0.4097. For the sake of simplicity, we may exploit the known results by taking $c = 1/2$ to state with certainty that

$$\sup_{s \in \mathbb{R}} |G_n(s) - \Phi(s)| \leq \frac{\mathrm{E}\,|Y_1^3|}{2\sqrt{n}}.$$

4.4.2 Edgeworth expansions

As in the previous section, Let us define $Y_i = (X_i - \mu)/\sigma$ and $S_n = \sqrt{n}\overline{Y}_n$. Furthermore, let

$$\gamma = \mathrm{E}\ Y_i^3 \quad \text{and} \quad \tau = \mathrm{E}\ Y_i^4$$

and suppose that $\tau < \infty$. The Central Limit Theorem says that for every real y,

$$P(S_n \leq y) = \Phi(y) + o(1) \quad \text{as } n \to \infty.$$

But we would like a better approximation to $P(S_n \leq y)$ than $\Phi(y)$, and we begin by constructing the characteristic function of S_n:

$$\psi_{S_n}(t) = \mathrm{E}\ \exp\left\{(\mathrm{i}t/\sqrt{n})\sum_{i=1}^{n} Y_i\right\} = \left[\psi_Y(t/\sqrt{n})\right]^n, \tag{4.25}$$

where $\psi_Y(t) = \mathrm{E}\ \exp\{\mathrm{i}tY\}$ is the characteristic function of Y_i.

Before proceeding with an examination of Equation (4.25), we first establish four preliminary facts:

1. **Sharpening a well-known limit:** We already know that $(1 + a/n)^n \to e^a$. But how good is this approximation? The binomial theorem shows (after quite a bit of algebra) that for a fixed nonnegative integer k,

$$\left(1 + \frac{a}{n}\right)^{n-k} = e^a\left(1 - \frac{a(a+2k)}{2n}\right) + o\left(\frac{1}{n}\right) \tag{4.26}$$

   as $n \to \infty$.

2. **Hermite polynomials:** If $\phi(x)$ denotes the standard normal density function, then we define the Hermite polynomials $H_k(x)$ by the equation

$$(-1)^k \frac{d^k}{dx^k}\phi(x) = H_k(x)\phi(x). \tag{4.27}$$

   Thus, by simply differentiating $\phi(x)$ repeatedly, we may verify that $H_1(x) = x$, $H_2(x) = x^2 - 1$, $H_3(x) = x^3 - 3x$, and so on. By differentiating Equation (4.27) itself, we obtain the recursive formula

$$\frac{d}{dx}\left[H_k(x)\phi(x)\right] = -H_{k+1}(x)\phi(x). \tag{4.28}$$

3. **An inversion formula for characteristic functions:** Suppose $Z \sim G(z)$ and $\psi_Z(t)$ denotes the characteristic function of Z. If $\int_{-\infty}^{\infty} |\psi_Z(t)|\, dt < \infty$, then $g(z) = G'(z)$ exists and

$$g(z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-\mathrm{i}tz} \psi_Z(t)\, dt. \tag{4.29}$$

We won't prove Equation (4.29) here, but a proof can be found in most books on theoretical probability.

4. **An identity involving $\phi(x)$:** For any positive integer k,

$$\begin{aligned} \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-\mathrm{i}tx} e^{-t^2/2} (it)^k\, dt &= \frac{(-1)^k}{2\pi} \frac{d^k}{dx^k} \int_{-\infty}^{\infty} e^{-\mathrm{i}tx} e^{-t^2/2}\, dt \\ &= (-1)^k \frac{d^k}{dx^k} \phi(x) \qquad (4.30) \\ &= H_k(x)\phi(x), \qquad (4.31) \end{aligned}$$

where (4.30) follows from (4.29) since $e^{-t^2/2}$ is the characteristic function for a standard normal distribution, and (4.31) follows from (4.27).

Returning to Equation (4.25), we next use a Taylor expansion of $\exp\{itY/\sqrt{n}\}$: As $n \to \infty$,

$$\begin{aligned} \psi_Y\left(\frac{t}{\sqrt{n}}\right) &= \mathrm{E}\left\{1 + \frac{\mathrm{i}tY}{\sqrt{n}} + \frac{(\mathrm{i}t)^2 Y^2}{2n} + \frac{(\mathrm{i}t)^3 Y^3}{6n\sqrt{n}} + \frac{(\mathrm{i}t)^4 Y^4}{24n^2}\right\} + o\left(\frac{1}{n^2}\right) \\ &= \left(1 - \frac{t^2}{2n}\right) + \frac{(\mathrm{i}t)^3 \gamma}{6n\sqrt{n}} + \frac{(\mathrm{i}t^4)\tau}{24n^2} + o\left(\frac{1}{n^2}\right). \end{aligned}$$

If we raise this tetranomial to the nth power, most terms are $o(1/n)$:

$$\begin{aligned} \left[\psi_Y\left(\frac{t}{\sqrt{n}}\right)\right]^n &= \left[\left(1 - \frac{t^2}{2n}\right)^n + \left(1 - \frac{t^2}{2n}\right)^{n-1} \left(\frac{(\mathrm{i}t)^3\gamma}{6\sqrt{n}} + \frac{(\mathrm{i}t)^4\tau}{24n}\right) \right. \\ &\quad \left. + \left(1 - \frac{t^2}{2n}\right)^{n-2} \frac{(n-1)(\mathrm{i}t)^6\gamma^2}{72n^2}\right] + o\left(\frac{1}{n}\right). \qquad (4.32) \end{aligned}$$

By Equations (4.26) and (4.32), we conclude that

$$\begin{aligned} \psi_{S_n}(t) &= e^{-t^2/2}\left[1 - \frac{t^4}{8n} + \frac{(\mathrm{i}t)^3\gamma}{6\sqrt{n}} + \frac{(\mathrm{i}t)^4\tau}{24n} + \frac{(\mathrm{i}t)^6\gamma^2}{72n}\right] + o\left(\frac{1}{n}\right) \\ &= e^{-t^2/2}\left[1 + \frac{(\mathrm{i}t)^3\gamma}{6\sqrt{n}} + \frac{(\mathrm{i}t)^4(\tau - 3)}{24n} + \frac{(\mathrm{i}t)^6\gamma^2}{72n}\right] + o\left(\frac{1}{n}\right). \qquad (4.33) \end{aligned}$$

If we apply these three approximations to equation (4.32), we obtain

$$\begin{aligned}\left[\psi_X\left(\frac{t}{\sqrt{n}}\right)\right]^n &= e^{-t^2/2}\left[1-\frac{t^4}{8n}+\frac{(it)^3\gamma}{6\sqrt{n}}+\frac{(it)^4\tau}{24n}+\frac{(it)^6\gamma^2}{72n}\right]+o\left(\frac{1}{n}\right)\\ &= e^{-t^2/2}\left[1+\frac{(it)^3\gamma}{6\sqrt{n}}+\frac{(it)^4(\tau-3)}{24n}+\frac{(it)^6\gamma^2}{72n}\right]+o\left(\frac{1}{n}\right).\end{aligned}$$

Putting (4.33) together with (4.29), we obtain the following density function as an approximation to the distribution of S_n:

$$\begin{aligned}g(y) &= \frac{1}{2\pi}\left(\int_{-\infty}^{\infty}e^{-\mathrm{i}ty}e^{-t^2/2}\,dt+\frac{\gamma}{6\sqrt{n}}\int_{-\infty}^{\infty}e^{-\mathrm{i}ty}e^{-t^2/2}(\mathrm{i}t)^3\,dt\right.\\ &\quad\left.+\frac{\tau-3}{24n}\int_{-\infty}^{\infty}e^{-\mathrm{i}tx}e^{-t^2/2}(\mathrm{i}t)^4\,dt+\frac{\gamma^2}{72n}\int_{-\infty}^{\infty}e^{-\mathrm{i}ty}e^{-t^2/2}(\mathrm{i}t)^6\,dt\right).\end{aligned}\tag{4.34}$$

Next, combine (4.34) with (4.31) to yield

$$g(y) = \phi(y)\left(1+\frac{\gamma H_3(y)}{6\sqrt{n}}+\frac{(\tau-3)H_4(y)}{24n}+\frac{\gamma^2H_6(y)}{72n}\right).\tag{4.35}$$

By (4.28), the antiderivative of $g(y)$ equals

$$\begin{aligned}G(y) &= \Phi(y)-\phi(y)\left(\frac{\gamma H_2(y)}{6\sqrt{n}}+\frac{(\tau-3)H_3(y)}{24n}+\frac{\gamma^2H_5(y)}{72n}\right)\\ &= \Phi(y)-\phi(y)\left(\frac{\gamma(y^2-1)}{6\sqrt{n}}+\frac{(\tau-3)(y^3-3y)}{24n}+\frac{\gamma^2(y^5-10y^3+15y)}{72n}\right).\end{aligned}$$

The expression above is called the second-order Edgeworth expansion. By carrying out the expansion in (4.33) to more terms, we may obtain higher-order Edgeworth expansions. On the other hand, the first-order Edgeworth expansion is

$$G(y) = \Phi(y)-\phi(y)\left(\frac{\gamma(y^2-1)}{6\sqrt{n}}\right).\tag{4.36}$$

(see Exercise 4.23). Thus, if the distribution of Y is symmetric, we obtain $\gamma = 0$ and therefore in this case, the usual (zero-order) central limit theorem approximation given by $\Phi(y)$ is already first-order accurate.

Incidentally, the second-order Edgeworth expansion explains why the standard definition of kurtosis of a distribution with mean 0 and variance 1 is the unusual-looking $\tau - 3$.

Exercises for Section 4.4

Exercise 4.22 Verify Equation (4.26).

Exercise 4.23 Verify that Equation (4.36) is the first-order Edgeworth approximation to the distribution function of S_n.

Chapter 5

The Delta Method and Applications

5.1 Local linear approximations

Suppose that a particular random sequence converges in distribution to a particular constant. The idea of using a first-order (linear) Taylor expansion of a known function, in the neighborhood of that constant limit, is a very useful technique known as the delta method. This chapter introduces the method, named for the Δ in $g(x + \Delta x) \approx g(x) + \Delta x g'(x)$, and discusses some of its applications.

5.1.1 Asymptotic distributions of transformed sequences

In the simplest form of the Central Limit Theorem, Theorem 4.19, we consider a sequence $X_1, X_2, \ldots$ of independent and identically distributed (univariate) random variables with finite variance σ^2. In this case, the Central Limit Theorem states that

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} \sigma Z, \tag{5.1}$$

where $\mu = \mathrm{E}\ X_1$ and Z is a standard normal random variable.

In this chapter, we wish to consider the asymptotic distribution of some function of $\overline{X}_n$. In the simplest case, the answer depends on results already known: Consider a linear function $g(t) = at + b$ for some known constants a and b. Since $\mathrm{E}\ \overline{X}_n = \mu$, clearly $\mathrm{E}\ g(\overline{X}_n) = a\mu + b = g(\mu)$ by the linearity of the expectation operator. Therefore, it is reasonable to ask whether $\sqrt{n}[g(\overline{X}_n) - g(\mu)]$ tends to some distribution as $n \to \infty$. But the linearity of $g(t)$ allows one to write

$$\sqrt{n}\left[g(\overline{X}_n) - g(\mu)\right] = a\sqrt{n}\left(\overline{X}_n - \mu\right).$$

We conclude by Theorem 2.27 that

$$\sqrt{n}\left[g(\overline{X}_n) - g(\mu)\right] \xrightarrow{d} a\sigma Z.$$

Of course, the distribution on the right hand side above is $N(0, a^2\sigma^2)$.

None of the preceding development is especially deep; one might even say that it is obvious that a linear transformation of the random variable $\overline{X}_n$ alters its asymptotic distribution by a constant multiple. Yet what if the function $g(t)$ is nonlinear? It is in this nonlinear case that a strong understanding of the argument above, as simple as it may be, pays real dividends. For if $\overline{X}_n$ is consistent for μ (say), then we know that, roughly speaking, $\overline{X}_n$ will be very close to μ for large n. Therefore, the only meaningful aspect of the behavior of $g(t)$ is its behavior in a small neighborhood of μ. *And in a small neighborhood of μ, $g(\mu)$ may be considered to be roughly a linear function if we use a first-order Taylor expansion.* In particular, we may approximate

$$g(t) \approx g(\mu) + g'(\mu)(t - \mu)$$

for t in a small neighborhood of μ. We see that $g'(\mu)$ is the multiple of t, and so the logic of the linear case above suggests

$$\sqrt{n}\left\{g(\overline{X}_n) - g(\mu)\right\} \xrightarrow{d} g'(\mu)\sigma Z. \tag{5.2}$$

Indeed, expression (5.2) is a special case of the powerful theorem known as the delta method, which we now state and prove:

Theorem 5.1 *Delta method:* If $g'(a)$ exists and $n^b(X_n - a) \xrightarrow{d} X$ for $b > 0$, then

$$n^b\left\{g(X_n) - g(a)\right\} \xrightarrow{d} g'(a)X.$$

Proof: By Slutsky's Theorem, $X_n - a \xrightarrow{P} 0$ because $X_n - a = n^{-b}n^b(X_n - a)$ and $n^{-b}(X) \xrightarrow{d} 0(X) = 0$. Therefore, we may apply Theorem 2.8, which is Taylor's theorem as it applies to random variables. Taking $d = 1$ in Equation (2.5) gives

$$n^b\left\{g(X_n) - g(a)\right\} \;=\; n^b(X_n - a)\left\{g'(a) + o_P(1)\right\}$$

as $n \to \infty$. Therefore, Slutsky's theorem together with the fact that $n^b(X_n - a) \xrightarrow{d} X$ proves Theorem 5.1. ∎

Expression (5.2) may be reexpressed as a corollary of Theorem 5.1:

Corollary 5.2 The often-used special case of Theorem 5.1 in which X is normally distributed states that if $g'(\mu)$ exists and $\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$, then

$$\sqrt{n}\left\{g(\overline{X}_n) - g(\mu)\right\} \xrightarrow{d} N\left\{0, \sigma^2 g'(\mu)^2\right\}.$$

Ultimately, we will extend Theorem 5.1 in two directions: Theorem 5.5 deals with the special case in which $g'(a) = 0$, and Theorem 5.6 is the multivariate version of the delta method. But we first apply the delta method to a couple of simple examples that illustrate a principle that we discussed in Section 4.1.4: When we speak of the "asymptotic distribution" of a sequence of random variables, we generally refer to a nontrivial (i.e., nonconstant) distribution. For example, in the case of an independent and identically distributed sequence $X_1, X_2, \ldots$ of random variables with finite variance, the phrase "asymptotic distribution of $\overline{X}_n$" generally refers to the fact that

$$\sqrt{n}\left(\overline{X}_n - \mathrm{E}\ X_1\right) \xrightarrow{d} N(0, \mathrm{Var}\ X_1),$$

not the fact that $\overline{X}_n \xrightarrow{P} \mathrm{E}\ X_1$.

Example 5.3 *Asymptotic distribution of $\overline{X}_n^2$* Suppose $X_1, X_2, \ldots$ are independent and identically distributed with mean μ and finite variance σ^2. Then by the central limit theorem,

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2).$$

Therefore, the delta method gives

$$\sqrt{n}(\overline{X}_n^2 - \mu^2) \xrightarrow{d} N(0, 4\mu^2\sigma^2). \tag{5.3}$$

However, this is not necessarily the end of the story. If $\mu = 0$, then the normal limit in (5.3) is degenerate—that is, expression (5.3) merely states that $\sqrt{n}(\overline{X}_n^2)$ converges in probability to the constant 0. This is not what we mean by the asymptotic distribution! Thus, we must treat the case $\mu = 0$ separately, noting in that case that $\sqrt{n}\overline{X}_n \xrightarrow{d} N(0, \sigma^2)$ by the central limit theorem, which implies that

$$n\overline{X}_n^2 \xrightarrow{d} \sigma^2\chi_1^2.$$

Example 5.4 *Estimating binomial variance:* Suppose $X_n \sim \mathrm{binomial}(n, p)$. Because X_n/n is the maximum likelihood estimator for p, the maximum likelihood estimator for $p(1-p)$ is $\delta_n = X_n(n - X_n)/n^2$. The central limit theorem tells us that $\sqrt{n}(X_n/n - p) \xrightarrow{d} N\{0, p(1-p)\}$, so the delta method gives

$$\sqrt{n}\left\{\delta_n - p(1-p)\right\} \xrightarrow{d} N\left\{0, p(1-p)(1-2p)^2\right\}.$$

Note that in the case $p = 1/2$, this does not give the asymptotic distribution of δ_n. Exercise 5.1 gives a hint about how to find the asymptotic distribution of δ_n in this case.

We have seen in the preceding examples that if $g'(a) = 0$, then the delta method gives something other than the asymptotic distribution we seek. However, by using more terms in the Taylor expansion, we obtain the following generalization of Theorem 5.1:

Theorem 5.5 If $g(t)$ has r derivatives at the point a and $g'(a) = g''(a) = \cdots = g^{(r-1)}(a) = 0$, then $n^b(X_n - a) \xrightarrow{d} X$ for $b > 0$ implies that

$$n^{rb}\left\{g(X_n) - g(a)\right\} \xrightarrow{d} \frac{1}{r!} g^{(r)}(a) X^r.$$

It is straightforward using the multivariate notion of differentiability discussed in Definition 1.36 to prove the following theorem:

Theorem 5.6 *Multivariate delta method:* If $\mathbf{g} : R^k \to R^\ell$ has a derivative $\nabla \mathbf{g}(\mathbf{a})$ at $\mathbf{a} \in R^k$ and

$$n^b\left(\mathbf{X}_n - \mathbf{a}\right) \xrightarrow{d} \mathbf{Y}$$

for some k-vector $\mathbf{Y}$ and some sequence $\mathbf{X}_1, \mathbf{X}_2, \ldots$ of k-vectors, where $b > 0$, then

$$n^b\left\{\mathbf{g}\left(\mathbf{X}_n\right) - \mathbf{g}\left(\mathbf{a}\right)\right\} \xrightarrow{d} \left[\nabla \mathbf{g}(\mathbf{a})\right]^\top \mathbf{Y}.$$

The proof of Theorem 5.6 involves a simple application of the multivariate Taylor expansion of Equation (1.31).

5.1.2 Variance stabilizing transformations

Often, if $\mathrm{E}\left(X_i\right) = \mu$ is the parameter of interest, the central limit theorem gives

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} N\{0, \sigma^2(\mu)\}.$$

In other words, the variance of the limiting distribution is a function of μ. This is a problem if we wish to do inference for μ, because ideally the limiting distribution should not depend on the unknown μ. The delta method gives a possible solution: Since

$$\sqrt{n}\left\{g(\overline{X}_n) - g(\mu)\right\} \xrightarrow{d} N\left\{0, \sigma^2(\mu) g'(\mu)^2\right\},$$

we may search for a transformation $g(x)$ such that $g'(\mu)\sigma(\mu)$ is a constant. Such a transformation is called a variance stabilizing transformation.

Example 5.7 Suppose that $X_1, X_2, \ldots$ are independent normal random variables with mean 0 and variance σ^2. Let us define $\tau^2 = \text{Var } X_i^2$, which for the normal distribution may be seen to be $2\sigma^4$. (To verify this, try showing that $\text{E } X_i^4 = 3\sigma^4$ by differentiating the normal characteristic function four times and evaluating at zero.) Thus, Example 4.11 shows that

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \sigma^2\right) \xrightarrow{d} N(0, 2\sigma^4).$$

To do inference for σ^2 when we believe that our data are truly independent and identically normally distributed, it would be helpful if the limiting distribution did not depend on the unknown σ^2. Therefore, it is sensible in light of Corollary 5.2 to search for a function $g(t)$ such that $2[g'(\sigma^2)]^2\sigma^4$ is not a function of σ^2. In other words, we want $g'(t)$ to be proportional to $\sqrt{t^{-2}} = |t|^{-1}$. Clearly $g(t) = \log t$ is such a function. Therefore, we call the logarithm function a variance-stabilizing function in this example, and Corollary 5.2 shows that

$$\sqrt{n}\left\{\log\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - \log\left(\sigma^2\right)\right\} \xrightarrow{d} N(0, 2).$$

Exercises for Section 5.1

Exercise 5.1 Let δ_n be defined as in Example 5.4. Find the asymptotic distribution of δ_n in the case $p = 1/2$. That is, find real-valued sequences a_n and b_n and a nontrivial random variable X such that $a_n(\delta_n - b_n) \xrightarrow{d} X$.

Hint: Let $Y_n = X_n - (n/2)$. Apply the central limit theorem to Y_n, then transform both sides of the resulting limit statement so that a statement involving δ_n results.

Exercise 5.2 Prove Theorem 5.5.

Exercise 5.3 Suppose $X_n \sim \text{binomial}(n, p)$, where $0 < p < 1$.

(a) Find the asymptotic distribution of $g(X_n/n) - g(p)$, where $g(x) = \min\{x, 1 - x\}$.

(b) Show that $h(x) = \sin^{-1}(\sqrt{x})$ is a variance-stabilizing transformation for X_n/n. This is called the *arcsine transformation* of a sample proportion.

Hint: $(d/du)\sin^{-1}(u) = 1/\sqrt{1 - u^2}$.

Exercise 5.4 Let $X_1, X_2, \ldots$ be independent from $N(\mu, \sigma^2)$ where $\mu \neq 0$. Let

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

Find the asymptotic distribution of the coefficient of variation $S_n/\overline{X}_n$.

Exercise 5.5 Let $X_n \sim$ binomial(n, p), where $p \in (0, 1)$ is unknown. Obtain confidence intervals for p in two different ways:

(a) Since $\sqrt{n}(X_n/n - p) \xrightarrow{d} N[0, p(1-p)]$, the variance of the limiting distribution depends only on p. Use the fact that $X_n/n \xrightarrow{P} p$ to find a consistent estimator of the variance and use it to derive a 95% confidence interval for p.

(b) Use the result of problem 5.3(b) to derive a 95% confidence interval for p.

(c) Evaluate the two confidence intervals in parts (a) and (b) numerically for all combinations of $n \in \{10, 100, 1000\}$ and $p \in \{.1, .3, .5\}$ as follows: For 1000 realizations of $X \sim \text{bin}(n, p)$, construct both 95% confidence intervals and keep track of how many times (out of 1000) that the confidence intervals contain p. Report the observed proportion of successes for each (n, p) combination. Does your study reveal any differences in the performance of these two competing methods?

5.2 Sample Moments

The weak law of large numbers tells us that If $X_1, X_2, \ldots$ are independent and identically distributed with $\text{E}\,|X_1|^k < \infty$, then

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k \xrightarrow{P} \text{E}\,X_1^k.$$

That is, sample moments are (weakly) consistent. For example, the sample variance, which we define as

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\overline{X}_n)^2, \tag{5.4}$$

is consistent for $\text{Var}\,X_i = \text{E}\,X_i^2 - (\text{E}\,X_i)^2$.

However, consistency is not the end of the story. The central limit theorem and the delta method will prove very useful in deriving asymptotic distribution results about sample moments. We consider two very important examples involving the sample variance of Equation (5.4).

Example 5.8 *Distribution of sample T statistic:* Suppose $X_1, X_2, \ldots$ are independent and identically distributed with $\mathrm{E}(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2 < \infty$. Define s_n^2 as in Equation (5.4), and let

$$T_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{s_n}.$$

Letting

$$A_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}$$

and $B_n = \sigma/s_n$, we obtain $T_n = A_n B_n$. Therefore, since $A_n \xrightarrow{d} N(0,1)$ by the central limit theorem and $B_n \xrightarrow{P} 1$ by the weak law of large numbers, Slutsky's theorem implies that $T_n \xrightarrow{d} N(0,1)$. In other words, T statistics are asymptotically normal under the null hypothesis.

Example 5.9 Let $X_1, X_2, \ldots$ be independent and identically distributed with mean μ, variance σ^2, third central moment $\mathrm{E}(X_i - \mu)^3 = \gamma$, and $\mathrm{Var}(X_i - \mu)^2 = \tau^2 < \infty$. Define S_n^2 as in Equation (4.6). We have shown earlier that $\sqrt{n}(S_n^2 - \sigma^2) \xrightarrow{d} N(0, \tau^2)$. The same fact may be proven using Theorem 4.9 as follows.

First, let $Y_i = X_i - \mu$ and $Z_i = Y_i^2$. We may use the multivariate central limit theorem to find the joint asymptotic distribution of $\overline{Y}_n$ and $\overline{Z}_n$, namely

$$\sqrt{n}\left\{ \begin{pmatrix} \overline{Y}_n \\ \overline{Z}_n \end{pmatrix} - \begin{pmatrix} 0 \\ \sigma^2 \end{pmatrix} \right\} \xrightarrow{d} N_2\left\{ \mathbf{0}, \begin{pmatrix} \sigma^2 & \gamma \\ \gamma & \tau^2 \end{pmatrix} \right\}.$$

Note that the above result uses the fact that $\mathrm{Cov}(Y_1, Z_1) = \gamma$.

We may write $S_n^2 = \overline{Z}_n - (\overline{Y}_n)^2$. Therefore, define the function $g(a,b) = b - a^2$ and observe that this gives $\nabla g(a,b) = (-2a, 1)^\top$. To use the delta method, we should evaluate

$$\nabla g(0,\sigma^2)^\top \begin{pmatrix} \sigma^2 & \gamma \\ \gamma & \tau^2 \end{pmatrix} \nabla g(0,\sigma^2) \;=\; \begin{pmatrix} 0 & 1 \end{pmatrix} \begin{pmatrix} \sigma^2 & \gamma \\ \gamma & \tau^2 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} \;=\; \tau^2$$

We conclude that

$$\sqrt{n}\left\{ g\begin{pmatrix} \overline{Y}_n \\ \overline{Z}_n \end{pmatrix} - g\begin{pmatrix} 0 \\ \sigma^2 \end{pmatrix} \right\} = \sqrt{n}(S_n^2 - \sigma^2) \xrightarrow{d} N(0, \tau^2)$$

as we found earlier (using a different argument) in Example 4.11.

Exercises for Section 5.2

Exercise 5.6 Suppose that $X_1, X_2, \ldots$ are independent and identically distributed Normal $(0, \sigma^2)$ random variables.

(a) Based on the result of Example 5.7, Give an approximate test at $\alpha = .05$ for $H_0 : \sigma^2 = \sigma_0^2$ vs. $H_a : \sigma^2 \neq \sigma_0^2$.

(b) For $n = 25$, estimate the true level of the test in part (a) for $\sigma_0^2 = 1$ by simulating 5000 samples of size $n = 25$ from the null distribution. Report the proportion of cases in which you reject the null hypothesis according to your test (ideally, this proportion will be about .05).

5.3 Sample Correlation

Suppose that $(X_1, Y_1), (X_2, Y_2), \ldots$ are independent and identically distributed vectors with $\mathrm{E}\ X_i^4 < \infty$ and $\mathrm{E}\ Y_i^4 < \infty$. For the sake of simplicity, we will assume without loss of generality that $\mathrm{E}\ X_i = \mathrm{E}\ Y_i = 0$ (alternatively, we could base all of the following derivations on the centered versions of the random variables).

We wish to find the asymptotic distribution of the sample correlation coefficient, r. If we let

$$\begin{pmatrix} m_x \\ m_y \\ m_{xx} \\ m_{yy} \\ m_{xy} \end{pmatrix} = \frac{1}{n} \begin{pmatrix} \sum_{i=1}^n X_i \\ \sum_{i=1}^n Y_i \\ \sum_{i=1}^n X_i^2 \\ \sum_{i=1}^n Y_i^2 \\ \sum_{i=1}^n X_i Y_i \end{pmatrix} \tag{5.5}$$

and

$$s_x^2 = m_{xx} - m_x^2, s_y^2 = m_{yy} - m_y^2, \text{ and } s_{xy} = m_{xy} - m_x m_y, \tag{5.6}$$

then $r = s_{xy}/(s_x s_y)$. According to the central limit theorem,

$$\sqrt{n} \left\{ \begin{pmatrix} m_x \\ m_y \\ m_{xx} \\ m_{yy} \\ m_{xy} \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ \sigma_x^2 \\ \sigma_y^2 \\ \sigma_{xy} \end{pmatrix} \right\} \xrightarrow{d} N_5 \left\{ \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \mathrm{Cov}\,(X_1, X_1) & \cdots & \mathrm{Cov}\,(X_1, X_1 Y_1) \\ \mathrm{Cov}\,(Y_1, X_1) & \cdots & \mathrm{Cov}\,(Y_1, X_1 Y_1) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}\,(X_1 Y_1, X_1) & \cdots & \mathrm{Cov}\,(X_1 Y_1, X_1 Y_1) \end{pmatrix} \right\} \tag{5.7}$$

Let Σ denote the covariance matrix in expression (5.7). Define a function $\mathbf{g} : \mathbb{R}^5 \to \mathbb{R}^3$ such that $\mathbf{g}$ applied to the vector of moments in Equation (5.5) yields the vector (s_x^2, s_y^2, s_{xy}) as

defined in expression (5.6). Then

$$\nabla \mathbf{g} \begin{pmatrix} a \\ b \\ c \\ d \\ e \end{pmatrix} = \begin{pmatrix} -2a & 0 & -b \\ 0 & -2b & -a \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Therefore, if we let

$$\begin{aligned} \Sigma^* &= \left[\nabla \mathbf{g} \begin{pmatrix} 0 \\ 0 \\ \sigma_x^2 \\ \sigma_y^2 \\ \sigma_{xy} \end{pmatrix} \right]^\top \Sigma \left[\nabla \mathbf{g} \begin{pmatrix} 0 \\ 0 \\ \sigma_x^2 \\ \sigma_y^2 \\ \sigma_{xy} \end{pmatrix} \right] \\ &= \begin{pmatrix} \mathrm{Cov}\,(X_1^2, X_1^2) & \mathrm{Cov}\,(X_1^2, Y_1^2) & \mathrm{Cov}\,(X_1^2, X_1 Y_1) \\ \mathrm{Cov}\,(Y_1^2, X_1^2) & \mathrm{Cov}\,(Y_1^2, Y_1^2) & \mathrm{Cov}\,(Y_1^2, X_1 Y_1) \\ \mathrm{Cov}\,(X_1 Y_1, X_1^2) & \mathrm{Cov}\,(X_1 Y_1, Y_1^2) & \mathrm{Cov}\,(X_1 Y_1, X_1 Y_1) \end{pmatrix}, \end{aligned}$$

then by the delta method,

$$\sqrt{n} \left\{ \begin{pmatrix} s_x^2 \\ s_y^2 \\ s_{xy} \end{pmatrix} - \begin{pmatrix} \sigma_x^2 \\ \sigma_y^2 \\ \sigma_{xy} \end{pmatrix} \right\} \xrightarrow{d} N_3(\mathbf{0}, \Sigma^*). \tag{5.8}$$

As an aside, note that expression (5.8) gives the same marginal asymptotic distribution for $\sqrt{n}(s_x^2 - \sigma_x^2)$ as was derived using a different approach in Example 4.11, since $\mathrm{Cov}\,(X_1^2, X_1^2)$ is the same as τ^2 in that example.

Next, define the function $h(a, b, c) = c/\sqrt{ab}$, so that we have $h(s_x^2, s_y^2, s_{xy}) = r$. Then

$$[\nabla h(a, b, c)]^\top = \frac{1}{2} \left(\frac{-c}{\sqrt{a^3 b}}, \frac{-c}{\sqrt{a b^3}}, \frac{2}{\sqrt{ab}} \right),$$

so that

$$[\nabla h(\sigma_x^2, \sigma_y^2, \sigma_{xy})]^\top = \left(\frac{-\sigma_{xy}}{2\sigma_x^3 \sigma_y}, \frac{-\sigma_{xy}}{2\sigma_x \sigma_y^3}, \frac{1}{\sigma_x \sigma_y} \right) = \left(\frac{-\rho}{2\sigma_x^2}, \frac{-\rho}{2\sigma_y^2}, \frac{1}{\sigma_x \sigma_y} \right). \tag{5.9}$$

Therefore, if A denotes the 1×3 matrix in Equation (5.9), using the delta method once again yields

$$\sqrt{n}(r - \rho) \xrightarrow{d} N(0, A \Sigma^* A^\top).$$

To recap, we have used the basic tools of the multivariate central limit theorem and the multivariate delta method to obtain a *univariate* result. This derivation of univariate facts via multivariate techniques is common practice in statistical large-sample theory.

Example 5.10 Consider the special case of bivariate normal (X_i, Y_i). In this case, we may derive

$$\Sigma^* = \begin{pmatrix} 2\sigma_x^4 & 2\rho^2\sigma_x^2\sigma_y^2 & 2\rho\sigma_x^3\sigma_y \\ 2\rho^2\sigma_x^2\sigma_y^2 & 2\sigma_y^4 & 2\rho\sigma_x\sigma_y^3 \\ 2\rho\sigma_x^3\sigma_y & 2\rho\sigma_x\sigma_y^3 & (1+\rho^2)\sigma_x^2\sigma_y^2 \end{pmatrix}. \tag{5.10}$$

In this case, $A\Sigma^* A^\top = (1-\rho^2)^2$, which implies that

$$\sqrt{n}(r-\rho) \xrightarrow{d} N\{0, (1-\rho^2)^2\}. \tag{5.11}$$

In the normal case, we may derive a variance-stabilizing transformation. According to Equation (5.11), we should find a function $f(x)$ satisfying $f'(x) = (1-x^2)^{-1}$. Since

$$\frac{1}{1-x^2} = \frac{1}{2(1-x)} + \frac{1}{2(1+x)},$$

we integrate to obtain

$$f(x) = \frac{1}{2}\log\frac{1+x}{1-x}.$$

This is called Fisher's transformation; we conclude that

$$\sqrt{n}\left(\frac{1}{2}\log\frac{1+r}{1-r} - \frac{1}{2}\log\frac{1+\rho}{1-\rho}\right) \xrightarrow{d} N(0,1).$$

Exercises for Section 5.3

Exercise 5.7 Verify expressions (5.10) and (5.11).

Exercise 5.8 Assume $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent and identically distributed from some bivariate normal distribution. Let ρ denote the population correlation coefficient and r the sample correlation coefficient.

(a) Describe a test of $H_0 : \rho = 0$ against $H_1 : \rho \neq 0$ based on the fact that

$$\sqrt{n}[f(r) - f(\rho)] \quad \xrightarrow{d} \quad N(0,1),$$

where $f(x)$ is Fisher's transformation $f(x) = (1/2)\log[(1+x)/(1-x)]$. Use $\alpha = .05$.

(b) Based on 5000 repetitions each, estimate the actual level for this test in the case when $\mathrm{E}(X_i) = \mathrm{E}(Y_i) = 0$, $\mathrm{Var}(X_i) = \mathrm{Var}(Y_i) = 1$, and $n \in \{3, 5, 10, 20\}$.

Exercise 5.9 Suppose that X and Y are jointly distributed such that X and Y are Bernoulli $(1/2)$ random variables with $P(XY = 1) = \theta$ for $\theta \in (0, 1/2)$. Let $(X_1, Y_1), (X_2, Y_2), \ldots$ be independent and identically distributed with (X_i, Y_i) distributed as (X, Y).

(a) Find the asymptotic distribution of $\sqrt{n}\left[(\overline{X}_n, \overline{Y}_n) - (1/2, 1/2)\right]$.

(b) If r_n is the sample correlation coefficient for a sample of size n, find the asymptotic distribution of $\sqrt{n}(r_n - \rho)$.

(c) Find a variance stabilizing transformation for r_n.

(d) Based on your answer to part (c), construct a 95% confidence interval for θ.

(e) For each combination of $n \in \{5, 20\}$ and $\theta \in \{.05, .25, .45\}$, estimate the true coverage probability of the confidence interval in part (d) by simulating 5000 samples and the corresponding confidence intervals. One problem you will face is that in some samples, the sample correlation coefficient is undefined because with positive probability each of the X_i or Y_i will be the same. In such cases, consider the confidence interval to be undefined and the true parameter therefore not contained therein.

Hint: To generate a sample of (X, Y), first simulate the X's from their marginal distribution, then simulate the Y's according to the conditional distribution of Y given X. To obtain this conditional distribution, find $P(Y = 1 \mid X = 1)$ and $P(Y = 1 \mid X = 0)$.

Chapter 6

Order Statistics and Quantiles

Consider an "ordering" function on n real numbers, a vector-valued function $\mathbf{f}_n$ that maps $\mathbb{R}^n \to \mathbb{R}^n$ so that if we let $\mathbf{y} = \mathbf{f}_n(\mathbf{x})$, then the values $y_1, \ldots, y_n$ are simply a permutation of the values $x_1, \ldots, x_n$ such that $y_1 \leq \cdots \leq y_n$. In this chapter, we consider the order statistics, which are the result of applying this ordering function to a simple random sample $X_1, \ldots, X_n$.

We introduce a specialized notation for these random variables, which we call the order statistics. Given a finite sample $X_1, \ldots, X_n$, define the values $X_{(1)}, \ldots, X_{(n)}$ to be a permutation of $X_1, \ldots, X_n$ such that $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$. We call $X_{(i)}$ the ith order statistic of the sample.

Even though the notation $X_{(i)}$ does not explicitly use the sample size n, the distribution of $X_{(i)}$ depends essentially on n. For this reason, some textbooks use slightly more complicated notation such as

$$X_{(1:n)}, X_{(2;n)}, \ldots, X_{(n;n)}$$

for the order statistics of a sample. We choose to use the simpler notation here, though it is important to remember that we will always understand the sample size to be n.

6.1 Extreme Order Statistics

The asymptotic distributions of order statistics at the extremes of a sample may be derived without any specialized knowledge other than the limit formula

$$\left(1 + \frac{c}{n}\right)^n \to e^c \quad \text{as } n \to \infty \tag{6.1}$$

and its generalization

$$\left(1+\frac{c_n}{b_n}\right)^{b_n} \to e^c \quad \text{if } c_n \to c \text{ and } b_n \to \infty \tag{6.2}$$

(see Example 1.20). Recall that by "asymptotic distribution of $X_{(1)}$," we mean sequences a_n and b_n, along with a nondegenerate random variable X, such that $a_n(X_{(1)} - b_n) \xrightarrow{d} X$. This section consists mostly of a series of illustrative examples.

Example 6.1 Suppose $X_1, \ldots, X_n$ are independent and identically distributed uniform(0,1) random variables. What is the asymptotic distribution of $X_{(n)}$?

Since $X_{(n)} \leq t$ if and only if $X_1 \leq t$, $X_2 \leq t$, ..., and $X_n \leq t$, by independence we have

$$P(X_{(n)} \leq t) = \begin{cases} 0 & \text{if } t \leq 0 \\ t^n & \text{if } 0 < t < 1 \\ 1 & \text{if } t \geq 1. \end{cases} \tag{6.3}$$

From Equation (6.3), it is apparent that $X_{(n)} \xrightarrow{P} 1$, though this limit statement does not fully reveal the asymptotic distribution of $X_{(n)}$. We desire sequences a_n and b_n such that $a_n(X_{(n)} - b_n)$ has a nondegenerate limiting distribution. Evidently, we should expect $b_n = 1$, a fact we shall rederive below.

Computing the distribution function of $a_n(X_{(n)} - b_n)$ directly, we find

$$F(u) = P\{a_n(X_{(n)} - b_n) \leq u\} = P\left\{X_{(n)} \leq \frac{u}{a_n} + b_n\right\}$$

as long as $a_n > 0$. Therefore, we see that

$$F(u) = \left(\frac{u}{a_n} + b_n\right)^n \quad \text{for} \quad 0 < \frac{u}{a_n} + b_n < 1. \tag{6.4}$$

We would like this expression to tend to a limit involving only u as $n \to \infty$. Keeping expression 6.2 in mind, we take $b_n = 1$ and $a_n = n$ so that $F(u) = (1 + u/n)^n$, which tends to e^u.

However, we are not quite finished, since we have not determined which values of u make the above limit valid. Equation 6.4 required that $0 < b_n + (u/a_n) < 1$, which in this case becomes $-1 < u/n < 0$. This means u may be any negative real number, since for any $u < 0$, $-1 < u/n < 0$ for all $n > |u|$. We conclude that if the random variable U has distribution function

$$F(u) = \begin{cases} \exp(u) & \text{if } u \leq 0 \\ 1 & \text{if } u > 0, \end{cases}$$

then $n(X_{(n)}-1) \xrightarrow{d} U$. Since $-U$ is simply a standard exponential random variable, we may also write

$$n(1 - X_{(n)}) \xrightarrow{d} \text{Exponential}(1).$$

Example 6.2 Suppose $X_1, X_2, \ldots$ are independent and identically distributed exponential random variables with mean 1. What is the asymptotic distribution of $X_{(n)}$?

As in Equation 6.3, if $u/a_n + b_n > 0$ then

$$P\left\{a_n(X_{(n)} - b_n) \leq u\right\} = P\left(X_{(n)} \leq \frac{u}{a_n} + b_n\right) = \left\{1 - \exp\left(-b_n - \frac{u}{a_n}\right)\right\}^n.$$

Taking $b_n = \log n$ and $a_n = 1$, the rightmost expression above simplifies to

$$\left\{1 - \frac{e^{-u}}{n}\right\}^n,$$

which has limit $\exp(-e^{-u})$. The condition $u/a_n + b_n > 0$ becomes $u + \log n > 0$, which is true for all $u \in \mathbb{R}$ as long as $n > \exp(-u)$. Therefore, we conclude that $X_{(n)} - \log n \xrightarrow{d} U$, where

$$P(U \leq u) \stackrel{\text{def}}{=} \exp\{-\exp(-u)\} \text{ for all } u. \tag{6.5}$$

The distribution of U in Equation 6.5 is known as the extreme value distribution or the Gumbel distribution.

In Examples 6.1 and 6.2, we derived the asymptotic distribution of a maximum from a simple random sample. We did this using only the definition of convergence in distribution without relying on any results other than expression 6.2. In a similar way, we may derive the joint asymptotic distribution of multiple order statistics, as in the following example.

Example 6.3 *Range of uniform sample:* Let $X_1, \ldots, X_n$ be a simple random sample from Uniform$(0, 1)$. Let $R_n = X_{(n)} - X_{(1)}$ denote the range of the sample. What is the asymptotic distribution of R_n?

To answer this question, we begin by finding the joint asymptotic distribution of $(X_{(n)}, X_{(1)})$, as follows. For sequences a_n and b_n, as yet unspecified, consider

$$\begin{aligned} P(a_n X_{(1)} > x \text{ and } b_n(1 - X_{(n)}) > y) &= P(X_{(1)} > x/a_n \text{ and } X_{(n)} < 1 - y/b_n) \\ &= P(x/a_n < X_{(1)} < \cdots < X_{(n)} < 1 - y/b_n), \end{aligned}$$

where we have assumed that a_n and b_n are positive. Since the probability above is simply the probability that the entire sample is to be found in the interval $(x/a_n, 1 - y/b_n)$, we conclude that as long as

$$0 < \frac{x}{a_n} < 1 - \frac{y}{b_n} < 1, \tag{6.6}$$

we have

$$P(a_n X_{(1)} > x \text{ and } b_n(1 - X_{(n)}) > y) \;=\; \left(1 - \frac{y}{b_n} - \frac{x}{a_n}\right)^n.$$

Expression (6.1) suggests that we set $a_n = b_n = n$, resulting in

$$P(nX_{(1)} > x \text{ and } n(1 - X_{(n)}) > y) \;=\; \left(1 - \frac{y}{n} - \frac{x}{n}\right)^n.$$

Expression 6.6 becomes

$$0 < \frac{x}{n} < 1 - \frac{y}{n} < 1,$$

which is satisfied for large enough n if and only if x and y are both positive. We conclude that for $x > 0$, $y > 0$,

$$P(nX_{(1)} > x \text{ and } n(1 - X_{(n)}) > y) \to e^{-x}e^{-y}.$$

Since this is the joint distribution of independent standard exponential random variables, say, Y_1 and Y_2, we conclude that

$$\begin{pmatrix} nX_{(1)} \\ n(1 - X_{(n)}) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}.$$

Therefore, applying the continuous function $f(a, b) = a + b$ to both sides gives

$$n(1 - X_{(n)} + X_{(1)}) = n(1 - R_n) \xrightarrow{d} Y_1 + Y_2 \sim \text{Gamma}(2, 1).$$

Let us consider a different example in which the asymptotic joint distribution does not involve independent random variables.

Example 6.4 As in Example 6.3, let $X_1, \ldots, X_n$ be independent and identically distributed from uniform$(0, 1)$. What is the joint asymptotic distribution of $X_{(n-1)}$ and $X_{(n)}$?

Proceeding as in Example 6.3, we obtain

$$P\left[\begin{pmatrix} n(1 - X_{(n-1)}) \\ n(1 - X_{(n)}) \end{pmatrix} > \begin{pmatrix} x \\ y \end{pmatrix}\right] = P\left(X_{(n-1)} < 1 - \frac{x}{n} \text{ and } X_{(n)} < 1 - \frac{y}{n}\right). \tag{6.7}$$

We consider two separate cases: If $0 < x < y$, then the right hand side of (6.7) is simply $P(X_{(n)} < 1 - y/n)$, which converges to e^{-y} as in Example 6.1. On the other hand, if $0 < y < x$, then

$$\begin{aligned} P\left(X_{(n-1)} < 1 - \frac{x}{n} \text{ and } X_{(n)} < 1 - \frac{y}{n}\right) &= P\left(X_{(n)} < 1 - \frac{x}{n}\right) \\ &\quad + P\left(X_{(n-1)} < 1 - \frac{x}{n} < X_{(n)} < 1 - \frac{y}{n}\right) \\ &= \left(1 - \frac{x}{n}\right)^n + n\left(1 - \frac{x}{n}\right)^{n-1}\left(\frac{x}{n} - \frac{y}{n}\right) \\ &\to e^{-x}(1 + x - y). \end{aligned}$$

The second equality above arises because $X_{(n-1)} < a < X_{(n)} < b$ if and only if exactly $n-1$ of the X_i are less than a and exactly one is between a and b. We now know the joint asymptotic distribution of $n(1 - X_{(n-1)})$ and $n(1 - X_{(n-1)})$; but can we describe this joint distribution in a simple way? Suppose that Y_1 and Y_2 are independent standard exponential variables. Consider the joint distribution of Y_1 and $Y_1 + Y_2$: If $0 < x < y$, then

$$P(Y_1 + Y_2 > x \text{ and } Y_1 > y) = P(Y_1 > y) = e^{-y}.$$

On the other hand, if $0 < y < x$, then

$$\begin{aligned} P(Y_1 + Y_2 > x \text{ and } Y_1 > y) &= P(Y_1 > \max\{y, x - Y_2\}) \;=\; \mathrm{E}\; e^{-\max\{y, x - Y_2\}} \\ &= e^{-y}P(y > x - Y_2) + \int_0^{x-y} e^{t-x}e^{-t}\,dt \\ &= e^{-x}(1 + x - y). \end{aligned}$$

Therefore, we conclude that

$$\begin{pmatrix} n(1 - X_{(n-1)}) \\ n(1 - X_{(n)}) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} Y_1 + Y_2 \\ Y_1 \end{pmatrix}.$$

Notice that marginally, we have shown that $n(1 - X_{(n-1)}) \xrightarrow{d} \text{Gamma}(2, 1)$.

Recall that if F is a continuous, invertible distribution function and U is a standard uniform random variable, then $F^{-1}(U) \sim F$. The proof is immediate, since $P\{F^{-1}(U) \le t\} = P\{U \le F(t)\} = F(t)$. We may use this fact in conjunction with the result of Example 6.4 as in the following example.

Example 6.5 Suppose $X_1, \ldots, X_n$ are independent standard exponential random variables. What is the joint asymptotic distribution of $(X_{(n-1)}, X_{(n)})$?

The distribution function of a standard exponential distribution is $F(t) = 1 - e^{-t}$, whose inverse is $F^{-1}(u) = -\log(1-u)$. Therefore,

$$\begin{pmatrix} -\log(1-U_{(n-1)}) \\ -\log(1-U_{(n)}) \end{pmatrix} \stackrel{d}{=} \begin{pmatrix} X_{(n-1)} \\ X_{(n)} \end{pmatrix},$$

where $\stackrel{d}{=}$ means "has the same distribution". Thus,

$$\begin{pmatrix} -\log\left[n(1-U_{(n-1)})\right] \\ -\log\left[n(1-U_{(n)})\right] \end{pmatrix} \stackrel{d}{=} \begin{pmatrix} X_{(n-1)} - \log n \\ X_{(n)} - \log n \end{pmatrix}.$$

We conclude by the result of Example 6.4 that

$$\begin{pmatrix} X_{(n-1)} - \log n \\ X_{(n)} - \log n \end{pmatrix} \stackrel{d}{\to} \begin{pmatrix} -\log(Y_1 + Y_2) \\ -\log Y_1 \end{pmatrix},$$

where Y_1 and Y_2 are independent standard exponential variables.

Exercises for Section 6.1

Exercise 6.1 For a given n, let $X_1, \ldots, X_n$ be independent and identically distributed with distribution function

$$P(X_i \le t) \;=\; \frac{t^3 + \theta^3}{2\theta^3} \qquad \text{for } t \in [-\theta, \theta].$$

Let $X_{(1)}$ denote the first order statistic from the sample of size n; that is, $X_{(1)}$ is the smallest of the X_i.

(a) Prove that $-X_{(1)}$ is consistent for θ.

(b) Prove that

$$n(\theta + X_{(1)}) \;\stackrel{d}{\to}\; Y,$$

where Y is a random variable with an exponential distribution. Find $\mathrm{E}\,(Y)$ in terms of θ.

(c) For a fixed α, define

$$\delta_{\alpha,n} \;=\; -\left(1 + \frac{\alpha}{n}\right) X_{(1)}.$$

Find, with proof, α^* such that

$$n\,(\theta - \delta_{\alpha^*,n}) \;\stackrel{d}{\to}\; Y - \mathrm{E}\,(Y),$$

where Y is the same random variable as in part (b).

(d) Compare the two consistent θ-estimators $\delta_{\alpha^*,n}$ and $-X_{(1)}$ empirically as follows. For $n \in \{10^2, 10^3, 10^4\}$, take $\theta = 1$ and simulate 1000 samples of size n from the distribution of X_i. From these 1000 samples, estimate the bias and mean squared error of each estimator. Which of the two appears better? Do your empirical results agree with the theoretical results in parts (c) and (d)?

Exercise 6.2 Let $X_1, X_2, \ldots$ be independent uniform $(0, \theta)$ random variables. Let $X_{(n)} = \max\{X_1, \ldots, X_n\}$ and consider the three estimators

$$\delta_n^0 = X_{(n)} \qquad \delta_n^1 = \frac{n}{n-1} X_{(n)} \qquad \delta_n^2 = \left(\frac{n}{n-1}\right)^2 X_{(n)}.$$

(a) Prove that each estimator is consistent for θ.

(b) Perform an empirical comparison of these three estimators for $n = 10^2, 10^3, 10^4$. Use $\theta = 1$ and simulate 1000 samples of size n from uniform $(0, 1)$. From these 1000 samples, estimate the bias and mean squared error of each estimator. Which one of the three appears to be best?

(c) Find the asymptotic distribution of $n(\theta - \delta_n^i)$ for $i = 0, 1, 2$. Based on your results, which of the three appears to be the best estimator and why? (For the latter question, don't attempt to make a rigorous mathematical argument; simply give an educated guess.)

Exercise 6.3 Find, with proof, the asymptotic distribution of $X_{(n)}$ if $X_1, \ldots, X_n$ are independent and identically distributed with each of the following distributions. (That is, find a_n, b_n, and a nondegenerate random variable X such that $a_n(X_{(n)} - b_n) \xrightarrow{d} X$.)

(a) Beta$(3, 1)$ with distribution function $F(x) = x^2$ for $x \in (0, 1)$.

(b) Standard logistic with distribution function $F(x) = e^x/(1 + e^x)$.

Exercise 6.4 Let $X_1, \ldots, X_n$ be independent uniform$(0, 1)$ random variables. Find the joint asymptotic distribution of $\left[nX_{(2)}, n(1 - X_{(n-1)})\right]$.

Hint: To find a probability such as $P(a < X_{(2)} < X_{(n)} < b)$, consider the trinomial distribution with parameters $[n; (a, b - a, 1 - b)]$ and note that the probability in question is the same as the probability that the numbers in the first and third categories are each ≤ 1.

Exercise 6.5 Let $X_1, \ldots, X_n$ be a simple random sample from the distribution function $F(x) = [1 - (1/x)]I\{x > 1\}$.

(a) Find the joint asymptotic distribution of $(X_{(n-1)}/n, X_{(n)}/n)$.

Hint: Proceed as in Example 6.5.

(b) Find the asymptotic distribution of $X_{(n-1)}/X_{(n)}$.

Exercise 6.6 If $X_1, \ldots, X_n$ are independent and identically distributed uniform$(0, 1)$ variables, prove that $X_{(1)}/X_{(2)} \xrightarrow{d}$ uniform$(0, 1)$.

Exercise 6.7 Let $X_1, \ldots, X_n$ be a simple random sample from a logistic distribution with distribution function $F(t) = e^{t/\theta}/(1 + e^{t/\theta})$ for all t.

(a) Find the asymptotic distribution of $X_{(n)} - X_{(n-1)}$.

Hint: Use the fact that $\log U_{(n)}$ and $\log U_{(n-1)}$ both converge in probability to zero.

(b) Based on part (a), construct an approximate 95% confidence interval for θ. Use the fact that the .025 and .975 quantiles of the standard exponential distribution are 0.0253 and 3.6889, respectively.

(c) Simulate 1000 samples of size $n = 40$ with $\theta = 2$. How many confidence intervals contain θ?

6.2 Sample Quantiles

To derive the distribution of sample quantiles, we begin by obtaining the exact distribution of the order statistics of a random sample from a uniform distribution. To facilitate this derivation, we begin with a quick review of changing variables. Suppose $\mathbf{X}$ has density $f_{\mathbf{X}}(\mathbf{x})$ and $\mathbf{Y} = \mathbf{g}(\mathbf{X})$, where $\mathbf{g} : \mathbb{R}^k \to \mathbb{R}^k$ is differentiable and has a well-defined inverse, which we denote by $\mathbf{h} : \mathbb{R}^k \to \mathbb{R}^k$. (In particular, we have $\mathbf{X} = \mathbf{h}[\mathbf{Y}]$.) The density for $\mathbf{Y}$ is

$$f_{\mathbf{Y}}(\mathbf{y}) = |\text{Det}[\nabla \mathbf{h}(\mathbf{y})]| \, f_{\mathbf{X}}[\mathbf{h}(\mathbf{y})], \tag{6.8}$$

where $|\text{Det}[\nabla \mathbf{h}(\mathbf{y})]|$ is the absolute value of the determinant of the $k \times k$ matrix $\nabla \mathbf{h}(\mathbf{y})$.

6.2.1 Uniform Order Statistics

We now show that the order statistics of a uniform distribution may be obtained using ratios of gamma random variables. Suppose $X_1, \ldots, X_{n+1}$ are independent standard exponential,

or Gamma$(1,1)$, random variables. For $j = 1, \ldots, n$, define

$$Y_j = \frac{\sum_{i=1}^{j} X_i}{\sum_{i=1}^{n+1} X_i}. \tag{6.9}$$

We will show that the joint distribution of $(Y_1, \ldots, Y_n)$ is the same as the joint distribution of the order statistics $(U_{(1)}, \ldots, U_{(n)})$ of a simple random sample from uniform$(0,1)$ by demonstrating that their joint density function is the same as that of the uniform order statistics, namely $n! I\{0 < u_{(1)} < \ldots < u_{(n)} < 1\}$.

We derive the joint density of $(Y_1, \ldots, Y_n)$ as follows. As an intermediate step, define $Z_j = \sum_{i=1}^{j} X_i$ for $j = 1, \ldots, n+1$. Then

$$X_i = \begin{cases} Z_i & \text{if } i = 1 \\ Z_i - Z_{i-1} & \text{if } i > 1, \end{cases}$$

which means that the gradient of the transformation from $\mathbf{Z}$ to $\mathbf{X}$ is upper triangular with ones on the diagonal, a matrix whose determinant is one. This implies that the density for $\mathbf{Z}$ is

$$f_{\mathbf{Z}}(\mathbf{z}) = \exp\{-z_{n+1}\} I\{0 < z_1 < z_2 < \cdots < z_{n+1}\}.$$

Next, if we define $Y_{n+1} = Z_{n+1}$, then we may express $\mathbf{Z}$ in terms of $\mathbf{Y}$ as

$$Z_i = \begin{cases} Y_{n+1} Y_i & \text{if } i < n+1 \\ Y_{n+1} & \text{if } i = n+1. \end{cases} \tag{6.10}$$

The gradient of the transformation in Equation (6.10) is lower triangular, with y_{n+1} along the diagonal except for a 1 in the lower right corner. The determinant of this matrix is y_{n+1}^n, so the density of $\mathbf{Y}$ is

$$f_{\mathbf{Y}}(\mathbf{y}) = y_{n+1}^n \exp\{-y_{n+1}\} I\{y_{n+1} > 0\} I\{0 < y_1 < \cdots < y_n < 1\}. \tag{6.11}$$

Equation (6.11) reveals several things: First, $(Y_1, \ldots, Y_n)$ is independent of Y_{n+1} and the marginal distribution of Y_{n+1} is Gamma$(n+1, 1)$. More important for our purposes, the marginal joint density of $(Y_1, \ldots, Y_n)$ is proportional to $I\{0 < y_1 < \cdots < y_n < 1\}$, which is exactly what we needed to prove. We conclude that the vector $\mathbf{Y}$ defined in Equation (6.9) has the same distribution as the vector of order statistics of a simple random sample from uniform$(0,1)$.

6.2.2 Uniform Sample Quantiles

Using the result of Section 6.2.1, we may derive the joint asymptotic distribution of a set of sample quantiles for a uniform simple random sample.

Suppose we are interested in the p_1 and p_2 quantiles, where $0 < p_1 < p_2 < 1$. The following argument may be generalized to obtain the joint asymptotic distribution of any finite number of quantiles. If $U_1, \ldots, U_n$ are independent uniform$(0,1)$ random variables, then the p_1 and p_2 sample quantiles may be taken to be the a_nth and b_nth order statistics, respectively, where $a_n \stackrel{\text{def}}{=} \lfloor .5 + np_1 \rfloor$ and $b_n \stackrel{\text{def}}{=} \lfloor .5 + np_2 \rfloor$ ($\lfloor .5 + x \rfloor$ is simply x rounded to the nearest integer).

Next, let

$$A_n \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{a_n} X_i, \quad B_n \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=a_n+1}^{b_n} X_i, \quad \text{and} \quad C_n \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=b_n+1}^{n+1} X_i.$$

We proved in Section 6.2.1 that $(U_{(a_n)}, U_{(b_n)})$ has the same distribution as

$$\mathbf{g}(A_n, B_n, C_n) \stackrel{\text{def}}{=} \left(\frac{A_n}{A_n + B_n + C_n}, \frac{A_n + B_n}{A_n + B_n + C_n} \right). \tag{6.12}$$

The asymptotic distribution of $\mathbf{g}(A_n, B_n, C_n)$ may be determined using the delta method if we can determine the joint asymptotic distribution of (A_n, B_n, C_n).

A bit of algebra reveals that

$$\sqrt{n}\,(A_n - p_1) \;=\; \sqrt{\frac{a_n}{n}} \sqrt{a_n} \left(\frac{nA_n}{a_n} - \frac{np_1}{a_n} \right) \;=\; \sqrt{\frac{a_n}{n}} \sqrt{a_n} \left(\frac{nA_n}{a_n} - 1 \right) + \frac{a_n - np_1}{\sqrt{n}}.$$

By the central limit theorem, $\sqrt{a_n}(nA_n/a_n - 1) \xrightarrow{d} N(0,1)$ because the X_i have mean 1 and variance 1. Furthermore, $a_n/n \to p_1$ and the rightmost term above goes to 0, so Slutsky's theorem gives

$$\sqrt{n}\,(A_n - p_1) \xrightarrow{d} N(0, p_1).$$

Similar arguments apply to B_n and to C_n. Because A_n and B_n and C_n are independent of one another, we may stack them as in Exercise 2.23 to obtain

$$\sqrt{n} \left\{ \begin{pmatrix} A_n \\ B_n \\ C_n \end{pmatrix} - \begin{pmatrix} p_1 \\ p_2 - p_1 \\ 1 - p_2 \end{pmatrix} \right\} \xrightarrow{d} N_3 \left\{ \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} p_1 & 0 & 0 \\ 0 & p_2 - p_1 & 0 \\ 0 & 0 & 1 - p_2 \end{pmatrix} \right\}$$

by Slutsky's theorem. For $\mathbf{g} : \mathbb{R}^3 \to \mathbb{R}^2$ defined in Equation (6.12), we obtain

$$\nabla \mathbf{g}(a,b,c) = \frac{1}{(a+b+c)^2} \begin{pmatrix} b+c & c \\ -a & c \\ -a & -a-b \end{pmatrix}.$$

Therefore,

$$[\nabla g(p_1, p_2 - p_1, 1 - p_2)]^\top = \begin{pmatrix} 1-p_1 & -p_1 & -p_1 \\ 1-p_2 & 1-p_2 & -p_2 \end{pmatrix},$$

so the delta method gives

$$\sqrt{n}\left\{ \begin{pmatrix} U_{(a_n)} \\ U_{(b_n)} \end{pmatrix} - \begin{pmatrix} p_1 \\ p_2 \end{pmatrix} \right\} \xrightarrow{d} N_2 \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} p_1(1-p_1) & p_1(1-p_2) \\ p_1(1-p_2) & p_2(1-p_2) \end{pmatrix} \right\}. \tag{6.13}$$

The method used above to derive the joint distribution (6.13) of two sample quantiles may be extended to any number of quantiles; doing so yields the following theorem:

Theorem 6.6 Suppose that for given constants $p_1, \ldots, p_k$ with $0 < p_1 < \cdots < p_k < 1$, we define sequences $\{a_{1n}\}, \ldots, \{a_{kn}\}$ such that for all $1 \le i \le k$,

$$\sqrt{n}\left(\frac{a_{in}}{n} - p_i\right) \to 0.$$

Then if $U_1, \ldots, U_n$ is a sample from Uniform(0,1),

$$\sqrt{n}\left\{ \begin{pmatrix} U_{(a_{1n})} \\ \vdots \\ U_{(a_{kn})} \end{pmatrix} - \begin{pmatrix} p_1 \\ \vdots \\ p_k \end{pmatrix} \right\} \xrightarrow{d} N_k \left\{ \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} p_1(1-p_1) & \cdots & p_1(1-p_k) \\ \vdots & & \vdots \\ p_1(1-p_k) & \cdots & p_k(1-p_k) \end{pmatrix} \right\}.$$

Note that for $i < j$, both the (i, j) and (j, i) entries in the covariance matrix above equal $p_i(1 - p_j)$ and $p_j(1 - p_i)$ never occurs in the matrix.

6.2.3 General sample quantiles

Let $F(x)$ be the distribution function for a random variable X. The quantile function $F^-(u)$ of Definition 3.18 is nondecreasing on $(0, 1)$ and it has the property that $F^-(U)$ has the same distribution as X for $U \sim$ uniform$(0, 1)$. This property follows from Lemma 3.19, since

$$P[F^-(U) \le x] = P[U \le F(x)] = F(x)$$

for all x. Since $F^-(u)$ is nondecreasing, it preserves ordering; thus, if $X_1, \ldots, X_n$ is a random sample from $F(x)$, then

$$(X_{(1)}, \ldots, X_{(n)}) \stackrel{d}{=} \left[F^-(U_{(1)}), \ldots, F^-(U_{(n)})\right].$$

(The symbol $\stackrel{d}{=}$ means "has the same distribution as".)

Now, suppose that at some point ξ, the derivative $F'(\xi)$ exists and is positive. Then $F(x)$ must be continuous and strictly increasing in a neighborhood of ξ. This implies that in this neighborhood, $F(x)$ has a well-defined inverse, which must be differentiable at the point $p \stackrel{\text{def}}{=} F(\xi)$. If $F^{-1}(u)$ denotes the inverse that exists in a neighborhood of p, then

$$\frac{dF^{-1}(p)}{dp} = \frac{1}{F'(\xi)}. \tag{6.14}$$

Equation (6.14) may be derived by differentiating the equation $F[F^{-1}(p)] = p$. Note also that whenever the inverse $F^{-1}(u)$ exists, it must coincide with the quantile function $F^{-}(u)$. Thus, the condition that $F'(\xi)$ exists and is positive is a sufficient condition to imply that $F^{-}(u)$ is differentiable at p. This differentiability is important: If we wish to transform the uniform order statistics $U_{(a_{in})}$ of Theorem 6.6 into order statistics $X_{(a_{in})}$ using the quantile function $F^{-}(u)$, the delta method requires the differentiability of $F^{-}(u)$ at each of the points $p_1, \ldots, p_k$.

The delta method, along with Equation (6.14), yields the following corollary of Theorem 6.6:

Theorem 6.7 Let $X_1, \ldots, X_n$ be a simple random sample from a distribution function $F(x)$ such that $F(x)$ is differentiable at each of the points $\xi_1 < \cdots < \xi_k$ and $F'(\xi_i) > 0$ for all i. Denote $F(\xi_i)$ by p_i. Then under the assumptions of Theorem 6.6,

$$\sqrt{n}\left\{\begin{pmatrix} X_{(a_{1n})} \\ \vdots \\ X_{(a_{kn})} \end{pmatrix} - \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_k \end{pmatrix}\right\} \xrightarrow{d} N_k\left\{\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{p_1(1-p_1)}{F'(\xi_1)^2} & \cdots & \frac{p_1(1-p_k)}{F'(\xi_1)F'(\xi_k)} \\ \vdots & & \vdots \\ \frac{p_1(1-p_k)}{F'(\xi_1)F'(\xi_k)} & \cdots & \frac{p_k(1-p_k)}{F'(\xi_k)^2} \end{pmatrix}\right\}.$$

Exercises for Section 6.2

Exercise 6.8 Let $X_1, \ldots, X_n$ be independent uniform$(0, 2\theta)$ random variables.

(a) Let $M = (X_{(1)} + X_{(n)})/2$. Find the asymptotic distribution of $n(M - \theta)$.

(b) Compare the asymptotic performance of the three estimators M, $\overline{X}_n$, and the sample median $\tilde{X}_n$ by considering their relative efficiencies.

(c) For $n \in \{101, 1001, 10001\}$, generate 500 samples of size n, taking $\theta = 1$. Keep track of M, $\overline{X}_n$, and $\tilde{X}_n$ for each sample. Construct a 3×3 table in which you report the sample variance of each estimator for each value of n. Do your simulation results agree with your theoretical results in part (b)?

Exercise 6.9 Let X_1 be Uniform$(0, 2\pi)$ and let X_2 be standard exponential, independent of X_1. Find the joint distribution of $(Y_1, Y_2) = (\sqrt{2X_2}\cos X_1, \sqrt{2X_2}\sin X_1)$.

Note: Since $-\log U$ has a standard exponential distribution if $U \sim$ uniform$(0, 1)$, this problem may be used to simulate normal random variables using simulated uniform random variables.

Exercise 6.10 Suppose $X_1, \ldots, X_n$ is a simple random sample from a distribution that is symmetric about θ, which is to say that $P(X_i \leq x) = F(x - \theta)$, where $F(x)$ is the distribution function for a distribution that is symmetric about zero. We wish to estimate θ by $(Q_p + Q_{1-p})/2$, where Q_p and Q_{1-p} are the p and $1 - p$ sample quantiles, respectively. Find the smallest possible asymptotic variance for the estimator and the p for which it is achieved for each of the following forms of $F(x)$:

(a) Standard Cauchy

(b) Standard normal

(c) Standard double exponential

Hint: For at least one of the three parts of this question, you will have to solve for a minimizer numerically.

Exercise 6.11 When we use a boxplot to assess the symmetry of a distribution, one of the main things we do is visually compare the lengths of $Q_3 - Q_2$ and $Q_2 - Q_1$, where Q_i denotes the ith sample quartile.

(a) Given a random sample of size n from $N(0, 1)$, find the asymptotic distribution of $(Q_3 - Q_2) - (Q_2 - Q_1)$.

(b) Repeat part (a) if the sample comes from a standard logistic distribution.

(c) Using 1000 simulations from each distribution, use graphs to assess the accuracy of each of the asymptotic approximations above for $n = 5$ and $n = 13$. (For a sample of size $4k + 1$, define Q_i to be the $(ik + 1)$th order statistic.) For each value of n and each distribution, plot the empirical distribution function against the theoretical limiting distribution function.

Exercise 6.12 Let $X_1, \ldots, X_n$ be a random sample from Uniform$(0, 2\theta)$. Find the asymptotic distributions of the median, the midquartile range, and $\frac{2}{3}Q_3$, where Q_3 denotes the third quartile and the midquartile range is the mean of the 1st and 3rd quartiles. Compare these three estimates of θ based on their asymptotic variances.

Chapter 7

Maximum Likelihood Estimation

7.1 Consistency

If X is a random variable (or vector) with density or mass function $f_\theta(x)$ that depends on a parameter θ, then the function $f_\theta(X)$ viewed as a function of θ is called the likelihood function of θ. We often denote this function by $L(\theta)$. Note that $L(\theta) = f_\theta(X)$ is implicitly a function of X, but we suppress this fact in the notation. Since repeated references to the "density or mass function" would be awkward, we will use the term "density" to refer to $f_\theta(x)$ throughout this chapter, even if the distribution function of X is not continuous. (Allowing noncontinuous distributions to have density functions may be made technically rigorous; however, this is a measure theoretic topic beyond the scope of this book.)

Let the set of possible values of θ be the set Ω. If $L(\theta)$ has a maximizer in Ω, say $\hat{\theta}$, then $\hat{\theta}$ is called a maximum likelihood estimator or MLE. Since the logarithm function is a strictly increasing function, any maximizer of $L(\theta)$ also maximizes $\ell(\theta) \stackrel{\text{def}}{=} \log L(\theta)$. It is often much easier to maximize $\ell(\theta)$, called the loglikelihood function, than $L(\theta)$.

Example 7.1 Suppose $\Omega = (0, \infty)$ and $X \sim \text{binomial}(n, e^{-\theta})$. Then

$$\ell(\theta) = \log \binom{n}{X} - X\theta + (n - X)\log(1 - e^{-\theta}),$$

so

$$\ell'(\theta) = -X + \frac{X - n}{1 - e^{\theta}}.$$

Thus, setting $\ell'(\theta) = 0$ yields $\theta = -\log(X/n)$. It isn't hard to verify that $\ell''(\theta) < 0$, so that $-\log(X/n)$ is in fact a maximizer of $\ell(\theta)$.

As the preceding example demonstrates, it is not always the case that a MLE exists—for if $X = 0$ or $X = n$, then $-\log(X/n)$ is not contained in Ω. This is just one of the technical details that we will consider. Ultimately, we will show that the maximum likelihood estimator is, in many cases, asymptotically normal. However, this is not always the case; in fact, it is not even necessarily true that the MLE is consistent, as shown in Problem 7.1.

We begin the discussion of the consistency of the MLE by defining the so-called Kullback-Leibler divergence.

Definition 7.2 If $f_{\theta_0}(x)$ and $f_{\theta_1}(x)$ are two densities, the Kullback-Leibler divergence from f_{θ_0} to f_{θ_1} equals

$$K(f_{\theta_0}, f_{\theta_1}) = \mathrm{E}_{\,\theta_0} \log \frac{f_{\theta_0}(X)}{f_{\theta_1}(X)}.$$

If $P_{\theta_0}(f_{\theta_0}(X) > 0 \text{ and } f_{\theta_1}(X) = 0) > 0$, then $K(f_{\theta_0}, f_{\theta_1})$ is defined to be ∞.

The Kullback-Leibler divergence is sometimes called the Kullback-Leibler information number or the relative entropy of f_{θ_1} with respect to f_{θ_0}. Although it is nonnegative, and takes the value zero if and only if $f_{\theta_1}(x) = f_{\theta_0}(x)$ except possibly for a set of x values having measure zero, The K-L divergence is not a true distance because $K(f_{\theta_0}, f_{\theta_1})$ is not necessarily the same as $K(f_{\theta_1}, f_{\theta_0})$.

We may show that the Kullback-Leibler information must be nonnegative by noting that

$$\mathrm{E}_{\,\theta_0} \frac{f_{\theta_1}(X)}{f_{\theta_0}(X)} = \mathrm{E}_{\,\theta_1}\, I\{f_{\theta_0}(X) > 0\} \leq 1.$$

Therefore, by Jensen's inequality (1.37) and the strict convexity of the function $-\log x$,

$$K(f_{\theta_0}, f_{\theta_1}) = \mathrm{E}_{\,\theta_0} - \log \frac{f_{\theta_1}(X)}{f_{\theta_0}(X)} \geq -\log \mathrm{E}_{\,\theta_0} \frac{f_{\theta_1}(X)}{f_{\theta_0}(X)} \geq 0, \tag{7.1}$$

with equality if and only if $P_{\theta_0}\{f_{\theta_0}(X) = f_{\theta_1}(X)\} = 1$. Inequality (7.1) is sometimes called the Shannon-Kolmogorov information inequality.

In the (admittedly somewhat bizarre) case in which the parameter space Ω contains only finitely many points, the Shannon-Kolmogorov information inequality may be used to prove the consistency of the maximum likelihood estimator. For the proof of the following theorem, note that if $X_1, \ldots, X_n$ are independent and identically distributed with density $f_{\theta_0}(x)$, then the loglikelihood is $\ell(\theta) = \sum_{i=1}^{n} \log f_{\theta_0}(x_i)$.

Theorem 7.3 Suppose Ω contains finitely many elements and that $X_1, \ldots, X_n$ are independent and identically distributed with density $f_{\theta_0}(x)$. Furthermore, suppose that the model parameter is identifiable, which is to say that different values of θ lead to different distributions. Then if $\hat{\theta}_n$ denotes the maximum likelihood estimator, $\hat{\theta}_n \xrightarrow{P} \theta_0$.

Proof: The Weak Law of Large Numbers (Theorem 2.19) implies that

$$\frac{1}{n}\sum_{i=1}^{n}\log\frac{f_\theta(X_i)}{f_{\theta_0}(X_i)} \xrightarrow{P} \mathrm{E}_{\theta_0}\log\frac{f_\theta(X_i)}{f_{\theta_0}(X_i)} = -K(f_{\theta_0}, f_\theta) \tag{7.2}$$

for all $\theta \in \Omega$. The value of $-K(f_{\theta_0}, f_\theta)$ is strictly negative for $\theta \neq \theta_0$ by the identifiability of θ. Therefore, since $\theta = \hat{\theta}_n$ is the maximizer of the left hand side of Equation (7.2),

$$P(\hat{\theta}_n \neq \theta_0) = P\left(\max_{\theta\neq\theta_0}\frac{1}{n}\sum_{i=1}^{n}\log\frac{f_\theta(X_i)}{f_{\theta_0}(X_i)} > 0\right) \leq \sum_{\theta\neq\theta_0} P\left(\frac{1}{n}\sum_{i=1}^{n}\log\frac{f_\theta(X_i)}{f_{\theta_0}(X_i)} > 0\right) \to 0.$$

This implies that $\hat{\theta}_n \xrightarrow{P} \theta_0$. ■

The result of Theorem 7.3 may be extended in several ways; however, it is unfortunately *not* true in general that a maximum likelihood estimator is consistent, as demonstrated by the example of Problem 7.1.

If we return to the simple Example 7.1, we found that the MLE was found by solving the equation

$$\ell'(\theta) = 0. \tag{7.3}$$

Equation (7.3) is called the likelihood equation, and naturally a root of the likelihood equation is a good candidate for a maximum likelihood estimator. However, there may be no root and there may be more than one. It turns out the probability that at least one root exists goes to 1 as $n \to \infty$. Consider Example 7.1, in which no MLE exists whenever $X = 0$ or $X = n$. In that case, both $P(X = 0) = (1 - e^{-\theta})^n$ and $P(X = n) = e^{-n\theta}$ go to zero as $n \to \infty$. In the case of multiple roots, one of these roots is typically consistent for θ_0, as stated in the following theorem.

Theorem 7.4 Suppose that $X_1, \ldots, X_n$ are independent and identically distributed with density $f_{\theta_0}(x)$ for θ_0 in an open interval $\Omega \subset R$, where the parameter is identifiable (i.e., different values of $\theta \in \Omega$ give different distributions). Furthermore, suppose that the loglikelihood function $\ell(\theta)$ is differentiable and that the support $\{x : f_\theta(x) > 0\}$ does not depend on θ. Then with probability approaching 1 as $n \to \infty$, there exists $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ such that $\ell'(\hat{\theta}_n) = 0$ and $\hat{\theta}_n \xrightarrow{P} \theta_0$.

Stated succinctly, Theorem 7.4 says that under certain regularity conditions, there is a consistent root of the likelihood equation. It is important to note that there is no guarantee that this consistent root is the MLE. However, if the likelihood equation only has a single root, we can be more precise:

Corollary 7.5 Under the conditions of Theorem 7.4, if for every n there is a unique root of the likelihood equation, and this root is a local maximum, then this root is the MLE and the MLE is consistent.

Proof: The only thing that needs to be proved is the assertion that the unique root is the MLE. Denote the unique root by $\hat{\theta}_n$ and suppose there is some other point θ such that $\ell(\theta) \geq \ell(\hat{\theta}_n)$. Then there must be a local minimum between $\hat{\theta}_n$ and θ, which contradicts the assertion that $\hat{\theta}_n$ is the unique root of the likelihood equation. ■

Exercises for Section 7.1

Exercise 7.1 In this problem, we explore an example in which the MLE is not consistent. Suppose that for $\theta \in (0, 1)$, X is a continuous random variable with density

$$f_\theta(x) = \theta g(x) + \frac{1-\theta}{\delta(\theta)} h\left(\frac{x-\theta}{\delta(\theta)}\right), \tag{7.4}$$

where $\delta(\theta) > 0$ for all θ, $g(x) = I\{-1 < x < 1\}/2$, and

$$h(x) = \frac{3(1-x^2)}{4} I\{-1 < x < 1\}.$$

(a) What condition on $\delta(\theta)$ ensures that $\{x : f_\theta(x) > 0\}$ does not depend on θ?

(b) With $\delta(\theta) = \exp\{-(1-\theta)^{-4}\}$, let $\theta = .2$. Take samples of sizes $n \in \{50, 250, 500\}$ from $f_\theta(x)$. In each case, graph the loglikelihood function and find the MLE. Also, try to identify the consistent root of the likelihood equation in each case.

Hints: To generate a sample from $f_\theta(x)$, note that $f_\theta(x)$ is a mixture density, which means you can start by generating a standard uniform random variable. If it's less than θ, generate a uniform variable on $(-1, 1)$. Otherwise, generate a variable with density $3(\delta^2 - x^2)/4\delta^3$ on $(-\delta, \delta)$ and then add θ. You should be able to do this by inverting the distribution function or by using appropriately scaled and translated beta$(2, 2)$ variables. If you use the inverse distribution function method, verify that

$$H^{-1}(u) = 2\cos\left\{\frac{4\pi}{3} + \frac{1}{3}\cos^{-1}(1-2u)\right\}.$$

Be very careful when graphing the loglikelihood and finding the MLE. In particular, make sure you evaluate the loglikelihood *analytically* at each of the sample

points in $(0, 1)$ and incorporate these analytical calculations into your code; if you fail to do this, you'll miss the point of the problem and you'll get the MLE incorrect. This is because the correct loglikelihood graph will have tall, extremely narrow spikes.

Exercise 7.2 In the situation of Exercise 7.1, prove that the MLE is inconsistent.

Exercise 7.3 Suppose that $X_1, \ldots, X_n$ are independent and identically distributed with density $f_\theta(x)$, where $\theta \in (0, \infty)$. For each of the following forms of $f_\theta(x)$, prove that the likelihood equation has a unique solution and that this solution maximizes the likelihood function.

(a) *Weibull:* For some constant $a > 0$,

$$f_\theta(x) = a\theta^a x^{a-1} \exp\{-(\theta x)^a\} I\{x > 0\}$$

(b) *Cauchy:*

$$f_\theta(x) = \frac{\theta}{\pi} \frac{1}{x^2 + \theta^2}$$

(c)

$$f_\theta(x) = \frac{3\theta^2\sqrt{3}}{2\pi(x^3 + \theta^3)} I\{x > 0\}$$

Exercise 7.4 Find the MLE and its asymptotic distribution given a random sample of size n from $f_\theta(x) = (1 - \theta)\theta^x$, $x = 0, 1, 2, \ldots$, $\theta \in (0, 1)$.

Hint: For the asymptotic distribution, use the central limit theorem.

7.2 Asymptotic normality of the MLE

As seen in the preceding section, the MLE is not necessarily even consistent, let alone asymptotically normal, so the title of this section is slightly misleading—however, "Asymptotic normality of the consistent root of the likelihood equation" is a bit too long! It will be necessary to review a few facts regarding Fisher information before we proceed.

Definition 7.6 *Fisher information:* For a density (or mass) function $f_\theta(x)$, the Fisher information function is given by

$$I(\theta) = \mathrm{E}_\theta \left\{ \frac{d}{d\theta} \log f_\theta(X) \right\}^2. \tag{7.5}$$

If $\eta = g(\theta)$ for some invertible and differentiable function $g(\cdot)$, then since

$$\frac{d}{d\eta} = \frac{d\theta}{d\eta}\frac{d}{d\theta} = \frac{1}{g'(\theta)}\frac{d}{d\theta}$$

by the chain rule, we conclude that

$$I(\eta) = \frac{I(\theta)}{\{g'(\theta)\}^2}. \tag{7.6}$$

Loosely speaking, $I(\theta)$ is the amount of information about θ contained in a single observation from the density $f_\theta(x)$. However, this interpretation doesn't always make sense—for example, it is possible to have $I(\theta) = 0$ for a very informative observation. See Exercise 7.6.

Although we do not dwell on this fact in this course because it has measure-theoretic underpinnings, expectation may be viewed as integration (even when, say, the distribution is discrete and the "density" is actually a mass function). Suppose that $f_\theta(x)$ is twice differentiable with respect to θ and that the operations of differentiation and integration may be interchanged in the following sense:

$$\frac{d}{d\theta}\int f_\theta(x)\,dx = \int \frac{d}{d\theta} f_\theta(x)\,dx \tag{7.7}$$

and

$$\frac{d^2}{d\theta^2}\int f_\theta(x)\,dx = \int \frac{d^2}{d\theta^2} f_\theta(x)\,dx. \tag{7.8}$$

(It is awkward to express the above ideas using our usual E_θ operator!) Since $\int f_\theta(x)\,dx = 1$, the left-hand sides of Equations (7.7) and (7.8) are both zero and this fact leads to two additional expressions for $I(\theta)$. From Equation (7.7) follows

$$I(\theta) = \mathrm{Var}_\theta\left\{\frac{d}{d\theta}\log f_\theta(X)\right\}, \tag{7.9}$$

and Equation (7.8) implies

$$I(\theta) = -\,\mathrm{E}_\theta\left\{\frac{d^2}{d\theta^2}\log f_\theta(X)\right\}; \tag{7.10}$$

see Exercise 7.5. In many cases, Equation (7.10) is the easiest form of the information to work with.

Equations (7.9) and (7.10) make clear a helpful property of the information, namely that for independent random variables, the information about θ contained in the joint sample is

simply the sum of the individual information components. In particular, if we have a simple random sample of size n from $f_\theta(x)$, then the information about θ equals $nI(\theta)$.

The reason that we need the Fisher information is that we will show that under certain assumptions (often called "regularity conditions"),

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\left\{0, \frac{1}{I(\theta_0)}\right\}, \tag{7.11}$$

where $\hat{\theta}_n$ is the consistent root of the likelihood equation guaranteed to exist by Theorem 7.4.

Example 7.7 Suppose that $X_1, \ldots, X_n$ are independent Poisson(θ_0) random variables. Then the likelihood equation has a unique root, namely $\hat{\theta}_n = \overline{X}_n$, and we know that by the central limit theorem $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \theta_0)$. Furthermore, the Fisher information for a single observation in this case is

$$-\operatorname{E}_\theta\left\{\frac{d^2}{d\theta^2}\log f_\theta(X)\right\} = \operatorname{E}_\theta \frac{X}{\theta^2} = \frac{1}{\theta}.$$

Thus, in this example, Equation (7.11) holds.

Rather than stating all of the regularity conditions necessary to prove Equation (7.11), we work backwards, figuring out the conditions as we go through the steps of the proof. The first step is to expand $\ell'(\hat{\theta}_n)$ in a Taylor series around θ_0. Let us introduce the notation $\ell_i(\theta)$ to denote the contribution to the loglikelihood from the ith observation; that is, $\ell_i(\theta) = \log f_\theta(X_i)$. Thus, we obtain

$$\ell(\theta) = \sum_{i=1}^{n} \ell_i(\theta).$$

For the Taylor expansion of $\ell'(\hat{\theta}_n)$, let $e_i(\hat{\theta}_n, \theta_0)$ denote the remainder for a first-order expansion of $\ell_i'(\hat{\theta}_n)$. That is, we define $e_i(\hat{\theta}_n, \theta_0)$ so that

$$\ell_i'(\hat{\theta}_n) = \ell_i'(\theta_0) + (\hat{\theta}_n - \theta_0)\ell_i''(\theta_0) + e_i(\hat{\theta}_n, \theta_0).$$

Summing over i, we obtain

$$\ell'(\hat{\theta}_n) = \ell'(\theta_0) + (\hat{\theta}_n - \theta_0)\left[\ell''(\theta_0) + E_n\right], \tag{7.12}$$

where $E_n = \sum_{i=1}^{n} e_i(\theta_n, \theta_0)/(\hat{\theta}_n - \theta_0)$, or, in the event $\hat{\theta}_n = \theta_0$, $E_n = 0$. (Remember, $\hat{\theta}_n$ and E_n are random variables.)

Rewriting Equation (7.12) gives

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{\sqrt{n}\{\ell'(\hat{\theta}_n) - \ell'(\theta_0)\}}{\ell''(\theta_0) + E_n} = \frac{\frac{1}{\sqrt{n}}\{\ell'(\theta_0) - \ell'(\hat{\theta}_n)\}}{-\frac{1}{n}\ell''(\theta_0) - \frac{1}{n}E_n}. \tag{7.13}$$

Let's consider the pieces of Equation (7.13) individually. If Equation (7.7) holds and $I(\theta_0) < \infty$, then

$$\frac{1}{\sqrt{n}}\ell'(\theta_0) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{d}{d\theta}\log f_{\theta_0}(X_i)\right) \xrightarrow{d} N\{0, I(\theta_0)\}$$

by the central limit theorem and Equation (7.9). If Equation (7.8) holds, then

$$-\frac{1}{n}\ell''(\theta_0) = -\frac{1}{n}\sum_{i=1}^{n}\frac{d^2}{d\theta^2}\log f_{\theta_0}(X_i) \xrightarrow{P} I(\theta_0)$$

by the weak law of large numbers and Equation (7.10). And we relied on the conditions of Theorem 7.4 to guarantee the existence of $\hat{\theta}_n$ such that $\ell'(\hat{\theta}_n) = 0$ with probability approaching one and $\hat{\theta}_n \to \theta_0$ (do you see where we used the latter fact?).

Finally, we need a condition that ensures that $\frac{1}{n}E_n \xrightarrow{P} 0$. One way this is often done is as follows: If we assume that the third derivative $\ell_i'''(\theta)$ exists and is uniformly bounded in a neighborhood of θ_0, say by the constant K_0, we may write the Taylor theorem remainder $e_i(\hat{\theta}_n, \theta_0)$ in the form of equation (1.7) to obtain

$$\frac{1}{n}E_n = \frac{1}{n}\sum_{i=1}^{n}\frac{e_i(\hat{\theta}_n, \theta_0)}{\hat{\theta}_n - \theta_0} = \frac{\hat{\theta}_n - \theta_0}{2n}\sum_{i=1}^{n}\ell_i'''(\theta_{in}^*),$$

where each θ_{in}^* is between $\hat{\theta}_n$ and θ_0. Therefore, since $\hat{\theta}_n \xrightarrow{P} \theta_0$, we know that with probability approaching 1 as $n \to \infty$,

$$\left|\frac{1}{n}E_n\right| \leq \frac{\hat{\theta}_n - \theta_0}{2n}\sum_{i=1}^{n}K_0 = \frac{K_0}{2}(\hat{\theta}_n - \theta_0),$$

which means that $\frac{1}{n}E_n \xrightarrow{P} 0$.

In conclusion, if all of our assumptions hold, then the numerator of (7.13) converges in distribution to $N\{0, I(\theta_0)\}$ by Slutsky's theorem. Furthermore, the denominator in (7.13) converges to $I(\theta_0)$, so a second use of Slutsky's theorem gives the following theorem.

Theorem 7.8 Suppose that the conditions of Theorem 7.4 are satisfied, and let $\hat{\theta}_n$ denote a consistent root of the likelihood equation. Assume also that Equations

(7.7) and (7.8) hold and that $0 < I(\theta_0) < \infty$. Finally, assume that $\ell_i(\theta)$ has three derivatives in some neighborhood of θ_0 and that $\ell_i'''(\theta)$ is uniformly bounded in this neighborhood. Then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\left\{0, \frac{1}{I(\theta_0)}\right\}.$$

Sometimes, it is not possible to find an exact zero of $\ell'(\theta)$. One way to get a numerical approximation to a zero of $\ell'(\theta)$ is to use Newton's method, in which we start at a point θ_0 and then set

$$\theta_1 = \theta_0 - \frac{\ell'(\theta_0)}{\ell''(\theta_0)}. \tag{7.14}$$

Ordinarily, after finding θ_1 we would set θ_0 equal to θ_1 and apply Equation (7.14) iteratively.

However, we may show that by using a *single step* of Newton's method, starting from a $\sqrt{n}$-consistent estimator of θ_0, we may obtain an estimator with the same asymptotic distribution as $\hat{\theta}_n$. A $\sqrt{n}$-consistent estimator is an estimator of θ_0, say $\tilde{\theta}_n$, with the property that $\sqrt{n}(\tilde{\theta}_n - \theta_0)$ is bounded in probability. For the full definition of bounded in probability, refer to Exercise 2.2, but a sufficient condition is that $\sqrt{n}(\tilde{\theta}_n - \theta_0)$ converges in distribution to *any* random variable.

The proof of the following theorem is left as an exercise:

Theorem 7.9 Suppose that $\tilde{\theta}_n$ is any $\sqrt{n}$-consistent estimator of θ_0. Then under the conditions of Theorem 7.8, if we set

$$\delta_n = \tilde{\theta}_n - \frac{\ell'(\tilde{\theta}_n)}{\ell''(\tilde{\theta}_n)}, \tag{7.15}$$

then

$$\sqrt{n}(\delta_n - \theta_0) \xrightarrow{d} N\left(0, \frac{1}{I(\theta_0)}\right).$$

Exercises for Section 7.2

Exercise 7.5 Show how assumptions (7.7) and (7.8) establish Equations (7.9) and (7.10), respectively.

Exercise 7.6 Suppose that X is a normal random variable with mean θ^3 and known variance σ^2. Calculate $I(\theta)$, then argue that the Fisher information can be zero in a case in which there is information about θ in the observed value of X.

Exercise 7.7 **(a)** Show that under the conditions of Theorem 7.8, then if $\hat{\theta}_n$ is a consistent root of the likelihood equation, $P_{\theta_0}(\hat{\theta}_n \text{ is a local maximum}) \to 1$.

(b) Using the result of part (a), show that for any two sequences $\hat{\theta}_{1n}$ and $\hat{\theta}_{2n}$ of consistent roots of the likelihood equation, $P_{\theta_0}(\hat{\theta}_{1n} = \hat{\theta}_{2n}) \to 1$.

Exercise 7.8 Prove Theorem 7.9.

Hint: Start with $\sqrt{n}(\delta_n - \theta_0) = \sqrt{n}(\delta_n - \tilde{\theta}_n) + \sqrt{n}(\tilde{\theta}_n - \theta_0)$, then expand $\ell'(\tilde{\theta}_n)$ in a Taylor series about θ_0 and substitute the result into Equation (7.15). After simplifying, use the result of Exercise 2.2 along with arguments similar to those leading up to Theorem 7.8.

Exercise 7.9 Suppose that the following is a random sample from a logistic density with distribution function $F_\theta(x) = (1+\exp\{\theta - x\})^{-1}$ (I'll cheat and tell you that I used $\theta = 2$.)

```
1.0944  6.4723  3.1180  3.8318  4.1262
1.2853  1.0439  1.7472  4.9483  1.7001
1.0422  0.1690  3.6111  0.9970  2.9438
```

(a) Evaluate the unique root of the likelihood equation numerically. Then, taking the sample median as our known $\sqrt{n}$-consistent estimator $\tilde{\theta}_n$ of θ, evaluate the estimator δ_n in Equation (7.15) numerically.

(b) Find the asymptotic distributions of $\sqrt{n}(\tilde{\theta}_n - 2)$ and $\sqrt{n}(\delta_n - 2)$. Then, simulate 200 samples of size $n = 15$ from the logistic distribution with $\theta = 2$. Find the sample variances of the resulting sample medians and δ_n-estimators. How well does the asymptotic theory match reality?

7.3 Asymptotic Efficiency and Superefficiency

In Theorem 7.8, we showed that a consistent root $\hat{\theta}_n$ of the likelihood equation satisfies

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\left(0, \frac{1}{I(\theta_0)}\right).$$

In Theorem 7.9, we stated that if $\tilde{\theta}_n$ is a $\sqrt{n}$-consistent estimator of θ_0 and $\delta_n = \tilde{\theta}_n - \ell'(\tilde{\theta}_n)/\ell''(\tilde{\theta}_n)$, then

$$\sqrt{n}(\delta_n - \theta_0) \xrightarrow{d} N\left(0, \frac{1}{I(\theta_0)}\right). \tag{7.16}$$

Note the similarity in the last two asymptotic limit statements. There seems to be something special about the limiting variance $1/I(\theta_0)$, and in fact this is true.

Much like the Cramér-Rao lower bound states that (under some regularity conditions) an unbiased estimator of θ_0 cannot have a variance smaller than $1/I(\theta_0)$, the following result is true:

Theorem 7.10 Suppose that the conditions of theorem 7.8 are satisfied and that δ_n is an estimator satisfying

$$\sqrt{n}(\delta_n - \theta_0) \xrightarrow{d} N\{0, v(\theta_0)\}$$

for all θ_0, where $v(\theta)$ is continuous. Then $v(\theta) \geq 1/I(\theta)$ for all θ.

In other words, $1/I(\theta)$ is, in the sense of Theorem 7.10, the smallest possible asymptotic variance for an estimator. For this reason, we refer to any estimator δ_n satisfying (7.16) for all θ_0 an **efficient** estimator.

One condition in Theorem 7.10 that may be a bit puzzling is the condition that $v(\theta)$ be continuous. If this condition is dropped, then a well-known counterexample, due to Hodges, exists:

Example 7.11 Suppose that δ_n is an efficient estimator of θ_0. Then if we define

$$\delta_n^* = \begin{cases} 0 & \text{if } n(\delta_n)^4 < 1; \\ \delta_n & \text{otherwise,} \end{cases}$$

it is possible to show (see Exercise 7.10) that δ_n^* is superefficient in the sense that

$$\sqrt{n}(\delta_n^* - \theta_0) \xrightarrow{d} N\left(0, \frac{1}{I(\theta_0)}\right)$$

for all $\theta_0 \neq 0$ but $\sqrt{n}(\delta_n^* - \theta_0) \xrightarrow{d} 0$ if $\theta_0 = 0$. In other words, when the true value of the parameter θ_0 is 0, then δ_n^* does much better than an efficient estimator; yet when $\theta_0 \neq 0$, δ_n^* does just as well.

Just as the invariance property of maximum likelihood estimation states that the MLE of a function of θ equals the same function applied to the MLE of θ, a function of an efficient estimator is itself efficient:

Theorem 7.12 If δ_n is efficient for θ_0, and if $g(\theta)$ is a differentiable and invertible function with $g'(\theta_0) \neq 0$, $g(\delta_n)$ is efficient for $g(\theta_0)$.

The proof of the above theorem follows immediately from the delta method, since the Fisher information for $g(\theta)$ is $I(\theta)/\{g'(\theta)\}^2$ by Equation (7.6). In fact, if one simply remembers

the content of Theorem 7.12, then it is not necessary to memorize Equation (7.6), for it is always possible to rederive this equation quickly using the delta method!

We have already noted that (under suitable regularity conditions) if $\tilde{\theta}_n$ is a $\sqrt{n}$-consistent estimator of θ_0 and

$$\delta_n = \tilde{\theta}_n - \frac{\ell'(\tilde{\theta}_n)}{\ell''(\tilde{\theta}_n)}, \tag{7.17}$$

then δ_n is an efficient estimator of θ_0. Alternatively, we may set

$$\delta_n^* = \tilde{\theta}_n + \frac{\ell'(\tilde{\theta}_n)}{nI(\tilde{\theta}_n)} \tag{7.18}$$

and δ_n^* is also an efficient estimator of θ_0. Problem 7.8 asked you to prove the former fact regarding δ_n; the latter fact regarding δ_n^* is proved in nearly the same way because $-\frac{1}{n}\ell''(\tilde{\theta}_n) \xrightarrow{P} I(\theta_0)$. In Equation (7.17), as already remarked earlier, δ_n results from a single step of Newton's method; in Equation (7.18), δ_n^* results from a similar method called Fisher scoring. As is clear from comparing Equations (7.17) and (7.18), scoring differs from Newton's method in that the Fisher information is used in place of the negative second derivative of the loglikelihood function. In some examples, scoring and Newton's method are equivalent.

A note on terminology: The derivative of $\ell(\theta)$ is sometimes called the **score function**. Furthermore, $nI(\theta)$ and $-\ell''(\theta)$ are sometimes referred to as the **expected information** and the **observed information**, respectively.

Example 7.13 Suppose $X_1, \ldots, X_n$ are independent from a Cauchy location family with density

$$f_\theta(x) = \frac{1}{\pi\{1 + (x-\theta)^2\}}.$$

Then

$$\ell'(\theta) = \sum_{i=1}^{n} \frac{2(x_i - \theta)}{1 + (x_i - \theta)^2},$$

so the likelihood equation is very difficult to solve. However, an efficient estimator may still be created by starting with some $\sqrt{n}$-consistent estimator $\tilde{\theta}_n$, say the sample median, and using either Equation (7.17) or Equation (7.18). In the latter case, we obtain the simple estimator

$$\delta_n^* = \tilde{\theta}_n + \frac{2}{n}\ell'(\tilde{\theta}_n), \tag{7.19}$$

verification of which is the subject of Problem 7.11.

For the remainder of this section, we turn our attention to Bayes estimators, which give yet another source of efficient estimators. A Bayes estimator is the expected value of the posterior distribution, which of course depends on the prior chosen. Although we do not prove this fact here (see Ferguson §21 for details), any Bayes estimator is efficient under some very general conditions. The conditions are essentially those of Theorem 7.8 along with the stipulation that the prior density is positive everywhere on Ω. (Note that if the prior density is not positive on Ω, the Bayes estimator may not even be consistent.)

Example 7.14 Consider the binomial distribution with beta prior, say $X \sim \text{binomial}(n, p)$ and $p \sim \text{beta}(a, b)$. Then the posterior density of p is proportional to the product of the likelihood and the prior, which (ignoring multiplicative constants not involving p) equals

$$p^{a-1}(1-p)^{b-1} \times p^X(1-p)^{n-X}.$$

Therefore, the posterior distribution of p is $\text{beta}(a+X, b+n-X)$. The Bayes estimator is the expectation of this distribution, which equals $(a+X)/(a+b+n)$. If we let γ_n denote the Bayes estimator here, then

$$\sqrt{n}(\gamma_n - p) = \sqrt{n}\left(\frac{X}{n} - p\right) + \sqrt{n}\left(\gamma_n - \frac{X}{n}\right).$$

We may see that the rightmost term converges to zero in probability by writing

$$\sqrt{n}\left(\gamma_n - \frac{X}{n}\right) = \frac{\sqrt{n}}{a+b+n}\left(a - (a+b)\frac{X}{n}\right),$$

since $a - (a+b)X/n \xrightarrow{P} a - (a+b)p$ by the weak law of large numbers. In other words, the Bayes estimator in this example has the same limiting distribution as the MLE, X/n. It is possible to verify that the MLE is efficient in this case.

The central question when constructing a Bayes estimator is how to choose the prior distribution. We consider one class of prior distributions, called **Jeffreys priors**. (A common grammatical mistake is to write "Jeffrey's priors," but this is incorrect because they are named for Harold Jeffreys and the letter s is not possessive. Analogously, Bayes estimators are named for Thomas Bayes.)

For a Bayes estimator $\hat{\theta}_n$ of θ_0, we have

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\left(0, \frac{1}{I(\theta_0)}\right).$$

Since the limiting variance of $\hat{\theta}_n$ depends on $I(\theta_0)$, if $I(\theta)$ is not a constant, then some values of θ may be estimated more precisely than others. In analogy with the idea of variance-stabilizing transformations seen in Section 5.1.2, we might consider a reparameterization

$\eta = g(\theta)$ such that $\{g'(\theta)\}^2/I(\theta)$ is a constant. More precisely, if $g'(\theta) = c\sqrt{I(\theta)}$, then

$$\sqrt{n}\left\{g(\hat{\theta}_n) - \eta_0\right\} \xrightarrow{d} N(0, c^2).$$

So as not to influence the estimation of η, we choose as the Jeffreys prior a uniform prior on η. Therefore, the Jeffreys prior density on θ is proportional to $g'(\theta)$, which is proportional to $\sqrt{I(\theta)}$. Note that this may lead to an improper prior.

Example 7.15 In the case of Example 7.14, we may verify that for a Bernoulli(p) observation,

$$\begin{aligned} I(p) &= -\,\mathrm{E}\,\frac{d^2}{dp^2}\{X\log p + (1-X)\log(1-p)\} \\ &= \mathrm{E}\left(\frac{X}{p^2} + \frac{1-X}{(1-p)^2}\right) = \frac{1}{p(1-p)}. \end{aligned}$$

Thus, the Jeffreys prior on p in this case has a density proportional to $p^{-1/2}(1-p)^{-1/2}$. In other words, the prior is beta($\frac{1}{2}, \frac{1}{2}$). Therefore, the Bayes estimator corresponding to the Jeffreys prior is

$$\gamma_n = \frac{X + \frac{1}{2}}{n+1}.$$

Exercises for Section 7.3

Exercise 7.10 Verify the claim made in Example 7.11.

Exercise 7.11 If $f_\theta(x)$ forms a location family, so that $f_\theta(x) = f(x-\theta)$ for some density $f(x)$, then the Fisher information $I(\theta)$ is a constant (you may assume this fact without proof).

(a) Verify that for the Cauchy location family,

$$f_\theta(x) = \frac{1}{\pi\{1 + (x-\theta)^2\}},$$

we have $I(\theta) = \frac{1}{2}$.

(b) For 500 samples of size $n = 51$ from a standard Cauchy distribution, calculate the sample median $\tilde{\theta}_n$ and the efficient estimator δ_n^* of Equation (7.19). Compare the variances of $\tilde{\theta}_n$ and δ_n^* with their theoretical asymptotic limits.

Exercise 7.12 **(a)** Derive the Jeffreys prior on θ for a random sample from Poisson(θ). Is this prior proper or improper?

(b) What is the Bayes estimator of θ for the Jeffreys prior? Verify directly that this estimator is efficient.

Exercise 7.13 **(a)** Derive the Jeffreys prior on σ^2 for a random sample from $N(0, \sigma^2)$. Is this prior proper or improper?

(b) What is the Bayes estimator of σ^2 for the Jeffreys prior? Verify directly that this estimator is efficient.

7.4 The multiparameter case

Suppose now that the parameter is the vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$. If $X \sim f_{\boldsymbol{\theta}}(x)$, then $I(\boldsymbol{\theta})$, the information matrix, is the $k \times k$ matrix

$$I(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}} \left\{ \mathbf{h}_{\boldsymbol{\theta}}(X) \mathbf{h}_{\boldsymbol{\theta}}^{\top}(X) \right\},$$

where $\mathbf{h}_{\boldsymbol{\theta}}(x) = \nabla_{\boldsymbol{\theta}} [\log f_{\boldsymbol{\theta}}(x)]$. This is a rare instance in which it's probably clearer to use component-wise notation than vector notation:

Definition 7.16 Given a k-dimensional parameter vector θ and a density function $f_{\boldsymbol{\theta}}(x)$, the Fisher information matrix $I(\boldsymbol{\theta})$ is the $k \times k$ matrix whose (i, j) element equals

$$I_{ij}(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}} \left\{ \frac{\partial}{\partial \theta_i} \log f_{\boldsymbol{\theta}}(X) \frac{\partial}{\partial \theta_j} \log f_{\boldsymbol{\theta}}(X) \right\},$$

as long as the above quantity is defined for all (i, j).

Note that the one-dimensional Definition 7.6 is a special case of Definition 7.16.

Example 7.17 Let $\boldsymbol{\theta} = (\mu, \sigma^2)$ and suppose $X \sim N(\mu, \sigma^2)$. Then

$$\log f_{\boldsymbol{\theta}}(x) = -\frac{1}{2} \log \sigma^2 - \frac{(x - \mu)^2}{2\sigma^2} - \log \sqrt{2\pi},$$

so

$$\frac{\partial}{\partial \mu} \log f_{\boldsymbol{\theta}}(x) = \frac{x - \mu}{\sigma^2} \quad \text{and} \quad \frac{\partial}{\partial \sigma^2} \log f_{\boldsymbol{\theta}}(x) = \frac{1}{2\sigma^2} \left(\frac{(x - \mu)^2}{\sigma^2} - 1 \right).$$

Thus, the entries in the information matrix are as follows:

$$\begin{aligned} I_{11}(\boldsymbol{\theta}) &= \mathrm{E}_{\boldsymbol{\theta}}\left(\frac{(X-\mu)^2}{\sigma^4}\right) = \frac{1}{\sigma^2}, \\ I_{21}(\boldsymbol{\theta}) = I_{12}(\boldsymbol{\theta}) &= \mathrm{E}_{\boldsymbol{\theta}}\left(\frac{(X-\mu)^3}{2\sigma^6} - \frac{X-\mu}{2\sigma^4}\right) = 0, \\ I_{22}(\boldsymbol{\theta}) &= \mathrm{E}_{\boldsymbol{\theta}}\left(\frac{1}{4\sigma^2} - \frac{(X-\mu)^2}{2\sigma^6} + \frac{(X-\mu)^4}{4\sigma^8}\right) = \frac{1}{4\sigma^4} - \frac{\sigma^2}{2\sigma^6} + \frac{3\sigma^4}{4\sigma^8} = \frac{1}{2\sigma^4}. \end{aligned}$$

As in the one-dimensional case, Definition 7.16 is often not the easiest form of $I(\boldsymbol{\theta})$ to work with. This fact is illustrated by Example 7.17, in which the calculation of the information matrix requires the evaluation of a fourth moment as well as a lot of algebra. However, in analogy with Equations (7.7) and (7.8), if

$$0 = \frac{\partial}{\partial\theta_i}\mathrm{E}_{\boldsymbol{\theta}}\frac{f_{\boldsymbol{\theta}}(X)}{f_{\boldsymbol{\theta}}(X)} = \mathrm{E}_{\boldsymbol{\theta}}\frac{\frac{\partial}{\partial\theta_i}f_{\boldsymbol{\theta}}(X)}{f_{\boldsymbol{\theta}}(X)} \quad \text{and} \quad 0 = \frac{\partial^2}{\partial\theta_i\partial\theta_j}\mathrm{E}_{\boldsymbol{\theta}}\frac{f_{\boldsymbol{\theta}}(X)}{f_{\boldsymbol{\theta}}(X)} = \mathrm{E}_{\boldsymbol{\theta}}\frac{\frac{\partial^2}{\partial\theta_i\partial\theta_j}f_{\boldsymbol{\theta}}(X)}{f_{\boldsymbol{\theta}}(X)} \tag{7.20}$$

for all i and j, then the following alternative forms of $I(\boldsymbol{\theta})$ are valid:

$$\begin{aligned} I_{ij}(\boldsymbol{\theta}) &= \mathrm{Cov}_{\boldsymbol{\theta}}\left(\frac{\partial}{\partial\theta_i}\log f_{\boldsymbol{\theta}}(X), \frac{\partial}{\partial\theta_j}\log f_{\boldsymbol{\theta}}(X)\right) && (7.21) \\ &= -\mathrm{E}_{\boldsymbol{\theta}}\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f_{\boldsymbol{\theta}}(X)\right). && (7.22) \end{aligned}$$

Example 7.18 In the normal case of Example 7.17, the information matrix is perhaps a bit easier to compute using Equation (7.22), since we obtain

$$\frac{\partial^2}{\partial\mu^2}\log f_{\boldsymbol{\theta}}(x) = -\frac{1}{\sigma^2}, \quad \frac{\partial^2}{\partial\mu\partial\sigma^2}\log f_{\boldsymbol{\theta}}(x) = -\frac{x-\mu}{\sigma^4},$$

and

$$\frac{\partial^2}{\partial(\sigma^2)^2}\log f_{\boldsymbol{\theta}}(x) = \frac{1}{2\sigma^4} - \frac{(x-\mu)^2}{\sigma^6}.$$

Taking expectations gives

$$I(\boldsymbol{\theta}) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}$$

as before but without requiring any fourth moments.

By Equation (7.21), the Fisher information matrix is nonnegative definite, just as the Fisher information is nonnegative in the one-parameter case. A further fact the generalizes into the multiparameter case is the additivity of information: If X and Y are independent, then $I_X(\boldsymbol{\theta}) + I_Y(\boldsymbol{\theta}) = I_{(X,Y)}(\boldsymbol{\theta})$. Finally, suppose that $\boldsymbol{\eta} = g(\boldsymbol{\theta})$ is a reparameterization, where $g(\boldsymbol{\theta})$ is invertible and differentiable. Then if J is the Jacobian matrix of the inverse transformation (i.e., $J_{ij} = \partial\theta_i/\partial\eta_j$), then the information about $\boldsymbol{\eta}$ is

$$I(\boldsymbol{\eta}) = J^\top I(\boldsymbol{\theta}) J.$$

As we'll see later, almost all of the same efficiency results that applied to the one-parameter case apply to the multiparameter case as well. In particular, we will see that an efficient estimator $\hat{\boldsymbol{\theta}}$ is one that satisfies

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) \xrightarrow{d} N_k\{\mathbf{0}, I(\boldsymbol{\theta}^0)^{-1}\}.$$

Note that the formula for the information under a reparameterization implies that if $\hat{\boldsymbol{\eta}}$ is an efficient estimator for $\boldsymbol{\eta}^0$ and the matrix J is invertible, then $g^{-1}(\hat{\boldsymbol{\eta}})$ is an efficient estimator for $\boldsymbol{\theta}^0 = g^{-1}(\boldsymbol{\eta}^0)$, since

$$\sqrt{n}\{g^{-1}(\hat{\boldsymbol{\eta}}) - g^{-1}(\boldsymbol{\eta}^0)\} \xrightarrow{d} N_k\{\mathbf{0}, J(J^\top I(\boldsymbol{\theta}^0) J)^{-1} J^\top\},$$

and the covariance matrix above equals $I(\boldsymbol{\theta}^0)^{-1}$.

As in the one-parameter case, the likelihood equation is obtained by setting the derivative of $\ell(\boldsymbol{\theta})$ equal to zero. In the multiparameter case, though, the gradient $\nabla\ell(\boldsymbol{\theta})$ is a $1 \times k$ vector, so the likelihood equation becomes $\nabla\ell(\boldsymbol{\theta}) = \mathbf{0}$. Since there are really k univariate equations implied by this likelihood equation, it is common to refer to the likelihood equations (plural), which are

$$\frac{\partial}{\partial\theta_i}\ell(\boldsymbol{\theta}) = 0 \quad \text{for } i = 1, \ldots, k.$$

In the multiparameter case, we have essentially the same theorems as in the one-parameter case.

Theorem 7.19 Suppose that $X_1, \ldots, X_n$ are independent and identically distributed random variables (or vectors) with density $f_{\boldsymbol{\theta}^0}(x)$ for $\boldsymbol{\theta}^0$ in an open subset Ω of R^k, where distinct values of $\boldsymbol{\theta}^0$ yield distinct distributions for X_1 (i.e., the parameter is identifiable). Furthermore, suppose that the support set $\{x : f_{\boldsymbol{\theta}}(x) > 0\}$ does not depend on $\boldsymbol{\theta}$. Then with probability approaching 1 as $n \to \infty$, there exists $\hat{\boldsymbol{\theta}}$ such that $\nabla\ell(\hat{\boldsymbol{\theta}}) = \mathbf{0}$ and $\hat{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}^0$.

As in the one-parameter case, we shall refer to the $\hat{\boldsymbol{\theta}}$ guaranteed by Theorem 7.19 as a consistent root of the likelihood equations. Unlike Theorem 7.4, however, Corollary 7.5 does

not generalize to the multiparameter case because it is possible that $\hat{\boldsymbol{\theta}}$ is the unique solution of the likelihood equations and a local maximum but not the MLE. The best we can say is the following:

Corollary 7.20 Under the conditions of Theorem 7.19, if there is a unique root of the likelihood equations, then this root is consistent for $\boldsymbol{\theta}^0$.

The asymptotic normality of a consistent root of the likelihood equation holds in the multiparameter case just as in the single-parameter case:

Theorem 7.21 Suppose that the conditions of Theorem 7.19 are satisfied and that $\hat{\boldsymbol{\theta}}$ denotes a consistent root of the likelihood equations. Assume also that Equation (7.20) is satisfied for all i and j and that $I(\boldsymbol{\theta}^0)$ is positive definite with finite entries. Finally, assume that $\partial^3 \log f_{\boldsymbol{\theta}}(x)/\partial\theta_i\partial\theta_j\partial\theta_k$ exists and is bounded in a neighborhood of $\boldsymbol{\theta}^0$ for all i, j, k. Then

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) \xrightarrow{d} N_k\{\mathbf{0}, I^{-1}(\boldsymbol{\theta}^0)\}.$$

As in the one-parameter case, there is some terminology associated with the derivatives of the loglikelihood function. The gradient vector $\nabla\ell(\boldsymbol{\theta})$ is called the score vector. The negative second derivative $-\nabla^2\ell(\boldsymbol{\theta})$ is called the observed information, and $nI(\boldsymbol{\theta})$ is sometimes called the expected information. And the second derivative of a real-valued function of a k-vector, such as the loglikelihood function $\ell(\boldsymbol{\theta})$, is called the Hessian matrix.

Newton's method (often called the Newton-Raphson method in the multivariate case) and scoring work just as they do in the one-parameter case. Starting from the point $\tilde{\boldsymbol{\theta}}_n$, one step of Newton-Raphson gives

$$\boldsymbol{\delta}_n = \tilde{\boldsymbol{\theta}}_n - \left\{\nabla^2\ell(\tilde{\boldsymbol{\theta}}_n)\right\}^{-1} \nabla\ell(\tilde{\boldsymbol{\theta}}_n) \tag{7.23}$$

and one step of scoring gives

$$\boldsymbol{\delta}_n^* = \tilde{\boldsymbol{\theta}}_n + \frac{1}{n} I^{-1}(\tilde{\boldsymbol{\theta}}_n)\nabla\ell(\tilde{\boldsymbol{\theta}}_n). \tag{7.24}$$

Theorem 7.22 Under the assumptions of Theorem 7.21, if $\tilde{\boldsymbol{\theta}}_n$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\theta}^0$, then the one-step Newton-Raphson estimator $\boldsymbol{\delta}_n$ defined in Equation (7.23) satisfies

$$\sqrt{n}(\boldsymbol{\delta}_n - \boldsymbol{\theta}^0) \xrightarrow{d} N_k\{\mathbf{0}, I^{-1}(\boldsymbol{\theta}^0)\}$$

and the one-step scoring estimator $\boldsymbol{\delta}_n^*$ defined in Equation (7.24) satisfies

$$\sqrt{n}(\boldsymbol{\delta}_n^* - \boldsymbol{\theta}^0) \xrightarrow{d} N_k\{\mathbf{0}, I^{-1}(\boldsymbol{\theta}^0)\}.$$

As in the one-parameter case, we define an efficient estimator $\hat{\boldsymbol{\theta}}_n$ as one that satisfies

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0) \xrightarrow{d} N_k\{\mathbf{0}, I^{-1}(\boldsymbol{\theta}^0)\}.$$

This definition is justified by the fact that if $\boldsymbol{\delta}_n$ is any estimator satisfying

$$\sqrt{n}(\boldsymbol{\delta}_n - \boldsymbol{\theta}) \xrightarrow{d} N_k\{\mathbf{0}, \Sigma(\boldsymbol{\theta})\},$$

where $I^{-1}(\boldsymbol{\theta})$ and $\Sigma(\boldsymbol{\theta})$ are continuous, then $\Sigma(\boldsymbol{\theta}) - I^{-1}(\boldsymbol{\theta})$ is nonnegative definite for all $\boldsymbol{\theta}$. This is analagous to saying that $\sigma^2(\theta) - I^{-1}(\theta)$ is nonnegative in the univariate case, which is to say that $I^{-1}(\theta)$ is the smallest possible variance.

Finally, we note that Bayes estimators are efficient, just as in the one-parameter case. This means that the same three types of estimators that are efficient in the one-parameter case—the consistent root of the likelihood equation, the one-step scoring and Newton-Raphson estimators, and Bayes estimators—are also efficient in the multiparameter case.

Exercises for Section 7.4

Exercise 7.14 Let $\mathbf{X} \sim$ multinomial$(1, \mathbf{p})$, where $\mathbf{p}$ is a k-vector for $k > 2$. Let $\mathbf{p}^* = (p_1, \ldots, p_{k-1})$. Find $I(\mathbf{p}^*)$.

Exercise 7.15 Suppose that $\boldsymbol{\theta} \in R \times R_+$ (that is, $\theta_1 \in R$ and $\theta_2 \in (0, \infty)$) and

$$f_{\boldsymbol{\theta}}(x) = \frac{1}{\theta_2} f\left(\frac{x - \theta_1}{\theta_2}\right)$$

for some continuous, differentiable density $f(x)$ that is symmetric about the origin. Find $I(\boldsymbol{\theta})$.

Exercise 7.16 Prove Theorem 7.21.

Hint: Use Theorem 1.40.

Exercise 7.17 Prove Theorem 7.22.

Hint: Use Theorem 1.40.

Exercise 7.18 The multivariate generalization of a beta distribution is a Dirichlet distribution, which is the natural prior distribution for a multinomial likelihood. If $\mathbf{p}$ is a random $(k+1)$-vector constrained so that $p_i > 0$ for all i and $\sum_{i=1}^{k+1} p_i = 1$, then $(p_1, \ldots, p_k)$ has a Dirichlet distribution with parameters $a_1 > 0, \ldots, a_{k+1} > 0$ if its density is proportional to

$$p_1^{a_1 - 1} \cdots p_k^{a_k - 1}(1 - p_1 - \cdots - p_k)^{a_{k+1} - 1} I\left\{\min_i p_i > 0\right\} I\left\{\sum_{i=1}^{k} p_i < 1\right\}.$$

Prove that if $G_1, \ldots, G_{k+1}$ are independent random variables with $G_i \sim \text{gamma}(a_i, 1)$, then

$$\frac{1}{G_1 + \cdots + G_{k+1}}(G_1, \ldots, G_k)$$

has a Dirichlet distribution with parameters $a_1, \ldots, a_{k+1}$.

7.5 Nuisance parameters

This section is the one intrinsically multivariate section in this chapter; it does not have an analogue in the one-parameter setting. Here we consider how efficiency of an estimator is affected by the presence of nuisance parameters.

Suppose $\boldsymbol{\theta}$ is the parameter vector but θ_1 is the only parameter of interest, so that $\theta_2, \ldots, \theta_k$ are nuisance parameters. We are interested in how the asymptotic precision with which we may estimate θ_1 is influenced by the presence of the nuisance parameters. In other words, if $\hat{\boldsymbol{\theta}}$ is efficient for $\boldsymbol{\theta}$, then how does $\hat{\theta}_1$ as an estimator of θ_1 compare to an efficient estimator of θ_1, say θ^*, under the assumption that all of the nuisance parameters are known?

Assume $I(\boldsymbol{\theta})$ is positive definite. Let σ_{ij} denote the (i,j) entry of $I(\boldsymbol{\theta})$ and let γ_{ij} denote the (i,j) entry of $I^{-1}(\boldsymbol{\theta})$. If all of the nuisance parameters are known, then $I(\theta_1) = \sigma_{11}$, which means that the asymptotic variance of $\sqrt{n}(\theta^* - \theta_1)$ is $1/\sigma_{11}$. On the other hand, if the nuisance parameters are not known then the asymptotic variance of $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is $I^{-1}(\boldsymbol{\theta})$, which means that the marginal asymptotic variance of $\sqrt{n}(\hat{\theta}_1 - \theta_1)$ is γ_{11}. Of interest here is the comparison between γ_{11} and $1/\sigma_{11}$.

The following theorem may be interpreted to mean that the presence of nuisance parameters always increases the variance of an efficient estimator.

Theorem 7.23 $\gamma_{11} \geq 1/\sigma_{11}$, with equality if and only if $\gamma_{12} = \cdots = \gamma_{1k} = 0$.

Proof: Partition $I(\boldsymbol{\theta})$ as follows:

$$I(\boldsymbol{\theta}) = \begin{pmatrix} \sigma_{11} & \boldsymbol{\rho}^\top \\ \boldsymbol{\rho} & \Sigma \end{pmatrix},$$

where $\boldsymbol{\rho}$ and Σ are $(k-1) \times 1$ and $(k-1) \times (k-1)$, respectively. Let $\tau = \sigma_{11} - \boldsymbol{\rho}^\top \Sigma^{-1} \boldsymbol{\rho}$. We may verify that if $\tau > 0$, then

$$I^{-1}(\boldsymbol{\theta}) = \frac{1}{\tau} \begin{pmatrix} 1 & -\boldsymbol{\rho}^\top \Sigma^{-1} \\ -\Sigma^{-1}\boldsymbol{\rho} & \Sigma^{-1}\boldsymbol{\rho}\boldsymbol{\rho}^\top \Sigma^{-1} + \tau \Sigma^{-1} \end{pmatrix}.$$

This proves the result, because the positive definiteness of $I(\boldsymbol{\theta})$ implies that Σ^{-1} is positive definite, which means that

$$\gamma_{11} = \frac{1}{\sigma_{11} - \boldsymbol{\rho}^\top \Sigma^{-1} \boldsymbol{\rho}} \geq \frac{1}{\sigma_{11}},$$

with equality if and only if $\boldsymbol{\rho} = \mathbf{0}$. Thus, it remains only to show that $\tau > 0$. But this is immediate from the positive definiteness of $I(\boldsymbol{\theta})$, since if we set

$$\mathbf{v} = \begin{pmatrix} 1 \\ -\boldsymbol{\rho}^\top \Sigma^{-1} \end{pmatrix},$$

then $\tau = \mathbf{v}^\top I(\boldsymbol{\theta}) \mathbf{v}$. ■

The above result shows that it is important to take nuisance parameters into account in estimation. However, it is not necessary to estimate the entire parameter vector all at once, since $(\hat{\theta}_1, \ldots, \hat{\theta}_k)$ is efficient for $\boldsymbol{\theta}$ if and only if each of the $\hat{\theta}_i$ is efficient for θ_i in the presence of the other nuisance parameters (see problem 7.19).

Exercises for Section 7.5

Exercise 7.19 Letting γ_{ij} denote the (i, j) entry of $I^{-1}(\boldsymbol{\theta})$, we say that $\hat{\theta}_i$ is efficient for θ_i in the presence of the nuisance parameters $\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_k$ if the asymptotic variance of $\sqrt{n}(\hat{\theta}_i - \theta_i)$ is γ_{ii}.

Prove that $(\hat{\theta}_1, \ldots, \hat{\theta}_k)$ is efficient for $\boldsymbol{\theta}$ if and only if for all i, the estimator $\hat{\theta}_i$ is efficient for θ_i in the presence of nuisance parameters $\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_k$.

Chapter 8

Hypothesis Testing

8.1 Wald, Rao, and Likelihood Ratio Tests

Suppose we wish to test $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$. The likelihood-based results of Chapter 7 give rise to several possible tests.

To this end, let $\ell(\theta)$ denote the loglikelihood and $\hat{\theta}_n$ the consistent root of the likelihood equation. Intuitively, the farther θ_0 is from $\hat{\theta}_n$, the stronger the evidence against the null hypothesis. But how far is "far enough"? Note that if θ_0 is close to $\hat{\theta}_n$, then $\ell(\theta_0)$ should also be close to $\ell(\hat{\theta}_n)$ and $\ell'(\theta_0)$ should be close to $\ell'(\hat{\theta}_n) = 0$.

- If we base a test upon the value of $(\hat{\theta}_n - \theta_0)$, we obtain a **Wald test**.
- If we base a test upon the value of $\ell(\hat{\theta}_n) - \ell(\theta_0)$, we obtain a **likelihood ratio test**.
- If we base a test upon the value of $\ell'(\theta_0)$, we obtain a **(Rao) score test**.

Recall that in order to prove Theorem 7.8, we argued that under certain regularity conditions, the following facts are true under H_0:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\left(0, \frac{1}{I(\theta_0)}\right); \tag{8.1}$$

$$\frac{1}{\sqrt{n}}\ell'(\theta_0) \xrightarrow{d} N\{0, I(\theta_0)\}; \tag{8.2}$$

$$-\frac{1}{n}\ell''(\theta_0) \xrightarrow{P} I(\theta_0). \tag{8.3}$$

Equation (8.2) is proved using the central limit theorem; Equation (8.3) is proved using the weak law of large numbers; and Equation (8.1) is the result of a Taylor expansion together with Equations (8.2) and (8.3). The three equations above will help us to justify the definitions of Wald, score, and likelihood ratio tests to follow.

Equation (8.1) suggests that if we define

$$W_n = \sqrt{nI(\theta_0)}(\hat{\theta}_n - \theta_0), \tag{8.4}$$

then W_n, called a Wald statistic, should converge in distribution to standard normal under H_0. Note that this fact remains true if we define

$$W_n = \sqrt{n\hat{I}}(\hat{\theta}_n - \theta_0), \tag{8.5}$$

where $\hat{I}$ is consistent for θ_0; for example, $\hat{I}$ could be $I(\hat{\theta}_n)$.

Definition 8.1 A **Wald test** is any test that rejects $H_0 : \theta = \theta_0$ in favor of $H_1 : \theta \neq \theta_0$ when $|W_n| \geq u_{\alpha/2}$ for W_n defined as in Equation (8.4) or Equation (8.5). As usual, $u_{\alpha/2}$ denotes the $1 - \frac{\alpha}{2}$ quantile of the standard normal distribution.

Equation (8.2) suggests that if we define

$$R_n = \frac{1}{\sqrt{nI(\theta_0)}}\ell'(\theta_0), \tag{8.6}$$

then R_n, called a Rao score statistic or simply a score statistic, converges in distribution to standard normal under H_0. We could also replace $I(\theta_0)$ by a consistent estimator $\hat{I}$ as in Equation (8.5), but usually this is not done: One of the main benefits of the score statistic is that it is not necessary to compute $\hat{\theta}_n$, and using $I(\hat{\theta}_n)$ instead of $I(\theta_0)$ would defeat this purpose.

Definition 8.2 A **score test**, sometimes called a **Rao score test**, is any test that rejects $H_0 : \theta = \theta_0$ in favor of $H_1 : \theta \neq \theta_0$ when $|R_n| \geq u_{\alpha/2}$ for R_n defined as in Equation (8.6).

The third type of test, the likelihood ratio test, requires a bit of development. It is a test based on the statistic

$$\Delta_n = \ell(\hat{\theta}_n) - \ell(\theta_0). \tag{8.7}$$

If we Taylor expand $\ell'(\hat{\theta}_n) = 0$ around the point θ_0, we obtain

$$\ell'(\theta_0) = -(\hat{\theta}_n - \theta_0)\left\{\ell''(\theta_0) + \frac{\hat{\theta}_n - \theta_0}{2}\ell'''(\theta^*)\right\}. \tag{8.8}$$

We now use a second Taylor expansion, this time of $\ell(\hat{\theta}_n)$, to obtain

$$\Delta_n = (\hat{\theta}_n - \theta_0)\ell'(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)^2 \left[\ell''(\theta_0) + \frac{\hat{\theta}_n - \theta_0}{3}\ell'''(\theta^{**}) \right]. \tag{8.9}$$

If we now substitute Equation (8.8) into Equation (8.9), we obtain

$$\Delta_n = n(\hat{\theta}_n - \theta_0)^2 \left\{ -\frac{1}{n}\ell''(\theta_0) + \frac{1}{2n}\ell''(\theta_0) + (\hat{\theta}_n - \theta_0)O_P(1) \right\}, \tag{8.10}$$

where the $O_P(1)$ term consists of the sum of $-\ell'''(\theta^*)/(2n)$ and $\ell'''(\theta^{**})/(6n)$, which is bounded in probability under the third-derivative assumption of Theorem 7.8.

By Equations (8.1) and (8.3) and Slutsky's theorem, this implies that $2\Delta_n \xrightarrow{d} \chi^2_1$ under the null hypothesis. Noting that the $1-\alpha$ quantile of a χ^2_1 distribution is $u^2_{\alpha/2}$, we make the following definition.

Definition 8.3 A **likelihood ratio test** is any test that rejects $H_0 : \theta = \theta_0$ in favor of $H_1 : \theta \neq \theta_0$ when $2\Delta_n \geq u^2_{\alpha/2}$ or, equivalently, when $\sqrt{2\Delta_n} \geq u_{\alpha/2}$.

Since it may be shown that $\sqrt{2\Delta_n} - |W_n| \xrightarrow{P} 0$ and $W_n - R_n \xrightarrow{P} 0$, the three tests defined above — Wald tests, score tests, and likelihood ratio tests — are asymptotically equivalent in the sense that under H_0, they reach the same decision with probability approaching 1 as $n \to \infty$. However, they may be quite different for a fixed sample size n, and furthermore they have some relative advantages and disadvantages with respect to one another. For example,

- It is straightforward to create one-sided Wald and score tests (i.e., tests of $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$ or $H_1 : \theta < \theta_0$), but this is more difficult with a likelihood ratio test.
- The score test does not require $\hat{\theta}_n$ whereas the other two tests do.
- The Wald test is most easily interpretable and yields immediate confidence intervals.
- The score test and likelihood ratio test are invariant under reparameterization, whereas the Wald test is not.

Example 8.4 Suppose that $X_1, \ldots, X_n$ are independent with density $f_\theta(x) = \theta e^{-x\theta} I\{x > 0\}$. Then $\ell(\theta) = n(\log\theta - \theta\overline{X}_n)$, which yields

$$\ell'(\theta) = n\left(\frac{1}{\theta} - \overline{X}_n\right) \quad \text{and} \quad \ell''(\theta) = -\frac{n}{\theta^2}.$$

From the form of $\ell''(\theta)$, we see that $I(\theta) = \theta^{-2}$, and setting $\ell'(\theta) = 0$ yields $\hat{\theta}_n = 1/\overline{X}_n$. From these facts, we obtain as the Wald, score, and likelihood ratio statistics

$$\begin{aligned} W_n &= \frac{\sqrt{n}}{\theta_0}\left(\frac{1}{\overline{X}_n} - \theta_0\right), \\ R_n &= \theta_0\sqrt{n}\left(\frac{1}{\theta_0} - \overline{X}_n\right) = \frac{W_n}{\theta_0\overline{X}_n}, \quad \text{and} \\ \Delta_n &= n\left\{\overline{X}_n\theta_0 - 1 - \log(\theta_0\overline{X}_n)\right\}. \end{aligned}$$

Thus, we reject $H_0 : \theta = \theta_0$ in favor of $H_1 : \theta \neq \theta_0$ whenever $|W_n| \geq u_{\alpha/2}$, $|R_n| \geq u_{\alpha/2}$, or $\sqrt{2\Delta_n} \geq u_{\alpha/2}$, depending on which test we're using.

Generalizing the three tests to the multiparameter setting is straightforward. Suppose we wish to test $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}^0$ against $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}^0$, where $\boldsymbol{\theta} \in R^k$. Then

$$W_n \stackrel{\text{def}}{=} n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0)^\top I(\boldsymbol{\theta}^0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) \stackrel{d}{\to} \chi^2_k, \tag{8.11}$$

$$R_n \stackrel{\text{def}}{=} \frac{1}{n}\nabla\ell(\boldsymbol{\theta}^0)^\top I^{-1}(\boldsymbol{\theta}^0)\nabla\ell(\boldsymbol{\theta}^0) \stackrel{d}{\to} \chi^2_k, \quad \text{and} \tag{8.12}$$

$$\Delta_n \stackrel{\text{def}}{=} \ell(\hat{\boldsymbol{\theta}}) - \ell(\boldsymbol{\theta}^0) \stackrel{d}{\to} \frac{1}{2}\chi^2_k. \tag{8.13}$$

Therefore, if c^k_α denotes the $1-\alpha$ quantile of the χ^2_k distribution, then the multivariate Wald test, score test, and likelihood ratio test reject H_0 when $W_n \geq c^k_\alpha$, $R_n \geq c^k_\alpha$, and $2\Delta_n \geq c^k_\alpha$, respectively. As in the one-parameter case, the Wald test may also be defined with $I(\boldsymbol{\theta}^0)$ replaced by a consistent estimator $\hat{I}$.

Exercises for Section 8.1

Exercise 8.1 Let $X_1, \ldots, X_n$ be a simple random sample from a Pareto distribution with density

$$f(x) = \theta c^\theta x^{-(\theta+1)} I\{x > c\}$$

for a known constant $c > 0$ and parameter $\theta > 0$. Derive the Wald, Rao, and likelihood ratio tests of $\theta = \theta_0$ against a two-sided alternative.

Exercise 8.2 Suppose that $\mathbf{X}$ is multinomial$(n, \mathbf{p})$, where $\mathbf{p} \in R^k$. In order to satisfy the regularity condition that the parameter space be an open set, define $\boldsymbol{\theta} = (p_1, \ldots, p_{k-1})$. Suppose that we wish to test $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}^0$ against $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}^0$.

(a) Prove that the Wald and score tests are the same as the usual Pearson chi-square test.

(b) Derive the likelihood ratio statistic $2\Delta_n$.

8.2 Contiguity and Local Alternatives

Suppose that we wish to test the hypotheses

$$H_0 : \theta = \theta_0 \quad \text{and} \quad H_1 : \theta > \theta_0. \tag{8.14}$$

The test is to be based on a statistic T_n, where as always n denotes the sample size, and we shall decide to

$$\text{reject } H_0 \text{ in (8.14) if } T_n \geq C_n \tag{8.15}$$

for some constant C_n. The test in question may be one of the three types of tests introduced in Section 8.1, or it may be an entirely different test. We may define some basic asymptotic concepts regarding tests of this type.

Definition 8.5 If $P_{\theta_0}(T_n \geq C_n) \to \alpha$ for test (8.15), then test (8.15) is said to have asymptotic level α.

Definition 8.6 If two different tests of the same hypotheses reach the same conclusion with probability approaching 1 under the null hypothesis as $n \to \infty$, the tests are said to be asymptotically equivalent.

The power of test (8.15) under the alternative θ is defined to be

$$\beta_n(\theta) = P_\theta(T_n \geq C_n).$$

We expect that the power should approach 1.

Definition 8.7 A test (or, more precisely, a sequence of tests) is said to be consistent against the alternative θ if $\beta_n(\theta) \to 1$.

Note that in some contexts, β is used to denote the type II error probability, which is actually one minus the power. We admit that the inconsistent usage of β in the literature is confusing, but we hope that bringing attention to this inconsistency will help to allay much of this confusion. Here, β will always refer to power.

Unfortunately, the concepts we have defined so far are of limited usefulness. If we wish to compare two different tests of the same hypotheses, then if the tests are both sensible they should be asymptotically equivalent and consistent. Thus, consistency is nice but it doesn't tell us much; asymptotic equivalence is nice but it doesn't allow us to compare tests.

We make things more interesting by considering, instead of a fixed alternative θ, a sequence of alternatives $\theta_1, \theta_2, \ldots$. Let us make some assumptions about the asymptotic distribution

of the test statistic T_n in (8.15). First, define $\mu(\theta)$ and $\tau^2(\theta)/n$ to be the mean and variance, respectively, of T_n when θ is the true value of the parameter. In other words, let

$$\mu(\theta) = \mathrm{E}_\theta(T_n) \text{ and } \tau^2(\theta) = n \operatorname{Var}_\theta(T_n).$$

We assume that *if the null hypothesis is true*, which means θ is fixed and equal to θ_0, then

$$\frac{\sqrt{n}\{T_n - \mu(\theta_0)\}}{\sqrt{\tau^2(\theta_0)}} \xrightarrow{d} N(0,1) \tag{8.16}$$

as $n \to \infty$. Furthermore, we assume that *if the alternatives are true*, which means that the distribution of T_n is determined by $\theta = \theta_n$ for each n, then

$$\frac{\sqrt{n}\{T_n - \mu(\theta_n)\}}{\sqrt{\tau^2(\theta_n)}} \xrightarrow{d} N(0,1) \tag{8.17}$$

as $n \to \infty$. The limit in (8.17) is trickier than the one in (8.16) because it assumes that the underlying parameter is changing along with n. Consider Example 8.8.

Example 8.8 Suppose we have a sequence of independently and identically distributed random variables $X_1, X_2, \ldots$ with common distribution F_θ. In a one-sample t-test, the test statistic T_n may be taken to be the sample mean $\overline{X}_n$. In this case, the limit in (8.16) follows immediately from the central limit theorem. Yet to verify (8.17), it is necessary to consider a triangular array of random variables:

$$\begin{aligned} X_{11} &\sim F_{\theta_1} \\ X_{21}, X_{22} &\sim F_{\theta_2} \\ X_{31}, X_{32}, X_{33} &\sim F_{\theta_3} \\ &\ldots \end{aligned}$$

We may often check that the Lyapunov or Lindeberg condition is satisfied, then use the results of Section 4.2 to establish (8.17). In fact, the existence of, say, a finite third absolute central moment, $\gamma(\theta) = \mathrm{E}_\theta\,|X_1 - \mathrm{E}\,X_1|^3$, is generally sufficient because

$$\frac{1}{(\sqrt{n\tau^2(\theta_n)})^3} \sum_{i=1}^{n} \mathrm{E}\,|X_{ni} - \mathrm{E}\,X_{ni}|^3 = \frac{\gamma(\theta_n)}{\sqrt{n}\tau^3(\theta_n)}$$

and the Lyapunov condition holds as long as $\gamma(\theta_n)/\tau^3(\theta_n)$ tends to some finite limit. We generally assume that $\theta_n \to \theta_0$, so as long as $\gamma(\theta)$ and $\tau(\theta)$ are continuous, $\gamma(\theta_n)/\tau^3(\theta_n)$ tends to $\gamma(\theta_0)/\tau^3(\theta_0)$.

If (8.16) and (8.17) are true, we may calculate the power of the test against the sequence of alternatives $\theta_1, \theta_2, \ldots$ in a straightforward way using normal distribution calculations. As mentioned in Example 8.8, we typically assume that this sequence converges to θ_0. Such a sequence is often referred to as a *contiguous sequence* of alternatives, since "contiguous" means "next to"; the idea of contiguity is that we choose not a single alternative hypothesis but a sequence of alternatives next to the null hypothesis.

First, we should determine a value for C_n so that test (8.15) has asymptotic level α. Define u_α to be the $1-\alpha$ quantile of the standard normal distribution. By limit (8.16),

$$P_{\theta_0}\left\{T_n - \mu(\theta_0) \geq \frac{\tau(\theta_0)u_\alpha}{\sqrt{n}}\right\} \to \alpha;$$

therefore, we define a new test, namely

$$\text{reject } H_0 \text{ in (8.14) if } T_n \geq \mu(\theta_0) + \frac{\tau(\theta_0)u_\alpha}{\sqrt{n}} \tag{8.18}$$

and conclude that test (8.18) has asymptotic level α as desired.

We now calculate the power of test (8.18) against the alternative θ_n:

$$\begin{aligned}
\beta_n(\theta_n) &= P_{\theta_n}\left\{T_n \geq \mu(\theta_0) + \frac{\tau(\theta_0)u_\alpha}{\sqrt{n}}\right\} \\
&= P_{\theta_n}\left\{\frac{\sqrt{n}\{T_n - \mu(\theta_n)\}}{\tau(\theta_n)} \cdot \frac{\tau(\theta_n)}{\tau(\theta_0)} + \frac{\sqrt{n}(\theta_n - \theta_0)}{\tau(\theta_0)} \cdot \frac{\mu(\theta_n) - \mu(\theta_0)}{\theta_n - \theta_0} \geq u_\alpha\right\} \quad (8.19)
\end{aligned}$$

Thus, $\beta_n(\theta_n)$ tends to an interesting limit (i.e., a limit between α and 1) if $\tau(\theta_n) \to \tau(\theta_0)$; $\sqrt{n}(\theta_n - \theta_0)$ tends to a nonzero, finite limit; and $\mu(\theta)$ is differentiable at θ_0. This fact is summarized in the following theorem:

Theorem 8.9 Let $\theta_n > \theta_0$ for all n. Suppose that limits (8.17) and (8.16) hold, $\tau(\theta)$ is continuous at θ_0, $\mu(\theta)$ is differentiable at θ_0, and $\sqrt{n}(\theta_n - \theta_0) \to \Delta$ for some finite $\Delta > 0$. If $\mu'(\theta_0)$ or $\tau(\theta_0)$ depends on n, then suppose that $\mu'(\theta_0)/\tau(\theta_0)$ tends to a nonzero, finite limit. Then if $\beta_n(\theta_n)$ denotes the power of test (8.18) against the alternative θ_n,

$$\beta_n(\theta_n) \to \lim_{n\to\infty} \Phi\left(\frac{\Delta\mu'(\theta_0)}{\tau(\theta_0)} - u_\alpha\right).$$

The proof of Theorem 8.9 merely uses Equation (8.19) and Slutsky's theorem, since the hypotheses of the theorem imply that $\tau(\theta_n)/\tau(\theta_0) \to 1$ and $\{\mu(\theta_n) - \mu(\theta_0)\}/(\theta_n - \theta_0) \to \mu'(\theta_0)$.

Example 8.10 Let $X \sim \text{Binomial}(n, p_n)$, where $p_n = p_0 + \Delta/\sqrt{n}$ and $T_n = X/n$. To test $H_0 : p = p_0$ against $H_1 : p > p_0$, note that

$$\frac{\sqrt{n}(T_n - p_0)}{\sqrt{p_0(1-p_0)}} \xrightarrow{d} N(0,1)$$

under H_0. Thus, test (8.18) says to reject H_0 whenever $T_n \geq p_0 + u_\alpha \sqrt{p_0(1-p_0)/n}$. This test has asymptotic level α. Since $\tau(p) = \sqrt{p(1-p)}$ is continuous and $\mu(p) = p$ is differentiable, Theorem 8.9 applies in this case as long as we can verify the limit (8.17).

Let $X_{n1}, \ldots, X_{nn}$ be independent Bernoulli(p_n) random variables. Then if $X_{n1} - p_n, \ldots, X_{nn} - p_n$ can be shown to satisfy the Lyapunov condition, we have

$$\frac{\sqrt{n}(T_n - p_n)}{\tau(p_n)} \xrightarrow{d} N(0,1)$$

and so Theorem 8.9 applies. The Lyapunov condition follows since $|X_{ni} - p_n| \leq 1$ implies

$$\frac{1}{\{\text{Var}\,(nT_n)\}^2} \sum_{i=1}^{n} \text{E}\,|X_{ni} - p_n|^4 \leq \frac{n}{\{np_n(1-p_n)\}^2} \to 0.$$

Thus, we conclude by Theorem 8.9 that

$$\beta_n(p_n) \to \Phi\left(\frac{\Delta}{\sqrt{p_0(1-p_0)}} - u_\alpha \right).$$

To apply this result, suppose that we wish to test whether a coin is fair by flipping it 100 times. We reject $H_0 : p = 1/2$ in favor of $H_1 : p > 1/2$ if the number of successes divided by 100 is at least as large as $1/2 + u_{.05}/20$, or 0.582. The power of this test against the alternative $p = 0.6$ is approximately

$$\Phi\left(\frac{\sqrt{100}(0.6 - 0.5)}{\sqrt{0.5^2}} - 1.645 \right) = \Phi(2 - 1.645) = 0.639.$$

Compare this asymptotic approximation with the exact power: The probability of at least 59 successes out of 100 for a binomial$(100, 0.6)$ random variable is 0.623.

Starting from Equation (8.19) and using the fact that

$$\frac{\sqrt{n}\{T_n - \mu(\theta_n)\}}{\tau(\theta_n)} \cdot \frac{\tau(\theta_n)}{\tau(\theta_0)}$$

is approximately distributed as standard normal, we may obtain the following approximation to the power for a fixed sample size n and a fixed alternative θ as follows:

$$\beta_n(\theta) \approx \Phi\left(\frac{\sqrt{n}\{\mu(\theta) - \mu(\theta_0)\}}{\tau(\theta_0)} - u_\alpha\right). \tag{8.20}$$

Note that in the case of Example 8.10, the approximation of Equation (8.20) is the same as the approximation obtained from Theorem 8.9 by setting $\Delta = \sqrt{n}(p - p_0)$.

There is an alternative formulation that yields a slightly different approximation. Starting from

$$\beta_n(\theta) = P_\theta\left\{\frac{\sqrt{n}\{T_n - \mu(\theta)\}}{\tau(\theta)} \geq u_\alpha\frac{\tau(\theta_0)}{\tau(\theta)} - \frac{\sqrt{n}\{\mu(\theta) - \mu(\theta_0)\}}{\tau(\theta)}\right\},$$

we obtain

$$\beta_n(\theta) \approx \Phi\left(\frac{\sqrt{n}\{\mu(\theta) - \mu(\theta_0)\}}{\tau(\theta)} - u_\alpha\frac{\tau(\theta_0)}{\tau(\theta)}\right). \tag{8.21}$$

Applying approximation (8.21) to the binomial case of Example 8.10, we obtain 0.641 instead of 0.639 for the approximate power.

We may invert approximations (8.20) and (8.21) to obtain approximate sample sizes required to achieve desired power β against alternative θ. From (8.20) we obtain

$$\sqrt{n} \approx \frac{(u_\alpha - u_\beta)\tau(\theta_0)}{\mu(\theta) - \mu(\theta_0)} \tag{8.22}$$

and from (8.21) we obtain

$$\sqrt{n} \approx \frac{u_\alpha\tau(\theta_0) - u_\beta\tau(\theta)}{\mu(\theta) - \mu(\theta_0)}. \tag{8.23}$$

We may compare tests by considering the relative sample sizes necessary to achieve the same power at the same level against the same alternative.

Definition 8.11 Given tests 1 and 2 of the same hypotheses with asymptotic level α and a sequence of alternatives $\{\theta_k\}$, suppose that

$$\beta^{(1)}_{m_k}(\theta_k) \to \beta \quad \text{and} \quad \beta^{(2)}_{n_k}(\theta_k) \to \beta$$

as $k \to \infty$ for some sequences $\{m_k\}$ and $\{n_k\}$ of sample sizes. Then the asymptotic relative efficiency (ARE) of test 1 with respect to test 2 is

$$e_{1,2} = \lim_{k\to\infty}\frac{n_k}{m_k},$$

assuming this limit exists.

In Examples 8.12 and 8.13, we consider two different tests for the same hypotheses. Then, in Example 8.15, we compute their asymptotic relative efficiency.

Example 8.12 Suppose we have paired data $(X_1, Y_1), \ldots, (X_n, Y_n)$. Let $Z_i = Y_i - X_i$ for all i. Assume that the Z_i are independent and identically distributed with distribution function $P(Z_i \le z) = F(z - \theta)$ for some θ, where $f(z) = F'(z)$ exists and is symmetric about 0. Let $W_1, \ldots, W_n$ be a permutation of $Z_1, \ldots, Z_n$ such that $|W_1| \le |W_2| \le \cdots \le |W_n|$.

We wish to test $H_0 : \theta = 0$ against $H_1 : \theta > 0$. First, consider the Wilcoxon signed rank test. Define

$$R_n = \sum_{i=1}^{n} i I\{W_i > 0\}.$$

Then under H_0, the $I\{W_i > 0\}$ variables are independent Bernoulli(1/2) random variables. Thus,

$$\text{E } R_n = \sum_{i=1}^{n} \frac{i}{2} = \frac{n(n+1)}{4} \quad \text{and} \quad \text{Var } R_n = \sum_{i=1}^{n} \frac{i^2}{4} = \frac{n(n+1)(2n+1)}{24}.$$

Furthermore, one may prove that

$$\frac{R_n - \text{E } R_n}{\sqrt{\text{Var } R_n}} \xrightarrow{d} N(0,1)$$

under H_0 by verifying, say, the Lindeberg condition. Thus, a test with asymptotic level α rejects H_0 when

$$R_n \ge \frac{n(n+1)}{4} + \frac{u_\alpha \tau(0)}{\sqrt{n}},$$

where $\tau(0) = n\sqrt{(n+1)(2n+1)/24}$. Furthermore, it is possible to prove that

$$\frac{\sqrt{n}[R_n - \mu(\theta_n)]}{\sqrt{\tau^2(\theta_n)}} \xrightarrow{d} N(0,1),$$

where $\mu(\theta_n)$ and $\tau^2(\theta_n)$ are the mean and n times the variance of R_n under the alternatives $\theta_n = \Delta/\sqrt{n}$, though we will not prove this fact here (it is not as simple as checking a Lindeberg condition because under $\theta_n > 0$, the $I\{W_i > 0\}$ variables are not quite independent). This means that it is possible to apply Theorem 8.9.

Now we must find E R_n under the alternative $\theta_n = \Delta/\sqrt{n}$. First, we note that since $|W_i| \le |W_j|$ for $i \le j$, $W_i + W_j > 0$ if and only if $W_j > 0$. Therefore, $\sum_{i=1}^{j} I\{W_i + W_j > 0\} = jI\{W_j > 0\}$ and so we may rewrite R_n in the form

$$R_n \;=\; \sum_{j=1}^{n}\sum_{i=1}^{j} I\{W_i + W_j > 0\} \;=\; \sum_{j=1}^{n}\sum_{i=1}^{j} I\{Z_i + Z_j > 0\}.$$

Therefore, we obtain

$$\mu(\theta_n) \;=\; \sum_{j=1}^{n}\sum_{i=1}^{j} P_{\theta_n}(Z_i + Z_j > 0) \;=\; nP_{\theta_n}(Z_1 > 0) + \binom{n}{2} P_{\theta_n}(Z_1 + Z_2 > 0).$$

Since $P_{\theta_n}(Z_1 > 0) = P_{\theta_n}(Z_1 - \theta_n > -\theta_n) = 1 - F(-\theta_n)$ and

$$\begin{aligned} P_{\theta_n}(Z_1 + Z_2 > 0) &= P_{\theta_n}\{(Z_1 - \theta_n) + (Z_2 - \theta_n) > -2\theta_n\} \\ &= \mathrm{E}\; P_{\theta_n}\{Z_1 - \theta_n > -2\theta_n - (Z_2 - \theta_n) \mid Z_2\} \\ &= \int_{-\infty}^{\infty} \{1 - F(-2\theta_n - z)\} f(z)\, dz, \end{aligned}$$

we conclude that

$$\mu'(\theta) = nf(\theta) + \binom{n}{2} \int_{-\infty}^{\infty} 2f(-2\theta - z) f(z)\, dz.$$

Thus, because $f(-z) = f(z)$ by assumption,

$$\mu'(0) = nf(0) + n(n-1) \int_{-\infty}^{\infty} f^2(z)\, dz.$$

Letting

$$K = \int_{-\infty}^{\infty} f^2(z)\, dz,$$

we obtain

$$\lim_{n\to\infty} \frac{\mu'(0)}{\tau(0)} \;=\; \lim_{n\to\infty} \frac{\sqrt{24}\{f(0) + (n-1)K\}}{\sqrt{(n+1)(2n+1)}} \;=\; K\sqrt{12}.$$

Therefore, Theorem 8.9 shows that

$$\beta_n(\theta_n) \to \Phi(\Delta K\sqrt{12} - u_\alpha). \tag{8.24}$$

Example 8.13 As in Example 8.12, suppose we have paired data $(X_1, Y_1), \ldots, (X_n, Y_n)$ and $Z_i = Y_i - X_i$ for all i. The Z_i are independent and identically distributed with distribution function $P(Z_i \leq z) = F(z - \theta)$ for some θ, where $f(z) = F'(z)$ exists and is symmetric about 0. Suppose that the variance of Z_i is σ^2.

Since the t-test (unknown variance) and z-test (known variance) have the same asymptotic properties, let's consider the z-test for simplicity. Then $\tau(\theta) = \sigma$ for all θ. The relevant statistic is merely $\overline{Z}_n$, and the central limit theorem implies $\sqrt{n}\overline{Z}_n/\sigma \xrightarrow{d} N(0,1)$ under the null hypothesis. Therefore, the z-test in this case rejects $H_0 : \theta = 0$ in favor of $H_1 : \theta > 0$ whenever $\overline{Z}_n > u_\alpha \sigma/\sqrt{n}$. A check of the Lindeberg condition on the triangular array given by Z_1 under θ_1; Z_1, Z_2 under θ_2; Z_1, Z_2, Z_3 under θ_3; and so on shows that

$$\frac{\sqrt{n}(\overline{Z}_n - \theta_n)}{\sigma} \xrightarrow{d} N(0,1)$$

under the alternatives $\theta_n = \Delta/\sqrt{n}$. Therefore, by Theorem 8.9, we obtain

$$\beta_n(\theta_n) \to \Phi(\Delta/\sigma - u_\alpha) \tag{8.25}$$

since $\mu'(\theta) = 1$ for all θ.

Before finding the asymptotic relative efficiency (ARE) of the Wilcoxon signed rank test and the t-test, we prove a lemma that enables this calculation.

Suppose that for two tests, called test 1 and test 2, we use sample sizes m and n, respectively. We want m and n to tend to infinity together, an idea we make explicit by setting $m = m_k$ and $n = n_k$ for $k = 1, 2, \ldots$. Suppose that we wish to apply both tests to the same sequence of alternative hypotheses $\theta_1, \theta_2, \ldots$. As usual, we make the assumption that $(\theta_k - \theta_0)$ times the square root of the sample size tends to a finite, nonzero limit as $k \to \infty$. Thus, we assume

$$\sqrt{m_k}(\theta_k - \theta_0) \to \Delta_1 \quad \text{and} \quad \sqrt{n_k}(\theta_k - \theta_0) \to \Delta_2.$$

Then if Theorem 8.9 may be applied to both tests, define $c_1 = \lim \mu_1'(\theta_0)/\tau_1(\theta_0)$ and $c_2 = \lim \mu_2'(\theta_0)/\tau_2(\theta_0)$. The theorem says that

$$\beta_{m_k}(\theta_k) \to \Phi\left\{\Delta_1 c_1 - u_\alpha\right\} \quad \text{and} \quad \beta_{n_k}(\theta_k) \to \Phi\left\{\Delta_2 c_2 - u_\alpha\right\}. \tag{8.26}$$

To find the ARE, then, Definition 8.11 specifies that we assume that the two limits in (8.26) are the same, which implies $\Delta_1 c_1 = \Delta_2 c_2$, or

$$\frac{n_k}{m_k} \to \frac{c_1^2}{c_2^2}.$$

Thus, the ARE of test 1 with respect to test 2 equals $(c_1/c_2)^2$. This result is summed up in the following lemma, which defines a new term, efficacy.

Lemma 8.14 For a test to which Theorem 8.9 applies, define the **efficacy** of the test to be

$$c = \lim_{n\to\infty} \mu'(\theta_0)/\tau(\theta_0). \tag{8.27}$$

Suppose that Theorem 8.9 applies to each of two tests, called test 1 and test 2. Then the ARE of test 1 with respect to test 2 equals $(c_1/c_2)^2$.

Example 8.15 Using the results of Examples 8.12 and 8.13, we conclude that the efficacies of the Wilcoxon signed rank test and the t-test are

$$\sqrt{12}\int_{-\infty}^{\infty} f^2(z)\,dz \quad \text{and} \quad \frac{1}{\sigma},$$

respectively. Thus, Lemma 8.14 implies that the ARE of the signed rank test to the t-test equals

$$12\sigma^2\left(\int_{-\infty}^{\infty} f^2(z)\,dz\right)^2.$$

In the case of normally distributed data, we may verify without too much difficulty that the integral above equals $(2\sigma\sqrt{\pi})^{-1}$, so the ARE is $3/\pi \approx 0.9549$. Notice how close this is to one, suggesting that for normal data, we lose very little efficiency by using a signed rank test instead of a t-test. In fact, it may be shown that this asymptotic relative efficiency has a lower bound of 0.864. However, there is no upper bound on the ARE in this case, which means that examples exist for which the t-test is arbitrarily inefficient compared to the signed rank test.

Exercises for Section 8.2

Exercise 8.3 For the hypotheses considered in Examples 8.12 and 8.13, the sign test is based on the statistic $N_+ = \#\{i : Z_i > 0\}$. Since $2\sqrt{n}(N_+/n - \frac{1}{2}) \xrightarrow{d} N(0,1)$ under the null hypothesis, the sign test (with continuity correction) rejects H_0 when

$$N_+ - \frac{1}{2} \geq \frac{u_\alpha\sqrt{n}}{2} + \frac{n}{2}.$$

(a) Find the efficacy of the sign test. Make sure to indicate how you go about verifying Equation (8.17).

(b) Find the ARE of the sign test with respect to the signed rank test and the t-test. Evaluate each of these for the case of normal data.

Exercise 8.4 Suppose $X_1, \ldots, X_n$ are a simple random sample from a uniform $(0, 2\theta)$ distribution. We wish to test $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$ at $\alpha = .05$.

Define Q_1 and Q_3 to be the first and third quartiles of the sample. Consider test A, which rejects when

$$Q_3 - Q_1 - \theta_0 \quad \geq \quad A_n,$$

and test B, which rejects when

$$\overline{X} - \theta_0 \quad \geq \quad B_n.$$

Based on the asymptotic distribution of $\overline{X}$ and the joint asymptotic distribution of (Q_1, Q_3), find the values of A_n and B_n that correspond with the test in (8.18). Then find the asymptotic relative efficiency of test A relative to test B.

Exercise 8.5 Let P_θ be a family of probability distributions indexed by a real parameter θ. If $X \sim P_\theta$, define $\mu(\theta) = \mathrm{E}\,(X)$ and $\sigma^2(\theta) = \mathrm{Var}\,(X)$. Now let $\theta_1, \theta_2, \ldots$ be a sequence of parameter values such that $\theta_n \to \theta_0$ as $n \to \infty$. Suppose that $\mathrm{E}_{\,\theta_n} |X|^{2+\delta} < M$ for all n for some positive δ and M. Also suppose that for each n, $X_{n1}, \ldots, X_{nn}$ are independent with distribution P_{θ_n} and define $\overline{X}_n = \sum_{i=1}^n X_{ni}/n$. Prove that if $\sigma^2(\theta_0) < \infty$ and $\sigma^2(\theta)$ is continuous at the point θ_0, then

$$\sqrt{n}[\overline{X}_n - \mu(\theta_n)] \quad \xrightarrow{d} \quad N(0, \sigma^2(\theta_0))$$

as $n \to \infty$.

Hint:

$$|a + b|^{2+\delta} \leq 2^{2+\delta} \left(|a|^{2+\delta} + |b|^{2+\delta} \right).$$

Exercise 8.6 Suppose $X_1, X_2, \ldots$ are independent exponential random variables with mean θ. Consider the test of $H_0 : \theta = 1$ vs $H_1 : \theta > 1$ in which we reject H_0 when

$$\overline{X}_n \quad \geq \quad 1 + \frac{u_\alpha}{\sqrt{n}},$$

where $\alpha = .05$.

(a) Derive an asymptotic approximation to the power of the test for a fixed sample size n and alternative θ. Tell where you use the result of Problem 8.5.

(b) Because the sum of independent exponential random variables is a gamma random variable, it is possible to compute the power exactly in this case. Create a table in which you compare the exact power of the test against the alternative $\theta = 1.2$ to the asymptotic approximation in part (a) for $n \in \{5, 10, 15, 20\}$.

Exercise 8.7 Let $X_1, \ldots, X_n$ be independent from Poisson (λ). Create a table in which you list the exact power along with approximations (8.20) and (8.21) for the test that rejects $H_0 : \lambda = 1$ in favor of $H_1 : \lambda > 1$ when

$$\frac{\sqrt{n}(\overline{X}_n - \lambda_0)}{\sqrt{\lambda_0}} \geq u_\alpha,$$

where $n = 20$ and $\alpha = .05$, against each of the alternatives 1.1, 1.5, and 2.

Exercise 8.8 Let $X_1, ..., X_n$ be an independent sample from an exponential distribution with mean λ, and $Y_1, ..., Y_n$ be an independent sample from an exponential distribution with mean μ. Assume that X_i and Y_i are independent. We are interested in testing the hypothesis $H_0 : \lambda = \mu$ versus $H_1 : \lambda > \mu$. Consider the statistic

$$T_n = 2\sum_{i=1}^{n}(I_i - 1/2)/\sqrt{n},$$

where I_i is the indicator variable $I_i = I(X_i > Y_i)$.

(a) Derive the asymptotic distribution of T_n under the null hypothesis.

(b) Use the Lindeberg Theorem to show that, under the local alternative hypothesis $(\lambda_n, \mu_n) = (\lambda + n^{-1/2}\delta, \lambda)$, where $\delta > 0$,

$$\frac{\sum_{i=1}^{n}(I_i - \rho_n)}{\sqrt{n\rho_n(1-\rho_n)}} \xrightarrow{\mathcal{L}} N(0,1), \quad \text{where} \quad \rho_n = \frac{\lambda_n}{\lambda_n + \mu_n} = \frac{\lambda + n^{-1/2}\delta}{2\lambda + n^{-1/2}\delta}.$$

(c) Using the conclusion of part (b), derive the asymptotic distribution of T_n under the local alternative specified in (b).

Exercise 8.9 Suppose $X_1, \ldots, X_m$ is a simple random sample and $Y_1, \ldots, Y_n$ is another simple random sample independent of the X_i, with $P(X_i \leq t) = t^2$ for $t \in [0, 1]$ and $P(Y_i \leq t) = (t - \theta)^2$ for $t \in [\theta, \theta + 1]$. Assume $m/(m + n) \to \rho$ as $m, n \to \infty$ and $0 < \theta < 1$.

Find the asymptotic distribution of $\sqrt{m+n}[g(\overline{Y} - \overline{X}) - g(\theta)]$.

8.3 The Wilcoxon Rank-Sum Test

Suppose that $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ are two independent simple random samples, with

$$P(X_i \leq t) = P(Y_j \leq t + \theta) = F(t) \tag{8.28}$$

for some continuous distribution function $F(t)$ with $f(t) = F'(t)$. Thus, the distribution of the Y_j is shifted by θ from the distribution of the X_i. We wish to test $H_0 : \theta = 0$ against $H_1 : \theta > 0$.

To do the asymptotics here, we will assume that n and m are actually both elements of separate sequences of sample sizes, indexed by a third variable, say k. Thus, $m = m_k$ and $n = n_k$ both go to ∞ as $k \to \infty$, and we suppress the subscript k on m and n for convenience of notation. Suppose that we combine the X_i and Y_j into a single sample of size $m + n$. Define the Wilcoxon rank-sum statistic to be

$$S_k = \sum_{j=1}^{n} \text{Rank of } Y_j \text{ among combined sample.}$$

Letting $Y_{(1)}, \ldots, Y_{(n)}$ denote the order statistics for the sample of Y_j as usual, we may rewrite S_k in the following way:

$$\begin{aligned} S_k &= \sum_{j=1}^{n} \text{Rank of } Y_{(j)} \text{ among combined sample} \\ &= \sum_{j=1}^{n} \left(j + \#\{i : X_i < Y_{(j)}\} \right) \\ &= \frac{n(n+1)}{2} + \sum_{j=1}^{n} \sum_{i=1}^{m} I\{X_i < Y_{(j)}\} \\ &= \frac{n(n+1)}{2} + \sum_{j=1}^{n} \sum_{i=1}^{m} I\{X_i < Y_j\}. \end{aligned} \tag{8.29}$$

Let $N = m + n$, and suppose that $m/N \to \rho$ as $k \to \infty$ for some constant $\rho \in (0, 1)$. First, we will establish the asymptotic behavior of S_k under the null hypothesis. Define

$$\mu(\theta) = \mathrm{E}_\theta \, S_k \quad \text{and} \quad \tau(\theta) = \sqrt{N \operatorname{Var} S_k}.$$

To evaluate $\mu(\theta_0)$ and $\tau(\theta_0)$, where $\theta_0 = 0$, let $Z_i = \sum_{j=1}^{n} I\{X_i < Y_j\}$. Then the Z_i are identically distributed but not independent, and we have $\mathrm{E}_{\theta_0} Z_i = n/2$ and

$$\begin{aligned} \operatorname{Var}_{\theta_0} Z_i &= \frac{n}{4} + n(n-1) \operatorname{Cov}_{\theta_0} \left(I\{X_i < Y_1\}, I\{X_i < Y_2\} \right) \\ &= \frac{n}{4} + \frac{n(n-1)}{3} - \frac{n(n-1)}{4} \\ &= \frac{n(n+2)}{12}. \end{aligned}$$

Furthermore,

$$\mathrm{E}_{\theta_0} Z_i Z_j = \sum_{r=1}^{n}\sum_{s=1}^{n} P_{\theta_0}(X_i < Y_r \text{ and } X_j < Y_s) = \frac{n(n-1)}{4} + nP_{\theta_0}(X_i < Y_1 \text{ and } X_j < Y_1),$$

which implies

$$\mathrm{Cov}_{\theta_0}(Z_i, Z_j) = \frac{n(n-1)}{4} + \frac{n}{3} - \frac{n^2}{4} = \frac{n}{12}.$$

Therefore,

$$\mu(\theta_0) = \frac{n(n+1)}{2} + \frac{mn}{2} = \frac{n(N+1)}{2}$$

and

$$\begin{aligned} \tau^2(\theta_0) &= Nm\,\mathrm{Var}\; Z_1 + Nm(m-1)\,\mathrm{Cov}\,(Z_1, Z_2) \\ &= [Nmn(n+2) + Nm(m-1)n]\,/12 \\ &= [Nmn(N+1)]\,/12. \end{aligned}$$

Let $\theta_1, \theta_2, \ldots$ be a sequence of alternatives such that $\sqrt{N}(\theta_k - \theta_0) \to \Delta$ for a positive, finite constant Δ. It is possible to show that

$$\frac{\sqrt{N}\{S_k - \mu(\theta_0)\}}{\tau(\theta_0)} \xrightarrow{d} N(0,1) \tag{8.30}$$

under H_0 (see Exercise 8.10) and

$$\frac{\sqrt{N}\{S_k - \mu(\theta_k)\}}{\tau(\theta_k)} \xrightarrow{d} N(0,1) \tag{8.31}$$

under the alternatives $\{\theta_k\}$. As in the case of the signed-rank test of the previous section, we will not prove the asymptotic normality under the sequence of alternatives here because it is not a direct consequence of any of the results we have seen thus far. Yet it may be proven using the Hoeffding projection idea described in Chapter 10, by which S_k may be expressed as a sum of independent random variables plus an asymptotically negligible term.

By expression (8.30), the test based on S_k with asymptotic level α rejects $H_0 : \theta = 0$ in favor of $H_a : \theta > 0$ whenever $S_k \geq \mu(\theta_0) + u_\alpha \tau(\theta_0)/\sqrt{N}$, or

$$S_k \geq \frac{n(N+1)}{2} + u_\alpha \sqrt{\frac{mn(N+1)}{12}}.$$

The Wilcoxon rank-sum test is sometimes referred to as the Mann-Whitney test. This alternative name helps to distinguish this test from the similarly named Wilcoxon signed rank test.

To find the limiting power of the rank-sum test, we may use Theorem 8.9 to conclude that

$$\beta_k(\theta_k) \to \lim_{k\to\infty} \Phi\left(\frac{\Delta\mu'(\theta_0)}{\tau(\theta_0)} - u_\alpha\right). \tag{8.32}$$

According to expression (8.32), we must evaluate $\mu'(\theta_0)$. To this end, note that

$$\begin{aligned} P_\theta(X_1 < Y_1) &= \mathrm{E}_\theta\left\{P_\theta(X_1 < Y_1 \mid Y_1)\right\} \\ &= \mathrm{E}_\theta F(Y_1) \\ &= \int_{-\infty}^{\infty} F(y) f(y-\theta)\, dy \\ &= \int_{-\infty}^{\infty} F(y+\theta) f(y)\, dy. \end{aligned}$$

Therefore,

$$\frac{d}{d\theta} P_\theta(X_1 < Y_1) = \int_{-\infty}^{\infty} f(y+\theta) f(y)\, dy.$$

This gives

$$\mu'(0) = mn \int_{-\infty}^{\infty} f^2(y)\, dy.$$

Thus, the efficacy of the Wilcoxon rank-sum test is

$$\lim_{k\to\infty} \frac{\mu'(\theta_0)}{\tau(\theta_0)} = \lim_{k\to\infty} \frac{mn\sqrt{12}\int_{-\infty}^{\infty} f^2(y)\, dy}{\sqrt{mnN(N+1)}} = \sqrt{12\rho(1-\rho)} \int_{-\infty}^{\infty} f^2(y)\, dy.$$

The asymptotic power of the test follows immediately from (8.32).

Exercises for Section 8.3

Exercise 8.10 Prove expression (8.30) under the null hypothesis $H_0 : \theta = 0$.

Hint: Verify either the Lindeberg condition or the Lyapunov condition.

Exercise 8.11 Suppose Var X_i = Var $Y_i = \sigma^2 < \infty$ and we wish to test the hypotheses $H_0 : \theta = 0$ vs. $H_1 : \theta > 0$ using the two-sample Z-statistic

$$\frac{\overline{Y} - \overline{X}}{\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}}.$$

Note that this Z-statistic is s/σ times the usual T-statistic, where s is the pooled sample standard deviation, so the asymptotic properties of the T-statistic are the same as those of the Z-statistic.

(a) Find the efficacy of the Z test. Justify your use of Theorem 8.9.

(b) Find the ARE of the Z test with respect to the rank-sum test for normally distributed data.

(c) Find the ARE of the Z test with respect to the rank-sum test if the data come from a double exponential distribution with $f(t) = \frac{1}{2\lambda}e^{-|t/\lambda|}$.

(d) Prove by example that the ARE of the Z-test with respect to the rank-sum test can be arbitrarily close to zero.

Chapter 9

Pearson's chi-square test

9.1 Null hypothesis asymptotics

Let $\mathbf{X}_1, \mathbf{X}_2, \cdots$ be independent from a multinomial$(1, \mathbf{p})$ distribution, where $\mathbf{p}$ is a k-vector with nonnegative entries that sum to one. That is,

$$P(X_{ij} = 1) = 1 - P(X_{ij} = 0) = p_j \text{ for all } 1 \leq j \leq k \tag{9.1}$$

and each $\mathbf{X}_i$ consists of exactly $k-1$ zeros and a single one, where the one is in the component of the "success" category at trial i. Note that the multinomial distribution is a generalization of the binomial distribution to the case in which there are k categories of outcome instead of only 2.

The purpose of this section is to derive the asymptotic distribution of the Pearson chi-square statistic

$$\chi^2 = \sum_{j=1}^{k} \frac{(n_j - np_j)^2}{np_j}, \tag{9.2}$$

where n_j is the random variable $n\overline{X}_j$, the number of successes in the jth category for trials $1, \ldots, n$. In a real application, the true value of $\mathbf{p}$ is not known, but instead we assume that $\mathbf{p} = \mathbf{p}^0$ for some null value $\mathbf{p}^0$. We will show that χ^2 converges in distribution to the chi-square distribution on $k-1$ degrees of freedom, which yields to the familiar chi-square test of goodness of fit for a multinomial distribution.

Equation (9.1) implies that Var $X_{ij} = p_j(1-p_j)$. Furthermore, Cov $(X_{ij}, X_{i\ell}) =$ E $X_{ij}X_{i\ell} -$

$p_j p_\ell = -p_j p_\ell$ for $j \neq \ell$. Therefore, the random vector $\mathbf{X}_i$ has covariance matrix

$$\Sigma = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_k \\ -p_1p_2 & p_2(1-p_2) & \cdots & -p_2p_k \\ \vdots & & \ddots & \vdots \\ -p_1p_k & -p_2p_k & \cdots & p_k(1-p_k) \end{pmatrix}. \tag{9.3}$$

Since $\mathrm{E}\ \mathbf{X}_i = \mathbf{p}$, the central limit theorem implies

$$\sqrt{n}\left(\overline{\mathbf{X}}_n - \mathbf{p}\right) \xrightarrow{d} N_k(\mathbf{0}, \Sigma). \tag{9.4}$$

Note that the sum of the jth column of Σ is $p_j - p_j(p_1 + \cdots + p_k) = 0$, which is to say that the sum of the rows of Σ is the zero vector, so Σ is not invertible.

We now present two distinct derivations of this asymptotic distribution of the χ^2 statistic in equation (9.2), because each derivation is instructive. One derivation avoids dealing with the singular matrix Σ, whereas the other does not.

In the first approach, define for each i $\mathbf{Y}_i = (X_{i1}, \ldots, X_{i,k-1})$. That is, let $\mathbf{Y}_i$ be the $k-1$-vector consisting of the first $k-1$ components of $\mathbf{X}_i$. Then the covariance matrix of $\mathbf{Y}_i$ is the upper-left $(k-1) \times (k-1)$ submatrix of Σ, which we denote by Σ^*. Similarly, let $\mathbf{p}^*$ denote the vector $(p_1, \ldots, p_{k-1})$.

One may verify that Σ^* is invertible and that

$$(\Sigma^*)^{-1} = \begin{pmatrix} \frac{1}{p_1} + \frac{1}{p_k} & \frac{1}{p_k} & \cdots & \frac{1}{p_k} \\ \frac{1}{p_k} & \frac{1}{p_2} + \frac{1}{p_k} & \cdots & \frac{1}{p_k} \\ \vdots & & \ddots & \vdots \\ \frac{1}{p_k} & \frac{1}{p_k} & \cdots & \frac{1}{p_{k-1}} + \frac{1}{p_k} \end{pmatrix}. \tag{9.5}$$

Furthermore, the χ^2 statistic of equation (9.2) by be rewritten as

$$\chi^2 = n(\overline{\mathbf{Y}} - \mathbf{p}^*)^\top (\Sigma^*)^{-1} (\overline{\mathbf{Y}} - \mathbf{p}^*). \tag{9.6}$$

The facts in Equations (9.5) and (9.6) are checked in Problem 9.2. If we now define

$$\mathbf{Z}_n = \sqrt{n}(\Sigma^*)^{-1/2}(\overline{\mathbf{Y}} - \mathbf{p}^*),$$

then the central limit theorem implies $\mathbf{Z}_n \xrightarrow{d} N_{k-1}(\mathbf{0}, I)$. By definition, the χ^2_{k-1} distribution is the distribution of the sum of the squares of $k-1$ independent standard normal random variables. Therefore,

$$\chi^2 = (\mathbf{Z}_n)^\top \mathbf{Z}_n \xrightarrow{d} \chi^2_{k-1}, \tag{9.7}$$

which is the result that leads to the familiar chi-square test.

In a second approach to deriving the limiting distribution (9.7), we use some properties of projection matrices.

Definition 9.1 A symmetric matrix P is called a projection matrix if it is idempotent; that is, if $P^2 = P$.

The following lemmas, to be proven in Problem 9.3, give some basic facts about projection matrices.

Lemma 9.2 Suppose P is a projection matrix. Then every eigenvalue of P equals 0 or 1. Suppose that r denotes the number of eigenvalues of P equal to 1. Then if $\mathbf{Z} \sim N_k(\mathbf{0}, P)$, $\mathbf{Z}^\top \mathbf{Z} \sim \chi^2_r$.

Lemma 9.3 The trace of a square matrix M, $\mathrm{Tr}\,(M)$, is equal to the sum of its diagonal entries. For matrices A and B whose sizes allow them to be multiplied in either order, $\mathrm{Tr}\,(AB) = \mathrm{Tr}\,(BA)$.

Recall (Lemma 4.8) that if a square matrix M is symmetric, then there exists an orthogonal matrix Q such that $QMQ^\top$ is a diagonal matrix whose entries consist of the eigenvalues of M. By Lemma 9.3, $\mathrm{Tr}\,(QMQ^\top) = \mathrm{Tr}\,(Q^\top QM) = \mathrm{Tr}\,(M)$, which proves yet another lemma:

Lemma 9.4 If M is symmetric, then $\mathrm{Tr}\,(M)$ equals the sum of the eigenvalues of M.

Define $\Gamma = \mathrm{diag}\,(\mathbf{p})$, and let Σ be defined as in Equation (9.3). Equation (9.4) implies

$$\sqrt{n}\Gamma^{-1/2}(\overline{\mathbf{X}} - \mathbf{p}) \xrightarrow{d} N_k(\mathbf{0}, \Gamma^{-1/2}\Sigma\Gamma^{-1/2}).$$

Since Σ may be written in the form $\Gamma - \mathbf{p}\mathbf{p}^\top$,

$$\Gamma^{-1/2}\Sigma\Gamma^{-1/2} \;=\; I - \Gamma^{-1/2}\mathbf{p}\mathbf{p}^\top\Gamma^{-1/2} \;=\; I - \sqrt{\mathbf{p}}\sqrt{\mathbf{p}}^\top \tag{9.8}$$

has trace $k-1$; furthermore,

$$(I - \sqrt{\mathbf{p}}\sqrt{\mathbf{p}}^\top)(I - \sqrt{\mathbf{p}}\sqrt{\mathbf{p}}^\top) \;=\; I - 2\sqrt{\mathbf{p}}\sqrt{\mathbf{p}}^\top + \sqrt{\mathbf{p}}\sqrt{\mathbf{p}}^\top\sqrt{\mathbf{p}}\sqrt{\mathbf{p}}^\top \;=\; I - \sqrt{\mathbf{p}}\sqrt{\mathbf{p}}^\top$$

because $\sqrt{\mathbf{p}}^\top\sqrt{\mathbf{p}} = 1$, so the covariance matrix (9.8) is a projection matrix.

Define $\mathbf{A}_n = \sqrt{n}\Gamma^{-1/2}(\overline{\mathbf{X}} - \mathbf{p})$. Then we may check (in problem 9.3) that

$$\chi^2 = (\mathbf{A}_n)^\top \mathbf{A}_n. \tag{9.9}$$

Therefore, since the covariance matrix (9.8) is a projection with trace $k-1$, Lemma 9.4 and Lemma 9.2 prove that $\chi^2 \xrightarrow{d} \chi^2_{k-1}$ as desired.

Exercises for Section 9.1

Exercise 9.1 *Hotelling's T^2.* Suppose $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots$ are independent and identically distributed from some k-dimensional distribution with mean $\boldsymbol{\mu}$ and finite nonsingular covariance matrix Σ. Let S_n denote the sample covariance matrix

$$S_n \;=\; \frac{1}{n-1}\sum_{j=1}^{n}(\mathbf{X}^{(j)} - \overline{\mathbf{X}})(\mathbf{X}^{(j)} - \overline{\mathbf{X}})^\top.$$

To test $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}^0$ against $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}^0$, define the statistic

$$T^2 \;=\; (\mathbf{V}^{(n)})^\top S_n^{-1}(\mathbf{V}^{(n)}),$$

where $\mathbf{V}^{(n)} = \sqrt{n}(\overline{\mathbf{X}} - \boldsymbol{\mu}^0)$. This is called Hotelling's T^2 statistic.

[Notes: This is a generalization of the square of a unidimensional t-statistic. If the sample is multivariate normal, then $[(n-k)/(nk-k)]T^2$ is distributed as $F_{k,n-k}$. A Pearson chi square statistic may be shown to be a special case of Hotelling's T^2.]

(a) You may assume that $S_n^{-1} \xrightarrow{P} \Sigma^{-1}$, which follows from the Weak Law of Large Numbers since $P(S_n \text{ is nonsingular}) \to 1$. Prove that under the null hypothesis, $T^2 \xrightarrow{d} \chi^2_k$.

(b) An approximate $1-\alpha$ confidence set for $\boldsymbol{\mu}$ based on the result in part (a) may be formed by plotting the elliptical set

$$\{\boldsymbol{\mu} : n(\overline{\mathbf{X}} - \boldsymbol{\mu})^\top S_n^{-1}(\overline{\mathbf{X}} - \boldsymbol{\mu}) = c_\alpha\},$$

where c_α is defined by the equation $\mathrm{t}P(\chi^2_k > c_\alpha) = \alpha$. For a random sample of size 100 from $N_2(\mathbf{0}, \Sigma)$, where

$$\Sigma = \begin{pmatrix} 1 & 3/5 \\ 3/5 & 1 \end{pmatrix},$$

produce a scatterplot of the sample and plot 90% and 99% confidence sets on this scatterplot.

Hints: In part (b), to produce a random vector with the $N_2(\mathbf{0}, \Sigma)$ distribution, take a $N_2(\mathbf{0}, I)$ random vector and left-multiply by a matrix A such that $AA^\top = \Sigma$. It is not hard to find such an A (it may be taken to be lower triangular). One way to graph the ellipse is to find a matrix B such that $B^\top S_n^{-1} B = I$. Then note that

$$\{\boldsymbol{\mu} : n(\overline{\mathbf{X}} - \boldsymbol{\mu})^\top S_n^{-1}(\overline{\mathbf{X}} - \boldsymbol{\mu}) = c_\alpha\} \;=\; \{\overline{\mathbf{X}} - B\boldsymbol{\nu} : \boldsymbol{\nu}^\top\boldsymbol{\nu} = c_\alpha/n\},$$

so it remains only to find points $\boldsymbol{\nu}$, closely spaced, such that $\boldsymbol{\nu}^\top\boldsymbol{\nu}$ equals a constant. To find a matrix B such as the one specified, note that the matrix of eigenvectors of S_n, properly normalized, gives an orthogonal matrix that diagonalizes.

Exercise 9.2 Verify Equations (9.5) and (9.6).

Exercise 9.3 Prove Lemma 9.2 and Lemma 9.3, then verify Equation (9.9).

Exercise 9.4 Pearson's chi-square for a 2-way table: Product multinomial model. If A and B are categorical variables with 2 and k levels, respectively, and we collect random samples of size m and n from levels 1 and 2 of A, then classify each individual according to its level of the variable B, the results of this study may be summarized in a $2 \times k$ table. The standard test of the independence of variables A and B is the Pearson chi-square test, which may be written as

$$\sum_{\text{all cells in table}} \frac{(O_j - E_j)^2}{E_j},$$

where O_j is the observed count in cell j and E_j is the estimate of the expected count under the null hypothesis. Equivalently, we may set up the problem as follows: If $\mathbf{X}$ and $\mathbf{Y}$ are independent Multinomial$(m, \mathbf{p})$ and Multinomial $(n, \mathbf{p})$ random vectors, respectively, then the Pearson chi-square statistic is

$$\mathrm{W}^2 \;=\; \sum_{j=1}^{k} \left\{ \frac{(X_j - mZ_j/N)^2}{mZ_j/N} + \frac{(Y_j - nZ_j/N)^2}{nZ_j/N} \right\},$$

where $\mathbf{Z} = \mathbf{X} + \mathbf{Y}$ and $N = n + m$. (Note: I used W^2 to denote the chi-square statistic to avoid using yet another variable that looks like an X.)

Prove that if $N \to \infty$ in such a way that $n/N \to \alpha \in (0, 1)$, then

$$\mathrm{W}^2 \;\xrightarrow{d}\; \chi^2_{k-1}.$$

Exercise 9.5 Pearson's chi-square for a 2-way table: Multinomial model. Now consider the case in which $(\mathbf{X}, \mathbf{Y})$ is a single multinomial $(N, \mathbf{q})$ random $2k$-vector. X_i will still denote the $(1, i)$ entry in a $2 \times k$ table, and Y_i will still denote the $(2, i)$ entry.

(a) In this case, $\mathbf{q}$ is a $2k$-vector. Let $\alpha = q_1/(q_1 + q_{k+1})$ and define $\mathbf{p}$ to be the k-vector such that $(q_1, \ldots, q_k) = \alpha\mathbf{p}$. Prove that under the usual null hypothesis that variable A is independent of variable B (i.e., the row variable and the column variable are independent), $\mathbf{q} = (\alpha\mathbf{p}, (1 - \alpha)\mathbf{p})$ and $p_1 + \cdots + p_k = 1$.

(b) As in Problem 9.4, let $\mathbf{Z} = \mathbf{X} + \mathbf{Y}$. Assume the null hypothesis is true and suppose that for some reason α is known. The Pearson chi-square statistic may be written as

$$\mathrm{W}^2 = \sum_{j=1}^{k} \left\{ \frac{(X_j - \alpha Z_j)^2}{\alpha Z_j} + \frac{(Y_j - (1-\alpha)Z_j)^2}{(1-\alpha)Z_j} \right\}. \tag{9.10}$$

Find the joint asymptotic distribution of

$$\sqrt{N\alpha(1-\alpha)} \left(\frac{X_1}{N\alpha} - \frac{Y_1}{N(1-\alpha)}, \ldots, \frac{X_k}{N\alpha} - \frac{Y_k}{n(1-\alpha)} \right)$$

and use this result to prove that $\mathrm{W}^2 \xrightarrow{d} \chi^2_k$.

Exercise 9.6 In Problem 9.5(b), it was assumed that α was known. However, in most problems this assumption is unrealistic. Therefore, we replace all occurrences of α in Equation (9.10) by $\hat{\alpha} = \sum_{i=1}^{k} X_i/N$. This results in a different asymptotic distribution for the W^2 statistic. Suppose we are given the following multinomial probabilities for a 2×2 table with independent row and column variables:

$P(X_1 = 1) = .1$	$P(X_2 = 1) = .15$	.25
$P(Y_1 = 1) = .3$	$P(Y_2 = 1) = .45$	.75
.4	.6	1

Note that $\alpha = .25$ in the above table. Let $N = 50$ and simulate 1000 multinomial random vectors with the above probabilities. For each, calculate the value of W^2 using both the known value $\alpha = .25$ and the value $\hat{\alpha}$ estimated from the data. Plot the empirical distribution function of each of these two sets of 1000 values. Compare with the theoretical distribution functions for the χ^2_1 and χ^2_2 distributions.

Hint: To generate a multinomial random variable with expectation vector matching the table above, because of the independence inherent in the table you can generate two independent Bernoulli random variables with respective success probabilities equal to the margins: That is, let $P(A = 2) = 1 - P(A = 1) = .6$ and $P(B = 2) = 1 - P(B = 1) = .75$, then classify the multinomial observation into the correct cell based on the random values of A and B.

Exercise 9.7 The following example comes from genetics. There is a particular characteristic of human blood (the so-called MN blood group) that has three types: M, MN, and N. Under idealized circumstances known as Hardy-Weinberg equilibrium, these three types occur in the population with probabilities $p_1 = \pi^2_M$,

$p_2 = 2\pi_M\pi_N$, and $p_3 = \pi_N^2$, respectively, where π_M is the frequency of the M allele in the population and $\pi_N = 1 - \pi_M$ is the frequency of the N allele.

We observe data $\mathbf{X}_1, \ldots, \mathbf{X}_n$, where $\mathbf{X}_i$ has one of three possible values: $(1,0,0)^T$, $(0,1,0)^T$, or $(0,0,1)^T$, depending on whether the ith individual has the M, MN, or N blood type. Denote the total number of individuals of each of the three types by n_1, n_2, and n_3; in other words, $n_j = n\overline{X}_j$ for each j.

If the value of π_M were known, then the results of this section would show that the Pearson χ^2 statistic converges in distribution to a chi-square distribution on 2 degrees of freedom. However, of course we usually don't know π_M. Instead, we estimate it using the maximum likelihood estimator $\hat{\pi}_M = (2n_1 + n_2)/2n$. By the invariance principle of maximum likelihood estimation, this gives $\hat{\mathbf{p}} = (\hat{\pi}_M^2, 2\hat{\pi}_M\hat{\pi}_N, \hat{\pi}_N^2)^T$ as the maximum likelihood estimator of $\mathbf{p}$.

(a) Define $\mathbf{B}_n = \sqrt{n}(\overline{\mathbf{X}} - \hat{\mathbf{p}})$. Use the delta method to derive the asymptotic distribution of $\Gamma^{-1/2}\mathbf{B}_n$, where $\Gamma = \text{diag}(p_1, p_2, p_3)$.

(b) Define $\hat{\Gamma}$ to be the diagonal matrix with entries $\hat{p}_1, \hat{p}_2, \hat{p}_3$ along its diagonal. Derive the asymptotic distribution of $\hat{\Gamma}^{-1/2}\mathbf{B}_n$.

(c) Derive the asymptotic distribution of the Pearson chi-square statistic

$$\chi^2 \;=\; \sum_{j=1}^{3} \frac{(n_j - n\hat{p}_j)^2}{n\hat{p}_j}. \tag{9.11}$$

Exercise 9.8 Take $\pi_M = .75$ and $n = 100$ in the situation described in Problem 9.7. Simulate 500 realizations of the data.

(a) Compute

$$\sum_{j=1}^{3} \frac{(n_j - np_j)^2}{np_j}$$

for each of your 500 datasets. Compare the empirical distribution function of these statistics with both the χ_1^2 and χ_2^2 distribution functions. Comment on what you observe.

(b) Compute the χ^2 statistic of Equation (9.11) for each of your 500 datasets. Compare the empirical distribution function of these statistics with both the χ_1^2 and χ_2^2 distribution functions. Comment on what you observe.

9.2 Power of Pearson's chi-square test

Suppose the k-vector $\mathbf{X}$ is distributed as multinomial $(n, \mathbf{p})$, and we wish to test the null hypothesis $H_0 : \mathbf{p} = \mathbf{p}^0$ against the alternative $H_1 : \mathbf{p} \neq \mathbf{p}^0$ using the Pearson chi-square test. We are given a sequence of specific alternatives $\mathbf{p}^{(n)}$ satisfying $\sqrt{n}(\mathbf{p}^{(n)} - \mathbf{p}) \to \boldsymbol{\delta}$ for some constant matrix $\boldsymbol{\delta}$. Note that this means $\sum_{i=1}^{k} \delta_i = 0$, a fact that will be used later. Our task is to derive the limit of the power of the sequence of tests under the sequence of alternatives $\mathbf{p}^{(n)}$.

The notion of a noncentral chi-square distribution will be important in this development, so we first give a definition.

Definition 9.5 If $A_1, \ldots, A_n$ are independent random variables with $A_i \sim N(\mu_i, 1)$, then the distribution of $A_1^2 + A_2^2 + \cdots + A_n^2$ is noncentral chi-square with n degrees of freedom and noncentrality parameter $\phi = \mu_1^2 + \cdots + \mu_n^2$. (In particular, the distribution depends on the μ_i only through ϕ.) We denote this distribution $\chi_n^2(\phi)$. Equivalently, we can say that if $\mathbf{A} \sim N_n(\boldsymbol{\mu}, I)$, then $\mathbf{A}^\top \mathbf{A} \sim \chi_n^2(\phi)$ where $\phi = \boldsymbol{\mu}^\top \boldsymbol{\mu}$.

In some references, the noncentrality parameter is defined to be equal to $\boldsymbol{\mu}^\top \boldsymbol{\mu}/2$. The form of the actual parameter is not important, though it is of course necessary to know in a particular context which parameterization is used.

Actually, Definition (9.5) is not a valid definition unless we may prove that the distribution of $\mathbf{A}^\top \mathbf{A}$ depends on $\boldsymbol{\mu}$ only through $\phi = \boldsymbol{\mu}^\top \boldsymbol{\mu}$. We prove this as follows. First, note that if $\phi = 0$ then there is nothing to prove. Otherwise, define $\boldsymbol{\mu}^* = \boldsymbol{\mu}/\sqrt{\phi}$. Next, find an orthogonal matrix Q whose first row is $(\boldsymbol{\mu}^*)^\top$. (It is always possible to do this, though we do not explain the details here. One method is the process of Gram-Schmidt orthogonalization). Then $Q\mathbf{A} \sim N_k(Q\boldsymbol{\mu}, I)$. Since $Q\boldsymbol{\mu}$ is a vector with first element $\sqrt{\phi}$ and remaining elements 0, $Q\mathbf{A}$ has a distribution that depends on $\boldsymbol{\mu}$ only through ϕ. But $\mathbf{A}^\top \mathbf{A} = (Q\mathbf{A})^\top (Q\mathbf{A})$, proving that the distribution of $\mathbf{A}^\top \mathbf{A}$ depends on the μ_i only through ϕ.

We will derive the power of the chi-square test by adapting the projection matrix technique of Section 9.1. First, we prove a lemma that generalizes Lemma 9.2.

Lemma 9.6 Suppose $\mathbf{Z} \sim N_k(\boldsymbol{\mu}, P)$, where P is a projection matrix of rank $r \leq k$ and $P\boldsymbol{\mu} = \boldsymbol{\mu}$. Then $\mathbf{Z}^\top \mathbf{Z} \sim \chi_r^2(\boldsymbol{\mu}^\top \boldsymbol{\mu})$.

Proof: Since P is a covariance matrix, it is symmetric, which means that there exists an orthogonal matrix Q with $QPQ^{-1} = \operatorname{diag}(\boldsymbol{\lambda})$, where $\boldsymbol{\lambda}$ is the vector of eigenvalues of P. Since P is a projection matrix, all of its eigenvalues are 0 or 1. Since P has rank r, exactly r of the eigenvalues are 1. Without loss of generality, assume that the first r entries of $\boldsymbol{\lambda}$ are 1 and the last $k - r$ are 0. The random vector $Q\mathbf{Z}$ is $N_n(Q\boldsymbol{\mu}, \operatorname{diag}(\boldsymbol{\lambda}))$, which implies that

$\mathbf{Z}^\top\mathbf{Z} = (Q\mathbf{Z})^\top(Q\mathbf{Z})$ is by definition distributed as $\chi^2_r(\phi) + \varphi$, where

$$\phi \;=\; \sum_{i=1}^{r}(Q\boldsymbol{\mu})_i^2 \;\text{ and }\; \varphi \;=\; \sum_{i=r+1}^{k}(Q\boldsymbol{\mu})_i^2.$$

Note, however, that

$$Q\boldsymbol{\mu} \;=\; QP\boldsymbol{\mu} \;=\; QPQ^\top Q\boldsymbol{\mu} \;=\; \text{diag}\,(\boldsymbol{\lambda})Q\boldsymbol{\mu}. \tag{9.12}$$

Since entries $r+1$ through k of $\boldsymbol{\lambda}$ are zero, the corresponding entries of $Q\boldsymbol{\mu}$ must be zero because of Equation (9.12). This implies two things: First, $\varphi = 0$; and second,

$$\phi \;=\; \sum_{i=1}^{r}(Q\boldsymbol{\mu})_i^2 \;=\; \sum_{i=1}^{k}(Q\boldsymbol{\mu})_i^2 \;=\; (Q\boldsymbol{\mu})^\top(Q\boldsymbol{\mu}) \;=\; \boldsymbol{\mu}^\top\boldsymbol{\mu}.$$

Thus, $\mathbf{Z}^\top\mathbf{Z} \sim \chi^2_r(\boldsymbol{\mu}^\top\boldsymbol{\mu})$, which proves the result. ■

Define $\Gamma = \text{diag}\,(\mathbf{p}^0)$. Let $\Sigma = \Gamma - \mathbf{p}^0(\mathbf{p}^0)^\top$ be the usual multinomial covariance matrix under the null hypothesis; i.e., $\sqrt{n}(\mathbf{X}^{(n)}/n - \mathbf{p}^0) \xrightarrow{d} N_k(\mathbf{0}, \Sigma)$ if $\mathbf{X}^{(n)} \sim \text{multinomial}(n, \mathbf{p}^0)$. Consider $\mathbf{X}^{(n)}$ to have instead a multinomial $(n, \mathbf{p}^{(n)})$ distribution. Under the assumption made earlier that $\sqrt{n}(\mathbf{p}^{(n)} - \mathbf{p}^0) \to \boldsymbol{\delta}$, it may be shown that

$$\sqrt{n}(\mathbf{X}^{(n)}/n - \mathbf{p}^{(n)}) \;\xrightarrow{d}\; N_k(\mathbf{0}, \Sigma). \tag{9.13}$$

We claim that the limit (9.13) implies that the chi square statistic $n(\mathbf{X}^{(n)}/n - \mathbf{p}^0)^\top\Gamma^{-1}(\mathbf{X}^{(n)}/n - \mathbf{p}^0)$ converges in distribution to $\chi^2_{k-1}(\boldsymbol{\delta}^\top\Gamma^{-1}\boldsymbol{\delta})$, a fact that we now prove.

First, recall that we have already shown that $\Gamma^{-1/2}\Sigma\Gamma^{-1/2}$ is a projection matrix of rank $k-1$. Define $\mathbf{V}^{(n)} = \sqrt{n}(\mathbf{X}^{(n)}/n - \mathbf{p}^0)$. Then

$$\mathbf{V}^{(n)} \;=\; \sqrt{n}(\mathbf{X}^{(n)}/n - \mathbf{p}^{(n)}) + \sqrt{n}(\mathbf{p}^{(n)} - \mathbf{p}^0).$$

The first term on the right hand side converges in distribution to $N_k(\mathbf{0}, \Sigma)$ and the second term converges to $\boldsymbol{\delta}$. Therefore, Slutsky's theorem implies that $\mathbf{V}^{(n)} \xrightarrow{d} N_k(\boldsymbol{\delta}, \Sigma)$, which gives

$$\Gamma^{-1/2}\mathbf{V}^{(n)} \xrightarrow{d} N_k(\Gamma^{-1/2}\boldsymbol{\delta}, \Gamma^{-1/2}\Sigma\Gamma^{-1/2}).$$

Thus, if we can show that $(\Gamma^{-1/2}\Sigma\Gamma^{-1/2})(\Gamma^{-1/2}\boldsymbol{\delta}) = (\Gamma^{-1/2}\boldsymbol{\delta})$, then the result we wish to prove follows from Lemma 9.6. But

$$\Gamma^{-1/2}\Sigma\Gamma^{-1}\boldsymbol{\delta} \;=\; \Gamma^{-1/2}[\Gamma - \mathbf{p}^0(\mathbf{p}^0)^\top]\Gamma^{-1}\boldsymbol{\delta} \;=\; \Gamma^{-1/2}[\boldsymbol{\delta} - \mathbf{p}^0(\mathbf{1})^\top\boldsymbol{\delta}] \;=\; \Gamma^{-1/2}\boldsymbol{\delta}$$

since $\mathbf{1}^\top\boldsymbol{\delta} = \sum_{i=1}^k \delta_i = 0$. Thus, we conclude that the chi-square statistic converges in distribution to $\chi^2_{k-1}(\boldsymbol{\delta}^\top\Gamma^{-1}\boldsymbol{\delta})$ under the sequence of alternatives $\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \ldots$.

Example 9.7 For a particular trinomial experiment with $n = 200$, suppose the null hypothesis is $H_0 : \mathbf{p} = \mathbf{p}^0 = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$. (This hypothesis might arise in the context of a genetics experiment.) We may calculate the approximate power of the Pearson chi-square test at level $\alpha = 0.01$ against the alternative $\mathbf{p} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$.

First, set $\boldsymbol{\delta} = \sqrt{n}(\mathbf{p} - \mathbf{p}^0) = \sqrt{200}(\frac{1}{12}, -\frac{1}{6}, \frac{1}{12})$. Under the alternative $\mathbf{p}$, the chi square statistic is approximately noncentral χ^2_2 with noncentrality parameter

$$\boldsymbol{\delta}^\top \operatorname{diag}(\mathbf{p}^0)^{-1}\boldsymbol{\delta} \;=\; 200\left(\frac{4}{144} + \frac{2}{36} + \frac{4}{144}\right) \;=\; \frac{200}{9}.$$

Since the test rejects H_0 whenever the statistic is larger than the .99 quantile of χ^2_2, namely 9.210, the power is approximated by $P\{\chi^2_2(\frac{200}{9}) > 9.210\} = 0.965$. These values were found using R as follows:

```
> qchisq(.99,2)
[1] 9.21034
> 1-pchisq(.Last.value, 2, ncp=200/9)
[1] 0.965006
```

Exercises for Section 9.2

Exercise 9.9 Suppose we have a tetranomial experiment and wish to test the hypothesis $H_0 : \mathbf{p} = (1/4, 1/4, 1/4, 1/4)$ against the alternative $H_1 : \mathbf{p} \neq (1/4, 1/4, 1/4, 1/4)$ at the .05 level.

(a) Approximate the power of the test against the specific alternative $(1/10, 2/10, 3/10, 4/10)$ for a sample of size $n = 200$.

(b) Give the approximate sample size necessary to give power of 80% against the alternative in part (a).

Exercise 9.10 In Exercise 9.1, let $\{\boldsymbol{\mu}^{(n)}\}$ be alternatives such that $\sqrt{n}(\boldsymbol{\mu}^{(n)} - \boldsymbol{\mu}^0) \to \boldsymbol{\delta}$. You may assume that under $\{\boldsymbol{\mu}^{(n)}\}$,

$$\sqrt{n}(\overline{\mathbf{X}} - \boldsymbol{\mu}^{(n)}) \;\xrightarrow{d}\; N_k(\mathbf{0}, \Sigma).$$

Find (with proof) the limit of the power against the alternatives $\{\boldsymbol{\mu}^{(n)}\}$ of the test that rejects $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}^0$ when $T^2 \geq c_\alpha$, where $P(\chi^2_k > c_\alpha) = \alpha$.

Chapter 10

U-statistics

When one is willing to assume the existence of a simple random sample $X_1, \ldots, X_n$, U-statistics generalize common notions of unbiased estimation such as the sample mean and the unbiased sample variance (in fact, the "U" in "U-statistics" stands for "unbiased"). Even though U-statistics may be considered a bit of a special topic, their study in a large-sample theory course has side benefits that make them valuable pedagogically. The theory of U-statistics nicely demonstrates the application of some of the large-sample topics presented thus far. Furthermore, the study of U-statistics enables a theoretical discussion of statistical functionals, which gives insight into the common modern practice of bootstrapping.

10.1 Statistical Functionals and V-Statistics

Let S be a set of cumulative distribution functions and let T denote a mapping from S into the real numbers $\mathbb{R}$. Then T is called a statistical functional. If, say, we are given a simple random sample from a distribution with unknown distribution function F, we may want to learn the value of $\theta = T(F)$ for a (known) functional T. In this way, we may think of the value of a statistical functional as a parameter we wish to estimate. Some particular instances of statistical functionals are as follows:

- If $T(F) = F(c)$ for some constant c, then T is a statistical functional mapping each F to $P_F(Y \leq c)$.
- If $T(F) = F^{-1}(p)$ for some constant p, where $F^{-1}(p)$ is defined in Equation (3.18), then T maps F to its pth quantile.
- If $T(F) = \mathrm{E}_F(Y)$ or $T(F) = \mathrm{Var}_F(Y)$, then T maps F to its mean μ or its variance σ^2,

respectively.

Suppose $X_1, \ldots, X_n$ is an independent and identically distributed sequence — in other words, a simple random sample — with distribution function $F(x)$. We define the empirical distribution function $\hat{F}_n$ to be the distribution function for a discrete uniform distribution on $\{X_1, \ldots, X_n\}$. In other words,

$$\hat{F}_n(x) \;=\; \frac{1}{n}\#\{i : X_i \le x\} \;=\; \frac{1}{n}\sum_{i=1}^{n} I\{X_i \le x\}.$$

Since $\hat{F}_n(x)$ is a legitimate distribution function, a reasonable estimator of $T(F)$ is the so-called *plug-in* estimator $T(\hat{F}_n)$. For example, if $T(F) = \mathrm{E}_F(Y)$, then the plug-in estimator given a simple random sample $X_1, X_2, \ldots$ from F is

$$T(\hat{F}_n) \;=\; \mathrm{E}_{\hat{F}_n}(Y) \;=\; \frac{1}{n}\sum_{i=1}^{n} X_i \;=\; \overline{X}_n. \tag{10.1}$$

In Equation (10.1), Y is a random variable whose distribution is the same as the empirical distribution of the data, which means that the true *population* mean of Y equals the *sample* mean of $X_1, \ldots, X_n$. This equation illustrates how we distinguish between the notational use of X and Y in this chapter: We use X and X_i whenever we must refer specifically to the data $X_1, \ldots, X_n$; but we use Y and Y_i whenever we refer generally to a functional and no specific reference to the data is made.

As we will see later, a plug-in estimator, such as $\overline{X}_n$ above, is also known as a V-statistic or a V-estimator when the functional $T(F)$ is of a particular type called an *expectation functional*.

Suppose that for some real-valued function $\phi(y)$, we define $T(F) = \mathrm{E}_F\,\phi(Y)$. In this case, we find

$$T\{\alpha F_1 + (1-\alpha)F_2\} \;=\; \alpha\,\mathrm{E}_{F_1}\,\phi(Y) + (1-\alpha)\,\mathrm{E}_{F_2}\,\phi(Y) \;=\; \alpha T(F_1) + (1-\alpha)T(F_2).$$

For this reason, such a functional is sometimes called a linear functional; see Definition 10.1.

To generalize this idea, we consider a real-valued function taking more than one real argument, say $\phi(y_1, \ldots, y_a)$ for some $a > 1$, and define

$$T(F) = \mathrm{E}_F\,\phi(Y_1, \ldots, Y_a), \tag{10.2}$$

which we take to mean the expectation of $\phi(Y_1, \ldots, Y_a)$ where $Y_1, \ldots, Y_a$ is a simple random sample from the distribution function F. Letting π denote some permutation mapping $\{1, \ldots, a\}$ onto itself, the fact that the Y_i are independent and identically distributed

means that the joint distribution of $(Y_1, \ldots, Y_a)$ is the same as the joint distribution of $(Y_{\pi(1)}, \ldots, Y_{\pi(a)})$. Therefore,

$$\mathrm{E}_F\, \phi(Y_1, \ldots, Y_a) = \mathrm{E}_F\, \phi(Y_{\pi(1)}, \ldots, Y_{\pi(a)}).$$

Since there are $a!$ such permutations, consider the function

$$\phi^*(y_1, \ldots, y_a) \stackrel{\text{def}}{=} \frac{1}{a!} \sum_{\text{all } \pi} \phi(y_{\pi(1)}, \ldots, y_{\pi(a)}).$$

Since $\mathrm{E}_F\, \phi(Y_1, \ldots, Y_a) = \mathrm{E}_F\, \phi^*(Y_1, \ldots, Y_a)$ and ϕ^* is symmetric in its arguments, we see that in Equation (10.2) we may assume without loss of generality that ϕ is symmetric in its arguments. In other words, $\phi(y_1, \ldots, y_a) = \phi(y_{\pi(1)}, \ldots, y_{\pi(a)})$ for any permutation π of the integers 1 through a. A function defined as in Equation (10.2) is called an *expectation functional*, as summarized in the following definition:

Definition 10.1 For some integer $a \geq 1$, let $\phi\colon \mathbb{R}^a \to \mathbb{R}$ be a function symmetric in its a arguments. The expectation of $\phi(Y_1, \ldots, Y_a)$ under the assumption that $Y_1, \ldots, Y_a$ are independent and identically distributed from some distribution F will be denoted by $\mathrm{E}_F\, \phi(Y_1, \ldots, Y_a)$. Then the functional $T(F) = \mathrm{E}_F\, \phi(Y_1, \ldots, Y_a)$ is called an *expectation functional*. If $a = 1$, then T is also called a *linear functional*.

Expectation functionals are important in this chapter because they are precisely the functionals that give rise to V-statistics and U-statistics. The function $\phi(y_1, \ldots, y_a)$ in Definition 10.1 is used so frequently that we give it a special name:

Definition 10.2 Let $T(F) = \mathrm{E}_F\, \phi(Y_1, \ldots, Y_a)$ be an expectation functional, where $\phi\colon \mathbb{R}^a \to \mathbb{R}$ is a function that is symmetric in its arguments. Then ϕ is called the *kernel* function associated with $T(F)$.

Suppose $T(F)$ is an expectation functional defined according to Equation (10.2). If we have a simple random sample of size n from F, then as noted earlier, a natural way to estimate $T(F)$ is by the use of the plug-in estimator $T(\hat{F}_n)$. This estimator is called a V-estimator or a V-statistic. It is possible to write down a V-statistic explicitly: Since $\hat{F}_n$ assigns probability $\frac{1}{n}$ to each X_i, we have

$$V_n \;=\; T(\hat{F}_n) \;=\; \mathrm{E}_{\hat{F}_n}\, \phi(Y_1, \ldots, Y_a) \;=\; \frac{1}{n^a} \sum_{i_1=1}^{n} \cdots \sum_{i_a=1}^{n} \phi(X_{i_1}, \ldots, X_{i_a}). \tag{10.3}$$

In the case $a = 1$, Equation (10.3) becomes

$$V_n = \frac{1}{n} \sum_{i=1}^{n} \phi(X_i). \tag{10.4}$$

It is clear in Equation (10.4) that $\mathrm{E}\, V_n = T(F)$, which we denote by θ. Furthermore, if $\sigma^2 = \mathrm{Var}_F\, \phi(Y) < \infty$, then the central limit theorem implies that

$$\sqrt{n}(V_n - \theta) \xrightarrow{d} N(0, \sigma^2).$$

For $a > 1$, however, the sum in Equation (10.3) contains some terms in which $i_1, \ldots, i_a$ are not all distinct. The expectation of such terms is not necessarily equal to $\theta = T(F)$ because in Definition 10.1, θ requires a independent random variables from F. Thus, V_n is not necessarily unbiased for $a > 1$.

Example 10.3 Let $a = 2$ and $\phi(y_1, y_2) = |y_1 - y_2|$. It may be shown (Problem 10.2) that the functional $T(F) = \mathrm{E}_F\, |Y_1 - Y_2|$ is not linear in F. Furthermore, since $|Y_{i_1} - Y_{i_2}|$ is identically zero whenever $i_1 = i_2$, it may also be shown that the V-estimator of $T(F)$ is biased:

$$\mathrm{E}\, V_n = \frac{1}{n^2} \sum\sum_{i \neq j} E_F |Y_i - Y_j| = \frac{n-1}{n} T(F)$$

because there are exactly $n(n-1)$ pairs (i, j) for which $i \neq j$.

Since the bias in V_n is due to the duplication among the subscripts $i_1, \ldots, i_a$, one way to correct this bias is to restrict the summation in Equation (10.3) to sets of subscripts $i_1, \ldots, i_a$ that contain no duplication. For example, we might sum instead over all possible subscripts satisfying $i_1 < \cdots < i_a$. The result is the U-statistic, which is the topic of Section 10.2.

Exercises for Section 10.1

Exercise 10.1 Let $X_1, \ldots, X_n$ be a simple random sample from F. For a fixed t for which $0 < F(t) < 1$, find the asymptotic distribution of $\hat{F}_n(t)$.

Exercise 10.2 Let $T(F) = \mathrm{E}_F\, |Y_1 - Y_2|$. Show that $T(F)$ is not a linear functional by exhibiting distributions F_1 and F_2 and a constant $\alpha \in (0, 1)$ such that

$$T\{\alpha F_1 + (1-\alpha) F_2\} \neq \alpha T(F_1) + (1-\alpha) T(F_2).$$

Exercise 10.3 Let $X_1, \ldots, X_n$ be a random sample from a distribution F with finite third absolute moment.

(a) For $a = 2$, find $\phi(y_1, y_2)$ such that $\mathrm{E}_F\, \phi(Y_1, Y_2) = \mathrm{Var}_F\, Y$. Your ϕ function should be symmetric in its arguments.

Hint: The fact that $\theta = \mathrm{E}\, Y_1^2 - \mathrm{E}\, Y_1 Y_2$ leads immediately to a non-symmetric ϕ function. Symmetrize it.

(b) For $a = 3$, find $\phi(y_1, y_2, y_3)$ such that $\mathrm{E}_F\, \phi(Y_1, Y_2, Y_3) = \mathrm{E}_F(Y - \mathrm{E}_F\, Y)^3$. As in part (a), ϕ should be symmetric in its arguments.

10.2 Asymptotic Normality

Recall that $X_1, \ldots, X_n$ are independent and identically distributed random variables. Because the V-statistic

$$V_n \;=\; \frac{1}{n^a} \sum_{i_1=1}^{n} \cdots \sum_{i_a=1}^{n} \phi(X_{i_1}, \ldots, X_{i_a})$$

is in general a biased estimator of the expectation functional $T(F) = \mathrm{E}_F\, \phi(Y_1, \ldots, Y_a)$ due to the presence of summands in which there are duplicated indices on the X_{i_k}, one way to produce an unbiased estimator is to sum only over those $(i_1, \ldots, i_a)$ in which no duplicates occur. Because ϕ is assumed to be symmetric in its arguments, we may without loss of generality restrict attention to the cases in which $1 \le i_1 < \cdots < i_a \le n$. Doing this, we obtain the U-statistic U_n:

Definition 10.4 Let a be a positive integer and let $\phi(y_1, \ldots, y_a)$ be the kernel function associated with an expectation functional $T(F)$ (see Definitions 10.1 and 10.2). Then the U-statistic corresponding to this functional equals

$$U_n \;=\; \frac{1}{\binom{n}{a}} \sum_{1 \le i_1 < \cdots < i_a \le n} \cdots \sum \phi(X_{i_1}, \ldots, X_{i_a}), \tag{10.5}$$

where $X_1, \ldots, X_n$ is a simple random sample of size $n \ge a$.

The "U" in "U-statistic" stands for unbiased (the "V" in "V-statistic" stands for von Mises, who was one of the originators of this theory in the late 1940's). The unbiasedness of U_n follows since it is the average of $\binom{n}{a}$ terms, each with expectation $T(F) = \mathrm{E}_F\, \phi(Y_1, \ldots, Y_a)$.

Example 10.5 Consider a random sample $X_1, \ldots, X_n$ from F, and let

$$R_n = \sum_{j=1}^{n} j I\{W_j > 0\}$$

be the Wilcoxon signed rank statistic, where $W_1, \ldots, W_n$ are simply $X_1, \ldots, X_n$ reordered in increasing absolute value. We showed in Example 8.12 that

$$R_n = \sum_{i=1}^{n} \sum_{j=1}^{i} I\{X_i + X_j > 0\}.$$

Letting $\phi(a, b) = I\{a + b > 0\}$, we see that ϕ is symmetric in its arguments and thus it is a legitimate kernel function for an expectation functional. We find that

$$\frac{1}{\binom{n}{2}} R_n \;=\; U_n + \frac{1}{\binom{n}{2}} \sum_{i=1}^{n} I\{X_i > 0\} \;=\; U_n + O_P\left(\tfrac{1}{n}\right),$$

where U_n is the U-statistic corresponding to the expectation functional $T(F) = P_F(Y_1 + Y_2 > 0)$. Therefore, some asymptotic properties of the signed rank test that we have already derived elsewhere can also be obtained using the theory of U-statistics.

In the special case $a = 1$, the V-statistic and the U-statistic coincide. In this case, we have already seen that both U_n and V_n are asymptotically normal by the central limit theorem. However, for $a > 1$, the two statistics do not coincide in general. Furthermore, we may no longer use the central limit theorem to obtain asymptotic normality because the summands are not independent (each X_i appears in more than one summand).

To prove the asymptotic normality of U-statistics, we shall use a method sometimes known as the H-projection method after its inventor, Wassily Hoeffding. If $\phi(y_1, \ldots, y_a)$ is the kernel function of an expectation functional $T(F) = \mathrm{E}_F\, \phi(Y_1, \ldots, Y_a)$, suppose $X_1, \ldots, X_n$ is a simple random sample from the distribution F. Let $\theta = T(F)$ and let U_n be the U-statistic defined in Equation (10.5). For $1 \leq k \leq a$, suppose that the values of $Y_1, \ldots, Y_k$ are held constant, say, $Y_1 = y_1, \ldots, Y_k = y_k$. This may be viewed as projecting the random vector $(Y_1, \ldots, Y_a)$ onto the $(a-k)$-dimensional subspace in $\mathbb{R}^a$ given by $\{(y_1, \ldots, y_k, c_{k+1}, \ldots, c_a) : (c_{k+1}, \ldots, c_a) \in \mathbb{R}^{a-k}\}$. If we take the conditional expectation, the result will be a function of $y_1, \ldots, y_k$, which we will denote by ϕ_k. To summarize, for $k = 1, \ldots, a$ we shall define

$$\phi_k(y_1, \ldots, y_k) \;=\; \mathrm{E}_F\, \phi(y_1, \ldots, y_k, Y_{k+1}, \ldots, Y_a). \tag{10.6}$$

Equivalently, we may use conditional expectation notation to write

$$\phi_k(Y_1, \ldots, Y_k) = \mathrm{E}_F \left\{ \phi(Y_1, \ldots, Y_a) \mid Y_1, \ldots, Y_k \right\}. \tag{10.7}$$

From Equation (10.7), we see that $\mathrm{E}_F\, \phi_k(Y_1, \ldots, Y_k) = \mathrm{E}_F\, \phi(Y_1, \ldots, Y_a) = \theta$ for all k.

The variances of the ϕ_k functions will be useful in what follows. Therefore, we introduce new notation, letting

$$\sigma_k^2 = \mathrm{Var}_F\, \phi_k(Y_1, \ldots, Y_k). \tag{10.8}$$

The importance of the σ_k^2 values, particularly σ_1^2, is seen in the following theorem, which gives a closed-form expression for the variance of a U-statistic:

Theorem 10.6 The variance of a U-statistic is

$$\operatorname{Var}_F U_n = \frac{1}{\binom{n}{a}} \sum_{k=1}^{a} \binom{a}{k} \binom{n-a}{a-k} \sigma_k^2.$$

If $\sigma_1^2, \ldots, \sigma_a^2$ are all finite, then

$$\operatorname{Var}_F U_n = \frac{a^2 \sigma_1^2}{n} + O\left(\frac{1}{n^2}\right).$$

Theorem 10.6 is proved in Exercise 10.4. This theorem shows that the variance of $\sqrt{n}U_n$ tends to $a^2\sigma_1^2$, and indeed we may well wonder whether it is true that $\sqrt{n}(U_n - \theta)$ is asymptotically normal with this limiting variance. It is the goal of Hoeffding's H-projection method to prove exactly that fact.

We shall derive the asymptotic normality of U_n in a sequence of steps. The basic idea will be to show that $U_n - \theta$ has the same limiting distribution as the sum

$$\tilde{U}_n = \sum_{j=1}^{n} \mathrm{E}_F \left(U_n - \theta \mid X_j\right) \tag{10.9}$$

of projections. The asymptotic distribution of $\tilde{U}_n$ follows from the central limit theorem because $\tilde{U}_n$ is the sum of independent and identically distributed random variables.

Lemma 10.7 For all $1 \leq j \leq n$,

$$\mathrm{E}_F \left(U_n - \theta \mid X_j\right) = \frac{a}{n} \left\{\phi_1(X_j) - \theta\right\}.$$

Proof: Expanding U_n using the definition (10.5) gives

$$\mathrm{E}_F \left(U_n - \theta \mid X_j\right) = \frac{1}{\binom{n}{a}} \sum_{1 \leq i_1 < \cdots < i_a \leq n} \cdots \sum \mathrm{E}_F \left\{\phi(X_{i_1}, \ldots, X_{i_a}) - \theta \mid X_j\right\},$$

where from equation (10.7) we see that

$$\mathrm{E}_F \left\{\phi(X_{i_1}, \ldots, X_{i_a}) - \theta \mid X_j\right\} = \begin{cases} \phi_1(X_j) - \theta & \text{if } j \in \{i_1, \ldots, i_a\} \\ 0 & \text{otherwise.} \end{cases}$$

The number of ways to choose $\{i_1, \ldots, i_a\}$ so that j is among them is $\binom{n-1}{a-1}$, so we obtain

$$\mathrm{E}_F \left(U_n - \theta \mid X_j\right) \;=\; \frac{\binom{n-1}{a-1}}{\binom{n}{a}} \left\{\phi_1(X_j) - \theta\right\} \;=\; \frac{a}{n} \left\{\phi_1(X_j) - \theta\right\}.$$

Lemma 10.8 If $\sigma_1^2 < \infty$ and $\tilde{U}_n$ is defined as in Equation (10.9), then

$$\sqrt{n}\tilde{U}_n \xrightarrow{d} N(0, a^2\sigma_1^2).$$

Proof: Lemma 10.8 follows immediately from Lemma 10.7 and the central limit theorem since $a\phi_1(X_j)$ has mean $a\theta$ and variance $a^2\sigma_1^2$.

Now that we know the asymptotic distribution of $\tilde{U}_n$, it remains to show that $U_n - \theta$ and $\tilde{U}_n$ have the same asymptotic behavior.

Lemma 10.9

$$\mathrm{E}_F\left\{\tilde{U}_n(U_n - \theta)\right\} = \mathrm{E}_F\,\tilde{U}_n^2.$$

Proof: By Equation (10.9) and Lemma 10.7, $\mathrm{E}_F\,\tilde{U}_n^2 = a^2\sigma_1^2/n$. Furthermore,

$$\begin{aligned}
\mathrm{E}_F\left\{\tilde{U}_n(U_n - \theta)\right\} &= \frac{a}{n}\sum_{j=1}^{n}\mathrm{E}_F\left\{(\phi_1(X_j) - \theta)(U_n - \theta)\right\} \\
&= \frac{a}{n}\sum_{j=1}^{n}\mathrm{E}_F\,\mathrm{E}_F\left\{(\phi_1(X_j) - \theta)(U_n - \theta) \mid X_j\right\} \\
&= \frac{a^2}{n^2}\sum_{j=1}^{n}\mathrm{E}_F\left\{\phi_1(X_j) - \theta\right\}^2 \\
&= \frac{a^2\sigma_1^2}{n},
\end{aligned}$$

where the third equality above follows from Lemma 10.7.

Lemma 10.10 If $\sigma_k^2 < \infty$ for $k = 1, \ldots, a$, then

$$\sqrt{n}\left(U_n - \theta - \tilde{U}_n\right) \xrightarrow{P} 0.$$

Proof: Since convergence in quadratic mean implies convergence in probability (Theorem 2.17), it suffices to show that

$$\mathrm{E}_F\left\{\sqrt{n}(U_n - \theta - \tilde{U}_n)\right\}^2 \to 0.$$

By Lemma 10.9, $n\,\mathrm{E}_F\left(U_n - \theta - \tilde{U}_n\right)^2 = n\left(\mathrm{Var}_F\,U_n - \mathrm{E}_F\,\tilde{U}_n^2\right)$. But $n\,\mathrm{Var}_F\,U_n = a^2\sigma_1^2 + O(1/n)$ by Theorem 10.6, and $n\,\mathrm{E}_F\,\tilde{U}_n^2 = a^2\sigma_1^2$, proving the result.

Finally, since $\sqrt{n}(U_n - \theta) = \sqrt{n}\tilde{U}_n + \sqrt{n}(U_n - \theta - \tilde{U}_n)$, Lemmas 10.8 and 10.10 along with Slutsky's theorem result in the theorem we originally set out to prove:

Theorem 10.11 If $\sigma_k^2 < \infty$ for $k = 1, \ldots, a$, then

$$\sqrt{n}(U_n - \theta) \xrightarrow{d} N(0, a^2\sigma_1^2). \tag{10.10}$$

Example 10.12 Consider the expectation functional defined by the kernel function $\phi(y_1, y_2) = (y_1 - y_2)^2/2$. We obtain

$$T(F) = \mathrm{E}_F\, \phi(Y_1, Y_2) = \mathrm{E}_F(Y_1^2 + Y_2^2 - 2Y_1Y_2)/2 = E_F Y^2 - (\mathrm{E}_F\, Y)^2 = \mathrm{Var}_F\, Y.$$

Given a simple random sample $X_1, \ldots, X_n$ from F, let us derive the asymptotic distribution of the associated U-statistic. First, we obtain

$$U_n = \frac{1}{\binom{n}{2}} \sum\sum_{1 \le i < j \le n} \phi(X_i, X_j) = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j=1}^n \phi(X_i, X_j),$$

where we have used the fact that $\phi(X_i, X_j) = 0$ in this example whenever $i = j$ in order to allow both i and j to range from 1 to n. Continuing, we obtain

$$\begin{aligned} U_n &= \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j=1}^n (X_i^2 + X_j^2 - 2X_iX_j)^2/2 \\ &= \frac{1}{2(n-1)} \sum_{i=1}^n X_i^2 + \frac{1}{2(n-1)} \sum_{j=1}^n X_j^2 - \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j=1}^n X_iX_j. \end{aligned}$$

By observing that $\overline{X}_n^2 = \sum_i \sum_j X_iX_j/n^2$, we may now conclude that

$$U_n = \frac{1}{n-1} \left[\sum_{i=1}^n X_i^2 - n\overline{X}_n^2 \right] = \frac{1}{n-1} \sum_{i=1}^n (X_i - \overline{X}_n)^2,$$

which is the usual unbiased sample variance.

We have already derived the asymptotic distribution of the sample variance — twice! — in Examples 4.11 and 5.9, though we used a *biased* version of the sample variance in each of those examples. Now, we may obtain the same result a third time using the theory of U-statistics we just developed. Since $a = 2$ here, we know that

$$\sqrt{n}(U_n - \sigma^2) \xrightarrow{d} N(0, 4\sigma_1^2).$$

It remains to find σ_1^2. To this end, we must define $\phi_1(y)$. Letting $\mu = \mathrm{E}_F\, Y$ and $\sigma^2 = \mathrm{Var}_F\, Y$, we obtain

$$\phi_1(y) = \mathrm{E}_F\, \phi(y, Y_2) = \mathrm{E}_F(y^2 - 2yY_2 + Y_2^2)/2 = (y^2 - 2\mu y + \sigma^2 + \mu^2)/2.$$

Therefore (since adding the constant σ^2 does not change the variance),

$$\sigma_1^2 = \text{Var}_F\, \phi_1(Y) = \frac{1}{4}\,\text{Var}_F(Y^2 - 2\mu Y + \mu^2) = \frac{1}{4}\,\text{Var}_F(Y-\mu)^2.$$

We conclude that

$$\sqrt{n}(U_n - \sigma^2) \xrightarrow{d} N[0, \text{Var}_F(Y-\mu)^2].$$

This confirms the results obtained in Examples 4.11 and 5.9.

Exercises for Section 10.2

Exercise 10.4 Prove Theorem 10.6, as follows:

(a) Prove that for $1 \le k \le a$,

$$\text{E}_F\, \phi(Y_1, \ldots, Y_a)\phi(Y_1, \ldots, Y_k, Y_{a+1}, \ldots, Y_{a+(a-k)}) = \sigma_k^2 + \theta^2$$

and thus $\text{Cov}_F\{\phi(Y_1, \ldots, Y_a), \phi(Y_1, \ldots, Y_k, Y_{a+1}, \ldots, Y_{a+(a-k)})\} = \sigma_k^2$.

Hint: Use conditioning! In this case, it makes sense to condition on $Y_1, \ldots, Y_k$ because conditional on those random variables, the expression above is the product of independent realizations of ϕ_k.

(b) Show that

$$\text{Var}_F \binom{n}{a} U_n =$$
$$\binom{n}{a} \sum_{k=1}^{a} \binom{a}{k}\binom{n-a}{a-k} \text{Cov}_F\{\phi(X_1, \ldots, X_a), \phi(X_1, \ldots, X_k, X_{a+1}, \ldots, X_{a+(a-k)})\}$$

and then use part (a) to prove the first equation of theorem 10.6.

(c) Verify the second equation of theorem 10.6.

Exercise 10.5 Suppose a kernel function $\phi(y_1, \ldots, y_a)$ satifies $\text{E}_F\,|\phi(Y_{i_1}, \ldots, Y_{i_a})| < \infty$ for any (not necessarily distinct) $i_1, \ldots, i_a$. Prove that if U_n and V_n are the corresponding U- and V-statistics for a simple random sample $X_1, \ldots, X_n$, then $\sqrt{n}(V_n - U_n) \xrightarrow{P} 0$ so that V_n has the same asymptotic distribution as U_n.

Hint: Verify and use the equation

$$\begin{aligned} V_n - U_n &= \left[V_n - \frac{1}{n^a} \sum_{\text{all } i_j \text{ distinct}} \cdots \sum \phi(X_{i_1}, \ldots, X_{i_a}) \right] \\ &\quad + \left[\frac{1}{n^a} - \frac{1}{a!\binom{n}{a}} \right] \sum_{\text{all } i_j \text{ distinct}} \cdots \sum \phi(X_{i_1}, \ldots, X_{i_a}). \end{aligned}$$

Exercise 10.6 For the kernel function of Example 10.3, $\phi(a, b) = |a - b|$, the corresponding U-statistic is called Gini's mean difference and it is denoted G_n. For a random sample from uniform$(0, \tau)$, find the asymptotic distribution of G_n.

Exercise 10.7 Let $\phi(y_1, y_2, y_3)$ have the property

$$\phi(a + by_1, a + by_2, a + by_3) \quad = \quad \phi(y_1, y_2, y_3)\text{sgn}(b) \quad \text{for all } a, b. \quad (10.11)$$

Let $\theta = \text{E } \phi(Y_1, Y_2, Y_3)$. The function sgn$(b)$ is defined as the sign of b, which may be expressed as $I\{b > 0\} - I\{b < 0\}$.

(a) We define the distribution F to be symmetric if for $Y \sim F$, there exists some μ (the center of symmetry) such that $Y - \mu$ and $\mu - Y$ have the same distribution. Prove that if F is symmetric then $\theta = 0$.

(b) Let $\overline{y}$ and $\tilde{y}$ denote the mean and median of y_1, y_2, y_3. Let $\phi(y_1, y_2, y_3) = \text{sgn}(\overline{y} - \tilde{y})$. Show that this function satisfies criterion (10.11), then find the asymptotic distribution for the corresponding U-statistic if F is the standard uniform distribution.

Exercise 10.8 If the arguments of the kernel function $\phi(y_1, \ldots, y_a)$ of a U-statistic are vectors instead of scalars, note that Theorem 10.11 still applies with no modification. With this in mind, consider for $\mathbf{y}, \mathbf{z} \in R^2$ the kernel $\phi(\mathbf{y}, \mathbf{z}) = I\{(y_1 - z_1)(y_2 - z_2) > 0\}$.

(a) Given a simple random sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$, if U_n denotes the U-statistic corresponding to the kernel above, the statistic $2U_n - 1$ is called Kendall's tau statistic. Suppose the marginal distributions of X_{i1} and X_{i2} are both continuous, with X_{i1} and X_{i2} independent. Find the asymptotic distribution of $\sqrt{n}(U_n - \theta)$ for an appropriate value of θ.

(b) To test the null hypothesis that a sample $W_1, \ldots, W_n$ is independent and identically distributed against the alternative hypothesis that the W_i are stochastically increasing in i, suppose we reject the null hypothesis if the number of pairs (W_i, W_j) with $W_i < W_j$ and $i < j$ is greater than c_n. This test is called Mann's test against trend. Based on your answer to part (a), find c_n so that the test has asymptotic level .05.

(c) Estimate the true level of the test in part (b) for a simple random sample of size n from a standard normal distribution for each $n \in \{5, 15, 75\}$. Use 5000 samples in each case.

Exercise 10.9 Suppose that $X_1, \ldots, X_n$ is a simple random sample from a uniform$(0, \alpha)$ distribution. For some fixed a, let U_n be the U-statistic associated with the kernel

function

$$\phi(y_1, \ldots, y_a) = \max\{y_1, \ldots, y_a\}.$$

Find the asymptotic distribution of U_n.

10.3 Multivariate and multi-sample U-statistics

In this section, we generalize the idea of U-statistics in two different directions. First, we consider single U-statistics for situations in which there is more than one sample. Next, we consider the joint asymptotic distribution of two (single-sample) U-statistics.

We begin by generalizing the idea of U-statistics to the case in which we have more than one random sample. Suppose that $X_{i1}, \ldots, X_{in_i}$ is a simple random sample from F_i for all $1 \leq i \leq s$. In other words, we have s random samples, each potentially from a different distribution, and n_i is the size of the ith sample. We may define a statistical functional

$$\theta = \mathrm{E}\ \phi\left(Y_{11}, \ldots, Y_{1a_1}; Y_{21}, \ldots, Y_{2a_2}; \cdots; Y_{s1}, \ldots, Y_{sa_s}\right). \tag{10.12}$$

Notice that the kernel ϕ in Equation (10.12) has $a_1 + a_2 + \cdots + a_s$ arguments; furthermore, we assume that the first a_1 of them may be permuted without changing the value of ϕ, the next a_2 of them may be permuted without changing the value of ϕ, etc. In other words, there are s distinct blocks of arguments of ϕ, and ϕ is symmetric in its arguments within each of these blocks. Finally, notice that in Equation (10.12), we have dropped the subscripted F on the expectation operator used in the previous section, when we wrote E_F — this is because there are now s different distributions, F_1 through F_s, and writing $\mathrm{E}_{F_1,\ldots,F_s}$ would make a bad notational situation even worse!

Letting $N = n_1 + \cdots + n_s$ denote the total sample size, the U-statistic corresponding to the expectation functional (10.12) is

$$U_N = \frac{1}{\binom{n_1}{a_1}} \cdots \frac{1}{\binom{n_s}{a_s}} \sum_{\substack{1 \leq i_1 < \cdots < i_{a_1} \leq n_1 \\ \cdots \\ 1 \leq r_1 < \cdots < r_{a_s} \leq n_s}} \cdots \sum \phi\left(X_{1i_1}, \ldots, X_{1i_{a_1}}; \cdots; X_{sr_1}, \ldots, X_{sr_{a_s}}\right). \tag{10.13}$$

As we did in the case of single-sample U-statistics, define for $0 \leq k_1 \leq a_i, \ldots, 0 \leq k_s \leq a_s$

$$\begin{aligned} &\phi_{k_1 \cdots k_s}(Y_{11}, \ldots, Y_{1k_1}; \cdots; Y_{s1}, \ldots, Y_{sk_s}) \ = \\ &\mathrm{E}\left\{\phi(Y_{11}, \ldots, Y_{1a_1}; \cdots; Y_{s1}, \ldots, Y_{sa_s}) \mid Y_{11}, \ldots, Y_{1k_1}, \cdots, Y_{s1}, \ldots, Y_{sk_s}\right\} \end{aligned} \tag{10.14}$$

and

$$\sigma^2_{k_1 \cdots k_s} = \mathrm{Var}\ \phi_{k_1 \cdots k_s}(Y_{11}, \ldots, Y_{1k_1}; \cdots; Y_{s1}, \ldots, Y_{sk_s}). \tag{10.15}$$

By an argument similar to the one used in the proof of Theorem 10.6, but much more tedious notationally, we can show that

$$\begin{aligned}\sigma^2_{k_1\cdots k_s} &= \text{Cov } \{\phi(Y_{11},\ldots,Y_{1a_1};\cdots;Y_{s1},\ldots,Y_{sa_s}),\\ &\quad \phi(Y_{11},\ldots,Y_{1k_1},Y_{1,a_1+1},\ldots;\cdots;Y_{s1},\ldots,X_{sk_s},Y_{s,a_s+1},\ldots)\}. \qquad (10.16)\end{aligned}$$

Notice that some of the k_i may equal 0. This was not true in the single-sample case, since ϕ_0 would have merely been the constant θ, so σ_0^2 would have been 0.

In the special case when $s = 2$, Equations (10.14), (10.15) and (10.16) become

$$\begin{aligned}\phi_{jk}(Y_1,\ldots,Y_j;Z_1,\ldots,Z_k) &= \text{E } \{\phi(Y_1,\ldots,Y_{a_1};Z_1,\ldots,Z_{a_2}) \mid Y_1,\ldots,Y_j,Z_1,\ldots,Z_k\},\\ \sigma^2_{jk} &= \text{Var } \phi_{jk}(Y_1,\ldots,Y_j;Z_1\ldots,Z_k),\end{aligned}$$

and

$$\begin{aligned}\sigma^2_{jk} &= \text{Cov } \{\phi(Y_1,\ldots,Y_{a_1};Z_1,\ldots,Z_{a_2}),\\ &\quad \phi(Y_1,\ldots,Y_j,Y_{a_1+1},\ldots,Y_{a_1+(a_1-j)};Z_1,\ldots,Z_k,Z_{a_2+1},\ldots,Z_{a_2+(a_2-k)})\},\end{aligned}$$

respectively, for $0 \le j \le a_1$ and $0 \le k \le a_2$.

Although we will not derive it here as we did for the single-sample case, there is an analagous asymptotic normality result for multisample U-statistics, as follows.

Theorem 10.13 Suppose that for $i = 1,\ldots,s$, $X_{i1},\ldots,X_{in_i}$ is a random sample from the distribution F_i and that these s samples are independent of each other. Suppose further that there exist constants $\rho_1,\ldots,\rho_s$ in the interval $(0,1)$ such that $n_i/N \to \rho_i$ for all i and that $\sigma^2_{a_1\cdots a_s} < \infty$. Then

$$\sqrt{N}(U_N - \theta) \xrightarrow{d} N(0,\sigma^2),$$

where

$$\sigma^2 = \frac{a_1^2}{\rho_1}\sigma^2_{10\cdots00} + \cdots + \frac{a_s^2}{\rho_s}\sigma^2_{00\cdots01}.$$

Although the notation required for the multisample U-statistic theory is nightmarish, life becomes considerably simpler in the case $s = 2$ and $a_1 = a_2 = 1$, in which case we obtain

$$U_N = \frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\phi(X_{1i};X_{2j}).$$

Equivalently, we may assume that $X_1, \ldots, X_m$ are a simple random sample from F and $Y_1, \ldots, Y_n$ are a simple random sample from G, which gives

$$U_N = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \phi(X_i; Y_j). \tag{10.17}$$

In the case of the U-statistic of Equation (10.17), Theorem 10.13 states that

$$\sqrt{N}(U_N - \theta) \xrightarrow{d} N\left(0, \frac{\sigma_{10}^2}{\rho} + \frac{\sigma_{01}^2}{1-\rho}\right),$$

where $\rho = \lim m/N$, $\sigma_{10}^2 = \text{Cov}\{\phi(X_1; Y_1), \phi(X_1; Y_2)\}$, and $\sigma_{01}^2 = \text{Cov}\{\phi(X_1; Y_1), \phi(X_2; Y_1)\}$.

Example 10.14 For independent random samples $X_1, \ldots X_m$ from F and $Y_1, \ldots, Y_n$ from G, consider the Wilcoxon rank-sum statistic W, defined to be the sum of the ranks of the Y_i among the combined sample. We may show that

$$W = \frac{1}{2}n(n+1) + \sum_{i=1}^{m} \sum_{j=1}^{n} I\{X_i < Y_j\}.$$

Therefore, if we let $\phi(a; b) = I\{a < b\}$, then the corresponding two-sample U-statistic U_N is related to W by $W = \frac{1}{2}n(n+1) + mnU_N$. Therefore, we may use Theorem 10.13 to obtain the asymptotic normality of U_N, and therefore of W. However, we make no assumption here that F and G are merely shifted versions of one another. Thus, we may now obtain in principle the asymptotic distribution of the rank-sum statistic for any two distributions F and G that we wish, so long as they have finite second moments.

The other direction in which we will generalize the development of U-statistics is consideration of the joint distribution of two single-sample U-statistics. Suppose that there are two kernel functions, $\phi(y_1, \ldots, y_a)$ and $\varphi(y_1, \ldots, y_b)$, and we define the two corresponding U-statistics

$$U_n^{(1)} = \frac{1}{\binom{n}{a}} \sum_{1 \le i_1 < \cdots < i_a \le n} \cdots \sum \phi(X_{i_1}, \ldots, X_{i_a})$$

and

$$U_n^{(2)} = \frac{1}{\binom{n}{b}} \sum_{1 \le j_1 < \cdots < j_b \le n} \cdots \sum \varphi(X_{j_1}, \ldots, X_{j_b})$$

for a single random sample $X_1, \ldots, X_n$ from F. Define $\theta_1 = \text{E}\, U_n^{(1)}$ and $\theta_2 = \text{E}\, U_n^{(2)}$. Furthermore, define γ_{jk} to be the covariance between $\phi_j(Y_1, \ldots, Y_j)$ and $\varphi_k(Y_1, \ldots, Y_k)$, where ϕ_j and φ_k are defined as in Equation (10.7). Letting $\ell = \min\{j, k\}$, it may be proved that

$$\gamma_{jk} = \text{Cov}\left\{\phi(Y_1, \ldots, Y_a), \varphi(Y_1, \ldots, Y_\ell, Y_{a+1}, \ldots, Y_{a+(b-\ell)})\right\}. \tag{10.18}$$

Note in particular that γ_{jk} depends only on the value of $\min\{j,k\}$.

The following theorem, stated without proof, gives the joint asymptotic distribution of $U_n^{(1)}$ and $U_n^{(2)}$.

Theorem 10.15 Suppose $X_1,\ldots,X_n$ is a random sample from F and that $\phi:\mathbb{R}^a\to\mathbb{R}$ and $\varphi:\mathbb{R}^b\to\mathbb{R}$ are two kernel functions satisfying Var $\phi(Y_1,\ldots,Y_a)<\infty$ and Var $\varphi(Y_1,\ldots,Y_b)<\infty$. Define $\tau_1^2=$ Var $\phi_1(Y_1)$ and $\tau_2^2=$ Var $\varphi_1(Y_1)$, and let γ_{jk} be defined as in Equation (10.18). Then

$$\sqrt{n}\left\{\begin{pmatrix}U_n^{(1)}\\U_n^{(2)}\end{pmatrix}-\begin{pmatrix}\theta_1\\\theta_2\end{pmatrix}\right\}\xrightarrow{d}N\left\{\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}a^2\tau_1^2 & ab\gamma_{11}\\ ab\gamma_{11} & b^2\tau_2^2\end{pmatrix}\right\}.$$

Exercises for Section 10.3

Exercise 10.10 Suppose $X_1,\ldots,X_m$ and $Y_1,\ldots,Y_n$ are independent random samples from distributions Unif$(0,\theta)$ and Unif$(\mu,\mu+\theta)$, respectively. Assume $m/N\to\rho$ as $m,n\to\infty$ and $0<\mu<\theta$.

(a) Find the asymptotic distribution of the U-statistic of Equation (10.17), where $\phi(x;y)=I\{x<y\}$. In so doing, find a function $g(x)$ such that $\mathrm{E}\,(U_N)=g(\mu)$.

(b) Find the asymptotic distribution of $g(\overline{Y}-\overline{X})$.

(c) Find the range of values of μ for which the Wilcoxon estimate of $g(\mu)$ is asymptotically more efficient than $g(\overline{Y}-\overline{X})$. (The asymptotic relative efficiency in this case is the ratio of asymptotic variances.)

Exercise 10.11 Solve each part of Problem 10.10, but this time under the assumptions that the independent random samples $X_1,\ldots,X_m$ and $Y_1,\ldots,Y_n$ satisfy $P(X_1\le t)=P(Y_1-\theta\le t)=t^2$ for $t\in[0,1]$ and $0<\theta<1$. As in Problem 10.10, assume $m/N\to\rho\in(0,1)$.

Exercise 10.12 Suppose $X_1,\ldots,X_m$ and $Y_1,\ldots,Y_n$ are independent random samples from distributions $N(0,1)$ and $N(\mu,1)$, respectively. Assume $m/(m+n)\to 1/2$ as $m,n\to\infty$. Let U_N be the U-statistic of Equation (10.17), where $\phi(x;y)=I\{x<y\}$. Suppose that $\theta(\mu)$ and $\sigma^2(\mu)$ are such that

$$\sqrt{N}[U_N-\theta(\mu)]\xrightarrow{d}N[0,\sigma^2(\mu)].$$

Calculate $\theta(\mu)$ and $\sigma^2(\mu)$ for $\mu\in\{.2,.5,1,1.5,2\}$.

Hint: This problem requires a bit of numerical integration. There are a couple of ways you might do this. A symbolic mathematics program like Mathematica or Maple will do it. There is a function called `integrate` in R and Splus and one called `quad` in MATLAB for integrating a function. If you cannot get any of these to work for you, let me know.

Exercise 10.13 Suppose $X_1, X_2, \ldots$ are independent and identically distributed with finite variance. Define

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

and let G_n be Gini's mean difference, the U-statistic defined in Problem 10.6. Note that S_n^2 is also a U-statistic, corresponding to the kernel function $\phi(x_1, x_2) = (x_1 - x_2)^2/2$.

(a) If X_i are distributed as Unif$(0, \theta)$, give the joint asymptotic distribution of G_n and S_n by first finding the joint asymptotic distribution of the U-statistics G_n and S_n^2. Note that the covariance matrix need not be positive definite; in this problem, the covariance matrix is singular.

(b) The singular asymptotic covariance matrix in this problem implies that as $n \to \infty$, the joint distribution of G_n and S_n becomes concentrated on a line. Does this appear to be the case? For 1000 samples of size n from Uniform$(0, 1)$, plot scatterplots of G_n against S_n. Take $n \in \{5, 25, 100\}$.

10.4 Introduction to the Bootstrap

This section does not use very much large-sample theory aside from the weak law of large numbers, and it is not directly related to the study of U-statistics. However, we include it here because of its natural relationship with the concepts of statistical functionals and plug-in estimators seen in Section 10.1, and also because it is an increasingly popular and often misunderstood method in statistical estimation.

Consider a statistical functional $T_n(F)$ that depends on n. For instance, $T_n(F)$ may be some property, such as bias or variance, of an estimator $\hat{\theta}_n$ of $\theta = \theta(F)$ based on a random sample of size n from some distribution F.

As an example, let $\theta(F) = F^{-1}\left(\frac{1}{2}\right)$ be the median of F. Take $\hat{\theta}_n$ to be the mth order statistic from a random sample of size $n = 2m - 1$ from F.

Consider the bias $T_n^B(F) = \mathrm{E}_F\,\hat{\theta}_n - \theta(F)$ and the variance $T_n^V(F) = \mathrm{E}_F\,\hat{\theta}_n^2 - (\mathrm{E}_F\,\hat{\theta}_n)^2$.

Theoretical properties of T_n^B and T_n^V are very difficult to obtain. Even asymptotics aren't very helpful, since $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N\{0, 1/(4f^2(\theta))\}$ tells us only that the bias goes to zero and the limiting variance may be very hard to estimate because it involves the unknown quantity $f(\theta)$, which is hard to estimate.

Consider the plug-in estimators $T_n^B(\hat{F}_n)$ and $T_n^V(\hat{F}_n)$. (Recall that $\hat{F}_n$ denotes the empirical distribution function, which puts a mass of $\frac{1}{n}$ on each of the n sample points.) In our median example,

$$T_n^B(\hat{F}_n) \;=\; \mathrm{E}_{\hat{F}_n}\, \hat{\theta}_n^* - \hat{\theta}_n$$

and

$$T_n^V(\hat{F}_n) \;=\; \mathrm{E}_{\hat{F}_n}(\hat{\theta}_n^*)^2 - (\mathrm{E}_{\hat{F}_n}\, \hat{\theta}_n^*)^2,$$

where $\hat{\theta}_n^*$ is the sample median from a random sample $X_1^*, \ldots, X_n^*$ from $\hat{F}_n$.

To see how difficult it is to calculate $T_n^B(\hat{F}_n)$ and $T_n^V(\hat{F}_n)$, consider the simplest nontrivial case, $n = 3$: Conditional on the order statistics $(X_{(1)}, X_{(2)}, X_{(3)})$, there are 27 equally likely possibilities for the value of (X_1^*, X_2^*, X_3^*), the sample of size 3 from $\hat{F}_n$, namely

$$(X_{(1)}, X_{(1)}, X_{(1)}), (X_{(1)}, X_{(1)}, X_{(2)}), \ldots, (X_{(3)}, X_{(3)}, X_{(3)}).$$

Of these 27 possibilities, exactly $1 + 6 = 7$ have the value $X_{(1)}$ occurring 2 or 3 times. Therefore, we obtain

$$P(\hat{\theta}_n^* = X_{(1)}) = \frac{7}{27}, P(\hat{\theta}_n^* = X_{(2)}) = \frac{13}{27}, \quad \text{and} \quad P(\hat{\theta}_n^* = X_{(3)}) = \frac{7}{27}.$$

This implies that

$$\mathrm{E}_{\hat{F}_n}\, \hat{\theta}_n^* \;=\; \frac{1}{27}(7X_{(1)} + 13X_{(2)} + 7X_{(3)}) \quad \text{and} \quad \mathrm{E}_{\hat{F}_n}(\hat{\theta}_n^*)^2 \;=\; \frac{1}{27}(7X_{(1)}^2 + 13X_{(2)}^2 + 7X_{(3)}^2).$$

Therefore, since $\hat{\theta}_n = X_{(2)}$, we obtain

$$T_n^B(\hat{F}_n) = \frac{1}{27}(7X_{(1)} - 14X_{(2)} + 7X_{(3)})$$

and

$$T_n^V(\hat{F}_n) = \frac{14}{729}(10X_{(1)}^2 + 13X_{(2)}^2 + 10X_{(3)}^2 - 13X_{(1)}X_{(2)} - 13X_{(2)}X_{(3)} - 7X_{(1)}X_{(3)}).$$

To obtain the sampling distribution of these estimators, of course, we would have to consider the joint distribution of $(X_{(1)}, X_{(2)}, X_{(3)})$. Naturally, the calculations become even more difficult as n increases.

Alternatively, we could use resampling in order to approximate $T_n^B(\hat{F}_n)$ and $T_n^V(\hat{F}_n)$. This is the bootstrapping idea, and it works like this: For some large number B, simulate B random samples from $\hat{F}_n$, namely

$$\begin{array}{c} X_{11}^*, \quad \ldots \quad , X_{1n}^*, \\ \vdots \\ X_{B1}^*, \quad \ldots \quad , X_{Bn}^*, \end{array}$$

and approximate a quantity like $\mathrm{E}_{\hat{F}_n} \hat{\theta}_n^*$ by the sample mean

$$\frac{1}{B} \sum_{i=1}^{B} \hat{\theta}_{in}^*,$$

where $\hat{\theta}_{in}^*$ is the sample median of the ith bootstrap sample $X_{i1}^*, \ldots, X_{in}^*$. Notice that the weak law of large numbers asserts that

$$\frac{1}{B} \sum_{i=1}^{B} \hat{\theta}_{in}^* \xrightarrow{P} \mathrm{E}_{\hat{F}_n} \hat{\theta}_n^*.$$

To recap, then, we wish to estimate some parameter $T_n(F)$ for an unknown distribution F based on a random sample from F. We estimate $T_n(F)$ by $T_n(\hat{F}_n)$, but it is not easy to evaluate $T_n(\hat{F}_n)$ so we approximate $T_n(\hat{F}_n)$ by resampling B times from $\hat{F}_n$ and obtain a bootstrap estimator $T_{B,n}^*$. Thus, there are two relevant issues:

1. How good is the approximation of $T_n(\hat{F}_n)$ by $T_{B,n}^*$? (Note that $T_n(\hat{F}_n)$ is NOT an unknown parameter; it is "known" but hard to evaluate.)
2. How precise is the estimation of $T_n(F)$ by $T_n(\hat{F}_n)$?

Question 1 is usually addressed using an asymptotic argument using the weak law or the central limit theorem and letting $B \to \infty$. For example, if we have an expectation functional $T_n(F) = \mathrm{E}_F\, h(X_1, \ldots, X_n)$, then

$$T_{B,n}^* \;=\; \frac{1}{B} \sum_{i=1}^{B} h(X_{i1}^*, \ldots, X_{in}^*) \;\xrightarrow{P}\; T_n(\hat{F}_n)$$

as $B \to \infty$.

Question 2, on the other hand, is often tricky; asymptotic results involve letting $n \to \infty$ and are handled case-by-case. We will not discuss these asymptotics here. On a related note,

however, there is an argument in Lehmann's book (on pages 432–433) about why a plug-in estimator may be better than an asymptotic estimator. That is, if it is possible to show $T_n(F) \to T$ as $n \to \infty$, then as an estimator of $T_n(F)$, $T_n(\hat{F}_n)$ may be preferable to T.

We conclude this section by considering the so-called parametric bootstrap. If we assume that the unknown distribution function F comes from a family of distribution functions indexed by a parameter μ, then $T_n(F)$ is really $T_n(F_\mu)$. Then, instead of the plug-in estimator $T_n(\hat{F}_n)$, we might consider the estimator $T_n(F_{\hat{\mu}})$, where $\hat{\mu}$ is an estimator of μ.

Everything proceeds as in the nonparametric version of bootstrapping. Since it may not be easy to evaulate $T_n(F_{\hat{\mu}})$ explicitly, we first find $\hat{\mu}$ and then take B random samples of size n, $X_{11}^*, \ldots, X_{1n}^*$ through $X_{B1}^*, \ldots, X_{Bn}^*$, from $F_{\hat{\mu}}$. These samples are used to approximate $T_n(F_{\hat{\mu}})$.

Example 10.16 Suppose $X_1, \ldots, X_n$ is a random sample from Poisson(μ). Take $\hat{\mu} = \overline{X}$. Suppose $T_n(F_\mu) = \text{Var}_{F_\mu} \hat{\mu}$. In this case, we happen to know that $T_n(F_\mu) = \mu/n$, but let's ignore this knowledge and apply a parametric bootstrap. For some large B, say 500, generate B samples from Poisson($\hat{\mu}$) and use the sample variance of $\hat{\mu}^*$ as an approximation to $T_n(F_{\hat{\mu}})$. In R, with $\mu = 1$ and $n = 20$ we obtain

```
 x <- rpois(20,1)  # Generate the sample from F
 muhat <- mean(x)
 muhat
[1] 0.85
 muhatstar <- rep(0,500) # Allocate the vector for muhatstar
 for(i in 1:500) muhatstar[i] <- mean(rpois(20,muhat))
 var(muhatstar)
[1] 0.04139177
```

Note that the estimate 0.041 is close to the known true value 0.05. This example is simplistic because we already know that $T_n(F) = \mu/n$, which makes $\hat{\mu}/n$ a more natural estimator. However, it is not always so simple to obtain a closed-form expression for $T_n(F)$.

Incidentally, we could also use a nonparametric bootstrap approach in this example:

```
 for (i in 1:500) muhatstar2[i] <- mean(sample(x,replace=T))
 var(muhatstar2)
[1] 0.0418454
```

Of course, 0.042 is an approximation to $T_n(\hat{F}_n)$ rather than $T_n(F_{\hat{\mu}})$. Furthermore, we can obtain a result arbitrarily close to $T_n(\hat{F}_n)$ by increasing the value of B:

```
muhatstar2_rep(0,100000)
for (i in 1:100000) muhatstar2[i] <- mean(sample(x,replace=T))
var(muhatstar2)
[1] 0.04136046
```

In fact, it is in principle possible to obtain an approximate variance for our estimates of $T_n(\hat{F}_n)$ and $T_n(F_{\hat{\mu}})$, and, using the central limit theorem, construct approximate confidence intervals for these quantities. This would allow us to specify the quantities to any desired level of accuracy.

Exercises for Section 10.4

Exercise 10.14 **(a)** Devise a nonparametric bootstrap scheme for setting confidence intervals for β in the linear regression model $Y_i = \alpha + \beta x_i + \epsilon_i$. There is more than one possible answer.

(b) Using $B = 1000$, implement your scheme on the following dataset to obtain a 95% confidence interval. Compare your answer with the standard 95% confidence interval.

Y	21	16	20	34	33	43	47
x	460	498	512	559	614	675	719

(In the dataset, Y is the number of manatee deaths due to collisions with powerboats in Florida and x is the number of powerboat registrations in thousands for even years from 1978-1990.)

Exercise 10.15 Consider the following dataset that lists the latitude and mean August temperature in degrees Fahrenheit for 7 US cities. The residuals are listed for use in part (b).

City	Latitude	Temperature	Residual
Miami	26	83	-5.696
Phoenix	33	92	10.116
Memphis	35	81	1.062
Baltimore	39	76	-0.046
Pittsburgh	40	71	-4.073
Boston	42	72	-1.127
Portland, OR	46	69	-0.235

Minitab gives the following output for a simple linear regression:

```
Predictor        Coef     SE Coef          T        P
Constant       113.99       13.01       8.76    0.000
latitude      -0.9730       0.3443     -2.83    0.037

S = 5.546       R-Sq = 61.5%     R-Sq(adj) = 53.8%
```

Note that this gives an asymptotic estimate of the variance of the slope parameter as $.3443^2 = .1185$.

In (a) through (c) below, use the described method to simulate $B = 500$ bootstrap samples $(x^*_{b1}, y^*_{b1}), \ldots, (x^*_{b7}, y^*_{b7})$ for $1 \leq b \leq B$. For each b, refit the model to obtain $\hat{\beta}^*_b$. Report the sample variance of $\hat{\beta}^*_1, \ldots, \hat{\beta}^*_B$ and compare with the asymptotic estimate of .1185.

(a) Parametric bootstrap. Take $x^*_{bi} = x_i$ for all b and i. Let $y^*_{bi} = \hat{\beta}_0 + \hat{\beta}_1 x_i + \epsilon_i$, where $\epsilon_i \sim N(0, \hat{\sigma}^2)$. Obtain $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\sigma}^2$ from the above output.

(b) Nonparametric bootstrap I. Take $x^*_{bi} = x_i$ for all b and i. Let $y^*_{bi} = \hat{\beta}_0 + \hat{\beta}_1 x_i + r^*_{bi}$, where $r^*_{b1}, \ldots, r^*_{b7}$ is an iid sample from the empirical distribution of the residuals from the original model (you may want to refit the original model to find these residuals).

(c) Nonparametric bootstrap II. Let $(x^*_{b1}, y^*_{b1}), \ldots, (x^*_{b7}, y^*_{b7})$ be an iid sample from the empirical distribution of $(x_1, y_1), \ldots, (x_7, y_7)$.

Note: In R or Splus, you can obtain the slope coefficient of the linear regression of the vector `y` on the vector `x` using `lm(y~x)$coef[2]`.

Exercise 10.16 The same resampling idea that is exploited in the bootstrap can be used to approximate the value of difficult integrals by a technique sometimes called Monte Carlo integration. Suppose we wish to compute

$$\theta = 2 \int_0^1 e^{-x^2} \cos^3(x)\, dx.$$

(a) Use numerical integration (e.g., the `integrate` function in R and Splus) to verify that $\theta = 1.070516$.

(b) Define $g(t) = 2e^{-t^2} \cos^3(t)$. Let $U_1, \ldots, U_n$ be an iid uniform(0,1) sample. Let

$$\hat{\theta}_1 = \frac{1}{n} \sum_{i=1}^{n} g(U_i).$$

Prove that $\hat{\theta}_1 \xrightarrow{P} \theta$.

(c) Define $h(t) = 2 - 2t$. Prove that if we take $V_i = 1 - \sqrt{U_i}$ for each i, then V_i is a random variable with density $h(t)$. Prove that with

$$\hat{\theta}_2 = \frac{1}{n} \sum_{i=1}^{n} \frac{g(V_i)}{h(V_i)},$$

we have $\hat{\theta}_2 \xrightarrow{P} \theta$.

(d) For $n = 1000$, simulate $\hat{\theta}_1$ and $\hat{\theta}_2$. Give estimates of the variance for each estimator by reporting $\hat{\sigma}^2/n$ for each, where $\hat{\sigma}^2$ is the sample variance of the $g(U_i)$ or the $g(V_i)/h(V_i)$ as the case may be.

(e) Plot, on the same set of axes, $g(t)$, $h(t)$, and the standard uniform density for $t \in [0, 1]$. From this plot, explain why the variance of $\hat{\theta}_2$ is smaller than the variance of $\hat{\theta}_1$. [Incidentally, the technique of drawing random variables from a density h whose shape is close to the function g of interest is a variance-reduction technique known as *importance sampling*.]

Note: This was sort of a silly example, since numerical methods yield an exact value for θ. However, with certain high-dimensional integrals, the "curse of dimensionality" makes exact numerical methods extremely time-consuming computationally; thus, Monte Carlo integration does have a practical use in such cases.