Chapter 5

Properties of a Random Sample

"I'm afraid that I rather give myself away when I explain," said he. "Results without causes are much more impressive."

Sherlock Holmes
The Stock-Broker's Clerk

5.1 Basic Concepts of Random Samples

Often, the data collected in an experiment consist of several observations on a variable of interest. We discussed examples of this at the beginning of Chapter 4. In this chapter, we present a model for data collection that is often used to describe this situation, a model referred to as random sampling. The following definition explains mathematically what is meant by the random sampling method of data collection.

Definition 5.1.1 The random variables $X_1, \ldots, X_n$ are called a *random sample of size n from the population* $f(x)$ if $X_1, \ldots, X_n$ are mutually independent random variables and the marginal pdf or pmf of each X_i is the same function $f(x)$. Alternatively, $X_1, \ldots, X_n$ are called *independent and identically distributed random variables with pdf or pmf* $f(x)$. This is commonly abbreviated to iid random variables.

The random sampling model describes a type of experimental situation in which the variable of interest has a probability distribution described by $f(x)$. If only one observation X is made on this variable, then probabilities regarding X can be calculated using $f(x)$. In most experiments there are $n > 1$ (a fixed, positive integer) repeated observations made on the variable, the first observation is X_1, the second is X_2, and so on. Under the random sampling model each X_i is an observation on the same variable and each X_i has a marginal distribution given by $f(x)$. Furthermore, the observations are taken in such a way that the value of one observation has no effect on or relationship with any of the other observations; that is, $X_1, \ldots, X_n$ are *mutually independent.* (See Exercise 5.4 for a generalization of independence.)

From Definition 4.6.5, the joint pdf or pmf of $X_1, \ldots, X_n$ is given by

$$f(x_1, \ldots, x_n) = f(x_1)f(x_2)\cdots f(x_n) = \prod_{i=1}^{n} f(x_i). \tag{5.1.1}$$

This joint pdf or pmf can be used to calculate probabilities involving the sample. Since $X_1, \ldots, X_n$ are identically distributed, all the marginal densities $f(x)$ are the

same function. In particular, if the population pdf or pmf is a member of a parametric family, say one of those introduced in Chapter 3, with pdf or pmf given by $f(x|\theta)$, then the joint pdf or pmf is

$$f(x_1,\ldots,x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta), \tag{5.1.2}$$

where the same parameter value θ is used in each of the terms in the product. If, in a statistical setting, we assume that the population we are observing is a member of a specified parametric family but the true parameter value is unknown, then a random sample from this population has a joint pdf or pmf of the above form with the value of θ unknown. By considering different possible values of θ, we can study how a random sample would behave for different populations.

Example 5.1.2 (Sample pdf–exponential) Let $X_1,\ldots,X_n$ be a random sample from an exponential(β) population. Specifically, $X_1,\ldots,X_n$ might correspond to the times until failure (measured in years) for n identical circuit boards that are put on test and used until they fail. The joint pdf of the sample is

$$f(x_1,\ldots,x_n|\beta) = \prod_{i=1}^{n} f(x_i|\beta) = \prod_{i=1}^{n} \frac{1}{\beta} e^{-x_i/\beta} = \frac{1}{\beta^n} e^{-(x_1+\cdots+x_n)/\beta}.$$

This pdf can be used to answer questions about the sample. For example, what is the probability that all the boards last more than 2 years? We can compute

$$\begin{aligned}
&P(X_1 > 2,\ldots,X_n > 2) \\
&\quad = \int_2^\infty \cdots \int_2^\infty \prod_{i=1}^{n} \frac{1}{\beta} e^{-x_i/\beta} dx_1 \cdots dx_n \\
&\quad = e^{-2/\beta} \int_2^\infty \cdots \int_2^\infty \prod_{i=2}^{n} \frac{1}{\beta} e^{-x_i/\beta}\, dx_2 \cdots dx_n \quad \text{(integrate out } x_1\text{)} \\
&\quad \vdots \qquad\qquad \text{(integrate out the remaining } x_i\text{s successively)} \\
&\quad = (e^{-2/\beta})^n \\
&\quad = e^{-2n/\beta}.
\end{aligned}$$

If β, the average lifelength of a circuit board, is large relative to n, we see that this probability is near 1.

The previous calculation illustrates how the pdf of a random sample defined by (5.1.1) or, more specifically, by (5.1.2) can be used to calculate probabilities about the sample. Realize that the independent and identically distributed property of a random sample can also be used directly in such calculations. For example, the above calculation can be done like this:

$$
\begin{aligned}
&P(X_1 > 2, \ldots, X_n > 2) \\
&\quad = P(X_1 > 2) \cdots P(X_n > 2) && \text{(independence)} \\
&\quad = [P(X_1 > 2)]^n && \text{(identical distributions)} \\
&\quad = (e^{-2/\beta})^n && \text{(exponential calculation)} \\
&\quad = e^{-2n/\beta}. && \|
\end{aligned}
$$

The random sampling model in Definition 5.1.1 is sometimes called sampling from an *infinite* population. Think of obtaining the values of $X_1, \ldots, X_n$ sequentially. First, the experiment is performed and $X_1 = x_1$ is observed. Then, the experiment is repeated and $X_2 = x_2$ is observed. The assumption of independence in random sampling implies that the probability distribution for X_2 is unaffected by the fact that $X_1 = x_1$ was observed first. "Removing" x_1 from the infinite population does not change the population, so $X_2 = x_2$ is still a random observation from the same population.

When sampling is from a *finite* population, Definition 5.1.1 may or may not be relevant depending on how the data collection is done. A finite population is a finite set of numbers, $\{x_1, \ldots, x_N\}$. A sample $X_1, \ldots, X_n$ is to be drawn from this population. Four ways of drawing this sample are described in Section 1.2.3. We will discuss the first two.

Suppose a value is chosen from the population in such a way that each of the N values is equally likely (probability $= 1/N$) to be chosen. (Think of drawing numbers from a hat.) This value is recorded as $X_1 = x_1$. Then the process is repeated. Again, each of the N values is equally likely to be chosen. The second value chosen is recorded as $X_2 = x_2$. (If the same number is chosen, then $x_1 = x_2$.) This process of drawing from the N values is repeated n times, yielding the sample $X_1, \ldots, X_n$. This kind of sampling is called *with replacement* because the value chosen at any stage is "replaced" in the population and is available for choice again at the next stage. For this kind of sampling, the conditions of Definition 5.1.1 are met. Each X_i is a discrete random variable that takes on each of the values $x_1, \ldots, x_N$ with equal probability. The random variables $X_1, \ldots, X_n$ are independent because the process of choosing any X_i is the same, regardless of the values that are chosen for any of the other variables. (This type of sampling is used in the *bootstrap*—see Section 10.1.4.)

A second method for drawing a random sample from a finite population is called sampling *without replacement*. Sampling without replacement is done as follows. A value is chosen from $\{x_1, \ldots, x_N\}$ in such a way that each of the N values has probability $1/N$ of being chosen. This value is recorded as $X_1 = x_1$. Now a second value is chosen from the remaining $N - 1$ values. Each of the $N - 1$ values has probability $1/(N - 1)$ of being chosen. The second chosen value is recorded as $X_2 = x_2$. Choice of the remaining values continues in this way, yielding the sample $X_1, \ldots, X_n$. But once a value is chosen, it is unavailable for choice at any later stage.

A sample drawn from a finite population without replacement does not satisfy all the conditions of Definition 5.1.1. The random variables $X_1, \ldots, X_n$ are not mutually independent. To see this, let x and y be distinct elements of $\{x_1, \ldots, x_N\}$. Then $P(X_2 = y | X_1 = y) = 0$, since the value y cannot be chosen at the second stage if it was already chosen at the first. However, $P(X_2 = y | X_1 = x) = 1/(N - 1)$. The

probability distribution for X_2 depends on the value of X_1 that is observed and, hence, X_1 and X_2 are not independent. However, it is interesting to note that $X_1, \ldots, X_n$ are identically distributed. That is, the marginal distribution of X_i is the same for each $i = 1, \ldots, n$. For X_1 it is clear that the marginal distribution is $P(X_1 = x) = 1/N$ for each $x \in \{x_1, \ldots, x_N\}$. To compute the marginal distribution for X_2, use Theorem 1.2.11(a) and the definition of conditional probability to write

$$P(X_2 = x) = \sum_{i=1}^{N} P(X_2 = x | X_1 = x_i) P(X_1 = x_i).$$

For one value of the index, say k, $x = x_k$ and $P(X_2 = x | X_1 = x_k) = 0$. For all other $j \neq k$, $P(X_2 = x | X_1 = x_j) = 1/(N-1)$. Thus,

$$P(X_2 = x) = (N-1)\left(\frac{1}{N-1}\frac{1}{N}\right) = \frac{1}{N}. \tag{5.1.3}$$

Similar arguments can be used to show that each of the X_is has the same marginal distribution.

Sampling without replacement from a finite population is sometimes called *simple random sampling*. It is important to realize that this is not the same sampling situation as that described in Definition 5.1.1. However, if the population size N is large compared to the sample size n, $X_1, \ldots, X_n$ are nearly independent and some approximate probability calculations can be made assuming they are independent. By saying they are "nearly independent" we simply mean that the conditional distribution of X_i given $X_1, \ldots, X_{i-1}$ is not too different from the marginal distribution of X_i. For example, the conditional distribution of X_2 given X_1 is

$$P(X_2 = x_1 | X_1 = x_1) = 0 \quad \text{and} \quad P(X_2 = x | X_1 = x_1) = \frac{1}{N-1} \quad \text{for } x \neq x_1.$$

This is not too different from the marginal distribution of X_2 given in (5.1.3) if N is large. The nonzero probabilities in the conditional distribution of X_i given $X_1, \ldots, X_{i-1}$ are $1/(N-i+1)$, which are close to $1/N$ if $i \leq n$ is small compared with N.

Example 5.1.3 (Finite population model) As an example of an approximate calculation using independence, suppose $\{1, \ldots, 1000\}$ is the finite population, so $N = 1000$. A sample of size $n = 10$ is drawn without replacement. What is the probability that all ten sample values are greater than 200? If $X_1, \ldots, X_{10}$ were mutually independent we would have

$$\begin{aligned} P(X_1 > 200, \ldots, X_{10} > 200) &= P(X_1 > 200) \cdot \cdots \cdot P(X_{10} > 200) \\ &= \left(\frac{800}{1000}\right)^{10} = .107374. \end{aligned} \tag{5.1.4}$$

To calculate this probability exactly, let Y be a random variable that counts the number of items in the sample that are greater than 200. Then Y has a hypergeometric ($N = 1000$, $M = 800$, $K = 10$) distribution. So

$$\begin{aligned} P(X_1 > 200, \ldots, X_{10} > 200) &= P(Y = 10) \\ &= \frac{\binom{800}{10}\binom{200}{0}}{\binom{1000}{10}} \\ &= .106164. \end{aligned}$$

Thus, (5.1.4) is a reasonable approximation to the true value. ||

Throughout the remainder of the book, we will use Definition 5.1.1 as our definition of a random sample from a population.

5.2 Sums of Random Variables from a Random Sample

When a sample $X_1, \ldots, X_n$ is drawn, some summary of the values is usually computed. Any well-defined summary may be expressed mathematically as a function $T(x_1, \ldots, x_n)$ whose domain includes the sample space of the random vector $(X_1, \ldots, X_n)$. The function T may be real-valued or vector-valued; thus the summary is a random variable (or vector), $Y = T(X_1, \ldots, X_n)$. This definition of a random variable as a function of others was treated in detail in Chapter 4, and the techniques in Chapter 4 can be used to describe the distribution of Y in terms of the distribution of the population from which the sample was obtained. Since the random sample $X_1, \ldots, X_n$ has a simple probabilistic structure (because the X_is are independent and identically distributed), the distribution of Y is particularly tractable. Because this distribution is usually derived from the distribution of the variables in the random sample, it is called the *sampling distribution* of Y. This distinguishes the probability distribution of Y from the distribution of the population, that is, the marginal distribution of each X_i. In this section, we will discuss some properties of sampling distributions, especially for functions $T(x_1, \ldots, x_n)$ defined by sums of random variables.

Definition 5.2.1 Let $X_1, \ldots, X_n$ be a random sample of size n from a population and let $T(x_1, \ldots, x_n)$ be a real-valued or vector-valued function whose domain includes the sample space of $(X_1, \ldots, X_n)$. Then the random variable or random vector $Y = T(X_1, \ldots, X_n)$ is called a *statistic*. The probability distribution of a statistic Y is called the *sampling distribution of* Y.

The definition of a statistic is very broad, with the only restriction being that a statistic cannot be a function of a parameter. The sample summary given by a statistic can include many types of information. For example, it may give the smallest or largest value in the sample, the average sample value, or a measure of the variability in the sample observations. Three statistics that are often used and provide good summaries of the sample are now defined.

Definition 5.2.2 The *sample mean* is the arithmetic average of the values in a random sample. It is usually denoted by

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Definition 5.2.3 The *sample variance* is the statistic defined by

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

The *sample standard deviation* is the statistic defined by $S = \sqrt{S^2}$.

As is commonly done, we have suppressed the functional notation in the above definitions of these statistics. That is, we have written S rather than $S(X_1, \ldots, X_n)$. The dependence of the statistic on the sample is understood. As before, we will denote observed values of statistics with lowercase letters. So $\bar{x}$, s^2, and s denote observed values of $\bar{X}$, S^2, and S.

The sample mean is certainly familiar to all. The sample variance and standard deviation are measures of variability in the sample that are related to the population variance and standard deviation in ways that we shall see below. We begin by deriving some properties of the sample mean and variance. In particular, the relationship for the sample variance given in Theorem 5.2.4 is related to (2.3.1), a similar relationship for the population variance.

Theorem 5.2.4 *Let $x_1, \ldots, x_n$ be any numbers and $\bar{x} = (x_1 + \cdots + x_n)/n$. Then*
a. $\min_a \sum_{i=1}^n (x_i - a)^2 = \sum_{i=1}^n (x_i - \bar{x})^2$,
b. $(n-1)s^2 = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2$.

Proof: To prove part (a), add and subtract $\bar{x}$ to get

$$\begin{aligned}\sum_{i=1}^{n}(x_i - a)^2 &= \sum_{i=1}^{n}(x_i - \bar{x} + \bar{x} - a)^2 \\ &= \sum_{i=1}^{n}(x_i - \bar{x})^2 + 2\sum_{i=1}^{n}(x_i - \bar{x})(\bar{x} - a) + \sum_{i=1}^{n}(\bar{x} - a)^2 \\ &= \sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(\bar{x} - a)^2. \qquad \text{(cross term is 0)}\end{aligned}$$

It is now clear that the right-hand side is minimized at $a = \bar{x}$. (Notice the similarity to Example 2.2.6 and Exercise 4.13.)

To prove part (b), take $a = 0$ in the above. □

The expression in Theorem 5.2.4(b) is useful both computationally and theoretically because it allows us to express s^2 in terms of sums that are easy to handle.

We will begin our study of sampling distributions by considering the expected values of some statistics. The following result is quite useful.

Lemma 5.2.5 *Let $X_1, \ldots, X_n$ be a random sample from a population and let $g(x)$ be a function such that $\mathrm{E}g(X_1)$ and $\mathrm{Var}\ g(X_1)$ exist. Then*

$$\mathrm{E}\left(\sum_{i=1}^{n} g(X_i)\right) = n\left(\mathrm{E}g(X_1)\right) \tag{5.2.1}$$

and

$$\mathrm{Var}\left(\sum_{i=1}^{n} g(X_i)\right) = n\left(\mathrm{Var}\ g(X_1)\right). \tag{5.2.2}$$

Proof: To prove (5.2.1), note that

$$\mathrm{E}\left(\sum_{i=1}^{n} g(X_i)\right) = \sum_{i=1}^{n} \mathrm{E}g(X_i) = n\left(\mathrm{E}g(X_1)\right).$$

Since the X_is are identically distributed, the second equality is true because $\mathrm{E}g(X_i)$ is the same for all i. Note that the independence of $X_1, \ldots, X_n$ is not needed for (5.2.1) to hold. Indeed, (5.2.1) is true for any collection of n identically distributed random variables.

To prove (5.2.2), note that

$$\begin{aligned}
\mathrm{Var}\left(\sum_{i=1}^{n} g(X_i)\right) &= \mathrm{E}\left[\sum_{i=1}^{n} g(X_i) - \mathrm{E}\left(\sum_{i=1}^{n} g(X_i)\right)\right]^2 && \text{(definition of variance)} \\
&= \mathrm{E}\left[\sum_{i=1}^{n} \left(g(X_i) - \mathrm{E}g(X_i)\right)\right]^2. && \begin{pmatrix}\text{expectation property and} \\ \text{rearrangement of terms}\end{pmatrix}
\end{aligned}$$

In this last expression there are n^2 terms. First, there are n terms $\left(g(X_i) - \mathrm{E}g(X_i)\right)^2$, $i = 1, \ldots, n$, and for each, we have

$$\begin{aligned}
\mathrm{E}\left(g(X_i) - \mathrm{E}g(X_i)\right)^2 &= \mathrm{Var}\ g(X_i) && \text{(definition of variance)} \\
&= \mathrm{Var}\ g(X_1). && \text{(identically distributed)}
\end{aligned}$$

The remaining $n(n-1)$ terms are all of the form $\left(g(X_i) - \mathrm{E}g(X_i)\right)\left(g(X_j) - \mathrm{E}g(X_j)\right)$, with $i \neq j$. For each term,

$$\begin{aligned}
\mathrm{E}\left[\left(g(X_i) - \mathrm{E}g(X_i)\right)\left(g(X_j) - \mathrm{E}g(X_j)\right)\right] &= \mathrm{Cov}\left(g(X_i), g(X_j)\right) && \begin{pmatrix}\text{definition of} \\ \text{covariance}\end{pmatrix} \\
&= 0. && \begin{pmatrix}\text{independence} \\ \text{Theorem 4.5.5}\end{pmatrix}
\end{aligned}$$

Thus, we obtain equation (5.2.2). □

Theorem 5.2.6 *Let $X_1, \ldots, X_n$ be a random sample from a population with mean μ and variance $\sigma^2 < \infty$. Then*

a. $E\bar{X} = \mu$,

b. $\text{Var } \bar{X} = \dfrac{\sigma^2}{n}$,

c. $ES^2 = \sigma^2$.

Proof: To prove (a), let $g(X_i) = X_i/n$, so $Eg(X_i) = \mu/n$. Then, by Lemma 5.2.5,

$$E\bar{X} = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}nEX_1 = \mu.$$

Similarly for (b), we have

$$\text{Var } \bar{X} = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}n\,\text{Var } X_1 = \frac{\sigma^2}{n}.$$

For the sample variance, using Theorem 5.2.4, we have

$$\begin{aligned} ES^2 &= E\left(\frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right]\right) \\ &= \frac{1}{n-1}\left(nEX_1^2 - nE\bar{X}^2\right) \\ &= \frac{1}{n-1}\left(n(\sigma^2+\mu^2) - n\left(\frac{\sigma^2}{n}+\mu^2\right)\right) = \sigma^2, \end{aligned}$$

establishing part (c) and proving the theorem. □

The relationships (a) and (c) in Theorem 5.2.6, relationships between a statistic and a population parameter, are examples of *unbiased* statistics. These are discussed in Chapter 7. The statistic $\bar{X}$ is an *unbiased estimator* of μ, and S^2 is an *unbiased estimator* of σ^2. The use of $n-1$ in the definition of S^2 may have seemed unintuitive. Now we see that, with this definition, $ES^2 = \sigma^2$. If S^2 were defined as the usual average of the squared deviations with n rather than $n-1$ in the denominator, then ES^2 would be $\frac{n-1}{n}\sigma^2$ and S^2 would not be an unbiased estimator of σ^2.

We now discuss in more detail the sampling distribution of $\bar{X}$. The methods from Sections 4.3 and 4.6 can be used to derive this sampling distribution from the population distribution. But because of the special probabilistic structure of a random sample (iid random variables), the resulting sampling distribution of $\bar{X}$ is simply expressed.

First we note some simple relationships. Since $\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$, if $f(y)$ is the pdf of $Y = (X_1 + \cdots + X_n)$, then $f_{\bar{X}}(x) = nf(nx)$ is the pdf of $\bar{X}$ (see Exercise 5.5). Thus, a result about the pdf of Y is easily transformed into a result about the pdf of $\bar{X}$. A similar relationship holds for mgfs:

$$M_{\bar{X}}(t) = Ee^{t\bar{X}} = Ee^{t(X_1+\cdots+X_n)/n} = Ee^{(t/n)Y} = M_Y(t/n).$$

Since $X_1, \ldots, X_n$ are identically distributed, $M_{X_i}(t)$ is the same function for each i. Thus, by Theorem 4.6.7, we have the following.

Theorem 5.2.7 *Let $X_1, \ldots, X_n$ be a random sample from a population with mgf $M_X(t)$. Then the mgf of the sample mean is*

$$M_{\bar{X}}(t) = [M_X(t/n)]^n.$$

Of course, Theorem 5.2.7 is useful only if the expression for $M_{\bar{X}}(t)$ is a familiar mgf. Cases when this is true are somewhat limited, but the following example illustrates that, when this method works, it provides a very slick derivation of the sampling distribution of $\bar{X}$.

Example 5.2.8 (Distribution of the mean) Let $X_1, \ldots, X_n$ be a random sample from a n(μ, σ^2) population. Then the mgf of the sample mean is

$$\begin{aligned} M_{\bar{X}}(t) &= \left[\exp\left(\mu\frac{t}{n} + \frac{\sigma^2(t/n)^2}{2}\right)\right]^n \\ &= \exp\left(n\left(\mu\frac{t}{n} + \frac{\sigma^2(t/n)^2}{2}\right)\right) = \exp\left(\mu t + \frac{(\sigma^2/n)t^2}{2}\right). \end{aligned}$$

Thus, $\bar{X}$ has a n$(\mu, \sigma^2/n)$ distribution.

Another simple example is given by a gamma(α, β) random sample (see Example 4.6.8). Here, we can also easily derive the distribution of the sample mean. The mgf of the sample mean is

$$M_{\bar{X}}(t) = \left[\left(\frac{1}{1-\beta(t/n)}\right)^{\alpha}\right]^n = \left(\frac{1}{1-(\beta/n)t}\right)^{n\alpha},$$

which we recognize as the mgf of a gamma$(n\alpha, \beta/n)$, the distribution of $\bar{X}$. ‖

If Theorem 5.2.7 is not applicable, because either the resulting mgf of $\bar{X}$ is unrecognizable or the population mgf does not exist, then the transformation method of Sections 4.3 and 4.6 might be used to find the pdf of $Y = (X_1 + \cdots + X_n)$ and $\bar{X}$. In such cases, the following *convolution formula* is useful.

Theorem 5.2.9 *If X and Y are independent continuous random variables with pdfs $f_X(x)$ and $f_Y(y)$, then the pdf of $Z = X + Y$ is*

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(w) f_Y(z-w)\, dw. \tag{5.2.3}$$

Proof: Let $W = X$. The Jacobian of the transformation from (X, Y) to (Z, W) is 1. So using (4.3.2), we obtain the joint pdf of (Z, W) as

$$f_{Z,W}(z, w) = f_{X,Y}(w, z-w) = f_X(w) f_Y(z-w).$$

Integrating out w, we obtain the marginal pdf of Z as given in (5.2.3). □

The limits of integration in (5.2.3) might be modified if f_X or f_Y or both are positive for only some values. For example, if f_X and f_Y are positive for only positive

values, then the limits of integration are 0 and z because the integrand is 0 for values of w outside this range. Equations similar to the convolution formula of (5.2.3) can be derived for operations other than summing; for example, formulas for differences, products, and quotients are also obtainable (see Exercise 5.6).

Example 5.2.10 (Sum of Cauchy random variables) As an example of a situation where the mgf technique fails, consider sampling from a Cauchy distribution. We will eventually derive the distribution of $\bar{Z}$, the mean of $Z_1, \ldots, Z_n$, iid Cauchy$(0,1)$ observations. We start, however, with the distribution of the sum of two independent Cauchy random variables and apply formula (5.2.3).

Let U and V be independent Cauchy random variables, $U \sim$ Cauchy$(0,\sigma)$ and $V \sim$ Cauchy$(0,\tau)$; that is,

$$f_U(u) = \frac{1}{\pi\sigma}\frac{1}{1+(u/\sigma)^2}, \qquad f_V(v) = \frac{1}{\pi\tau}\frac{1}{1+(v/\tau)^2}, \qquad \begin{matrix} -\infty < u < \infty, \\ -\infty < v < \infty. \end{matrix}$$

Based on formula (5.2.3), the pdf of $Z = U + V$ is given by

$$(5.2.4)\quad f_Z(z) = \int_{-\infty}^{\infty} \frac{1}{\pi\sigma}\frac{1}{1+(w/\sigma)^2}\frac{1}{\pi\tau}\frac{1}{1+((z-w)/\tau)^2}\,dw, \quad -\infty < z < \infty.$$

This integral is somewhat involved but can be solved by a partial fraction decomposition and some careful antidifferentiation (see Exercise 5.7). The result is

$$(5.2.5)\qquad f_Z(z) = \frac{1}{\pi(\sigma+\tau)}\frac{1}{1+(z/(\sigma+\tau))^2}, \quad -\infty < z < \infty.$$

Thus, the sum of two independent Cauchy random variables is again a Cauchy, with the scale parameters adding. It therefore follows that if $Z_1, \ldots, Z_n$ are iid Cauchy$(0,1)$ random variables, then $\sum Z_i$ is Cauchy$(0,n)$ and also $\bar{Z}$ is Cauchy$(0,1)$! The sample mean has the same distribution as the individual observations. (See Example A.0.5 in Appendix A for a computer algebra version of this calculation.) ||

If we are sampling from a location–scale family or if we are sampling from certain types of exponential families, the sampling distribution of sums of random variables, and in particular of $\bar{X}$, is easy to derive. We will close this section by discussing these two situations.

We first treat the location–scale case discussed in Section 3.5. Suppose $X_1, \ldots, X_n$ is a random sample from $(1/\sigma)f((x-\mu)/\sigma)$, a member of a location–scale family. Then the distribution of $\bar{X}$ has a simple relationship to the distribution of $\bar{Z}$, the sample mean from a random sample from the standard pdf $f(z)$. To see the nature of this relationship, note that from Theorem 3.5.6 there exist random variables $Z_1, \ldots, Z_n$ such that $X_i = \sigma Z_i + \mu$ and the pdf of each Z_i is $f(z)$. Furthermore, we see that $Z_1, \ldots, Z_n$ are mutually independent. Thus $Z_1, \ldots, Z_n$ is a random sample from $f(z)$. The sample means $\bar{X}$ and $\bar{Z}$ are related by

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\sum_{i=1}^{n}(\sigma Z_i + \mu) = \frac{1}{n}\left(\sigma\sum_{i=1}^{n} Z_i + n\mu\right) = \sigma\bar{Z} + \mu.$$

Thus, again applying Theorem 3.5.6, we find that if $g(z)$ is the pdf of $\bar{Z}$, then $(1/\sigma)g((x-\mu)/\sigma)$ is the pdf of $\bar{X}$. It may be easier to work first with $Z_1, \ldots, Z_n$ and $f(z)$ to find the pdf $g(z)$ of $\bar{Z}$. If this is done, the parameters μ and σ do not have to be dealt with, which may make the computations less messy. Then we immediately know that the pdf of $\bar{X}$ is $(1/\sigma)g((x-\mu)/\sigma)$.

In Example 5.2.10, we found that if $Z_1, \ldots, Z_n$ is a random sample from a Cauchy(0, 1) distribution, then $\bar{Z}$ also has a Cauchy(0, 1) distribution. Now we can conclude that if $X_1, \ldots, X_n$ is a random sample from a Cauchy(μ, σ) distribution, then $\bar{X}$ also has a Cauchy(μ, σ) distribution. It is important to note that the dispersion in the distribution of $\bar{X}$, as measured by σ, is the same, regardless of the sample size n. This is in sharp contrast to the more common situation in Theorem 5.2.6 (the population has finite variance), where $\text{Var}\,\bar{X} = \sigma^2/n$ decreases as the sample size increases.

When sampling is from an exponential family, some sums from a random sample have sampling distributions that are easy to derive. The statistics $T_1, \ldots, T_k$ in the next theorem are important summary statistics, as will be seen in Section 6.2.

Theorem 5.2.11 *Suppose $X_1, \ldots, X_n$ is a random sample from a pdf or pmf $f(x|\theta)$, where*

$$f(x|\theta) = h(x)c(\theta)\exp\left(\sum_{i=1}^{k} w_i(\theta)t_i(x)\right)$$

is a member of an exponential family. Define statistics $T_1, \ldots, T_k$ by

$$T_i(X_1, \ldots, X_n) = \sum_{j=1}^{n} t_i(X_j), \quad i = 1, \ldots, k.$$

If the set $\{(w_1(\theta), w_2(\theta), \ldots, w_k(\theta)), \theta \in \Theta\}$ contains an open subset of $\Re^k$, then the distribution of $(T_1, \ldots, T_k)$ is an exponential family of the form

$$f_T(u_1, \ldots, u_k|\theta) = H(u_1, \ldots, u_k)[c(\theta)]^n \exp\left(\sum_{i=1}^{k} w_i(\theta)u_i\right). \tag{5.2.6}$$

The open set condition eliminates a density such as the n(θ, θ^2) and, in general, eliminates curved exponential families from Theorem 5.2.11. Note that in the pdf or pmf of $(T_1, \ldots, T_k)$, the functions $c(\theta)$ and $w_i(\theta)$ are the same as in the original family although the function $H(u_1, \ldots, u_k)$ is, of course, different from $h(x)$. We will not prove this theorem but will only illustrate the result in a simple case.

Example 5.2.12 (Sum of Bernoulli random variables) Suppose $X_1, \ldots, X_n$ is a random sample from a Bernoulli(p) distribution. From Example 3.4.1 (with $n = 1$) we see that a Bernoulli(p) distribution is an exponential family with $k = 1, c(p) = (1-p), w_1(p) = \log(p/(1-p))$, and $t_1(x) = x$. Thus, in the previous theorem, $T_1 = T_1(X_1, \ldots, X_n) = X_1 + \cdots + X_n$. From the definition of the binomial distribution in Section 3.2, we know that T_1 has a binomial(n, p) distribution. From Example 3.4.1 we also see that a binomial(n, p) distribution is an exponential family with the same $w_1(p)$ and $c(p) = (1-p)^n$. Thus expression (5.2.6) is verified for this example. ||

5.3 Sampling from the Normal Distribution

This section deals with the properties of sample quantities drawn from a normal population—still one of the most widely used statistical models. Sampling from a normal population leads to many useful properties of sample statistics and also to many well-known sampling distributions.

5.3.1 Properties of the Sample Mean and Variance

We have already seen how to calculate the means and variances of $\bar{X}$ and S^2 in general. Now, under the additional assumption of normality, we can derive their full distributions, and more. The properties of $\bar{X}$ and S^2 are summarized in the following theorem.

Theorem 5.3.1 *Let $X_1, \ldots, X_n$ be a random sample from a $\mathrm{n}(\mu, \sigma^2)$ distribution, and let $\bar{X} = (1/n)\sum_{i=1}^n X_i$ and $S^2 = [1/(n-1)]\sum_{i=1}^n (X_i - \bar{X})^2$. Then*

a. *$\bar{X}$ and S^2 are independent random variables.*

b. *$\bar{X}$ has a $\mathrm{n}(\mu, \sigma^2/n)$ distribution,*

c. *$(n-1)S^2/\sigma^2$ has a chi squared distribution with $n-1$ degrees of freedom.*

Proof: First note that, from Section 3.5 on location–scale families, we can assume, without loss of generality, that $\mu = 0$ and $\sigma = 1$. (Also see the discussion preceding Theorem 5.2.11.) Furthermore, part (b) has already been established in Example 5.2.8, leaving us to prove parts (a) and (c).

To prove part (a) we will apply Theorem 4.6.12, and show that $\bar{X}$ and S^2 are functions of independent random vectors. Note that we can write S^2 as a function of $n-1$ deviations as follows:

$$\begin{aligned} S^2 &= \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2 \\ &= \frac{1}{n-1}\left((X_1 - \bar{X})^2 + \sum_{i=2}^n (X_i - \bar{X})^2\right) \\ &= \frac{1}{n-1}\left(\left[\sum_{i=2}^n (X_i - \bar{X})\right]^2 + \sum_{i=2}^n (X_i - \bar{X})^2\right), \quad \left(\text{since } \sum_{i=1}^n (X_i - \bar{X}) = 0\right) \end{aligned}$$

Thus, S^2 can be written as a function only of $(X_2 - \bar{X}, \ldots, X_n - \bar{X})$. We will now show that these random variables are independent of $\bar{X}$. The joint pdf of the sample $X_1, \ldots, X_n$ is given by

$$f(x_1, \ldots, x_n) = \frac{1}{(2\pi)^{n/2}} e^{-(1/2)\sum_{i=1}^n x_i^2}, \quad -\infty < x_i < \infty.$$

Make the transformation

$$\begin{aligned} y_1 &= \bar{x}, \\ y_2 &= x_2 - \bar{x}, \\ &\vdots \\ y_n &= x_n - \bar{x}. \end{aligned}$$

This is a linear transformation with a Jacobian equal to $1/n$. We have

$$\begin{aligned} &f(y_1, \ldots, y_n) \\ &\quad = \frac{n}{(2\pi)^{n/2}} e^{-(1/2)(y_1 - \sum_{i=2}^n y_i)^2} e^{-(1/2)\sum_{i=2}^n (y_i + y_1)^2}, \quad -\infty < y_i < \infty \\ &\quad = \left[\left(\frac{n}{2\pi} \right)^{1/2} e^{(-ny_1^2)/2} \right] \left[\frac{n^{1/2}}{(2\pi)^{(n-1)/2}} e^{-(1/2)\left[\sum_{i=2}^n y_i^2 + (\sum_{i=2}^n y_i)^2\right]} \right], \quad -\infty < y_i < \infty. \end{aligned}$$

Since the joint pdf of $Y_1, \ldots, Y_n$ factors, it follows from Theorem 4.6.11 that Y_1 is independent of $Y_2, \ldots, Y_n$ and, hence, from Theorem 4.6.12 that X is independent of S^2. □

To finish the proof of the theorem we must now derive the distribution of S^2. Before doing so, however, we digress a little and discuss the chi squared distribution, whose properties play an important part in the derivation of the pdf of S^2. Recall from Section 3.3 that the chi squared pdf is a special case of the gamma pdf and is given by

$$f(x) = \frac{1}{\Gamma(p/2)\, 2^{p/2}} x^{(p/2)-1} e^{-x/2}, \quad 0 < x < \infty,$$

where p is called the *degrees of freedom*. We now summarize some pertinent facts about the chi squared distribution.

Lemma 5.3.2 (Facts about chi squared random variables) *We use the notation χ_p^2 to denote a chi squared random variable with p degrees of freedom.*

a. *If Z is a $n(0,1)$ random variable, then $Z^2 \sim \chi_1^2$; that is, the square of a standard normal random variable is a chi squared random variable.*

b. *If $X_1, \ldots, X_n$ are independent and $X_i \sim \chi_{p_i}^2$, then $X_1 + \cdots + X_n \sim \chi_{p_1 + \cdots + p_n}^2$; that is, independent chi squared variables add to a chi squared variable, and the degrees of freedom also add.*

Proof: We have encountered these facts already. Part (a) was established in Example 2.1.7. Part (b) is a special case of Example 4.6.8, which has to do with sums of independent gamma random variables. Since a χ_p^2 random variable is a gamma$(p/2, 2)$, application of the example gives part (b). □

Proof of Theorem 5.3.1(c): We will employ an induction argument to establish the distribution of S^2, using the notation $\bar{X}_k$ and S_k^2 to denote the sample mean and variance based on the first k observations. (Note that the actual ordering of the

observations is immaterial—we are just considering them to be ordered to facilitate the proof.) It is straightforward to establish (see Exercise 5.15) that

$$(n-1)S_n^2 = (n-2)S_{n-1}^2 + \left(\frac{n-1}{n}\right)(X_n - \bar{X}_{n-1})^2. \tag{5.3.1}$$

Now consider $n = 2$. Defining $0 \times S_1^2 = 0$, we have from (5.3.1) that

$$S_2^2 = \frac{1}{2}(X_2 - X_1)^2.$$

Since the distribution of $(X_2 - X_1)/\sqrt{2}$ is n(0, 1), part (a) of Lemma 5.3.2 shows that $S_2^2 \sim \chi_1^2$. Proceeding with the induction, we assume that for $n = k$, $(k-1)S_k^2 \sim \chi_{k-1}^2$. For $n = k+1$ we have from (5.3.1)

$$kS_{k+1}^2 = (k-1)S_k^2 + \left(\frac{k}{k+1}\right)(X_{k+1} - \bar{X}_k)^2. \tag{5.3.2}$$

According to the induction hypothesis, $(k-1)S_k^2 \sim \chi_{k-1}^2$. If we can establish that $(k/(k+1))(X_{k+1} - \bar{X}_k)^2 \sim \chi_1^2$, independent of S_k^2, it will follow from part (b) of Lemma 5.3.2 that $kS_{k+1}^2 \sim \chi_k^2$, and the theorem will be proved.

The independence of $(X_{k+1} - \bar{X}_k)^2$ and S_k^2 again follows from Theorem 4.6.12. The vector $(X_{k+1}, \bar{X}_k)$ is independent of S_k^2 and so is any function of the vector. Furthermore, $X_{k+1} - \bar{X}_k$ is a normal random variable with mean 0 and variance

$$\text{Var}\,(X_{k+1} - \bar{X}_k) = \frac{k+1}{k},$$

and therefore $(k/(k+1))(X_{k+1} - \bar{X}_k)^2 \sim \chi_1^2$, and the theorem is established. □

The independence of $\bar{X}$ and S^2 can be established in a manner different from that used in the proof of Theorem 5.3.1. Rather than show that the joint pdf factors, we can use the following lemma, which ties together independence and correlation for normal samples.

Lemma 5.3.3 *Let $X_j \sim \text{n}(\mu_j, \sigma_j^2)$, $j = 1, \ldots, n$, independent. For constants a_{ij} and b_{rj} $(j = 1, \ldots, n; i = 1, \ldots, k; r = 1, \ldots, m)$, where $k + m \le n$, define*

$$U_i = \sum_{j=1}^{n} a_{ij} X_j, \quad i = 1, \ldots, k.$$

$$V_r = \sum_{j=1}^{n} b_{rj} X_j, \quad r = 1, \ldots, m.$$

a. *The random variables U_i and V_r are independent if and only if $\text{Cov}(U_i, V_r) = 0$. Furthermore, $\text{Cov}(U_i, V_r) = \sum_{j=1}^{n} a_{ij} b_{rj} \sigma_j^2$.*

b. *The random vectors $(U_1, \ldots, U_k)$ and $(V_1, \ldots, V_m)$ are independent if and only if U_i is independent of V_r for all pairs i, r $(i = 1, \ldots, k; r = 1, \ldots, m)$.*

Proof: It is sufficient to prove the lemma for $\mu_i = 0$ and $\sigma_i^2 = 1$, since the general statement of the lemma then follows quickly. Furthermore, the implication from in-

dependence to 0 covariance is immediate (Theorem 4.5.5) and the expression for the covariance is easily verified (Exercise 5.14). Note also that Corollary 4.6.10 shows that U_i and V_r are normally distributed.

Thus, we are left with proving that if the constants satisfy the above restriction (equivalently, the covariance is 0), then we have independence under normality. We prove the lemma only for $n = 2$, since the proof for general n is similar but necessitates a detailed n-variate transformation.

To prove part (a) start with the joint pdf of X_1 and X_2,

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

Make the transformation (we can suppress the double subscript in the $n = 2$ case)

$$u = a_1x_1 + a_2x_2, \qquad v = b_1x_1 + b_2x_2,$$

so

$$x_1 = \frac{b_2u - a_2v}{a_1b_2 - b_1a_2}, \qquad x_2 = \frac{a_1v - b_1u}{a_1b_2 - b_1a_2},$$

with Jacobian

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial u} & \frac{\partial x_1}{\partial v} \\ \frac{\partial x_2}{\partial u} & \frac{\partial x_2}{\partial v} \end{vmatrix} = \frac{1}{a_1b_2 - b_1a_2}.$$

Thus, the pdf of U and V is

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

$-\infty < u, v < \infty$. Expanding the squares in the exponent, we can write

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

The assumption on the constants shows that the cross-term is identically 0. Hence, the pdf factors so, by Lemma 4.2.7, U and V are independent and part (a) is established.

A similar type of argument will work for part (b), the details of which we will not go into. If the appropriate transformation is made, the joint pdf of the vectors $(U_1, \ldots, U_k)$ and $(V_1, \ldots, V_m)$ can be obtained. By an application of Theorem 4.6.11, the vectors are independent if the joint pdf factors. From the form of the normal pdf, this will happen if and only if U_i is independent of V_r for all pairs i, r $(i = 1, \ldots, k; r = 1, \ldots, m)$. □

This lemma shows that, if we start with independent normal random variables, covariance and independence are equivalent for linear functions of these random vari-

ables. Thus, we can check independence for normal variables by merely checking the covariance term, a much simpler calculation. There is nothing magic about this; it just follows from the form of the normal pdf. Furthermore, part (b) allows us to infer overall independence of normal vectors by just checking pairwise independence, a property that does not hold for general random variables.

We can use Lemma 5.3.3 to provide an alternative proof of the independence of $\bar{X}$ and S^2 in normal sampling. Since we can write S^2 as a function of $n-1$ deviations $(X_2 - \bar{X}, \ldots, X_n - \bar{X})$, we must show that these random variables are uncorrelated with $\bar{X}$. The normality assumption, together with Lemma 5.3.3, will then allow us to conclude independence.

As an illustration of the application of Lemma 5.3.3, write

$$\bar{X} = \sum_{i=1}^{n} \left(\frac{1}{n}\right) X_i,$$

$$X_j - \bar{X} = \sum_{i=1}^{n} \left(\delta_{ij} - \frac{1}{n}\right) X_i,$$

where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise. It is then easy to show that

$$\mathrm{Cov}(\bar{X}, X_j - \bar{X}) = \sum_{i=1}^{n} \left(\frac{1}{n}\right)\left(\delta_{ij} - \frac{1}{n}\right) = 0,$$

showing that $\bar{X}$ and $X_j - \bar{X}$ are independent (as long as the X_is have the same variance).

5.3.2 The Derived Distributions: Student's t and Snedecor's F

The distributions derived in Section 5.3.1 are, in a sense, the first step in a statistical analysis that assumes normality. In particular, in most practical cases the variance, σ^2, is unknown. Thus, to get any idea of the variability of $\bar{X}$ (as an estimate of μ), it is necessary to estimate this variance. This topic was first addressed by W. S. Gosset (who published under the pseudonym of Student) in the early 1900s. The landmark work of Student resulted in Student's t distribution or, more simply, the t distribution.

If $X_1, \ldots, X_n$ are a random sample from a $\mathrm{n}(\mu, \sigma^2)$, we know that the quantity

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \tag{5.3.3}$$

is distributed as a $\mathrm{n}(0,1)$ random variable. If we knew the value of σ and we measured $\bar{X}$, then we could use (5.3.3) as a basis for inference about μ, since μ would then be the only unknown quantity. Most of the time, however, σ is unknown. Student did the obvious thing—he looked at the distribution of

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}, \tag{5.3.4}$$

a quantity that could be used as a basis for inference about μ when σ was unknown.

The distribution of (5.3.4) is easy to derive, provided that we first notice a few simplifying maneuvers. Multiply (5.3.4) by σ/σ and rearrange slightly to obtain

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{(\bar{X}-\mu)/(\sigma/\sqrt{n})}{\sqrt{S^2/\sigma^2}}. \tag{5.3.5}$$

The numerator of (5.3.5) is a $\mathrm{n}(0,1)$ random variable, and the denominator is $\sqrt{\chi^2_{n-1}/(n-1)}$, *independent* of the numerator. Thus, the distribution of (5.3.4) can be found by solving the simplified problem of finding the distribution of $U/\sqrt{V/p}$, where U is $\mathrm{n}(0,1)$, V is χ^2_p, and U and V are independent. This gives us Student's t distribution.

Definition 5.3.4 Let $X_1, \ldots, X_n$ be a random sample from a $\mathrm{n}(\mu, \sigma^2)$ distribution. The quantity $(\bar{X}-\mu)/(S/\sqrt{n})$ has *Student's t distribution with* $n-1$ *degrees of freedom*. Equivalently, a random variable T has Student's t distribution with p degrees of freedom, and we write $T \sim t_p$ if it has pdf

$$f_T(t) = \frac{\Gamma\left(\frac{p+1}{2}\right)}{\Gamma\left(\frac{p}{2}\right)} \frac{1}{(p\pi)^{1/2}} \frac{1}{(1+t^2/p)^{(p+1)/2}}, \quad -\infty < t < \infty. \tag{5.3.6}$$

Notice that if $p = 1$, then (5.3.6) becomes the pdf of the Cauchy distribution, which occurs for samples of size 2. Once again the Cauchy distribution has appeared in an ordinary situation.

The derivation of the t pdf is straightforward. If we start with U and V defined above, it follows from (5.3.5) that the joint pdf of U and V is

$$f_{U,V}(u,v) = \frac{1}{(2\pi)^{1/2}} e^{-u^2/2} \frac{1}{\Gamma\left(\frac{p}{2}\right) 2^{p/2}} v^{(p/2)-1} e^{-v/2}, \quad -\infty < u < \infty, \quad 0 < v < \infty.$$

(Recall that U and V are independent.) Now make the transformation

$$t = \frac{u}{\sqrt{v/p}}, \qquad w = v.$$

The Jacobian of the transformation is $(w/p)^{1/2}$, and the marginal pdf of T is given by

$$\begin{aligned} f_T(t) &= \int_0^\infty f_{U,V}\left(t\left(\frac{w}{p}\right)^{1/2}, w\right)\left(\frac{w}{p}\right)^{1/2} dw \\ &= \frac{1}{(2\pi)^{1/2}} \frac{1}{\Gamma\left(\frac{p}{2}\right) 2^{p/2}} \int_0^\infty e^{-(1/2)t^2 w/p} w^{(p/2)-1} e^{-w/2} \left(\frac{w}{p}\right)^{1/2} dw \\ &= \frac{1}{(2\pi)^{1/2}} \frac{1}{\Gamma\left(\frac{p}{2}\right) 2^{p/2} p^{1/2}} \int_0^\infty e^{-(1/2)(1+t^2/p)w} w^{((p+1)/2)-1} dw. \end{aligned}$$

Recognize the integrand as the kernel of a gamma$((p+1)/2, 2/(1+t^2/p))$ pdf. We therefore have

$$f_T(t) = \frac{1}{(2\pi)^{1/2}} \frac{1}{\Gamma\left(\frac{p}{2}\right) 2^{p/2} p^{1/2}} \Gamma\left(\frac{p+1}{2}\right) \left[\frac{2}{1+t^2/p}\right]^{(p+1)/2},$$

which is equal to (5.3.6).

Student's t has no mgf because it does not have moments of all orders. In fact, if there are p degrees of freedom, then there are only $p-1$ moments. Hence, a t_1 has no mean, a t_2 has no variance, etc. It is easy to check (see Exercise 5.18) that if T_p is a random variable with a t_p distribution, then

$$\begin{aligned} &\mathrm{E}T_p = 0, \quad \text{if } p > 1, \\ &\mathrm{Var}\, T_p = \frac{p}{p-2}, \quad \text{if } p > 2. \end{aligned} \tag{5.3.7}$$

Another important derived distribution is Snedecor's F, whose derivation is quite similar to that of Student's t. Its motivation, however, is somewhat different. The F distribution, named in honor of Sir Ronald Fisher, arises naturally as the distribution of a ratio of variances.

Example 5.3.5 (Variance ratio distribution) Let $X_1, \ldots, X_n$ be a random sample from a n(μ_X, σ_X^2) population, and let $Y_1, \ldots, Y_m$ be a random sample from an independent n(μ_Y, σ_Y^2) population. If we were interested in comparing the variability of the populations, one quantity of interest would be the ratio σ_X^2/σ_Y^2. Information about this ratio is contained in S_X^2/S_Y^2, the ratio of sample variances. The F distribution allows us to compare these quantities by giving us a distribution of

$$\frac{S_X^2/S_Y^2}{\sigma_X^2/\sigma_Y^2} = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}. \tag{5.3.8}$$

Examination of (5.3.8) shows us how the F distribution is derived. The ratios S_X^2/σ_X^2 and S_Y^2/σ_Y^2 are each scaled chi squared variates, and they are independent. ‖

Definition 5.3.6 Let $X_1, \ldots, X_n$ be a random sample from a n(μ_X, σ_X^2) population, and let $Y_1, \ldots, Y_m$ be a random sample from an independent n(μ_Y, σ_Y^2) population. The random variable $F = (S_X^2/\sigma_X^2)/(S_Y^2/\sigma_Y^2)$ has *Snedecor's F distribution with $n-1$ and $m-1$ degrees of freedom.* Equivalently, the random variable F has the F distribution with p and q degrees of freedom if it has pdf

$$f_F(x) = \frac{\Gamma\left(\frac{p+q}{2}\right)}{\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{q}{2}\right)} \left(\frac{p}{q}\right)^{p/2} \frac{x^{(p/2)-1}}{[1+(p/q)x]^{(p+q)/2}}, \quad 0 < x < \infty. \tag{5.3.9}$$

The F distribution can be derived in a more general setting than is done here. A variance ratio may have an F distribution even if the parent populations are not normal. Kelker (1970) has shown that as long as the parent populations have a certain type of symmetry (*spherical symmetry*), then the variance ratio will have an F distribution.

The derivation of the F pdf, starting from normal distributions, is similar to the derivation of Student's t. In fact, in one special case the F is a transform of the t. (See Theorem 5.3.8.) Similar to what we did for the t, we can reduce the task of deriving the F pdf to that of finding the pdf of $(U/p)/(V/q)$, where U and V are independent, $U \sim \chi^2_p$ and $V \sim \chi^2_q$. (See Exercise 5.17.)

Example 5.3.7 (Continuation of Example 5.3.5) To see how the F distribution may be used for inference about the true ratio of population variances, consider the following. The quantity $(S_X^2/\sigma_X^2)/(S_Y^2/\sigma_Y^2)$ has an $F_{n-1,m-1}$ distribution. (In general, we use the notation $F_{p,q}$ to denote an F random variable with p and q degrees of freedom.) We can calculate

$$\begin{aligned}
\mathrm{E}F_{n-1,m-1} &= \mathrm{E}\left(\frac{\chi^2_{n-1}/(n-1)}{\chi^2_{m-1}/(m-1)}\right) && \text{(by definition)} \\
&= \mathrm{E}\left(\frac{\chi^2_{n-1}}{n-1}\right)\mathrm{E}\left(\frac{m-1}{\chi^2_{m-1}}\right) && \text{(independence)} \\
&= \left(\frac{n-1}{n-1}\right)\left(\frac{m-1}{m-3}\right) && \text{(chi squared calculations)} \\
&= \frac{m-1}{m-3}.
\end{aligned}$$

Note that this last expression is finite and positive only if $m > 3$. We have that

$$\mathrm{E}\left(\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}\right) = \mathrm{E}F_{n-1,m-1} = \frac{m-1}{m-3},$$

and, removing expectations, we have for reasonably large m,

$$\frac{S_X^2/S_Y^2}{\sigma_X^2/\sigma_Y^2} \approx \frac{m-1}{m-3} \approx 1,$$

as we might expect. ||

The F distribution has many interesting properties and is related to a number of other distributions. We summarize some of these facts in the next theorem, whose proof is left as an exercise. (See Exercises 5.17 and 5.18.)

Theorem 5.3.8

a. *If $X \sim F_{p,q}$, then $1/X \sim F_{q,p}$; that is, the reciprocal of an F random variable is again an F random variable.*

b. *If $X \sim t_q$, then $X^2 \sim F_{1,q}$.*

c. *If $X \sim F_{p,q}$, then $(p/q)X/(1+(p/q)X) \sim beta(p/2, q/2)$.*

5.4 Order Statistics

Sample values such as the smallest, largest, or middle observation from a random sample can provide additional summary information. For example, the highest flood waters or the lowest winter temperature recorded during the last 50 years might be useful data when planning for future emergencies. The median price of houses sold during the previous month might be useful for estimating the cost of living. These are all examples of *order statistics*.

Definition 5.4.1 The *order statistics* of a random sample $X_1, \ldots, X_n$ are the sample values placed in ascending order. They are denoted by $X_{(1)}, \ldots, X_{(n)}$.

The order statistics are random variables that satisfy $X_{(1)} \leq \cdots \leq X_{(n)}$. In particular,

$$X_{(1)} = \min_{1 \leq i \leq n} X_i,$$

$$X_{(2)} = \text{second smallest } X_i,$$

$$\vdots$$

$$X_{(n)} = \max_{1 \leq i \leq n} X_i.$$

Since they are random variables, we can discuss the probabilities that they take on various values. To calculate these probabilities we need the pdfs or pmfs of the order statistics. The formulas for the pdfs of the order statistics of a random sample from a continuous population will be the main topic later in this section, but first, we will mention some statistics that are easily defined in terms of the order statistics.

The *sample range*, $R = X_{(n)} - X_{(1)}$, is the distance between the smallest and largest observations. It is a measure of the dispersion in the sample and should reflect the dispersion in the population.

The *sample median*, which we will denote by M, is a number such that approximately one-half of the observations are less than M and one-half are greater. In terms of the order statistics, M is defined by

$$M = \begin{cases} X_{((n+1)/2)} & \text{if } n \text{ is odd} \\ \left(X_{(n/2)} + X_{(n/2+1)}\right)/2 & \text{if } n \text{ is even.} \end{cases} \tag{5.4.1}$$

The median is a measure of location that might be considered an alternative to the sample mean. One advantage of the sample median over the sample mean is that it is less affected by extreme observations. (See Section 10.2 for details.)

Although related, the mean and median usually measure different things. For example, in recent baseball salary negotiations a major point of contention was the owners' contributions to the players' pension fund. The owners' view could be paraphrased as, "The average baseball player's annual salary is \$433,659 so, with that kind of money, the current pension is adequate." But the players' view was, "Over half of the players make less than \$250,000 annually and, because of the short professional life of most

players, need the security of a larger pension." (These figures are for the 1988 season, not the year of the dispute.) Both figures were correct, but the owners were discussing the mean while the players were discussing the median. About a dozen players with salaries over $2 million can raise the average salary to $433,659 while the majority of the players make less than $250,000, including all rookies who make $62,500. When discussing salaries, prices, or any variable with a few extreme values, the median gives a better indication of "typical" values than the mean. Other statistics that can be defined in terms of order statistics and are less sensitive to extreme values (such as the *α-trimmed* mean discussed in Exercise 10.20) are discussed in texts such as Tukey (1977).

For any number p between 0 and 1, the $(100p)$th sample percentile is the observation such that approximately np of the observations are less than this observation and $n(1-p)$ of the observations are greater. The 50th sample percentile ($p = .5$) is the sample median. For other values of p, we can more precisely define the sample percentiles in terms of the order statistics in the following way.

Definition 5.4.2 The notation $\{b\}$, when appearing in a subscript, is defined to be *the number b rounded to the nearest integer in the usual way.* More precisely, if i is an integer and $i - .5 \leq b < i + .5$, then $\{b\} = i$.

The $(100p)$th sample percentile is $X_{(\{np\})}$ if $\frac{1}{2n} < p < .5$ and $X_{(n+1-\{n(1-p)\})}$ if $.5 < p < 1 - \frac{1}{2n}$. For example, if $n = 12$ and the 65th percentile is wanted, we note that $12 \times (1 - .65) = 4.2$ and $12 + 1 - 4 = 9$. Thus the 65th percentile is $X_{(9)}$. There is a restriction on the range of p because the size of the sample limits the range of sample percentiles.

The cases $p < .5$ and $p > .5$ are defined separately so that the sample percentiles exhibit the following symmetry. If the $(100p)$th sample percentile is the ith smallest observation, then the $(100(1-p))$th sample percentile should be the ith largest observation and the above definition achieves this. For example, if $n = 11$, the 30th sample percentile is $X_{(3)}$ and the 70th sample percentile is $X_{(9)}$.

In addition to the median, two other sample percentiles are commonly identified. These are the *lower quartile* (25th percentile) and *upper quartile* (75th percentile). A measure of dispersion that is sometimes used is the *interquartile range*, the distance between the lower and upper quartiles.

Since the order statistics are functions of the sample, probabilities concerning order statistics can be computed in terms of probabilities for the sample. If $X_1, \ldots, X_n$ are iid discrete random variables, then the calculation of probabilities for the order statistics is mainly a counting task. These formulas are derived in Theorem 5.4.3. If $X_1, \ldots, X_n$ are a random sample from a continuous population, then convenient expressions for the pdf of one or more order statistics are derived in Theorems 5.4.4 and 5.4.6. These can then be used to derive the distribution of functions of the order statistics.

Theorem 5.4.3 *Let $X_1, \ldots, X_n$ be a random sample from a discrete distribution with pmf $f_X(x_i) = p_i$, where $x_1 < x_2 < \cdots$ are the possible values of X in ascending*

order. Define

$$P_0 = 0$$
$$P_1 = p_1$$
$$P_2 = p_1 + p_2$$
$$\vdots$$
$$P_i = p_1 + p_2 + \cdots + p_i$$
$$\vdots$$

Let $X_{(1)}, \ldots, X_{(n)}$ denote the order statistics from the sample. Then

$$P(X_{(j)} \le x_i) = \sum_{k=j}^{n} \binom{n}{k} P_i^k (1 - P_i)^{n-k} \tag{5.4.2}$$

and

$$P(X_{(j)} = x_i) = \sum_{k=j}^{n} \binom{n}{k} \left[P_i^k (1 - P_i)^{n-k} - P_{i-1}^k (1 - P_{i-1})^{n-k}\right]. \tag{5.4.3}$$

Proof: Fix i, and let Y be a random variable that counts the number of $X_1, \ldots, X_n$ that are less than or equal to x_i. For each of $X_1, \ldots, X_n$, call the event $\{X_j \le x_i\}$ a "success" and $\{X_j > x_i\}$ a "failure." Then Y is the number of successes in n trials. The probability of a success is the same value, namely $P_i = P(X_j \le x_i)$, for each trial, since $X_1, \ldots, X_n$ are identically distributed. The success or failure of the jth trial is independent of the outcome of any other trial, since X_j is independent of the other X_is. Thus, $Y \sim$ binomial(n, P_i).

The event $\{X_{(j)} \le x_i\}$ is equivalent to the event $\{Y \ge j\}$; that is, at least j of the sample values are less than or equal to x_i. Equation (5.4.2) expresses this binomial probability,

$$P(X_{(j)} \le x_i) = P(Y \ge j).$$

Equation (5.4.3) simply expresses the difference,

$$P(X_{(j)} = x_i) = P(X_{(j)} \le x_i) - P(X_{(j)} \le x_{i-1}).$$

The case $i = 1$ is exceptional in that $P(X_{(j)} = x_1) = P(X_{(j)} \le x_1)$. The definition of $P_0 = 0$ takes care of this exception in (5.4.3). □

If $X_1, \ldots, X_n$ are a random sample from a continuous population, then the situation is simplified slightly by the fact that the probability is 0 that any two X_js are equal, freeing us from worrying about ties. Thus $P(X_{(1)} < X_{(2)} < \cdots < X_{(n)}) = 1$ and the sample space for $(X_{(1)}, \ldots, X_{(n)})$ is $\{(x_1, \ldots, x_n) : x_1 < x_2 < \cdots < x_n\}$. In Theorems 5.4.4 and 5.4.6 we derive the pdf for one and the joint pdf for two order statistics, again using binomial arguments.

Theorem 5.4.4 *Let $X_{(1)}, \ldots, X_{(n)}$ denote the order statistics of a random sample, $X_1, \ldots, X_n$, from a continuous population with cdf $F_X(x)$ and pdf $f_X(x)$. Then the pdf of $X_{(j)}$ is*

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!} f_X(x)[F_X(x)]^{j-1}[1-F_X(x)]^{n-j}. \tag{5.4.4}$$

Proof: We first find the cdf of $X_{(j)}$ and then differentiate it to obtain the pdf. As in Theorem 5.4.3, let Y be a random variable that counts the number of $X_1, \ldots, X_n$ less than or equal to x. Then, defining a "success" as the event $\{X_j \le x\}$, we see that $Y \sim$ binomial$(n, F_X(x))$. (Note that we can write $P_i = F_X(x_i)$ in Theorem 5.4.3. Also, although $X_1, \ldots, X_n$ are continuous random variables, the counting variable Y is discrete.) Thus,

$$F_{X_{(j)}}(x) = P(Y \ge j) = \sum_{k=j}^{n} \binom{n}{k} [F_X(x)]^k [1-F_X(x)]^{n-k},$$

and the pdf of $X_{(j)}$ is

$$\begin{aligned}
f_{X_{(j)}}(x) &= \frac{d}{dx} F_{X_{(j)}}(x) \\
&= \sum_{k=j}^{n} \binom{n}{k} \Big(k\,[F_X(x)]^{k-1}\,[1-F_X(x)]^{n-k}\, f_X(x) \\
&\quad - (n-k)[F_X(x)]^k [1-F_X(x)]^{n-k-1} f_X(x) \Big) && \text{(chain rule)} \\
&= \binom{n}{j} j f_X(x)[F_X(x)]^{j-1}[1-F_X(x)]^{n-j} \\
&\quad + \sum_{k=j+1}^{n} \binom{n}{k} k[F_X(x)]^{k-1}[1-F_X(x)]^{n-k} f_X(x) \\
&\quad - \sum_{k=j}^{n-1} \binom{n}{k} (n-k)[F_X(x)]^k [1-F_X(x)]^{n-k-1} f_X(x) && \left(\begin{smallmatrix} k=n \text{ term} \\ \text{is } 0 \end{smallmatrix}\right) \\
&= \frac{n!}{(j-1)!(n-j)!} f_X(x)[F_X(x)]^{j-1}[1-F_X(x)]^{n-j} \\
&\quad + \sum_{k=j}^{n-1} \binom{n}{k+1} (k+1)\,[F_X(x)]^k\,[1-F_X(x)]^{n-k-1} f_X(x) && \left(\begin{smallmatrix} \text{change} \\ \text{dummy} \\ \text{variable} \end{smallmatrix}\right) \qquad (5.4.5) \\
&\quad - \sum_{k=j}^{n-1} \binom{n}{k} (n-k)\,[F_X(x)]^k\,[1-F_X(x)]^{n-k-1}\, f_X(x).
\end{aligned}$$

Noting that

$$\binom{n}{k+1}(k+1) = \frac{n!}{k!(n-k-1)!} = \binom{n}{k}(n-k), \tag{5.4.6}$$

we see that the last two sums in (5.4.5) cancel. Thus, the pdf $f_{X_{(j)}}(x)$ is given by the expression in (5.4.4). □

Example 5.4.5 (Uniform order statistic pdf) Let $X_1, \ldots, X_n$ be iid uniform$(0,1)$, so $f_X(x) = 1$ for $x \in (0,1)$ and $F_X(x) = x$ for $x \in (0,1)$. Using (5.4.4), we see that the pdf of the jth order statistic is

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!} x^{j-1}(1-x)^{n-j} \quad \text{for } x \in (0,1)$$
$$= \frac{\Gamma(n+1)}{\Gamma(j)\Gamma(n-j+1)} x^{j-1}(1-x)^{(n-j+1)-1}.$$

Thus, the jth order statistic from a uniform$(0,1)$ sample has a beta$(j, n-j+1)$ distribution. From this we can deduce that

$$\mathrm{E}X_{(j)} = \frac{j}{n+1} \quad \text{and} \quad \mathrm{Var}\, X_{(j)} = \frac{j(n-j+1)}{(n+1)^2(n+2)}. \qquad \|$$

The joint distribution of two or more order statistics can be used to derive the distribution of some of the statistics mentioned at the beginning of this section. The joint pdf of any two order statistics is given in the following theorem, whose proof is left to Exercise 5.26.

Theorem 5.4.6 *Let $X_{(1)}, \ldots, X_{(n)}$ denote the order statistics of a random sample, $X_1, \ldots, X_n$, from a continuous population with cdf $F_X(x)$ and pdf $f_X(x)$. Then the joint pdf of $X_{(i)}$ and $X_{(j)}$, $1 \le i < j \le n$, is*

$$(5.4.7) \quad f_{X_{(i)},X_{(j)}}(u,v) = \frac{n!}{(i-1)!(j-1-i)!(n-j)!} f_X(u) f_X(v) [F_X(u)]^{i-1}$$
$$\times [F_X(v) - F_X(u)]^{j-1-i} [1 - F_X(v)]^{n-j}$$

for $-\infty < u < v < \infty$.

The joint pdf of three or more order statistics could be derived using similar but even more involved arguments. Perhaps the other most useful pdf is $f_{X_{(1)},\ldots,X_{(n)}}(x_1, \ldots, x_n)$, the joint pdf of all the order statistics, which is given by

$$f_{X_{(1)},\ldots,X_{(n)}}(x_1, \ldots, x_n) = \begin{cases} n! f_X(x_1) \cdot \cdots \cdot f_X(x_n) & -\infty < x_1 < \cdots < x_n < \infty \\ 0 & \text{otherwise.} \end{cases}$$

The $n!$ naturally comes into this formula because, for any set of values $x_1, \ldots, x_n$, there are $n!$ equally likely assignments for these values to $X_1, \ldots, X_n$ that all yield the same values for the order statistics. This joint pdf and the techniques from Chapter 4 can be used to derive marginal and conditional distributions and distributions of other functions of the order statistics. (See Exercises 5.27 and 5.28.)

We now use the joint pdf (5.4.7) to derive the distribution of some of the functions mentioned at the beginning of this section.

Example 5.4.7 (Distribution of the midrange and range) Let $X_1, \ldots, X_n$ be iid uniform$(0, a)$ and let $X_{(1)}, \ldots, X_{(n)}$ denote the order statistics. The range was earlier defined as $R = X_{(n)} - X_{(1)}$. The *midrange*, a measure of location like the sample median or the sample mean, is defined by $V = (X_{(1)} + X_{(n)})/2$. We will derive the joint pdf of R and V from the joint pdf of $X_{(1)}$ and $X_{(n)}$. From (5.4.7) we have that

$$f_{X_{(1)},X_{(n)}}(x_1, x_n) = \frac{n(n-1)}{a^2}\left(\frac{x_n}{a} - \frac{x_1}{a}\right)^{n-2}$$
$$= \frac{n(n-1)(x_n - x_1)^{n-2}}{a^n}, \quad 0 < x_1 < x_n < a.$$

Solving for $X_{(1)}$ and $X_{(n)}$, we obtain $X_{(1)} = V - R/2$ and $X_{(n)} = V + R/2$. The Jacobian for this transformation is -1. The transformation from $(X_{(1)}, X_{(n)})$ to (R, V) maps $\{(x_1, x_n): 0 < x_1 < x_n < a\}$ onto the set $\{(r, v): 0 < r < a, r/2 < v < a - r/2\}$. To see this, note that obviously $0 < r < a$ and for a fixed value of r, v ranges from $r/2$ (corresponding to $x_1 = 0, x_n = r$) to $a - r/2$ (corresponding to $x_1 = a - r, x_n = a$). Thus, the joint pdf of (R, V) is

$$f_{R,V}(r, v) = \frac{n(n-1)r^{n-2}}{a^n}, \quad 0 < r < a, \quad r/2 < v < a - r/2.$$

The marginal pdf of R is thus

$$f_R(r) = \int_{r/2}^{a-r/2} \frac{n(n-1)r^{n-2}}{a^n}\, dv$$
$$= \frac{n(n-1)r^{n-2}(a-r)}{a^n}, \quad 0 < r < a. \tag{5.4.8}$$

If $a = 1$, we see that r has a beta$(n-1, 2)$ distribution. Or, for arbitrary a, it is easy to deduce from (5.4.8) that R/a has a beta distribution. Note that the constant a is a scale parameter.

The set where $f_{R,V}(r, v) > 0$ is shown in Figure 5.4.1, where we see that the range of integration of r depends on whether $v > a/2$ or $v \le a/2$. Thus, the marginal pdf of V is given by

$$f_V(v) = \int_0^{2v} \frac{n(n-1)r^{n-2}}{a^n} dr = \frac{n(2v)^{n-1}}{a^n}, \quad 0 < v \le a/2,$$

and

$$f_V(v) = \int_0^{2(a-v)} \frac{n(n-1)r^{n-2}}{a^n} dr = \frac{n[2(a-v)]^{n-1}}{a^n}, \quad a/2 < v \le a.$$

This pdf is symmetric about $a/2$ and has a peak at $a/2$. ‖

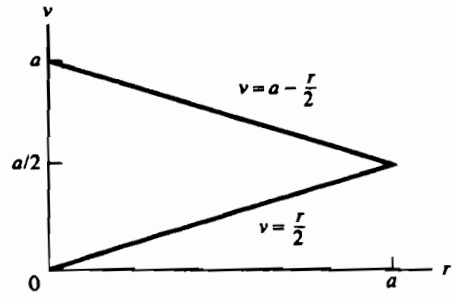

Figure 5.4.1. *Region on which* $f_{R,V}(r, v) > 0$ *for Example 5.4.7*

5.5 Convergence Concepts

This section treats the somewhat fanciful idea of allowing the sample size to approach infinity and investigates the behavior of certain sample quantities as this happens. Although the notion of an infinite sample size is a theoretical artifact, it can often provide us with some useful approximations for the finite-sample case, since it usually happens that expressions become simplified in the limit.

We are mainly concerned with three types of convergence, and we treat them in varying amounts of detail. (A full treatment of convergence is given in Billingsley 1995 or Resnick 1999, for example.) In particular, we want to look at the behavior of $\bar{X}_n$, the mean of n observations, as $n \to \infty$.

5.5.1 Convergence in Probability

This type of convergence is one of the weaker types and, hence, is usually quite easy to verify.

Definition 5.5.1 A sequence of random variables, $X_1, X_2, \ldots$, *converges in probability* to a random variable X if, for every $\epsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0 \quad \text{or, equivalently,} \quad \lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1.$$

The $X_1, X_2, \ldots$ in Definition 5.5.1 (and the other definitions in this section) are typically not independent and identically distributed random variables, as in a random sample. The distribution of X_n changes as the subscript changes, and the convergence concepts discussed in this section describe different ways in which the distribution of X_n converges to some limiting distribution as the subscript becomes large.

Frequently, statisticians are concerned with situations in which the limiting random variable is a constant and the random variables in the sequence are sample means (of some sort). The most famous result of this type is the following.

Theorem 5.5.2 (Weak Law of Large Numbers) *Let* $X_1, X_2, \ldots$ *be iid random variables with* $\mathrm{E}X_i = \mu$ *and* $\mathrm{Var}\, X_i = \sigma^2 < \infty$. *Define* $\bar{X}_n = (1/n)\sum_{i=1}^{n} X_i$. *Then,*

for every $\epsilon > 0$,

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| < \epsilon) = 1;$$

that is, $\bar{X}_n$ *converges in probability to* μ.

Proof: The proof is quite simple, being a straightforward application of Chebychev's Inequality. We have, for every $\epsilon > 0$,

$$P(|\bar{X}_n - \mu| \geq \epsilon) = P((\bar{X}_n - \mu)^2 \geq \epsilon^2) \leq \frac{\mathrm{E}(\bar{X}_n - \mu)^2}{\epsilon^2} = \frac{\mathrm{Var}\ \bar{X}_n}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}.$$

Hence, $P(|\bar{X}_n - \mu| < \epsilon) = 1 - P(|\bar{X}_n - \mu| \geq \epsilon) \geq 1 - \sigma^2/(n\epsilon^2) \to 1$, as $n \to \infty$. □

The Weak Law of Large Numbers (WLLN) quite elegantly states that, under general conditions, the sample mean approaches the population mean as $n \to \infty$. In fact, there are more general versions of the WLLN, where we need assume only that the mean is finite. However, the version stated in Theorem 5.5.2 is applicable in most practical situations.

The property summarized by the WLLN, that a sequence of the "same" sample quantity approaches a constant as $n \to \infty$, is known as *consistency*. We will examine this property more closely in Chapter 7.

Example 5.5.3 (Consistency of S^2) Suppose we have a sequence $X_1, X_2, \ldots$ of iid random variables with $\mathrm{E}X_i = \mu$ and $\mathrm{Var}\ X_i = \sigma^2 < \infty$. If we define

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2,$$

can we prove a WLLN for S_n^2? Using Chebychev's Inequality, we have

$$P(|S_n^2 - \sigma^2| \geq \epsilon) \leq \frac{\mathrm{E}(S_n^2 - \sigma^2)^2}{\epsilon^2} = \frac{\mathrm{Var}\ S_n^2}{\epsilon^2}$$

and thus, a sufficient condition that S_n^2 converges in probability to σ^2 is that $\mathrm{Var}\ S_n^2 \to 0$ as $n \to \infty$. ‖

A natural extension of Definition 5.5.1 relates to functions of random variables. That is, if the sequence $X_1, X_2, \ldots$ converges in probability to a random variable X or to a constant a, can we make any conclusions about the sequence of random variables $h(X_1), h(X_2), \ldots$ for some reasonably behaved function h? This next theorem shows that we can. (See Exercise 5.39 for a proof.)

Theorem 5.5.4 *Suppose that* $X_1, X_2, \ldots$ *converges in probability to a random variable* X *and that* h *is a continuous function. Then* $h(X_1), h(X_2), \ldots$ *converges in probability to* $h(X)$.

Example 5.5.5 (Consistency of S) If S_n^2 is a consistent estimator of σ^2, then by Theorem 5.5.4, the sample standard deviation $S_n = \sqrt{S_n^2} = h(S_n^2)$ is a consistent estimator of σ. Note that S_n is, in fact, a biased estimator of σ (see Exercise 5.11), but the bias disappears asymptotically. ‖

5.5.2 Almost Sure Convergence

A type of convergence that is stronger than convergence in probability is almost sure convergence (sometimes confusingly known as *convergence with probability 1*). This type of convergence is similar to pointwise convergence of a sequence of functions, except that the convergence need not occur on a set with probability 0 (hence the "almost" sure).

Definition 5.5.6 A sequence of random variables, $X_1, X_2, \ldots$, *converges almost surely* to a random variable X if, for every $\epsilon > 0$,

$$P(\lim_{n\to\infty} |X_n - X| < \epsilon) = 1.$$

Notice the similarity in the statements of Definitions 5.5.1 and 5.5.6. Although they look similar, they are very different statements, with Definition 5.5.6 much stronger. To understand almost sure convergence, we must recall the basic definition of a random variable as given in Definition 1.4.1. A random variable is a real-valued function defined on a sample space S. If a sample space S has elements denoted by s, then $X_n(s)$ and $X(s)$ are all functions defined on S. Definition 5.5.6 states that X_n converges to X almost surely if the functions $X_n(s)$ converge to $X(s)$ for all $s \in S$ except perhaps for $s \in N$, where $N \subset S$ and $P(N) = 0$. Example 5.5.7 illustrates almost sure convergence. Example 5.5.8 illustrates the difference between convergence in probability and almost sure convergence.

Example 5.5.7 (Almost sure convergence) Let the sample space S be the closed interval $[0, 1]$ with the uniform probability distribution. Define random variables $X_n(s) = s + s^n$ and $X(s) = s$. For every $s \in [0, 1)$, $s^n \to 0$ as $n \to \infty$ and $X_n(s) \to s = X(s)$. However, $X_n(1) = 2$ for every n so $X_n(1)$ does not converge to $1 = X(1)$. But since the convergence occurs on the set $[0, 1)$ and $P([0, 1)) = 1$, X_n converges to X almost surely. ||

Example 5.5.8 (Convergence in probability, not almost surely) In this example we describe a sequence that converges in probability, but not almost surely. Again, let the sample space S be the closed interval [0,1] with the uniform probability distribution. Define the sequence $X_1, X_2, \ldots$ as follows:

$$X_1(s) = s + I_{[0,1]}(s), \quad X_2(s) = s + I_{[0,\frac{1}{2}]}(s), \quad X_3(s) = s + I_{[\frac{1}{2},1]}(s),$$
$$X_4(s) = s + I_{[0,\frac{1}{3}]}(s), \quad X_5(s) = s + I_{[\frac{1}{3},\frac{2}{3}]}(s), \quad X_6(s) = s + I_{[\frac{2}{3},1]}(s),$$

etc. Let $X(s) = s$. It is straightforward to see that X_n converges to X in probability. As $n \to \infty, P(|X_n - X| \geq \epsilon)$ is equal to the probability of an interval of s values whose length is going to 0. However, X_n does not converge to X almost surely. Indeed, there is no value of $s \in S$ for which $X_n(s) \to s = X(s)$. For every s, the value $X_n(s)$ alternates between the values s and $s + 1$ infinitely often. For example, if $s = \frac{3}{8}, X_1(s) = 1\frac{3}{8}, X_2(s) = 1\frac{3}{8}, X_3(s) = \frac{3}{8}, X_4(s) = \frac{3}{8}, X_5(s) = 1\frac{3}{8}, X_6(s) = \frac{3}{8}$, etc. No pointwise convergence occurs for this sequence. ||

As might be guessed, convergence almost surely, being the stronger criterion, implies convergence in probability. The converse is, of course, false, as Example 5.5.8 shows. However, if a sequence converges in probability, it is possible to find a *subsequence* that converges almost surely. (Resnick 1999, Section 6.3, has a thorough treatment of the connections between the two types of convergence.)

Again, statisticians are often concerned with convergence to a constant. We now state, without proof, the stronger analog of the WLLN, the Strong Law of Large Numbers (SLLN). See Miscellanea 5.8.4 for an outline of a proof.

Theorem 5.5.9 (Strong Law of Large Numbers) *Let $X_1, X_2, \ldots$ be iid random variables with $\mathrm{E}X_i = \mu$ and $\mathrm{Var}\, X_i = \sigma^2 < \infty$, and define $\bar{X}_n = (1/n)\sum_{i=1}^n X_i$. Then, for every $\epsilon > 0$,*

$$P(\lim_{n\to\infty} |\bar{X}_n - \mu| < \epsilon) = 1;$$

that is, $\bar{X}_n$ converges almost surely to μ.

For both the Weak and Strong Laws of Large Numbers we had the assumption of a finite variance. Although such an assumption is true (and desirable) in most applications, it is, in fact, a stronger assumption than is needed. Both the weak and strong laws hold without this assumption. The only moment condition needed is that $\mathrm{E}|X_i| < \infty$ (see Resnick 1999, Chapter 7, or Billingsley 1995, Section 22).

5.5.3 Convergence in Distribution

We have already encountered the idea of convergence in distribution in Chapter 2. Remember the properties of moment generating functions (mgfs) and how their convergence implies convergence in distribution (Theorem 2.3.12).

Definition 5.5.10 A sequence of random variables, $X_1, X_2, \ldots$, *converges in distribution* to a random variable X if

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

at all points x where $F_X(x)$ is continuous.

Example 5.5.11 (Maximum of uniforms) If $X_1, X_2, \ldots$ are iid uniform$(0,1)$ and $X_{(n)} = \max_{1\le i\le n} X_i$, let us examine if (and to where) $X_{(n)}$ converges in distribution.

As $n \to \infty$, we expect $X_{(n)}$ to get close to 1 and, as $X_{(n)}$ must necessarily be less than 1, we have for any $\varepsilon > 0$,

$$\begin{aligned} P(|X_{(n)} - 1| \ge \varepsilon) &= P(X_{(n)} \ge 1 + \varepsilon) + P(X_{(n)} \le 1 - \varepsilon) \\ &= 0 + P(X_{(n)} \le 1 - \varepsilon). \end{aligned}$$

Next using the fact that we have an *iid* sample, we can write

$$P(X_{(n)} \le 1 - \varepsilon) = P(X_i \le 1 - \varepsilon, i = 1, \ldots n) = (1 - \varepsilon)^n,$$

which goes to 0. So we have proved that $X_{(n)}$ converges to 1 in probability. However, if we take $\varepsilon = t/n$, we then have

$$P(X_{(n)} \le 1 - t/n) = (1 - t/n)^n \to e^{-t},$$

which, upon rearranging, yields

$$P(n(1 - X_{(n)}) \le t) \to 1 - e^{-t};$$

that is, the random variable $n(1-X_{(n)})$ converges in distribution to an exponential(1) random variable. ‖

Note that although we talk of a sequence of random variables converging in distribution, it is really the cdfs that converge, not the random variables. In this very fundamental way convergence in distribution is quite different from convergence in probability or convergence almost surely. However, it is implied by the other types of convergence.

Theorem 5.5.12 *If the sequence of random variables, $X_1, X_2, \ldots$, converges in probability to a random variable X, the sequence also converges in distribution to X.*

See Exercise 5.40 for a proof. Note also that, from Section 5.5.2, convergence in distribution is also implied by almost sure convergence.

In a special case, Theorem 5.5.12 has a converse that turns out to be useful. See Example 10.1.13 for an illustration and Exercise 5.41 for a proof.

Theorem 5.5.13 *The sequence of random variables, $X_1, X_2, \ldots$, converges in probability to a constant μ if and only if the sequence also converges in distribution to μ. That is, the statement*

$$P(|X_n - \mu| > \varepsilon) \to 0 \text{ for every } \varepsilon > 0$$

is equivalent to

$$P(X_n \le x) \to \begin{cases} 0 & \text{if } x < \mu \\ 1 & \text{if } x > \mu. \end{cases}$$

The sample mean is one statistic whose large-sample behavior is quite important. In particular, we want to investigate its limiting distribution. This is summarized in one of the most startling theorems in statistics, the Central Limit Theorem (CLT).

Theorem 5.5.14 (Central Limit Theorem) *Let $X_1, X_2, \ldots$ be a sequence of iid random variables whose mgfs exist in a neighborhood of 0 (that is, $M_{X_i}(t)$ exists for $|t| < h$, for some positive h). Let $\mathrm{E}X_i = \mu$ and $\mathrm{Var}\, X_i = \sigma^2 > 0$. (Both μ and σ^2 are finite since the mgf exists.) Define $\bar{X}_n = (1/n)\sum_{i=1}^n X_i$. Let $G_n(x)$ denote the cdf of $\sqrt{n}(\bar{X}_n - \mu)/\sigma$. Then, for any x, $-\infty < x < \infty$,*

$$\lim_{n\to\infty} G_n(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy;$$

that is, $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ has a limiting standard normal distribution.

Before we prove this theorem (the proof is somewhat anticlimactic) we first look at its implications. Starting from virtually no assumptions (other than independence and finite variances), we end up with normality! The point here is that normality comes from sums of "small" (finite variance), independent disturbances. The assumption of finite variances is essentially necessary for convergence to normality. Although it can be relaxed somewhat, it cannot be eliminated. (Recall Example 5.2.10, dealing with the Cauchy distribution, where there is no convergence to normality.)

While we revel in the wonder of the CLT, it is also useful to reflect on its limitations. Although it gives us a useful general approximation, we have no automatic way of knowing how good the approximation is in general. In fact, the goodness of the approximation is a function of the original distribution, and so must be checked case by case. Furthermore, with the current availability of cheap, plentiful computing power, the importance of approximations like the Central Limit Theorem is somewhat lessened. However, despite its limitations, it is still a marvelous result.

Proof of Theorem 5.5.14: We will show that, for $|t| < h$, the mgf of $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ converges to $e^{t^2/2}$, the mgf of a n(0, 1) random variable.

Define $Y_i = (X_i - \mu)/\sigma$, and let $M_Y(t)$ denote the common mgf of the Y_is, which exists for $|t| < \sigma h$ and is given by Theorem 2.3.15. Since

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i, \tag{5.5.1}$$

we have, from the properties of mgfs (see Theorems 2.3.15 and 4.6.7),

$$\begin{aligned} M_{\sqrt{n}(\bar{X}_n - \mu)/\sigma}(t) &= M_{\Sigma_{i=1}^n Y_i/\sqrt{n}}(t) \\ &= M_{\Sigma_{i=1}^n Y_i}\left(\frac{t}{\sqrt{n}}\right) && \text{(Theorem 2.3.15)} \\ &= \left(M_Y\left(\frac{t}{\sqrt{n}}\right)\right)^n. && \text{(Theorem 4.6.7)} \end{aligned} \tag{5.5.2}$$

We now expand $M_Y(t/\sqrt{n})$ in a Taylor series (power series) around 0. (See Definition 5.5.20.) We have

$$M_Y\left(\frac{t}{\sqrt{n}}\right) = \sum_{k=0}^{\infty} M_Y^{(k)}(0) \frac{(t/\sqrt{n})^k}{k!}, \tag{5.5.3}$$

where $M_Y^{(k)}(0) = \left(d^k/dt^k\right) M_Y(t)|_{t=0}$. Since the mgfs exist for $|t| < h$, the power series expansion is valid if $t < \sqrt{n}\sigma h$.

Using the facts that $M_Y^{(0)} = 1$, $M_Y^{(1)} = 0$, and $M_Y^{(2)} = 1$ (by construction, the mean and variance of Y are 0 and 1), we have

$$M_Y\left(\frac{t}{\sqrt{n}}\right) = 1 + \frac{(t/\sqrt{n})^2}{2!} + R_Y\left(\frac{t}{\sqrt{n}}\right), \tag{5.5.4}$$

where R_Y is the remainder term in the Taylor expansion,

$$R_Y\left(\frac{t}{\sqrt{n}}\right) = \sum_{k=3}^{\infty} M_Y^{(k)}(0)\frac{(t/\sqrt{n})^k}{k!}.$$

An application of Taylor's Theorem (Theorem 5.5.21) shows that, for fixed $t \neq 0$, we have

$$\lim_{n\to\infty} \frac{R_Y(t/\sqrt{n})}{(t/\sqrt{n})^2} = 0.$$

Since t is fixed, we also have

$$\lim_{n\to\infty} \frac{R_Y(t/\sqrt{n})}{(1/\sqrt{n})^2} = \lim_{n\to\infty} nR_Y\left(\frac{t}{\sqrt{n}}\right) = 0, \tag{5.5.5}$$

and (5.5.5) is also true at $t = 0$ since $R_Y(0/\sqrt{n}) = 0$. Thus, for any fixed t, we can write

$$\begin{aligned}\lim_{n\to\infty}\left(M_Y\left(\frac{t}{\sqrt{n}}\right)\right)^n &= \lim_{n\to\infty}\left[1 + \frac{(t/\sqrt{n})^2}{2!} + R_Y\left(\frac{t}{\sqrt{n}}\right)\right]^n \\ &= \lim_{n\to\infty}\left[1 + \frac{1}{n}\left(\frac{t^2}{2} + nR_Y\left(\frac{t}{\sqrt{n}}\right)\right)\right]^n \\ &= e^{t^2/2}\end{aligned} \tag{5.5.6}$$

by an application of Lemma 2.3.14, where we set $a_n = (t^2/2) + nR_Y(t/\sqrt{n})$. (Note that (5.5.5) implies that $a_n \to t^2/2$ as $n \to \infty$.) Since $e^{t^2/2}$ is the mgf of the n(0, 1) distribution, the theorem is proved. □

The Central Limit Theorem is valid in much more generality than is stated in Theorem 5.5.14 (see Miscellanea 5.8.1). In particular, all of the assumptions about mgfs are not needed—the use of characteristic functions (see Miscellanea 2.6.2) can replace them. We state the next theorem without proof. It is a version of the Central Limit Theorem that is general enough for almost all statistical purposes. Notice that the only assumption on the parent distribution is that it has finite variance.

Theorem 5.5.15 (Stronger form of the Central Limit Theorem) *Let $X_1, X_2, \ldots$ be a sequence of iid random variables with $\mathrm{E}X_i = \mu$ and $0 < \mathrm{Var}\, X_i = \sigma^2 < \infty$. Define $\bar{X}_n = (1/n)\sum_{i=1}^n X_i$. Let $G_n(x)$ denote the cdf of $\sqrt{n}(\bar{X}_n - \mu)/\sigma$. Then, for any x, $-\infty < x < \infty$,*

$$\lim_{n\to\infty} G_n(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy;$$

that is, $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ has a limiting standard normal distribution.

The proof is almost identical to that of Theorem 5.5.14, except that characteristic functions are used instead of mgfs. Since the characteristic function of a distribution always exists, it is not necessary to mention them in the assumptions of the theorem. The proof is more delicate, however, since functions of *complex variables* must be dealt with. Details can be found in Billingsley (1995, Section 27).

The Central Limit Theorem provides us with an all-purpose approximation (but remember the warning about the goodness of the approximation). In practice, it can always be used for a first, rough calculation.

Example 5.5.16 (Normal approximation to the negative binomial) Suppose $X_1, \ldots, X_n$ are a random sample from a negative binomial(r, p) distribution. Recall that

$$\mathrm{E}X = \frac{r(1-p)}{p}, \quad \mathrm{Var}\, X = \frac{r(1-p)}{p^2},$$

and the Central Limit Theorem tells us that

$$\frac{\sqrt{n}(\bar{X} - r(1-p)/p)}{\sqrt{r(1-p)/p^2}}$$

is approximately n$(0, 1)$. The approximate probability calculations are much easier than the exact calculations. For example, if $r = 10$, $p = \frac{1}{2}$, and $n = 30$, an exact calculation would be

$$\begin{aligned} P(\bar{X} \le 11) &= P\left(\sum_{i=1}^{30} X_i \le 330\right) \\ &= \sum_{x=0}^{330} \binom{300+x-1}{x} \left(\frac{1}{2}\right)^{300} \left(\frac{1}{2}\right)^{x} \qquad \left(\begin{array}{c}\sum X \text{ is negative} \\ \text{binomial}(nr, p)\end{array}\right) \\ &= .8916, \end{aligned}$$

which is a very difficult calculation. (Note that this calculation is difficult even with the aid of a computer—the magnitudes of the factorials cause great difficulty. Try it if you don't believe it!) The CLT gives us the approximation

$$\begin{aligned} P(\bar{X} \le 11) &= P\left(\frac{\sqrt{30}(\bar{X} - 10)}{\sqrt{20}} \le \frac{\sqrt{30}(11 - 10)}{\sqrt{20}}\right) \\ &\approx P(Z \le 1.2247) = .8888. \end{aligned}$$

See Exercise 5.37 for some further refinement. ||

An approximation tool that can be used in conjunction with the Central Limit Theorem is known as Slutsky's Theorem.

Theorem 5.5.17 (Slutsky's Theorem) *If $X_n \to X$ in distribution and $Y_n \to a$, a constant, in probability, then*

a. $Y_n X_n \to aX$ *in distribution.*

b. $X_n + Y_n \to X + a$ *in distribution.*

The proof of Slutsky's Theorem is omitted, since it relies on a characterization of convergence in distribution that we have not discussed. A typical application is illustrated by the following example.

Example 5.5.18 (Normal approximation with estimated variance) Suppose that

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \to \mathrm{n}(0,1),$$

but the value of σ is unknown. We have seen in Example 5.5.3 that, if $\lim_{n\to\infty} \mathrm{Var}\, S_n^2 = 0$, then $S_n^2 \to \sigma^2$ in probability. By Exercise 5.32, $\sigma/S_n \to 1$ in probability. Hence, Slutsky's Theorem tells us

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} = \frac{\sigma}{S_n}\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \to \mathrm{n}(0,1). \qquad \|$$

5.5.4 The Delta Method

The previous section gives conditions under which a standardized random variable has a limit normal distribution. There are many times, however, when we are not specifically interested in the distribution of the random variable itself, but rather some function of the random variable.

Example 5.5.19 (Estimating the odds) Suppose we observe $X_1, X_2, \ldots, X_n$ independent Bernoulli(p) random variables. The typical parameter of interest is p, the success probability, but another popular parameter is $\frac{p}{1-p}$, the *odds*. For example, if the data represent the outcomes of a medical treatment with $p = 2/3$, then a person has odds 2 : 1 of getting better. Moreover, if there were another treatment with success probability r, biostatisticians often estimate the *odds ratio* $\frac{p}{1-p}/\frac{r}{1-r}$, giving the relative odds of one treatment over another.

As we would typically estimate the success probability p with the observed success probability $\hat{p} = \sum_i X_i/n$, we might consider using $\frac{\hat{p}}{1-\hat{p}}$ as an estimate of $\frac{p}{1-p}$. But what are the properties of this estimator? How might we estimate the variance of $\frac{\hat{p}}{1-\hat{p}}$? Moreover, how can we approximate its sampling distribution?

Intuition abandons us, and exact calculation is relatively hopeless, so we have to rely on an approximation. The Delta Method will allow us to obtain reasonable, approximate answers to our questions. $\|$

One method of proceeding is based on using a Taylor series approximation, which allows us to approximate the mean and variance of a function of a random variable. We will also see that these rather straightforward approximations are good enough to obtain a CLT. We begin with a short review of Taylor series.

Definition 5.5.20 If a function $g(x)$ has derivatives of order r, that is, $g^{(r)}(x) = \frac{d^r}{dx^r}g(x)$ exists, then for any constant a, the *Taylor polynomial of order r about a* is

$$T_r(x) = \sum_{i=0}^{r} \frac{g^{(i)}(a)}{i!}(x-a)^i.$$

Taylor's major theorem, which we will not prove here, is that the *remainder* from the approximation, $g(x) - T_r(x)$, always tends to 0 faster than the highest-order explicit term.

Theorem 5.5.21 (Taylor) *If* $g^{(r)}(a) = \frac{d^r}{dx^r}g(x)\big|_{x=a}$ *exists, then*

$$\lim_{x\to a}\frac{g(x) - T_r(x)}{(x-a)^r} = 0.$$

In general, we will not be concerned with the explicit form of the remainder. Since we are interested in approximations, we are just going to ignore the remainder. There are, however, many explicit forms, one useful one being

$$g(x) - T_r(x) = \int_a^x \frac{g^{(r+1)}(t)}{r!}(x-t)^r dt.$$

For the statistical application of Taylor's Theorem, we are most concerned with the *first-order* Taylor series, that is, an approximation using just the first derivative (taking $r = 1$ in the above formulas). Furthermore, we will also find use for a multivariate Taylor series. Since the above detail is univariate, some of the following will have to be accepted on faith.

Let $T_1, \ldots, T_k$ be random variables with means $\theta_1, \ldots, \theta_k$, and define $\mathbf{T} = (T_1, \ldots, T_k)$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$. Suppose there is a differentiable function $g(\mathbf{T})$ (an estimator of some parameter) for which we want an approximate estimate of variance. Define

$$g_i'(\boldsymbol{\theta}) = \frac{\partial}{\partial t_i}g(\mathbf{t})\Big|_{t_1=\theta_1,\ldots,t_k=\theta_k}.$$

The first-order Taylor series expansion of g about $\boldsymbol{\theta}$ is

$$g(\mathbf{t}) = g(\boldsymbol{\theta}) + \sum_{i=1}^{k} g_i'(\boldsymbol{\theta})(t_i - \theta_i) + \text{Remainder}.$$

For our statistical approximation we forget about the remainder and write

$$g(\mathbf{t}) \approx g(\boldsymbol{\theta}) + \sum_{i=1}^{k} g_i'(\boldsymbol{\theta})(t_i - \theta_i). \tag{5.5.7}$$

Now, take expectations on both sides of (5.5.7) to get

$$\begin{aligned} \mathrm{E}_{\boldsymbol{\theta}} g(\mathbf{T}) &\approx g(\boldsymbol{\theta}) + \sum_{i=1}^{k} g_i'(\boldsymbol{\theta})\mathrm{E}_{\boldsymbol{\theta}}(T_i - \theta_i) \\ &= g(\boldsymbol{\theta}). \qquad (T_i \text{ has mean } \theta_i) \end{aligned} \tag{5.5.8}$$

We can now approximate the variance of $g(\mathbf{T})$ by

$$\begin{aligned}
\mathrm{Var}_{\boldsymbol{\theta}}\, g(\mathbf{T}) &\approx \mathrm{E}_{\boldsymbol{\theta}}\left([g(\mathbf{T}) - g(\boldsymbol{\theta})]^2\right) && \text{(using (5.5.8))} \\
&\approx \mathrm{E}_{\boldsymbol{\theta}}\left(\left(\sum\nolimits_{i=1}^{k} g_i'(\boldsymbol{\theta})(T_i - \theta_i)\right)^2\right) && \text{(using (5.5.7))} \\
&= \sum_{i=1}^{k}[g_i'(\boldsymbol{\theta})]^2 \mathrm{Var}_{\theta}\, T_i + 2\sum_{i>j} g_i'(\boldsymbol{\theta}) g_j'(\boldsymbol{\theta}) \mathrm{Cov}_{\theta}(T_i, T_j), && (5.5.9)
\end{aligned}$$

where the last equality comes from expanding the square and using the definition of variance and covariance (similar to Exercise 4.44). Approximation (5.5.9) is very useful because it gives us a variance formula for a general function, using only simple variances and covariances. Here are two examples.

Example 5.5.22 (Continuation of Example 5.5.19) Recall that we are interested in the properties of $\frac{\hat{p}}{1-\hat{p}}$ as an estimate of $\frac{p}{1-p}$, where p is a binomial success probability. In our above notation, take $g(p) = \frac{p}{1-p}$ so $g'(p) = \frac{1}{(1-p)^2}$ and

$$\begin{aligned}
\mathrm{Var}\left(\frac{\hat{p}}{1-\hat{p}}\right) &\approx [g'(p)]^2\, \mathrm{Var}(\hat{p}) \\
&= \left[\frac{1}{(1-p)^2}\right]^2 \frac{p(1-p)}{n} = \frac{p}{n(1-p)^3},
\end{aligned}$$

giving us an approximation for the variance of our estimator. ‖

Example 5.5.23 (Approximate mean and variance) Suppose X is a random variable with $\mathrm{E}_{\mu} X = \mu \neq 0$. If we want to estimate a function $g(\mu)$, a first-order approximation would give us

$$g(X) = g(\mu) + g'(\mu)(X - \mu).$$

If we use $g(X)$ as an estimator of $g(\mu)$, we can say that approximately

$$\begin{aligned}
\mathrm{E}_{\mu} g(X) &\approx g(\mu), \\
\mathrm{Var}_{\mu}\, g(X) &\approx [g'(\mu)]^2 \mathrm{Var}_{\mu}\, X.
\end{aligned}$$

For a specific example, take $g(\mu) = 1/\mu$. We estimate $1/\mu$ with $1/X$, and we can say

$$\begin{aligned}
\mathrm{E}_{\mu}\left(\frac{1}{X}\right) &\approx \frac{1}{\mu}, \\
\mathrm{Var}_{\mu}\left(\frac{1}{X}\right) &\approx \left(\frac{1}{\mu}\right)^4 \mathrm{Var}_{\mu} X.
\end{aligned}$$

‖

Using these Taylor series approximations for the mean and variance, we get the following useful generalization of the Central Limit Theorem, known as the *Delta Method*.

Theorem 5.5.24 (Delta Method) *Let Y_n be a sequence of random variables that satisfies $\sqrt{n}(Y_n - \theta) \to \mathrm{n}(0, \sigma^2)$ in distribution. For a given function g and a specific value of θ, suppose that $g'(\theta)$ exists and is not 0. Then*

$$\sqrt{n}[g(Y_n) - g(\theta)] \to \mathrm{n}(0, \sigma^2[g'(\theta)]^2) \text{ in distribution.} \tag{5.5.10}$$

Proof: The Taylor expansion of $g(Y_n)$ around $Y_n = \theta$ is

$$g(Y_n) = g(\theta) + g'(\theta)(Y_n - \theta) + \text{ Remainder}, \tag{5.5.11}$$

where the remainder $\to 0$ as $Y_n \to \theta$. Since $Y_n \to \theta$ in probability it follows that the remainder $\to 0$ in probability. By applying Slutsky's Theorem (Theorem 5.5.17) to

$$\sqrt{n}[g(Y_n) - g(\theta)] = g'(\theta)\sqrt{n}(Y_n - \theta),$$

the result now follows. See Exercise 5.43 for details. □

Example 5.5.25 (Continuation of Example 5.5.23) Suppose now that we have the mean of a random sample $\bar{X}$. For $\mu \neq 0$, we have

$$\sqrt{n}\left(\frac{1}{\bar{X}} - \frac{1}{\mu}\right) \to \mathrm{n}\left(0, \left(\frac{1}{\mu}\right)^4 \mathrm{Var}_\mu X_1\right)$$

in distribution.

If we do not know the variance of X_1, to use the above approximation requires an estimate, say S^2. Moreover, there is the question of what to do with the $1/\mu$ term, as we also do not know μ. We can estimate everything, which gives us the approximate variance

$$\widehat{\mathrm{Var}}\left(\frac{1}{\bar{X}}\right) \approx \left(\frac{1}{\bar{X}}\right)^4 S^2.$$

Furthermore, as both $\bar{X}$ and S^2 are consistent estimators, we can again apply Slutsky's Theorem to conclude that for $\mu \neq 0$,

$$\frac{\sqrt{n}\left(\frac{1}{\bar{X}} - \frac{1}{\mu}\right)}{\left(\frac{1}{\bar{X}}\right)^2 S} \to \mathrm{n}(0, 1)$$

in distribution.

Note how we wrote this latter quantity, dividing through by the estimated standard deviation and making the limiting distribution a standard normal. This is the only way that makes sense if we need to estimate any parameters in the limiting distribution. We also note that there is an alternative approach when there are parameters to estimate, and here we can actually avoid using an estimate for μ in the variance (see the score test in Section 10.3.2). ‖

There are two extensions of the basic Delta Method that we need to deal with to complete our treatment. The first concerns the possibility that $g'(\mu) = 0$. This could

happen, for example, if we were interested in estimating the variance of a binomial variance (see Exercise 5.44).

If $g'(\theta) = 0$, we take one more term in the Taylor expansion to get

$$g(Y_n) = g(\theta) + g'(\theta)(Y_n - \theta) + \frac{g''(\theta)}{2}(Y_n - \theta)^2 + \text{ Remainder.}$$

If we do some rearranging (setting $g' = 0$), we have

$$g(Y_n) - g(\theta) = \frac{g''(\theta)}{2}(Y_n - \theta)^2 + \text{ Remainder.} \tag{5.5.12}$$

Now recall that the square of a $n(0,1)$ is a χ^2_1 (Example 2.1.9), which implies that

$$\frac{n(Y_n - \theta)^2}{\sigma^2} \to \chi^2_1$$

in distribution. Therefore, an argument similar to that used in Theorem 5.5.24 will establish the following theorem.

Theorem 5.5.26 (Second-order Delta Method) *Let Y_n be a sequence of random variables that satisfies $\sqrt{n}(Y_n - \theta) \to n(0, \sigma^2)$ in distribution. For a given function g and a specific value of θ, suppose that $g'(\theta) = 0$ and $g''(\theta)$ exists and is not 0. Then*

$$n[g(Y_n) - g(\theta)] \to \sigma^2 \frac{g''(\theta)}{2}\chi^2_1 \text{ in distribution.} \tag{5.5.13}$$

Approximation techniques are very useful when more than one parameter makes up the function to be estimated and more than one random variable is used in the estimator. One common example is in growth studies, where a ratio of weight/height is a variable of interest. (Recall that in Chapter 3 we saw that a ratio of two *normal* random variables has a Cauchy distribution. The ratio problem, while being important to experimenters, is nasty in theory.)

This brings us to the second extension of the Delta Method, to the multivariate case. As we already have Taylor's Theorem for the multivariate case, this extension contains no surprises.

Example 5.5.27 (Moments of a ratio estimator) Suppose X and Y are random variables with nonzero means μ_X and μ_Y, respectively. The parametric function to be estimated is $g(\mu_X, \mu_Y) = \mu_X/\mu_Y$. It is straightforward to calculate

$$\frac{\partial}{\partial \mu_X} g(\mu_X, \mu_Y) = \frac{1}{\mu_Y}$$

and

$$\frac{\partial}{\partial \mu_Y} g(\mu_X, \mu_Y) = \frac{-\mu_X}{\mu_Y^2}.$$

The first-order Taylor approximations (5.5.8) and (5.5.9) give

$$\mathrm{E}\left(\frac{X}{Y}\right) \approx \frac{\mu_X}{\mu_Y}$$

and

$$\begin{aligned}\mathrm{Var}\left(\frac{X}{Y}\right) &\approx \frac{1}{\mu_Y^2}\mathrm{Var}\,X + \frac{\mu_X^2}{\mu_Y^4}\mathrm{Var}\,Y - 2\frac{\mu_X}{\mu_Y^3}\mathrm{Cov}(X,Y) \\ &= \left(\frac{\mu_X}{\mu_Y}\right)^2\left(\frac{\mathrm{Var}\,X}{\mu_X^2} + \frac{\mathrm{Var}\,Y}{\mu_Y^2} - 2\frac{\mathrm{Cov}(X,Y)}{\mu_X\mu_Y}\right).\end{aligned}$$

Thus, we have an approximation for the mean and variance of the ratio estimator, and the approximations use only the means, variances, and covariance of $\bar{X}$ and Y. Exact calculations would be quite hopeless, with closed-form expressions being unattainable. ||

We next present a CLT to cover an estimator such as the ratio estimator. Note that we must deal with multiple random variables although the ultimate CLT is a univariate one. Suppose the vector-valued random variable $\mathbf{X} = (X_1, \ldots, X_p)$ has mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)$ and covariances $\mathrm{Cov}(X_i, X_j) = \sigma_{ij}$, and we observe an independent random sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ and calculate the means $\bar{X}_i = \sum_{k=1}^n X_{ik}$, $i = 1, \ldots, p$. For a function $g(\mathbf{x}) = g(x_1, \ldots, x_p)$ we can use the development after (5.5.7) to write

$$g(\bar{x}_1, \ldots, \bar{x}_p) = g(\mu_1, \ldots, \mu_p) + \sum_{k=1}^{p} g_k'(\mathbf{x})(\bar{x}_k - \mu_k),$$

and we then have the following theorem.

Theorem 5.5.28 (Multivariate Delta Method) *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a random sample with $\mathrm{E}(X_{ij}) = \mu_i$ and $\mathrm{Cov}(X_{ik}, X_{jk}) = \sigma_{ij}$. For a given function g with continuous first partial derivatives and a specific value of $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)$ for which $\tau^2 = \Sigma\Sigma\sigma_{ij}\frac{\partial g(\boldsymbol{\mu})}{\partial \mu_i} \cdot \frac{\partial g(\boldsymbol{\mu})}{\partial \mu_j} > 0$,*

$$\sqrt{n}[g(\bar{X}_1, \ldots, \bar{X}_s) - g(\mu_1, \ldots, \mu_p)] \to \mathrm{n}(0, \tau^2) \text{ in distribution .}$$

The proof necessitates dealing with the convergence of multivariate random variables, and we will not deal with such multivariate intricacies here, but will take Theorem 5.5.28 on faith. The interested reader can find more details in Lehmann and Casella (1998, Section 1.8).

5.6 Generating a Random Sample

Thus far we have been concerned with the many methods of describing the behavior of random variables—transformations, distributions, moment calculations, limit theorems. In practice, these random variables are used to describe and model real phenomena, and observations on these random variables are the data that we collect.

Thus, typically, we observe random variables $X_1, \ldots, X_n$ from a distribution $f(x|\theta)$ and are most concerned with using properties of $f(x|\theta)$ to describe the behavior of the random variables. In this section we are, in effect, going to turn that strategy around. Here we are concerned with *generating* a random sample $X_1, \ldots, X_n$ from a given distribution $f(x|\theta)$.

Example 5.6.1 (Exponential lifetime) Suppose that a particular electrical component is to be modeled with an exponential(λ) lifetime. The manufacturer is interested in determining the probability that, out of c components, at least t of them will last h hours. Taking this one step at a time, we have

$$\begin{aligned} p_1 &= P(\text{component lasts at least } h \text{ hours}) \\ &= P(X \geq h|\lambda), \end{aligned} \tag{5.6.1}$$

and assuming that the components are independent, we can model the outcomes of the c components as Bernoulli trials, so

$$\begin{aligned} p_2 &= P(\text{at least } t \text{ components last } h \text{ hours}) \\ &= \sum_{k=t}^{c} \binom{c}{k} p_1^k (1-p_1)^{c-k}. \end{aligned} \tag{5.6.2}$$

Although calculation of (5.6.2) is straightforward, it can be computationally burdensome, especially if both t and c are large numbers. Moreover, the exponential model has the advantage that p_1 can be expressed in closed form, that is,

$$p_1 = \int_h^\infty \frac{1}{\lambda} e^{-x/\lambda}\, dx = e^{-h/\lambda}. \tag{5.6.3}$$

However, if each component were modeled with, say, a gamma distribution, then p_1 may not be expressible in closed form. This would make calculation of p_2 even more involved. ||

A simulation approach to the calculation of expressions such as (5.6.2) is to generate random variables with the desired distribution and then use the Weak Law of Large Numbers (Theorem 5.5.2) to validate the simulation. That is, if Y_i, $i = 1, \ldots, n$, are iid, then a consequence of that theorem (provided the assumptions hold) is

$$\frac{1}{n} \sum_{i=1}^{n} h(Y_i) \longrightarrow \mathrm{E}h(Y) \tag{5.6.4}$$

in probability, as $n \to \infty$. (Expression (5.6.4) also holds almost everywhere, a consequence of Theorem 5.5.9, the Strong Law of Large Numbers.)

Example 5.6.2 (Continuation of Example 5.6.1) The probability p_2 can be calculated using the following steps. For $j = 1, \ldots, n$:

a. Generate $X_1, \ldots, X_c$ iid $\sim$ exponential(λ).
b. Set $Y_j = 1$ if at least t X_is are $\geq h$; otherwise, set $Y_j = 0$.
Then, because $Y_j \sim$ Bernoulli(p_2) and $EY_j = p_2$,

$$\frac{1}{n}\sum_{j=1}^{n} Y_j \to p_2 \text{ as } n \to \infty. \qquad \|$$

Examples 5.6.1 and 5.6.2 highlight the major concerns of this section. First, we must examine how to generate the random variables that we need, and second, we then use a version of the Law of Large Numbers to validate the simulation approximation.

Since we have to start somewhere, we start with the assumption that we can generate iid uniform random variables $U_1, \ldots, U_m$. (This problem of generating uniform random numbers has been worked on, with great success, by computer scientists.) There exist many algorithms for generating *pseudo-random* numbers that will pass almost all tests for uniformity. Moreover, most good statistical packages have a reasonable uniform random number generator. (See Devroye 1985 or Ripley 1987 for more on generating pseudo-random numbers.)

Since we are starting from the uniform random variables, our problem here is really not the problem of generating the desired random variables, but rather of *transforming* the uniform random variables to the desired distribution. In essence, there are two general methodologies for doing this, which we shall (noninformatively) call *direct* and *indirect* methods.

5.6.1 Direct Methods

A *direct method* of generating a random variable is one for which there exists a closed-form function $g(u)$ such that the transformed variable $Y = g(U)$ has the desired distribution when $U \sim$ uniform(0, 1). As might be recalled, this was already accomplished for continuous random variables in Theorem 2.1.10, the Probability Integral Transform, where any distribution was transformed to the uniform. Hence the inverse transformation solves our problem.

Example 5.6.3 (Probability Integral Transform) If Y is a continuous random variable with cdf F_Y, then Theorem 2.1.10 implies that the random variable $F_Y^{-1}(U)$, where $U \sim$ uniform(0, 1), has distribution F_Y. If $Y \sim$ exponential(λ), then

$$F_Y^{-1}(U) = -\lambda \log(1 - U)$$

is an exponential(λ) random variable (see Exercise 5.49).

Thus, if we generate $U_1, \ldots, U_n$ as iid uniform random variables, $Y_i = -\lambda \log(1 - U_i)$, $i = 1, \ldots, n$, are iid exponential(λ) random variables. As an example, for $n =$ 10,000, we generate $u_1, u_2, \ldots, u_{10,000}$ and calculate

$$\frac{1}{n}\sum u_i = .5019 \quad \text{and} \quad \frac{1}{n-1}\sum (u_i - \bar{u})^2 = .0842.$$

From (5.6.4), which follows from the WLLN (Theorem 5.5.2), we know that $\bar{U} \to EU = 1/2$ and, from Example 5.5.3, $S^2 \to \text{Var}\, U = 1/12 = .0833$, so our estimates

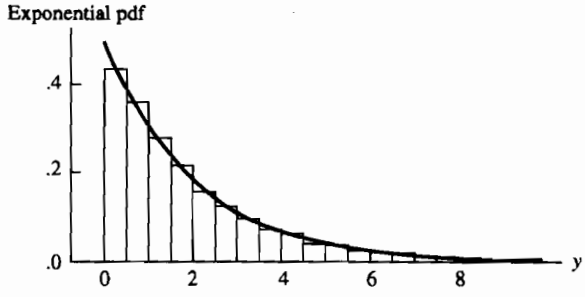

Figure 5.6.1. *Histogram of* 10,000 *observations from an exponential pdf with* $\lambda = 2$, *together with the pdf*

are quite close to the true parameters. The transformed variables $Y_i = -2\log(1-u_i)$ have an exponential(2) distribution, and we find that

$$\frac{1}{n}\sum y_i = 2.0004 \quad \text{and} \quad \frac{1}{n-1}\sum (y_i - \bar{y})^2 = 4.0908,$$

in close agreement with $EY = 2$ and $\text{Var}\, Y = 4$. Figure 5.6.1 illustrates the agreement between the sample histogram and the population pdf. ||

The relationship between the exponential and other distributions allows the quick generation of many random variables. For example, if U_j are iid uniform$(0,1)$ random variables, then $Y_j = -\lambda\log(u_j)$ are iid exponential (λ) random variables and

$$Y = -2\sum_{j=1}^{\nu}\log(U_j) \sim \chi^2_{2\nu},$$

(5.6.5)
$$Y = -\beta\sum_{j=1}^{a}\log(U_j) \sim \text{ gamma}(a,\beta),$$

$$Y = \frac{\sum_{j=1}^{a}\log(U_j)}{\sum_{j=1}^{a+b}\log(U_j)} \sim \text{ beta}(a,b).$$

Many other variations are possible (see Exercises 5.47–5.49), but all are being driven by the exponential-uniform transformation.

Unfortunately, there are limits to this transformation. For example, we cannot use it to generate χ^2 random variables with odd degrees of freedom. Hence, we cannot get a χ^2_1, which would in turn get us a normal$(0,1)$ — an extremely useful variable to be able to generate. We will return to this problem in the next subsection.

Recall that the basis of Example 5.6.3 and hence the transformations in (5.6.5) was the Probability Integral Transform, which, in general, can be written as

(5.6.6)
$$F_Y^{-1}(u) = y \leftrightarrow u = \int_{-\infty}^{y} f_y(t)dt.$$

Application of this formula to the exponential distribution was particularly handy, as the integral equation had a simple solution (see also Exercise 5.56). However, in many cases no closed-form solution for (5.6.6) will exist. Thus, each random variable generation will necessitate a solution of an integral equation, which, in practice, could be prohibitively long and complicated. This would be the case, for example, if (5.6.6) were used to generate a χ^2_1.

When no closed-form solution for (5.6.6) exists, other options should be explored. These include other types of generation methods and indirect methods. As an example of the former, consider the following.

Example 5.6.4 (Box-Muller algorithm) Generate U_1 and U_2, two independent uniform$(0, 1)$ random variables, and set

$$R = \sqrt{-2 \log U_1} \quad \text{and} \quad \theta = 2\pi U_2.$$

Then

$$X = R \cos\theta \quad \text{and} \quad Y = R \sin\theta$$

are independent normal$(0, 1)$ random variables. Thus, although we had no quick transformation for generating a single n$(0, 1)$ random variable, there is such a method for generating two variables. (See Exercise 5.50.) ||

Unfortunately, solutions such as those in Example 5.6.4 are not plentiful. Moreover, they take advantage of the specific structure of certain distributions and are, thus, less applicable as general strategies. It turns out that, for the most part, generation of other continuous distributions (than those already considered) is best accomplished through indirect methods. Before exploring these, we end this subsection with a look at where (5.6.6) is quite useful: the case of discrete random variables.

If Y is a discrete random variable taking on values $y_1 < y_2 < \cdots < y_k$, then analogous to (5.6.6) we can write

$$P[F_Y(y_i) < U \le F_Y(y_{i+1})] = F_Y(y_{i+1}) - F_Y(y_i) = P(Y = y_{i+1}). \tag{5.6.7}$$

Implementation of (5.6.7) to generate discrete random variables is actually quite straightforward and can be summarized as follows. To generate $Y_i \sim F_Y(y)$,

a. Generate $U \sim$ uniform$(0, 1)$.

b. If $F_y(y_i) < U \le F_y(y_{i+1})$, set $Y = y_{i+1}$.

We define $y_0 = -\infty$ and $F_Y(y_0) = 0$.

Example 5.6.5 (Binomial random variable generation) To generate a $Y \sim$ binomial$(4, \frac{5}{8})$, for example, generate $U \sim$ uniform$(0, 1)$ and set

$$Y = \begin{cases} 0 & \text{if } 0 < U \le .020, \\ 1 & \text{if } .020 < U \le .152 \\ 2 & \text{if } .152 < U \le .481 \\ 3 & \text{if } .481 < U \le .847 \\ 4 & \text{if } .847 < U \le 1. \end{cases} \tag{5.6.8}$$

||

The algorithm (5.6.8) also works if the range of the discrete random variable is infinite, say Poisson or negative binomial. Although, theoretically, this could require a large number of evaluations, in practice this does not happen because there are simple and clever ways of speeding up the algorithm. For example, instead of checking each y_i in the order $1, 2, \ldots$, it can be much faster to start checking y_is near the mean. (See Ripley 1987, Section 3.3, and Exercise 5.55.)

We will see many uses of simulation methodology. To start off, consider the following exploration of the Poisson distribution, which is a version of the *parametric bootstrap* that we will see in Section 10.1.4.

Example 5.6.6 (Distribution of the Poisson variance) If $X_1, \ldots, X_n$ are iid Poisson(λ), then by either Theorem 5.2.7 or 5.2.11 the distribution of $\sum X_i$ is Poisson($n\lambda$). Thus, it is quite easy to describe the distribution of the sample mean $\bar{X}$. However, describing the distribution of the sample variance, $S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$, is not a simple task.

The distribution of S^2 is quite simple to simulate, however. Figure 5.6.2 shows such a histogram. Moreover, the simulated samples can also be used to calculate probabilities about S^2. If S_i^2 is the value calculated from the ith simulated sample, then

$$\frac{1}{M}\sum_{i=1}^{M} I(S_i^2 \geq a) \to P_\lambda(S^2 \geq a)$$

as $M \to \infty$.

To illustrate the use of such methodology consider the following sample of bay anchovy larvae counts taken from the Hudson River in late August 1984:

(5.6.9) $$19, 32, 29, 13, 8, 12, 16, 20, 14, 17, 22, 18, 23.$$

If it is assumed that the larvae are distributed randomly and uniformly in the river, then the number that are collected in a fixed size net should follow a Poisson distribution. Such an argument follows from a spatial version of the Poisson postulates (see the Miscellanea of Chapter 2). To see if such an assumption is tenable, we can check whether the mean and variance of the observed data are consistent with the Poisson assumptions.

For the data in (5.6.9) we calculate $\bar{x} = 18.69$ and $s^2 = 44.90$. Under the Poisson assumptions we expect these values to be the same. Of course, due to sampling variability, they will not be exactly the same, and we can use a simulation to get some idea of what to expect. In Figure 5.6.2 we simulated 5,000 samples of size $n = 13$ from a Poisson distribution with $\lambda = 18.69$, and constructed the relative frequency histogram of S^2. Note that the observed value of $S^2 = 44.90$ falls into the tail of the distribution. In fact, since 27 of the values of S^2 were greater than 44.90, we can estimate

$$P(S^2 > 44.90|\lambda = 18.69) \approx \frac{1}{5000}\sum_{i=1}^{5000} I(S_i^2 > 44.90) = \frac{27}{5000} = .0054,$$

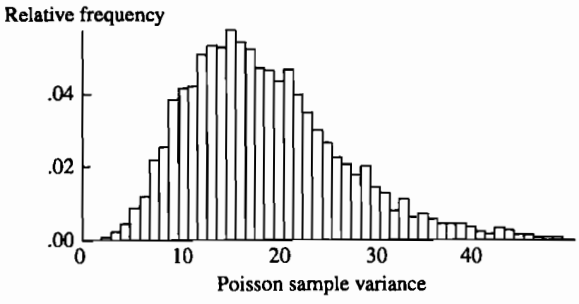

Figure 5.6.2. *Histogram of the sample variances, S^2, of* 5,000 *samples of size* 13 *from a Poisson distribution with* $\lambda = 18.69$. *The mean and standard deviation of the* 5,000 *values are* 18.86 *and* 7.68.

which leads us to question the Poisson assumption; see Exercise 5.54. (Such findings spawned the extremely bad bilingual pun "Something is fishy in the Hudson—the Poisson has failed.") ‖

5.6.2 Indirect Methods

When no easily found direct transformation is available to generate the desired random variable, an extremely powerful indirect method, the Accept/Reject Algorithm, can often provide a solution. The idea behind the Accept/Reject Algorithm is, perhaps, best explained through a simple example.

Example 5.6.7 (Beta random variable generation—I) Suppose the goal is to generate $Y \sim \text{beta}(a, b)$. If both a and b are integers, then the direct transformation method (5.6.5) can be used. However, if a and b are not integers, then that method will not work. For definiteness, set $a = 2.7$ and $b = 6.3$. In Figure 5.6.3 we have put the beta density $f_Y(y)$ inside a box with sides 1 and $c \geq \max_y f_Y(y)$. Now consider the following method of calculating $P(Y \leq y)$. If (U, V) are independent uniform$(0, 1)$ random variables, then the probability of the shaded area is

$$
\begin{aligned}
P\left(V \leq y, U \leq \frac{1}{c} f_Y(V)\right) &= \int_0^y \int_0^{f_Y(v)/c} du\, dv \\
&= \frac{1}{c} \int_0^y f_Y(v) dv \qquad (5.6.10) \\
&= \frac{1}{c} P(Y \leq y).
\end{aligned}
$$

So we can calculate the beta probabilities from the uniform probabilities, which suggests that we can generate the beta random variable from the uniform random variables.

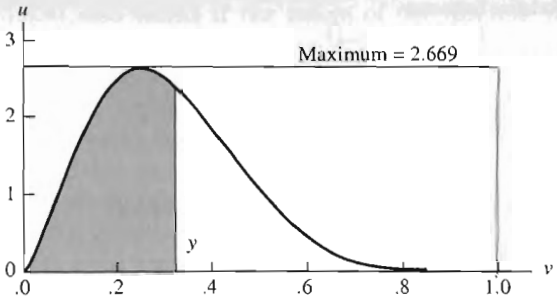

Figure 5.6.3. *The beta distribution with $a = 2.7$ and $b = 6.3$ with $c = \max_y f_Y(y) = 2.669$. The uniform random variable V gives the x-coordinate, and we use U to test if we are under the density.*

From (5.6.10), if we set $y = 1$, then we have $\frac{1}{c} = P(U < \frac{1}{c} f_Y(V))$, so

$$P(Y \le y) = \frac{P(V \le y, U \le \frac{1}{c} f_Y(V))}{P(U \le \frac{1}{c} f_Y(V))}$$

$$= P\left(V \le y | U \le \frac{1}{c} f_Y(V)\right), \tag{5.6.11}$$

which suggests the following algorithm.

To generate $Y \sim \text{beta}(a, b)$:

a. Generate (U, V) independent uniform$(0, 1)$.

b. If $U < \frac{1}{c} f_Y(V)$, set $Y = V$; otherwise, return to step (a).

This algorithm generates a beta(a, b) random variable as long as $c \ge \max_y f_Y(y)$ and, in fact, can be generalized to any bounded density with bounded support (Exercises 5.59 and 5.60). ‖

It should be clear that the optimal choice of c is $c = \max_y f_Y(y)$. To see why this is so, note that the algorithm is open-ended in the sense that we do not know how many (U, V) pairs will be needed in order to get one Y variable. However, if we define the random variable

$$N = \text{ number of } (U, V) \text{ pairs required for one } Y, \tag{5.6.12}$$

then, recalling that $\frac{1}{c} = P(U \le \frac{1}{c} f_Y(V))$, we see that N is a geometric$(1/c)$ random variable. Thus to generate one Y we expect to need $E(N) = c$ pairs (U, V), and in this sense minimizing c will optimize the algorithm.

Examining Figure 5.6.3 we see that the algorithm is wasteful in the area where $U > \frac{1}{c} f_Y(V)$. This is because we are using a uniform random variable V to get a beta random variable Y. To improve, we might start with something closer to the beta.

The testing step, step (b) of the algorithm, can be thought of as testing whether the random variable V "looks like" it could come from the density f_Y. Suppose that $V \sim f_V$, and we compute

$$M = \sup_y \frac{f_Y(y)}{f_V(y)} < \infty.$$

A generalization of step (b) is to compare $U \sim \text{uniform}(0,1)$ to $\frac{1}{M} f_Y(V)/f_V(V)$. The larger this ratio is, the more V "looks like" a random variable from the density f_Y, and the more likely it is that $U < \frac{1}{M} f_Y(V)/f_V(V)$. This is the basis of the general *Accept/Reject Algorithm*.

5.6.3 The Accept/Reject Algorithm

Theorem 5.6.8 *Let $Y \sim f_Y(y)$ and $V \sim f_V(V)$, where f_Y and f_V have common support with*

$$M = \sup_y f_Y(y)/f_V(y) < \infty.$$

To generate a random variable $Y \sim f_Y$:

a. *Generate $U \sim$ uniform$(0,1)$, $V \sim f_V$, independent.*

b. *If $U < \frac{1}{M} f_Y(V)/f_V(V)$, set $Y = V$; otherwise, return to step (a).*

Proof: The generated random variable Y has cdf

$$\begin{aligned}
P(Y \le y) &= P(V \le y | \text{stop}) \\
&= P\left(V \le y \middle| U < \frac{1}{M} f_Y(V)/f_V(V)\right) \\
&= \frac{P(V \le y, U < \frac{1}{M} f_Y(V)/f_V(V))}{P(U < \frac{1}{M} f_Y(V)/f_V(V))} \\
&= \frac{\int_{-\infty}^{y} \int_0^{\frac{1}{M} f_Y(v)/f_V(v)} du f_V(v) dv}{\int_{-\infty}^{\infty} \int_0^{\frac{1}{M} f_Y(v)/f_V(v)} du f_V(v) dv} \\
&= \int_{-\infty}^{y} f_Y(v) dv,
\end{aligned}$$

which is the desired cdf. □

Note also that

$$\begin{aligned}
M &= \sup_y f_Y(y)/f_V(y) \\
&= \left[P\left(U < \frac{1}{M} f_Y(V)/f_V(V)\right)\right]^{-1} \\
&= \frac{1}{P(\text{Stop})},
\end{aligned}$$

so the number of trials needed to generate one Y is a geometric$(1/M)$ random variable, and M is the expected number of trials.

Example 5.6.9 (Beta random variable generation–II) To generate $Y \sim$ beta$(2.7, 6.3)$ consider the algorithm:

a. Generate $U \sim$ uniform$(0,1)$, $V \sim$ beta$(2,6)$.

b. If $U < \frac{1}{M}\frac{f_Y(V)}{f_V(V)}$, set $Y = V$; otherwise, return to step (a).

This Accept/Reject Algorithm will generate the required Y as long as $\sup_y f_Y(y)/f_V(y) \leq M < \infty$. For the given densities we have $M = 1.67$, so the requirement is satisfied. (See Exercise 5.63.)

For this algorithm $EN = 1.67$, while for the algorithm of Example 5.6.7, which uses the uniform V, we have $EN = 2.67$. Although this seems to indicate that the latter algorithm is faster, remember that generating a beta$(2,6)$ random variable will need eight uniform random variables. Thus, comparison of algorithms is not always straightforward and will include consideration of both computer speed and programming ease. ‖

The importance of the requirement that $M < \infty$ should be stressed. This can be interpreted as requiring the density of V (often called the *candidate density*) to have heavier tails than the density of Y (often called the *target density*). This requirement tends to ensure that we will obtain a good representation of the values of Y, even those values that are in the tails. For example, if $V \sim$ Cauchy and $Y \sim$ n$(0,1)$, then we expect the range of V samples to be wider than that of Y samples, and we should get good performance from an Accept/Reject Algorithm based on these densities. However, it is much more difficult to change n$(0,1)$ random variables into Cauchy random variables because the extremes will be underrepresented.

There are cases, however, where the target density has heavy tails, and it is difficult to get candidate densities that will result in finite values of M. In such cases the Accept/Reject Algorithm will no longer apply, and one is led to another class of methods known as Markov Chain Monte Carlo (MCMC) methods. Special cases of such methods are known as the *Gibbs Sampler* and the *Metropolis Algorithm*. We state the latter.

Metropolis Algorithm Let $Y \sim f_Y(y)$ and $V \sim f_V(v)$, where f_Y and f_V have common support. To generate $Y \sim f_Y$:

0. Generate $V \sim f_V$. Set $Z_0 = V$.

For $i = 1, 2, \ldots$:

1. Generate $U_i \sim$ uniform$(0,1)$, $V_i \sim f_V$, and calculate

$$\rho_i = \min\left\{\frac{f_Y(V_i)}{f_V(V_i)} \cdot \frac{f_V(Z_{i-1})}{f_Y(Z_{i-1})}, 1\right\}.$$

2. Set

$$Z_i = \begin{cases} V_i & \text{if } U_i \leq \rho_i \\ Z_{i-1} & \text{if } U_i > \rho_i. \end{cases}$$

Then, as $i \to \infty$, Z_i converges to Y in distribution.

Although the algorithm does not require a finite M, it does not produce a random variable with exactly the density f_Y, but rather a convergent sequence. In practice, after the algorithm is run for a while (i gets big), the Zs that are produced behave very much like variables from f_Y. (See Chib and Greenberg 1995 for an elementary introduction to the Metropolis Algorithm.)

Although MCMC methods can be traced back to at least Metropolis *et al.* (1953), they entered real prominence with the work of Gelfand and Smith (1990), building on that of Geman and Geman (1984). See Miscellanea 5.8.5 for more details.

5.7 Exercises

5.1 Color blindness appears in 1% of the people in a certain population. How large must a sample be if the probability of its containing a color-blind person is to be .95 or more? (Assume that the population is large enough to be considered infinite, so that sampling can be considered to be with replacement.)

5.2 Suppose $X_1, X_2, \ldots$ are jointly continuous and independent, each distributed with marginal pdf $f(x)$, where each X_i represents annual rainfall at a given location.

(a) Find the distribution of the number of years until the first year's rainfall, X_1, is exceeded for the first time.

(b) Show that the mean number of years until X_1 is exceeded for the first time is infinite.

5.3 Let $X_1, \ldots, X_n$ be iid random variables with continuous cdf F_X, and suppose $EX_i = \mu$. Define the random variables $Y_1, \ldots, Y_n$ by

$$Y_i = \begin{cases} 1 & \text{if } X_i > \mu \\ 0 & \text{if } X_i \leq \mu. \end{cases}$$

Find the distribution of $\sum_{i=1}^n Y_i$.

5.4 A generalization of iid random variables is *exchangeable* random variables, an idea due to deFinetti (1972). A discussion of exchangeability can also be found in Feller (1971). The random variables $X_1, \ldots, X_n$ are *exchangeable* if any permutation of any subset of them of size k ($k \leq n$) has the same distribution. In this exercise we will see an example of random variables that are exchangeable but not iid. Let $X_i|P \sim$ iid Bernoulli(P), $i = 1, \ldots, n$, and let $P \sim$ uniform$(0, 1)$.

(a) Show that the marginal distribution of any k of the Xs is the same as

$$P(X_1 = x_1, \ldots, X_k = x_k) = \int_0^1 p^t(1-p)^{k-t}dp = \frac{t!(k-t)!}{(k+1)!},$$

where $t = \sum_{i=1}^k x_i$. Hence, the Xs are exchangeable.

(b) Show that, marginally,

$$P(X_1 = x_1, \ldots, X_n = x_n) \neq \Pi_{i=1}^n P(X_i = x_i),$$

so the distribution of the Xs is exchangeable but not iid.

(deFinetti proved an elegant characterization theorem for an infinite sequence of exchangeable random variables. He proved that any such sequence of exchangeable random variables is a mixture of iid random variables.)

5.5 Let $X_1, \ldots, X_n$ be iid with pdf $f_X(x)$, and let $\bar{X}$ denote the sample mean. Show that

$$f_{\bar{X}}(x) = n f_{X_1+\cdots+X_n}(nx),$$

even if the mgf of X does not exist.

5.6 If X has pdf $f_X(x)$ and Y, independent of X, has pdf $f_Y(y)$, establish formulas, similar to (5.2.3), for the random variable Z in each of the following situations.

(a) $Z = X - Y$

(b) $Z = XY$

(c) $Z = X/Y$

5.7 In Example 5.2.10, a partial fraction decomposition is needed to derive the distribution of the sum of two independent Cauchy random variables. This exercise provides the details that are skipped in that example.

(a) Find the constants A, B, C, and D that satisfy

$$\frac{1}{1+(w/\sigma)^2}\,\frac{1}{1+((z-w)/\tau)^2} = \frac{Aw}{1+(w/\sigma)^2} + \frac{B}{1+(w/\sigma)^2} - \frac{Cw}{1+((z-w)/\tau)^2} - \frac{D}{1+((z-w)/\tau)^2},$$

where A, B, C, and D may depend on z but not on w.

(b) Using the facts that

$$\int \frac{t}{1+t^2}\,dt = \frac{1}{2}\log(1+t^2) + \textit{constant} \quad \text{and} \quad \int \frac{1}{1+t^2}\,dt = \arctan(t) + \textit{constant},$$

evaluate (5.2.4) and hence verify (5.2.5).

(Note that the integration in part (*b*) is quite delicate. Since the mean of a Cauchy does not exist, the integrals $\int_{-\infty}^{\infty} \frac{Aw}{1+(w/\sigma)^2}\,dw$ and $\int_{-\infty}^{\infty} \frac{Cw}{1+((z-w)/\tau)^2}\,dw$ do not exist. However, the integral of the difference *does exist*, which is all that is needed.)

5.8 Let $X_1, \ldots, X_n$ be a random sample, where $\bar{X}$ and S^2 are calculated in the usual way.

(a) Show that

$$S^2 = \frac{1}{2n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} (X_i - X_j)^2.$$

Assume now that the X_is have a finite fourth moment, and denote $\theta_1 = \mathrm{E}X_i, \theta_j = \mathrm{E}(X_i - \theta_1)^j, j = 2, 3, 4$.

(b) Show that $\mathrm{Var}\, S^2 = \frac{1}{n}(\theta_4 - \frac{n-3}{n-1}\theta_2^2)$.

(c) Find $\mathrm{Cov}(\bar{X}, S^2)$ in terms of $\theta_1, \ldots, \theta_4$. Under what conditions is $\mathrm{Cov}(\bar{X}, S^2) = 0$?

5.9 Establish the *Lagrange Identity*, that for any numbers $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$,

$$\left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right) - \left(\sum_{i=1}^{n} a_i b_i\right)^2 = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (a_i b_j - a_j b_i)^2.$$

Use the identity to show that the correlation coefficient is equal to 1 if and only if all of the sample points lie on a straight line (Wright 1992). (*Hint*: Establish the identity for $n = 2$; then induct.)

5.10 Let $X_1, \ldots, X_n$ be a random sample from a $\mathrm{n}(\mu, \sigma^2)$ population.

(a) Find expressions for $\theta_1, \ldots, \theta_4$, as defined in Exercise 5.8, in terms of μ and σ^2.

(b) Use the results of Exercise 5.8, together with the results of part (a), to calculate $\mathrm{Var}\, S^2$.

(c) Calculate $\mathrm{Var}\, S^2$ a completely different (and easier) way: Use the fact that $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$.

5.11 Suppose $\bar{X}$ and S^2 are calculated from a random sample $X_1, \ldots, X_n$ drawn from a population with finite variance σ^2. We know that $\mathrm{E}S^2 = \sigma^2$. Prove that $\mathrm{E}S \leq \sigma$, and if $\sigma^2 > 0$, then $\mathrm{E}S < \sigma$.

5.12 Let $X_1, \ldots, X_n$ be a random sample from a $\mathrm{n}(0, 1)$ population. Define

$$Y_1 = \left|\frac{1}{n}\sum_{i=1}^{n} X_i\right|, \quad Y_2 = \frac{1}{n}\sum_{i=1}^{n} |X_i|.$$

Calculate $\mathrm{E}Y_1$ and $\mathrm{E}Y_2$, and establish an inequality between them.

5.13 Let $X_1, \ldots, X_n$ be iid $\mathrm{n}(\mu, \sigma^2)$. Find a function of S^2, the sample variance, say $g(S^2)$, that satisfies $\mathrm{E}g(S^2) = \sigma$. (*Hint*: Try $g(S^2) = c\sqrt{S^2}$, where c is a constant.)

5.14 (a) Prove that the statement of Lemma 5.3.3 follows from the special case of $\mu_i = 0$ and $\sigma_i^2 = 1$. That is, show that if $X_j = \sigma_j Z_j + \mu_j$ and $Z_j \sim \mathrm{n}(0, 1), j = 1, \ldots, n$, all independent, a_{ij}, b_{rj} are constants, and

$$\mathrm{Cov}\left(\sum_{j=1}^{n} a_{ij} Z_j, \sum_{j=1}^{n} b_{rj} Z_j\right) = 0 \Rightarrow \sum_{j=1}^{n} a_{ij} Z_j \text{ and } \sum_{j=1}^{n} b_{rj} Z_j \text{ are independent,}$$

then

$$\mathrm{Cov}\left(\sum_{j=1}^{n} a_{ij} X_j, \sum_{j=1}^{n} b_{rj} X_j\right) = 0 \Rightarrow \sum_{j=1}^{n} a_{ij} X_j \text{ and } \sum_{j=1}^{n} b_{rj} X_j \text{ are independent.}$$

(b) Verify the expression for $\mathrm{Cov}\left(\sum_{j=1}^{n} a_{ij} X_j, \sum_{j=1}^{n} b_{rj} X_j\right)$ in Lemma 5.3.3.

5.15 Establish the following recursion relations for means and variances. Let $\bar{X}_n$ and S_n^2 be the mean and variance, respectively, of $X_1, \ldots, X_n$. Then suppose another observation, X_{n+1}, becomes available. Show that

(a) $\bar{X}_{n+1} = \frac{X_{n+1} + n\bar{X}_n}{n+1}$.

(b) $nS_{n+1}^2 = (n-1)S_n^2 + \left(\frac{n}{n+1}\right)(X_{n+1} - \bar{X}_n)^2$.

5.16 Let $X_i, i = 1, 2, 3$, be independent with $\mathrm{n}(i, i^2)$ distributions. For each of the following situations, use the X_is to construct a statistic with the indicated distribution.

(a) chi squared with 3 degrees of freedom

(b) t distribution with 2 degrees of freedom

(c) F distribution with 1 and 2 degrees of freedom

5.17 Let X be a random variable with an $F_{p,q}$ distribution.

(a) Derive the pdf of X.

(b) Derive the mean and variance of X.

(c) Show that $1/X$ has an $F_{q,p}$ distribution.

(d) Show that $(p/q)X/[1 + (p/q)X]$ has a beta distribution with parameters $p/2$ and $q/2$.

5.18 Let X be a random variable with a Student's t distribution with p degrees of freedom.

(a) Derive the mean and variance of X.

(b) Show that X^2 has an F distribution with 1 and p degrees of freedom.

(c) Let $f(x|p)$ denote the pdf of X. Show that

$$\lim_{p\to\infty} f(x|p) \to \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

at each value of x, $-\infty < x < \infty$. This correctly suggests that as $p \to \infty, X$ converges in distribution to a $\mathrm{n}(0, 1)$ random variable. (*Hint*: Use Stirling's Formula.)

(d) Use the results of parts (a) and (b) to argue that, as $p \to \infty, X^2$ converges in distribution to a χ_1^2 random variable.

(e) What might you conjecture about the distributional limit, as $p \to \infty$, of $qF_{q,p}$?

5.19 (a) Prove that the χ^2 distribution is *stochastically increasing* in its degrees of freedom; that is, if $p > q$, then for any a, $P(\chi_p^2 > a) \geq P(\chi_q^2 > a)$, with strict inequality for some a.

(b) Use the results of part (a) to prove that for any ν, $kF_{k,\nu}$ is stochastically increasing in k.

(c) Show that for any k, ν, and α, $kF_{\alpha,k,\nu} > (k-1)F_{\alpha,k-1,\nu}$. (The notation $F_{\alpha,k-1,\nu}$ denotes a level-α *cutoff point*; see Section 8.3.1. Also see Miscellanea 8.5.1 and Exercise 11.15.)

5.20 (a) We can see that the t distribution is a mixture of normals using the following argument:

$$P(T_\nu \leq t) = P\left(\frac{Z}{\sqrt{\chi_\nu^2/\nu}} \leq t\right) = \int_0^\infty P\left(Z \leq t\sqrt{x}/\sqrt{\nu}\right) P\left(\chi_\nu^2 = x\right) dx,$$

where T_ν is a t random variable with ν degrees of freedom. Using the Fundamental Theorem of Calculus and interpreting $P(\chi^2_\nu = \nu x)$ as a pdf, we obtain

$$f_{T_\nu}(t) = \int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2 x/2\nu} \frac{\sqrt{x}}{\sqrt{\nu}} \frac{1}{\Gamma(\nu/2)2^{\nu/2}} (x)^{(\nu/2)-1} e^{-x/2}\, dx,$$

a scale mixture of normals. Verify this formula by direct integration.

(b) A similar formula holds for the F distribution; that is, it can be written as a mixture of chi squareds. If $F_{1,\nu}$ is an F random variable with 1 and ν degrees of freedom, then we can write

$$P(F_{1,\nu} \le \nu t) = \int_0^\infty P\left(\chi^2_1 \le ty\right) f_\nu(y)\, dy,$$

where $f_\nu(y)$ is a χ^2_ν pdf. Use the Fundamental Theorem of Calculus to obtain an integral expression for the pdf of $F_{1,\nu}$, and show that the integral equals the pdf.

(c) Verify that the generalization of part (b),

$$P\left(F_{m,\nu} \le \frac{\nu}{m} t\right) = \int_0^\infty P\left(\chi^2_m \le ty\right) f_\nu(y)\, dy,$$

is valid for all integers $m > 1$.

5.21 What is the probability that the larger of two continuous iid random variables will exceed the population median? Generalize this result to samples of size n.

5.22 Let X and Y be iid n(0, 1) random variables, and define $Z = \min(X, Y)$. Prove that $Z^2 \sim \chi^2_1$.

5.23 Let $U_i, i = 1, 2, \ldots$, be independent uniform(0, 1) random variables, and let X have distribution

$$P(X = x) = \frac{c}{x!}, \quad x = 1, 2, 3, \ldots,$$

where $c = 1/(e - 1)$. Find the distribution of

$$Z = \min\{U_1, \ldots, U_X\}.$$

(*Hint*: Note that the distribution of $Z|X = x$ is that of the first-order statistic from a sample of size x.)

5.24 Let $X_1, \ldots, X_n$ be a random sample from a population with pdf

$$f_X(x) = \begin{cases} 1/\theta & \text{if } 0 < x < \theta \\ 0 & \text{otherwise.} \end{cases}$$

Let $X_{(1)} < \cdots < X_{(n)}$ be the order statistics. Show that $X_{(1)}/X_{(n)}$ and $X_{(n)}$ are independent random variables.

5.25 As a generalization of the previous exercise, let $X_1, \ldots, X_n$ be iid with pdf

$$f_X(x) = \begin{cases} \frac{a}{\theta^a} x^{a-1} & \text{if } 0 < x < \theta \\ 0 & \text{otherwise.} \end{cases}$$

Let $X_{(1)} < \cdots < X_{(n)}$ be the order statistics. Show that $X_{(1)}/X_{(2)}, X_{(2)}/X_{(3)}, \ldots, X_{(n-1)}/X_{(n)}$, and $X_{(n)}$ are mutually independent random variables. Find the distribution of each of them.

5.26 Complete the proof of Theorem 5.4.6.

(a) Let U be a random variable that counts the number of $X_1, \ldots, X_n$ less than or equal to u, and let V be a random variable that counts the number of $X_1, \ldots, X_n$ greater than u and less than or equal to v. Show that $(U, V, n - U - V)$ is a multinomial random vector with n trials and cell probabilities $(F_X(u), F_X(v) - F_X(u), 1 - F_X(v))$.

(b) Show that the joint cdf of $X_{(i)}$ and $X_{(j)}$ can be expressed as

$$\begin{aligned} F_{X_{(i)},X_{(j)}}(u,v) &= P(U \geq i, U+V \geq j) \\ &= \sum_{k=i}^{j-1} \sum_{m=j-k}^{n-k} P(U=k, V=m) + P(U \geq j) \\ &= \sum_{k=i}^{j-1} \sum_{m=j-k}^{n-k} \frac{n!}{k!m!(n-k-m)!} [F_X(u)]^k [F_X(v) - F_X(u)]^m \\ &\quad \times [1 - F_X(v)]^{n-k-m} + P(U \geq j). \end{aligned}$$

(c) Find the joint pdf by computing the mixed partial as indicated in (4.1.4). (The mixed partial of $P(U \geq j)$ is 0 since this term depends only on u, not v. For the other terms, there is much cancellation using relationships like (5.4.6).)

5.27 Let $X_1, \ldots, X_n$ be iid with pdf $f_X(x)$ and cdf $F_X(x)$, and let $X_{(1)} < \cdots < X_{(n)}$ be the order statistics.

(a) Find an expression for the conditional pdf of $X_{(i)}$ given $X_{(j)}$ in terms of f_X and F_X.

(b) Find the pdf of $V|R = r$, where V and R are defined in Example 5.4.7.

5.28 Let $X_1, \ldots, X_n$ be iid with pdf $f_X(x)$ and cdf $F_X(x)$, and let $X_{(i_1)} < \cdots < X_{(i_l)}$ and $X_{(j_1)} < \cdots < X_{(j_m)}$ be any two disjoint groups of order statistics. In terms of the pdf $f_X(x)$ and the cdf $F_X(x)$, find expressions for

(a) The marginal cdf and pdf of $X_{(i_1)}, \ldots, X_{(i_l)}$.

(b) The conditional cdf and pdf of $X_{(i_1)}, \ldots, X_{(i_l)}$ given $X_{(j_1)}, \ldots, X_{(j_m)}$.

5.29 A manufacturer of booklets packages them in boxes of 100. It is known that, on the average, the booklets weigh 1 ounce, with a standard deviation of .05 ounce. The manufacturer is interested in calculating

$$P(\text{100 booklets weigh more than 100.4 ounces}),$$

a number that would help detect whether too many booklets are being put in a box. Explain how you would calculate the (approximate?) value of this probability. Mention any relevant theorems or assumptions needed.

5.30 If $\bar{X}_1$ and $\bar{X}_2$ are the means of two independent samples of size n from a population with variance σ^2, find a value for n so that $P(|\bar{X}_1 - \bar{X}_2| < \sigma/5) \approx .99$. Justify your calculations.

5.31 Suppose $\bar{X}$ is the mean of 100 observations from a population with mean μ and variance $\sigma^2 = 9$. Find limits between which $\bar{X} - \mu$ will lie with probability at least .90. Use both Chebychev's Inequality and the Central Limit Theorem, and comment on each.

5.32 Let $X_1, X_2, \ldots$ be a sequence of random variables that converges in probability to a constant a. Assume that $P(X_i > 0) = 1$ for all i.

(a) Verify that the sequences defined by $Y_i = \sqrt{X_i}$ and $Y_i' = a/X_i$ converge in probability.

(b) Use the results in part (a) to prove the fact used in Example 5.5.18, that σ/S_n converges in probability to 1.

5.33 Let X_n be a sequence of random variables that converges in distribution to a random variable X. Let Y_n be a sequence of random variables with the property that for any finite number c,

$$\lim_{n\to\infty} P(Y_n > c) = 1.$$

Show that for any finite number c,

$$\lim_{n\to\infty} P(X_n + Y_n > c) = 1.$$

(This is the type of result used in the discussion of the power properties of the tests described in Section 10.3.2.)

5.34 Let $X_1, \ldots, X_n$ be a random sample from a population with mean μ and variance σ^2. Show that

$$\mathrm{E}\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} = 0 \quad \text{and} \quad \mathrm{Var}\,\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} = 1.$$

Thus, the normalization of $\bar{X}_n$ in the Central Limit Theorem gives random variables that have the same mean and variance as the limiting n(0, 1) distribution.

5.35 Stirling's Formula (derived in Exercise 1.28), which gives an approximation for factorials, can be easily derived using the CLT.

(a) Argue that, if $X_i \sim$ exponential(1), $i = 1, 2, \ldots$, all independent, then for every x,

$$P\left(\frac{\bar{X}_n - 1}{1/\sqrt{n}} \le x\right) \to P(Z \le x),$$

where Z is a standard normal random variable.

(b) Show that differentiating both sides of the approximation in part (a) suggests

$$\frac{\sqrt{n}}{\Gamma(n)}(x\sqrt{n} + n)^{n-1} e^{-(x\sqrt{n}+n)} \approx \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

and that $x = 0$ gives Stirling's Formula.

5.36 Given that $N = n$, the conditional distribution of Y is χ^2_{2n}. The unconditional distribution of N is Poisson(θ).

(a) Calculate EY and Var Y (unconditional moments).

(b) Show that, as $\theta \to \infty$, $(Y - \mathrm{E}Y)/\sqrt{\mathrm{Var}\,Y} \to$ n(0, 1) in distribution.

5.37 In Example 5.5.16, a normal approximation to the negative binomial distribution was given. Just as with the normal approximation to the binomial distribution given in Example 3.3.2, the approximation might be improved with a "continuity correction." For X_is defined as in Example 5.5.16, let $V_n = \sum_{i=1}^{n} X_i$. For $n = 10, p = .7$, and $r = 2$, calculate $P(V_n = v)$ for $v = 0, 1, \ldots, 10$ using each of the following three methods.

(a) exact calculations

(b) normal approximation as given in Example 5.5.16

(c) normal approximation with continuity correction

5.38 The following extensions of the inequalities established in Exercise 3.45 are useful in establishing a SLLN (see Miscellanea 5.8.4). Let $X_1, X_2, \ldots, X_n$ be iid with mgf $M_X(t)$, $-h < t < h$, and let $S_n = \sum_{i=1}^n X_i$ and $\bar{X}_n = S_n/n$.

(a) Show that $P(S_n > a) \leq e^{-at}[M_X(t)]^n$, for $0 < t < h$, and $P(S_n \leq a) \leq e^{-at}[M_X(t)]^n$, for $-h < t < 0$.

(b) Use the facts that $M_X(0) = 1$ and $M'_X(0) = \mathrm{E}(X)$ to show that, if $\mathrm{E}(X) < 0$, then there is a $0 < c < 1$ with $P(S_n > a) \leq c^n$. Establish a similar bound for $P(S_n \leq a)$.

(c) Define $Y_i = X_i - \mu - \varepsilon$ and use the above argument, with $a = 0$, to establish that $P(\bar{X}_n - \mu > \varepsilon) \leq c^n$.

(d) Now define $Y_i = -X_i + \mu - \varepsilon$, establish an equality similar to part (c), and combine the two to get

$$P(|\bar{X}_n - \mu| > \varepsilon) \leq 2c^n \text{ for some } 0 < c < 1.$$

5.39 This exercise, and the two following, will look at some of the mathematical details of convergence.

(a) Prove Theorem 5.5.4. (*Hint*: Since h is continuous, given $\varepsilon > 0$ we can find a δ such that $|h(x_n) - h(x)| < \varepsilon$ whenever $|x_n - x| < \delta$. Translate this into probability statements.)

(b) In Example 5.5.8, find a subsequence of the X_is that converges almost surely, that is, that converges pointwise.

5.40 Prove Theorem 5.5.12 for the case where X_n and X are continuous random variables.

(a) Given t and ε, show that $P(X \leq t - \varepsilon) \leq P(X_n \leq t) + P(|X_n - X| \geq \varepsilon)$. This gives a lower bound on $P(X_n \leq t)$.

(b) Use a similar strategy to get an upper bound on $P(X_n \leq t)$.

(c) By pinching, deduce that $P(X_n \leq t) \to P(X \leq t)$.

5.41 Prove Theorem 5.5.13; that is, show that

$$P(|X_n - \mu| > \varepsilon) \to 0 \text{ for every } \varepsilon \quad \Leftrightarrow \quad P(X_n \leq x) \to \begin{cases} 0 & \text{if } x < \mu \\ 1 & \text{if } x \geq \mu. \end{cases}$$

(a) Set $\varepsilon = |x - \mu|$ and show that if $x > \mu$, then $P(X_n \leq x) \geq P(|X_n - \mu| \leq \varepsilon)$, while if $x < \mu$, then $P(X_n \leq x) \leq P(|X_n - \mu| \geq \varepsilon)$. Deduce the $\Rightarrow$ implication.

(b) Use the fact that $\{x : |x - \mu| > \varepsilon\} = \{x : x - \mu < -\varepsilon\} \cup \{x : x - \mu > \varepsilon\}$ to deduce the $\Leftarrow$ implication.

(See Billingsley 1995, Section 25, for a detailed treatment of the above results.)

5.42 Similar to Example 5.5.11, let $X_1, X_2, \ldots$ be iid and $X_{(n)} = \max_{1 \leq i \leq n} X_i$.

(a) If X_i beta$(1, \beta)$, find a value of ν so that $n^\nu(1 - X_{(n)})$ converges in distribution.

(b) If X_i exponential(1), find a sequence a_n so that $X_{(n)} - a_n$ converges in distribution.

5.43 Fill in the details in the proof of Theorem 5.5.24.

(a) Show that if $\sqrt{n}(Y_n - \mu) \to \text{n}(0, \sigma^2)$ in distribution, then $Y_n \to \mu$ in probability.

(b) Give the details for the application of Slutsky's Theorem (Theorem 5.5.17).

5.44 Let $X_i, i = 1, 2, \ldots$, be independent Bernoulli(p) random variables and let $Y_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

(a) Show that $\sqrt{n}\,(Y_n - p) \to n\,[0, p(1-p)]$ in distribution.

(b) Show that for $p \neq 1/2$, the estimate of variance $Y_n(1-Y_n)$ satisfies $\sqrt{n}[Y_n(1-Y_n) - p(1-p)] \to n\left[0, (1-2p)^2 p(1-p)\right]$ in distribution.

(c) Show that for $p = 1/2$, $n\left[Y_n(1-Y_n) - \frac{1}{4}\right] \to -\frac{1}{4}\chi_1^2$ in distribution. (If this appears strange, note that $Y_n(1-Y_n) \leq 1/4$, so the left-hand side is always negative. An equivalent form is $2n\left[\frac{1}{4} - Y_n(1-Y_n)\right] \to \chi_1^2$.)

5.45 For the situation of Example 5.6.1, calculate the probability that at least 75% of the components last 150 hours when

(a) $c = 300, X \sim \text{gamma}(a, b), a = 4, b = 5$.

(b) $c = 100, X \sim \text{gamma}(a, b), a = 20, b = 5$.

(c) $c = 100, X \sim \text{gamma}(a, b), a = 20.7, b = 5$.

(Hint: In parts (a) and (b) it is possible to evaluate the gamma integral in closed form, although it probably isn't worth the effort in (b). There is no closed-form expression for the integral in part (c), which has to be evaluated through either numerical integration or simulation.)

5.46 Referring to Exercise 5.45, compare your answers to what is obtained from a normal approximation to the binomial (see Example 3.3.2).

5.47 Verify the distributions of the random variables in (5.6.5).

5.48 Using strategies similar to (5.6.5), show how to generate an $F_{m,n}$ random variable, where both m and n are even integers.

5.49 Let $U \sim \text{uniform}(0, 1)$.

(a) Show that both $-\log U$ and $-\log(1-U)$ are exponential random variables.

(b) Show that $X = log\frac{u}{1-u}$ is a logistic(0, 1) random variable.

(c) Show how to generate a logistic(μ, β) random variable.

5.50 The Box-Muller method for generating normal *pseudo-random variables* (Example 5.6.4) is based on the transformation

$$X_1 = \cos(2\pi U_1)\sqrt{-2\log U_2}, \qquad X_2 = \sin(2\pi U_1)\sqrt{-2\log U_2},$$

where U_1 and U_2 are iid uniform(0,1). Prove that X_1 and X_2 are independent n(0, 1) random variables.

5.51 One of the earlier methods (not one of the better ones) of generating pseudo-random standard normal random variables from uniform random variables is to take $X = \sum_{i=1}^{12} U_i - 6$, where the U_is are iid uniform(0, 1).

(a) Justify the fact that X is approximately n(0, 1).

(b) Can you think of any obvious way in which the approximation fails?

(c) Show how good (or bad) the approximation is by comparing the first four moments. (The fourth moment is 29/10 and is a lengthy calculation—mgfs and computer algebra would help; see Example A.0.6 in Appendix A.)

5.52 For each of the following distributions write down an algorithm for generating the indicated random variables.

(a) $Y \sim$ binomial$(8, \frac{2}{3})$
(b) $Y \sim$ hypergeometric $N = 10, M = 8, K = 4$
(c) $Y \sim$ negative binomial$(5, \frac{1}{3})$

5.53 For each of the distributions in the previous exercise:

(a) Generate 1,000 variables from the indicated distribution.
(b) Compare the mean, variance, and histogram of the generated random variables with the theoretical values.

5.54 Refer to Example 5.6.6. Another sample of bay anchovy larvae counts yielded the data

$$158, 143, 106, 57, 97, 80, 109, 109, 350, 224, 109, 214, 84.$$

(a) Use the technique of Example 5.6.6 to construct a simulated distribution of S^2 to see if the assumption of Poisson counts is tenable.
(b) A possible explanation of the failure of the Poisson assumptions (and increased variance) is the failure of the assumption that the larvae are uniformly distributed in the river. If the larvae tend to clump, the negative binomial(r, p) distribution (with mean $\mu = r\frac{1-p}{p}$ and variance $\mu + \frac{\mu^2}{r}$) is a reasonable alternative model. For $\mu = \bar{x}$, what values of r lead to simulated distributions that are consistent with the data?

5.55 Suppose the method of (5.6.7) is used to generate the random variable Y, where $y_i = i$, $i = 0, 1, 2, \ldots$. Show that the expected number of comparisons is $E(Y + 1)$. (*Hint:* See Exercise 2.14.)

5.56 Let Y have the Cauchy distribution, $f_Y(y) = \frac{1}{1+y^2}, -\infty < y < \infty$.

(a) Show that $F_Y(y) = \tan^{-1}(y)$.
(b) Show how to simulate a Cauchy(a, b) random variable starting from a uniform$(0, 1)$ random variable.

(See Exercise 2.12 for a related result.)

5.57 Park *et al.* (1996) describe a method for generating correlated binary variables based on the follow scheme. Let X_1, X_2, X_3 be independent Poisson random variables with mean $\lambda_1, \lambda_2, \lambda_3$, respectively, and create the random variables

$$Y_1 = X_1 + X_3 \quad \text{and} \quad Y_2 = X_2 + X_3.$$

(a) Show that $\text{Cov}(Y_1, Y_2) = \lambda_3$.
(b) Define $Z_i = I(Y_i = 0)$ and $p_i = e^{-(\lambda_i + \lambda_3)}$. Show that Z_i are Bernoulli(p_i) with

$$\text{Corr}(Z_1, Z_2) = \frac{p_1 p_2 (e^{\lambda_3} - 1)}{\sqrt{p_1(1-p_1)}\sqrt{p_2(1-p_2)}}.$$

(c) Show that the correlation of Z_1 and Z_2 is not unrestricted in the range $[-1, 1]$, but

$$\text{Corr}(Z_1, Z_2) \leq \min\left\{\sqrt{\frac{p_2(1-p_1)}{p_1(1-p_2)}}, \sqrt{\frac{p_1(1-p_2)}{p_2(1-p_1)}}\right\}.$$

5.58 Suppose that $U_1, U_2, \ldots, U_n$ are iid uniform(0, 1) random variables, and let $S_n = \sum_{i=1}^{n} U_i$. Define the random variable N by

$$N = \min\{k : S_k > 1\}.$$

(a) Show that $P(S_k \leq t) = t^k/k!$.

(b) Show that $P(N = n) = P(S_{n-1} < 1) - P(S_n < 1)$ and, surprisingly, that $\mathrm{E}(N) = e$, the base of the natural logarithms.

(c) Use the result of part (b) to calculate the value of e by simulation.

(d) How large should n be so that you are 95% confident that you have the first four digits of e correct?

(Russell 1991, who attributes the problem to Gnedenko 1978, describes such a simulation experiment.)

5.59 Prove that the algorithm of Example 5.6.7 generates a beta(a, b) random variable.

5.60 Generalize the algorithm of Example 5.6.7 to apply to any bounded pdf; that is, for an arbitrary bounded pdf $f(x)$ on $[a, b]$, define $c = \max_{a \leq x \leq b} f(x)$. Let X and Y be independent, with $X \sim$ uniform(a, b) and $Y \sim$ uniform$(0, c)$. Let d be a number greater than b, and define a new random variable

$$W = \begin{cases} X & \text{if } Y < f(X) \\ d & \text{if } Y \geq f(X). \end{cases}$$

(a) Show that $P(W \leq w) = \int_a^w f(t)dt/[c(b-a)]$ for $a \leq w \leq b$.

(b) Using part (a), explain how a random variable with pdf $f(x)$ can be generated. (*Hint:* Use a geometric argument; a picture will help.)

5.61 (a) Suppose it is desired to generate $Y \sim$ beta(a, b), where a and b are not integers. Show that using $V \sim$ beta$([a], [b])$ will result in a finite value of $M = \sup_y f_Y(y)/f_V(y)$.

(b) Suppose it is desired to generate $Y \sim$ gamma(a, b), where a and b are not integers. Show that using $V \sim$ gamma$([a], b)$ will result in a finite value of $M = \sup_y f_Y(y)/f_V(y)$.

(c) Show that, in each of parts (a) and (b), if V had parameter $[a] + 1$, then M would be infinite.

(d) In each of parts (a) and (b) find optimal values for the parameters of V in the sense of minimizing $E(N)$ (see (5.6.12)).

(Recall that $[a] =$ greatest integer $\leq a$.)

5.62 Find the values of M so that an Accept/Reject Algorithm can generate $Y \sim$ n$(0, 1)$ using $U \sim$ uniform$(0, 1)$ and

(a) $V \sim$ Cauchy.

(b) $V \sim$ double exponential.

(c) Compare the algorithms. Which one do you recommend?

5.63 For generating $Y \sim$ n$(0, 1)$ using an Accept/Reject Algorithm, we could generate $U \sim$ uniform, $V \sim$ exponential(λ) , and attach a random sign to V ($\pm$ each with equal probability). What value of λ will optimize this algorithm?

5.64 A technique similar to Accept/Reject is *importance sampling*, which is quite useful for calculating features of a distribution. Suppose that $X \sim f$, but the pdf f is difficult

to simulate from. Generate $Y_1, Y_2, \ldots, Y_m$, iid from g, and, for any function h, calculate $\frac{1}{m}\sum_{i=1}^{m}\frac{f(Y_i)}{g(Y_i)}h(Y_i)$. We assume that the supports of f and g are the same and $\text{Var}\, h(X) < \infty$.

(a) Show that $\text{E}\left(\frac{1}{m}\sum_{i=1}^{m}\frac{f(Y_i)}{g(Y_i)}h(Y_i)\right) = \text{E}h(X)$.

(b) Show that $\frac{1}{m}\sum_{i=1}^{m}\frac{f(Y_i)}{g(Y_i)}h(Y_i) \to \text{E}h(X)$ in probability.

(c) Although the estimator of part (a) has the correct expectation, in practice the estimator

$$\sum_{i=1}^{m}\left(\frac{f(Y_i)/g(Y_i)}{\sum_{j=1}^{m}f(Y_j)/g(Y_j)}\right)h(Y_i)$$

is preferred. Show that this estimator converges in probability to $\text{E}h(X)$. Moreover, show that if h is constant, this estimator is superior to the one in part (a). (Casella and Robert 1996 further explore the properties of this estimator.)

5.65 A variation of the importance sampling algorithm of Exercise 5.64 can actually produce an approximate sample from f. Again let $X \sim f$ and generate $Y_1, Y_2, \ldots, Y_m$, iid from g. Calculate $q_i = [f(Y_i)/g(Y_i)]/[\sum_{j=1}^{m} f(Y_j)/g(Y_j)]$. Then generate random variables X^* from the discrete distribution on $Y_1, Y_2, \ldots, Y_m$, where $P(X^* = Y_k) = q_k$. Show that $X_1^*, X_2^*, \ldots, X_r^*$ is approximately a random sample from f.
(*Hint*: Show that $P(X^* \le x) = \sum_{i=1}^{m} q_i I(Y_i \le x)$, let $m \to \infty$, and use the WLLN in the numerator and denominator.)
This algorithm is called the *Sampling/Importance Resampling (SIR)* algorithm by Rubin (1988) and is referred to as the *weighted bootstrap* by Smith and Gelfand (1992).

5.66 If $X_1, \ldots, X_n$ are iid $\text{n}(\mu, \sigma^2)$, we have seen that the distribution of the sample mean $\bar{X}$ is $\text{n}(\mu, \sigma^2/n)$. If we are interested in using a more robust estimator of location, such as the median (5.4.1), it becomes a more difficult task to derive its distribution.

(a) Show that M is the median of the X_is if and only if $(M - \mu)/\sigma$ is the median of $(X_i - \mu)/\sigma$. Thus, we only need consider the distribution of the median from a $\text{n}(0, 1)$ sample.

(b) For a sample of size $n = 15$ from a $\text{n}(0, 1)$, simulate the distribution of the median M.

(c) Compare the distribution in part (b) to the asymptotic distribution of the median $\sqrt{n}(M - \mu) \sim \text{n}[0, 1/4f^2(0)]$, where f is the pdf. Is $n = 15$ large enough for the asymptotics to be valid?

5.67 In many instances the Metropolis Algorithm is the algorithm of choice because either (i) there are no obvious candidate densities that satisfy the Accept/Reject supremum condition, or (ii) the supremum condition is difficult to verify, or (iii) laziness leads us to substitute computing power for brain power.
For each of the following situations show how to implement the Metropolis Algorithm to generate a sample of size 100 from the specified distribution.

(a) $X \sim \frac{1}{\sigma}f[(x - \mu)/\sigma]$, f = Student's t with ν degrees of freedom, ν, μ, and σ known

(b) $X \sim \text{lognormal}(\mu, \sigma^2)$, μ, σ^2 known

(c) $X \sim \text{Weibull}(\alpha, \beta)$, α, β known

5.68 If we use the Metropolis Algorithm rather than the Accept/Reject Algorithm we are freed from verifying the supremum condition of the Accept/Reject Algorithm. Of

course, we give up the property of getting random variables with exactly the distribution we want and must settle for an approximation.

(a) Show how to use the Metropolis Algorithm to generate a random variable with an approximate Student's t distribution with ν degrees of freedom, starting from $n(0,1)$ random variables.

(b) Show how to use the Accept/Reject Algorithm to generate a random variable with a Student's t distribution with ν degrees of freedom, starting from Cauchy random variables.

(c) Show how to use transformations to generate directly a random variable with a Student's t distribution with ν degrees of freedom.

(d) For $\nu = 2, 10, 25$ compare the methods by generating a sample of size 100. Which method do you prefer? Why?

(Mengersen and Tweedie 1996 show that the convergence of the Metropolis Algorithm is much faster if the supremum condition is satisfied, that is, if $\sup f/g \leq M < \infty$, where f is the target density and g is the candidate density.)

5.69 Show that the pdf $f_Y(y)$ is a *stable point* of the Metropolis Algorithm. That is, if $Z_i \sim f_Y(y)$, then $Z_{i+1} \sim f_Y(y)$.

5.8 Miscellanea

5.8.1 More on the Central Limit Theorem

For the case of a sequence of iid random variables, necessary and sufficient conditions for convergence to normality are known, with probably the most famous result due to Lindeberg and Feller. The following special case is due to Lévy. Let $X_1, X_2, \ldots$ be an iid sequence with $EX_i = \mu < \infty$, and let $V_n = \sum_{i=1}^n X_i$. The sequence V_n will converge to a $n(0,1)$ random variable (when suitably normalized) if and only if

$$\lim_{t\to\infty} \frac{t^2 P(|X_1 - \mu| > t)}{E\left((X_1 - \mu)^2 I_{[-t,t]}(X_1 - \mu)\right)} = 0.$$

Note that the condition is a variance condition. While it does not quite require that the variances be finite, it does require that they be "almost" finite. This is an important point in the convergence to normality—normality comes from summing up small disturbances.

Other types of central limit theorems abound—in particular, ones aimed at relaxing the independence assumption. While this assumption cannot be done away with, it can be made less restrictive (see Billingsley 1995, Section 27, or Resnick 1999, Chapter 8).

5.8.2 The Bias of S^2

Most of the calculations that we have done in this chapter have assumed that the observations are independent, and calculations of some expectations have relied on this fact. David (1985) pointed out that, if the observations are dependent, then S^2 may be a biased estimate of σ^2. That is, it may not happen that $ES^2 = \sigma^2$. However, the range of the possible bias is easily calculated. If $X_1, \ldots, X_n$ are

random variables (not necessarily independent) with mean μ and variance σ^2, then

$$(n-1)\mathrm{E}S^2 = \mathrm{E}\left(\sum_{i=1}^{n}(X_i-\mu)^2 - n(\bar{X}-\mu)^2\right) = n\sigma^2 - n\,\mathrm{Var}\,\bar{X}.$$

Var $\bar{X}$ can vary, according to the amount and type of dependence, from 0 (if all of the variables are constant) to σ^2 (if all of the variables are copies of X_1). Substituting these values in the above equation, we get the range of $\mathrm{E}S^2$ under dependence as

$$0 \le \mathrm{E}S^2 \le \frac{n}{n-1}\sigma^2.$$

5.8.3 Chebychev's Inequality Revisited

In Section 3.6 we looked at Chebychev's Inequality (see also Miscellanea 3.8.2), and in Example 3.6.2 we saw a particularly useful form. That form still requires knowledge of the mean and variance of a random variable, and in some cases we might be interested in bounds using estimated values for the mean and variance.

If $X_1, \ldots, X_n$ is a random sample from a population with mean μ and variance σ^2, Chebychev's Inequality says

$$P(|X-\mu| \ge k\sigma) \le \frac{1}{k^2}.$$

Saw *et al.* (1984) showed that if we substitute $\bar{X}$ for μ and S^2 for σ^2, we obtain

$$P\big(|X-\bar{X}| \ge kS\big) \le \frac{1}{n+1}g\left(\frac{n(n+1)k^2}{n-1+(n+1)k^2}\right),$$

where

$$g(t) = \begin{cases} \nu & \text{if } \nu \text{ is even} \\ \nu & \text{if } \nu \text{ is odd and } t < a \\ \nu - 1 & \text{if } \nu \text{ is odd and } t > a \end{cases}$$

and

$$\nu = \text{largest integer} < \frac{n+1}{t}, \quad a = \frac{(n+1)(n+1-\nu)}{1+\nu(n+1-\nu)}.$$

5.8.4 More on the Strong Law

As mentioned, the Strong Law of Large Numbers, Theorem 5.5.9, can be proved under the less restrictive condition that the random variables have only a finite mean (see, for example, Resnick 1999, Chapter 7, or Billingsley 1995, Section 22). However, under the assumptions of the existence of an mgf, Koopmans (1993) has presented a proof that uses only calculus.

The type of convergence asserted in the SLLN is the convergence that we are most familiar with; it is *pointwise* convergence of the sequence of random variables $\bar{X}_n$

to their common mean μ. As we saw in Example 5.5.8, this is stronger form of convergence than convergence in probability, the convergence of the weak law.

The conclusion of the SLLN is that

$$P(\lim_{n\to\infty} |\bar{X}_n - \mu| < \epsilon) = 1;$$

that is, with probability 1, the limit of the sequence of $\{\bar{X}_n\}$ is μ. Alternatively, the set where the sequence diverges has probability 0. For the sequence to diverge, there must exist $\delta > 0$ such that for every n there is a $k > n$ with $|\bar{X}_k - \mu| > \delta$. The set of all $\bar{X}_k$ that satisfy this is a divergent sequence and is represented by the set

$$A_\delta = \cap_{n=1}^{\infty} \cup_{k=n}^{\infty} \{|\bar{X}_k - \mu| > \delta\}.$$

We can get an upper bound on $P(A_\delta)$ by dropping the intersection term, and then the probability of the set where the sequence $\{\bar{X}_n\}$ diverges is bounded above by

$$\begin{aligned} P(A_\delta) &\le P(\cup_{k=n}^{\infty}\{|\bar{X}_k - \mu| > \delta\}) \\ &\le \sum_{k=n}^{\infty} P(\{|\bar{X}_k - \mu| > \delta\}) \quad \text{(Boole's Inequality, Theorem 1.2.11)} \\ &\le 2\sum_{k=n}^{\infty} c^k, \quad 0 < c < 1, \end{aligned}$$

where Exercise 5.38(d) can be used to establish the last inequality. We then note that we are summing the geometric series, and it follows from (1.5.4) that

$$P(A_\delta) \le 2\sum_{k=n}^{\infty} c^k = 2\frac{c^n}{1-c} \to 0 \text{ as } n \to \infty,$$

and, hence, the set where the sequence $\{\bar{X}_n\}$ diverges has probability 0 and the SLLN is established.

5.8.5 Markov Chain Monte Carlo

Methods that are collectively known as Markov Chain Monte Carlo (MCMC) methods are used in the generation of random variables and have proved extremely useful for doing complicated calculations, most notably, calculations involving integrations and maximizations. The Metropolis Algorithm (see Section 5.6) is an example of an MCMC method.

As the name suggests, these methods are based on Markov chains, a probabilistic structure that we haven't explored (see Chung 1974 or Ross 1988 for an introduction). The sequence of random variables $X_1, X_2, \ldots$ is a *Markov chain* if

$$P(X_{k+1} \in A | X_1, \ldots, X_k) = P(X_{k+1} \in A | X_k);$$

that is, the distribution of the present random variable depends, at most, on the immediate past random variable. Note that this is a generalization of independence.

The *Ergodic Theorem*, which is a generalization of the Law of Large Numbers, says that if the Markov chain $X_1, X_2, \ldots$ satisfies some regularity conditions (which are often satisfied in statistical problems), then

$$\frac{1}{n}\sum_{i=1}^{n} h(X_i) \to \mathrm{E}h(X) \text{ as } n \to \infty,$$

provided the expectation exists. Thus, the calculations of Section 5.6 can be extended to Markov chains and MCMC methods.

To fully understand MCMC methods it is really necessary to understand more about Markov chains, which we will not do here. There is already a vast literature on MCMC methods, encompassing both theory and applications. Tanner (1996) provides a good introduction to computational methods in statistics, as does Robert (1994, Chapter 9), who provides a more theoretical treatment with a Bayesian flavor. An easier introduction to this topic via the Gibbs sampler (a particular MCMC method) is given by Casella and George (1992). The Gibbs sampler is, perhaps, the MCMC method that is still the most widely used and is responsible for the popularity of this method (due to the seminal work of Gelfand and Smith 1990 expanding on Geman and Geman 1984). The list of references involving MCMC methods is prohibitively long. Some other introductions to this literature are through the papers of Gelman and Rubin (1992), Geyer and Thompson (1992), and Smith and Roberts (1993), with a particularly elegant theoretical introduction given by Tierney (1994). Robert and Casella (1999) is a textbook-length treatment of this field.