Chapter 4

Central Limit Theorems

The main result of this chapter, in Section 4.2, is the Lindeberg-Feller Central Limit Theorem, from which we obtain the result most commonly known as "The Central Limit Theorem" as a corollary. As in Chapter 3, we mix univariate and multivariate results here. As a general summary, much of Section 4.1 is multivariate and most of the remainder of the chapter is univariate. The interplay between univariate and multivariate results is exemplified by the Central Limit Theorem itself, Theorem 4.9, which is stated for the multivariate case but whose proof is a simple combination of the analagous univariate result with Theorem 4.12, the Cramér-Wold theorem.

Before we discuss central limit theorems, we include one section of background material for the sake of completeness. Section 4.1 introduces the powerful Continuity Theorem, Theorem 4.3, which is the basis for proofs of various important results including the Lindeberg-Feller Theorem. This section also defines multivariate normal distributions.

4.1 Characteristic Functions and Normal Distributions

While it may seem odd to group two such different-sounding topics into the same section, there are actually many points of overlap between characteristic function theory and the multivariate normal distribution. Characteristic functions are essential for proving the Central Limit Theorems of this chapter, which are fundamentally statements about normal distributions. Furthermore, the simplest way to define normal distributions is by using their characteristic functions. The standard univariate method of defining a normal distribution by writing its density does not work here (at least not in a simple way), since not all normal distributions have densities in the usual sense. We even provide a proof of an important result—that characteristic functions determine their distributions uniquely—that uses normal distributions in an essential way. Thus, the study of characteristic functions and the study of normal distributions are so closely related in statistical large-sample theory that it is perfectly natural for us to introduce them together.

4.1.1 The Continuity Theorem

Definition 4.1 For a random vector $\mathbf{X}$, we define the characteristic function $\phi_{\mathbf{X}} : \mathbb{R}^k \to \mathbb{C}$ by

$$\phi_{\mathbf{X}}(\mathbf{t}) \;=\; \mathrm{E}\exp(\mathrm{i}\mathbf{t}^\top\mathbf{X}) \;=\; \mathrm{E}\cos(\mathbf{t}^\top\mathbf{X}) + \mathrm{i}\,\mathrm{E}\sin(\mathbf{t}^\top\mathbf{X}),$$

where $\mathrm{i}^2 = -1$ and $\mathbb{C}$ denotes the complex numbers.

The characteristic function, which is defined on all of $\mathbb{R}^k$ for *any* $\mathbf{X}$ (unlike the moment generating function, which requires finite moments), has some basic properties. For instance, $\phi_{\mathbf{X}}(\mathbf{t})$ is always a continuous function with $\phi_{\mathbf{X}}(\mathbf{0}) = 1$ and $|\phi_{\mathbf{X}}(\mathbf{t})| \le 1$. Also, inspection of Definition 4.1 reveals that for any constant vector $\mathbf{a}$ and scalar b,

$$\phi_{\mathbf{X}+\mathbf{a}}(\mathbf{t}) = \exp(\mathrm{i}\mathbf{t}^\top\mathbf{a})\phi_{\mathbf{X}}(\mathbf{t}) \;\text{ and }\; \phi_{b\mathbf{X}}(\mathbf{t}) = \phi_{\mathbf{X}}(b\mathbf{t}). \tag{4.1}$$

Also, if $\mathbf{X}$ and $\mathbf{Y}$ are independent,

$$\phi_{\mathbf{X}+\mathbf{Y}}(\mathbf{t}) = \phi_{\mathbf{X}}(\mathbf{t})\phi_{\mathbf{Y}}(\mathbf{t}). \tag{4.2}$$

One of the main reasons that characteristic functions are so useful is the fact that they uniquely determine the distributions from which they are derived. This fact is so important that we state it as a theorem:

Theorem 4.2 The random vectors $\mathbf{X}_1$ and $\mathbf{X}_2$ have the same distribution if and only if $\phi_{\mathbf{X_1}}(\mathbf{t}) = \phi_{\mathbf{X_2}}(\mathbf{t})$ for all $\mathbf{t}$.

Now suppose that $\mathbf{X}_n \stackrel{d}{\to} \mathbf{X}$, which implies $\mathbf{t}^\top\mathbf{X}_n \stackrel{d}{\to} \mathbf{t}^\top\mathbf{X}$. Since both $\sin x$ and $\cos x$ are bounded continuous functions, Theorem 2.28 implies that $\phi_{\mathbf{X}_n}(\mathbf{t}) \to \phi_{\mathbf{X}}(\mathbf{t})$. The converse, which is much harder to prove, is also true:

Theorem 4.3 *Continuity Theorem:* $\mathbf{X}_n \stackrel{d}{\to} \mathbf{X}$ if and only if $\phi_{\mathbf{X}_n}(\mathbf{t}) \to \phi_{\mathbf{X}}(\mathbf{t})$ for all $\mathbf{t}$.

Here is a partial proof that $\phi_{\mathbf{X}_n}(\mathbf{t}) \to \phi_{\mathbf{X}}(\mathbf{t})$ implies $\mathbf{X}_n \stackrel{d}{\to} \mathbf{X}$. First, we note that the distribution functions F_n must contain a convergent subsequence, say $F_{n_k} \to G$ as $k \to \infty$, where $G : \mathbb{R} \to [0, 1]$ must be a nondecreasing function but G is not necessarily a true distribution function (and, of course, convergence is guaranteed only at continuity points of G). It is possible to define the characteristic function of G—though we will not prove this assertion—and it must follow that $\phi_{F_{n_k}}(t) \to \phi_G(t)$. But this implies that $\phi_G(t) = \phi_{\mathbf{X}}(t)$ because it was assumed that $\phi_{\mathbf{X}_n}(t) \to \phi_{\mathbf{X}}(t)$. By Theorem 4.2, G must be the distribution function of $\mathbf{X}$. Therefore, every convergent subsequence of $\{\mathbf{X}_n\}$ converges to $\mathbf{X}$, which gives the result.

Theorem 4.3 is an extremely useful tool for proving facts about convergence in distribution. Foremost among these will be the Lindeberg-Feller Theorem in Section 4.2, but other results follow as well. For example, a quick proof of the Cramér-Wold Theorem, Theorem 4.12, is possible (see Exercise 4.3).

4.1.2 Moments

One of the facts that allows us to prove results about distributions using results about characteristic functions is the relationship between the moments of a distribution and the derivatives of a characteristic function. We emphasize here that *all* random variables have well-defined characteristic functions, even if they do not have any moments. What we will see is that existence of moments is related to differentiability of the characteristic function.

We derive $\partial\phi_{\mathbf{X}}(\mathbf{t})/\partial t_j$ directly by considering the limit, if it exists, of

$$\frac{\phi_{\mathbf{X}}(\mathbf{t}+h\mathbf{e}_j) - \phi_{\mathbf{X}}(\mathbf{t})}{h} = \mathrm{E}\left[\exp\{i\mathbf{t}^\top\mathbf{X}\}\left(\frac{\exp\{ihX_j\}-1}{h}\right)\right]$$

as $h \to 0$, where $\mathbf{e}_j$ denotes the jth unit vector with 1 in the jth component and 0 elsewhere. Note that

$$\left|\exp\{i\mathbf{t}^\top\mathbf{X}\}\left(\frac{\exp\{ihX_j\}-1}{h}\right)\right| = \left|\int_0^{X_j} \exp\{iht\}\,dt\right| \le |X_j|,$$

so if $\mathrm{E}\,|X_j| < \infty$ then the dominated convergence theorem, Theorem 3.22, implies that

$$\frac{\partial}{\partial t_j}\phi_{\mathbf{X}}(\mathbf{t}) = \mathrm{E}\lim_{h\to 0}\left[\exp\{i\mathbf{t}^\top\mathbf{X}\}\left(\frac{\exp\{ihX_j\}-1}{h}\right)\right] = i\,\mathrm{E}\left[X_j\exp\{i\mathbf{t}^\top\mathbf{X}\}\right].$$

We conclude that

Lemma 4.4 If $\mathrm{E}\,\|\mathbf{X}\| < \infty$, then $\nabla\phi_{\mathbf{X}}(\mathbf{0}) = i\,\mathrm{E}\,\mathbf{X}$.

A similar argument gives

Lemma 4.5 If $\mathrm{E}\,\mathbf{X}^\top\mathbf{X} < \infty$, then $\nabla^2\phi_{\mathbf{X}}(\mathbf{0}) = -\,\mathrm{E}\,\mathbf{X}\mathbf{X}^\top$.

It is possible to relate higher-order moments of X to higher-order derivatives of $\phi_{\mathbf{X}}(\mathbf{t})$ using the same logic, but for our purposes, only Lemmas 4.4 and 4.5 are needed.

4.1.3 The Multivariate Normal Distribution

It is easy to define a univariate normal distribution. If μ and σ^2 are the mean and variance, respectively, then if $\sigma^2 > 0$ the corresponding normal distribution is by definition the distribution whose density is the well-known function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\}.$$

If $\sigma^2 = 0$, on the other hand, we simply take the corresponding normal distribution to be the constant μ. However, it is not quite so easy to define a multivariate normal distribution. This is due to the fact that not all nonconstant multivariate normal distributions have densities on $\mathbb{R}^k$ in the usual sense. It turns out to be much simpler to define multivariate normal distributions using their characteristic functions:

Definition 4.6 Let Σ be any symmetric, nonnegative definite, $k \times k$ matrix and let $\boldsymbol{\mu}$ be any vector in $\mathbb{R}^k$. Then the normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ is defined to be the distribution with characteristic function

$$\phi_{\mathbf{X}}(\mathbf{t}) = \exp\left(i\mathbf{t}^\top \boldsymbol{\mu} - \frac{\mathbf{t}^\top \Sigma \mathbf{t}}{2} \right). \tag{4.3}$$

Definition 4.6 has a couple of small flaws. First, because it does not stipulate $k \neq 1$, it offers a definition of *univariate* normality that might compete with the already-established definition. However, Exercise 4.1(a) verifies that the two definitions coincide. Second, Definition 4.6 asserts without proof that equation (4.3) actually defines a legitimate characteristic function. How do we know that a distribution with this characteristic function really exists for all possible Σ and $\boldsymbol{\mu}$? There are at least two ways to mend this flaw. One way is to establish sufficient conditions for a particular function to be a legitimate characteristic function, then prove that the function in Equation (4.3) satisfies them. This is possible, but it would take us too far from the aim of this section, which is to establish just enough background to aid the study of statistical large-sample theory. Another method is to construct a random variable whose characteristic function coincides with equation (4.3); yet to do this requires that we delve into some linear algebra. Since this linear algebra will prove useful later, this is the approach we now take.

Before constructing a multivariate normal random vector in full generality, we first consider the case in which Σ is diagonal, say $\Sigma = D = \text{diag}(d_1, \ldots, d_k)$. The stipulation in Definition 4.6 that Σ be nonnegative definite means in this special case that $d_i \geq 0$ for all i. Now take $X_1, \ldots, X_k$ to be independent, univariate normal random variables with zero means and $\text{Var}\, X_i = d_i$. We assert without proof—the assertion will be proven later—that $\mathbf{X} = (X_1, \ldots, X_k)$ is then a multivariate normal random vector, according to Definition 4.6, with mean $\mathbf{0}$ and covariance matrix D.

To define a multivariate normal random vector with a general (non-diagonal) covariance matrix Σ, we make use of the fact that any symmetric matrix may be diagonalized by an orthogonal matrix. We first define orthogonal, then state the diagonalizability result as a lemma that will not be proven here.

Definition 4.7 A square matrix Q is orthogonal if Q^{-1} exists and is equal to $Q^\top$.

Lemma 4.8 If A is a symmetric $k \times k$ matrix, then there exists an orthogonal matrix Q such that $QAQ^\top$ is diagonal.

Note that the diagonal elements of the matrix $QAQ^\top$ in the matrix above must be the eigenvalues of A. This follows since if λ is a diagonal element of $QAQ^\top$, then it is an eigenvalue of $QAQ^\top$. Hence, there exists a vector x such that $QAQ^\top x = \lambda x$, which implies that $A(Q^\top x) = \lambda(Q^\top x)$ and so λ is an eigenvalue of A.

Taking Σ and $\boldsymbol{\mu}$ as in Definition 4.6, Lemma 4.8 implies that there exists an orgthogonal matrix Q such that $Q\Sigma Q^\top$ is diagonal. Since we know that every diagonal entry in $Q\Sigma Q^\top$ is nonnegative, we may define $\mathbf{Y} = (Y_1, \ldots, Y_k)$, where $Y_1, \ldots, Y_k$ are independent normal random vectors with mean zero and Var Y_i equal to the ith diagonal entry of $Q\Sigma Q^\top$. Then the random vector

$$\mathbf{X} = \boldsymbol{\mu} + Q^\top \mathbf{Y} \tag{4.4}$$

has the characteristic function in equation (4.3), a fact whose proof is the subject of Exercise 4.1. Thus, Equation (4.3) of Definition 4.6 always gives the characteristic function of an actual distribution. We denote this multivariate normal distribution by $N_k(\boldsymbol{\mu}, \Sigma)$, or simply $N(\mu, \sigma^2)$ if $k = 1$.

To conclude this section, we point out that in case Σ *is* invertible, then $N_k(\boldsymbol{\mu}, \Sigma)$ has a density in the usual sense on $\mathbb{R}^k$:

$$f(\mathbf{x}) = \frac{1}{\sqrt{2^k \pi^k |\Sigma|}} \exp\left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}, \tag{4.5}$$

where $|\Sigma|$ denotes the determinant of Σ. However, this density will be of little value in the large-sample topics to follow.

4.1.4 Asymptotic Normality

Now that $N_k(\boldsymbol{\mu}, \Sigma)$ is defined, we may use it to state one of the most useful theorems in all of statistical large-sample theory, the Central Limit Theorem for independent and identically distributed (iid) sequences of random vectors. We defer the proof of this theorem to the next section, where we establish a much more general result called the Lindeberg-Feller Central Limit Theorem.

Theorem 4.9 *Central Limit Theorem for independent and identically distributed multivariate sequences:* If $\mathbf{X}_1, \mathbf{X}_2, \ldots$ are independent and identically distributed with mean $\boldsymbol{\mu} \in R^k$ and covariance Σ, where Σ has finite entries, then

$$\sqrt{n}(\overline{\mathbf{X}}_n - \boldsymbol{\mu}) \stackrel{d}{\rightarrow} N_k(\mathbf{0}, \Sigma).$$

Although we refer to several different theorems in this chapter as central limit theorems of one sort or another, we also employ the standard statistical usage in which the phrase "The Central Limit Theorem," with no modifier, refers to Theorem 4.9 or its univariate analogue.

Before exhibiting some examples that apply Theorem 4.9, we discuss what is generally meant by the phrase "asymptotic distribution". Suppose we are given a sequence $X_1, X_2, \ldots$ of random variables and asked to determine the asymptotic distribution of this sequence. This might mean to find X such that $X_n \stackrel{d}{\rightarrow} X$. However, depending on the context, this might not be the case; for example, if $X_n \stackrel{d}{\rightarrow} c$ for a constant c, then we mean something else by "asymptotic distribution".

In general, the "asymptotic distribution of X_n" means a *nonconstant* random variable X, along with real-number sequences $\{a_n\}$ and $\{b_n\}$, such that $a_n(X_n - b_n) \stackrel{d}{\rightarrow} X$. In this case, the distribution of X might be referred to as the asymptotic or limiting distribution of either X_n or of $a_n(X_n - b_n)$, depending on the context.

Example 4.10 Suppose that X_n is the sum of n independent Bernoulli(p) random variables, so that $X_n \sim \text{binomial}(n, p)$. Even though we know that $X_n/n \stackrel{P}{\rightarrow} p$ by the weak law of large numbers, this is not generally what we mean by the asymptotic distribution of X_n/n. Instead, the asymptotic distribution of X_n/n is expressed by

$$\sqrt{n}\left(\frac{X_n}{n} - p\right) \stackrel{d}{\rightarrow} N\{0, p(1-p)\},$$

which follows from the Central Limit Theorem because a Bernoulli(p) random variable has mean p and variance $p(1-p)$.

Example 4.11 *Asymptotic distribution of sample variance:* Suppose that $X_1, X_2, \ldots$ are independent and identically distributed with $\mathrm{E}\,(X_i) = \mu$, $\mathrm{Var}\,(X_i) = \sigma^2$, and $\mathrm{Var}\,\{(X_i - \mu)^2\} = \tau^2 < \infty$. Define

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2. \qquad (4.6)$$

We wish to determine the asymptotic distribution of S_n^2.

Since the distribution of $X_i - \overline{X}_n$ does not change if we replace each X_i by $X_i - \mu$, we may assume without loss of generality that $\mu = 0$. By the Central Limit Theorem, we know that

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \sigma^2\right) \xrightarrow{d} N(0, \tau^2).$$

Furthermore, the Central Limit Theorem and the Weak Law imply $\sqrt{n}(\overline{X}_n) \xrightarrow{d} N(0, \sigma^2)$ and $\overline{X}_n \xrightarrow{P} 0$, respectively, so Slutsky's theorem implies $\sqrt{n}\left(\overline{X}_n^2\right) \xrightarrow{P} 0$. Therefore, since

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

Slutsky's theorem implies that $\sqrt{n}\,(S_n^2 - \sigma^2) \xrightarrow{d} N(0, \tau^2)$, which is the desired result.

Note that the definition of S_n^2 in Equation 4.6 is not the usual unbiased sample variance, which uses the denominator $n - 1$ instead of n. However, since

$$\sqrt{n}\left(\frac{n}{n-1}S_n^2 - \sigma^2\right) = \sqrt{n}(S_n^2 - \sigma^2) + \frac{\sqrt{n}}{n-1}S_n^2$$

and $\sqrt{n}/(n-1) \to 0$, we see that the simpler choice of n does not change the asymptotic distribution at all.

4.1.5 The Cramér-Wold Theorem

Suppose that $\mathbf{X}_1, \mathbf{X}_2, \ldots$ is a sequence of random k-vectors. By Theorem 2.34, we see immediately that

$$\mathbf{X}_n \xrightarrow{d} \mathbf{X} \text{ implies } \mathbf{a}^\top\mathbf{X}_n \xrightarrow{d} \mathbf{a}^\top\mathbf{X} \text{ for any } \mathbf{a} \in \mathbb{R}^k. \tag{4.7}$$

This is because multiplication by a constant vector $\mathbf{a}^T$ is a continuous transformation from $\mathbb{R}^k$ to $\mathbb{R}$. It is not clear, however, whether the converse of statement (4.7) is true. Such a converse would be useful because it would give a means for proving multivariate convergence in distribution using only univariate methods. As the counterexample in Example 2.38 shows, multivariate convergence in distribution does *not* follow from the mere fact that each of the components converges in distribution. Yet the converse of statement (4.7) is much stronger than the statement that each component converges in distribution; could it be true that requiring *all* linear combinations to converge in distribution is strong enough to guarantee multivariate convergence? The answer is yes:

Theorem 4.12 *Cramér-Wold Theorem:* $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$ if and only if $\mathbf{a}^\top \mathbf{X}_n \xrightarrow{d} \mathbf{a}^\top \mathbf{X}$ for all $\mathbf{a} \in \mathbb{R}^k$.

Using the machinery of characteristic functions, to be presented in Section 4.1, the proof of the Cramér-Wold Theorem is immediate; see Exercise 4.3. This theorem in turn provides a straightforward method for proving cerain multivariate theorems using univariate results. For instance, once we establish the univariate Central Limit Theorem (Theorem 4.19), we will show how to use the Cramér-Wold Theorem to prove the multivariate CLT, Theorem 4.9.

Exercises for Section 4.1

Exercise 4.1 **(a)** Prove that if $Y \sim N(0, \sigma^2)$ with $\sigma^2 > 0$, then $\phi_Y(t) = \exp\left(-\frac{1}{2}t^2\sigma^2\right)$. Argue that this demonstrates that Definition 4.6 is valid in the case $k = 1$.

Hint: Verify and solve the differential equation $\phi_Y'(t) = -t\sigma^2\phi_Y(t)$. Use integration by parts.

(b) Using part (a), prove that if $\mathbf{X}$ is defined as in Equation (4.4), then $\phi_{\mathbf{X}}(\mathbf{t}) = \exp\left(i\mathbf{t}^\top\boldsymbol{\mu} - \frac{1}{2}\mathbf{t}^\top\Sigma\mathbf{t}\right)$.

Exercise 4.2 We will prove Theorem 4.2, which states that chacteristic functions uniquely determine their distributions.

(a) First, prove the *Parseval relation* for random $\mathbf{X}$ and $\mathbf{Y}$:

$$\mathrm{E}\left[\exp(-\mathrm{i}\mathbf{a}^\top\mathbf{Y})\phi_{\mathbf{X}}(\mathbf{Y})\right] = \mathrm{E}\ \phi_{\mathbf{Y}}(\mathbf{X} - \mathbf{a}).$$

Hint: Use conditioning to evaluate $\mathrm{E}\exp\{\mathrm{i}(\mathbf{X} - \mathbf{a})^\top\mathbf{Y}\}$.

(b) Suppose that $\mathbf{Y} = (Y_1, \ldots, Y_k)$, where $Y_1, \ldots, Y_k$ are independent and identically distributed normal random variables with mean 0 and variance σ^2. That is, $\mathbf{Y}$ has density

$$f_{\mathbf{Y}}(\mathbf{y}) = (\sqrt{2\pi\sigma^2})^{-k}\exp(-\mathbf{y}^\top\mathbf{y}/2\sigma^2).$$

Show that $\mathbf{X} + \mathbf{Y}$ has density

$$f_{\mathbf{X}+\mathbf{Y}}(\mathbf{s}) = \mathrm{E}\ f_{\mathbf{Y}}(\mathbf{s} - \mathbf{X}).$$

(c) Use the result of Exercise 4.1 along with part (b) to show that

$$f_{\mathbf{X}+\mathbf{Y}}(\mathbf{s}) = (\sqrt{2\pi\sigma^2})^{-k}\,\mathrm{E}\ \phi_{\mathbf{Y}}\left(\frac{\mathbf{X}}{\sigma^2} - \frac{s}{\sigma^2}\right).$$

Argue that this fact proves $\phi_{\mathbf{X}}(\mathbf{t})$ uniquely determines the distribution of $\mathbf{X}$.

Hint: Use parts (a) and (b) to show that the distribution of $\mathbf{X}+\mathbf{Y}$ depends on $\mathbf{X}$ only through $\phi_{\mathbf{X}}$. Then note that $\mathbf{X}+\mathbf{Y} \stackrel{d}{\to} \mathbf{X}$ as $\sigma^2 \to 0$.

Exercise 4.3 Use the Continuity Theorem to prove the Cramér-Wold Theorem, Theorem 4.12.

Hint: $\mathbf{a}^\top \mathbf{X}_n \stackrel{d}{\to} \mathbf{a}^\top \mathbf{X}$ implies that $\phi_{\mathbf{a}^\top \mathbf{X}_n}(1) \to \phi_{\mathbf{a}^\top \mathbf{X}}(1)$.

Exercise 4.4 Suppose $\mathbf{X} \sim N_k(\boldsymbol{\mu}, \Sigma)$, where Σ is invertible. Prove that

$$(\mathbf{X}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{X}-\boldsymbol{\mu}) \sim \chi^2_k.$$

Hint: If Q diagonalizes Σ, say $Q\Sigma Q^\top = \Lambda$, let $\Lambda^{1/2}$ be the diagonal, nonnegative matrix satisfying $\Lambda^{1/2}\Lambda^{1/2} = \Lambda$ and consider $\mathbf{Y}^\top\mathbf{Y}$, where $\mathbf{Y} = (\Lambda^{1/2})^{-1}Q(\mathbf{X}-\boldsymbol{\mu})$.

Exercise 4.5 Let $X_1, X_2, \ldots$ be independent Poisson random variables with mean $\lambda = 1$. Define $Y_n = \sqrt{n}(\overline{X}_n - 1)$.

(a) Find $\mathrm{E}\,(Y_n^+)$, where $Y_n^+ = Y_n I\{Y_n > 0\}$.

(b) Find, with proof, the limit of $\mathrm{E}\,(Y_n^+)$ and prove Stirling's formula

$$n! \;\sim\; \sqrt{2\pi}\, n^{n+1/2} e^{-n}.$$

Hint: Use the result of Exericse 3.12.

Exercise 4.6 Use the Continuity Theorem to prove Theorem 2.19, the univariate Weak Law of Large Numbers.

Hint: Use a Taylor expansion (1.5) with $d = 2$ for both the real and imaginary parts of the characteristic function of $\overline{X}_n$.

Exercise 4.7 Use the Cramér-Wold Theorem along with the univariate Central Limit Theorem (from Example 2.12) to prove Theorem 4.9.

4.2 The Lindeberg-Feller Central Limit Theorem

The Lindeberg-Feller Central Limit Theorem states in part that sums of independent random variables, properly standardized, converge in distribution to standard normal as long as a certain condition, called the Lindeberg Condition, is satisfied. Since these random variables do not have to be identically distributed, this result generalizes the Central Limit Theorem for independent and identically distributed sequences.

4.2.1 The Lindeberg and Lyapunov Conditions

Suppose that $X_1, X_2, \ldots$ are independent random variables such that $\mathrm{E}\, X_n = \mu_n$ and $\mathrm{Var}\, X_n = \sigma_n^2 < \infty$. Define

$$\begin{aligned} Y_n &= X_n - \mu_n, \\ T_n &= \textstyle\sum_{i=1}^n Y_i, \\ s_n^2 &= \mathrm{Var}\, T_n = \textstyle\sum_{i=1}^n \sigma_i^2. \end{aligned}$$

Instead of defining Y_n to be the centered version of X_n, we could have simply taken μ_n to be zero without loss of generality. However, when these results are used in practice, it is easy to forget the centering step, so we prefer to make it explicit here.

Note that T_n/s_n has mean zero and variance 1. We wish to give sufficient conditions that ensure $T_n/s_n \xrightarrow{d} N(0,1)$. We give here two separate conditions, one called the Lindeberg condition and the other called the Lyapunov condition. The *Lindeberg Condition* for sequences states that

$$\text{for every } \epsilon > 0,\ \frac{1}{s_n^2}\sum_{i=1}^n \mathrm{E}\left(Y_i^2 I\{|Y_i| \ge \epsilon s_n\}\right) \to 0 \text{ as } n \to \infty; \tag{4.8}$$

the *Lyapunov Condition* for sequences states that

$$\text{there exists } \delta > 0 \text{ such that } \frac{1}{s_n^{2+\delta}}\sum_{i=1}^n \mathrm{E}\left(|Y_i|^{2+\delta}\right) \to 0 \text{ as } n \to \infty. \tag{4.9}$$

We shall see later (in Theorem 4.16, the Lindeberg-Feller Theorem) that Condition (4.8) implies $T_n/s_n \to N(0,1)$. For now, we show only that Condition (4.9)—the Lyapunov Condition—is stronger than Condition (4.8). Thus, the Lyapunov Condition also implies $T_n/s_n \to N(0,1)$:

Theorem 4.13 The Lyapunov Condition (4.9) implies the Lindeberg Condition (4.8).

Proof: Assume that the Lyapunov Condition is satisfied and fix $\epsilon > 0$. Since $|Y_i| \ge \epsilon s_n$ implies $|Y_i/\epsilon s_n|^\delta \ge 1$, we obtain

$$\begin{aligned} \frac{1}{s_n^2}\sum_{i=1}^n \mathrm{E}\left(Y_i^2 I\{|Y_i| \ge \epsilon s_n\}\right) &\le \frac{1}{\epsilon^\delta s_n^{2+\delta}}\sum_{i=1}^n \mathrm{E}\left(|Y_i|^{2+\delta} I\{|Y_i| \ge \epsilon s_n\}\right) \\ &\le \frac{1}{\epsilon^\delta s_n^{2+\delta}}\sum_{i=1}^n \mathrm{E}\left(|Y_i|^{2+\delta}\right). \end{aligned}$$

Since the right hand side tends to 0, the Lindeberg Condition is satisfied. ■

Example 4.14 Suppose that we perform a series of independent Bernoulli trials with possibly different success probabilities. Under what conditions will the proportion of successes, properly standardized, tend to a normal distribution?

Let $X_n \sim \text{Bernoulli}(p_n)$, so that $Y_n = X_n - p_n$ and $\sigma_n^2 = p_n(1-p_n)$. As we shall see later (Theorem 4.16), either the Lindeberg Condition (4.8) or the Lyapunov Condition (4.9) will imply that $\sum_{i=1}^n Y_i/s_n \stackrel{d}{\to} N(0,1)$.

Let us check the Lyapunov Condition for, say, $\delta = 1$. First, verify that

$$\mathrm{E}\,|Y_n|^3 = p_n(1-p_n)^3 + (1-p_n)p_n^3 = \sigma_n^2[(1-p_n)^2 - p_n^2] \le \sigma_n^2.$$

Using this upper bound on $\mathrm{E}\,|Y_n|^3$, we obtain $\sum_{i=1}^n \mathrm{E}\,|Y_i|^3 \le s_n^2$. Therefore, the Lyapunov condition is satisfied whenever $s_n^2/s_n^3 \to 0$, which implies $s_n \to \infty$. We conclude that the proportion of successes tends to a normal distribution whenever

$$s_n^2 = \sum_{i=1}^n p_n(1-p_n) \to \infty,$$

which will be true as long as $p_n(1-p_n)$ does not tend to 0 too fast.

4.2.2 Independent and Identically Distributed Variables

We now set the stage for proving a central limit theorem for independent and identically distributed random variables by showing that the Lindeberg Condition is satisfied by such a sequence as long as the common variance is finite.

Example 4.15 Suppose that $X_1, X_2, \ldots$ are independent and identically distributed with $\mathrm{E}\,(X_i) = \mu$ and $\mathrm{Var}\,(X_i) = \sigma^2 < \infty$. The case $\sigma^2 = 0$ is uninteresting, so we assume $\sigma^2 > 0$.

Let $Y_i = X_i - \mu$ and $s_n^2 = \mathrm{Var}\,\sum_{i=1}^n Y_i = n\sigma^2$. Fix $\epsilon > 0$. The Lindeberg Condition states that

$$\frac{1}{n\sigma^2}\sum_{i=1}^n \mathrm{E}\,\left(Y_i^2 I\{|Y_i| \ge \epsilon\sigma\sqrt{n}\}\right) \to 0 \text{ as } n \to \infty. \tag{4.10}$$

Since the Y_i are identically distributed, the left hand side of expression (4.10) simplifies to

$$\frac{1}{\sigma^2}\mathrm{E}\,\left(Y_1^2 I\{|Y_1| \ge \epsilon\sigma\sqrt{n}\}\right). \tag{4.11}$$

To simplify notation, let Z_n denote the random variable $Y_1^2 I\{|Y_1| \geq \epsilon\sigma\sqrt{n}\}$. Thus, we wish to prove that $\mathrm{E}\, Z_n \to 0$. Note that Z_n is nonzero if and only if $|Y_1| \geq \epsilon\sigma\sqrt{n}$. Since this event has probability tending to zero as $n \to \infty$, we conclude that $Z_n \xrightarrow{P} 0$ by the definition of convergence in probability. We can also see that $|Z_n| \leq Y_1^2$, and we know that $\mathrm{E}\, Y_1^2 < \infty$. Therefore, we may apply the Dominated Convergence Theorem, Theorem 3.22, to conclude that $\mathrm{E}\, Z_n \to 0$. This demonstrates that the Lindeberg Condition is satisfied.

The preceding argument, involving the Dominated Convergence Theorem, is quite common in proofs that the Lindeberg Condition is satisfied. Any beginning student is well-advised to study this argument carefully.

Note that the assumptions of Example 4.15 are not strong enough to ensure that the Lyapunov Condition (4.9) is satisfied. This is because there are some random variables that have finite variances but no finite $2 + \delta$ moment for any $\delta > 0$. Construction of such an example is the subject of Exercise 4.10. However, such examples are admittedly somewhat pathological, and if one is willing to assume that $X_1, X_2, \ldots$ are independent and identically distributed with $\mathrm{E}\, |X_1|^{2+\delta} = \gamma < \infty$ for some $\delta > 0$, then the Lyapunov Condition is much easier to check than the Lindeberg Condition. Indeed, because $s_n = \sigma\sqrt{n}$, the Lyapunov Condition reduces to

$$\frac{n\gamma}{(n\sigma^2)^{1+\delta/2}} = \frac{\gamma}{n^{\delta/2}\sigma^{2+\delta}} \to 0,$$

which follows immediately.

4.2.3 Triangular Arrays

It is sometimes the case that $X_1, \ldots, X_n$ are independent random variables—possibly even identically distributed—but their distributions depend on n. Take the simple case of the binomial(n, p_n) distribution as an example, where the probability p_n of success on any trial changes as n increases. What can we say about the asymptotic distribution in such a case? It seems that what we need is some way of dealing with a sequence of sequences, say, $X_{n1}, \ldots, X_{nn}$ for $n \geq 1$. This is exactly the idea of a triangular array of random variables.

Generalizing the concept of "sequence of independent random variables," a triangular array or random variables may be visualized as follows:

$$\begin{array}{llll} X_{11} & & & \leftarrow \text{independent} \\ X_{21} & X_{22} & & \leftarrow \text{independent} \\ X_{31} & X_{32} & X_{33} & \leftarrow \text{independent} \\ & \vdots & & \end{array}$$

Thus, we assume that for each n, $X_{n1}, \ldots, X_{nn}$ are independent. Carrying over the notation from before, we assume E $X_{ni} = \mu_{ni}$ and Var $X_{ni} = \sigma_{ni}^2 < \infty$. Let

$$\begin{aligned} Y_{ni} &= X_{ni} - \mu_{ni}, \\ T_n &= \textstyle\sum_{i=1}^n Y_{ni}, \\ s_n^2 &= \text{Var } T_n = \textstyle\sum_{i=1}^n \sigma_{ni}^2. \end{aligned}$$

As before, T_n/s_n has mean 0 and variance 1; our goal is to give conditions under which

$$\frac{T_n}{s_n} \xrightarrow{d} N(0,1). \tag{4.12}$$

Such conditions are given in the Lindeberg-Feller Central Limit Theorem. The key to this theorem is the *Lindeberg condition* for triangular arrays:

$$\text{For every } \epsilon > 0,\ \frac{1}{s_n^2} \sum_{i=1}^n \text{E}\left(Y_{ni}^2 I\left\{|Y_{ni}| \geq \epsilon s_n\right\}\right) \to 0 \text{ as } n \to \infty. \tag{4.13}$$

Before stating the Lindeberg-Feller theorem, we need a technical condition that says essentially that the contribution of each X_{ni} to s_n^2 should be negligible:

$$\frac{1}{s_n^2} \max_{i \leq n} \sigma_{ni}^2 \to 0 \text{ as } n \to \infty. \tag{4.14}$$

Now that Conditions (4.12), (4.13), and (4.14) have been written, the main result may be stated in a single line:

Theorem 4.16 *Lindeberg-Feller Central Limit Theorem:* Condition (4.13) holds if and only if Conditions (4.12) and (4.14) hold.

A proof of the Lindeberg-Feller Theorem is the subject of Exercises 4.8 and 4.9. In most practical applications of this theorem, the Lindeberg Condition (4.13) is used to establish asymptotic normality (4.12); the remainder of the theorem's content is less useful.

Example 4.17 As an extension of Example 4.14, suppose $X_n \sim \text{binomial}(n, p_n)$. The calculations here are not substantially different from those in Example 4.14, so we use the Lindeberg Condition here for the purpose of illustration. We claim that

$$\frac{X_n - np_n}{\sqrt{np_n(1-p_n)}} \xrightarrow{d} N(0,1) \tag{4.15}$$

whenever $np_n(1-p_n) \to \infty$ as $n \to \infty$. In order to use Theorem 4.16 to prove this result, let $Y_{n1}, \ldots, Y_{nn}$ be independent and identically distributed with

$$P(Y_{ni} = 1 - p_n) = 1 - P(Y_{ni} = -p_n) = p_n.$$

Then with $X_n = np_n + \sum_{i=1}^n Y_{ni}$, we obtain $X_n \sim \text{binomial}(n, p_n)$ as specified. Furthermore, $\text{E } Y_{ni} = 0$ and $\text{Var } Y_{ni} = p_n(1-p_n)$, so the Lindeberg condition says that for any $\epsilon > 0$,

$$\frac{1}{np_n(1-p_n)} \sum_{i=1}^n \text{E}\left(Y_{ni}^2 I\left\{ |Y_{ni}| \geq \epsilon\sqrt{np_n(1-p_n)} \right\}\right) \to 0. \tag{4.16}$$

Since $|Y_{ni}| \leq 1$, the left hand side of expression (4.16) will be identically zero whenever $\epsilon\sqrt{np_n(1-p_n)} > 1$. Thus, a sufficient condition for (4.15) to hold is that $np_n(1-p_n) \to \infty$. One may show that this is also a necessary condition (this is Exercise 4.11).

Note that any independent sequence $X_1, X_2, \ldots$ may be considered a triangular array by simply taking $X_{n1} = X_1$ for all $n \geq 1$, $X_{n2} = X_2$ for all $n \geq 2$, and so on. Therefore, Theorem 4.16 applies equally to the Lindeberg Condition (4.8) for sequences. Furthermore, the proof of Theorem 4.13 is unchanged if the sequence Y_i is replaced by the array Y_{ni}. Therefore, we obtain an alternative means for checking asymptotic normality:

Corollary 4.18 Asymptotic normality (4.12) follows if the triangular array above satisfies the *Lyapunov Condition* for triangular arrays:

$$\text{there exists } \delta > 0 \text{ such that } \frac{1}{s_n^{2+\delta}} \sum_{i=1}^n \text{E}\left(|Y_{ni}|^{2+\delta}\right) \to 0 \text{ as } n \to \infty. \tag{4.17}$$

Combining Theorem 4.16 with Example 4.15, in which the Lindeberg condition is verified for a sequence of independent and identically distributed variables with finite positive variance, gives the result commonly referred to simply as "The Central Limit Theorem":

Theorem 4.19 *Univariate Central Limit Theorem for iid sequences:* Suppose that $X_1, X_2, \ldots$ are independent and identically distributed with $\text{E}(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2 < \infty$. Then

$$\sqrt{n}\left(\overline{X}_n - \mu\right) \xrightarrow{d} N(0, \sigma^2). \tag{4.18}$$

The case $\sigma^2 = 0$ is not covered by Example 4.15, but in this case limit (4.18) holds automatically.

We conclude this section by generalizing Theorem 4.19 to the multivariate case, Theorem 4.9. The proof is straightforward using theorem 4.19 along with the Cramér-Wold theorem, theorem 4.12. Recall that the Cramér-Wold theorem allows us to establish multivariate convergence in distribution by proving univariate convergence in distribution for arbitrary linear combinations of the vector components.

Proof of Theorem 4.9: Let $\mathbf{X} \sim N_k(\mathbf{0}, \Sigma)$ and take any vector $\mathbf{a} \in \mathbb{R}^k$. We wish to show that

$$\mathbf{a}^\top \left[\sqrt{n}\left(\overline{\mathbf{X}}_n - \boldsymbol{\mu}\right)\right] \stackrel{d}{\to} \mathbf{a}^\top \mathbf{X}.$$

But this follows immediately from the univariate Central Limit Theorem, since $\mathbf{a}^\top(\mathbf{X}_1 - \boldsymbol{\mu}), \mathbf{a}^\top(\mathbf{X}_2 - \boldsymbol{\mu}), \ldots$ are independent and identically distributed with mean 0 and variance $\mathbf{a}^\top \Sigma \mathbf{a}$. ■

We will see many, many applications of the univariate and multivariate Central Limit Theorems in the chapters that follow.

Exercises for Section 4.2

Exercise 4.8 Prove that (4.13) implies both (4.12) and (4.14) (the "forward half" of the Lindeberg-Feller Theorem). Use the following steps:

(a) Prove that for any complex numbers $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ with $|a_i| \le 1$ and $|b_i| \le 1$,

$$|a_1 \cdots a_n - b_1 \cdots b_n| \le \sum_{i=1}^n |a_i - b_i|. \tag{4.19}$$

Hint: First prove the identity when $n = 2$, which is the key step. Then use mathematical induction.

(b) Prove that

$$\left|\phi_{Y_{ni}}\left(\frac{t}{s_n}\right) - \left(1 - \frac{t^2\sigma_{ni}^2}{2s_n^2}\right)\right| \le \frac{\epsilon |t|^3 \sigma_{ni}^2}{s_n^2} + \frac{t^2}{s_n^2} \mathrm{E}\left(Y_{ni}^2 I\{|Y_{ni}| \ge \epsilon s_n\}\right). \tag{4.20}$$

Hint: Use the results of Exercise 1.43, parts (c) and (d), to argue that for any Y,

$$\left|\exp\left\{\frac{\mathrm{i}tY}{s_n}\right\} - \left(1 + \frac{\mathrm{i}tY}{s_n} - \frac{t^2Y^2}{2s_n^2}\right)\right| \le \left|\frac{tY}{s_n}\right|^3 I\left\{\left|\frac{Y}{s_n}\right| < \epsilon\right\} + \left(\frac{tY}{s_n}\right)^2 I\{|Y| \ge \epsilon s_n\}.$$

(c) Prove that (4.13) implies (4.14).

Hint: For any i, show that

$$\frac{\sigma_{ni}^2}{s_n^2} < \epsilon^2 + \frac{\mathrm{E}\left(Y_{ni}^2 I\{|Y_{ni}| \ge \epsilon s_n\}\right)}{s_n^2}.$$

(d) Use parts (a) and (b) to prove that, for n large enough so that $t^2 \max_i \sigma_{ni}^2/s_n^2 \leq 1$,

$$\left| \prod_{i=1}^n \phi_{Y_{ni}}\left(\frac{t}{s_n}\right) - \prod_{i=1}^n \left(1 - \frac{t^2\sigma_{ni}^2}{2s_n^2}\right) \right| \leq \epsilon |t|^3 + \frac{t^2}{s_n^2} \sum_{i=1}^n \mathrm{E}\,\left(Y_{ni}^2 I\left\{|Y_{ni}| \geq \epsilon s_n\right\}\right).$$

(e) Use part (a) to prove that

$$\left| \prod_{i=1}^n \left(1 - \frac{t^2\sigma_{ni}^2}{2s_n^2}\right) - \prod_{i=1}^n \exp\left(-\frac{t^2\sigma_{ni}^2}{2s_n^2}\right) \right| \leq \frac{t^4}{4s_n^4} \sum_{i=1}^n \sigma_{ni}^4 \leq \frac{t^4}{4s_n^2} \max_{1\leq i\leq n} \sigma_{ni}^2.$$

Hint: Prove that for $x \leq 0$, $|1 + x - \exp(x)| \leq x^2$.

(f) Now put it all together. Show that

$$\left| \prod_{i=1}^n \phi_{Y_{ni}}\left(\frac{t}{s_n}\right) - \prod_{i=1}^n \exp\left(-\frac{t^2\sigma_{ni}^2}{2s_n^2}\right) \right| \to 0,$$

proving (4.12).

Exercise 4.9 In this problem, we prove the converse of Exercise 4.8, which is the part of the Lindeberg-Feller Theorem due to Feller: Under the assumptions of the Exercise 4.8, the variance condition (4.14) and the asymptotic normality (4.12) together imply the Lindeberg condition (4.13).

(a) Define

$$\alpha_{ni} = \phi_{Y_{ni}}(t/s_n) - 1.$$

Prove that

$$\max_{1\leq i\leq n} |\alpha_{ni}| \leq 2 \max_{1\leq i\leq n} P\left(|Y_{ni}| \geq \epsilon s_n\right) + 2\epsilon|t|$$

and thus

$$\max_{1\leq i\leq n} |\alpha_{ni}| \to 0 \text{ as } n \to \infty.$$

Hint: Use the result of Exercise 1.43(a) to show that $|\exp\{it\} - 1| \leq 2\min\{1, |t|\}$ for $t \in \mathbb{R}$. Then use Chebyshev's inequality along with condition (4.14).

(b) Use part (a) to prove that

$$\sum_{i=1}^{n} |\alpha_{ni}|^2 \to 0$$

as $n \to \infty$.

Hint: Use the result of Exercise 1.43(b) to show that $|\alpha_{ni}| \le t^2\sigma_{ni}^2/s_n^2$. Then write $|\alpha_{ni}|^2 \le |\alpha_{ni}| \max_i |\alpha_{ni}|$.

(c) Prove that for n large enough so that $\max_i |\alpha_{ni}| \le 1/2$,

$$\left| \prod_{i=1}^{n} \exp(\alpha_{ni}) - \prod_{i=1}^{n} (1+\alpha_{ni}) \right| \le \sum_{i=1}^{n} |\alpha_{ni}|^2.$$

Hint: Use the fact that $|\exp(z-1)| = \exp(\text{Re } z - 1) \le 1$ for $|z| \le 1$ to argue that Inequality (4.19) applies. Also use the fact that $|\exp(z) - 1 - z| \le |z|^2$ for $|z| \le 1/2$.

(d) From part (c) and condition (4.12), conclude that

$$\sum_{i=1}^{n} \text{Re}\,(\alpha_{ni}) \to -\frac{1}{2}t^2.$$

(e) Show that

$$0 \le \sum_{i=1}^{n} \mathrm{E}\left(\cos \frac{tY_{ni}}{s_n} - 1 + \frac{t^2 Y_{ni}^2}{2s_n^2} \right) \to 0.$$

(f) For arbitrary $\epsilon > 0$, choose t large enough so that $t^2/2 > 2/\epsilon^2$. Show that

$$\left(\frac{t^2}{2} - \frac{2}{\epsilon^2} \right) \frac{1}{s_n^2} \sum_{i=1}^{n} \mathrm{E}\left(Y_{ni}^2 I\{|Y_{ni}| \ge \epsilon s_n\} \right) \le \sum_{i=1}^{n} \mathrm{E}\left(\cos \frac{tY_{ni}}{s_n} - 1 + \frac{t^2 Y_{ni}^2}{2s_n^2} \right),$$

which completes the proof.

Hint: Bound the expression in part (e) below by using the fact that -1 is a lower bound for $\cos x$. Also note that $|Y_{ni}| \ge \epsilon s_n$ implies $-2 \ge -2Y_{ni}^2/(\epsilon^2 s_n^2)$.

Exercise 4.10 Give an example of an independent and identically distributed sequence to which the Central Limit Theorem 4.19 applies but for which the Lyapunov condition is not satisfied.

Exercise 4.11 In Example 4.17, we show that $np_n(1-p_n) \to \infty$ is a sufficient condition for (4.15) to hold. Prove that it is also a necessary condition. You may assume that $p_n(1-p_n)$ is always nonzero.

Hint: Use the Lindeberg-Feller Theorem.

Exercise 4.12 **(a)** Suppose that $X_1, X_2, \ldots$ are independent and identically distributed with $\mathrm{E}\, X_i = \mu$ and $0 < \mathrm{Var}\, X_i = \sigma^2 < \infty$. Let $a_{n1}, \ldots, a_{nn}$ be constants satisfying

$$\frac{\max_{i \le n} a_{ni}^2}{\sum_{j=1}^n a_{nj}^2} \to 0 \quad \text{as } n \to \infty.$$

Let $T_n = \sum_{i=1}^n a_{ni} X_i$, and prove that $(T_n - \mathrm{E}\, T_n)/\sqrt{\mathrm{Var}\, T_n} \xrightarrow{d} N(0,1)$.

(b) Reconsider Example 2.22, the simple linear regression case in which

$$\hat\beta_{0n} = \sum_{i=1}^n v_i^{(n)} Y_i \quad \text{and} \quad \hat\beta_{1n} = \sum_{i=1}^n w_i^{(n)} Y_i,$$

where

$$w_i^{(n)} = \frac{z_i - \overline{z}_n}{\sum_{j=1}^n (z_j - \overline{z}_n)^2} \quad \text{and} \quad v_i^{(n)} = \frac{1}{n} - \overline{z}_n w_i^{(n)}$$

for constants $z_1, z_2, \ldots$. Using part (a), state and prove sufficient conditions on the constants z_i that ensure the asymptotic normality of $\sqrt{n}(\hat\beta_{0n} - \beta_0)$ and $\sqrt{n}(\hat\beta_{1n} - \beta_1)$. You may assume the results of Example 2.22, where it was shown that $\mathrm{E}\, \hat\beta_{0n} = \beta_0$ and $\mathrm{E}\, \hat\beta_{1n} = \beta_1$.

Exercise 4.13 Give an example (with proof) of a sequence of independent random variables $Z_1, Z_2, \ldots$ with $\mathrm{E}\,(Z_i) = 0$, $\mathrm{Var}\,(Z_i) = 1$ such that $\sqrt{n}(\overline{Z}_n)$ does not converge in distribution to $N(0,1)$.

Exercise 4.14 Let $(a_1, \ldots, a_n)$ be a random permutation of the integers $1, \ldots, n$. If $a_j < a_i$ for some $i < j$, then the pair (i, j) is said to form an inversion. Let X_n be the total number of inversions:

$$X_n = \sum_{j=2}^n \sum_{i=1}^{j-1} I\{a_j < a_i\}.$$

For example, if $n = 3$ and we consider the permutation $(3, 1, 2)$, there are 2 inversions since $1 = a_2 < a_1 = 3$ and $2 = a_3 < a_1 = 3$. This problem asks you to find the asymptotic distribution of X_n.

(a) Define $Y_1 = 0$ and for $j > 1$, let

$$Y_j = \sum_{i=1}^{j-1} I\{a_j < a_i\}$$

be the number of a_i greater than a_j to the left of a_j. Then the Y_j are independent (you don't have to show this; you may wish to think about why, though). Find $\mathrm{E}\,(Y_j)$ and $\mathrm{Var}\, Y_j$.

(b) Use $X_n = Y_1 + Y_2 + \cdots + Y_n$ to prove that

$$\frac{3}{2}\sqrt{n}\left(\frac{4X_n}{n^2} - 1\right) \xrightarrow{d} N(0, 1).$$

(c) For $n = 10$, evaluate the distribution of inversions as follows. First, simulate 1000 permutations on $\{1, 2, \ldots, 10\}$ and for each permutation, count the number of inversions. Plot a histogram of these 1000 numbers. Use the results of the simulation to estimate $P(X_{10} \le 24)$. Second, estimate $P(X_{10} \le 24)$ using a normal approximation. Can you find the exact integer c such that $10!P(X_{10} \le 24) = c$?

Exercise 4.15 Suppose that X_1, X_2, X_3 is a sample of size 3 from a beta $(2, 1)$ distribution.

(a) Find $P(X_1 + X_2 + X_3 \le 1)$ exactly.

(b) Find $P(X_1 + X_2 + X_3 \le 1)$ using a normal approximation derived from the central limit theorem.

(c) Let $Z = I\{X_1 + X_2 + X_3 \le 1\}$. Approximate $\mathrm{E}\, Z = P(X_1 + X_2 + X_3 \le 1)$ by $\overline{Z} = \sum_{i=1}^{1000} Z_i/1000$, where $Z_i = I\{X_{i1} + X_{i2} + X_{i3} \le 1\}$ and the X_{ij} are independent beta $(2, 1)$ random variables. In addition to $\overline{Z}$, report $\mathrm{Var}\, Z$ for your sample. (To think about: What is the theoretical value of $\mathrm{Var}\, Z$?)

(d) Approximate $P(X_1 + X_2 + X_3 \le \frac{3}{2})$ using the normal approximation and the simulation approach. (Don't compute the exact value, which is more difficult to than in part (a); do you see why?)

Exercise 4.16 Lindeberg and Lyapunov impose conditions on moments so that asymptotic normality occurs. However, it is possible to have asymptotic normality even if there are no moments at all. Let X_n assume the values $+1$ and -1 with probability $(1-2^{-n})/2$ each and the value 2^k with probability 2^{-k} for $k > n$.

(a) Show that $\mathrm{E}\,(X_n^j) = \infty$ for all positive integers j and n.

(b) Show that $\sqrt{n}\left(\overline{X}_n\right) \xrightarrow{d} N(0,1)$.

Exercise 4.17 Assume that elements ("coupons") are drawn from a population of size n, randomly and with replacement, until the number of distinct elements that have been sampled is r_n, where $1 \le r_n \le n$. Let S_n be the drawing on which this first happens. Suppose that $r_n/n \to \rho$, where $0 < \rho < 1$.

(a) Suppose $k-1$ distinct coupons have thus far entered the sample. Let X_{nk} be the waiting time until the next distinct one appears, so that

$$S_n = \sum_{k=1}^{r_n} X_{nk}.$$

Find the expectation and variance of X_{nk}.

(b) Let $m_n = \mathrm{E}\,(S_n)$ and $\tau_n^2 = \mathrm{Var}\,(S_n)$. Show that

$$\frac{S_n - m_n}{\tau_n} \xrightarrow{d} N(0,1).$$

Tip: One approach is to apply Lyapunov's condition with $\delta = 2$. This involves demonstrating an asymptotic expression for τ_n^2 and a bound on $\mathrm{E}\,\left[X_{nk} - E(X_{nk})\right]^4$. There are several ways to go about this.

Exercise 4.18 Suppose that $X_1, X_2, \ldots$ are independent binomial$(2,p)$ random variables. Define $Y_i = I\{X_i = 0\}$.

(a) Find $\mathbf{a}$ such that the joint asymptotic distribution of

$$\sqrt{n}\left[\begin{pmatrix}\overline{X}_n \\ \overline{Y}_n\end{pmatrix} - \mathbf{a}\right]$$

is nontrivial, and find this joint asymptotic distribution.

(b) Using the Cramér-Wold Theorem, Theorem 4.12, find the asymptotic distribution of $\sqrt{n}(\overline{X}_n + \overline{Y}_n - 1 - p^2)$.

4.3 Stationary m-Dependent Sequences

Here we consider sequences that are identically distributed but not independent. In fact, we make a stronger assumption than identically distributed; namely, we assume that $X_1, X_2, \ldots$ is a stationary sequence. (Stationary is defined in Definition 2.24.) Denote E X_i by μ and let $\sigma^2 = \text{Var } X_i$.

We seek sufficient conditions for the asymptotic normality of $\sqrt{n}(\overline{X}_n - \mu)$. The variance of $\overline{X}_n$ for a stationary sequence is given by Equation (2.20). Letting $\gamma_k = \text{Cov}\,(X_1, X_{1+k})$, we conclude that

$$\text{Var }\left\{\sqrt{n}(\overline{X}_n - \mu)\right\} = \sigma^2 + \frac{2}{n}\sum_{k=1}^{n-1}(n-k)\gamma_k. \tag{4.21}$$

Suppose that

$$\frac{2}{n}\sum_{k=1}^{n-1}(n-k)\gamma_k \to \gamma \tag{4.22}$$

as $n \to \infty$. Then based on Equation (4.21), it seems reasonable to ask whether

$$\sqrt{n}(\overline{X}_n - \mu) \stackrel{d}{\to} N(0, \sigma^2 + \gamma).$$

The answer, in many cases, is yes. This section explores one such case.

Recall from Definition 2.26 that $X_1, X_2, \ldots$ is m-dependent for some $m \geq 0$ if the vector $(X_1, \ldots, X_i)$ is independent of $(X_{i+j}, X_{i+j+1}, \ldots)$ whenever $j > m$. Therefore, for an m-dependent sequence we have $\gamma_k = 0$ for all $k > m$, so limit (4.22) becomes

$$\frac{2}{n}\sum_{k=1}^{n-1}(n-k)\gamma_k \to 2\sum_{k=1}^{m}\gamma_k.$$

For a stationary m-dependent sequence, the following theorem asserts the asymptotic normality of $\overline{X}_n$ as long as the X_i are bounded:

Theorem 4.20 If for some $m \geq 0$, $X_1, X_2, \ldots$ is a stationary m-dependent sequence of bounded random variables with E $X_i = \mu$ and Var $X_i = \sigma^2$, then

$$\sqrt{n}(\overline{X}_n - \mu) \stackrel{d}{\to} N\left(0, \sigma^2 + 2\sum_{k=1}^{m}\text{Cov}\,[X_1, X_{1+k}]\right).$$

The assumption in Theorem 4.20 that the X_i are bounded is not necessary, as long as $\sigma^2 < \infty$. However, the proof of the theorem is quite tricky without the boundedness assumption, and the theorem is strong enough for our purposes as it stands. See, for instance, Ferguson (1996) for a complete proof. The theorem may be proved using the following strategy: For some integer k_n, define random variables $V_1, V_2, \ldots$ and $W_1, W_2, \ldots$ as follows:

$$\begin{array}{ll} V_1 = X_1 + \cdots + X_{k_n}, & W_1 = X_{k_n+1} + \cdots + X_{k_n+m}, \\ V_2 = X_{k_n+m+1} + \cdots + X_{2k_n+m}, & W_2 = X_{2k_n+m+1} + \cdots + X_{2k_n+2m}, \\ \vdots & \end{array} \tag{4.23}$$

In other words, each V_i is the sum of k_n of the X_i and each W_i is the sum of m of the X_i. Because the sequence of X_i is m-dependent, we conclude that the V_i are independent. For this reason, we may apply the Lindeberg-Feller theorem to the V_i. If k_n is defined carefully, then the contribution of the W_i may be shown to be negligible. This strategy is implemented in Exercise 4.19, where a proof of Theorem 4.20 is outlined.

Example 4.21 *Runs of successes:* Suppose $X_1, X_2, \ldots$ are independent Bernoulli(p) variables. Let T_n denote the number of runs of successes in $X_1, \ldots, X_n$, where a run of successes is defined as a sequence of consecutive X_i, all of which equal 1, that is both preceded and followed by zeros (unless the run begins with X_1 or ends with X_n). What is the asymptotic distribution of T_n?

We note that

$$\begin{aligned} T_n &= \sum_{i=1}^n I\{\text{run starts at } i\text{th position}\} \\ &= X_1 + \sum_{i=2}^n X_i(1 - X_{i-1}), \end{aligned}$$

since a run starts at the ith position for $i > 1$ if and only if $X_i = 1$ and $X_{i-1} = 0$.

Letting $Y_i = X_{i+1}(1 - X_i)$, we see immediately that $Y_1, Y_2, \ldots$ is a stationary 1-dependent sequence with $\mathrm{E}\, Y_i = p(1-p)$, so that by Theorem 4.20, $\sqrt{n}\{\overline{Y}_n - p(1-p)\} \stackrel{d}{\to} N(0, \tau^2)$, where

$$\begin{aligned} \tau^2 &= \mathrm{Var}\, Y_1 + 2\,\mathrm{Cov}\,(Y_1, Y_2) \\ &= \mathrm{E}\, Y_1^2 - (\mathrm{E}\, Y_1)^2 + 2\,\mathrm{E}\, Y_1 Y_2 - 2(\mathrm{E}\, Y_1)^2 \\ &= \mathrm{E}\, Y_1 - 3(\mathrm{E}\, Y_1)^2 \;=\; p(1-p) - 3p^2(1-p)^2. \end{aligned}$$

Since

$$\frac{T_n - np(1-p)}{\sqrt{n}} = \sqrt{n}\{\overline{Y}_n - p(1-p)\} + \frac{X_1 - Y_n}{\sqrt{n}},$$

we conclude that

$$\frac{T_n - np(1-p)}{\sqrt{n}} \xrightarrow{d} N(0, \tau^2).$$

Exercises for Section 4.3

Exercise 4.19 We wish to prove theorem 4.20. Suppose $X_1, X_2, \ldots$ is a stationary m-dependent sequence of bounded random variables such that $\text{Var}\, X_i = \sigma^2$. Without loss of generality, assume $\text{E}\, X_i = 0$. We wish to prove that $\sqrt{n}(\overline{X}_n) \xrightarrow{d} N(0, \tau^2)$, where

$$\tau^2 = \sigma^2 + 2\sum_{k=1}^{m} \text{Cov}\,(X_1, X_{1+k}).$$

For all n, define $k_n = \lfloor n^{1/4} \rfloor$ and $\ell_n = \lfloor n/(k_n + m) \rfloor$ and $t_n = \ell_n(k_n + m)$. Define $V_1, \ldots, V_{\ell_n}$ and $W_1, \ldots, W_{\ell_n}$ as in Equation (4.23). Then

$$\sqrt{n}(\overline{X}_n) = \frac{1}{\sqrt{n}}\sum_{i=1}^{\ell_n} V_i + \frac{1}{\sqrt{n}}\sum_{i=1}^{\ell_n} W_i + \frac{1}{\sqrt{n}}\sum_{i=t_n+1}^{n} X_i.$$

(a) Prove that

$$\frac{1}{\sqrt{n}}\sum_{i=t_n+1}^{n} X_i \xrightarrow{P} 0. \tag{4.24}$$

Hint: Bound the left hand side of expression (4.24) using Markov's inequality (1.35) with $r = 1$. What is the greatest possible number of summands?

(b) Prove that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{\ell_n} W_i \xrightarrow{P} 0.$$

Hint: For $k_n > m$, the W_i are independent and identically distributed with distributions that do not depend on n. Use the central limit theorem on $(1/\sqrt{\ell_n})\sum_{i=1}^{\ell_n} W_i$.

(c) Prove that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{\ell_n} V_i \xrightarrow{d} N(0, \tau^2),$$

then use Slutsky's theorem to prove theorem 4.20.

Hint: Use the Lindeberg-Feller theorem.

Exercise 4.20 Suppose $X_0, X_1, \ldots$ is an independent sequence of Bernoulli trials with success probability p. Suppose X_i is the indicator of your team's success on rally i in a volleyball game. Your team scores a point each time it has a success that follows another success. Let $S_n = \sum_{i=1}^{n} X_{i-1}X_i$ denote the number of points your team scores by time n.

(a) Find the asymptotic distribution of S_n.

(b) Simulate a sequence $X_0, X_1, \ldots, X_{1000}$ as above and calculate S_{1000} for $p = .4$. Repeat this process 100 times, then graph the empirical distribution of S_{1000} obtained from simulation on the same axes as the theoretical asymptotic distribution from (a). Comment on your results.

Exercise 4.21 Let $X_0, X_1, \ldots$ be independent and identically distributed random variables from a continuous distribution $F(x)$. Define $Y_i = I\{X_i < X_{i-1} \text{ and } X_i < X_{i+1}\}$. Thus, Y_i is the indicator that X_i is a relative minimum. Let $S_n = \sum_{i=1}^{n} Y_i$.

(a) Find the asymptotic distribution of S_n.

(b) Let $n = 5000$. For a sample $X_0, \ldots, X_{5001}$ of size 5002 from the uniform $(0, 1)$ random number generator in R, compute an approximate two-sided p-value based on the observed value of S_n and the answer to part (a). The null hypothesis is that the sequence of "random" numbers generated is independent and identically distributed. (Naturally, the "random" numbers are not random at all, but are generated by a deterministic formula that is supposed to mimic randomness.)