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 $h(x)$. The requirement that the sample space of $(T_1, \ldots, T_k)$ contain an open subset of $\Re^k$ usually is equivalent to the requirement that $n \geq k$. We will not prove this theorem but will only illustrate the result in a simple case.

**Example 5.2.3:** Suppose $X_1, \ldots, X_n$ is a random sample from a Bernoulli($p$) distribution. From Example 3.3.1 (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 3.1, we know that $T_1$ has a binomial($n, p$) distribution. From Example 3.3.1 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 (5.2.4) is verified for this example. ||

# 5.3 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 treat them in varying amounts of detail. (A full treatment of convergence is given in Feller (1968, 1971) and Chung (1974), for example.) In particular, we want to look at the behavior of $\bar{X}_n$, the mean of $n$ observations, as $n \to \infty$.

## 5.3.1 Convergence in Probability

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

DEFINITION 5.3.1: 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 5.3.1 (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 5.3.1 (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 $\overline{X}_n = (1/n)\sum_{i=1}^n X_i$. Then, for every $\epsilon > 0$,

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

that is, $\overline{X}_n$ converges in probability to $\mu$.

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

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

Hence, $P(|\overline{X}_n - \mu| < \epsilon) = 1 - P(|\overline{X}_n - \mu| \geq \epsilon) \geq 1 - \sigma^2/(n\epsilon^2) \to 1$, 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 5.3.1 is applicable in most practical situations. (See Exercise 5.13 for one way of weakening the hypotheses of the WLLN.)

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 5.3.1:** 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

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

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 $\sigma^2$ is that $\mathrm{Var}\, S_n^2 \to 0$ as $n \to \infty$. ||

## 5.3.2 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 5.3.2: 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 5.3.1 and 5.3.2. Although they look similar, they are very different statements with Definition 5.3.2 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 5.3.2 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 5.3.2 illustrates almost sure convergence. Example 5.3.3 illustrates the difference between convergence in probability and almost sure convergence.

**Example 5.3.2:** 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 5.3.3:** 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:

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

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 5.3.3 shows. However, if a sequence converges in probability, it is possible to find a *subsequence* that converges almost surely (see Chung (1974)) for theorems or Exercise 5.14 for an example).

Again, statisticians are often concerned with convergence to a constant. We now state, without proof, the stronger analogue of the WLLN, the Strong Law of Large Numbers (SLLN).

**THEOREM 5.3.2 (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 $\overline{X}_n = (1/n)\sum_{i=1}^n X_i$. Then, for every $\epsilon > 0$,

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

that is, $\overline{X}_n$ converges almost surely to $\mu$. □

## 5.3.3 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 2.3.4).

DEFINITION 5.3.3: 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.

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.

We again want to look at the large-sample behavior of the sample mean and, in particular, investigate its limiting distribution. We begin by proving one of the most startling theorems in statistics, the Central Limit Theorem (CLT).

**THEOREM 5.3.3 (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 $\mu$ and $\sigma^2$ are finite since the mgf exists.) Define $\overline{X}_n = (1/n)\sum_{i=1}^n X_i$. Let $G_n(x)$ denote the cdf of $\sqrt{n}(\overline{X}_n - \mu)/\sigma$. 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, $\sqrt{n}(\overline{X}_n - \mu)/\sigma$ 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 5.2.2, dealing with the Cauchy distribution, where there is no convergence to normality.)

While reveling 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 way of knowing how good this approximation is. 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* 5.3.3: 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 $Y_i$s, which exists for $|t| < \sigma h$ and is given by Theorem 2.3.5. Since

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i, \tag{5.3.1}$$

we have, from the properties of mgfs (see Theorems 2.3.5 and 4.6.3)

$$\begin{aligned} M_{\sqrt{n}(\bar{X}_n-\mu)/\sigma}(t) &= M_{\sum_{i=1}^n Y_i/\sqrt{n}}(t) \\ &= M_{\sum_{i=1}^n Y_i}\left(\frac{t}{\sqrt{n}}\right) && \text{(Theorem 2.3.5)} \\ &= \left(M_Y\left(\frac{t}{\sqrt{n}}\right)\right)^n. && \text{(Theorem 4.6.3)} \end{aligned} \tag{5.3.2}$$

We now expand $M_Y(t/\sqrt{n})$ in a Taylor series (power series) around 0. (See Definition 7.4.2.) We have

$$M_Y\left(\frac{t}{\sqrt{n}}\right) = \sum_{k=0}^{\infty} M_Y^{(k)}(0)\frac{(t/\sqrt{n})^k}{k!}, \tag{5.3.3}$$

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

$$M_Y\left(\frac{t}{\sqrt{n}}\right) = 1 + \frac{(t/\sqrt{n})^2}{2!} + R_Y\left(\frac{t}{\sqrt{n}}\right), \tag{5.3.4}$$

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 7.4.1) 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

$$\lim_{n\to\infty} \frac{R_Y(t/\sqrt{n})}{(1/\sqrt{n})^2} = \lim_{n\to\infty} nR_Y\left(\frac{t}{\sqrt{n}}\right) = 0, \tag{5.3.5}$$

and (5.3.5) is also true at $t = 0$ since $R_Y(0/\sqrt{n}) = 0$. Thus, for any fixed $t$, we can write

$$\begin{aligned} \lim_{n\to\infty} \left(M_Y\left(\frac{t}{\sqrt{n}}\right)\right)^n &= \lim_{n\to\infty}\left[1 + \frac{(t/\sqrt{n})^2}{2!} + R_Y\left(\frac{t}{\sqrt{n}}\right)\right]^n \\ &= \lim_{n\to\infty}\left[1 + \frac{1}{n}\left(\frac{t^2}{2} + nR_Y\left(\frac{t}{\sqrt{n}}\right)\right)\right]^n \\ &= e^{t^2/2}, \end{aligned} \tag{5.3.6}$$

by an application of Lemma 2.3.1, where we set $a_n = (t^2/2) + nR_Y(t/\sqrt{n})$. (Note that (5.3.5) 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 5.3.3 (see the Miscellanea section for a discussion). In particular, all of the assumptions about mgfs are not needed—the use of characteristic functions (Chapter 2 Miscellanea) 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 5.3.4 (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 5.3.3, 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 Chung (1974) or Feller (1971).

It is also possible to prove this theorem without recourse to characteristic functions, using only elementary arguments. By doing careful analysis and being clever with Taylor series expansions, it can be shown directly that probabilities involving $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ converge to normal probabilities (Brown, 1988). (Brown's proof is similar in spirit to, but much more involved than, the proof of the *Demoivre–Laplace Limit Theorem* given in Feller (1968). The Demoivre–Laplace Limit Theorem is a special case of the CLT, that binomials converge to normals as $n \to \infty$.)

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 5.3.4:** 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 $\mathrm{n}(0,1)$. 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

$$P(\bar{X} \le 11) = P\left(\frac{\sqrt{30}(\bar{X} - 10)}{\sqrt{20}} \le \frac{\sqrt{30}(11 - 10)}{\sqrt{20}}\right)$$

$$\approx P(Z \leq 1.2247)$$
$$= .8888. \qquad \|$$

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

**THEOREM 5.3.5 (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 5.3.5:** Suppose that

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

but the value of $\sigma$ is unknown. We have seen in Example 5.3.1 that, if $\lim_{n\to\infty} \mathrm{Var}\, S_n^2 = 0$, then $S_n^2 \to \sigma^2$ in probability. By Exercise 5.15, $\sigma/S_n \to 1$ in probability. Hence, Slutsky's Theorem tells us

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

## 5.4 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.

### 5.4.1 Properties of the Sample Mean and Variance

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

**THEOREM 5.4.1:** Let $X_1, \ldots, X_n$ be a random sample from a $\mathrm{n}(\mu, \sigma^2)$ distribution, and let $\overline{X} = (1/n)\sum_{i=1}^n X_i$ and $S^2 = [1/(n-1)]\sum_{i=1}^n (X_i - \overline{X})^2$. Then

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