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 defining 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

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

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 \( \infty \) if for every real number \( M \), there exists \( N \) such that

\[
a_n > M \quad \text{for all } n > N.
\]

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 \( \infty \)”.

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 \( \epsilon \). 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 \( \infty \). 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

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

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 
\[
|a_n + b_n - (a + b)| \leq |a_n - a| + |b_n - b|,
\]
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 \( a_nb_n \to ab \); 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 minimum. 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 \( \not\in S \). 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, \( \inf_{k\leq n} a_k \) 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 \( \lim \sup_n a_n \), is the limit of the nonincreasing sequence

\[
\sup_{k\geq 1} a_k, \quad \sup_{k\geq 2} a_k, \quad \ldots
\]

The limit inferior, denoted \( \liminf_n a_n \) or sometimes \( \lim \inf_n a_n \), is the limit of the nondecreasing sequence

\[
\inf_{k\geq 1} a_k, \quad \inf_{k\geq 2} a_k, \ldots
\]

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 \( \liminf \) and \( \limsup \) 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 $\limsup$ and $\liminf$ 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 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 formulation).

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 analogously: $f(x_0^-) = a$ 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

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

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 $f(x_0^-) = \pm\infty$. 

7
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.](image)
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 $\epsilon$ was arbitrary, the proof is finished. \qed

**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 $a_nb_n \to ab$

**Hint:** Show that $|a_nb_n - ab| \leq |(a_n - a)(b_n - b)| + |a(b_n - b)| + |b(a_n - a)|$ 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 Cesàro, which means that

$$\frac{1}{n} \sum_{k=1}^{n} a_k \to c \quad \text{as } n \to \infty.$$  \hfill (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 $k \leq N$ and one for $k > N$. The case $c = \pm \infty$ follows a similar line of reasoning.

**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 $s \leq M$ 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 $x \not\in \{0, 1, 2\}$, define

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

10
(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 $\lim \sup_n F(a_n) \neq F(a)$.

(b) Change your answer to part (a) so that $a_n \to a$ and $\lim \sup_n F(a_n) = F(a)$, 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 $\lim \inf_n F(a_n) = F(a)$, 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

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

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 $k$th 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 \( f^{(k)}(x_0) \):

\[
\left. \frac{d^k}{dx^k} f(x) \right|_{x=x_0}
\]

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

\[
f(x) = f(a) + (x-a)f'(a) + \cdots + \frac{(x-a)^d}{d!}f^{(d)}(a) + r_d(x,a), \quad (1.5)
\]

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

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

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

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

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  \textit{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
\[
\lim_{x \to c} \frac{f(x)}{g(x)} = \lim_{x \to c} \frac{f'(x)}{g'(x)},
\] (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 \((a, \infty)\) or \((-\infty, a)\).

Example 1.20  Example 1.4 states that
\[
\lim_{n \to \infty} \left(1 + \frac{c}{n}\right)^n = \exp(c)
\] (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
\[
\lim_{x \to 0} \frac{\log(1+cx)}{x} = \lim_{x \to 0} \frac{c}{1+cx} = c.
\]
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
\[
\lim_{n \to \infty} \frac{\log(1+cx_n)}{x_n} = \lim_{n \to \infty} n \log \left(1 + \frac{c}{n}\right) = c,
\] (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
\[
\lim_{x \to x_0} \frac{f(x) - f(2x_0 - x)}{2(x - x_0)}.
\] (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
\[
f'(a) \approx \frac{f(a + h) - f(a)}{h}.
\] (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

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

\[
\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt
\] (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

\[
\Gamma(x + 1) = x\Gamma(x)
\] (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

\[
\Psi(x) \approx \frac{1}{2} \log [x(x-1)]
\] (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 \texttt{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, \texttt{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

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

the above fraction might equal $0/0$ or $\infty/\infty$, 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

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

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 $k > -1$,

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

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

$$\sum_{i=1}^{n} \frac{1}{i} \sim \log n. \quad (1.21)$$

**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

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

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

$$1 \leq \liminf_{n} \sum_{i=1}^{n} \frac{1}{\log n} \leq \limsup_{n} \sum_{i=1}^{n} \frac{1}{\log n} \leq 1.$$ 

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

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

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

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

we conclude that $(\log \log n) / \log n$ must also tend to 0 as $n$ tends to $\infty$ 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

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

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

\[ \sum_{i=1}^{n} \sqrt{i} \geq \sum_{i=\lfloor n/2 \rfloor}^{n} \sqrt{i} \geq \frac{n}{2} \sqrt{\lfloor n/2 \rfloor}. \]

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

\[ \sum_{i=1}^{n} \sqrt{i} \sim \frac{2n^{3/2}}{3}. \] (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 \( \infty \) 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 $\infty$ that are so common that they have names, such as logarithmic, polynomial, and exponential growth. If $\alpha$, $\beta$, and $\gamma$ are arbitrary positive constants, then the sequences $\log n^\alpha$, $n^\beta$, and $(1 + \gamma)^n$ exhibit logarithmic, polynomial, and exponential growth, respectively. Furthermore, we always have

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

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 $\alpha$ and $\beta$. Since $\exp(x)$ is a convex function, Theorem 1.31 gives

$$(\log n)^\alpha = o(n^\beta). \quad (1.27)$$

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 $\beta$ and $\gamma$. 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 $f(x)$. 

21
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 \( \delta_n(X) = \frac{1}{n+2} \sum_{i=1}^{n} X_i \) of \( g(\theta) = 1/\theta \). Show that for some constants \( c_1 \) and \( c_2 \) depending on \( \theta \),

\[
\text{bias of } \delta_n \sim c_1 \quad \text{(variance of } \delta_n) \sim \frac{c_2}{n}
\]

as \( n \to \infty \). The bias of \( \delta_n \) 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 \( \delta_n \) be the posterior mean of \( \theta \), assuming a standard exponential prior for \( \theta \) (i.e., \( p(\theta) = e^{-\theta I\{\theta > 0\}} \)). Compute \( \delta_n \).

**Hints:** The posterior distribution of \( \theta \) 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 \( \alpha/\beta \). To determine \( \alpha \) and \( \beta \) for the posterior distribution of \( \theta \), 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 \( \alpha \) and \( \beta \).

(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 \( \delta_n \) 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 \( \delta_n \)). 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 $n^b$ 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 $\delta_n$ 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 $\theta$ 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 $\delta_n$.

**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 \( \infty \) 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.

\[
\begin{align*}
\sqrt{n} \log n! & \quad \sum_{i=1}^{n} \sqrt{i} & 2\log n & \quad (\log n)^{\log\log n}
\end{align*}
\]

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.

\[
\begin{align*}
\log(n!) & \quad n^2 & \quad n^n & \quad 3^n \\
\log(\log n) & \quad n & \quad \log n & \quad 2^3\log n & \quad n^{n/2} \\
n! & \quad 2^n & \quad n^{\log n} & \quad (\log n)^n
\end{align*}
\]

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 \( j \)th component of the \( i \)th 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:

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

We may now write down the analogue of Definition 1.1.

**Definition 1.32** The sequence $a_1, a_2, \ldots$ is said to have limit $c \in \mathbb{R}^k$, written $a_n \to c$ as $n \to \infty$ or $\lim_{n \to \infty} a_n = c$, if $\|a_n - c\| \to 0$ as $n \to \infty$. That is, $a_n \to c$ means that for any $\epsilon > 0$ there exists $N$ such that $\|a_n - c\| < \epsilon$ 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 $a_n \to c$:

**Lemma 1.33** $a_n \to c$ if and only if $a_{nj} \to c_j$ for all $1 \leq j \leq k$.

**Proof:** Since

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

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 $|a_{nj} - c_j| \leq \|a_n - c\|$ for each $j$.

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

25
Suppose that the function $f(x)$ maps vectors in some open subset $U$ of $\mathbb{R}^k$ to vectors in $\mathbb{R}^\ell$, a property denoted by $f : U \to \mathbb{R}^\ell$. In order to define continuity, we first extend Definition 1.11 to the multivariate case:

**Definition 1.34** For a function $f : U \to \mathbb{R}^\ell$, where $U$ is open in $\mathbb{R}^k$, we write

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

for some $a \in U$ and $c \in \mathbb{R}^\ell$ if for every $\epsilon > 0$ there exists a $\delta > 0$ such that $\|f(x) - c\| < \epsilon$ whenever $x \in U$ and $0 < \|x - a\| < \delta$.

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

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

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

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

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

$$\lim_{h \to 0} \frac{f(a + h) - f(a) - J_f(a)h}{\|h\|} = 0. \quad (1.29)$$

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

The assertion in Definition 1.36 that $J_f(a)$ is unique may be proved as follows: Suppose that $J_f^{(1)}(a)$ and $J_f^{(2)}(a)$ are two versions of the Jacobian matrix. Then Equation (1.29) implies
that
\[
\lim_{h \to 0} \frac{(J_f^{(1)}(a) - J_f^{(2)}(a)) h}{\|h\|} = 0;
\]
but $h/\|h\|$ is an arbitrary unit vector, which means that $(J_f^{(1)}(a) - J_f^{(2)}(a))$ 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 $f'(x)$, $f(x)$, $Df(x)$, or $\nabla f(x)$ 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 $\nabla f(x)$ denote the transpose of the Jacobian matrix $J_f(x)$. We often refer to $\nabla f$ as the gradient of $f$.

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

**Definition 1.37** Let $g(x)$ be a real-valued function defined on a neighborhood of $a$ in $\mathbb{R}^k$. For $1 \leq i \leq k$, let $e_i$ denote the $i$th standard basis vector in $\mathbb{R}^k$, consisting of a one in the $i$th component and zeros elsewhere. We define the $i$th partial derivative of $g(x)$ at $a$ to be
\[
\frac{\partial g(x)}{\partial x_i} \bigg|_{x=a} \overset{\text{def}}{=} \lim_{h \to 0} \frac{g(a + he_i) - g(a)}{h},
\]
if this limit exists.

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

**Theorem 1.38** Suppose $f(x)$ is differentiable at $a$ in the sense of Definition 1.36. Define the gradient matrix $\nabla f(a)$ to be the transpose of the Jacobian matrix $J_f(a)$. Then
\[
\nabla f(a) = \left(\begin{array}{c}
\frac{\partial f_1(x)}{\partial x_1} & \cdots & \frac{\partial f_\ell(x)}{\partial x_1} \\
\vdots & \ddots & \vdots \\
\frac{\partial f_1(x)}{\partial x_k} & \cdots & \frac{\partial f_\ell(x)}{\partial x_k}
\end{array}\right) \bigg|_{x=a}.
\]  

(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 $f$ maps $k$-vectors to $\ell$-vectors, $\nabla f(x)$ 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 \( \nabla f(x) \), defined in Equation (1.30),
throughout these notes.

By Definition 1.36, the gradient matrix satisfies the first-order Taylor formula
\[
f(x) = f(a) + \nabla f(a)^\top (x - a) + r(x, a),
\]
where \( r(x, a)/\|x - a\| \to 0 \) as \( x \to a \).

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
\( f(x) \) is real-valued but its argument \( x \) is a vector. To this end, suppose that \( U \subset \mathbb{R}^k \) is open
and that \( f(x) \) maps \( U \) into \( \mathbb{R} \). Then according to Equation (1.30), \( \nabla f(x) \) is a \( k \times 1 \) vector of
partial derivatives, which means that \( \nabla f(x) \) maps \( k \)-vectors to \( k \)-vectors. If we differentiate
once more and evaluate the result at \( a \), denoting the result by \( \nabla^2 f(a) \), then Equation (1.30)
with \( \partial/\partial x_i f(x) \) substituted for \( f_i(x) \) gives
\[
\nabla^2 f(a) = \begin{pmatrix}
\frac{\partial^2 f(x)}{\partial x_1^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_k} \\
\vdots & \ddots & \vdots \\
\frac{\partial^2 f(x)}{\partial x_k \partial x_1} & \cdots & \frac{\partial^2 f(x)}{\partial x_k^2}
\end{pmatrix}
\]
\[
\bigg|_{x=a}.
\]

**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 \( f(x) \) at \( a \).

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 \( f(x) \) is twice differentiable at
some point \( a \in \mathbb{R}^k \). Then \( \nabla^2 f(a) \) is a symmetric matrix, and
\[
f(x) = f(a) + \nabla f(a)^\top (x - a) + \frac{1}{2}(x - a)^\top \nabla^2 f(a)(x - a) + r_2(x, a),
\]
where \( r_2(x, a)/\|x - a\|^2 \to 0 \) as \( x \to a \).
Exercises for Section 1.4

Exercise 1.28  (a) Suppose that \( f(x) \) is continuous at \( 0 \). Prove that \( f(te_i) \) is continuous as a function of \( t \) at \( t = 0 \) for each \( i \), where \( e_i \) is the \( i \)th standard basis vector.

(b) Prove that the converse of (a) is not true by inventing a function \( f(x) \) that is not continuous at \( 0 \) but such that \( f(te_i) \) 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 \( c \), then \( f(a_n) \to f(c) \). 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 \( x = a + te_i \) and let \( t \to 0 \), where \( e_i \) 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

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

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 \( \mu \) and \( \sigma^2 \).

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 \( X_1, \ldots, X_n, \mu, \) and \( \sigma^2 \). Setting the score vector equal to zero and solving for \( \mu \) and \( \sigma^2 \) gives the well-known maximum likelihood estimators of \( \mu \) and \( \sigma^2 \), namely \( \bar{X} \) and \( \frac{1}{n} \sum_i (X_i - \bar{X})^2 \).

Exercise 1.33  Define

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

\[
\begin{array}{cccccc}
2.946 & 0.975 & 1.333 & 4.484 & 1.711 \\
2.627 & -0.628 & 2.476 & 2.599 & 2.143
\end{array}
\]

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(\( \alpha, \beta \)) distribution, for which the density function is

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

where \( \alpha \) and \( \beta \) 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

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

(a) Calculate the score vector for an independent and identically distributed gamma(\( \alpha, \beta \)) 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
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 $\Omega$ 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 $E X$. 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 $E X$ (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 $E X$ 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 $E X = \pm \infty$: 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 $E X^+$ and $E X^-$ are always well-defined if we allow $\infty$ as a possibility, but if both $X^+$ and $X^-$ have infinite expectation, then there is no sensible way to define $E X$. It is easy to find examples of random variables $X$ for which $E X$ 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
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 $\Omega$ and $E X$ and $E Y$ are well-defined.

1. **Linearity:** For any real numbers $a$ and $b$, $E(aX + bY) = aE(X) + bE(Y)$ (and if $aE(X) + bE(Y)$ is undefined, then so is $E(aX + bY)$).

2. **Monotonicity:** If $X(\omega) \leq Y(\omega)$ for all $\omega \in \Omega$, then $E X \leq E Y$.

3. **Conditioning:** If $E(X|Y)$ denotes the conditional expectation of $X$ given $Y$ (which, as a function of $Y$, is itself a random variable), then $E X = E \{E(X|Y)\}$.

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

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

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

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

$$\text{Cov}(X, Y) \overset{\text{def}}{=} E XY - (E X)(E Y)$$

and $\text{Var}(X) \overset{\text{def}}{=} \text{Cov}(X, X)$. The linearity property above extends to random vectors: For scalars $a$ and $b$ we have $E(aX + bY) = aE(X) + bE(Y)$, and for matrices $P$ and $Q$ with dimensions such that $PX + QY$ is well-defined, $E(PX + QY) = PE(X) + QE(Y)$. The covariance between two random vectors is

$$\text{Cov}(X, Y) \overset{\text{def}}{=} E XY^\top - (E X)(E Y)^\top,$$

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

$$\text{Var}(PX) = P \text{Var}(X)P^\top$$

for any constant matrix $P$ with as many columns as $X$ 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

$$|X|^r \geq |X|^r 1\{|X| \geq a\} \geq a^r 1\{|X| \geq a\},$$
where throughout these notes, \( I\{\cdot\} \) denotes the indicator function

\[
I\{\text{expression}\} \overset{\text{def}}{=} \begin{cases} 
1 & \text{if expression is true} \\
0 & \text{if expression is not true.}
\end{cases} \tag{1.34}
\]

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

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

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

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

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

\[
f(E X) \leq E f(X) \tag{1.37}
\]

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 \([E X, f(E X)]\) such that \( g(x) \leq f(x) \) for all \( x \) (see Figure 1.2). By monotonicity, we know that \( E g(X) \leq E f(X) \). We now invoke the linearity of the expectation operator to conclude that

\[
E g(X) = g(E X) = f(E X),
\]

which proves inequality (1.37).

**Exercises for Section 1.5**

33
Figure 1.2: The solid curve is a convex function $f(x)$ and the dotted line is a supporting hyperplane $g(x)$, tangent at $x = E X$. 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

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

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 $p + q = 1$,

$$E |XY| \leq (E |X|^{1/p})^p (E |Y|^{1/q})^q.$$  \hspace{1cm} (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
\[
(E |X|)^\alpha \leq E |X|^\alpha.
\]

**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
\[
S_k = \sum_{i=1}^{k} (X_i - E X_i)
\]
to be the centered \( k \)th partial sum for \( 1 \leq k \leq n \). Then for \( a > 0 \), Kolmogorov’s inequality states that
\[
P \left( \max_{1 \leq k \leq n} |S_k| \geq a \right) \leq \frac{\text{Var} S_n}{a^2}. \tag{1.40}
\]

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

**Hint:** Argue that
\[
\sum_{i=1}^{n} E I \{ A_i \} = P \left( \max_{1 \leq k \leq n} |S_k| \geq a \right)
\]
and \( E \left[ I \{ A_k \} S_k^2 \right] \geq a^2 E I \{ A_k \} \).

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

$$E \{ I\{A_k\} S_k (S_n - S_k) \} = E \{ I\{A_k\} S_k \} E (S_n - S_k).$$

What is $E (S_n - S_k)$?

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 $E X_i = 0$ and $\text{Var} X_i = 9 - i$. 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 $x + iy$, where $x$ and $y$ are real numbers and $i = \sqrt{-1}$. The elegant result known as Euler’s formula relates the points on the unit circle to the complex exponential function:

$$\exp\{it\} = \cos t + i \sin t \quad \text{for all } t \in \mathbb{R}. \quad (1.41)$$

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

$$\left| \int_{0}^{t} g(x) e^{ix} \, dx \right| \leq \left| \int_{0}^{t} g(x) \, dx \right| = \left| \int_{0}^{t} |g(x)| \, dx \right|. \quad (1.42)$$

The inequalities below in parts (a) through (d) involving $\exp\{it\}$ 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 $e^{it}$ 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 $t \in \mathbb{R}$,

$$|\exp\{it\} - 1| \leq |t|. $$

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

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

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

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

(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

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

Exercise 1.44 Refer to Exercise 1.43. Graph the functions $|\exp\{it\} - 1 - it + \frac{1}{2} t^2|$, $|t|^3/6$, 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 $z = x + iy$ equals $\sqrt{x^2 + y^2}$. Refer to Equation (1.41) to deal with the expression $\exp\{it\}$.

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

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

(1.43)

Hint: First, prove that equality holds if $Y$ is supported on the nonnegative integers. Then note for a general $Y$ that $E \lfloor Y \rfloor \leq E Y$, 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,

$$\int_0^{\infty} P(Y \geq t) \, dt = E Y. $$
(This equation remains true if $E Y = \infty$.) To sketch a proof, note that if we can prove $\int E f(Y, t) \, dt = E \int f(Y, t) \, dt$, the result follows immediately by taking $f(Y, t) = I\{Y \geq t\}$.