BUSS975 Causal Inference in Financial Research

Ji-Woong Chung
chung_jiwoong@korea.ac.kr
Korea University Business School

Chapter 4

Properties of Estimators

Introduction and Motivation

In earlier chapters, we reviewed probability theory as a language for characterizing uncertainty. We introduced random variables and their distributions, and discussed concepts to describe features of random variables (e.g. expectation, variance). We also considered restrictions on joint distributions of random variables. Armed with this probabilistic toolbox, we can now turn to the central task of **statistical inference**: using data to learn about unknown quantities of interest.

For example, in a returns-to-education study, a parameter of interest might be the difference in average earnings between those who attended college and those who did not. In notation, one such causal estimand is

$$\tau = E[\, Y_i(1) - Y_i(0) \mid D_i = 1\,] = E[\, Y_i \mid D_i = 1\,] - E[\, Y_i \mid D_i = 0\,],$$

where $Y_i(1)$ and $Y_i(0)$ denote person i's potential earnings with and without college education, and D_i is an indicator for attending college. The quantity $E[Y \mid D = 1] - E[Y \mid D = 0]$ is a feature of the joint distribution of (Y, D) in the population. In reality, however, we do not know this distribution exactly; we only observe a finite sample of data. *Statistics forms a bridge between probability models and data*, allowing us to use observed data to infer the values of theoretical quantities like τ.

In this chapter, we introduce the formal concepts of *estimators* and their

properties. We first define what an estimator is and explore various examples of how one can construct estimators for a given parameter. We then examine **finite-sample properties** of estimators, such as bias, variance, and mean squared error (MSE), which describe the behavior of an estimator for a fixed sample size. Next, we discuss **large-sample (asymptotic) properties**, including consistency (whether an estimator approaches the true value as the sample size grows) and the asymptotic distribution (how the estimator behaves for large n, often characterized by the Central Limit Theorem). We also introduce tools like the Continuous Mapping Theorem and Slutsky's Theorem, which help derive properties of complex estimators. Finally, we show how large-sample results lead to practical measures of uncertainty (standard errors and confidence intervals), and we clarify the correct interpretation of such intervals and estimates.

4.1 Estimators and Random Samples

4.1.1 Random Sampling and IID Data

Statistical inference typically assumes that our data come from a *random sample* of the population of interest. Formally, we say random variables $X_1, X_2, \ldots, X_n$ are *independent* if knowing the value of some X_i provides no information about X_j for $i \neq j$, and they are *identically distributed* if all X_i share the same probability distribution F. When both conditions hold, we write

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} F,$$

meaning the sample $X_1, \ldots, X_n$ consists of n independent draws from the distribution F. This is the mathematical idealization of drawing n observations at random from a population described by F. We will primarily (though not exclusively) consider the iid sampling framework in this text.

Example 4.1 (IID vs. Merely Independent or Identical)**.** Suppose $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ are two independent normal random variables. If $\mu_1 = \mu_2$ and $\sigma_1^2 = \sigma_2^2$, then X_1 and X_2 are not only independent but also identically distributed (in fact, $X_1, X_2 \overset{iid}{\sim} N(\mu_1, \sigma_1^2)$). If, however, $\mu_1 \neq \mu_2$ (or $\sigma_1^2 \neq \sigma_2^2$), then while X_1 and X_2 are independent, they are not identically distributed. Conversely, if X_1 and X_2 share the same distribution but are

not independent, then they are identically distributed but not independent. The term *iid* requires both properties to hold.

Notation: Sometimes we abuse notation by writing $X_1, \ldots, X_n \overset{iid}{\sim} X$ to mean that each X_i has the same distribution as some random variable X. This is just a shorthand indicating iid sampling from the distribution of X. We will also often omit explicit mention of the underlying distribution or probability measure when the context is clear (e.g. writing $E[X]$ instead of $E_F[X]$ to denote expectation under distribution F).

4.1.2 Parameters, Estimands, and Estimators

In many situations, we are interested not in the entirety of the distribution F, but in some specific numerical feature of it. Such features are called *parameters* (or *estimands*). For instance, the population mean $\mu = E[X]$ is a parameter of the distribution of a random variable X. Other examples of parameters include a population variance $\sigma^2 = \mathrm{Var}(X)$, a population median, a regression coefficient in a linear model, or the causal effect τ mentioned earlier. In general, think of a parameter θ as a fixed (but unknown) number that we want to learn about, which is defined as some functional of the population distribution.

An *estimator* is a rule or formula that produces a guess for the value of a parameter, using sample data. Formally, an estimator $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ is a function of the observed sample $(X_1, \ldots, X_n)$. Because it is a function of random variables, an estimator $\hat{\theta}_n$ is itself a random variable. The outcome of applying the estimator to a particular observed dataset is called an *estimate* (a realized value of the estimator).

It is important to distinguish conceptually between the true parameter θ and an estimator $\hat{\theta}_n$. The parameter θ is a fixed, non-random number describing the population (often unknown to us), whereas the estimator $\hat{\theta}_n$ is random due to its dependence on the random sample. We use the estimator's observed value (the estimate) as our best guess for the true θ. In notation, it is common to denote parameters by Greek letters (e.g. θ, μ, β) and use a "hat" to denote the estimator (e.g. $\hat{\theta}_n$) based on a sample of size n.

Example 4.2 (Estimators for the Population CDF)**.** A fundamental example

of an estimator is the *empirical cumulative distribution function (empirical CDF)*. Let $F(x) = P(X \leq x)$ be the true cumulative distribution function of a random variable X. Given an iid sample $X_1, \ldots, X_n \sim F$, we can estimate $F(x)$ by the proportion of sample points $\leq x$. Define

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{X_i \leq x\},$$

where $\mathbf{1}\{\cdot\}$ is the indicator function. $\hat{F}_n(x)$ is the sample CDF: for each x, it counts the fraction of observations that do not exceed x. $\hat{F}_n(x)$ is an estimator of $F(x)$ because intuitively, the fraction of the sample less than or equal to x should be close to the probability of a random observation being $\leq x$.

The empirical CDF $\hat{F}_n$ provides a non-parametric estimate of the entire distribution F. Moreover, it illustrates the **sample analogue principle**: to estimate a feature of the distribution, use the analogous feature of the empirical distribution. For example, if our parameter of interest is $F(x)$ itself, we plug in the empirical distribution to get $\hat{F}_n(x)$. If the parameter of interest is something like $P(X \in A)$ for some event A, the sample analogue would be the proportion of the sample falling in A. In the next example, we apply this principle to estimate mean and variance.

Example 4.3 (Sample Mean and Sample Variance)**.** Consider again an iid sample $X_1, \ldots, X_n \sim F$ with unknown population mean $\mu = E[X]$ and variance $\sigma^2 = Var(X)$. The *sample mean*

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

is the estimator obtained by taking the expectation with respect to the empirical CDF $\hat{F}_n$. In other words, $\hat{\mu}_n = E_n[X]$, where E_n denotes expectation under the empirical distribution (assigning probability $1/n$ to each observed value). This is exactly the sample analogue of the population mean.

Similarly, the *sample variance*

$$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\mu}_n)^2$$

is the sample analogue of the population variance (it is the variance of the empirical distribution $\hat{F}_n$). $\hat{\sigma}_n^2$ is a natural estimator for $\sigma^2 = E[(X-\mu)^2]$. Note that we divide by n here (the definition following directly from the empirical distribution). In practice, an adjusted version with $n-1$ in the denominator is often used; we will discuss the reason for that adjustment later.

Both $\hat{\mu}_n$ and $\hat{\sigma}_n^2$ are functions of the sample and hence are random variables. Before seeing any data, $\hat{\mu}_n$ might take different values for different samples. After observing data, we obtain concrete numbers (the estimates) for $\hat{\mu}_n$ and $\hat{\sigma}_n^2$.

The sample analogue principle is a powerful guideline, but it is not the only way to construct estimators. Another common approach is to define an estimator as the value that optimizes some criterion (often related to a loss or likelihood function). Such *extremum estimators* include least-squares and maximum likelihood estimators.

For instance, suppose we want to estimate the mean $\mu = E[X]$. We could define an estimator as the number that minimizes the sum of squared deviations of the data:

$$\hat{\mu}_n = \arg\min_{m \in \mathbb{R}} \sum_{i=1}^{n} (X_i - m)^2.$$

This is a least-squares estimator for μ. To find the minimizer, we set the derivative to zero:

$$\frac{\partial}{\partial m} \sum_{i=1}^{n} (X_i - m)^2 = -2 \sum_{i=1}^{n} (X_i - m) = 0,$$

which yields $\sum_{i=1}^{n} X_i - nm = 0$ and hence $m = \frac{1}{n}\sum_{i=1}^{n} X_i$. Therefore the solution is $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, which is exactly the sample mean. In this case, the extremum approach agrees with the sample analogue principle.

Extremum estimation is a broad class: many estimators in econometrics and machine learning (e.g. OLS, MLE, GMM) are defined as the minimizer or maximizer of some sample criterion. They often coincide with intuitive sample analogues of population defining conditions (such as "moments" or likelihood equations).

4.1.3 Multiple Estimators for the Same Parameter

For any given parameter, there can be many different reasonable estimators. Not all estimators are equally good, and one goal of statistics is to develop criteria to compare and choose among estimators.

Example 4.4 (Four Estimators for a Mean)**.** Suppose $X_1, \ldots, X_n \overset{iid}{\sim} F$ with mean $\mu = E[X]$. Consider the following four estimators for μ:

1. $\hat{\mu}_n^{(1)} = 0$. (Ignore the data completely and always estimate μ as 0.)
2. $\hat{\mu}_n^{(2)} = X_1$. (Use the first observation as the estimate.)
3. $\hat{\mu}_n^{(3)} = \frac{1}{n}\sum_{i=1}^{n} X_i$. (The sample mean, as before.)
4. $\hat{\mu}_n^{(4)} = \frac{1}{n+\lambda}\sum_{i=1}^{n} X_i$ for some fixed $\lambda > 0$. (A "shrunk" version of the sample mean that slightly down-weights the data.)

Each of these is a legitimate estimator in the sense that it is a function of the sample. Intuitively, $\hat{\mu}_n^{(1)}$ will be useful only in degenerate cases (perhaps if we *a priori* know μ is near 0). $\hat{\mu}_n^{(2)}$ uses minimal information from the sample, so we might expect it to be inefficient. $\hat{\mu}_n^{(3)}$ is the natural estimator. $\hat{\mu}_n^{(4)}$ introduces a small bias on purpose by adding λ to the divisor; this kind of adjustment might reduce variance.

Which estimator is *best*? We need formal criteria to compare them. In the next section, we introduce several finite-sample properties (bias, variance, mean squared error) that help to evaluate and rank estimators. Ultimately, one often prefers an estimator that balances small bias and small variance.

4.1.4 The Sampling Distribution of an Estimator

Because an estimator $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ is a random variable, it has a probability distribution of its own. This is known as the *sampling distribution* of the estimator. The sampling distribution describes how the estimator would vary if we repeated the data-generating process many times (each time producing a new sample and hence a new estimate).

For example, if $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the sample mean of iid draws from F, then the sampling distribution of $\hat{\mu}_n$ is the distribution of $\frac{1}{n}\sum_{i=1}^{n} X_i$. We will often summarize or approximate this distribution to make inferences (as we will see with the Central Limit Theorem).

Sampling distributions often depend on n. For instance, the variance of $\hat{\mu}_n$ for iid data is $Var(\hat{\mu}_n) = \sigma^2/n$, which shrinks with n. In general, **finite-sample properties** of an estimator refer to exact properties of its distribution for each fixed n. **Large-sample (asymptotic) properties** refer to limiting or approximate properties as $n \to \infty$. Both are important: finite-sample analysis tells us about performance in a given sample, while asymptotic analysis tells us the long-run behavior as data become plentiful.

The next sections discuss key finite-sample properties (bias, variance, MSE) and then asymptotic properties (consistency and distributional approximation).

4.2 Finite-Sample Properties of Estimators

4.2.1 Bias

We first consider the notion of *bias*, which measures the systematic error in an estimator.

Definition 4.5 (Bias)**.** The **bias** of an estimator $\hat{\theta}_n$ for a parameter θ is defined as the difference between its expectation and the true value:

$$Bias(\hat{\theta}_n) = E[\hat{\theta}_n] - \theta.$$

If $Bias(\hat{\theta}_n) = 0$, we say $\hat{\theta}_n$ is **unbiased**. If $E[\hat{\theta}_n] < \theta$, the estimator is **downward biased** (it underestimates on average). If $E[\hat{\theta}_n] > \theta$, it is **upward biased**.

Unbiasedness is a desirable property: an unbiased estimator, on average, hits the true parameter. However, as we will see later, unbiasedness is not the only consideration for a good estimator (an unbiased estimator might have a large variance, making it unreliable in any given sample).

Example 4.6 (Bias of the Four Mean Estimators)**.** Consider the four estimators from Example 4.4 for the mean $\mu = E[X]$ of an iid sample. We can compute their bias as follows:

$$
\begin{aligned}
Bias(\hat{\mu}_n^{(1)}) &= E[0] - \mu = 0 - \mu = -\mu, \\
Bias(\hat{\mu}_n^{(2)}) &= E[X_1] - \mu = \mu - \mu = 0, \\
Bias(\hat{\mu}_n^{(3)}) &= E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] - \mu = \frac{1}{n}\sum_{i=1}^{n} E[X_i] - \mu = \frac{1}{n}(n\mu) - \mu = 0, \\
Bias(\hat{\mu}_n^{(4)}) &= E\left[\frac{1}{n+\lambda}\sum_{i=1}^{n} X_i\right] - \mu = \frac{1}{n+\lambda}\sum_{i=1}^{n} E[X_i] - \mu \\
&= \frac{n\mu}{n+\lambda} - \mu = -\frac{\lambda}{n+\lambda}\mu.
\end{aligned}
$$

Thus $\hat{\mu}_n^{(1)}$ is biased (unless $\mu = 0$); it underestimates μ on average (for $\mu > 0$) by an amount $|\mu|$. $\hat{\mu}_n^{(2)}$ and $\hat{\mu}_n^{(3)}$ are unbiased estimators of μ. The estimator $\hat{\mu}_n^{(4)}$ is biased: it underestimates μ by $-\frac{\lambda}{n+\lambda}\mu$ (note this bias is negative if $\mu > 0$ and positive if $\mu < 0$). The bias of $\hat{\mu}_n^{(4)}$ depends on the unknown μ itself, which is often the case for biased estimators introduced via a constant like λ. However, observe that the bias of $\hat{\mu}_n^{(4)}$ shrinks as n grows (for fixed λ). In fact, $Bias(\hat{\mu}_n^{(4)}) \to 0$ as $n \to \infty$, which foreshadows that $\hat{\mu}_n^{(4)}$ can still be useful in large samples.

This example shows that it is easy to create an unbiased estimator (e.g. $\hat{\mu}_n^{(2)}$ or $\hat{\mu}_n^{(3)}$). Unbiasedness alone does not tell us which unbiased estimator is better.

Example 4.7 (Bias of the Sample Variance)**.** Consider the sample variance $\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu}_n)^2$ from Example 4.3, which is an estimator for

$Var(X) = \sigma^2$. We can expand $\hat{\sigma}_n^2$ as follows:

$$\begin{aligned}\hat{\sigma}_n^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu}_n)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(X_i^2 - 2X_i\hat{\mu}_n + \hat{\mu}_n^2) \\ &= \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \frac{2\hat{\mu}_n}{n}\sum_{i=1}^{n}X_i + \frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_n^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \hat{\mu}_n^2.\end{aligned}$$

(The cross term simplified because $\frac{1}{n}\sum_{i=1}^{n}X_i = \hat{\mu}_n$.) Now, taking expectation:

$$E[\hat{\sigma}_n^2] = E\left[\frac{1}{n}\sum_{i=1}^{n}X_i^2\right] - E[\hat{\mu}_n^2] = \frac{1}{n}\sum_{i=1}^{n}E[X_i^2] - Var(\hat{\mu}_n) - [E(\hat{\mu}_n)]^2.$$

Since $E[X_i^2] = Var(X) + [E(X)]^2 = \sigma^2 + \mu^2$, and $E[\hat{\mu}_n] = \mu$, $Var(\hat{\mu}_n) = \sigma^2/n$, we get

$$E[\hat{\sigma}_n^2] = \sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2 = \sigma^2\Big(1 - \frac{1}{n}\Big).$$

Thus

$$Bias(\hat{\sigma}_n^2) = E[\hat{\sigma}_n^2] - \sigma^2 = -\frac{1}{n}\sigma^2.$$

The sample variance with denominator n is *biased downward*: its expectation is slightly less than the true variance σ^2. The bias is $-\sigma^2/n$, which for finite n is nonzero (unless $\sigma^2 = 0$), but the bias tends to 0 as $n \to \infty$.

Can we construct an unbiased estimator for $Var(X)$? Yes: noting the above result, if we multiply $\hat{\sigma}_n^2$ by $\frac{n}{n-1}$, the expectation will become σ^2. Specifically, define

$$\tilde{\sigma}_n^2 = \frac{n}{n-1}\hat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu}_n)^2.$$

This $\tilde{\sigma}_n^2$ is the usual *unbiased sample variance* formula (with $n-1$ in the denominator). Indeed $E[\tilde{\sigma}_n^2] = \frac{n}{n-1}E[\hat{\sigma}_n^2] = \frac{n}{n-1}\sigma^2(1 - \frac{1}{n}) = \sigma^2$. In practice, many statistical software packages use $n-1$ in the denominator for sample variance by default for exactly this reason.

4.2.2 Variance of an Estimator

Bias measures the accuracy of an estimator in terms of its average (expected) value. The other side of the coin is the estimator's *variance*, which measures the precision or reliability of the estimator across different samples.

Definition 4.8 (Variance of an Estimator)**.** The **variance** of an estimator $\hat{\theta}_n$ is

$$Var(\hat{\theta}_n) = E\Big[(\hat{\theta}_n - E[\hat{\theta}_n])^2\Big].$$

The positive square root of $Var(\hat{\theta}_n)$ is called the **standard deviation** of the estimator. When referring to estimates, this is often called the *standard error* (SE).

The variance of an estimator describes how much the estimator fluctuates around its expected value. If $\hat{\theta}_n$ has high variance, different samples would give very different estimates. If it has low variance, the estimates would be tightly clustered around $E[\hat{\theta}_n]$. Note that if an estimator is unbiased, $E[\hat{\theta}_n] = \theta$, so $Var(\hat{\theta}_n) = E[(\hat{\theta}_n - \theta)^2]$ in that case (the variance equals the mean squared error, as we will formalize soon).

Example 4.9 (Variance of the Four Mean Estimators)**.** For the four estimators of μ in Example 4.4, we have:

$$\begin{aligned}
Var(\hat{\mu}_n^{(1)}) &= Var(0) = 0, \\
Var(\hat{\mu}_n^{(2)}) &= Var(X_1) = \sigma^2, \\
Var(\hat{\mu}_n^{(3)}) &= Var\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}, \\
Var(\hat{\mu}_n^{(4)}) &= Var\Big(\frac{1}{n+\lambda}\sum_{i=1}^{n} X_i\Big) = \frac{1}{(n+\lambda)^2}(n\sigma^2) = \frac{n\,\sigma^2}{(n+\lambda)^2}.
\end{aligned}$$

We used the fact that $Var(\sum_{i=1}^{n} X_i) = n\sigma^2$ for independent draws. (Strictly speaking, the last formula for $Var(\hat{\mu}_n^{(4)})$ is correct when X_i's are independent and λ is treated as fixed. If λ were random or depending on data, the formula would be more complex, but here λ is a fixed constant.)

A few observations: $\hat{\mu}_n^{(1)}$ (the always-0 estimator) has zero variance (it does not vary at all—it is completely certain in its guess, albeit wrong unless

$\mu = 0$). $\hat{\mu}_n^{(2)}$ (just one observation) has variance σ^2, the same as an individual X. $\hat{\mu}_n^{(3)}$ (sample mean) has variance σ^2/n, which is much smaller for large n—this reflects the benefit of averaging many observations, smoothing out the noise. $\hat{\mu}_n^{(4)}$ has variance $\frac{n}{(n+\lambda)^2}\sigma^2$. For large n, this is approximately σ^2/n (since $n/(n+\lambda)^2 \approx 1/n$ for large n). For small n, $\hat{\mu}_n^{(4)}$ actually has slightly smaller variance than $\hat{\mu}_n^{(3)}$ because $\frac{n}{(n+\lambda)^2} < \frac{1}{n}$ for any $\lambda > 0$. For example, if $n = 10$ and $\lambda = 1$, $Var(\hat{\mu}_n^{(3)}) = \sigma^2/10$ whereas $Var(\hat{\mu}_n^{(4)}) = 10\sigma^2/11^2 \approx 0.826\,(\sigma^2/10)$, about 17% smaller variance than $\hat{\mu}_n^{(3)}$. Of course, we saw earlier that $\hat{\mu}_n^{(4)}$ is biased, so we have a bias–variance trade-off at play.

Notably, $Var(\hat{\mu}_n^{(2)})$, $Var(\hat{\mu}_n^{(3)})$, and $Var(\hat{\mu}_n^{(4)})$ all involve the unknown σ^2 (the population variance). In practice, we might plug in an estimate of σ^2 (like $\hat{\sigma}_n^2$) to estimate the variances of these estimators. But conceptually, when comparing the theoretical performance of estimators, we treat σ^2 as given and see how the variances scale with n.

To visualize why considering both bias and variance is important, imagine the "sampling distribution" of an estimator as a target (with the true value as the bullseye). An estimator with low bias but high variance will on average hit near the bullseye, but any single shot might be far off (points widely scattered around the target). An estimator with low variance but high bias will hit close together, but consistently off-center from the bullseye. Ideally, we want both low bias and low variance.

Figure 4.1 provides an illustration of the sampling distributions for three of the estimators for μ discussed above (except the degenerate $\hat{\mu}_n^{(1)}$). In the figure, we assume $\mu = 1$ and $\sigma^2 = 1$ for the population, and $n = 10$. The blue curve is the distribution of $\hat{\mu}_n^{(2)}$ (just X_1), which is $N(1, 1)$ in this case (mean 1, variance 1) – fairly spread out. The red curve is the distribution of $\hat{\mu}_n^{(3)}$ (sample mean), which is $N(1, 0.1)$ (mean 1, variance 0.1), a much narrower distribution. The green curve is for $\hat{\mu}_n^{(4)}$ with $\lambda = 1$. If we account for bias exactly, $\hat{\mu}_n^{(4)}$ would be centered at $\frac{10}{11} \approx 0.909$ with variance 0.0826. In the figure, for simplicity of illustration, we have drawn it roughly as $N(1, 0.09)$ – essentially the same center as the others, but slightly smaller variance than the red curve. In reality, its distribution is shifted slightly left due to bias. The figure highlights how $\hat{\mu}_n^{(3)}$ and $\hat{\mu}_n^{(4)}$ concentrate more tightly around the true mean compared to $\hat{\mu}_n^{(2)}$.

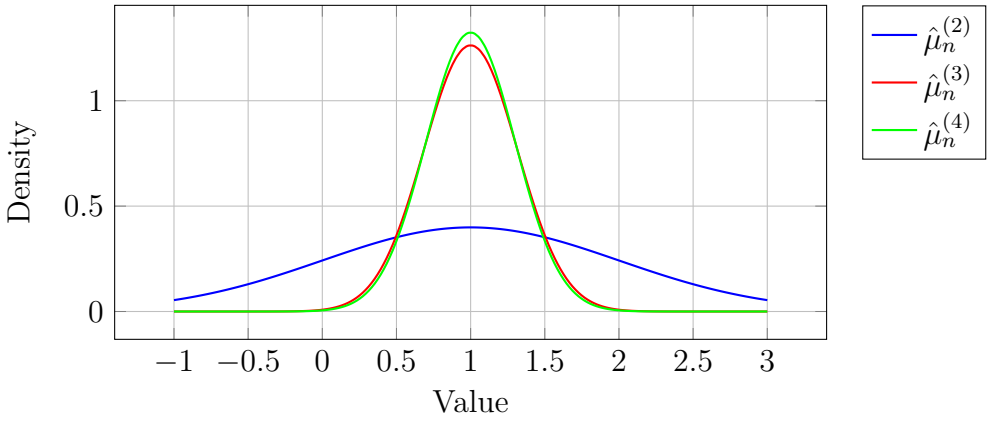

Figure 4.1: Sampling distributions of three estimators for μ, with $n = 10$, $\mu = 1$, $\sigma^2 = 1$. The blue distribution (widest) is for $\hat{\mu}_n^{(2)} = X_1$; the red (narrower) is for $\hat{\mu}_n^{(3)} = \bar{X}$; the green (narrowest) is for $\hat{\mu}_n^{(4)}$ with $\lambda = 1$. The biased estimator $\hat{\mu}_n^{(4)}$ is drawn here centered at 1 for visual comparison, though in reality its mean is slightly below 1. This figure illustrates that using more data (or shrinking the estimate) concentrates the estimator closer to the true value, reducing variance.

We see that $\hat{\mu}_n^{(3)}$ (sample mean) yields a much tighter distribution around the true μ than $\hat{\mu}_n^{(2)}$ does, without any bias. $\hat{\mu}_n^{(4)}$ has a tiny bias but an even tighter distribution in this example. To decide which estimator is preferable, we might consider a combined measure of error that accounts for both bias and variance. One such measure is the mean squared error, discussed next.

4.2.3 Mean Squared Error (MSE)

A commonly used overall measure of an estimator's quality is the **mean squared error**.

Definition 4.10 (Mean Squared Error)**.** The **mean squared error** of an estimator $\hat{\theta}_n$ for θ is defined as

$$MSE(\hat{\theta}_n) = E\Big[(\hat{\theta}_n - \theta)^2\Big].$$

MSE is the expected squared deviation of the estimator from the true parameter value. It captures in one number both the variance and the bias of

the estimator.

Expanding the square, we can express the MSE in terms of bias and variance:

Proposition 4.11 (Bias–Variance Decomposition)**.** *For any estimator* $\hat{\theta}_n$*,*

$$MSE(\hat{\theta}_n) = Bias(\hat{\theta}_n)^2 + Var(\hat{\theta}_n).$$

Proof. By definition $(\hat{\theta}_n - \theta)^2 = (\hat{\theta}_n - E[\hat{\theta}_n] + E[\hat{\theta}_n] - \theta)^2$. Expanding this, we get $(\hat{\theta}_n - E[\hat{\theta}_n])^2 + 2(\hat{\theta}_n - E[\hat{\theta}_n])(E[\hat{\theta}_n] - \theta) + (E[\hat{\theta}_n] - \theta)^2$. Taking expectation: $E[(\hat{\theta}_n - E[\hat{\theta}_n])^2] + 2(E[\hat{\theta}_n] - \theta)E[\hat{\theta}_n - E[\hat{\theta}_n]] + (E[\hat{\theta}_n] - \theta)^2$. The middle term is zero (because $E[\hat{\theta}_n - E[\hat{\theta}_n]] = 0$). So we have $Var(\hat{\theta}_n) + \left(E[\hat{\theta}_n] - \theta\right)^2 = Var(\hat{\theta}_n) + Bias(\hat{\theta}_n)^2$. □

Thus MSE combines the variance and the square of the bias. For an unbiased estimator, MSE $= Var(\hat{\theta}_n)$, since bias is zero. If an estimator has bias, MSE includes that penalty as well.

Example 4.12 (MSE Comparison)**.** Let's revisit the hypothetical estimators from a previous example to see the importance of the bias–variance trade-off. Consider two estimators of some parameter $\theta = 0$: - $\hat{\theta}_1$ takes value -100 or 100, each with probability 0.5. - $\hat{\theta}_2$ always takes the value 1 (with probability 1).

$\hat{\theta}_1$ is unbiased: $E[\hat{\theta}_1] = 0 \cdot 0.5 + 0 \cdot 0.5 = 0 = \theta$. $\hat{\theta}_2$ is biased: $E[\hat{\theta}_2] = 1 \neq 0$ (bias = 1). Now consider their MSE:

$$\begin{aligned} MSE(\hat{\theta}_1) &= E[(\hat{\theta}_1 - 0)^2] = 0.5 \cdot (-100)^2 + 0.5 \cdot (100)^2 \\ &= 0.5(10000) + 0.5(10000) = 10000, \\ MSE(\hat{\theta}_2) &= E[(\hat{\theta}_2 - 0)^2] = (1 - 0)^2 = 1. \end{aligned}$$

According to MSE, $\hat{\theta}_2$ is far superior (MSE 1 vs 10000), even though $\hat{\theta}_2$ is biased and $\hat{\theta}_1$ is unbiased. The intuition is clear: $\hat{\theta}_1$ occasionally makes enormous errors (off by 100), which is devastating for MSE, whereas $\hat{\theta}_2$ is always off by 1 (small but systematic error). In most practical situations, we would prefer an estimator like $\hat{\theta}_2$ (with small, possibly nonzero bias and small variance) over $\hat{\theta}_1$ (with zero bias but wildly high variance).

This toy example highlights that unbiasedness alone isn't a guarantee of a good estimator; one must account for variance as well.

The bias–variance trade-off is a fundamental concept. Often, by allowing a slight bias in an estimator, one can substantially reduce its variance, yielding a lower MSE. Conversely, forcing an estimator to be unbiased can sometimes result in higher variance. Figure 4.2 conceptually illustrates how the total error (MSE) can sometimes be minimized at an intermediate model complexity or estimator sophistication, where neither bias nor variance is extreme. (In machine learning, for instance, a very simple model has high bias/low variance, while a very flexible model has low bias/high variance; an intermediate model can minimize prediction error.)

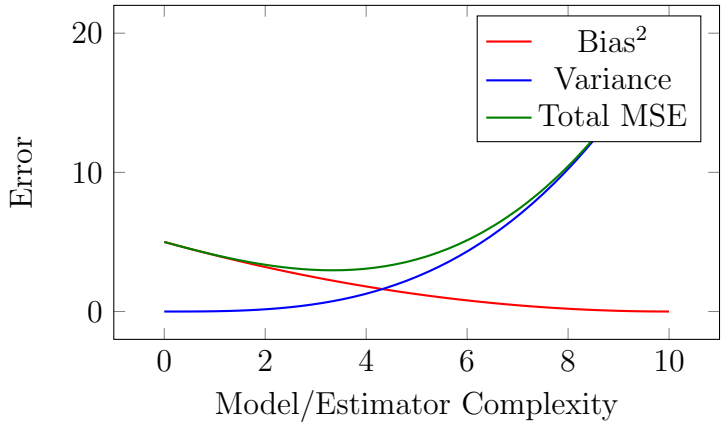

Figure 4.2: Illustration of Bias–Variance Trade-off. As the complexity of a model or estimator increases, bias typically decreases but variance increases. The total mean squared error is the sum of bias^2 and variance, and it is minimized at an intermediate level of complexity (neither too simple nor too flexible).

In the context of our earlier estimators for μ, $\hat{\mu}_n^{(2)}$ (one observation) and $\hat{\mu}_n^{(3)}$ (sample mean) were both unbiased, but $\hat{\mu}_n^{(3)}$ had much smaller variance, so it would be preferred (lower MSE). When we compared $\hat{\mu}_n^{(4)}$ (slightly biased) to $\hat{\mu}_n^{(3)}$, we saw that for $n = 10, \lambda = 1$ in a particular scenario, $\hat{\mu}_n^{(4)}$ had a slightly lower MSE than $\hat{\mu}_n^{(3)}$ (roughly 0.09 vs 0.10 in that example). That suggests $\hat{\mu}_n^{(4)}$ might be preferred if we only care about MSE. However, this advantage depends on the true parameter values (here μ and σ^2) and on λ. In general, when designing an estimator, one might tune a parameter (like λ) to minimize MSE.

Exercise 4.13. For the estimator $\hat{\mu}_n^{(4)}$ in Example 4.4, derive an expression for its MSE in terms of μ, σ^2, n, and λ. Find the value of λ (in terms of μ, σ^2, n) that minimizes this MSE. Discuss how the optimal λ depends on μ and σ^2. (Hint: You will find that the optimal λ is 0 if $\mu = 0$, but becomes larger as $|\mu|$ grows relative to σ.)

4.3 Large-Sample Properties of Estimators

In finite samples, we saw that bias and variance can depend on unknown population parameters, which can make it hard to definitively choose the "best" estimator without knowing θ. For example, in the last section:

- $Bias(\hat{\mu}_n^{(4)}) = -\frac{\lambda}{n+\lambda}\mu$ depends on μ,
- $Var(\hat{\mu}_n^{(2)})$ and $Var(\hat{\mu}_n^{(3)})$ depend on σ^2,
- $Var(\hat{\mu}_n^{(4)})$ depends on both μ and σ^2.

Without knowing μ or σ^2, we cannot directly compute or compare the exact MSEs of these estimators. This is a common dilemma: the performance of an estimator often depends on the very quantity it is trying to estimate (or other unknown aspects of the distribution).

Rather than asking "Which estimator is best for this fixed sample size n?" (which might require knowledge we don't have), we often ask a slightly different question: *Which estimator will perform best (or adequately well) in the long run as we gather more and more data?* In other words, we shift focus to asymptotic properties: what happens as $n \to \infty$? An estimator that performs arbitrarily well with enough data can be deemed a sensible choice, even if we cannot be sure it is optimal for small n.

In the large-sample analysis, two concepts are fundamental:

1. **Consistency:** Does $\hat{\theta}_n$ converge to the true θ as $n \to \infty$? (In probability, or almost surely, etc.)
2. **Asymptotic distribution:** If we rescale $\hat{\theta}_n$ appropriately, does it have a well-defined distributional limit as $n \to \infty$ (often a normal distribu-

tion)? This provides approximations to the sampling distribution for large n.

To discuss these, we first need to introduce the formal notions of convergence for random variables.

4.3.1 Convergence in Probability and Consistency

We start with the idea of a sequence of random variables converging in probability to some value. This concept parallels the usual definition of convergence for deterministic sequences, but now the "closeness" must happen with probability approaching 1.

Definition 4.14 (Convergence in Probability)**.** Let $X_1, X_2, \ldots$ be a sequence of random variables, and let X be another random variable (often a constant degenerate random variable). We say X_n **converges in probability to** X, written

$$X_n \xrightarrow{p} X,$$

if for every $\epsilon > 0$,

$$P(|X_n - X| > \epsilon) \to 0 \quad \text{as } n \to \infty.$$

In words, $X_n \xrightarrow{p} X$ means that the probability that X_n differs from X by more than an arbitrary small threshold ϵ goes to zero as n becomes large. Equivalently, for large n, X_n is very likely to be within an ϵ-neighborhood of X. Convergence in probability is sometimes called *weak convergence in probability* and is one of several modes of convergence (another common one is almost sure convergence, which is stronger, but we'll not delve into that here).

Now, applying this concept to estimators, we define consistency:

Definition 4.15 (Consistency)**.** An estimator $\hat{\theta}_n$ for parameter θ is **consistent** if $\hat{\theta}_n$ converges in probability to θ:

$$\hat{\theta}_n \xrightarrow{p} \theta \qquad (n \to \infty).$$

Thus, $\hat{\theta}_n$ is consistent if for any $\epsilon > 0$, no matter how small, eventually with enough data the estimator will be within ϵ of the true θ with high probability. Intuitively, a consistent estimator "homes in" on the right answer as n grows.

Consistency is often viewed as a minimal requirement for an estimator: an inconsistent estimator will not give the right answer even with infinite data, so it is usually not acceptable. If an estimator is not consistent, we typically discard it in favor of one that is (assuming one exists).

Example 4.16 (Consistency of Simple Estimators)**.** Consider again the four estimators of μ from Example 4.4:

- $\hat{\mu}_n^{(1)} = 0$ is *not* consistent for μ (unless μ really equals 0). For any $\epsilon < |\mu|$, we have $P(|\hat{\mu}_n^{(1)} - \mu| > \epsilon) = P(|\mu| > \epsilon)$, which is either 0 (if $\mu = 0$ or if $\epsilon > |\mu|$) or 1 (if $0 < \epsilon < |\mu|$). If $\mu \neq 0$, this probability does not go to 0 as $n \to \infty$; it is constantly 1. So $\hat{\mu}_n^{(1)} \not\xrightarrow{p} \mu$ for $\mu \neq 0$.
- $\hat{\mu}_n^{(2)} = X_1$ is also *not* consistent for μ. Since X_1 is an *iid* draw from distribution with mean μ, $P(|X_1 - \mu| > \epsilon)$ is some number less than 1 (assuming the distribution has some spread). For example, if $X_1 \sim N(\mu, \sigma^2)$, then $P(|X_1 - \mu| > \epsilon) = 2[1 - \Phi(\epsilon/\sigma)]$, which is > 0 for any fixed ϵ. As n increases, X_1 does not change (it's always the first observation), so $P(|X_1 - \mu| > \epsilon)$ does not tend to 0; it remains whatever it is. Thus X_1 is not converging to μ in probability. Intuitively, one data point does not "magically" become equal to the true mean just because we could have collected more data (but didn't use them).
- $\hat{\mu}_n^{(3)} = \frac{1}{n}\sum_{i=1}^n X_i$ *is* consistent for μ, under very general conditions (for example, finite mean suffices). This is guaranteed by the *Weak Law of Large Numbers*, stated below.
- $\hat{\mu}_n^{(4)} = \frac{1}{n+\lambda}\sum_{i=1}^n X_i$. We expect this to also be consistent, since it is very close to $\hat{\mu}_n^{(3)}$ for large n (the extra λ becomes negligible). We will show consistency of $\hat{\mu}_n^{(4)}$ after introducing a couple more tools.

The upshot: using no data or only a fixed number of data points generally gives inconsistent estimators. Using all the data and averaging in a reasonable way can yield consistency.

The most famous result establishing consistency of the sample mean (and analogues) is the Law of Large Numbers. We present the weak version (convergence in probability).

Theorem 4.17 (Weak Law of Large Numbers (WLLN))**.** *Let $X_1, X_2, \ldots, X_n$ be iid random variables with finite mean $E[X_i] = \mu$ (and finite variance, though finite mean is enough for the weak law). Then:*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mu.$$

That is, the sample average converges in probability to the true mean.

Proof (using Chebyshev's Inequality). First, $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \mu$ by linearity, and $Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} n Var(X_1) = \frac{\sigma^2}{n}$ if $Var(X_1) = \sigma^2 < \infty$. Now for any $\epsilon > 0$, by Chebyshev's inequality (which states $P(|Y - E[Y]| > \epsilon) \leq Var(Y)/\epsilon^2$ for any random variable Y with finite variance):

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \epsilon\right) \;\leq\; \frac{Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)}{\epsilon^2} \;=\; \frac{\sigma^2/n}{\epsilon^2} \;=\; \frac{\sigma^2}{n\epsilon^2}.$$

As $n \to \infty$, the right-hand side $\frac{\sigma^2}{n\epsilon^2} \to 0$. Thus $P(|\bar{X}_n - \mu| > \epsilon) \to 0$, which by definition means $\bar{X}_n \xrightarrow{p} \mu$. □

The Weak Law of Large Numbers formalizes the intuitive idea that averages become stable as sample size grows. In practice, it justifies using sample means to estimate expected values.

By the WLLN, $\hat{\mu}_n^{(3)} = \bar{X} \to \mu$ in probability, so $\hat{\mu}_n^{(3)}$ is consistent for μ. This is one reason the sample mean is so ubiquitous: it is a simple estimator that (under mild conditions) is unbiased and consistent for the population mean.

Now, what about $\hat{\mu}_n^{(4)} = \frac{1}{n+\lambda}\sum_{i=1}^{n} X_i$? We can express it as

$$\hat{\mu}_n^{(4)} = \frac{n}{n+\lambda} \cdot \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{n}{n+\lambda}\,\bar{X}_n.$$

We know $\bar{X}_n \xrightarrow{p} \mu$ (by WLLN). Also, the factor $\frac{n}{n+\lambda}$ is a sequence of constants (not random) that converges to 1 as $n \to \infty$. Thus $\frac{n}{n+\lambda} \to 1$. The product of

these two sequences should converge to $1 \cdot \mu = \mu$. This intuition is formalized by the Continuous Mapping Theorem (introduced below), but we can argue directly: for any $\epsilon > 0$,

$$|\hat{\mu}_n^{(4)} - \mu| = \left| \frac{n}{n+\lambda} \bar{X}_n - \mu \right| \leq \left| \frac{n}{n+\lambda} \bar{X}_n - \bar{X}_n \right| + |\bar{X}_n - \mu| = \left| \frac{\lambda}{n+\lambda} \right| \cdot |\bar{X}_n| + |\bar{X}_n - \mu|.$$

As $n \to \infty$, $\frac{\lambda}{n+\lambda} \to 0$ and typically $|\bar{X}_n|$ will not grow without bound (in fact, it converges in probability to a finite number μ). More rigorously, one can show $|\frac{\lambda}{n+\lambda} \bar{X}_n|$ converges to 0 in probability (since it's bounded by $\frac{\lambda}{n+\lambda}(|\mu| + |\bar{X}_n - \mu|)$ and that goes to 0 in probability). Meanwhile $|\bar{X}_n - \mu|$ goes to 0 in probability by WLLN. Hence the sum goes to 0 in probability. Therefore, $\hat{\mu}_n^{(4)} \xrightarrow{p} \mu$ as well. (A more straightforward method is to invoke the Continuous Mapping Theorem as we will soon.)

To proceed systematically for more complex scenarios, we introduce a couple of general results: one for joint convergence of multiple random quantities, and one for continuous transformations of convergent sequences.

4.3.2 Joint Convergence and Continuous Mapping

Often we deal with multiple estimators or multiple components simultaneously. For instance, $(\bar{X}, \widehat{Var}(X))$ might be a pair of estimators for $(E[X], Var(X))$. We might want to say this pair converges in probability to (μ, σ^2). The following result helps: it states that convergence in probability of each component implies convergence of the vector of components.

Theorem 4.18. *If $X_{1,n} \xrightarrow{p} \theta_1$, $X_{2,n} \xrightarrow{p} \theta_2$, ..., $X_{k,n} \xrightarrow{p} \theta_k$ as $n \to \infty$, then the random vector $\big(X_{1,n}, X_{2,n}, \ldots, X_{k,n}\big)$ converges in probability to $(\theta_1, \theta_2, \ldots, \theta_k)$. Symbolically:*

$$X_{j,n} \xrightarrow{p} \theta_j \text{ for each } j = 1, \ldots, k \qquad \Longrightarrow \qquad (X_{1,n}, \ldots, X_{k,n}) \xrightarrow{p} (\theta_1, \ldots, \theta_k).$$

We will not prove this formally (it follows from definitions and a union bound inequality). Intuitively, if each component gets arbitrarily close to the corresponding true value with high probability, then the whole collection will get close to the true vector with high probability.

Next, the Continuous Mapping Theorem (CMT) allows us to take a convergent sequence of random variables and apply a continuous function to it, preserving convergence.

Theorem 4.19 (Continuous Mapping Theorem)**.** *Let $g : \mathbb{R}^k \to \mathbb{R}^m$ be a continuous function, and suppose $\mathbf{X}_n$ is a random vector in $\mathbb{R}^k$ such that $\mathbf{X}_n \xrightarrow{p} \mathbf{a}$ (a constant vector in $\mathbb{R}^k$). Then*

$$g(\mathbf{X}_n) \xrightarrow{p} g(\mathbf{a}).$$

In particular, if $X_n \xrightarrow{p} a \in \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ is continuous at a, then $g(X_n) \xrightarrow{p} g(a)$.

The idea is straightforward: continuous functions preserve limits. The requirement that g is continuous (at the point in question) is important; if g has a discontinuity, convergence might not carry through (e.g., X_n might converge to a point where g jumps).

Example 4.20 (Applying CMT)**.** Suppose $A_n \xrightarrow{p} a$ and $B_n \xrightarrow{p} b$ for some constants a, b with $b \neq 0$. Consider the function $g(x, y) = \frac{x}{y}$. This g is continuous at (a, b) (assuming $b \neq 0$ so there's no division-by-zero issue, which is why we require $b \neq 0$). By the CMT,

$$\frac{A_n}{B_n} = g(A_n, B_n) \xrightarrow{p} g(a, b) = \frac{a}{b}.$$

Other simple corollaries: If $A_n \xrightarrow{p} a$ and $B_n \xrightarrow{p} b$, then $A_n + B_n \to a + b$, $A_n B_n \to ab$, $\sqrt{A_n} \to \sqrt{a}$ (assuming $a \geq 0$ so $\sqrt{\cdot}$ is continuous at a), and so on.

Using joint convergence (Theorem 4.18) and CMT, we can now easily establish the consistency of $\hat{\mu}_n^{(4)}$ and $\hat{\sigma}_n^2$, as promised:

Example 4.21 (Consistency of $\hat{\mu}_n^{(4)}$)**.** We had $\hat{\mu}_n^{(4)} = \frac{n}{n+\lambda} \cdot \frac{1}{n} \sum_{i=1}^n X_i = g(A_n, B_n)$ where $A_n = \frac{n}{n+\lambda}$ and $B_n = \frac{1}{n} \sum_{i=1}^n X_i$, and $g(a, b) = a \cdot b$. Here A_n is a sequence of (non-random) numbers converging to 1, and by WLLN $B_n \xrightarrow{p} \mu$. Thus by Theorem 4.18, $(A_n, B_n) \xrightarrow{p} (1, \mu)$. The function $g(a, b) = ab$ is continuous, so by CMT:

$$\hat{\mu}_n^{(4)} = A_n B_n \xrightarrow{p} 1 \cdot \mu = \mu.$$

Therefore, $\hat{\mu}_n^{(4)}$ is consistent for μ.

Example 4.22 (Consistency of $\hat{\sigma}_n^2$)**.** We can write the sample variance (with denominator n) in the form

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big)^2.$$

Define $A_n = \frac{1}{n}\sum_{i=1}^{n} X_i^2$ and $B_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}_n$. Then $\hat{\sigma}_n^2 = g(A_n, B_n)$ where $g(a,b) = a - b^2$. By the Law of Large Numbers, $A_n \xrightarrow{p} E[X^2]$ and $B_n \xrightarrow{p} E[X] = \mu$. Thus $(A_n, B_n) \xrightarrow{p} (E[X^2],\, \mu)$. The function $g(a,b) = a - b^2$ is continuous everywhere. Therefore, by CMT:

$$\hat{\sigma}_n^2 = A_n - (B_n)^2 \xrightarrow{p} E[X^2] - \mu^2 = Var(X) = \sigma^2.$$

So the (uncorrected) sample variance is a consistent estimator of the true variance. And since multiplying by the constant $\frac{n}{n-1} \to 1$ (which is continuous in the limit) would not affect convergence, the unbiased sample variance $\tilde{\sigma}_n^2 = \frac{n}{n-1}\hat{\sigma}_n^2$ is also consistent for σ^2.

4.3.3 Asymptotic Distribution and the Central Limit Theorem

Consistency tells us that the estimator eventually hits the bullseye (in probability). But it does not tell us how *fast* it converges, nor does it give a way to quantify the uncertainty remaining when n is large but finite. For that, we study the *asymptotic distribution* of $\hat{\theta}_n$ after appropriate scaling or centering. The most celebrated result in this realm is the Central Limit Theorem (CLT), which describes the distribution of the sum or average of iid random variables for large n.

Theorem 4.23 (Central Limit Theorem (Lindeberg–Lévy form))**.** *Let $X_1, \ldots, X_n$ be iid with mean $\mu = E[X_i]$ and variance $0 < \sigma^2 = Var(X_i) < \infty$. Then*

$$\sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\Big) \xrightarrow{d} N(0, \sigma^2),$$

i.e. the normalized deviation of the sample mean from the true mean converges in distribution to a normal distribution with mean 0 and variance σ^2. Equivalently,

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1).$$

Here $\xrightarrow{d}$ denotes convergence in distribution (also called convergence in law). We define it as follows:

Definition 4.24 (Convergence in Distribution)**.** A sequence of random variables X_n is said to **converge in distribution** to a random variable X (written $X_n \xrightarrow{d} X$) if

$$\lim_{n\to\infty} P(X_n \leq t) = P(X \leq t)$$

for every real number t at which $P(X \leq t)$ is continuous in t. (In other words, the CDFs $F_{X_n}(t)$ converge pointwise to the CDF $F_X(t)$ at all continuity points of F_X.)

Convergence in distribution basically means the distribution of X_n gets closer and closer to that of X. In the CLT, the distribution of the standardized average tends to the standard normal distribution.

The CLT is remarkable because it holds regardless of the distribution of X_i (as long as the variance is finite): the bell curve emerges as a universal approximation. This justifies why so many statistical procedures assume normality for large samples.

Example 4.25 (Asymptotic Distribution of $\bar{X}$)**.** Applying the CLT result to $\hat{\mu}_n^{(3)} = \bar{X}_n$, we have:

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2).$$

Equivalently, for large n,

$$\bar{X}_n \approx N\Big(\mu,\ \frac{\sigma^2}{n}\Big).$$

In other words, the sampling distribution of the sample mean is approximately normal with mean μ and variance σ^2/n. This is an approximation that improves as n grows.

Typically, we don't know σ^2, but we can estimate it (for example, using $\hat{\sigma}_n^2$). If we plug the estimate in, we get

$$\bar{X}_n \approx N\Big(\mu,\ \frac{\hat{\sigma}_n^2}{n}\Big).$$

For large n, $\hat{\sigma}_n^2$ is close to σ^2 (by consistency), so this is a reasonable approximation. The quantity $\frac{\hat{\sigma}_n}{\sqrt{n}}$ is called the *standard error* of $\bar{X}_n$. We will formalize this shortly.

The CLT can be extended to vector-valued averages as well (the *multivariate CLT*). For instance, if we have two sample means (say sample mean of Y and sample mean of X in a bivariate sample), their joint distribution is asymptotically bivariate normal.

Theorem 4.26 (Multivariate Central Limit Theorem (Bivariate case))**.** *Let (Y_i, X_i), $i = 1, \ldots, n$ be iid bivariate random vectors with $E[Y_i] = \mu_Y$, $E[X_i] = \mu_X$, $Var(Y_i) = \sigma_Y^2$, $Var(X_i) = \sigma_X^2$, and $Cov(Y_i, X_i) = \sigma_{YX}$. Then as $n \to \infty$:*

$$\sqrt{n}\Big((\bar{Y}_n - \mu_Y),\ (\bar{X}_n - \mu_X)\Big) \xrightarrow{d} N\Big((0,0),\ \Sigma\Big),$$

where Σ is the 2×2 covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_Y^2 & \sigma_{YX} \\ \sigma_{YX} & \sigma_X^2 \end{pmatrix}.$$

That is, the vector $(\bar{Y}_n, \bar{X}_n)$ is asymptotically bivariate normal with mean (μ_Y, μ_X) and covariance matrix Σ/n.

A similar statement holds for any fixed-dimensional vector of sample averages. This result allows us to analyze linear combinations of estimates as well as ratios, etc., in large samples by leveraging properties of the multivariate normal.

Often we apply a linear combination to the result of a multivariate CLT. For example, if we are interested in $\bar{Y}_n - \bar{X}_n$, that is a linear combination $(1, -1)$ of $(\bar{Y}_n, \bar{X}_n)$. From the above theorem, one can show

$$\sqrt{n}\big((\bar{Y}_n - \mu_Y) - (\bar{X}_n - \mu_X)\big) = \sqrt{n}\big((\bar{Y}_n - \bar{X}_n) - (\mu_Y - \mu_X)\big) \xrightarrow{d} N(0,\ \sigma_Y^2 + \sigma_X^2 - 2\sigma_{YX}),$$

since the variance of $Y - X$ is $Var(Y - X) = \sigma_Y^2 + \sigma_X^2 - 2\sigma_{YX}$. We can derive this formally using a "bivariate Slutsky theorem" which parallels the univariate one we will discuss momentarily.

Example 4.27 (Difference in Means)**.** Suppose we have iid data (Y_i, D_i) where $D_i \in \{0, 1\}$ is an indicator (e.g. D_i might denote treatment status and Y_i an outcome). Let $p = P(D = 1)$ (assume $0 < p < 1$). Consider the estimators:

$$\bar{Y}_1 = \frac{1}{n}\sum_{i=1}^{n} Y_i D_i, \qquad \bar{Y}_0 = \frac{1}{n}\sum_{i=1}^{n} Y_i(1 - D_i).$$

These are the average of Y among those with $D = 1$ (times the proportion of ones) and the average of Y among those with $D = 0$ (times the proportion of zeros) respectively. More precisely, $\bar{Y}_1 = \frac{N_1}{n} \cdot \overline{Y|D=1}$ and $\bar{Y}_0 = \frac{N_0}{n} \cdot \overline{Y|D=0}$, where $N_1 = \sum D_i$, $N_0 = n - N_1$. But as n large, $N_1/n \approx p$ and $N_0/n \approx 1-p$ (in fact $N_1/n \to p$ in probability by LLN as well, since N_1/n is just the sample mean of D_i which has expectation p). So $\bar{Y}_1$ and $\bar{Y}_0$ are essentially $p \cdot \overline{Y|D=1}$ and $(1-p) \cdot \overline{Y|D=0}$.

What is the asymptotic joint distribution of $(\bar{Y}_1, \bar{Y}_0)$? This is a bit tricky because $\bar{Y}_1$ and $\bar{Y}_0$ are not independent (they share the overall sample and if one group has more Y, the other might have less, etc.). However, using the multivariate CLT on the bivariate vector $(YD,\ Y(1-D))$, we get:

$$\sqrt{n} \begin{pmatrix} \bar{Y}_1 - E[YD] \\ \bar{Y}_0 - E[Y(1-D)] \end{pmatrix} \xrightarrow{d} N\Big((0,0),\ \Sigma\Big),$$

where

$$E[YD] = p \cdot E[Y|D=1], \qquad E[Y(1-D)] = (1-p) \cdot E[Y|D=0],$$

and Σ can be written in terms of the variances and covariance of YD and $Y(1-D)$.

Now, if the goal is to estimate $E[Y|D=1] - E[Y|D=0]$ (the difference in means between the two groups), a natural estimator is $\frac{\bar{Y}_1}{\bar{D}} - \frac{\bar{Y}_0}{1-\bar{D}}$, where $\bar{D} = N_1/n$ is the sample proportion of $D = 1$. But a simpler (asymptotically equivalent) estimator is $\frac{1}{p}\bar{Y}_1 - \frac{1}{1-p}\bar{Y}_0$ (plugging in the true p). Either way, one can find (using a delta method or Slutsky argument) that:

$$\sqrt{n}\Big[\big(\overline{Y|D=1} - \overline{Y|D=0}\big) - \big(E[Y|D=1] - E[Y|D=0]\big)\Big] \xrightarrow{d} N\Big(0,\ \frac{Var(YD)}{p^2} + \frac{Var(Y(1-D))}{(1-p)^2} - \frac{2Cov(YD,\ Y(1-D))}{p(1-p)}\Big).$$

This looks complicated, but it simplifies because YD and $Y(1-D)$ never take nonzero values simultaneously for a given observation (if $D = 1$, then $Y(1-D) = 0$; if $D = 0$, then $YD = 0$). Under some algebra, one can show the asymptotic variance is $\frac{Var(Y|D=1)}{p} + \frac{Var(Y|D=0)}{1-p}$. The key takeaway: the difference in group means is asymptotically normal. We will often use such results to do inference on treatment effects.

The calculations for multi-component estimators can become intricate, but conceptually we handle them by combining CLT with continuous transformations (via CMT) or linear combinations (via Slutsky's theorem, next).

4.3.4 Slutsky's Theorem

Slutsky's theorem is a handy result that complements the CLT. It allows us to replace unknown parameters in the asymptotic distribution with consistent estimators, without affecting the limit. It also handles adding or multiplying asymptotically negligible quantities.

Theorem 4.28 (Slutsky's Theorem)**.** *Suppose $A_n \xrightarrow{d} A$ and $B_n \xrightarrow{p} b$, where A is a random variable and b is a constant. Then:*

1. $A_n + B_n \xrightarrow{d} A + b$,
2. $A_n B_n \xrightarrow{d} A \cdot b$,
3. *If $b \neq 0$, then* $\frac{A_n}{B_n} \xrightarrow{d} \frac{A}{b}$.

In words, if B_n converges in probability to a constant b, then asymptotically we can treat B_n like b when looking at the distribution of A_n. The intuition is that B_n is so tightly concentrated around b for large n that the randomness in B_n doesn't contribute in the limit.

One common use of Slutsky's theorem is to plug in a consistent estimator for an unknown variance or standard deviation in the CLT. For example, we might not know σ, but we know $\hat{\sigma}_n \xrightarrow{p} \sigma$. The CLT gave $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \to_d N(0,1)$. We can replace σ with $\hat{\sigma}_n$ in the denominator, since $\hat{\sigma}_n/\sigma \to 1$. Formally:

$$\frac{\bar{X}_n - \mu}{\hat{\sigma}_n/\sqrt{n}} = \frac{\hat{\sigma}_n}{\sigma} \cdot \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} 1 \cdot N(0,1) = N(0,1),$$

because $\frac{\hat{\sigma}_n}{\sigma} \xrightarrow{p} 1$ and $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)$. We used Slutsky's theorem with $A_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ (which converges in distribution to $N(0,1)$) and $B_n = \frac{\hat{\sigma}_n}{\sigma}$ (which converges in probability to 1). Thus:

$$\frac{\bar{X}_n - \mu}{\hat{\sigma}_n/\sqrt{n}} \xrightarrow{d} N(0,1).$$

This result is fundamental because it justifies using $\hat{\sigma}_n$ (sample standard deviation) in place of σ when constructing confidence intervals or test statistics for the mean.

Example 4.29 (Asymptotic Normal t-Statistic)**.** Continuing the above reasoning: define the statistic

$$T_n = \frac{\bar{X}_n - \mu}{\hat{\sigma}_n / \sqrt{n}},$$

which looks like a Student's t-statistic for testing $H_0 : \mu = \mu_0$ (in practice, μ might be replaced by μ_0 if testing a hypothesis). Even though T_n does not exactly follow a t-distribution unless X_i are normal, Slutsky's theorem shows that $T_n \xrightarrow{d} N(0,1)$ as $n \to \infty$ for *any* distribution of the X_i (with finite variance). Thus for large n, we can use the normal approximation for T_n. This is why large-sample procedures often do not require strict normality assumptions—the CLT kicks in.

Example 4.30 (Asymptotic Normality of $\hat{\mu}_n^{(4)}$)**.** Let's apply these ideas to the slightly biased estimator $\hat{\mu}_n^{(4)} = \frac{n}{n+\lambda}\bar{X}_n$. We want the asymptotic distribution of $\sqrt{n}(\hat{\mu}_n^{(4)} - \mu)$. We can write:

$$\sqrt{n}(\hat{\mu}_n^{(4)} - \mu) = \sqrt{n}\Big(\frac{n}{n+\lambda}\bar{X}_n - \mu\Big) = \underbrace{\frac{n}{n+\lambda}}_{\to 1}\sqrt{n}(\bar{X}_n - \mu) + \sqrt{n}\Big(\frac{n}{n+\lambda} - 1\Big)\mu.$$

The second term on the right is

$$\begin{aligned}\sqrt{n}\Big(\frac{n}{n+\lambda} - 1\Big)\mu &= \sqrt{n}\Big(\frac{n-(n+\lambda)}{n+\lambda}\Big)\mu = \sqrt{n}\Big(\frac{-\lambda}{n+\lambda}\Big)\mu \\ &= -\frac{\lambda}{1+\lambda/n}\sqrt{n}\frac{1}{n}\mu = -\frac{\lambda}{1+\lambda/n}\cdot\frac{\mu}{\sqrt{n}}.\end{aligned}$$

As $n \to \infty$, $-\frac{\lambda}{1+\lambda/n} \to -\lambda$, but importantly it is multiplied by $\frac{\mu}{\sqrt{n}}$ which goes to 0. In fact, $\sqrt{n}(\frac{n}{n+\lambda} - 1)\mu \to 0$ as $n \to \infty$. We can formalize this: $\sqrt{n}(\frac{n}{n+\lambda} - 1)\mu$ converges in probability to 0 (actually it goes to 0 deterministically). So it is asymptotically negligible.

The first term is $\frac{n}{n+\lambda}\sqrt{n}(\bar{X}_n - \mu)$. We have $\frac{n}{n+\lambda} \to 1$ and by CLT $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$. By Slutsky's theorem,

$$\frac{n}{n+\lambda}\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} 1 \cdot N(0,\sigma^2) = N(0,\sigma^2).$$

Now combining both terms:

$$\sqrt{n}(\hat{\mu}_n^{(4)} - \mu) = \frac{n}{n+\lambda}\sqrt{n}(\bar{X}_n - \mu) + \underbrace{\sqrt{n}\Big(\frac{n}{n+\lambda} - 1\Big)\mu}_{\to 0}.$$

The second term goes to 0 in probability, the first term converges in distribution to $N(0, \sigma^2)$. Slutsky's theorem (in a slightly extended form to sum of two sequences, where one goes to a constant) implies that the sum converges in distribution to $N(0, \sigma^2)$ as well (since adding a vanishing term doesn't change the limit). Thus we conclude:

$$\sqrt{n}(\hat{\mu}_n^{(4)} - \mu) \xrightarrow{d} N(0, \sigma^2).$$

As expected, $\hat{\mu}_n^{(4)}$ is asymptotically equivalent to $\hat{\mu}_n^{(3)}$—the small bias does not affect the first-order asymptotic distribution (it only appears as a $O(1/\sqrt{n})$ term which vanishes in the limit).

The above example shows an important principle: if an estimator's bias is $o(n^{-1/2})$ (smaller order than $1/\sqrt{n}$), then it will typically have the same asymptotic distribution as an unbiased or consistent estimator, meaning the bias is asymptotically negligible.

Now that we have asymptotic normality for many estimators (like sample means, differences in means, regression coefficients under suitable conditions, etc.), we can construct approximate confidence intervals and perform tests.

4.3.5 Standard Errors and Confidence Intervals

When we say an estimator $\hat{\theta}_n$ is asymptotically normal, we mean

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \Sigma),$$

for some variance Σ (which might depend on unknown parameters). Equivalently,

$$\hat{\theta}_n \approx N\Big(\theta,\ \frac{\Sigma}{n}\Big)$$

for large n. In practice, we estimate Σ by some $\hat{\Sigma}_n$ (a consistent estimator), and define the *standard error* of $\hat{\theta}_n$ as $\mathrm{se}(\hat{\theta}_n) = \sqrt{\hat{\Sigma}_n/n}$. More formally:

Definition 4.31 (Standard Error)**.** If $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \Sigma)$ and $\hat{\Sigma}_n$ is a consistent estimator of Σ, then

$$\mathrm{se}(\hat{\theta}_n) = \sqrt{\hat{\Sigma}_n / n}$$

is called the (estimated) **standard error** of $\hat{\theta}_n$. It is the estimated standard deviation of the sampling distribution of $\hat{\theta}_n$. Often we report $\hat{\theta}_n \pm 1.96\,\mathrm{se}(\hat{\theta}_n)$ as an approximate 95% confidence interval for θ.

For the sample mean example, $\Sigma = \sigma^2$ and $\hat{\Sigma}_n = \hat{\sigma}_n^2$, so $\mathrm{se}(\bar{X}_n) = \hat{\sigma}_n/\sqrt{n}$. For a difference in means, Σ would be the sum of variances as derived earlier, and $\hat{\Sigma}$ would plug in sample estimates of those variances.

Given asymptotic normality, constructing confidence intervals is straightforward:

If $\frac{\hat{\theta}_n - \theta}{\mathrm{se}(\hat{\theta}_n)} \xrightarrow{d} N(0,1)$, then approximately

$$P\Big(-z_{1-\alpha/2} \le \frac{\hat{\theta}_n - \theta}{\mathrm{se}(\hat{\theta}_n)} \le z_{1-\alpha/2} \Big) \approx 1 - \alpha,$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution. Rearranging,

$$P\Big(\hat{\theta}_n - z_{1-\alpha/2}\,\mathrm{se}(\hat{\theta}_n) \le \theta \le \hat{\theta}_n + z_{1-\alpha/2}\,\mathrm{se}(\hat{\theta}_n) \Big) \approx 1 - \alpha.$$

This leads to the $(1-\alpha)$ confidence interval:

$$C_n = \Big[\ \hat{\theta}_n \ \pm\ z_{1-\alpha/2}\,\mathrm{se}(\hat{\theta}_n)\ \Big].$$

Similarly, a one-sided upper $(1-\alpha)$ confidence bound is $\big(-\infty,\ \hat{\theta}_n + z_{1-\alpha}\,\mathrm{se}(\hat{\theta}_n)\big]$, etc.

More formally:

Theorem 4.32 (Asymptotic Confidence Interval)**.** *Under the assumptions that* $\frac{\hat{\theta}_n - \theta}{se(\hat{\theta}_n)} \xrightarrow{d} N(0,1)$*, the interval*

$$C_n = \Big[\ \hat{\theta}_n - z_{1-\alpha/2}\,se(\hat{\theta}_n), \quad \hat{\theta}_n + z_{1-\alpha/2}\,se(\hat{\theta}_n)\ \Big]$$

is an asymptotically $(1-\alpha)$ *confidence interval for* θ*. That is,*

$$\lim_{n\to\infty} P(\theta \in C_n) = 1-\alpha.$$

One-sided variants can be constructed similarly.

In words, for large n, the random interval C_n will contain the true parameter θ about $(1-\alpha)\times 100\%$ of the time in repeated samples.

It is crucial to interpret this correctly: before we collect data, C_n is random and θ is fixed, so $P(\theta \in C_n) = 1-\alpha$ (approximately). After we collect data and compute c_n (a specific numeric interval), either θ is in that interval or not (so the probability is 0 or 1 for that event, but we just don't know). We *do not* say "there is a 95% probability that θ lies in the interval $[a,b]$" once $[a,b]$ is computed; rather, we say "this procedure yields intervals that cover θ 95% of the time in the long run". In practice, people often loosely say "95

Interpretation: Estimators vs Estimates

Before concluding, let us emphasize the distinction between the random estimator and its realized value:

- An **estimator** $\hat{\theta}_n$ is a random variable, a function of the sample. We can talk about its distribution, variance, bias, etc., and make probability statements (like $P(|\hat{\theta}_n - \theta| < \epsilon)$).

- An **estimate** is the realized number you get after plugging in your actual observed data into the estimator. Once you have an estimate (a concrete number), it is no longer random from the perspective of the analysis (the randomness has "collapsed" to that single outcome). You cannot meaningfully assign a probability to a fixed estimate being right or wrong; the probability was in the process that led to it.

This is especially important for confidence intervals. The confidence level (say 95%) refers to the procedure, not to the specific interval you obtained. Over many hypothetical repetitions of the experiment and interval calculations, 95% of those intervals would contain θ. For the one interval you got, either it contains θ or it doesn't; there is no probability attached to that anymore (unless you bring in a Bayesian perspective and treat θ as random, which is a different framework).

Example 4.33 (Misinterpretation of Confidence Intervals)**.** Suppose a student takes the GRE test and scores 153 on the Quantitative section. The ETS reports that the standard error of measurement for an individual GRE Quant score is about 2.2 points. A 95% confidence interval for the test taker's "true" ability (the score they would get on average if they took many equivalent test forms) might be calculated as $153 \pm 1.96(2.2)$, which is approximately $[149, 157]$.

ETS might report this as: "We can be 95% confident that the test taker's true Quantitative score is between 149 and 157." This phrasing, while common, is slightly misleading. The correct interpretation is: "If many test takers with the same true ability took the test, 95% of them would score between 149 and 157." Or: "For 95% of test takers, the interval [score $\pm$ 4] will contain their true score."

For a specific individual's true score, it's either in $[149, 157]$ or not. We can't assign a probability to that after the fact (in the frequentist framework) because the true score is not random. The 95% refers to the success rate of the method, not a probability for this particular interval.

It's a subtle point that often confuses people (even official guides sometimes state it in the colloquial way that sounds like a probability about the parameter). The takeaway: Confidence intervals have a confidence level that describes the process, not the realized interval.

Summary

In this chapter, we introduced the concept of estimators and their key properties:

- We learned that an estimator is a function of sample data used to infer a population parameter. We saw examples of estimators derived via the sample analogue principle (e.g., sample mean for population mean, sample proportion for probability, etc.) and via optimization (least squares).

- We discussed finite-sample properties: - **Bias**: the difference between an estimator's expectation and the true parameter. Unbiased estimators have zero bias. - **Variance**: the dispersion of the estimator's sampling distribu-

tion. Together with bias, it determines reliability. - **Mean Squared Error (MSE)**: = $\text{Bias}^2 + \text{Var}$. It provides a single measure combining accuracy and precision. We saw the bias-variance trade-off: sometimes a small bias can greatly reduce variance and improve MSE.

- We introduced the notion of a **sampling distribution** of an estimator and illustrated how different estimators for the same parameter can have different distributions.

- We then moved to large-sample properties: - **Consistency**: an estimator that converges in probability to the true parameter as $n \to \infty$. This is a minimal requirement for an estimator to be useful. (No matter how large the sample, an inconsistent estimator won't get arbitrarily close to the truth.) - The **Law of Large Numbers** guarantees consistency of sample means (and analogous estimators). - We defined convergence in probability and used tools like **Continuous Mapping Theorem** to find limits of transformations of consistent estimators. - **Asymptotic (large-sample) distribution**: We defined convergence in distribution and presented the **Central Limit Theorem**, which shows sample means (and many other estimators) are approximately normally distributed for large n. This allowed us to approximate the sampling distribution without knowing the exact finite-sample distribution. - We introduced the **Delta Method** (via an example) as a way to get the asymptotic distribution of a smooth function of an estimator (using a Taylor expansion). - **Slutsky's Theorem** was presented to justify plugging in consistent estimates into asymptotic results (e.g., replacing unknown variance by sample variance in the CLT). - We extended CLT to multiple dimensions and saw how linear combinations of jointly normal estimates are normal (this is used, for example, in differences of means or regression).

- Armed with asymptotic normality, we defined **standard errors** as estimates of the standard deviation of an estimator, and we constructed **confidence intervals** using the normal approximation. We stressed the correct interpretation of confidence intervals.

The big picture is that under quite general conditions, even if we cannot derive the exact sampling distribution of an estimator, we can often rely on large-sample approximations (normal theory) to conduct inference. In the next chapter, we will leverage these ideas to formulate and test hypotheses about parameters (for example, testing if a parameter equals some value, or if two parameters are different, etc.), which is the realm of hypothesis testing.