Chapter 4

Sampling Distributions and Limits

CHAPTER OUTLINE

Section 1 Sampling Distributions
Section 2 Convergence in Probability
Section 3 Convergence with Probability 1
Section 4 Convergence in Distribution
Section 5 Monte Carlo Approximations
Section 6 Normal Distribution Theory
Section 7 Further Proofs (Advanced)

In many applications of probability theory, we will be faced with the following problem. Suppose that $X_1, X_2, \ldots, X_n$ is an identically and independently distributed (i.i.d.) sequence, i.e., $X_1, X_2, \ldots, X_n$ is a sample from some distribution, and we are interested in the distribution of a new random variable $Y = h(X_1, X_2, \ldots, X_n)$ for some function h. In particular, we might want to compute the distribution function of Y or perhaps its mean and variance. The distribution of Y is sometimes referred to as its *sampling distribution*, as Y is based on a sample from some underlying distribution.

We will see that some of the methods developed in earlier chapters are useful in solving such problems — especially when it is possible to compute an exact solution, e.g., obtain an exact expression for the probability or density function of Y. Section 4.6 contains a number of exact distribution results for a variety of functions of normal random variables. These have important applications in statistics.

Quite often, however, exact results are impossible to obtain, as the problem is just too complex. In such cases, we must develop an approximation to the distribution of Y.

For many important problems, a version of Y is defined for each sample size n (e.g., a sample mean or sample variance), so that we can consider a sequence of random variables $Y_1, Y_2, \ldots$, etc. This leads us to consider the limiting distribution of such a sequence so that, when n is large, we can approximate the distribution of Y_n by the

limit, which is often much simpler. This approach leads to a famous result, known as the central limit theorem, discussed in Section 4.4.

Sometimes we cannot even develop useful approximations for large n, due to the difficulty of the problem or perhaps because n is just too small in a particular application. Fortunately, however, we can then use the Monte Carlo approach where the power of the computer becomes available. This is discussed in Section 4.5.

In Chapter 5 we will see that, in statistical applications, we typically do not know much about the underlying distribution of the X_i from which we are sampling. We then collect a sample and a value, such as Y, that will serve as an estimate of a characteristic of the underlying distribution, e.g., the sample mean $\bar{X}$ will serve as an estimate of the mean of the distribution of the X_i. We then want to know what happens to these estimates as n grows. If we have chosen our estimates well, then the estimates will converge to the quantities we are estimating as n increases. Such an estimate is called *consistent*. In Sections 4.2 and 4.3, we will discuss the most important consistency theorems — namely, the weak and strong laws of large numbers.

4.1 Sampling Distributions

Let us consider a very simple example.

EXAMPLE 4.1.1

Suppose we obtain a sample X_1, X_2 of size $n = 2$ from the discrete distribution with probability function given by

$$p_X(x) = \begin{cases} 1/2 & x = 1 \\ 1/4 & x = 2 \\ 1/4 & x = 3 \\ 0 & \text{otherwise.} \end{cases}$$

Let us take $Y_2 = (X_1 X_2)^{1/2}$. This is the *geometric mean* of the sample values (the geometric mean of n positive numbers $x_1, \ldots, x_n$ is defined as $(x_1 \cdots x_n)^{1/n}$).

To determine the distribution of Y_2, we first list the possible values for Y_2, the samples that give rise to these values, and their probabilities of occurrence. The values of these probabilities specify the sampling distribution of Y. We have the following table.

y	Sample	$p_{Y_2}(y)$
1	$\{(1,1)\}$	$(1/2)(1/2) = 1/4$
$\sqrt{2}$	$\{(1,2),(2,1)\}$	$(1/2)(1/4) + (1/4)(1/2) = 1/4$
$\sqrt{3}$	$\{(1,3),(1,3)\}$	$(1/2)(1/4) + (1/4)(1/2) = 1/4$
2	$\{(2,2)\}$	$(1/4)(1/4) = 1/16$
$\sqrt{6}$	$\{(2,3),(3,2)\}$	$(1/4)(1/4) + (1/4)(1/4) = 1/8$
3	$\{(3,3)\}$	$(1/4)(1/4) = 1/16$

Now suppose instead we have a sample $X_1, \ldots, X_{20}$ of size $n = 20$, and we want to find the distribution of $Y_{20} = (X_1 \cdots X_{20})^{1/20}$. Obviously, we can proceed as above, but this time the computations are much more complicated, as there are now $3^{20} = 3{,}486{,}784{,}401$ possible samples, as opposed to the $3^2 = 9$ samples used to form the

previous table. Directly computing $p_{Y_{20}}$, as we have done for p_{Y_2}, would be onerous — even for a computer! So what can we do here?

One possibility is to look at the distribution of $Y_n = (X_1 \cdots X_n)^{1/n}$ when n is large and see if we can approximate this in some fashion. The results of Section 4.4.1 show that

$$\ln Y_n = \frac{1}{n} \sum_{i=1}^{n} \ln X_i$$

has an approximate normal distribution when n is large. In fact, the approximating normal distribution when $n = 20$ turns out to be an $N(0.447940, 0.105167)$ distribution. We have plotted this density in Figure 4.1.1.

Another approach is to use the methods of Section 2.10 to generate N samples of size $n = 20$ from p_X, calculate $\ln Y_{20}$ for each (ln is a 1-1 transformation, and we transform to avoid the potentially large values assumed by Y_{20}), and then use these N values to approximate the distribution of $\ln Y_{20}$. For example, in Figure 4.1.2 we have provided a plot of a density histogram (see Section 5.4.3 for more discussion of histograms) of $N = 10^4$ values of $\ln Y_{20}$ calculated from $N = 10^4$ samples of size $n = 20$ generated (using the computer) from p_X. The area of each rectangle corresponds to the proportion of values of $\ln Y_{20}$ that were in the interval given by the base of the rectangle. As we will see in Sections 4.2, 4.3, and 4.4, these areas approximate the actual probabilities that $\ln Y_{20}$ falls in these intervals. These approximations improve as we increase N.

Notice the similarity in the shapes of Figures 4.1.1 and 4.1.2. Figure 4.1.2 is not symmetrical about its center, however, as it is somewhat skewed. This is an indication that the normal approximation is not entirely adequate when $n = 20$. ∎

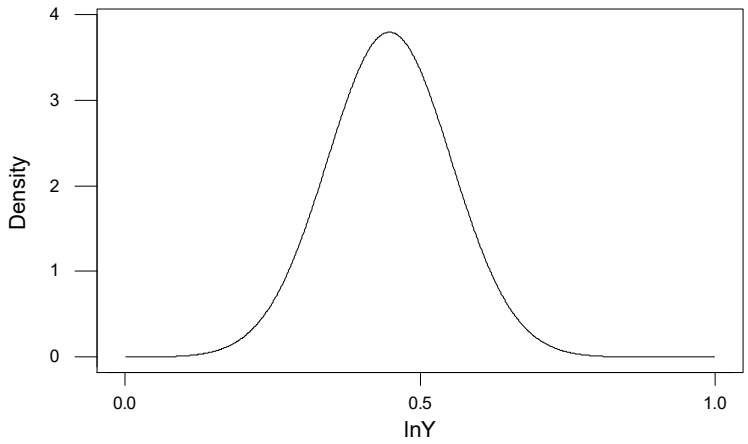

Figure 4.1.1: Plot of the approximating $N(0.447940, 0.105167)$ density to the distribution of $\ln Y_{20}$ in Example 4.1.1.

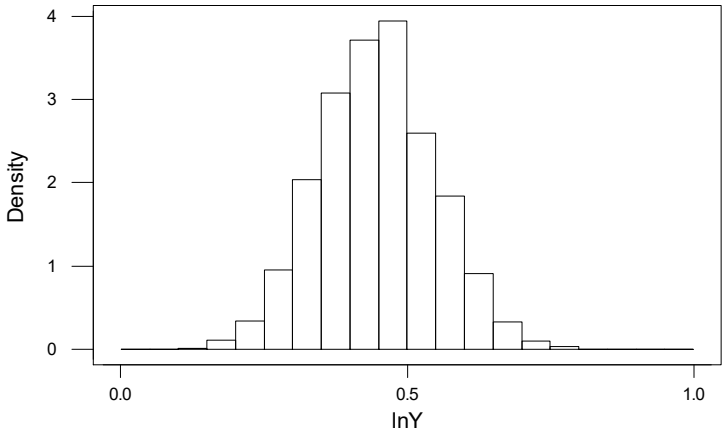

Figure 4.1.2: Plot of $N = 10^4$ values of $\ln Y_{20}$ obtained by generating $N = 10^4$ samples from p_X in Example 4.1.1.

Sometimes we are lucky and can work out the sampling distribution of

$$Y = h(X_1, X_2, \ldots, X_n)$$

exactly in a form useful for computing probabilities and expectations for Y. In general, however, when we want to compute $P(Y \in B) = P_Y(B)$, we will have to determine the set of samples $(X_1, X_2, \ldots, X_n)$ such that $Y \in B$, as given by

$$h^{-1}B = \{(x_1, x_2, \ldots, x_n) : h(x_1, x_2, \ldots, x_n) \in B\},$$

and then compute $P((X_1, X_2, \ldots, X_n) \in h^{-1}B)$. This is typically an intractable problem and approximations or simulation (Monte Carlo) methods will be essential. Techniques for deriving such approximations will be discussed in subsequent sections of this chapter. In particular, we will develop an important approximation to the sampling distribution of the sample mean

$$\bar{X} = h(X_1, X_2, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Summary of Section 4.1

- A sampling distribution is the distribution of a random variable corresponding to a function of some i.i.d. sequence.
- Sampling distributions can sometimes be computed by direct computation or by approximations such as the central limit theorem.

EXERCISES

4.1.1 Suppose that X_1, X_2, X_3 are i.i.d. from p_X in Example 4.1.1. Determine the exact distribution of $Y_3 = (X_1 X_2 X_3)^{1/3}$.

4.1.2 Suppose that a fair six-sided die is tossed $n = 2$ independent times. Compute the exact distribution of the sample mean.

4.1.3 Suppose that an urn contains a proportion p of chips labelled 0 and proportion $1 - p$ of chips labelled 1. For a sample of $n = 2$, drawn with replacement, determine the distribution of the sample mean.

4.1.4 Suppose that an urn contains N chips labelled 0 and M chips labelled 1. For a sample of $n = 2$, drawn without replacement, determine the distribution of the sample mean.

4.1.5 Suppose that a symmetrical die is tossed $n = 20$ independent times. Work out the exact sampling distribution of the maximum of this sample.

4.1.6 Suppose three fair dice are rolled, and let Y be the number of 6's showing. Compute the exact distribution of Y.

4.1.7 Suppose two fair dice are rolled, and let W be the *product* of the two numbers showing. Compute the exact distribution of W.

4.1.8 Suppose two fair dice are rolled, and let Z be the *difference* of the two numbers showing (i.e., the first number *minus* the second number). Compute the exact distribution of Z.

4.1.9 Suppose four fair coins are flipped, and let Y be the number of pairs of coins which land the same way (i.e., the number of pairs that are either both heads or both tails). Compute the exact distribution of Y.

COMPUTER EXERCISES

4.1.10 Generate a sample of $N = 10^3$ values of Y_{50} in Example 4.1.1. Calculate the mean and standard deviation of this sample.

4.1.11 Suppose that $X_1, X_2, \ldots, X_{10}$ is an i.i.d. sequence from an $N(0, 1)$ distribution. Generate a sample of $N = 10^3$ values from the distribution of $\max(X_1, X_2, \ldots, X_{10})$. Calculate the mean and standard deviation of this sample.

PROBLEMS

4.1.12 Suppose that $X_1, X_2, \ldots, X_n$ is a sample from the Poisson(λ) distribution. Determine the exact sampling distribution of $Y = X_1 + X_2 + \cdots + X_n$. (Hint: Determine the moment-generating function of Y and use the uniqueness theorem.)

4.1.13 Suppose that X_1, X_2 is a sample from the Uniform[0,1] distribution. Determine the exact sampling distribution of $Y = X_1 + X_2$. (Hint: Determine the density of Y.)

4.1.14 Suppose that X_1, X_2 is a sample from the Uniform[0,1] distribution. Determine the exact sampling distribution of $Y = (X_1 X_2)^{1/2}$. (Hint: Determine the density of $\ln Y$ and then transform.)

4.2 Convergence in Probability

Notions of *convergence* are fundamental to much of mathematics. For example, if $a_n = 1 - 1/n$, then $a_1 = 0$, $a_2 = 1/2$, $a_3 = 2/3$, $a_4 = 3/4$, etc. We see that the values of a_n are getting "closer and closer" to 1, and indeed we know from calculus that $\lim_{n\to\infty} a_n = 1$ in this case.

For random variables, notions of convergence are more complicated. If the values themselves are random, then how can they "converge" to anything? On the other hand, we can consider various *probabilities* associated with the random variables and see if *they* converge in some sense.

The simplest notion of convergence of random variables is convergence in probability, as follows. (Other notions of convergence will be developed in subsequent sections.)

Definition 4.2.1 Let $X_1, X_2, \ldots$ be an infinite sequence of random variables, and let Y be another random variable. Then the sequence $\{X_n\}$ *converges in probability* to Y, if for all $\epsilon > 0$, $\lim_{n\to\infty} P(|X_n - Y| \geq \epsilon) = 0$, and we write $X_n \xrightarrow{P} Y$.

In Figure 4.2.1, we have plotted the differences $X_n - Y$, for selected values of n, for 10 generated sequences $\{X_n - Y\}$ for a typical situation where the random variables X_n converge to a random variable Y in probability. We have also plotted the horizontal lines at $\pm\epsilon$ for $\epsilon = 0.25$. From this we can see the increasing concentration of the distribution of $X_n - Y$ about 0, as n increases, as required by Definition 4.2.1. In fact, the 10 observed values of $X_{100} - Y$ all satisfy the inequality $|X_{100} - Y| < 0.25$.

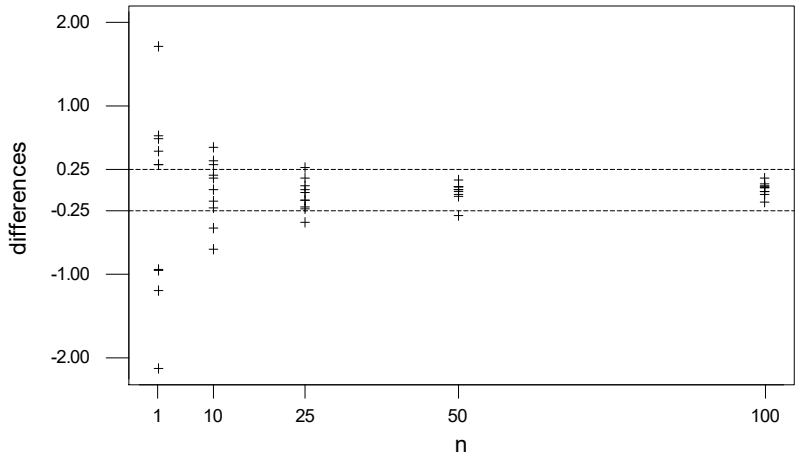

Figure 4.2.1: Plot of 10 replications of $\{X_n - Y\}$ illustrating the convergence in probability of X_n to Y.

We consider some applications of this definition.

EXAMPLE 4.2.1
Let Y be any random variable, and let $X_1 = X_2 = X_3 = \cdots = Y$. (That is, the random variables are all *identical* to each other.) In that case, $|X_n - Y| = 0$, so of course

$$\lim_{n\to\infty} P(|X_n - Y| \geq \epsilon) = 0$$

for all $\epsilon > 0$. Hence, $X_n \xrightarrow{P} Y$. ∎

EXAMPLE 4.2.2
Suppose $P(X_n = 1 - 1/n) = 1$ and $P(Y = 1) = 1$. Then $P(|X_n - Y| \geq \epsilon) = 0$ whenever $n > 1/\epsilon$. Hence, $P(|X_n - Y| \geq \epsilon) \to 0$ as $n \to \infty$ for all $\epsilon > 0$. Hence, the sequence $\{X_n\}$ converges in probability to Y. (Here, the distributions of X_n and Y are all *degenerate*.) ∎

EXAMPLE 4.2.3
Let $U \sim$ Uniform[0, 1]. Define X_n by

$$X_n = \begin{cases} 3 & U \leq \frac{2}{3} - \frac{1}{n} \\ 8 & \text{otherwise,} \end{cases}$$

and define Y by

$$Y = \begin{cases} 3 & U \leq \frac{2}{3} \\ 8 & \text{otherwise.} \end{cases}$$

Then

$$P(|X_n - Y| \geq \epsilon) \leq P(X_n \neq Y) = P\left(\frac{2}{3} - \frac{1}{n} < U \leq \frac{2}{3}\right) = \frac{1}{n}.$$

Hence, $P(|X_n - Y| \geq \epsilon) \to 0$ as $n \to \infty$ for all $\epsilon > 0$, and the sequence $\{X_n\}$ converges in probability to Y. (This time, the distributions of X_n and Y are *not* degenerate.) ∎

A common case is where the distributions of the X_n are not degenerate, but Y is just a constant, as in the following example.

EXAMPLE 4.2.4
Suppose $Z_n \sim$ Exponential(n) and let $Y = 0$. Then

$$P(|Z_n - Y| \geq \epsilon) = P(Z_n \geq \epsilon) = \int_\epsilon^\infty ne^{-nx}dx = e^{-n\epsilon}.$$

Hence, again $P(|Z_n - Y| \geq \epsilon) \to 0$ as $n \to \infty$ for all $\epsilon > 0$, so the sequence $\{Z_n\}$ converges in probability to Y. ∎

4.2.1 The Weak Law of Large Numbers

One of the most important applications of convergence in probability is the weak law of large numbers. Suppose $X_1, X_2, \ldots$ is a sequence of independent random variables that each have the same mean μ. For large n, what can we say about their average

$$M_n = \frac{1}{n}(X_1 + \cdots + X_n)?$$

We refer to M_n as the *sample average*, or *sample mean*, for $X_1, \ldots, X_n$. When the sample size n is fixed, we will often use $\bar{X}$ as a notation for sample mean instead of M_n.

For example, if we flip a sequence of fair coins, and if $X_i = 1$ or $X_i = 0$ as the ith coin comes up heads or tails, then M_n represents the *fraction* of the first n coins that came up heads. We might expect that for large n, this fraction will be close to $1/2$, i.e., to the expected value of the X_i.

The weak law of large numbers provides a precise sense in which average values M_n tend to get close to $E(X_i)$, for large n.

Theorem 4.2.1 (*Weak law of large numbers*) Let $X_1, X_2, \ldots$ be a sequence of independent random variables, each having the same mean μ and each having variance less than or equal to $v < \infty$. Then for all $\epsilon > 0$, $\lim_{n\to\infty} P(|M_n - \mu| \geq \epsilon) = 0$. That is, the averages converge in probability to the common mean μ or $M_n \xrightarrow{P} \mu$.

PROOF Using linearity of expected value, we see that $E(M_n) = \mu$. Also, using independence, we have

$$\begin{aligned} \text{Var}(M_n) &= \frac{1}{n^2}(\text{Var}(X_1) + \text{Var}(X_2) + \cdots + \text{Var}(X_n)) \\ &\leq \frac{1}{n^2}(v + v + \cdots + v) = \frac{1}{n^2}(nv) = v/n\,. \end{aligned}$$

Hence, by Chebychev's inequality (Theorem 3.6.2), we have

$$P(|M_n - \mu| \geq \epsilon) \leq \text{Var}(M_n)/\epsilon^2 \leq v/\epsilon^2 n.$$

This converges to 0 as $n \to \infty$, which proves the theorem. ∎

It is a fact that, in Theorem 4.2.1, if we require the X_i variables to be i.i.d. instead of merely independent, then we do not even need the X_i to have finite variance. But we will not discuss this result further here. Consider some applications of the weak law of large numbers.

EXAMPLE 4.2.5

Consider flipping a sequence of identical fair coins. Let M_n be the fraction of the first n coins that are heads. Then $M_n = (X_1 + \cdots + X_n)/n$, where $X_i = 1$ if the ith coin is heads, otherwise $X_i = 0$. Hence, by the weak law of large numbers, we have

$$\begin{aligned} \lim_{n\to\infty} P(M_n < 0.49) &= \lim_{n\to\infty} P(M_n - 0.5 < -0.01) \\ &\leq \lim_{n\to\infty} P(M_n - 0.5 < -0.01 \text{ or } M_n - 0.5 > 0.01\ ) \\ &= \lim_{n\to\infty} P(|M_n - 0.5| > 0.01) = 0 \end{aligned}$$

and, similarly, $\lim_{n\to\infty} P(M_n > 0.51) = 0$. This illustrates that for large n, it is very likely that M_n is very close to 0.5. ∎

EXAMPLE 4.2.6
Consider flipping a sequence of identical coins, each of which has probability p of coming up heads. Let M_n again be the fraction of the first n coins that are heads. Then by the weak law of large numbers, for any $\epsilon > 0$, $\lim_{n\to\infty} P(p-\epsilon < M_n < p+\epsilon) = 1$. We thus see that for large n, it is very likely that M_n is very close to p. (The previous example corresponds to the special case $p = 1/2$.) ∎

EXAMPLE 4.2.7
Let $X_1, X_2, \ldots$ be i.i.d. with distribution $N(3, 5)$. Then $E(M_n) = 3$, and by the weak law of large numbers, $P(3 - \epsilon < M_n < 3 + \epsilon) \to 1$ as $n \to \infty$. Hence, for large n, the average value M_n is very close to 3. ∎

EXAMPLE 4.2.8
Let $W_1, W_2, \ldots$ be i.i.d. with distribution Exponential(6). Then $E(M_n) = 1/6$, and by the weak law of large numbers, $P(1/6 - \epsilon < M_n < 1/6 + \epsilon) \to 1$ as $n \to \infty$. Hence, for large n, the average value M_n is very close to $1/6$. ∎

Summary of Section 4.2

- A sequence $\{X_n\}$ of random variables converges in probability to Y if
$$\lim_{n\to\infty} P(|X_n - Y| \geq \epsilon) = 0.$$
- The weak law of large numbers says that if $\{X_n\}$ is i.i.d. (or is independent with constant mean and bounded variance), then the averages $M_n = (X_1 + \cdots + X_n)/n$ converge in probability to $E(X_i)$.

EXERCISES

4.2.1 Let $U \sim$ Uniform[5, 10], and let $Z = I_{U\in[5,\,7)}$ and $Z_n = I_{U\in[5,\,7+1/n^2)}$. Prove that $Z_n \to Z$ in probability.

4.2.2 Let $Y \sim$ Uniform[0, 1], and let $X_n = Y^n$. Prove that $X_n \to 0$ in probability.

4.2.3 Let $W_1, W_2, \ldots$ be i.i.d. with distribution Exponential(3). Prove that for some n, we have $P(W_1 + W_2 + \cdots + W_n < n/2) > 0.999$.

4.2.4 Let $Y_1, Y_2, \ldots$ be i.i.d. with distribution $N(2, 5)$. Prove that for some n, we have $P(Y_1 + Y_2 + \cdots + Y_n > n) > 0.999$.

4.2.5 Let $X_1, X_2, \ldots$ be i.i.d. with distribution Poisson(8). Prove that for some n, we have $P(X_1 + X_2 + \cdots + X_n > 9n) < 0.001$.

4.2.6 Suppose $X \sim$ Uniform[0, 1], and let $Y_n = \frac{n-1}{n}X$. Prove that $Y_n \xrightarrow{P} X$.

4.2.7 Let H_n be the number of heads when flipping n fair coins, let $X_n = e^{-H_n}$, and let $Y = 0$. Prove that $X_n \xrightarrow{P} Y$.

4.2.8 Let $Z_n \sim$ Uniform[0, n], let $W_n = 5Z_n/(Z_n + 1)$, and let $W = 5$. Prove that $W_n \xrightarrow{P} W$.

4.2.9 Consider flipping n fair coins. Let H_n be the total number of heads, and let F_n be the number of heads on coins 1 through $n - 1$ (i.e., omitting the nth coin). Let $X_n = H_n/(H_n + 1)$, and $Y_n = F_n/(H_n + 1)$, and $Z = 0$. Prove that $X_n - Y_n \xrightarrow{P} Z$.

4.2.10 Let Z_n be the sum of the *squares* of the numbers showing when we roll n fair dice. Find (with proof) a number m such that $\frac{1}{n}Z_n \xrightarrow{P} m$. (Hint: Use the weak law of large numbers.)

4.2.11 Consider flipping n fair nickels and n fair dimes. Let X_n equal 4 times the number of nickels showing heads, plus 5 times the number of dimes showing heads. Find (with proof) a number r such that $\frac{1}{n}X_n \xrightarrow{P} r$.

COMPUTER EXERCISES

4.2.12 Generate i.i.d. $X_1, \ldots, X_n$ distributed Exponential(5) and compute M_n when $n = 20$. Repeat this N times, where N is large (if possible, take $N = 10^5$, otherwise as large as is feasible), and compute the proportion of values of M_n that lie between 0.19 and 0.21. Repeat this with $n = 50$. What property of convergence in probability do your results illustrate?

4.2.13 Generate i.i.d. $X_1, \ldots, X_n$ distributed Poisson(7) and compute M_n when $n = 20$. Repeat this N times, where N is large (if possible, take $N = 10^5$, otherwise as large as is feasible), and compute the proportion of values of M_n that lie between 6.99 and 7.01. Repeat this with $n = 100$. What property of convergence in probability do your results illustrate?

PROBLEMS

4.2.14 Give an example of random variables $X_1, X_2, \ldots$ such that $\{X_n\}$ converges to 0 in probability, but $E(X_n) = 1$ for all n. (Hint: Suppose $P(X_n = n) = 1/n$ and $P(X_n = 0) = 1 - 1/n$.)

4.2.15 Prove that $X_n \xrightarrow{P} 0$ if and only if $|X_n| \xrightarrow{P} 0$.

4.2.16 Prove or disprove that $X_n \xrightarrow{P} 5$ if and only if $|X_n| \xrightarrow{P} 5$.

4.2.17 Suppose $X_n \xrightarrow{P} X$, and $Y_n \xrightarrow{P} Y$. Let $Z_n = X_n + Y_n$ and $Z = X + Y$. Prove that $Z_n \xrightarrow{P} Z$.

CHALLENGES

4.2.18 Suppose $X_n \xrightarrow{P} X$, and f is a continuous function. Prove that $f(X_n) \xrightarrow{P} f(X)$.

4.3 Convergence with Probability 1

A notion of convergence for random variables that is closely associated with the convergence of a sequence of real numbers is provided by the concept of convergence with probability 1. This property is given in the following definition.

Definition 4.3.1 Let $X_1, X_2, \ldots$ be an infinite sequence of random variables. We shall say that the sequence $\{X_i\}$ *converges with probability 1* (or *converges almost surely (a.s.)*) to a random variable Y, if $P(\lim_{n\to\infty} X_n = Y) = 1$ and we write $X_n \xrightarrow{a.s.} Y$.

In Figure 4.3.1, we illustrate this convergence by graphing the sequence of differences $\{X_n - Y\}$ for a typical situation where the random variables X_n converge to a random variable Y with probability 1. We have also plotted the horizontal lines at $\pm\epsilon$ for $\epsilon = 0.1$. Notice that inevitably all the values $X_n - Y$ are in the interval $(-0.1, 0.1)$ or, in other words, the values of X_n are within 0.1 of the values of Y.

Definition 4.3.1 indicates that for any given $\epsilon > 0$, there will exist a value N_ϵ such that $|X_n - Y| < \epsilon$ for every $n \geq N_\epsilon$. The value of N_ϵ will vary depending on the observed value of the sequence $\{X_n - Y\}$, but it always exists. Contrast this with the situation depicted in Figure 4.2.1, which only says that the probability distribution $X_n - Y$ concentrates about 0 as n grows and not that the individual values of $X_n - Y$ will necessarily all be near 0 (also see Example 4.3.2).

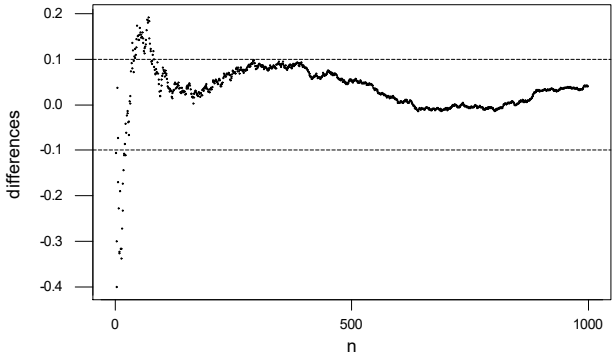

Figure 4.3.1: Plot of a single replication $\{X_n - Y\}$ illustrating the convergence with probability 1 of X_n to Y.

Consider an example of this.

EXAMPLE 4.3.1

Consider again the setup of Example 4.2.3, where $U \sim$ Uniform[0, 1],

$$X_n = \begin{cases} 3 & U \leq \frac{2}{3} - \frac{1}{n} \\ 8 & \text{otherwise} \end{cases}$$

and

$$Y = \begin{cases} 3 & U \leq \frac{2}{3} \\ 8 & \text{otherwise.} \end{cases}$$

If $U > 2/3$, then $Y = 8$ and also $X_n = 8$ for all n, so clearly $X_n \to Y$. If $U < 2/3$, then for large enough n we will also have

$$U \leq \frac{2}{3} - \frac{1}{n},$$

so again $X_n \to Y$. On the other hand, if $U = 2/3$, then we will always have $X_n = 8$, even though $Y = 3$. Hence, $X_n \to Y$ except when $U = 2/3$. Because $P(U = 2/3) = 0$, we do have $X_n \to Y$ with probability 1. ∎

One might wonder what the relationship is between convergence in probability and convergence with probability 1. The following theorem provides an answer.

Theorem 4.3.1 Let $Z, Z_1, Z_2, \ldots$ be random variables. Suppose $Z_n \to Z$ with probability 1. Then $Z_n \to Z$ in probability. That is, if a sequence of random variables converges almost surely, then it converges in probability to the same limit.

PROOF See Section 4.7 for the proof of this result. ∎

On the other hand, the converse to Theorem 4.3.1 is false, as the following example shows.

EXAMPLE 4.3.2
Let U have the uniform distribution on $[0, 1]$. We construct an infinite sequence of random variables $\{X_n\}$ by setting

$$\begin{aligned} X_1 &= I_{[0,1/2)}(U),\ X_2 = I_{[1/2,1]}(U), \\ X_3 &= I_{[0,1/4)}(U),\ X_4 = I_{[1/4,1/2)}(U),\ X_5 = I_{[1/2,3/4)}(U),\ X_6 = I_{[3/4,1]}(U), \\ X_7 &= I_{[0,1/8)}(U), X_8 = I_{[1/8,1/4)}(U), \ldots \\ &\vdots \end{aligned}$$

where I_A is the *indicator function* of the event A, i.e., $I_A(s) = 1$ if $s \in A$, and $I_A(s) = 0$ if $s \notin A$.

Note that we first subdivided $[0, 1]$ into two equal-length subintervals and defined X_1 and X_2 as the indicator functions for the two subintervals. Next we subdivided $[0, 1]$ into four equal-length subintervals and defined X_3, X_4, X_5, and X_6 as the indicator functions for the four subintervals. We continued this process by next dividing $[0, 1]$ into eight equal-length subintervals, then 16 equal-length subintervals, etc., to obtain an infinite sequence of random variables.

Each of these random variables X_n takes the values 0 and 1 only and so must follow a Bernoulli distribution. In particular, $X_1 \sim$ Bernoulli$(1/2)$, $X_2 \sim$ Bernoulli$(1/2)$, $X_3 \sim$ Bernoulli$(1/4)$, etc.

Then for $0 < \epsilon < 1$, we have that $P(|X_n - 0| \geq \epsilon) = P(X_n = 1)$. Because the intervals for U that make $X_n \neq 0$ are getting smaller and smaller, we see that $P(X_n = 1)$ is converging to 0. Hence, X_n converges to 0 in probability.

On the other hand, X_n does *not* converge to 0 almost surely. Indeed, no matter what value U takes on, there will always be infinitely many different n for which $X_n = 1$. Hence, we will have $X_n = 1$ infinitely often, so that we will *not* have X_n converging to 0 for any particular value of U. Thus, $P(\lim_{n\to\infty} X_n \to 0) = 0$, and X_n does *not* converge to 0 with probability 1. ∎

Theorem 4.3.1 and Example 4.3.2 together show that convergence with probability 1 is a *stronger* notion than convergence in probability.

Now, the weak law of large numbers (Section 4.2.1) concludes only that the averages M_n are converging in probability to $E(X_i)$. A stronger version of this result would instead conclude convergence with probability 1. We consider that now.

4.3.1 The Strong Law of Large Numbers

The following is a strengthening of the weak law of large numbers because it concludes convergence with probability 1 instead of just convergence in probability.

Theorem 4.3.2 (*Strong law of large numbers*) Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables, each having finite mean μ. Then

$$P\left(\lim_{n\to\infty} M_n = \mu\right) = 1.$$

That is, the averages converge with probability 1 to the common mean μ or $M_n \stackrel{a.s.}{\to} \mu$.

PROOF See *A First Look at Rigorous Probability Theory, Second Edition*, by J. S. Rosenthal (World Scientific Publishing Co., 2006) for a proof of this result. ∎

This result says that sample averages converge with probability 1 to μ.

Like Theorem 4.2.1, it says that for large n the averages M_n are usually close to $\mu = E(X_i)$ for large n. But it says in addition that if we wait long enough (i.e., if n is large enough), then eventually the averages will *all* be close to μ, for *all* sufficiently large n. In other words, the sample mean is consistent for μ.

Summary of Section 4.3

- A sequence $\{X_n\}$ of random variables converges with probability 1 (or converges almost surely) to Y if, $P(\lim_{n\to\infty} X_n = Y) = 1$.
- Convergence with probability 1 implies convergence in probability.
- The strong law of large numbers says that if $\{X_n\}$ is i.i.d., then the averages $M_n = (X_1 + \cdots + X_n)/n$ converge with probability 1 to $E(X_i)$.

EXERCISES

4.3.1 Let $U \sim$ Uniform[5, 10], and let $Z = I_{[5,\,7)}(U)$ (i.e., Z is the indicator function of $[5,\ 7)$) and $Z_n = I_{[5,\,7+1/n^2)}(U)$. Prove that $Z_n \to Z$ with probability 1.

4.3.2 Let $Y \sim$ Uniform[0, 1], and let $X_n = Y^n$. Prove that $X_n \to 0$ with probability 1.

4.3.3 Let $W_1, W_2, \ldots$ be i.i.d. with distribution Exponential(3). Prove that with probability 1, for some n, we have $W_1 + W_2 + \cdots + W_n < n/2$.

4.3.4 Let $Y_1, Y_2, \ldots$ be i.i.d. with distribution $N(2, 5)$. Prove that with probability 1, for some n, we have $Y_1 + Y_2 + \cdots + Y_n > n$.

4.3.5 Suppose $X_n \to X$ with probability 1, and also $Y_n \to Y$ with probability 1. Prove that $P(X_n \to X \text{ and } Y_n \to Y) = 1$.

4.3.6 Suppose $Z_1, Z_2, \ldots$ are i.i.d. with finite mean μ. Let $M_n = (Z_1 + \cdots + Z_n)/n$. Determine (with explanation) whether the following statements are true or false.

(a) With probability 1, $M_n = \mu$ for some n.
(b) With probability 1, $\mu - 0.01 < M_n < \mu + 0.01$ for some n.
(c) With probability 1, $\mu - 0.01 < M_n < \mu + 0.01$ for all but finitely many n.
(d) For any $x \in R^1$, with probability 1, $x - 0.01 < M_n < x + 0.01$ for some n.

4.3.7 Let $\{X_n\}$ be i.i.d., with $X_n \sim$ Uniform[3, 7]. Let $Y_n = (X_1 + X_2 + \ldots + X_n)/n$. Find (with proof) a number m such that $Y_n \overset{a.s.}{\to} m$. (Hint: Use the strong law of large numbers.)

4.3.8 Let Z_n be the sum of the squares of the numbers showing when we roll n fair dice. Find (with proof) a number m such that $\frac{1}{n} Z_n \overset{a.s.}{\to} m$.

4.3.9 Consider flipping n fair nickels and n fair dimes. Let X_n equal 4 times the number of nickels showing heads, plus 5 times the number of dimes showing heads. Find (with proof) a number r such that $\frac{1}{n} X_n \overset{a.s.}{\to} r$.

4.3.10 Suppose $Y_n \overset{a.s.}{\to} Y$. Does this imply that $P(|Y_5 - Y| > |Y_4 - Y|) = 0$? Explain.

4.3.11 Consider repeatedly flipping a fair coin. Let H_n be the number of heads on the first n flips, and let $Z_n = H_n/n$.

(a) Prove that, with probability 1, there is some m such that $|Z_n - 1/2| < 0.001$ for all $n \geq m$.

(b) Let r be the smallest positive integer satisfying $|Z_r - 1/2| < 0.001$. Must we have $|Z_n - 1/2| < 0.001$ for all $n \geq r$? Why or why not?

4.3.12 Suppose $P(X = 0) = P(X = 1) = 1/2$, and let $X_n = X$ for $n = 1, 2, 3, \ldots$. (That is, the random variables X_n are all *identical*.) Let $Y_n = (X_1 + X_2 + \ldots + X_n)/n$.

(a) Prove that $P(\lim_{n\to\infty} Y_n = 0) = P(\lim_{n\to\infty} Y_n = 1) = 1/2$.

(b) Prove that there is no number m such that $P(\lim_{n\to\infty} Y_n = m) = 1$.

(c) Why does part (b) not contradict the law of large numbers?

COMPUTER EXERCISES

4.3.13 Generate i.i.d. $X_1, \ldots, X_n$ distributed Exponential(5) with n large (take $n = 10^5$ if possible). Plot the values $M_1, M_2, \ldots, M_n$. To what value are they converging? How quickly?

4.3.14 Generate i.i.d. $X_1, \ldots, X_n$ distributed Poisson(7) with n large (take $n = 10^5$ if possible). Plot the values $M_1, M_2, \ldots, M_n$. To what value are they converging? How quickly?

4.3.15 Generate i.i.d. $X_1, X_2, \ldots, X_n$ distributed $N(-4, 3)$ with n large (take $n = 10^5$ if possible). Plot the values $M_1, M_2, \ldots, M_n$. To what value are they converging? How quickly?

PROBLEMS

4.3.16 Suppose for each positive integer k, there are random variables W_k, $X_{k,1}$, $X_{k,2}$, ... such that $P(\lim_{n\to\infty} X_{k,n} = W_k) = 1$. Prove that $P(\lim_{n\to\infty} X_{k,n} = W_k \text{ for all } k) = 1$.

4.3.17 Prove that $X_n \overset{a.s.}{\to} 0$ if and only if $|X_n| \overset{a.s.}{\to} 0$.

4.3.18 Prove or disprove that $X_n \xrightarrow{a.s.} 5$ if and only if $|X_n| \xrightarrow{a.s.} 5$.

4.3.19 Suppose $X_n \xrightarrow{a.s.} X$, and $Y_n \xrightarrow{a.s.} Y$. Let $Z_n = X_n + Y_n$ and $Z = X + Y$. Prove that $Z_n \xrightarrow{a.s.} Z$.

CHALLENGES

4.3.20 Suppose for each real number $r \in [0, 1]$, there are random variables $W_r, X_{r,1}, X_{r,2}, \ldots$ such that $P(\lim_{n\to\infty} X_{r,n} = W_r) = 1$. Prove or disprove that we must have $P(\lim_{n\to\infty} X_{n,r} = W_r \text{ for all } r \in [0, 1]) = 1$.

4.3.21 Give an example of random variables $X_1, X_2, \ldots$ such that $\{X_n\}$ converges to 0 with probability 1, but $E(X_n) = 1$ for all n.

4.3.22 Suppose $X_n \xrightarrow{a.s.} X$, and f is a continuous function. Prove that $f(X_n) \xrightarrow{a.s.} f(X)$.

4.4 Convergence in Distribution

There is yet another notion of convergence of a sequence of random variables that is important in applications of probability and statistics.

Definition 4.4.1 Let $X, X_1, X_2, \ldots$ be random variables. Then we say that the sequence $\{X_n\}$ *converges in distribution* to X, if for all $x \in R^1$ such that $P(X = x) = 0$ we have $\lim_{n\to\infty} P(X_n \le x) = P(X \le x)$, and we write $X_n \xrightarrow{D} X$.

Intuitively, $\{X_n\}$ converges in distribution to X if for large n, the distribution of X_n is close to that of X. The importance of this, as we will see, is that often the distribution of X_n is difficult to work with, while that of X is much simpler. With X_n converging in distribution to X, however, we can approximate the distribution of X_n by that of X.

EXAMPLE 4.4.1

Suppose $P(X_n = 1) = 1/n$, and $P(X_n = 0) = 1 - 1/n$. Let $X = 0$ so that $P(X = 0) = 1$. Then,

$$P(X_n \le x) = \begin{cases} 0 & x < 0 \\ 1 - 1/n & 0 \le x < 1 \\ 1 & 1 \le x \end{cases} \rightarrow P(X \le x) = \begin{cases} 0 & x < 0 \\ 1 & 0 \le x \end{cases}$$

as $n \to \infty$. As $P(X_n \le x) \to P(X \le x)$ for every x, and in particular at all x where $P(X = x) = 0$, we have that $\{X_n\}$ converges in distribution to X. Intuitively, as $n \to \infty$, it is more and more likely that X_n will equal 0. ∎

EXAMPLE 4.4.2

Suppose $P(X_n = 1) = 1/2 + 1/n$, and $P(X_n = 0) = 1/2 - 1/n$. Suppose further that $P(X = 0) = P(X = 1) = 1/2$. Then $\{X_n\}$ converges in distribution to X because $P(X_n = 1) \to 1/2$ and $P(X_n = 0) \to 1/2$ as $n \to \infty$. ∎

EXAMPLE 4.4.3

Let $X \sim$ Uniform[0, 1], and let $P(X_n = i/n) = 1/n$ for $i = 1, 2, \ldots, n$. Then X is absolutely continuous, while X_n is discrete. On the other hand, for any $0 \le x \le 1$, we

have $P(X \leq x) = x$, and letting $\lfloor x \rfloor$ denote the greatest integer less than or equal to x, we have

$$P(X_n \leq x) = \frac{\lfloor nx \rfloor}{n}.$$

Hence, $|P(X_n \leq x) - P(X \leq x)| \leq 1/n$ for all n. Because $\lim_{n\to\infty} 1/n = 0$, we do indeed have $X_n \to X$ in distribution. ∎

EXAMPLE 4.4.4
Suppose $X_1, X_2, \ldots$ are i.i.d. with finite mean μ, and $M_n = (X_1 + \cdots + X_n)/n$. Then the weak law of large numbers says that for any $\epsilon > 0$, we have

$$P(M_n \leq \mu - \epsilon) \to 0 \quad \text{and} \quad P(M_n \leq \mu + \epsilon) \to 1$$

as $n \to \infty$. It follows that $\lim_{n\to\infty} P(M_n \leq x) = P(M \leq x)$ for any $x \neq \mu$, where M is the constant random variable $M = \mu$. Hence, $M_n \to M$ in distribution. Note that it is *not* necessarily the case that $P(M_n \leq \mu) \to P(M \leq \mu) = 1$. However, this does not contradict the definition of convergence in distribution because $P(M = \mu) \neq 0$, so we do not need to worry about the case $x = \mu$. ∎

EXAMPLE 4.4.5 *Poisson Approximation to the Binomial*
Suppose $X_n \sim \text{Binomial}(n, \lambda/n)$ and $X \sim \text{Poisson}(\lambda)$. We have seen in Example 2.3.6 that

$$P(X_n = j) = \binom{n}{j}\left(\frac{\lambda}{n}\right)^j (1 - \frac{\lambda}{n})^{n-j} \to e^{-\lambda}\frac{\lambda^j}{j!}$$

as $n \to \infty$. This implies that $F_{X_n}(x) \to F_X(x)$ at every point $x \notin \{0, 1, 2, \ldots\}$, and these are precisely the points for which $P(X = x) = 0$. Therefore, $\{X_n\}$ converges in distribution to X. (Indeed, this was our original motivation for the Poisson distribution.) ∎

Many more examples of convergence in distribution are given by the central limit theorem, discussed in the next section. We first pause to consider the relationship of convergence in distribution to our previous notions of convergence.

Theorem 4.4.1 If $X_n \xrightarrow{P} X$, then $X_n \xrightarrow{D} X$.

PROOF See Section 4.7 for the proof of this result. ∎

The converse to Theorem 4.4.1 is false. Indeed, the fact that X_n converges in distribution to X says nothing about the underlying *relationship* between X_n and X, it says only something about their distributions. The following example illustrates this.

EXAMPLE 4.4.6
Suppose $X, X_1, X_2, \ldots$ are i.i.d., each equal to ± 1 with probability $1/2$ each. In this case, $P(X_n \leq x) = P(X \leq x)$ for all n and for all $x \in R^1$, so of course X_n converges in distribution to X. On the other hand, because X and X_n are independent,

$$P(|X - X_n| \geq 2) = \frac{1}{2}$$

for all n, which does *not* go to 0 as $n \to \infty$. Hence, X_n does *not* converge to X in probability (or with probability 1). So we can have convergence in distribution without having convergence in probability or convergence with probability 1. ∎

The following result, stated without proof, indicates how moment-generating functions can be used to check for convergence in distribution. (This generalizes Theorem 3.4.6.)

Theorem 4.4.2 Let X be a random variable, such that for some $s_0 > 0$, we have $m_X(s) < \infty$ whenever $s \in (-s_0, s_0)$. If $Z_1, Z_2, \ldots$ is a sequence of random variables with $m_{Z_n}(s) < \infty$ and $\lim_{n\to\infty} m_{Z_n}(s) = m_X(s)$ for all $s \in (-s_0, s_0)$, then $\{Z_n\}$ converges to X in distribution.

We will make use of this result to prove one of the most famous theorems of probability — the central limit theorem.

Finally, we note that combining Theorem 4.4.1 with Theorem 4.3.1 reveals the following.

Corollary 4.4.1 If $X_n \to X$ with probability 1, then $X_n \xrightarrow{D} X$

4.4.1 The Central Limit Theorem

We now present the central limit theorem, one of the most important results in all of probability theory. Intuitively, it says that a large sum of i.i.d. random variables, properly normalized, will always have approximately a *normal* distribution. This shows that the normal distribution is extremely fundamental in probability and statistics — even though its density function is complicated and its cumulative distribution function is intractable.

Suppose $X_1, X_2, \ldots$ is an i.i.d. sequence of random variables each having finite mean μ and finite variance σ^2. Let $S_n = X_1 + \cdots + X_n$ be the sample sum and $M_n = S_n/n$ be the sample mean. The central limit theorem is concerned with the distribution of the random variable

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma} = \frac{M_n - \mu}{\sigma/\sqrt{n}} = \sqrt{n}\left(\frac{M_n - \mu}{\sigma}\right),$$

where $\sigma = \sqrt{\sigma^2}$. We know $E(M_n) = \mu$ and $\text{Var}(M_n) = \sigma^2/n$, which implies that $E(Z_n) = 0$ and $\text{Var}(Z_n) = 1$. The variable Z_n is thus obtained from the sample mean (or sample sum) by subtracting its mean and dividing by its standard deviation. This transformation is referred to as *standardizing* a random variable, so that it has mean 0 and variance 1. Therefore, Z_n is the standardized version of the sample mean (sample sum).

Note that the distribution of Z_n shares two characteristics with the $N(0, 1)$ distribution, namely, it has mean 0 and variance 1. The central limit theorem shows that there is an even stronger relationship.

Theorem 4.4.3 (*The central limit theorem*) Let $X_1, X_2, \ldots$ be i.i.d. with finite mean μ and finite variance σ^2. Let $Z \sim N(0, 1)$. Then as $n \to \infty$, the sequence $\{Z_n\}$ converges in distribution to Z, i.e., $Z_n \xrightarrow{D} Z$.

PROOF See Section 4.7 for the proof of this result. ∎

The central limit theorem is so important that we shall restate its conclusions in several different ways.

Corollary 4.4.2 For each fixed $x \in R^1$, $\lim_{n\to\infty} P(Z_n \le x) = \Phi(x)$, where Φ is the cumulative distribution function for the standard normal distribution.

We can write this as follows.

Corollary 4.4.3 For each fixed $x \in R^1$,

$$\lim_{n\to\infty} P(S_n \le n\mu + x\sqrt{n}\sigma) = \Phi(x) \quad \text{and} \quad \lim_{n\to\infty} P(M_n \le \mu + x\sigma/\sqrt{n}) = \Phi(x).$$

In particular, S_n is approximately equal to $n\mu$, with deviations from this value of order $\sqrt{n}$, and M_n is approximately equal to μ, with deviations from this value of order $1/\sqrt{n}$.

We note that it is not essential in the central limit theorem to divide by σ, in which case the theorem asserts instead that $(S_n - n\mu)/\sqrt{n}$ (or $\sqrt{n}(M_n - \mu)$) converges in distribution to the $N(0, \sigma^2)$ distribution. That is, the limiting distribution will still be normal but will have variance σ^2 instead of variance 1.

Similarly, instead of dividing by exactly σ, it suffices to divide by any quantity σ_n, provided $\sigma_n \xrightarrow{a.s.} \sigma$. A simple modification of the proof of Theorem 4.4.2 leads to the following result.

Corollary 4.4.4 If

$$Z_n^* = \frac{S_n - n\mu}{\sqrt{n}\sigma_n} = \frac{M_n - \mu}{\sigma_n/\sqrt{n}} = \sqrt{n}\left(\frac{M_n - \mu}{\sigma_n}\right)$$

and $\lim_{n\to\infty} \sigma_n \xrightarrow{a.s.} \sigma$, then $Z_n^* \xrightarrow{D} Z$ as $n \to \infty$.

To illustrate the central limit theorem, we consider a simulation experiment.

EXAMPLE 4.4.7 *The Central Limit Theorem Illustrated in a Simulation*
Suppose we generate a sample $X_1, \ldots, X_n$ from the Uniform[0, 1] density. Note that the Uniform[0, 1] density is completely unlike a normal density. An easy calculation shows that when $X \sim$ Uniform[0, 1], then $E(X) = 1/2$ and $\text{Var}(X) = 1/12$.

Now suppose we are interested in the distribution of the sample average $M_n = S_n/n = (X_1 + \cdots + X_n)/n$ for various choices of n. The central limit theorem tells us that

$$Z_n = \frac{S_n - n/2}{\sqrt{n/12}} = \sqrt{n}\left(\frac{M_n - 1/2}{\sqrt{1/12}}\right)$$

converges in distribution to an $N(0, 1)$ distribution. But how large does n have to be for this approximation to be accurate?

To assess this, we ran a Monte Carlo simulation experiment. In Figure 4.4.1, we have plotted a density histogram of $N = 10^5$ values from the $N(0, 1)$ distribution based on 800 subintervals of $(-4, 4)$, each of length $l = 0.01$. Density histograms are more extensively discussed in Section 5.4.3, but for now we note that above each interval we have plotted the proportion of sampled values that fell in the interval, divided by the length of the interval. As we increase N and decrease l, these histograms will look more and more like the density of the distribution from which we are sampling. Indeed, Figure 4.4.1 looks very much like an $N(0, 1)$ density, as it should.

In Figure 4.4.2, we have plotted a density histogram (using the same values of N and l) of Z_1. Note that $Z_1 \sim \text{Uniform}[-\sqrt{12}/2, \sqrt{12}/2]$, and indeed the histogram does look like a uniform density. Figure 4.4.3 presents a density histogram of Z_2, which still looks very nonnormal — but note that the histogram of Z_3 in Figure 4.4.4 is beginning to look more like a normal distribution. The histogram of Z_{10} in Figure 4.4.5 looks very normal. In fact, the proportion of Z_{10} values in $(-\infty, 1.96]$, for this histogram, equals 0.9759, while the exact proportion for an $N(0, 1)$ distribution is 0.9750.

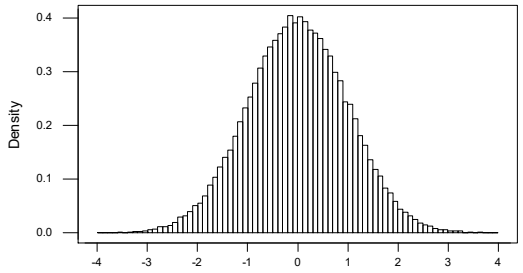

Figure 4.4.1: Density histogram of 10^5 standard normal values.

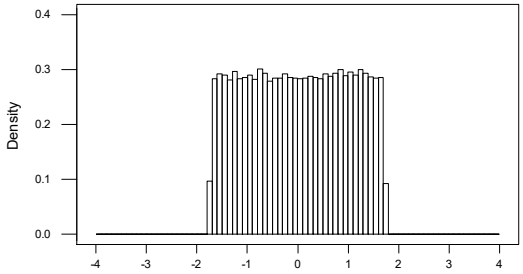

Figure 4.4.2: Density histogram for 10^5 values of Z_1 in Example 4.4.7.

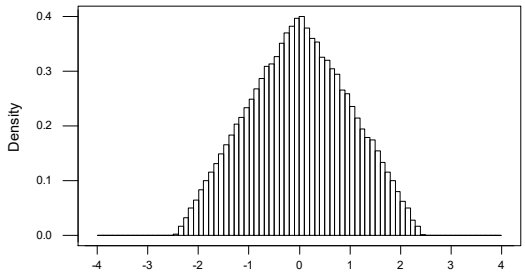

Figure 4.4.3: Density histogram for 10^5 values of Z_2 in Example 4.4.7.

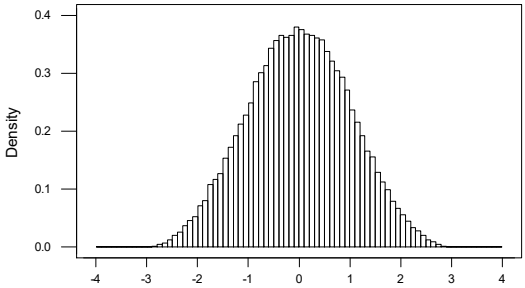

Figure 4.4.4: Density histogram for 10^5 values of Z_3 in Example 4.4.7.

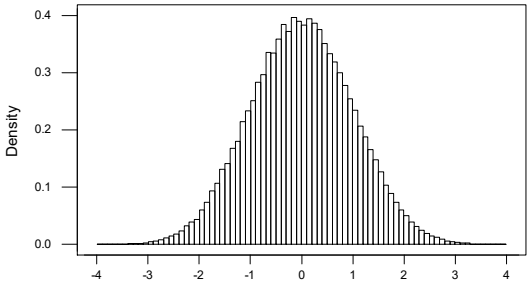

Figure 4.4.5: Density histogram for 10^5 values of Z_{10} in Example 4.4.7.

So in this example, the central limit theorem has taken effect very quickly, even though we are sampling from a very nonnormal distribution. As it turns out, it is

primarily the tails of a distribution that determine how large n has to be for the central limit theorem approximation to be accurate. When a distribution has tails no heavier than a normal distribution, we can expect the approximation to be quite accurate for relatively small sample sizes. ∎

We consider some further applications of the central limit theorem.

EXAMPLE 4.4.8
For example, suppose $X_1, X_2, \ldots$ are i.i.d. random variables, each with the Poisson(5) distribution. Recall that this implies that $\mu = E(X_i) = 5$ and $\sigma^2 = \text{Var}(X_i) = 5$. Hence, for each fixed $x \in R^1$, we have

$$P\left(S_n \leq 5n + x\sqrt{5n}\right) \to \Phi(x)$$

as $n \to \infty$. ∎

EXAMPLE 4.4.9 *Normal Approximation to the Binomial Distribution*
Suppose $X_1, X_2, \ldots$ are i.i.d. random variables, each with the Bernoulli(θ) distribution. Recall that this implies that $E(X_i) = \theta$ and $v = \text{Var}(X_i) = \theta(1-\theta)$. Hence, for each fixed $x \in R^1$, we have

$$P\left(S_n \leq n\theta + x\sqrt{n\theta(1-\theta)}\right) \to \Phi(x), \tag{4.4.1}$$

as $n \to \infty$.

But now note that we have previously shown that $Y_n = S_n \sim \text{Binomial}(n, \theta)$. So (4.4.1) implies that whenever we have a random variable $Y_n \sim \text{Binomial}(n, \theta)$, then

$$P(Y_n \leq y) = P\left(\frac{Y_n - n\theta}{\sqrt{n\theta(1-\theta)}} \leq \frac{y - n\theta}{\sqrt{n\theta(1-\theta)}}\right) \approx \Phi\left(\frac{y - n\theta}{\sqrt{n\theta(1-\theta)}}\right) \tag{4.4.2}$$

for large n.

Note that we are approximating a discrete distribution by a continuous distribution here. Reflecting this, a small improvement is often made to (4.4.2) when y is a nonnegative integer. Instead, we use

$$P(Y_n \leq y) \approx \Phi\left(\frac{y + 0.5 - n\theta}{\sqrt{n\theta(1-\theta)}}\right).$$

Adding 0.5 to y is called the *correction for continuity*. In effect, this allocates all the relevant normal probability in the interval $(y - 0.5, y + 0.5)$ to the nonnegative integer y. This has been shown to improve the approximation (4.4.2). ∎

EXAMPLE 4.4.10 *Approximating Probabilities Using the Central Limit Theorem*
While there are tables for the binomial distribution (Table D.6), we often have to compute binomial probabilities for situations the tables do not cover. We can always use statistical software for this, in fact, such software makes use of the normal approximation we derived from the central limit theorem.

For example, suppose that we have a biased coin, where the probability of getting a head on a single toss is $\theta = 0.6$. We will toss the coin $n = 1000$ times and then

calculate the probability of getting at least 550 heads and no more than 625 heads. If Y denotes the number of heads obtained in the 1000 tosses, we have that $Y \sim$ Binomial(1000, 0.6) , so

$$\begin{aligned} E(Y) &= 1000\,(0.6) = 600, \\ \text{Var}(Y) &= 1000\,(0.6)\,(0.4) = 240. \end{aligned}$$

Therefore, using the correction for continuity and Table D.2,

$$\begin{aligned} P(550 \le Y \le 625) &= P(550 - 0.5 \le Y \le 625 + 0.5) \\ &= P\left(\frac{549.5 - 600}{\sqrt{240}} \le \frac{Y - 600}{\sqrt{240}} \le \frac{625.5 - 600}{\sqrt{240}}\right) \\ &= P\left(-3.2598 \le \frac{Y - 600}{\sqrt{240}} \le 1.646\right) \\ &\approx \Phi(1.65) - \Phi(-3.26) = 0.9505 - 0.0006 = 0.9499. \end{aligned}$$

Note that it would be impossible to compute this probability using the formulas for the binomial distribution. ∎

One of the most important uses of the central limit theorem is that it leads to a method for assessing the error in an average when this is estimating or approximating some quantity of interest.

4.4.2 The Central Limit Theorem and Assessing Error

Suppose $X_1, X_2, \ldots$ is an i.i.d. sequence of random variables, each with finite mean μ and finite variance σ^2, and we are using the sample average M_n to approximate the mean μ. This situation arises commonly in many computational (see Section 4.5) and statistical (see Chapter 6) problems. In such a context, we can generate the X_i, but we do not know the value of μ.

If we approximate μ by M_n, then a natural question to ask is: How much error is there in the approximation? The central limit theorem tells us that

$$\begin{aligned} \Phi(3) - \Phi(-3) &= \lim_{n\to\infty} P\left(-3 < \frac{M_n - \mu}{\sigma/\sqrt{n}} < 3\right) \\ &= \lim_{n\to\infty} P\left(M_n - 3\frac{\sigma}{\sqrt{n}} < \mu < M_n + 3\frac{\sigma}{\sqrt{n}}\right). \end{aligned}$$

Using Table D.2 (or statistical software), we have that $\Phi(3) - \Phi(-3) = 0.9987 - (1 - 0.9987) = 0.9974$. So, for large n, we have that the interval

$$\left(M_n - 3\sigma/\sqrt{n},\, M_n + 3\sigma/\sqrt{n}\right)$$

contains the unknown value of μ with virtual certainty (actually with probability about 0.9974). Therefore, the half-length $3\sigma/\sqrt{n}$ of this interval gives us an assessment of the error in the approximation M_n. Note that $\text{Var}(M_n) = \sigma^2/n$, so the half-length of the interval equals 3 standard deviations of the estimate M_n.

Because we do not know μ, it is extremely unlikely that we will know σ (as its definition uses μ). But if we can find a consistent estimate σ_n of σ, then we can use Corollary 4.4.4 instead to construct such an interval.

As it turns out, the correct choice of σ_n depends on what we know about the distribution we are sampling from (see Chapter 6 for more discussion of this). For example, if $X_1 \sim \text{Bernoulli}(\theta)$, then $\mu = \theta$ and $\sigma^2 = \text{Var}(X_1) = \theta(1-\theta)$. By the strong law of large numbers (Theorem 4.3.2), $M_n \overset{a.s.}{\to} \mu = \theta$ and thus

$$\sigma_n = \sqrt{M_n(1-M_n)} \overset{a.s.}{\to} \sqrt{\theta(1-\theta)} = \sigma.$$

Then, using the same argument as above, we have that, for large n, the interval

$$\left(M_n - 3\sqrt{M_n(1-M_n)/n},\, M_n + 3\sqrt{M_n(1-M_n)/n}\right) \tag{4.4.3}$$

contains the true value of θ with virtual certainty (again, with probability about 0.9974). The half-length of (4.4.3) is a measure of the accuracy of the estimate M_n — notice that this can be computed from the values $X_1, \ldots, X_n$. We refer to the quantity $(M_n(1-M_n)/n)^{1/2}$ as the *standard error* of the estimate M_n.

For a general random variable X_1, let

$$\begin{aligned}\sigma_n^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_n)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2 - 2M_n\sum_{i=1}^{n}X_i + nM_n^2\right)\\ &= \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2 - 2M_n^2 + M_n^2\right) = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2 - M_n^2\right).\end{aligned}$$

By the strong law of large numbers, we have that $M_n \overset{a.s.}{\to} \mu$ and

$$\frac{1}{n}\sum_{i=1}^{n}X_i^2 \overset{a.s.}{\to} E\left(X_1^2\right) = \sigma^2 + \mu^2.$$

Because $n/(n-1) \to 1$ and $M_n^2 \overset{a.s.}{\to} \mu^2$ as well, we conclude that $\sigma_n^2 \overset{a.s.}{\to} \sigma^2$. This implies that $\sigma_n \overset{a.s.}{\to} \sigma$ hence σ_n is consistent for σ. It is common to call σ_n^2 the *sample variance* of the sample $X_1, \ldots, X_n$. When the sample size n is fixed, we will often denote this estimate of the variance by S^2.

Again, using the above argument, we have that, for large n, the interval

$$\left(M_n - 3\sigma_n/\sqrt{n},\, M_n + 3\sigma_n/\sqrt{n}\right) = \left(M_n - 3S/\sqrt{n},\, M_n + 3S/\sqrt{n}\right) \tag{4.4.4}$$

contains the true value of μ with virtual certainty (also with probability about 0.9974). Therefore, the half-length is a measure of the accuracy of the estimate M_n — notice that this can be computed from the values $X_1, \ldots, X_n$. The quantity $S/\sqrt{n}$ is referred to as the *standard error* of the estimate M_n.

We will make use of these estimates of the error in approximations in the following section.

Summary of Section 4.4

- A sequence $\{X_n\}$ of random variables converges in distribution to Y if, for all $y \in R^1$ with $P(Y = y) = 0$, we have $\lim_{n\to\infty} F_{X_n}(y) = F_Y(y)$, i.e., $\lim_{n\to\infty} P(X_n \le y) = P(Y \le y)$.
- If $\{X_n\}$ converges to Y in probability (or with probability 1), then $\{X_n\}$ converges to Y in distribution.
- The very important central limit theorem says that if $\{X_n\}$ are i.i.d. with finite mean μ and variance σ^2, then the random variables $Z_n = (S_n - n\mu)/\sqrt{n}\sigma$ converge in distribution to a standard normal distribution.
- The central limit theorem allows us to approximate various distributions by normal distributions, which is helpful in simulation experiments and in many other contexts. Table D.2 (or any statistical software package) provides values for the cumulative distribution function of a standard normal.

EXERCISES

4.4.1 Suppose $P(X_n = i) = (n + i)/(3n + 6)$ for $i = 1, 2, 3$. Suppose also that $P(X = i) = 1/3$ for $i = 1, 2, 3$. Prove that $\{X_n\}$ converges in distribution to X.

4.4.2 Suppose $P(Y_n = k) = (1 - 2^{-n-1})^{-1}/2^{k+1}$ for $k = 0, 1, \ldots, n$. Let $Y \sim$ Geometric$(1/2)$. Prove that $\{Y_n\}$ converges in distribution to Y.

4.4.3 Let Z_n have density $(n + 1)x^n$ for $0 < x < 1$, and 0 otherwise. Let $Z = 1$. Prove that $\{Z_n\}$ converges in distribution to Z.

4.4.4 Let W_n have density

$$\frac{1 + x/n}{1 + 1/2n}$$

for $0 < x < 1$, and 0 otherwise. Let $W \sim$ Uniform$[0, 1]$. Prove that $\{W_n\}$ converges in distribution to W.

4.4.5 Let $Y_1, Y_2, \ldots$ be i.i.d. with distribution Exponential(3). Use the central limit theorem and Table D.2 (or software) to estimate the probability $P(\sum_{i=1}^{1600} Y_i \le 540)$.

4.4.6 Let $Z_1, Z_2, \ldots$ be i.i.d. with distribution Uniform$[-20, 10]$. Use the central limit theorem and Table D.2 (or software) to estimate the probability $P(\sum_{i=1}^{900} Z_i \ge -4470)$.

4.4.7 Let $X_1, X_2, \ldots$ be i.i.d. with distribution Geometric$(1/4)$. Use the central limit theorem and Table D.2 (or software) to estimate the probability $P(\sum_{i=1}^{800} X_i \ge 2450)$.

4.4.8 Suppose $X_n \sim N(0, 1/n)$, i.e., X_n has a normal distribution with mean 0 and variance $1/n$. Does the sequence $\{X_n\}$ converge in distribution to some random variable? If yes, what is the distribution of the random variable?

4.4.9 Suppose $P(X_n = i/n) = 2i/n(n+1)$ for $i = 1, 2, 3, \ldots, n$. Let Z have density function given by $f(z) = 2z$ for $0 < z < 1$, otherwise $f(z) = 0$.
(a) Compute $P(Z \le y)$ for $0 < y < 1$.
(b) Compute $P(X_n \le m/n)$ for some integer $1 \le m \le n$. (Hint: Remember that $\sum_{i=1}^{m} i = m(m + 1)/2$.)

(c) Compute $P(X_n \leq y)$ for $0 < y < 1$.
(d) Prove that $X_n \xrightarrow{D} Z$.

4.4.10 Suppose $P(Y_n \leq y) = 1 - e^{-2ny/(n+1)}$ for all $y > 0$. Prove that $Y_n \xrightarrow{D} Y$ where $Y \sim$ Exponential(λ) for some $\lambda > 0$ and compute λ.

4.4.11 Suppose $P(Z_n \leq z) = 1 - (1 - \frac{3z}{n})^n$ for all $0 < z < n/3$. Prove that $Z_n \xrightarrow{D} Z$ where $Z \sim$ Exponential(λ) for some $\lambda > 0$ and compute λ. (Hint: Recall from calculus that $\lim_{n\to\infty}(1 + \frac{c}{n})^n = e^c$ for any real number c.)

4.4.12 Suppose the service time, in minutes, at a bank has the Exponential distribution with $\lambda = 1/2$. Use the central limit theorem to estimate the probability that the average service time of the first n customers is less than 2.5 minutes, when:
(a) $n = 16$.
(b) $n = 36$.
(c) $n = 100$.

4.4.13 Suppose the number of kilograms of a metal alloy produced by a factory each week is uniformly distributed between 20 and 30. Use the central limit theorem to estimate the probability that next year's output will be less than 1280 kilograms. (Assume that a year contains precisely 52 weeks.)

4.4.14 Suppose the time, in days, until a component fails has the Gamma distribution with $\alpha = 5$ and $\lambda = 1/10$. When a component fails, it is immediately replaced by a new component. Use the central limit theorem to estimate the probability that 40 components will together be sufficient to last at least 6 years. (Assume that a year contains precisely 365.25 days.)

COMPUTER EXERCISES

4.4.15 Generate N samples $X_1, X_2, \ldots, X_{20} \sim$ Exponential(3) for N large ($N = 10^4$, if possible). Use these samples to estimate the probability $P(1/6 \leq M_{20} \leq 1/2)$. How does your answer compare to what the central limit theorem gives as an approximation?

4.4.16 Generate N samples $X_1, X_2, \ldots, X_{30} \sim$ Uniform[$-20, 10$] for N large ($N = 10^4$, if possible). Use these samples to estimate the probability $P(M_{30} \leq -5)$. How does your answer compare to what the central limit theorem gives as an approximation?

4.4.17 Generate N samples $X_1, X_2, \ldots, X_{20} \sim$ Geometric(1/4) for N large ($N = 10^4$, if possible). Use these samples to estimate the probability $P(2.5 \leq M_{20} \leq 3.3)$. How does your answer compare to what the central limit theorem gives as an approximation?

4.4.18 Generate N samples $X_1, X_2, \ldots, X_{20}$ from the distribution of $\log Z$ where $Z \sim$ Gamma(4, 1) for N large ($N = 10^4$, if possible). Use these samples to construct a density histogram of the values of M_{20}. Comment on the shape of this graph.

4.4.19 Generate N samples $X_1, X_2, \ldots, X_{20}$ from the Binomial(10, 0.01) distribution for N large ($N = 10^4$, if possible). Use these samples to construct a density histogram of the values of M_{20}. Comment on the shape of this graph.

PROBLEMS

4.4.20 Let $a_1, a_2, \ldots$ be any sequence of nonnegative real numbers with $\sum_i a_i = 1$. Suppose $P(X = i) = a_i$ for every positive integer i. Construct a sequence $\{X_n\}$ of *absolutely continuous* random variables, such that $X_n \to X$ in distribution.

4.4.21 Let $f : [0, 1] \to (0, \infty)$ be a continuous positive function such that $\int_0^1 f(x)\,dx = 1$. Consider random variables X and $\{X_n\}$ such that $P(a \le X \le b) = \int_a^b f(x)\,dx$ for $a \le b$ and

$$P\left(X_n = \frac{i}{n}\right) = \frac{f(i/n)}{\sum_{j=1}^n f(j/n)}$$

for $i = 1, 2, 3, \ldots, n$. Prove that $X_n \to X$ in distribution.

4.4.22 Suppose that $Y_i = X_i^3$ and that $X_1, \ldots, X_n$ is a sample from an $N(0, 1)$ distribution. Indicate how you would approximate the probability $P(M_n \le m)$ where $M_n = (Y_1 + \cdots + Y_n)/n$.

4.4.23 Suppose $Y_i = \cos(2\pi U_i)$ and $U_1, \ldots, U_n$ is a sample from the Uniform[0, 1] distribution. Indicate how you would approximate the probability $P(M_n \le m)$, where $M_n = (Y_1 + \cdots + Y_n)/n$.

COMPUTER PROBLEMS

4.4.24 Suppose that $Y = X^3$ and $X \sim N(0, 1)$. By generating a large sample ($n = 10^4$, if possible) from the distribution of Y, approximate the probability $P(Y \le 1)$ and assess the error in your approximation. Compute this probability exactly and compare it with your approximation.

4.4.25 Suppose that $Y = X^3$ and $X \sim N(0, 1)$. By generating a large sample ($n = 10^4$, if possible) from the distribution of Y, approximate the expectation $E(\cos(X^3))$, and assess the error in your approximation.

CHALLENGES

4.4.26 Suppose $X_n \to C$ in distribution, where C is a constant. Prove that $X_n \to C$ in probability. (This proves that if X is *constant*, then the converse to Theorem 4.4.1 *does* hold, even though it does not hold for general X.)

4.5 Monte Carlo Approximations

The laws of large numbers say that if $X_1, X_2, \ldots$ is an i.i.d. sequence of random variables with mean μ, and

$$M_n = \frac{X_1 + \cdots + X_n}{n},$$

then for large n we will have $M_n \approx \mu$.

Suppose now that μ is *unknown*. Then, as discussed in Section 4.4.2, it is possible to change perspective and use M_n (for large n) as an *estimator* or approximation of μ. Any time we approximate or estimate a quantity, we must also say something about

how much error is in the estimate. Of course, we cannot say what this error is exactly, as that would require knowing the exact value of μ. In Section 4.4.2, however, we showed how the central limit theorem leads to a very natural approach to assessing this error, using three times the standard error of the estimate. We consider some examples.

EXAMPLE 4.5.1
Consider flipping a sequence of identical coins, each of which has probability θ of coming up heads, but where θ is unknown. Let M_n again be the fraction of the first n coins that are heads. Then we know that for large n, it is very likely that M_n is very close to θ. Hence, we can use M_n to *estimate* θ. Furthermore, the discussion in Section 4.4.2 indicates that (4.4.3) is the relevant interval to quote when assessing the accuracy of the estimate M_n. ∎

EXAMPLE 4.5.2
Suppose we believe a certain medicine lowers blood pressure, but we do not know by how much. We would like to know the mean amount μ, by which this medicine lowers blood pressure.

Suppose we observe n patients (chosen at random so they are i.i.d.), where patient i has blood pressure B_i before taking the medicine and blood pressure A_i afterwards. Let $X_i = B_i - A_i$. Then

$$M_n = \frac{1}{n}\sum_{i=1}^{n}(B_i - A_i)$$

is the average amount of blood pressure decrease. (Note that $B_i - A_i$ may be negative for some patients, and it is important to also include those negative terms in the sum.) Then for large n, the value of M_n is a good *estimate* of $E(X_i) = \mu$. Furthermore, the discussion in Section 4.4.2 indicates that (4.4.4) is the relevant interval to quote when assessing the accuracy of the estimate M_n. ∎

Such estimators can also be used to estimate purely mathematical quantities that do not involve any experimental data (such as coins or medical patients) but that are too difficult to compute directly. In this case, such estimators are called *Monte Carlo approximations* (named after the gambling casino in the principality of Monaco because they introduce randomness to solve nonrandom problems).

EXAMPLE 4.5.3
Suppose we wish to evaluate

$$I = \int_0^1 \cos(x^2)\sin(x^4)\,dx.$$

This integral cannot easily be solved exactly. But it can be approximately computed using a Monte Carlo approximation, as follows. We note that

$$I = E(\cos(U^2)\,\sin(U^4)),$$

where $U \sim$ Uniform[0, 1]. Hence, for large n, the integral I is approximately equal to $M_n = (T_1 + \cdots + T_n)/n$, where $T_i = \cos(U_i^2)\sin(U_i^4)$, and where $U_1, U_2, \ldots$ are i.i.d. Uniform[0, 1].

Putting this all together, we obtain an algorithm for approximating the integral I, as follows.

1. Select a large positive integer n.
2. Obtain $U_i \sim \text{Uniform}[0,1]$, independently for $i = 1, 2, \ldots, n$.
3. Set $T_i = \cos(U_i^2)\sin(U_i^4)$, for $i = 1, 2, \ldots, n$.
4. Estimate I by $M_n = (T_1 + \cdots + T_n)/n$.

For large enough n, this algorithm will provide a good estimate of the integral I.

For example, the following table records the estimates M_n and the intervals (4.4.4) based on samples of Uniform[0,1] variables for various choices of n.

n	M_n	$M_n - 3S/\sqrt{n}$	$M_n + 3S/\sqrt{n}$
10^3	0.145294	0.130071	0.160518
10^4	0.138850	0.134105	0.143595
10^5	0.139484	0.137974	0.140993

From this we can see that the value of I is approximately 0.139484, and the true value is almost certainly in the interval (0.137974, 0.140993). Notice how the lengths of the intervals decrease as we increase n. In fact, it can be shown that the exact value is $I = 0.139567$, so our approximation is excellent. ∎

EXAMPLE 4.5.4

Suppose we want to evaluate the integral

$$I = \int_0^\infty 25x^2 \cos(x^2) e^{-25x}\, dx.$$

This integral cannot easily be solved exactly, but it can also be approximately computed using a Monte Carlo approximation, as follows.

We note first that $I = E(X^2\cos(X^2))$, where $X \sim$ Exponential(25). Hence, for large n, the integral I is approximately equal to $M_n = (T_1 + \cdots + T_n)/n$, where $T_i = X_i^2\cos(X_i^2)$, with $X_1, X_2, \ldots$ i.i.d. Exponential(25).

Now, we know from Section 2.10 that we can simulate $X \sim$ Exponential(25) by setting $X = -\ln(U)/25$ where $U \sim$ Uniform[0, 1]. Hence, putting this all together, we obtain an algorithm for approximating the integral I, as follows.

1. Select a large positive integer n.
2. Obtain $U_i \sim \text{Uniform}[0,1]$, independently for $i = 1, 2, \ldots, n$.
3. Set $X_i = -\ln(U_i)/25$, for $i = 1, 2, \ldots, n$.
4. Set $T_i = X_i^2\cos(X_i^2)$, for $i = 1, 2, \ldots, n$.
5. Estimate I by $M_n = (T_1 + \cdots + T_n)/n$.

For large enough n, this algorithm will provide a good estimate of the integral I.

For example, the following table records the estimates M_n and the intervals (4.4.4) based on samples of Exponential(25) variables for various choices of n.

n	M_n	$M_n - 3S/\sqrt{n}$	$M_n + 3S/\sqrt{n}$
10^3	3.33846×10^{-3}	2.63370×10^{-3}	4.04321×10^{-3}
10^4	3.29933×10^{-3}	3.06646×10^{-3}	3.53220×10^{-3}
10^5	3.20629×10^{-3}	3.13759×10^{-3}	3.27499×10^{-3}

From this we can see that the value of I is approximately 3.20629×10^{-3} and that the true value is almost certainly in the interval $\left(3.13759 \times 10^{-3}, 3.27499 \times 10^{-3}\right)$. ∎

EXAMPLE 4.5.5

Suppose we want to evaluate the sum

$$S = \sum_{j=0}^{\infty} (j^2 + 3)^{-7} 5^{-j}.$$

Though this is very difficult to compute directly, it can be approximately computed using a Monte Carlo approximation.

Let us rewrite the sum as

$$S = \left(\frac{5}{4}\right) \sum_{j=0}^{\infty} (j^2 + 3)^{-7} \left(\frac{4}{5}\right) \left(1 - \frac{4}{5}\right)^j.$$

We then see that $S = (5/4)\, E((X^2 + 3)^{-7})$, where $X \sim$ Geometric$(4/5)$.

Now, we know from Section 2.10 that we can simulate $X \sim$ Geometric$(4/5)$ by setting $X = \lfloor \ln(1 - U)/\ln(1 - 4/5) \rfloor$ or, equivalently, $X = \lfloor \ln(U)/\ln(1 - 4/5) \rfloor$, where $U \sim$ Uniform$[0, 1]$ and where $\lfloor \cdot \rfloor$ means to round down to the next integer value. Hence, we obtain an algorithm for approximating the sum S, as follows.

1. `Select a large positive integer` n.
2. `Obtain` $U_i \sim$ `Uniform`$[0, 1]$, `independently for` $i = 1, 2, \ldots, n$.
3. `Set` $X_i = \lfloor \ln(U_i)/\ln(1 - 4/5) \rfloor$, `for` $i = 1, 2, \ldots, n$.
4. `Set` $T_i = (X_i^2 + 3)^{-7}$, `for` $i = 1, 2, \ldots, n$.
5. `Estimate` S `by` $M_n = (5/4)(T_1 + \cdots + T_n)/n$.

For large enough n, this algorithm will provide a good estimate of the sum S.

For example, the following table records the estimates M_n and the intervals (4.4.4) based on samples of Geometric$(4/5)$ variables for various choices of n.

n	M_n	$M_n - 3S/\sqrt{n}$	$M_n + 3S/\sqrt{n}$
10^3	4.66773×10^{-4}	4.47078×10^{-4}	4.86468×10^{-4}
10^4	4.73538×10^{-4}	4.67490×10^{-4}	4.79586×10^{-4}
10^5	4.69377×10^{-4}	4.67436×10^{-4}	4.71318×10^{-4}

From this we can see that the value of S is approximately 4.69377×10^{-4} and that the true value is almost certainly in the interval $\left(4.67436 \times 10^{-4}, 4.71318 \times 10^{-4}\right)$. ∎

Note that when using a Monte Carlo approximation, it is not necessary that the range of an integral or sum be the entire range of the corresponding random variable, as follows.

EXAMPLE 4.5.6

Suppose we want to evaluate the integral

$$J = \int_0^\infty \sin(x) e^{-x^2/2}\, dx.$$

Again, this is extremely difficult to evaluate exactly.

Here

$$J = \sqrt{2\pi}\, E(\sin(X) I_{\{X>0\}}),$$

where $X \sim N(0, 1)$ and $I_{\{X>0\}}$ is the indicator function of the event $\{X > 0\}$. We know from Section 2.10 that we can simulate $X \sim N(0, 1)$ by setting

$$X = \sqrt{2 \log(1/U)}\, \cos(2\pi V),$$

where U and V are i.i.d. Uniform[0, 1]. Hence, we obtain the following algorithm for approximating the integral J.

1. Select a large positive integer n.
2. Obtain $U_i, V_i \sim$ Uniform$[0, 1]$, independently for $i = 1, 2, \ldots, n$.
3. Set $X_i = \sqrt{2 \log(1/U_i)}\, \cos(2\pi V_i)$, for $i = 1, 2, \ldots, n$.
4. Set $T_i = \sin(X_i) I_{\{X_i>0\}}$, for $i = 1, 2, \ldots, n$. (That is, set $T_i = \sin(X_i)$ if $X_i > 0$, otherwise set $T_i = 0$.)
5. Estimate J by $M_n = \sqrt{2\pi}\,(T_1 + \cdots + T_n)/n$.

For large enough n, this algorithm will again provide a good estimate of the integral I.

For example, the following table records the estimates M_n and the intervals (4.4.4) based on samples of $N(0, 1)$ variables for various choices of n.

n	M_n	$M_n - 3S/\sqrt{n}$	$M_n + 3S/\sqrt{n}$
10^3	0.744037	0.657294	0.830779
10^4	0.733945	0.706658	0.761233
10^5	0.722753	0.714108	0.731398

From this we can see that the value of J is approximately 0.722753 and that the true value is almost certainly in the interval (0.714108, 0.731398). ∎

Now we consider an important problem for statistical applications of probability theory.

EXAMPLE 4.5.7 *Approximating Sampling Distributions Using Monte Carlo*
Suppose $X_1, X_2, \ldots, X_n$ is an i.i.d. sequence from the probability measure P. We want to find the distribution of a new random variable $Y = h(X_1, X_2, \ldots, X_n)$ for some function h. Provided we can generate from P, then Monte Carlo methods give us a way to approximate this distribution.

Denoting the cumulative distribution function of Y by F_Y, we have

$$F_Y(y) = P((-\infty, y]) = E_{P_Y}(I_{(-\infty,y]}(Y)) = E(I_{(-\infty,y]}(h(X_1, X_2, \ldots, X_n))).$$

So $F_Y(y)$ can be expressed as the expectation of the random variable

$$I_{(-\infty,y]}(h(X_1, X_2, \ldots, X_n))$$

based on sampling from P.

To estimate this, we generate N samples of size n

$$(X_{i1}, X_{i2}, \ldots, X_{in}),$$

for $i = 1, \ldots, N$ from P (note N is the Monte Carlo sample size and can be varied, whereas the sample size n is fixed here) and then calculate the proportion of values $h(X_{i1}, X_{i2}, \ldots, X_{in}) \in (-\infty, y]$. The estimate M_N is then given by

$$\hat{F}_Y(y) = \frac{1}{N}\sum_{i=1}^{N} I_{(-\infty,y]}(h(X_{i1}, X_{i2}, \ldots, X_{in})).$$

By the laws of large numbers, this converges to $F_Y(y)$ as $N \to \infty$. To evaluate the error in this approximation, we use (4.4.3), which now takes the form

$$\left(\hat{F}_Y(y) - 3\sqrt{\hat{F}_Y(y)\left(1 - \hat{F}_Y(y)\right)/n}, \hat{F}_Y(y) + 3\sqrt{\hat{F}_Y(y)\left(1 - \hat{F}_Y(y)\right)/n}\right).$$

We presented an application of this in Example 4.4.7. Note that if the base of a rectangle in the histogram of Figure 4.4.2 is given by $(a, b]$, then the height of this rectangle equals the proportion of values that fell in $(a, b]$ times $1/(b-a)$. This can be expressed as $(\hat{F}_Y(b) - \hat{F}_Y(a))/(b-a)$, which converges to $(F_Y(b) - F_Y(a))/(b-a)$ as $N \to \infty$. This proves that the areas of the rectangles in the histogram converge to $F_Y(b) - F_Y(a)$ as $N \to \infty$.

More generally, we can approximate an expectation $E(g(Y))$ using the average

$$\frac{1}{N}\sum_{i=1}^{N} g(h(X_{i1}, X_{i2}, \ldots, X_{in})).$$

By the laws of large numbers, this average converges to $E(g(Y))$ as $N \to \infty$. ∎

Typically, there is more than one possible Monte Carlo algorithm for estimating a quantity of interest. For example, suppose we want to approximate the integral $\int_a^b g(x)\,dx$, where we assume this integral is finite. Let f be a density on the interval

(a, b), such that $f(x) > 0$ for every $x \in (a, b)$, and suppose we have a convenient algorithm for generating $X_1, X_2, \ldots$ i.i.d. with distribution given by f. We have that

$$\int_a^b g(x)\,dx = \int_a^b \frac{g(x)}{f(x)} f(x)\,dx = E\left(\frac{g(X)}{f(X)}\right)$$

when X is distributed with density f. So we can estimate $\int_a^b g(x)\,dx$ by

$$M_n = \frac{1}{n}\sum_{i=1}^{n} \frac{g(X_i)}{f(X_i)} = \frac{1}{n}\sum_{i=1}^{n} T_i,$$

where $T_i = g(X_i)/f(X_i)$. In effect, this is what we did in Example 4.5.3 (f is the Uniform[0, 1] density), in Example 4.5.4 (f is the Exponential(25) density), and in Example 4.5.6 (f is the $N(0, 1)$ density). But note that there are many other possible choices. In Example 4.5.3, we could have taken f to be any beta density. In Example 4.5.4, we could have taken f to be any gamma density, and similarly in Example 4.5.6. Most statistical computer packages have commands for generating from these distributions. In a given problem, what is the best one to use?

In such a case, we would naturally use the algorithm that was most *efficient*. For the algorithms we have been discussing here, this means that if, based on a sample of n, algorithm 1 leads to an estimate with standard error $\sigma_1/\sqrt{n}$, and algorithm 2 leads to an estimate with standard error $\sigma_2/\sqrt{n}$, then algorithm 1 is more efficient than algorithm 2 whenever $\sigma_1 < \sigma_2$. Naturally, we would prefer algorithm 1 because the intervals (4.4.3) or (4.4.4) will tend to be shorter for algorithm 1 for the same sample size. Actually, a more refined comparison of efficiency would also take into account the total amount of computer time used by each algorithm, but we will ignore this aspect of the problem here. See Problem 4.5.21 for more discussion of efficiency and the choice of algorithm in the context of the integration problem.

Summary of Section 4.5

- An unknown quantity can be approximately computed using a Monte Carlo approximation, whereby independent replications of a random experiment (usually on a computer) are averaged to estimate the quantity.
- Monte Carlo approximations can be used to approximate complicated sums, integrals, and sampling distributions, all by choosing the random experiment appropriately.

EXERCISES

4.5.1 Describe a Monte Carlo approximation of $\int_{-\infty}^{\infty} \cos^2(x) e^{-x^2/2}\,dx$.

4.5.2 Describe a Monte Carlo approximation of $\sum_{j=0}^{m} j^6 \binom{m}{j} 2^j 3^{-m}$. (Hint: Remember the Binomial$(m, 2/3)$ distribution.)

4.5.3 Describe a Monte Carlo approximation of $\int_0^{\infty} e^{-5x-14x^2}\,dx$. (Hint: Remember the Exponential(5) distribution.)

4.5.4 Suppose $X_1, X_2, \ldots$ are i.i.d. with distribution Poisson(λ), where λ is unknown. Consider $M_n = (X_1 + X_2 + \cdots + X_n)/n$ as an estimate of λ. Suppose we know that $\lambda \leq 10$. How large must n be to guarantee that M_n will be within 0.1 of the true value of λ with virtual certainty, i.e., when is 3 standard deviations smaller than 0.1?

4.5.5 Describe a Monte Carlo approximation of $\sum_{j=0}^{\infty} \sin(j^2) 5^j / j!$. Assume you have available an algorithm for generating from the Poisson(5) distribution.

4.5.6 Describe a Monte Carlo approximation of $\int_0^{10} e^{-x^4} dx$. (Hint: Remember the Uniform[0, 10] distribution.)

4.5.7 Suppose we repeat a certain experiment 2000 times and obtain a sample average of -5 and a standard error of 17. In terms of this, specify an interval that is virtually certain to contain the experiment's (unknown) true mean μ.

4.5.8 Suppose we repeat a certain experiment 400 times and get i.i.d. response values $X_1, X_2, \ldots, X_{400}$. Suppose we compute that the sample average is $M_{400} = 6$ and furthermore that $\sum_{i=1}^{400} (X_i)^2 = 15{,}400$. In terms of this:
(a) Compute the standard error σ_n.
(b) Specify an interval that is virtually certain to contain the (unknown) true mean μ of the X_i.

4.5.9 Suppose a certain experiment has probability θ of success, where $0 < \theta < 1$ but θ is unknown. Suppose we repeat the experiment 1000 times, of which 400 are successes and 600 are failures. Compute an interval of values that are virtually certain to contain θ.

4.5.10 Suppose a certain experiment has probability θ of success, where $0 < \theta < 1$ but θ is unknown. Suppose we repeat the experiment n times, and let Y be the fraction of successes.
(a) In terms of θ, what is Var(Y)?
(b) For what value of θ is Var(Y) the largest?
(c) What is this largest possible value of Var(Y)?
(d) Compute the smallest integer n such that we can be sure that Var(Y) < 0.01, regardless of the value of θ.

4.5.11 Suppose X and Y are random variables with joint density given by $f_{X,Y}(x, y) = C\, g(x, y)$ for $0 \leq x, y \leq 1$ (with $f_{X,Y}(x, y) = 0$ for other x, y), for appropriate constant C, where

$$g(x, y) \;=\; x^2\, y^3\, \sin(xy)\, \cos(\sqrt{xy}\,)\, \exp(x^2 + y)\,.$$

(a) Explain why

$$E(X) \;=\; \int_0^1 \int_0^1 x\, f_{X,Y}(x, y)\, dx\, dy \;=\; \frac{\int_0^1 \int_0^1 x\, g(x, y)\, dx\, dy}{\int_0^1 \int_0^1 g(x, y)\, dx\, dy}\,.$$

(b) Describe a Monte Carlo algorithm to approximately compute $E(X)$.

4.5.12 Let $g(x, y) = \cos(\sqrt{xy}\,)$, and consider the integral $I = \int_0^5 \int_0^4 g(x, y)\, dy\, dx$.
(a) Prove that $I = 20\, E[g(X, Y)]$ where $X \sim$ Uniform[0, 5] and $Y \sim$ Uniform[0, 4].
(b) Use part (a) to describe a Monte Carlo algorithm to approximately compute I.

4.5.13 Consider the integral $J = \int_0^1 \int_0^\infty h(x, y)\, dy\, dx$, where

$$h(x, y) = e^{-y^2} \cos(\sqrt{xy}).$$

(a) Prove that $J = E[e^Y h(X, Y)]$, where $X \sim$ Uniform[0, 1] and $Y \sim$ Exponential(1).
(b) Use part (a) to describe a Monte Carlo algorithm to approximately compute J.
(c) If $X \sim$ Uniform[0, 1] and $Y \sim$ Exponential(5), then prove that

$$J = (1/5)\, E[e^{5Y} h(X, Y)].$$

(d) Use part (c) to describe a Monte Carlo algorithm to approximately compute J.
(e) Explain how you might use a computer to determine which is better, the algorithm in part (b) or the algorithm in part (d).

COMPUTER EXERCISES

4.5.14 Use a Monte Carlo algorithm to approximate $\int_0^1 \cos(x^3) \sin(x^4)\, dx$ based on a large sample (take $n = 10^5$, if possible). Assess the error in the approximation.
4.5.15 Use a Monte Carlo algorithm to approximate $\int_0^\infty 25 \cos(x^4) e^{-25x}\, dx$ based on a large sample (take $n = 10^5$, if possible). Assess the error in the approximation.
4.5.16 Use a Monte Carlo algorithm to approximate $\sum_{j=0}^\infty (j^2 + 3)^{-5} 5^{-j}$ based on a large sample (take $n = 10^5$, if possible). Assess the error in the approximation.
4.5.17 Suppose $X \sim N(0, 1)$. Use a Monte Carlo algorithm to approximate $P(X^2 - 3X + 2 \geq 0)$ based on a large sample (take $n = 10^5$, if possible). Assess the error in the approximation.

PROBLEMS

4.5.18 Suppose that $X_1, X_2, \ldots$ are i.i.d. Bernoulli(θ) where θ is unknown. Determine a lower bound on n so that the probability that the estimate M_n will be within δ of the unknown value of θ is about 0.9974. This allows us to run simulations with high confidence that the error in the approximation quoted is less than some prescribed value δ. (Hint: Use the fact that $x\,(1 - x) \leq 1/4$ for all $x \in [0, 1]$.)
4.5.19 Suppose that $X_1, X_2, \ldots$ are i.i.d. with unknown mean μ and unknown variance σ^2. Suppose we know, however, that $\sigma^2 \leq \sigma_0^2$, where σ_0^2 is a known value. Determine a lower bound on n so that the probability that the estimate M_n will be within δ of the unknown value of μ is about 0.9974. This allows us to run simulations with high confidence that the error in the approximation quoted is less than some prescribed value δ.
4.5.20 Suppose $X_1, X_2, \ldots$ are i.i.d. with distribution Uniform[$0, \theta$], where θ is unknown, and consider $Z_n = n^{-1}\,(n + 1)\, X_{(n)}$ as an estimate of θ (see Section 2.8.4 on order statistics).
(a) Prove that $E\,(Z_n) = \theta$ and compute Var(Z_n).

(b) Use Chebyshev's inequality to show that Z_n converges in probability to θ.
(c) Show that $E(2M_n) = \theta$ and compare M_n and Z_n with respect to their efficiencies as estimators of θ. Which would you use to estimate θ and why?

4.5.21 (*Importance sampling*) Suppose we want to approximate the integral $\int_a^b g(x)\,dx$, where we assume this integral is finite. Let f be a density on the interval (a, b) such that $f(x) > 0$ for every $x \in (a, b)$ and is such that we have a convenient algorithm for generating $X_1, X_2, \ldots$ i.i.d. with distribution given by f.
(a) Prove that

$$M_n(f) = \frac{1}{n}\sum_{i=1}^{n} \frac{g(X_i)}{f(X_i)} \overset{a.s.}{\to} \int_a^b g(x)\,dx.$$

(We refer to f as an *importance sampler* and note this shows that every f satisfying the above conditions, provides a consistent estimator $M_n(f)$ of $\int_a^b g(x)\,dx$.)
(b) Prove that

$$\mathrm{Var}(M_n(f)) = \frac{1}{n}\left\{\int_a^b \frac{g^2(x)}{f(x)}\,dx - \left(\int_a^b g(x)\,dx\right)^2\right\}.$$

(c) Suppose that $g(x) = h(x)f(x)$, where f is as described above. Show that importance sampling with respect to f leads to the estimator

$$M_n(f) = \frac{1}{n}\sum_{i=1}^{n} h(X_i).$$

(d) Show that if there exists c such that $|g(x)| \le cf(x)$ for all $x \in (a, b)$, then $\mathrm{Var}(M_n(f)) < \infty$.
(e) Determine the standard error of $M_n(f)$ and indicate how you would use this to assess the error in the approximation $M_n(f)$ when $\mathrm{Var}(M_n(f)) < \infty$.

COMPUTER PROBLEMS

4.5.22 Use a Monte Carlo algorithm to approximate $P(X^3 + Y^3 \le 3)$, where $X \sim N(1, 2)$ independently of $Y \sim \mathrm{Gamma}(1, 1)$ based on a large sample (take $n = 10^5$, if possible). Assess the error in the approximation. How large does n have to be to guarantee the estimate is within 0.01 of the true value with virtual certainty? (Hint: Problem 4.5.18.)

4.5.23 Use a Monte Carlo algorithm to approximate $E(X^3 + Y^3)$, where $X \sim N(1, 2)$ independently of $Y \sim \mathrm{Gamma}(1, 1)$ based on a large sample (take $n = 10^5$, if possible). Assess the error in the approximation.

4.5.24 For the integral of Exercise 4.5.3, compare the efficiencies of the algorithm based on generating from an Exponential(5) distribution with that based on generating from an $N(0, 1/7)$ distribution.

CHALLENGES

4.5.25 (*Buffon's needle*) Suppose you drop a needle at random onto a large sheet of lined paper. Assume the distance between the lines is exactly equal to the length of the needle.
(a) Prove that the probability that the needle lands touching a line is equal to $2/\pi$. (Hint: Let D be the distance from the higher end of the needle to the line just below it, and let A be the angle the needle makes with that line. Then what are the distributions of D and A? Under what conditions on D and A will the needle be touching a line?)
(b) Explain how this experiment could be used to obtain a Monte Carlo approximation for the value of π.

4.5.26 *(Optimal importance sampling)* Consider importance sampling as described in Problem 4.5.21.

(a) Prove that $\text{Var}(M_n(f))$ is minimized by taking

$$f(x) = |g(x)| \Big/ \int_a^b |g(x)|\, dx.$$

Calculate the minimum variance and show that the minimum variance is 0 when $g(x) \geq 0$ for all $x \in (a, b)$.

(b) Why is this optimal importance sampler typically not feasible? (The optimal importance sampler does indicate, however, that in our search for an efficient importance sampler, we look for an f that is large when $|g|$ is large and small when $|g|$ is small.)

DISCUSSION TOPICS

4.5.27 An integral like $\int_0^\infty x^2 \cos(x^2) e^{-x}\, dx$ can be approximately computed using a numerical integration computer package (e.g., using Simpson's rule). What are some advantages and disadvantages of using a Monte Carlo approximation instead of a numerical integration package?

4.5.28 Carry out the Buffon's needle Monte Carlo experiment, described in Challenge 4.5.25, by repeating the experiment at least 20 times. Present the estimate of π so obtained. How close is it to the true value of π? What could be done to make the estimate more accurate?

4.6 Normal Distribution Theory

Because of the central limit theorem (Theorem 4.4.3), the normal distribution plays an extremely important role in statistical theory. For this reason, we shall consider a number of important properties and distributions related to the normal distribution. These properties and distributions will be very important for the statistical theory in later chapters of this book.

We already know that if $X_1 \sim N(\mu_1, \sigma_1^2)$ independent of $X_2 \sim N(\mu_2, \sigma_2^2)$, then $cX_1 + d \sim N(c\mu_1 + d, c^2\sigma^2)$ (see Exercise 2.6.3) and $X_1 + X_2 \sim N(\mu_1 + \mu_2,\ \sigma_1^2 +$

σ_2^2) (see Problem 2.9.14). Combining these facts and using induction, we have the following result.

Theorem 4.6.1 Suppose $X_i \sim N(\mu_i, \sigma_i^2)$ for $i = 1, 2, \ldots, n$ and that they are independent random variables. Let $Y = (\sum_i a_i X_i) + b$ for some constants $\{a_i\}$ and b. Then

$$Y \sim N\left((\sum_i a_i \mu_i) + b, \sum_i a_i^2 \sigma_i^2\right).$$

This immediately implies the following.

Corollary 4.6.1 Suppose $X_i \sim N(\mu, \sigma^2)$ for $i = 1, 2, \ldots, n$ and that they are independent random variables. If $\bar{X} = (X_1 + \cdots + X_n)/n$, then $\bar{X} \sim N(\mu, \sigma^2/n)$.

A more subtle property of normal distributions is the following.

Theorem 4.6.2 Suppose $X_i \sim N(\mu_i, \sigma_i^2)$ for $i = 1, 2, \ldots, n$ and also that the $\{X_i\}$ are *independent*. Let $U = \sum_{i=1}^n a_i X_i$ and $V = \sum_{i=1}^n b_i X_i$ for some constants $\{a_i\}$ and $\{b_i\}$. Then $\text{Cov}(U, V) = \sum_i a_i b_i \sigma_i^2$. Furthermore, $\text{Cov}(U, V) = 0$ if and only if U and V are independent.

PROOF The formula for $\text{Cov}(U, V)$ follows immediately from the linearity of covariance (Theorem 3.3.2) because we have

$$\begin{aligned}\text{Cov}(U, V) &= \text{Cov}\left(\sum_{i=1}^n a_i X_i, \sum_{j=1}^n b_j X_j\right) = \sum_{i=1}^n \sum_{j=1}^n a_i b_j \,\text{Cov}(X_i, X_j) \\ &= \sum_{i=1}^n a_i b_i \,\text{Cov}(X_i, X_i) = \sum_{i=1}^n a_i b_i \,\text{Var}(X_i) = \sum_{i=1}^n a_i b_i \sigma_i^2\end{aligned}$$

(note that $\text{Cov}(X_i, X_j) = 0$ for $i \neq j$, by independence). Also, if U and V are independent, then we must have $\text{Cov}(U, V) = 0$ by Corollary 3.3.2.

It remains to prove that, if $\text{Cov}(U, V) = 0$, then U and V are independent. This involves a two-dimensional change of variable, as discussed in the advanced Section 2.9.2, so we refer the reader to Section 4.7 for this part of the proof. ∎

Theorem 4.6.2 says that, for the special case of linear combinations of independent normal distributions, if $\text{Cov}(U, V) = 0$, then U and V are independent. However, it is important to remember that this property is *not* true in general, and there are random variables X and Y such that $\text{Cov}(X, Y) = 0$ even though X and Y are not independent (see Example 3.3.10). Furthermore, this property is not even true of *normal* distributions in general (see Problem 4.6.13).

Note that using linear algebra, we can write the equations $U = \sum_{i=1}^{n} a_i X_i$ and $V = \sum_{i=1}^{n} b_i X_i$ of Theorem 4.6.2 in matrix form as

$$\begin{pmatrix} U \\ V \end{pmatrix} = A \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}, \tag{4.6.1}$$

where

$$A = \begin{pmatrix} a_1 & a_2 & \dots & a_n \\ b_1 & b_2 & \dots & b_n \end{pmatrix}.$$

Furthermore, the rows of A are *orthogonal* if and only if $\sum_i a_i b_i = 0$. Now, in the case $\sigma_i = 1$ for all i, we have that $\text{Cov}(U, V) = \sum_i a_i b_i$. Hence, if $\sigma_i = 1$ for all i, then Theorem 4.6.2 can be interpreted as saying that if U and V are given by (4.6.1), then U and V are independent if and only if the rows of A are orthogonal. Linear algebra is used extensively in more advanced treatments of these ideas.

4.6.1 The Chi-Squared Distribution

We now introduce another distribution, related to the normal distribution.

Definition 4.6.1 The *chi-squared distribution* with n degrees of freedom (or chi-squared(n) , or $\chi^2(n)$) is the distribution of the random variable

$$Z = X_1^2 + X_2^2 + \cdots + X_n^2,$$

where $X_1, \dots, X_n$ are i.i.d., each with the standard normal distribution $N(0, 1)$.

Most statistical packages have built-in routines for the evaluation of chi-squared probabilities (also see Table D.3 in Appendix D).

One property of the chi-squared distribution is easy.

Theorem 4.6.3 If $Z \sim \chi^2(n)$, then $E(Z) = n$.

PROOF Write $Z = X_1^2 + X_2^2 + \cdots + X_n^2$, where $\{X_i\}$ are i.i.d. $\sim N(0, 1)$. Then $E((X_i)^2) = 1$. It follows by linearity that $E(Z) = 1 + \cdots + 1 = n$. ∎

The density function of the chi-squared distribution is a bit harder to obtain. We begin with the case $n = 1$.

Theorem 4.6.4 Let $Z \sim \chi^2(1)$. Then

$$f_Z(z) = \frac{1}{\sqrt{2\pi z}} e^{-z/2} = \frac{(1/2)^{1/2}}{\Gamma(1/2)} z^{-1/2} e^{-z/2}$$

for $z > 0$, with $f_Z(z) = 0$ for $z < 0$. That is, $Z \sim \text{Gamma}(1/2, 1/2)$ (using $\Gamma(1/2) = \sqrt{\pi}$).

PROOF Because $Z \sim \chi^2(1)$, we can write $Z = X^2$, where $X \sim N(0, 1)$. We then compute that, for $z > 0$,

$$\int_{-\infty}^{z} f_Z(s)\,ds = P(Z \le z) = P(X^2 \le z) = P(-\sqrt{z} \le X \le \sqrt{z}).$$

But because $X \sim N(0, 1)$ with density function $\phi(s) = (2\pi)^{-1/2} e^{-s^2/2}$, we can rewrite this as

$$\int_{-\infty}^{z} f_Z(s)\,ds = \int_{-\sqrt{z}}^{\sqrt{z}} \phi(s)\,ds = \int_{-\infty}^{\sqrt{z}} \phi(s)\,ds - \int_{-\infty}^{-\sqrt{z}} \phi(s)\,ds.$$

Because this is true for all $z > 0$, we can differentiate with respect to z (using the fundamental theorem of calculus and the chain rule) to obtain

$$f_Z(z) = \frac{1}{2\sqrt{z}}\phi(\sqrt{z}) - \frac{-1}{2\sqrt{z}}\phi(-\sqrt{z}) = \frac{1}{\sqrt{z}}\phi(\sqrt{z}) = \frac{1}{\sqrt{2\pi z}} e^{-z/2},$$

as claimed. ∎

In Figure 4.6.1, we have plotted the $\chi^2(1)$ density. Note that the density becomes infinite at 0.

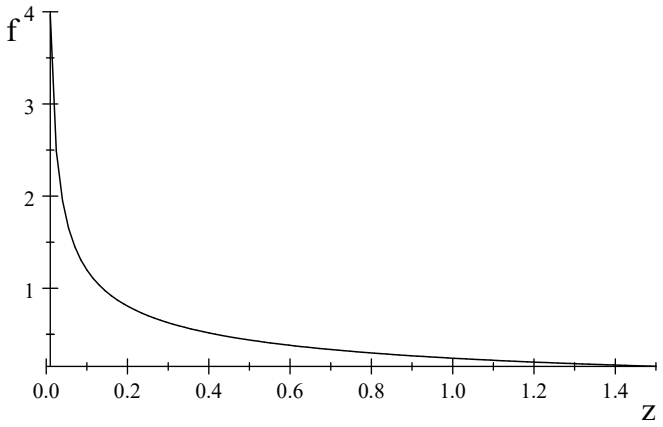

Figure 4.6.1: Plot of the $\chi^2(1)$ density.

Theorem 4.6.5 Let $Z \sim \chi^2(n)$. Then $Z \sim \text{Gamma}(n/2,\ 1/2)$. That is,

$$f_Z(z) = \frac{1}{2^{n/2}\Gamma(n/2)} z^{(n/2)-1} e^{-z/2}$$

for $z > 0$, with $f_Z(z) = 0$ for $z < 0$.

PROOF Because $Z \sim \chi^2(n)$, we can write $Z = X_1^2 + X_2^2 + \cdots + X_n^2$, where the X_i are i.i.d. $N(0, 1)$. But this means that X_i^2 are i.i.d. $\chi^2(1)$. Hence, by Theorem 4.6.4, we have X_i^2 i.i.d. Gamma$(1/2, 1/2)$ for $i = 1, 2, \ldots, n$. Therefore, Z is the sum of n independent random variables, each having distribution Gamma$(1/2, 1/2)$.

Now by Appendix C (see Problem 3.4.20), the moment-generating function of a Gamma(α, β) random variable is given by $m(s) = \beta^\alpha(\beta - s)^{-\alpha}$ for $s < \beta$. Putting $\alpha = 1/2$ and $\beta = 1/2$, and applying Theorem 3.4.5, the variable $Y = X_1^2 + X_2^2 + \cdots + X_n^2$ has moment-generating function given by

$$m_Y(s) = \prod_{i=1}^{n} m_{X_i^2}(s) = \prod_{i=1}^{n} \left(\frac{1}{2}\right)^{1/2} \left(\frac{1}{2} - s\right)^{-1/2} = \left(\frac{1}{2}\right)^{n/2} \left(\frac{1}{2} - s\right)^{-n/2}$$

for $s < 1/2$. We recognize this as the moment-generating function of the Gamma$(n/2, 1/2)$ distribution. Therefore, by Theorem 3.4.6, we have that $X_1^2 + X_2^2 + \cdots + X_n^2 \sim$ Gamma$(n/2, 1/2)$, as claimed.

This result can also be obtained using Problem 2.9.15 and induction. ∎

Note that the $\chi^2(2)$ density is the same as the Exponential(2) density. In Figure 4.6.2, we have plotted several χ^2 densities. Observe that the χ^2 are asymmetric and skewed to the right. As the degrees of freedom increase, the central mass of probability moves to the right.

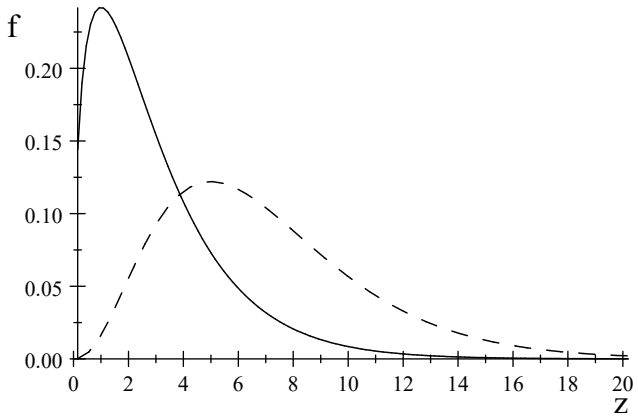

Figure 4.6.2: Plot of the $\chi^2(3)$ (solid line) and the $\chi^2(7)$ (dashed line) density functions.

One application of the chi-squared distribution is the following.

Theorem 4.6.6 Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$. Put

$$\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n) \text{ and } S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Then $(n-1)\,S^2/\sigma^2 \sim \chi^2(n-1)$, and furthermore, S^2 and $\bar{X}$ are independent.

PROOF See Section 4.7 for the proof of this result. ∎

Because the $\chi^2(n-1)$ distribution has mean $n-1$, we obtain the following.

Corollary 4.6.2 $E(S^2) = \sigma^2$.

PROOF Theorems 4.6.6 and 4.6.3 imply that $E((n-1)S^2/\sigma^2) = n - 1$ and that $E(S^2) = \sigma^2$. ∎

Theorem 4.6.6 will find extensive use in Chapter 6. For example, this result, together with Corollary 4.6.1, gives us the joint sampling distribution of the sample mean $\bar{X}$ and the sample variance S^2 when we are sampling from an $N(\mu, \sigma^2)$ distribution. If we do not know μ, then $\bar{X}$ is a natural estimator of this quantity and, similarly, S^2 is a natural estimator of σ^2, when it is unknown. Interestingly, we divide by $n-1$ in S^2, rather than n, precisely because we want $E(S^2) = \sigma^2$ to hold, as in Corollary 4.6.2. Actually, this property does not depend on sampling from a normal distribution. It can be shown that anytime $X_1, \ldots, X_n$ is a sample from a distribution with variance σ^2, then $E(S^2) = \sigma^2$.

4.6.2 The t Distribution

The t distribution also has many statistical applications.

Definition 4.6.2 The *t distribution* with n degrees of freedom (or Student(n), or $t(n)$), is the distribution of the random variable

$$Z = \frac{X}{\sqrt{(X_1^2 + X_2^2 + \cdots + X_n^2)/n}},$$

where $X, X_1, \ldots, X_n$ are i.i.d., each with the standard normal distribution $N(0, 1)$. (Equivalently, $Z = X/\sqrt{Y/n}$, where $Y \sim \chi^2(n)$.)

Most statistical packages have built-in routines for the evaluation of $t(n)$ probabilities (also see Table D.4 in Appendix D).

The density of the $t(n)$ distribution is given by the following result.

Theorem 4.6.7 Let $U \sim t(n)$. Then

$$f_U(u) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{u^2}{n}\right)^{-(n+1)/2} \frac{1}{\sqrt{n}}$$

for all $u \in R^1$.

PROOF For the proof of this result, see Section 4.7. ∎

The following result shows that, when n is large, the $t(n)$ distribution is very similar to the $N(0, 1)$ distribution.

Theorem 4.6.8 As $n \to \infty$, the $t(n)$ distribution converges in distribution to a standard normal distribution.

PROOF Let $Z_1, \ldots, Z_n, Z$ be i.i.d. $N(0, 1)$. As $n \to \infty$, by the strong law of large numbers, $(Z_1^2 + \cdots + Z_n^2)/n$ converges with probability 1 to the constant 1. Hence, the distribution of

$$\frac{Z}{\sqrt{\left(Z_1^2 + \cdots + Z_n^2\right)/n}} \tag{4.6.2}$$

converges to the distribution of Z, which is the standard normal distribution. By Definition 4.6.2, we have that (4.6.2) is distributed $t(n)$. ∎

In Figure 4.6.3, we have plotted several t densities. Notice that the densities of the t distributions are symmetric about 0 and look like the standard normal density.

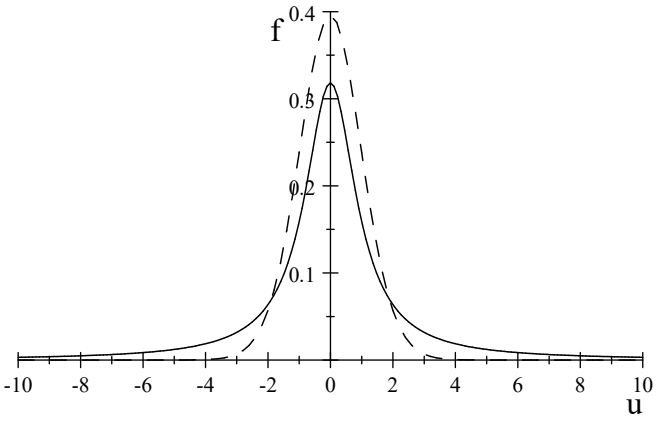

Figure 4.6.3: Plot of the $t(1)$ (solid line) and the $t(30)$ (dashed line) density functions.

The $t(n)$ distribution has longer tails than the $N(0, 1)$ distribution. For example, the $t(1)$ distribution (also known as the *Cauchy distribution*) has 0.9366 of its probability in the interval $(-10, 10)$, whereas the $N(0, 1)$ distribution has all of its probability there (at least to four decimal places). The $t(30)$ and the $N(0, 1)$ densities are very similar.

4.6.3 The F Distribution

Finally, we consider the F distribution.

Definition 4.6.3 The *F distribution* with m and n degrees of freedom (or $F(m, n)$) is the distribution of the random variable

$$Z = \frac{(X_1^2 + X_2^2 + \cdots + X_m^2)\,/\,m}{(Y_1^2 + Y_2^2 + \cdots + Y_n^2)\,/\,n},$$

where $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ are i.i.d., each with the standard normal distribution. (Equivalently, $Z = (X/m)/(Y/n)$, where $X \sim \chi^2(m)$ and $Y \sim \chi^2(n)$.)

Most statistical packages have built-in routines for the evaluation of $F(m, n)$ probabilities (also see Table D.5 in Appendix D).

The density of the $F(m, n)$ distribution is given by the following result.

Theorem 4.6.9 Let $U \sim F(m, n)$. Then

$$f_U(u) = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \left(\frac{m}{n}u\right)^{(m/2)-1} \left(1 + \frac{m}{n}u\right)^{-(m+n)/2} \frac{m}{n}$$

for $u > 0$, with $f_U(u) = 0$ for $u < 0$.

PROOF For the proof of this result, see Section 4.7. ∎

In Figure 4.6.4, we have plotted several $F(m, n)$ densities. Notice that these densities are skewed to the right.

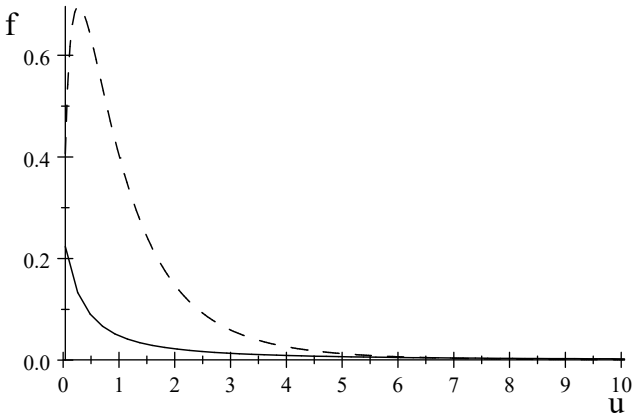

Figure 4.6.4: Plot of the $F(2, 1)$ (solid line) and the $F(3, 10)$ (dashed line) density functions.

The following results are useful when it is necessary to carry out computations with the $F(m, n)$ distribution.

Theorem 4.6.10 If $Z \sim F(m, n)$, then $1/Z \sim F(n, m)$.

PROOF Using Definition 4.6.3, we have

$$\frac{1}{Z} = \frac{(Y_1^2 + Y_2^2 + \cdots + Y_n^2)/n}{(X_1^2 + X_2^2 + \cdots + X_m^2)/m}$$

and the result is immediate from the definition. ∎

Therefore, if $Z \sim F(m, n)$, then $P(Z \leq z) = P(1/Z \geq 1/z) = 1 - P(1/Z \leq 1/z)$, and $P(1/Z \leq 1/z)$ is the cdf of the $F(n, m)$ distribution evaluated at $1/z$.

In many statistical applications, n can be very large. The following result then gives a useful approximation for that case.

Theorem 4.6.11 If $Z_n \sim F(m, n)$, then mZ_n converges in distribution to a $\chi^2(m)$ distribution as $n \to \infty$.

PROOF Using Definition 4.6.3, we have

$$mZ = \frac{X_1^2 + X_2^2 + \cdots + X_m^2}{(Y_1^2 + Y_2^2 + \cdots + Y_n^2)/n}.$$

By Definition 4.6.1, $X_1^2 + \cdots + X_m^2 \sim \chi^2(m)$. By Theorem 4.6.3, $E(Y_i^2) = 1$, so the strong law of large numbers implies that $(Y_1^2 + Y_2^2 + \cdots + Y_n^2)/n$ converges almost surely to 1. This establishes the result. ∎

Finally, Definitions 4.6.2 and 4.6.3 immediately give the following result.

Theorem 4.6.12 If $Z \sim t(n)$, then $Z^2 \sim F(1, n)$.

Summary of Section 4.6

- Linear combinations of independent normal random variables are also normal, with appropriate mean and variance.
- Two linear combinations of the *same* collection of independent normal random variables are independent if and only if their covariance equals 0.
- The chi-squared distribution with n degrees of freedom is the distribution corresponding to a sum of squares of n i.i.d. standard normal random variables. It has mean n. It is equal to the Gamma$(n/2,\ 1/2)$ distribution.
- The t distribution with n degrees of freedom is the distribution corresponding to a standard normal random variable, divided by the square-root of $1/n$ times an independent chi-squared random variable with n degrees of freedom. Its density function was presented. As $n \to \infty$, it converges in distribution to a standard normal distribution.
- The F distribution with m and n degrees of freedom is the distribution corresponding to m/n times a chi-squared distribution with m degrees of freedom, divided by an independent chi-squared distribution with n degrees of freedom.

Its density function was presented. If t has a $t(n)$ distribution, then t^2 is distributed $F(1, n)$.

EXERCISES

4.6.1 Let $X_1 \sim N(3, 2^2)$ and $X_2 \sim N(-8, 5^2)$ be independent. Let $U = X_1 - 5X_2$ and $V = -6X_1 + CX_2$, where C is a constant.
(a) What are the distributions of U and V?
(b) What value of C makes U and V be independent?

4.6.2 Let $X \sim N(3, 5)$ and $Y \sim N(-7, 2)$ be independent.
(a) What is the distribution of $Z = 4X - Y/3$?
(b) What is the covariance of X and Z?

4.6.3 Let $X \sim N(3, 5)$ and $Y \sim N(-7, 2)$ be independent. Find values of $C_1 \neq 0, C_2, C_3 \neq 0, C_4, C_5$ so that $C_1(X + C_2)^2 + C_3(Y + C_4)^2 \sim \chi^2(C_5)$.

4.6.4 Let $X \sim \chi^2(n)$ and $Y \sim N(0, 1)$ be independent. Prove that $X + Y^2 \sim \chi^2(n + 1)$.

4.6.5 Let $X \sim \chi^2(n)$ and $Y \sim \chi^2(m)$ be independent. Prove that $X + Y \sim \chi^2(n + m)$.

4.6.6 Let $X_1, X_2, \ldots, X_{4n}$ be i.i.d. with distribution $N(0, 1)$. Find a value of C such that

$$C \frac{X_1^2 + X_2^2 + \cdots + X_n^2}{X_{n+1}^2 + X_{n+2}^2 + \cdots + X_{4n}^2} \sim F(n, 3n).$$

4.6.7 Let $X_1, X_2, \ldots, X_{n+1}$ be i.i.d. with distribution $N(0, 1)$. Find a value of C such that

$$C \frac{X_1}{\sqrt{X_2^2 + \cdots + X_n^2 + X_{n+1}^2}} \sim t\,(n)\,.$$

4.6.8 Let $X \sim N(3, 5)$ and $Y \sim N(-7, 2)$ be independent. Find values of $C_1, C_2, C_3, C_4, C_5, C_6$ so that

$$\frac{C_1(X + C_2)^{C_3}}{((Y + C_4)^2)^{C_5}} \sim t(C_6).$$

4.6.9 Let $X \sim N(3, 5)$ and $Y \sim N(-7, 2)$ be independent. Find values of $C_1, C_2, C_3, C_4, C_5, C_6, C_7$ so that

$$\frac{C_1(X + C_2)^{C_3}}{(Y + C_4)^{C_5}} \sim F(C_6, C_7).$$

4.6.10 Let $X_1, X_2, \ldots, X_{100}$ be independent, each with the standard normal distribution.
(a) Compute the distribution of X_1^2.
(b) Compute the distribution of $X_3^2 + X_5^2$.
(c) Compute the distribution of $X_{10}/\sqrt{[X_{20}^2 + X_{30}^2 + X_{40}^2]/3}$.
(d) Compute the distribution of $3X_{10}^2 / [X_{20}^2 + X_{30}^2 + X_{40}^2]$.

(e) Compute the distribution of

$$\frac{30}{70}\,\frac{X_1^2 + X_2^2 + \cdots + X_{70}^2}{X_{71}^2 + X_{72}^2 + \cdots + X_{100}^2}.$$

4.6.11 Let $X_1, X_2, \ldots, X_{61}$ be independent, each distributed as $N(\mu, \sigma^2)$. Set $\bar{X} = (1/61)(X_1 + X_2 + \cdots + X_{61})$ and

$$S^2 = \frac{1}{60}\Big[(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_{61} - \bar{X})^2\Big]$$

as usual.

(a) For what values of K and m is it true that the quantity $Y \equiv K(\bar{X} - \mu)/\sqrt{S^2}$ has a t distribution with m degrees of freedom?

(b) With K as in part (a), find y such that $P(Y \geq y) = 0.05$.

(c) For what values of a and b and c is it true that the quantity $W \equiv a(\bar{X} - \mu)^2/S^2$ has an F distribution with b and c degrees of freedom?

(d) For those values of a and b and c, find a quantity w so that $P(W \geq w) = 0.05$.

4.6.12 Suppose the core temperature (in degrees celsius, when used intensively) of the latest Dell desktop computer is normally distributed with mean 40 and standard deviation 5, while for the latest Compaq it is normally distributed with mean 45 and standard deviation 8. Suppose we measure the Dell temperature 20 times (on separate days) and obtain measurements $D_1, D_2, \ldots, D_{20}$, and we also measure the Compaq temperature 30 times and obtain measurements $C_1, C_2, \ldots, C_{30}$.

(a) Compute the distribution of $\bar{D} \equiv (D_1 + \cdots + D_{20})/20$.

(b) Compute the distribution of $\bar{C} \equiv (C_1 + \cdots + C_{30})/30$.

(c) Compute the distribution of $Z \equiv \bar{C} - \bar{D}$.

(d) Compute $P(\bar{C} < \overline{D})$.

(e) Let $U = (D_1 - \bar{D})^2 + (D_2 - \bar{D})^2 + \cdots + (D_{20} - \bar{D})^2$. What is $P(U > 633.25)$?

PROBLEMS

4.6.13 Let $X \sim N(0, 1)$, and let $P(Y = 1) = P(Y = -1) = 1/2$. Assume X and Y are independent. Let $Z = XY$.

(a) Prove that $Z \sim N(0, 1)$.

(b) Prove that $\mathrm{Cov}(X, Z) = 0$.

(c) Prove directly that X and Z are *not* independent.

(d) Why does this not contradict Theorem 4.6.2?

4.6.14 Let $Z \sim t(n)$. Prove that $P(Z < -x) = P(Z > x)$ for $x \in R^1$, namely, prove that the $t(n)$ distribution is symmetric about 0.

4.6.15 Let $X_n \sim F(n, 2n)$ for $n = 1, 2, 3, \ldots$. Prove that $X_n \to 1$ in probability and with probability 1.

4.6.16 (*The general chi-squared distribution*) Prove that for $\alpha > 0$, the function

$$f(z) = \frac{1}{2^{\alpha/2}\Gamma(\alpha/2)} z^{(\alpha/2)-1} e^{-z/2}$$

defines a probability distribution on $(0, \infty)$. This distribution is known as the $\chi^2(\alpha)$ distribution, i.e., it generalizes the distribution in Section 4.6.2 by allowing the degrees of freedom to be an arbitrary positive real number. (Hint: The $\chi^2(\alpha)$ distribution is the same as a Gamma$(\alpha/2, 1/2)$ distribution.)

4.6.17 (MV) (*The general t distribution*) Prove that for $\alpha > 0$, the function

$$f(u) = \frac{\Gamma\left(\frac{\alpha+1}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{\alpha}{2}\right)}\left(1+\frac{u^2}{\alpha}\right)^{-(\alpha+1)/2}\frac{1}{\sqrt{\alpha}}$$

defines a probability distribution on $(-\infty, \infty)$ by showing that the random variable

$$U = \frac{X}{\sqrt{Y/\alpha}}$$

has this density when $X \sim N(0, 1)$ independent of $Y \sim \chi^2(\alpha)$, as in Problem 4.6.16. This distribution is known as the $t(\alpha)$ distribution, i.e., it generalizes the distribution in Section 4.6.3 by allowing the degrees of freedom to be an arbitrary positive real number. (Hint: The proof is virtually identical to that of Theorem 4.6.7.)

4.6.18 (MV) (*The general F distribution*) Prove that for $\alpha > 0, \beta > 0$, the function

$$f(u) = \frac{\Gamma\left(\frac{\alpha+\beta}{2}\right)}{\Gamma\left(\frac{\alpha}{2}\right)\Gamma\left(\frac{\beta}{2}\right)}\left(\frac{\alpha}{\beta}u\right)^{(\alpha/2)-1}\left(1+\frac{\alpha}{\beta}u\right)^{-(\alpha+\beta)/2}\frac{\alpha}{\beta}$$

defines a probability distribution on $(0, \infty)$ by showing that the random variable

$$U = \frac{X/\alpha}{Y/\beta}$$

has this density whenever $X \sim \chi^2(\alpha)$ independent of $Y \sim \chi^2(\beta)$, as in Problem 4.6.16. This distribution is known as the $F(\alpha, \beta)$ distribution, i.e., it generalizes the distribution in Section 4.6.4 by allowing the numerator and denominator degrees of freedom to be arbitrary positive real numbers. (Hint: The proof is virtually identical to that of Theorem 4.6.9).

4.6.19 Prove that when $X \sim t(\alpha)$, as defined in Problem 4.6.17, and $\alpha > 1$, then $E(X) = 0$. Further prove that when $\alpha > 2$, $\text{Var}(X) = \alpha/(\alpha - 2)$. You can assume the existence of these integrals — see Challenge 4.6.21. (Hint: To evaluate the second moment, use $Y = X^2 \sim F(1, \alpha)$ as defined in Problem 4.6.18.)

4.6.20 Prove that when $X \sim F(\alpha, \beta)$, then $E(X) = \beta/(\beta - 2)$ when $\beta > 2$ and $\text{Var}(X) = 2\beta^2(\alpha + \beta - 2)/\alpha(\beta - 2)^2(\beta - 4)$ when $\beta > 4$.

CHALLENGES

4.6.21 Following Problem 4.6.19, prove that the mean of X does not exist whenever $0 < \alpha \leq 1$. Further prove that the variance of X does not exist whenever $0 < \alpha \leq 1$ and is infinite when $1 < \alpha \leq 2$.

4.6.22 Prove the identity (4.7.1) in Section 4.7, which arises as part of the proof of Theorem 4.6.6.

4.7 Further Proofs (Advanced)

Proof of Theorem 4.3.1

We want to prove the following result. Let $Z, Z_1, Z_2, \ldots$ be random variables. Suppose $Z_n \to Z$ with probability 1. Then $Z_n \to Z$ in probability. That is, if a sequence of random variables converges almost surely, then it converges in probability to the same limit.

Assume $P(Z_n \to Z) = 1$. Fix $\epsilon > 0$, and let $A_n = \{s : |Z_m - Z| \geq \epsilon \text{ for some } m \geq n\}$. Then $\{A_n\}$ is a decreasing sequence of events. Furthermore, if $s \in \cap_{n=1}^{\infty} A_n$, then $Z_n(s) \nrightarrow Z(s)$ as $n \to \infty$. Hence,

$$P\left(\cap_{n=1}^{\infty} A_n\right) \leq P(Z_n \nrightarrow Z) = 0.$$

By continuity of probabilities, we have $\lim_{n\to\infty} P(A_n) = P(\cap_{n=1}^{\infty} A_n) = 0$. Hence, $P(|Z_n - Z| \geq \epsilon) \leq P(A_n) \to 0$ as $n \to \infty$. Because this is true for any $\epsilon > 0$, we see that $Z_n \to Z$ in probability. ∎

Proof of Theorem 4.4.1

We show that if $X_n \xrightarrow{P} X$, then $X_n \xrightarrow{D} X$.

Suppose $X_n \to X$ in probability and that $P(X = x) = 0$. We wish to show that $\lim_{n\to\infty} P(X_n \leq x) = P(X \leq x)$.

Choose any $\epsilon > 0$. Now, if $X_n \leq x$, then we must have either $X \leq x + \epsilon$ or $|X - X_n| \geq \epsilon$. Hence, by subadditivity,

$$P(X_n \leq x) \leq P(X \leq x + \epsilon) + P(|X - X_n| \geq \epsilon).$$

Replacing x by $x - \epsilon$ in this equation, we see also that

$$P(X \leq x - \epsilon) \leq P(X_n \leq x) + P(|X - X_n| \geq \epsilon).$$

Rearranging and combining these two inequalities, we have

$$P(X \leq x-\epsilon)-P(|X-X_n| \geq \epsilon) \leq P(X_n \leq x) \leq P(X \leq x+\epsilon)+P(|X-X_n| \geq \epsilon).$$

This is the key.

We next let $n \to \infty$. Because $X_n \to X$ in probability, we know that

$$\lim_{n\to\infty} P(|X - X_n| \geq \epsilon) = 0.$$

This means that $\lim_{n\to\infty} P(X_n \leq x)$ is "sandwiched" between $P(X \leq x - \epsilon)$ and $P(X \leq x + \epsilon)$.

We then let $\epsilon \searrow 0$. By continuity of probabilities,

$$\lim_{\epsilon\searrow 0} P(X \leq x + \epsilon) = P(X \leq x) \text{ and } \lim_{\epsilon\searrow 0} P(X \leq x - \epsilon) = P(X < x).$$

This means that $\lim_{n\to\infty} P(X_n \leq x)$ is "sandwiched" between $P(X < x)$ and $P(X \leq x)$.

But because $P(X = x) = 0$, we must have $P(X < x) = P(X \leq x)$. Hence, $\lim_{n\to\infty} P(X_n \leq x) = P(X \leq x)$, as required. ∎

Proof of Theorem 4.4.3 (*The central limit theorem*)

We must prove the following. Let $X_1, X_2, \ldots$ *be i.i.d. with finite mean* μ *and finite variance* σ^2. *Let* $Z \sim N(0,1)$. *Set* $S_n = X_1 + \cdots + X_n$, *and*

$$Z_n = \frac{S_n - n\mu}{\sqrt{n\sigma^2}}.$$

Then as $n \to \infty$, *the sequence* $\{Z_n\}$ *converges in distribution to the* Z, *i.e.,* $Z_n \xrightarrow{D} Z$.

Recall that the standard normal distribution has moment-generating function given by $m_Z(s) = \exp\left(s^2/2\right)$.

We shall now *assume* that $m_{Z_n}(s)$ is finite for $|s| < s_0$ for some $s_0 > 0$. (This assumption can be eliminated by using *characteristic functions* instead of moment-generating functions.) Assuming this, we will prove that for each real number s, we have $\lim_{n\to\infty} m_{Z_n}(s) = m_Z(s)$, where $m_{Z_n}(s)$ is the moment-generating function of Z_n. It then follows from Theorem 4.4.2 that Z_n converges to Z in distribution.

To proceed, let $Y_i = (X_i - \mu)/\sigma$. Then $E(Y_i) = 0$ and $E(Y_i^2) = \text{Var}(Y_i) = 1$. Also, we have

$$Z_n = \frac{1}{\sqrt{n}}(Y_1 + \cdots + Y_n).$$

Let $m_Y(s) = E(e^{sY_i})$ be the moment-generating function of Y_i (which is the same for all i, because they are i.i.d.). Then using independence, we compute that

$$\begin{aligned}
\lim_{n\to\infty} m_{Z_n}(s) &= \lim_{n\to\infty} E\left(e^{sZ_n}\right) = \lim_{n\to\infty} E\left(e^{s(Y_1+\cdots+Y_n)/\sqrt{n}}\right) \\
&= \lim_{n\to\infty} E\left(e^{sY_1/\sqrt{n}}\, e^{sY_2/\sqrt{n}} \cdots e^{sY_n/\sqrt{n}}\right) \\
&= \lim_{n\to\infty} E\left(e^{sY_1/\sqrt{n}}\right) E\left(e^{sY_2/\sqrt{n}}\right) \cdots E\left(e^{sY_n/\sqrt{n}}\right) \\
&= \lim_{n\to\infty} m_Y(s/\sqrt{n}) m_Y(s/\sqrt{n}) \cdots m_Y(s/\sqrt{n}) \\
&= \lim_{n\to\infty} m_Y(s/\sqrt{n})^n.
\end{aligned}$$

Now, we know from Theorem 3.5.3 that $m_Y(0) = E(e^0) = 1$. Also, $m'_Y(0) = E(Y_i) = 0$ and $m''_Y(0) = E(Y_i^2) = 1$. But then expanding $m_Y(s)$ in a Taylor series around $s = 0$, we see that

$$m_Y(s) = 1 + 0s + \frac{1}{2!}s^2 + o(s^2) = 1 + s^2/2 + o(s^2),$$

where $o(s^2)$ stands for a quantity that, as $s \to 0$, goes to 0 faster than s^2 does — namely, $o(s^2)/s \to 0$ as $s \to 0$. This means that

$$m_Y(s/\sqrt{n}) = 1 + (s/\sqrt{n})^2/2 + o((s/\sqrt{n})^2) = 1 + s^2/2n + o(1/n),$$

where now $o(1/n)$ stands for a quantity that, as $n \to \infty$, goes to 0 faster than $1/n$ does.

Finally, we recall from calculus that, for any real number c, $\lim_{n\to\infty}(1+c/n)^n = e^c$. It follows from this and the above that

$$\lim_{n\to\infty}(m_Y(s/2\sqrt{n}))^n = \lim_{n\to\infty}(1+s^2/2n)^n = e^{s^2/2}.$$

That is, $\lim_{n\to\infty} m_{Z_n}(s) = e^{s^2/2}$, as claimed. ∎

Proof of Theorem 4.6.2

We prove the following. Suppose $X_i \sim N(\mu_i, \sigma_i^2)$ for $i = 1, 2, \ldots, n$ and also that the $\{X_i\}$ are independent. Let $U = \sum_{i=1}^n a_i X_i$ and $V = \sum_{i=1}^n b_i X_i$, for some constants $\{a_i\}$ and $\{b_i\}$. Then $Cov(U, V) = \sum_i a_i b_i \sigma_i^2$. Furthermore, $Cov(U, V) = 0$ if and only if U and V are independent.

It was proved in Section 4.6 that $\text{Cov}(U, V) = \sum_i a_i b_i \sigma_i^2$ and that $\text{Cov}(U, V) = 0$ if U and V are independent. It remains to prove that, if $\text{Cov}(U, V) = 0$, then U and V are independent. For simplicity, we take $n = 2$ and $\mu_1 = \mu_2 = 0$ and $\sigma_1^2 = \sigma_2^2 = 1$; the general case is similar but messier. We therefore have

$$U = a_1X_1 + a_2X_2 \quad \text{and} \quad V = b_1X_1 + b_2X_2.$$

The Jacobian derivative of this transformation is

$$J(x_1, x_2) = \frac{\partial U}{\partial X_1}\frac{\partial V}{\partial X_2} - \frac{\partial V}{\partial X_1}\frac{\partial U}{\partial X_2} = a_1b_2 - b_1a_2.$$

Inverting the transformation gives

$$X_1 = \frac{b_2U - a_2V}{a_1b_2 - b_1a_2} \quad \text{and} \quad X_2 = \frac{a_1V - b_1U}{a_1b_2 - b_1a_2}.$$

Also,

$$f_{X_1,X_2}(x_1, x_2) = \frac{1}{2\pi}e^{-(x_1^2+x_2^2)/2}.$$

Hence, from the multidimensional change of variable theorem (Theorem 2.9.2), we have

$$\begin{aligned} f_{U,V}(u, v) &= f_{X_1,X_2}(x_1, x_2)\left(\frac{b_2u - a_2v}{a_1b_2 - b_1a_2}, \frac{a_1v - b_1u}{a_1b_2 - b_1a_2}\right)|J(x_1, x_2)|^{-1} \\ &= \frac{1}{2\pi}\frac{\exp\{-((b_2u - a_2v)^2 + (a_1v - b_1u)^2)/2(a_1b_2 - b_1a_2)^2\}}{|a_1b_2 - b_1a_2|}. \end{aligned}$$

But

$$(b_2u - a_2v)^2 + (a_1v - b_1u)^2 = (b_1^2 + b_2^2)u^2 + (a_1^2 + a_2^2)v^2 - 2(a_1b_1 + a_2b_2)uv$$

and $\text{Cov}(U, V) = a_1b_1 + a_2b_2$. Hence, if $\text{Cov}(U, V) = 0$, then

$$(b_2u - a_2v)^2 + (a_1v - b_1u)^2 = (b_1^2 + b_2^2)u^2 + (a_1^2 + a_2^2)v^2$$

and

$$f_{U,V}(u,v) = \frac{\exp\{-((b_1^2+b_2^2)u^2+(a_1^2+a_2^2)v^2)/2(a_1b_2-b_1a_2)^2\}}{2\pi|a_1b_2-b_1a_2|}$$
$$= \frac{\exp\{-(b_1^2+b_2^2)u^2/2(a_1b_2-b_1a_2)^2\}\exp\{-(a_1^2+a_2^2)v^2/2(a_1b_2-b_1a_2)^2\}}{2\pi|a_1b_2-b_1a_2|}.$$

It follows that we can *factor* $f_{U,V}(u,v)$ as a function of u times a function of v. But this implies (see Problem 2.8.19) that U and V are independent. ∎

Proof of Theorem 4.6.6

We want to prove that when $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$ and

$$\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n) \quad \text{and} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

then $(n-1)\,S^2/\sigma^2 \sim \chi^2(n-1)$ and, furthermore, that S^2 and $\bar{X}$ are independent.

We have

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

We rewrite this expression as (see Challenge 4.6.22)

$$\frac{n-1}{\sigma^2}S^2$$
$$= \left(\frac{X_1 - X_2}{\sigma\sqrt{2}}\right)^2 + \left(\frac{X_1 + X_2 - 2X_3}{\sigma\sqrt{2\cdot 3}}\right)^2 + \left(\frac{X_1 + X_2 + X_3 - 3X_4}{\sigma\sqrt{3\cdot 4}}\right)^2$$
$$+ \cdots + \left(\frac{X_1 + X_2 + \cdots + X_{n-1} - (n-1)X_n}{\sigma\sqrt{(n-1)n}}\right)^2. \tag{4.7.1}$$

Now, by Theorem 4.6.1, each of the $n-1$ expressions within brackets in (4.7.1) has the standard normal distribution. Furthermore, by Theorem 4.6.2, the expressions within brackets in (4.7.1) are all *independent* of one another and are also all independent of $\bar{X}$.

It follows that $(n-1)\,S^2/\sigma^2$ is independent of $\bar{X}$. It also follows, by the definition of the chi-squared distribution, that $(n-1)\,S^2/\sigma^2 \sim \chi^2(n-1)$. ∎

Proof of Theorem 4.6.7

We want to show that when $U \sim t(n)$, then

$$f_U(u) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{n}{2}\right)}\left(1 + \frac{u^2}{n}\right)^{-(n+1)/2}\frac{1}{\sqrt{n}}$$

for all $u \in R^1$.

Because $U \sim t(n)$, we can write $U = X/\sqrt{Y/n}$, where X and Y are independent with $X \sim N(0,1)$ and $Y \sim \chi^2(n)$. It follows that X and Y have joint density given by

$$f_{X,Y}(x,y) = \frac{e^{-x^2/2} y^{(n/2)-1} e^{-y/2}}{\sqrt{2\pi}\, 2^{n/2} \Gamma\left(\frac{n}{2}\right)}$$

when $y > 0$ (with $f_{X,Y}(x,y) = 0$ for $y < 0$).

Let $V = Y$. We shall use the multivariate change of variables formula (Theorem 2.9.2) to compute the joint density $f_{U,V}(u,v)$ of U and V. Because $U = X/\sqrt{Y/n}$ and $V = Y$, it follows that $X = U\sqrt{V/n}$ and $Y = V$. We compute the Jacobian term as

$$J(x,y) = \det \begin{pmatrix} \frac{\partial u}{\partial x} & \frac{\partial v}{\partial x} \\ \frac{\partial u}{\partial y} & \frac{\partial v}{\partial y} \end{pmatrix} = \det \begin{pmatrix} \frac{1}{\sqrt{y/n}} & 0 \\ \frac{-x\sqrt{n}}{y^{3/2}} & 1 \end{pmatrix} = \frac{1}{\sqrt{y/n}}.$$

Hence,

$$\begin{aligned} f_{U,V}(u,v) &= f_{X,Y}\left(u\sqrt{\frac{v}{n}}, v\right) J^{-1}\left(u\sqrt{\frac{v}{n}}, v\right) \\ &= \frac{e^{-(u^2 v)/2n} v^{(n/2)-1} e^{-v/2}}{\sqrt{2\pi}\, 2^{n/2} \Gamma\left(\frac{n}{2}\right)} \sqrt{\frac{v}{n}} \\ &= \frac{1}{\sqrt{\pi}\,\Gamma(n/2)} \frac{1}{2^{(n+1)/2}} \frac{1}{\sqrt{n}} v^{(n+1)/2-1} e^{-(v/2)\left(1+u^2/n\right)} \end{aligned}$$

for $v > 0$ (with $f_{U,V}(u,v) = 0$ for $v < 0$).

Finally, we compute the marginal density of U:

$$\begin{aligned} f_U(u) &= \int_{-\infty}^{\infty} f_{U,V}(u,v)\, dv \\ &= \frac{1}{\sqrt{\pi}\,\Gamma(n/2)} \frac{1}{2^{(n+1)/2}} \frac{1}{\sqrt{n}} \int_0^{\infty} v^{(n+1)/2-1} e^{-(v/2)\left(1+u^2/n\right)}\, dv \\ &= \frac{1}{\sqrt{\pi}\,\Gamma(n/2)} \left(1 + \frac{u^2}{n}\right)^{-(n+1)/2} \frac{1}{\sqrt{n}} \int_0^{\infty} w^{(n+1)/2-1} e^{-w/2}\, dw \\ &= \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi}\,\Gamma(n/2)} \left(1 + \frac{u^2}{n}\right)^{-(n+1)/2} \frac{1}{\sqrt{n}}, \end{aligned}$$

where we have made the substitution $w = \left(1 + u^2/n\right) v/2$ to get the third equality and then used the definition of the gamma function to obtain the result. ∎

Proof of Theorem 4.6.9

We want to show that when $U \sim F(m,n)$, then

$$f_U(u) = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \left(\frac{m}{n}u\right)^{(m/2)-1} \left(1 + \frac{m}{n}u\right)^{-(m+n)/2} \frac{m}{n}$$

for $u > 0$, with $f_U(u) = 0$ for $u < 0$.

Because $U \sim F(n, m)$, we can write $U = (X/m)/(Y/n)$, where X and Y are independent with $X \sim \chi^2(m)$ and $Y \sim \chi^2(n)$. It follows that X and Y have joint density given by

$$f_{X,Y}(x, y) = \frac{x^{(m/2)-1} e^{-x/2} y^{(n/2)-1} e^{-y/2}}{2^{m/2}\Gamma\left(\frac{m}{2}\right) 2^{n/2}\Gamma\left(\frac{n}{2}\right)}$$

when $x, y > 0$ (with $f_{X,Y}(x, y) = 0$ for $x < 0$ or $y < 0$).

Let $V = Y$, and use the multivariate change of variables formula (Theorem 2.9.2) to compute the joint density $f_{U,V}(u, v)$ of U and V. Because $U = (X/m)/(Y/n)$ and $V = Y$, it follows that $X = (m/n)UV$ and $Y = V$. We compute the Jacobian term as

$$J(x, y) = \det \begin{pmatrix} \frac{\partial u}{\partial x} & \frac{\partial v}{\partial x} \\ \frac{\partial u}{\partial y} & \frac{\partial v}{\partial y} \end{pmatrix} = \det \begin{pmatrix} \frac{n}{my} & 0 \\ \frac{-nX}{mY^2} & 1 \end{pmatrix} = \frac{n}{my}.$$

Hence,

$$\begin{aligned} f_{U,V}(u, v) &= f_{X,Y}((m/n)uv,\ v) J^{-1}((m/n)uv,\ v) \\ &= \frac{\left(\frac{m}{n}uv\right)^{(m/2)-1} e^{-(m/n)(uv/2)} v^{(n/2)-1} e^{-(v/2)}}{2^{m/2}\Gamma\left(\frac{m}{2}\right) 2^{n/2}\Gamma\left(\frac{n}{2}\right)} \frac{m}{n} v \\ &= \frac{1}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \left(\frac{m}{n}u\right)^{(m/2)-1} \frac{m}{n} \frac{1}{2^{(m+n)/2}} v^{(m+n)/2-1} e^{-(v/2)(1+mu/n)} \end{aligned}$$

for $u, v > 0$ (with $f_{U,V}(u, v) = 0$ for $u < 0$ or $v < 0$).

Finally, we compute the marginal density of U as

$$\begin{aligned} & f_U(u) \\ &= \int_{-\infty}^{\infty} f_{U,V}(u, v)\, dv \\ &= \frac{1}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \left(\frac{m}{n}u\right)^{(m/2)-1} \frac{m}{n} \frac{1}{2^{(m+n)/2}} \int_0^{\infty} v^{(m+n)/2-1} e^{-(v/2)(1+mu/n)}\, dv \\ &= \frac{1}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \left(\frac{m}{n}u\right)^{(m/2)-1} \left(1 + \frac{m}{n}u\right)^{-(n+m)/2} \frac{m}{n} \int_0^{\infty} w^{(m+n)/2-1} e^{-w}\, dw \\ &= \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \left(\frac{m}{n}u\right)^{(m/2)-1} \left(1 + \frac{m}{n}u\right)^{-(n+m)/2} \frac{m}{n}, \end{aligned}$$

where we have used the substitution $w = (1 + mu/n)\, v/2$ to get the third equality, and the final result follows from the definition of the gamma function. ∎