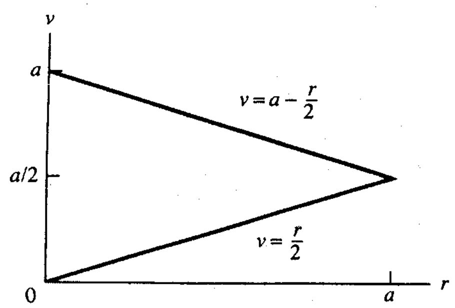

Figure 5.4.1. *Region on which $f_{R,V}(r,v) > 0$ for Example 5.4.7*

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 $\epsilon > 0$,

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

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 $\varepsilon > 0$,

$$P(|X_{(n)} - 1| \geq \varepsilon) = P(X_{(n)} \geq 1 + \varepsilon) + P(X_{(n)} \leq 1 - \varepsilon)$$
$$= 0 + P(X_{(n)} \leq 1 - \varepsilon).$$

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

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

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 $EX_i = \mu$ and $\text{Var } X_i = \sigma^2 > 0$. (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),

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

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

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

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 $\lim_{n\to\infty} \operatorname{Var} S_n^2 = 0$, 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

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

||

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

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

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

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

(5.5.9)
$$= \sum_{i=1}^k [g_i'(\boldsymbol{\theta})]^2 \text{Var}_\theta\, T_i + 2\sum_{i>j} g_i'(\boldsymbol{\theta}) g_j'(\boldsymbol{\theta}) \text{Cov}_\theta(T_i, T_j),$$

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

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

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 $\text{E}_\mu X = \mu \neq 0$. 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

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

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

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

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 $\sqrt{n}(Y_n - \theta) \to \text{n}(0, \sigma^2)$ in distribution. For a given function g and a specific value of θ, suppose that $g'(\theta)$ exists and is not 0. Then*

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

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

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

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

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

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

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

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

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

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 $\mathrm{n}(0,1)$ is a χ_1^2 (Example 2.1.9), which implies that

$$\frac{n(Y_n - \theta)^2}{\sigma^2} \to \chi_1^2$$

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 $\sqrt{n}(Y_n - \theta) \to \mathrm{n}(0, \sigma^2)$ 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*

$$n[g(Y_n) - g(\theta)] \to \sigma^2 \frac{g''(\theta)}{2}\chi_1^2 \text{ in distribution.} \tag{5.5.13}$$

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 $\tau^2 = \Sigma\Sigma\sigma_{ij}\frac{\partial g(\mu)}{\partial \mu_i} \cdot \frac{\partial g(\mu)}{\partial \mu_j} > 0$,*

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

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.