Chapter 6
An Introduction to Large Sample Asymptotics

6.1 Introduction

For **inference** (**confidence intervals** and **hypothesis testing**) on unknown parameters we need **sampling distributions** (either **exact** or **approximate**) of estimates and other statistics.

In Chapter 4, in the context of the **linear regression model**, we derived the **mean** and **variance** of the OLS estimator.
This is not a complete description of the sampling distribution and is not sufficient for inference.
The theory does not apply in the context of the linear projection model, which is more relevant for empirical applications.

In Chapter 5, in the context of the **normal regression model**, we derived the **exact sampling distribution** of the OLS estimator, t-statistics, and F-statistics, which are allowing for inference.
But these results are narrowly confined to the normal regression model, which requires the unrealistic assumption that the regression **error** is **normally distributed** and **independent of the regressors**.
Perhaps we can view these results as some sort of approximation to the sampling distributions without requiring the assumption of normality, but how can we be precise about this?

To illustrate the situation with an example, let y_i and x_i be drawn from the joint density

$$f(x,y) = \frac{1}{2\pi xy} \exp\left(-\frac{1}{2}(\log y - \log x)^{\mathbf{2}}\right) \exp\left(-\frac{1}{2}(\log x)^{\mathbf{2}}\right)$$

and let beta^ be the slope coefficient estimate from a least-squares regression of y_i on x_i and a constant. Using simulation methods, the density function of beta^ was computed and plotted in Figure 6.1 for sample sizes of n = 25; n = 100 and n = 800: The vertical line marks the true projection coefficient.

From the figure we can see that the density functions are **dispersed** and **highly non-normal**. As the **sample size increases** the density becomes **more concentrated about the population coefficient**. Is there a simple way to characterize the sampling distribution of beta^?

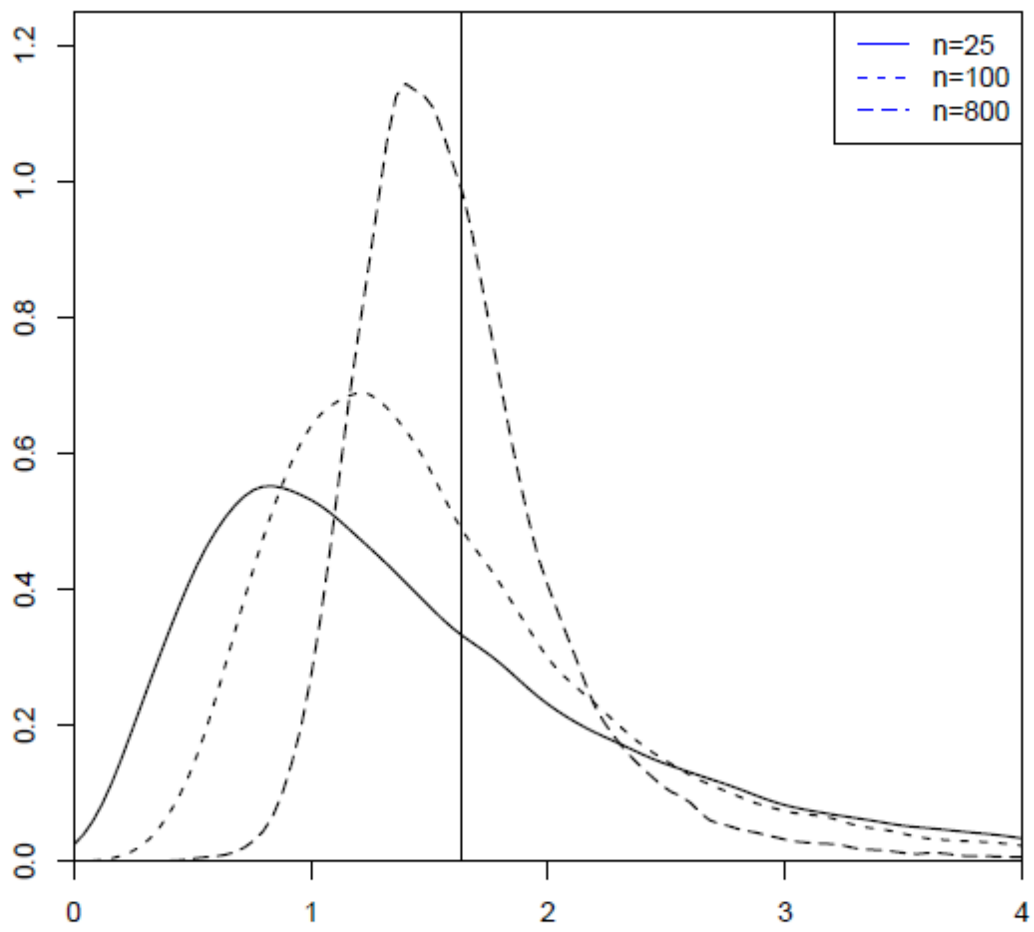

Figure 6.1: Sampling Density of $\widehat{\beta}$

In principle the sampling distribution of beta^ is a function of the joint distribution of (y_i; x_i) and the sample size n.
In practice this function is extremely complicated so **it is not feasible to analytically calculate the exact distribution** of beta^ except in very special cases. We typically rely on **approximation methods**.

In this chapter we introduce asymptotic theory, which **approximates by taking the limit of the finite sample distribution as the sample size n tends to infinity**.
It is important to understand that this is an approximation technique, as the asymptotic distributions are used to assess the finite sample distributions of our estimators in actual practical samples.

The primary tools of asymptotic theory are the **weak law of large numbers** (WLLN), **central limit theorem** (CLT), and **continuous mapping theorem** (CMT). With these tools we can approximate the sampling distributions of most econometric estimators.

6.2 Asymptotic Limits

Asymptotic analysis is a method of approximation obtained by taking a suitable limit.

There is more than one method to take limits, but the most common is to take the limit of the sequence of sampling distributions as the sample size tends to positive infinity, written as n → infinity. It is not meant to be interpreted literally, but rather as an approximating device.

The first building block for asymptotic analysis is the concept of a limit of a sequence.

Definition 6.1 A sequence a_n has the **limit** a, written $a_n \longrightarrow a$ as $n \to \infty$, or alternatively as $\lim_{n\to\infty} a_n = a$, if for all $\delta > 0$ there is some $n_\delta < \infty$ such that for all $n \geq n_\delta$, $|a_n - a| \leq \delta$.

In words, a_n has the limit a if the sequence gets closer and closer to a as n gets larger.

If a sequence has a limit, that limit is unique (a sequence cannot have two distinct limits). If an has the limit a; we also say that a_n converges to a as n → infinity.

Not all sequences have limits. For example, the sequence $\{1, 2, 1, 2, 1, 2, \ldots\}$ does not have a limit.

It is sometimes useful to have a more general definition of limits which always exist, and these are the limit superior and limit inferior of a sequence.

Definition 6.2 $\liminf_{n\to\infty} a_n \stackrel{def}{=} \lim_{n\to\infty} \inf_{m\geq n} a_m$

Definition 6.3 $\limsup_{n\to\infty} a_n \stackrel{def}{=} \lim_{n\to\infty} \sup_{m\geq n} a_m$

The limit inferior and limit superior always exist (including + and - infinity as possibilities), and **equal when the limit exists**. In the example given earlier, the limit inferior of $\{1, 2, 1, 2, 1, 2, \ldots\}$ is 1, and the limit superior is 2.

6.3 Convergence in Probability

For a sequence of random variables, we cannot directly apply the deterministic concept of a sequence of numbers. Instead, we require a definition of convergence which is appropriate for random variables. There are more than one such definition, but the most commonly used is called convergence in probability.

Definition 6.4 A random variable $z_n \in \mathbb{R}$ **converges in probability** to z as $n \to \infty$, denoted $z_n \xrightarrow{p} z$, or alternatively $\text{plim}_{n\to\infty} z_n = z$, if for all $\delta > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(|z_n - z| \leq \delta\right) = 1. \tag{6.1}$$

We call z the **probability limit** (or **plim**) of z_n.

Equation (6.1) states that P(|z_n - z|<= delta) approaches 1 as the sample size n increases. The definition of convergence in probability requires that this holds for any delta.

You may notice that the definition concerns **the distribution of the random variables** z_n, **not their realizations**. Furthermore, notice that the definition uses the concept of a conventional (deterministic) limit, but the latter is applied to a sequence of probabilities, not directly to the random variables z_n or their realizations.

A common mistake is to confuse convergence in probability with convergence in expectation:

$$\mathbb{E}(z_n) \longrightarrow \mathbb{E}(z).$$

They are related but distinct concepts. Neither convergence in probability nor convergence in expectation implies the other.

6.4 Weak Law of Large Numbers

In large samples, we expect parameter estimates to be close to the population values.

For example, in Section 4.3, we saw that the sample mean y^bar is unbiased for mu = E(y) and has variance sigma^2/n. As n gets large its variance decreases and thus the distribution of y concentrates about the population mean. It turns out that **this implies that the sample mean converges in probability to the population mean**.

When y has a **finite variance** there is a fairly straightforward proof by applying Chebyshev's inequality.

Theorem 6.1 Chebyshev's Inequality. For any random variable z_n and constant $\delta > 0$

$$\mathbb{P}\left(\left|z_n - \mathbb{E}\left(z_n\right)\right| \geq \delta\right) \leq \frac{\operatorname{var}(z_n)}{\delta^2}.$$

Chebyshev's inequality is terrifically important in asymptotic theory.

Applied to the sample mean y^bar which has variance sigma^2/n , Chebyshev's inequality shows that for delta,

$$\lim_{n\to\infty} \mathbb{P}\left(|\overline{y} - \mu| < \delta\right) = 1.$$

This result is called the weak law of large numbers. Our derivation assumed that y has a **finite variance**, but with a more careful derivation all that is necessary is a **finite mean**.

Theorem 6.2 Weak Law of Large Numbers (WLLN)
If y_i are independent and identically distributed and $\mathbb{E}|y| < \infty$, then as $n \longrightarrow \infty$

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \xrightarrow{p} \mathbb{E}(y).$$

The proof of Theorem 6.2 is presented in Section 6.26. Technically, the assumption that y_i are identically distributed can be replaced by the assumption that y_i is **uniformly integrable** (see Section 6.19 for the definition) but i.i.d. is sufficient for most applications.

In general, an estimator which converges in probability to the population value is called consistent.

Definition 6.5 An estimator $\widehat{\theta}$ of a parameter θ is **consistent** if $\widehat{\theta} \xrightarrow{p} \theta$ as $n \to \infty$.

Theorem 6.3 If y_i are independent and identically distributed and $\mathbb{E}|y| < \infty$, then $\widehat{\mu} = \overline{y}$ is consistent for the population mean μ.

Consistency is a good property for an estimator to possess. It means that for any given data distribution; there is a sample size n sufficiently large such that the estimator beta^ will be arbitrarily close to the true value beta with high probability. **The theorem does not tell us, however, how large this n has to be**. Thus the theorem does not give practical guidance for empirical practice. Still, it is a minimal property for an estimator to be considered a "good" estimator, and provides a foundation for more useful approximations.

6.5 Almost Sure Convergence and the Strong Law

Convergence in probability is sometimes called **weak convergence**. A related concept is almost sure convergence, also known as **strong convergence**. (In probability theory the term "almost sure" means "with probability equal to one". An event which is random but occurs with probability equal to one is said to be almost sure.)

Definition 6.6 A random variable $z_n \in \mathbb{R}$ **converges almost surely** to z as $n \to \infty$, denoted $z_n \xrightarrow{a.s.} z$, if for every $\delta > 0$

$$\mathbb{P}\left(\lim_{n\to\infty} |z_n - z| \le \delta\right) = 1. \tag{6.4}$$

The convergence (6.4) is stronger than (6.1) because it computes **the probability of a limit** rather than **the limit of a probability**. Almost sure convergence is stronger than convergence in probability in the sense that **almost convergence implies convergence in probability**.

In the example (6.3) of Section 6.3, the sequence z_n converges in probability to zero for any sequence a_n; but this is not sufficient for z_n to converge almost surely. In order for z_n to converge to zero almost surely, it is necessary that a_n converges to 0.

In the random sampling context the sample mean can be shown to converge almost surely to the population mean. This is called the strong law of large numbers.

Theorem 6.4 Strong Law of Large Numbers (SLLN)
If y_i are independent and identically distributed and $\mathbb{E}|y| < \infty$, then as $n \to \infty$,

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \xrightarrow{a.s.} \mathbb{E}(y).$$

The WLLN is sufficient for most purposes in econometrics, so we will not use the SLLN in this text.

6.6 Vector-Valued Moments

Nothing important changes if we generalize to the case where y in R^m is a vector.

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}.$$

$$\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{y}) = \begin{pmatrix} \mathbb{E}(y_1) \\ \mathbb{E}(y_2) \\ \vdots \\ \mathbb{E}(y_m) \end{pmatrix}.$$

When working with random vectors y it is convenient to measure their magnitude by their Euclidean length or Euclidean norm

$$\|\boldsymbol{y}\| = \left(y_1^2 + \cdots + y_m^2\right)^{1/2}.$$

$$\|\boldsymbol{y}\|^2 = \boldsymbol{y}'\boldsymbol{y}.$$

It is equivalent to describe finiteness of moments in terms of **the Euclidean norm of a vector** or **all individual components**.

Theorem 6.5 For $\boldsymbol{y} \in \mathbb{R}^m$, $\mathbb{E}\|\boldsymbol{y}\| < \infty$ if and only if $\mathbb{E}|y_j| < \infty$ for $j = 1, ..., m$.

The m-by-m variance matrix of y is

$$\boldsymbol{V} = \operatorname{var}(\boldsymbol{y}) = \mathbb{E}\left((\boldsymbol{y}-\boldsymbol{\mu})(\boldsymbol{y}-\boldsymbol{\mu})'\right).$$

V is often called a variance-covariance matrix. You can show that the elements of V are finite if

$$\mathbb{E}\left(\|\boldsymbol{y}\|^2\right) < \infty.$$

A random sample {y_1, ..., y_n} consists of n observations of independent and identically distributed draws from the distribution of y. The vector sample mean is

$$\overline{\boldsymbol{y}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y}_i = \begin{pmatrix} \overline{y}_1 \\ \overline{y}_2 \\ \vdots \\ \overline{y}_m \end{pmatrix}$$

Convergence in probability of a vector can be defined as convergence in probability of all elements in the vector.

$\overline{\boldsymbol{y}} \xrightarrow{p} \boldsymbol{\mu}$ if and only if $\overline{y}_j \xrightarrow{p} \mu_j$ for $j = 1, \ldots, m$.

Theorem 6.6 WLLN for random vectors
If $\boldsymbol{y}_i$ are independent and identically distributed and $\mathbb{E}\left\|\boldsymbol{y}\right\| < \infty$, then as $n \to \infty$,

$$\overline{\boldsymbol{y}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y}_i \xrightarrow{p} \mathbb{E}(\boldsymbol{y}).$$

6.7 Convergence in Distribution

The WLLN is a useful first step, but does not give an approximation to the distribution of an estimator. A large-sample or asymptotic approximation can be obtained using the concept of convergence in distribution.

Definition 6.7 Let $\boldsymbol{z}_n$ be a random vector with distribution $F_n(\boldsymbol{u}) = \mathbb{P}(\boldsymbol{z}_n \leq \boldsymbol{u})$. We say that $\boldsymbol{z}_n$ **converges in distribution** to $\boldsymbol{z}$ as $n \to \infty$, denoted $\boldsymbol{z}_n \xrightarrow{d} \boldsymbol{z}$, if for all $\boldsymbol{u}$ at which $F(\boldsymbol{u}) = \mathbb{P}(\boldsymbol{z} \leq \boldsymbol{u})$ is continuous, $F_n(\boldsymbol{u}) \to F(\boldsymbol{u})$ as $n \to \infty$.

Under these conditions, it is also said that **F_n converges weakly to F**. It is common to refer to z and its distribution F(u) as the **asymptotic distribution**, **large sample distribution**, or **limit distribution** of z_n.

When the limit distribution z is degenerate (that is, P (z = c) = 1 for some c) we can write the convergence as

$$\boldsymbol{z}_n \xrightarrow{d} \boldsymbol{c},$$

which is equivalent to convergence in probability,

$$\boldsymbol{z}_n \xrightarrow{p} \boldsymbol{c}.$$

Our definition of convergence in distribution is pointwise, in the sense that it is stated for each u. It turns out that the convergence is also uniform over u when F(u) is continuous.

Theorem 6.7 If $\boldsymbol{z}_n \xrightarrow{d} \boldsymbol{z}$ and $F(\boldsymbol{u})$ is continuous then

$$\sup_{-\infty<\boldsymbol{u}<\infty} |F_n(\boldsymbol{u}) - F(\boldsymbol{u})| \longrightarrow 0$$

as $n \to \infty$.

Technically, in most cases of interest it is difficult to establish the limit distributions of sample statistics z_n by working directly with their distribution function. It turns out that in most cases it is easier to use alternative yet equivalent characterizations.

Theorem 6.8 $\boldsymbol{z}_n \xrightarrow{d} \boldsymbol{z}$ if and only if $\mathbb{E}(f(\boldsymbol{z}_n)) \to \mathbb{E}(f(\boldsymbol{z}))$ for all bounded, continuous functions f.

A further specialization of the above theorem focuses on the **characteristic function** C_n, which is a transformation of the distribution. (See Section 2.32 for the definition.) **The characteristic function C_n(t) completely describes the distribution of z_n.** It therefore seems reasonable to expect that if C_n(t) converges to a limit function C(t), then the distribution of z_n converges as well. This turns out to be true, and is known as **Levy's continuity theorem**.

Theorem 6.9 Lévy's Continuity Theorem. $\boldsymbol{z}_n \xrightarrow{d} \boldsymbol{z}$ if and only if $\mathbb{E}(\exp(\mathrm{i}\boldsymbol{t}'\boldsymbol{z}_n)) \to \mathbb{E}(\exp(\mathrm{i}\boldsymbol{t}'\boldsymbol{z}))$ for every $\boldsymbol{t} \in \mathbb{R}^k$.

Finally, we mention a standard trick which is commonly used to establish **multivariate convergence** results.

Theorem 6.10 Cramér-Wold Device. $\boldsymbol{z}_n \xrightarrow{d} \boldsymbol{z}$ if and only if $\boldsymbol{\lambda}'\boldsymbol{z}_n \xrightarrow{d} \boldsymbol{\lambda}'\boldsymbol{z}$ for every $\boldsymbol{\lambda} \in \mathbb{R}^k$ with $\boldsymbol{\lambda}'\boldsymbol{\lambda} = 1$.

6.8 Central Limit Theorem

We would like to obtain a distributional approximation to the sample mean y^bar. We start under the random sampling assumption so that the observations are independent and identically distributed, and have a finite mean mu = E(y) and variance sigma^2 = var(y).

From the WLLN we know that y^bar converges in probability to mu. Since convergence in probability to a constant is the same as convergence in distribution, this means that y converges in distribution to mu as well.

This is not a useful distributional result as the limit distribution is a constant. To obtain a non-degenerate distribution we need to **rescale** y^bar.

Recall that

$$\operatorname{var}(\overline{y} - \mu) = \sigma^2/n,$$

which means that

$$\operatorname{var}\left(\sqrt{n}\,(\overline{y} - \mu)\right) = \sigma^2.$$

This suggests renormalizing the statistic as

$$z_n = \sqrt{n}\left(\overline{y} - \mu\right).$$

Notice that E(z_n) = 0 and var(z_n) = sigma^2. This shows that the mean and variance have been stabilized. We now seek to determine the asymptotic distribution of z_n.

The answer is provided by the central limit theorem (CLT) which states that standardized sample averages converge in distribution to normal random vectors. **There are several versions of the CLT.** The most basic is the case where the observations are independent and identically distributed.

Theorem 6.11 Lindeberg-Lévy Central Limit Theorem. If y_i are independent and identically distributed and $\mathbb{E}\left(y_i^2\right) < \infty$, then as $n \to \infty$

$$\sqrt{n}\left(\overline{y} - \mu\right) \xrightarrow{d} \mathrm{N}\left(0, \sigma^2\right)$$

where $\mu = \mathbb{E}\left(y\right)$ and $\sigma^2 = \mathbb{E}(y_i - \mu)^2$.

The proof of the CLT is rather technical (so is presented in Section 6.26) but at **the core is a quadratic approximation of the log of the characteristic function**.

As we discussed above, in finite samples the standardized sum z_n has mean zero and variance sigma^2. What the CLT adds is that z_n is also approximately normally distributed, and that the normal approximation improves as n increases.

We now state the simplest and most commonly used version of a heterogeneous CLT based on the Lindeberg CLT.

Theorem 6.13 Suppose y_{ni} are independent but not necessarily identically distributed. If (6.7) and (6.9) hold, then as $n \to \infty$

$$\sqrt{n}\left(\overline{y} - \mathbb{E}\left(\overline{y}\right)\right) \xrightarrow{d} \mathrm{N}\left(0, \sigma^2\right).$$

One advantage of Theorem 6.13 is that it allows sigma^2 = 0 (unlike Theorem 6.12).

12/13

6.11 Moments of Transformations

Often we want to estimate a parameter mu which is the expected value of a transformation of a random vector y. That is, mu can be written as

$$\boldsymbol{\mu} = \mathbb{E}\left(\boldsymbol{h}\left(\boldsymbol{y}\right)\right)$$

for some function h.

For example, the second moment of y is E(y^2); the rth moment is E(y^r); the moment generating function is E(exp(ty)), and the distribution function is E(1{y <= x}).

For example, the moment estimator of E(y^r) is

$$n^{-1}\sum_{i=1}^{n} y_i^r,$$

the moment estimator of the moment generating function is

$$n^{-1}\sum_{i=1}^{n} \exp\left(ty_i\right),$$

and the estimator of the distribution function is

$n^{-1} \sum_{i=1}^{n} 1 \{y_i \leq x\}.$

Since mu^ is a sample average, and transformations of i.i.d. variables are also i.i.d., the asymptotic results of the previous sections immediately apply.

A word of caution. Theorems 6.17 and 6.18 give

For example, consider the sample 8th moment and suppose for simplicity that y is N(0,1). Then we can calculate that var (mu_8^) = 12016000/n, which is immense, even for large n.

In general, **higher-order moments are challenging to estimate** because **their variance depends upon even higher moments which can be quite large**.

6.12 Smooth Function Model

We now expand our investigation and consider estimation of parameters which can be written as a **continuous** function of mu.

$$\boldsymbol{\theta} = \boldsymbol{g}(\boldsymbol{\mu}) = \boldsymbol{g}(\mathbb{E}(\boldsymbol{h}(\boldsymbol{y})))$$

for some function g and h. This is generally known as the smooth function model, and encompasses a wide variety of econometric estimators.

As one example, the geometric mean of wages w is

$$\gamma = \exp(\mathbb{E}(\log(w))).$$

$$g(u) = \exp(u) \text{ and } h(w) = \log(w).$$

Another simple yet common example is the variance

$$\begin{aligned}\sigma^2 &= \mathbb{E}\left(w - \mathbb{E}\left(w\right)\right)^2 \\ &= \mathbb{E}\left(w^2\right) - \left(\mathbb{E}\left(w\right)\right)^2 .\end{aligned}$$

$$\boldsymbol{h}(w) = \left(\begin{array}{c} w \\ w^2 \end{array} \right)$$

$$g\left(\mu_1, \mu_2\right) = \mu_2 - \mu_1^2.$$

Similarly, the skewness of the wage distribution is

$$sk = \frac{\mathbb{E}\left(\left(w - \mathbb{E}\left(w\right)\right)^3\right)}{\left(\mathbb{E}\left(\left(w - \mathbb{E}\left(w\right)\right)^2\right)\right)^{3/2}}.$$

$$\boldsymbol{h}(w) = \begin{pmatrix} w \\ w^2 \\ w^3 \end{pmatrix}$$

$$g\left(\mu_1, \mu_2, \mu_3\right) = \frac{\mu_3 - 3\mu_2\mu_1 + 2\mu_1^3}{\left(\mu_2 - \mu_1^2\right)^{3/2}}.$$

The parameter theta = g (mu) is not a population moment, so it does not have a **direct moment estimator**. Instead, it is common to use a **plug-in estimator** formed by replacing the unknown mu with its point estimator mu^ and then "plugging" this into the expression for theta. The first step is

$$\widehat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{h}(\boldsymbol{y}_i)$$

and the second step is

$$\widehat{\boldsymbol{\theta}} = \boldsymbol{g}(\widehat{\boldsymbol{\mu}}).$$

In the next three sections we present a large-sample theory for the plug-in estimator for the smooth function model.

6.13 Continuous Mapping Theorem

Continuous functions are limit-preserving.

Theorem 6.19 Continuous Mapping Theorem (CMT). If $\boldsymbol{z}_n \xrightarrow{p} \boldsymbol{c}$ as $n \to \infty$ and $\boldsymbol{g}(\cdot)$ is continuous at $\boldsymbol{c}$, then $\boldsymbol{g}(\boldsymbol{z}_n) \xrightarrow{p} \boldsymbol{g}(\boldsymbol{c})$ as $n \to \infty$.

For example, if z_n converges in probability to c as n goes to infinity, then

$$z_n + a \xrightarrow{p} c + a$$

$$a z_n \xrightarrow{p} ac$$

$$z_n^2 \xrightarrow{p} c^2$$

as the functions g (u) = u + a; g (u) = au; and g (u) = u^2 are continuous.

Also

$$\frac{a}{z_n} \xrightarrow{p} \frac{a}{c}$$

if c is not equal to 0: The condition that c is not equal to 0 is important as the function g(u) = a/u is not continuous at u = 0.

If y_i are independent and identically distributed, mu = E(h(y)) and E||h(y)|| < infinity, then for

$\widehat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{h}(\boldsymbol{y}_i)$, as $n \to \infty$, $\widehat{\boldsymbol{\mu}} \xrightarrow{p} \boldsymbol{\mu}$.

Applying the CMT,

$\widehat{\boldsymbol{\theta}} = \boldsymbol{g}(\widehat{\boldsymbol{\mu}}) \xrightarrow{p} \boldsymbol{g}(\boldsymbol{\mu}) = \boldsymbol{\theta}$.

Theorem 6.20 If $\boldsymbol{y}_i$ are i.i.d., $\boldsymbol{\theta} = \boldsymbol{g}(\mathbb{E}(\boldsymbol{h}(\boldsymbol{y})))$, $\mathbb{E}\|\boldsymbol{h}(\boldsymbol{y})\| < \infty$, and $\boldsymbol{g}(\boldsymbol{u})$ is continuous at $\boldsymbol{u} = \boldsymbol{\mu}$, for $\widehat{\boldsymbol{\theta}} = \boldsymbol{g}(\widehat{\boldsymbol{\mu}})$ with $\widehat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{h}(\boldsymbol{y}_i)$, then $\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}$ as $n \to \infty$.

To apply Theorem 6.20 it is necessary to check if the function g is continuous at mu.

In our first example g(u) = exp(u) is continuous everywhere. It therefore follows from Theorem 6.6 and Theorem 6.20 that if E|log(w)| < infintiy then gamma^ converges in probability to gamma as n goes to infinity.

In the example of the variance, g is continuous for all mu. Thus if E(w^2) < infinity, then (sigma^)^2 converges in probability to sigma^2.

In our third example g defined in (6.18) is continuous for all mu such that var(w) = mu_2 - mu_1^2 > 0, which holds unless w has a degenerate distribution. Thus if E|w|^3 < infinity and var(w) > 0 then as sk^ converges in probability to sk as n goes to infinity.

6.14 Delta Method

In this section we introduce two tools - **an extended version of the CMT** and the **Delta Method** - which allow us to calculate the asymptotic distribution of the plug-in estimator.

> **Theorem 6.21 Continuous Mapping Theorem**
> If $\boldsymbol{z}_n \xrightarrow{d} \boldsymbol{z}$ as $n \to \infty$ and $\boldsymbol{g} : \mathbb{R}^m \to \mathbb{R}^k$ has the set of discontinuity points D_g such that $\mathbb{P}(\boldsymbol{z} \in D_g) = 0$, then $\boldsymbol{g}(\boldsymbol{z}_n) \xrightarrow{d} \boldsymbol{g}(\boldsymbol{z})$ as $n \to \infty$.

It was first proved by Mann and Wald (1943) and is therefore sometimes referred to as the **Mann-Wald Theorem**.

Theorem 6.21 allows the function g to be **discontinuous** only if **the probability at being at a discontinuity point is zero**.

For example, the function g(u) = 1/u is discontinuous at u = 0; but if z_n converges in distribution to z ~ n(0,1), then P(z=0) = 0 so 1/z_n converges in distribution to 1/z.

A special case of the Continuous Mapping Theorem is known as Slutsky's Theorem.

Theorem 6.22 Slutsky's Theorem

If $z_n \xrightarrow{d} z$ and $c_n \xrightarrow{p} c$ as $n \to \infty$, then

1. $z_n + c_n \xrightarrow{d} z + c$
2. $z_n c_n \xrightarrow{d} zc$
3. $\dfrac{z_n}{c_n} \xrightarrow{d} \dfrac{z}{c}$ if $c \neq 0$.

Even though Slutsky's Theorem is a special case of the CMT, it is a useful statement as it focuses on the most common applications - **addition**, **multiplication**, and **division**.

Despite the fact that the plug-in estimator theta^ is a function of mu^ for which we have an asymptotic distribution, Theorem 6.21 does not directly give us an asymptotic distribution for theta^.

This is because theta^ = g(mu^) is written as a function of mu^, not of the standardized sequence n^(1/2)(mu^-mu).

We need an intermediate step - **a first order Taylor series expansion**. This step is so critical to statistical theory that it has its own name - The Delta Method.

Theorem 6.23 Delta Method:
If $\sqrt{n}\left(\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu}\right) \xrightarrow{d} \boldsymbol{\xi}$, where $\boldsymbol{g}(\boldsymbol{u})$ is continuously differentiable in a neighborhood of $\boldsymbol{\mu}$ then as $n \to \infty$

$$\sqrt{n}\left(\boldsymbol{g}\left(\widehat{\boldsymbol{\mu}}\right)-\boldsymbol{g}(\boldsymbol{\mu})\right) \xrightarrow{d} \boldsymbol{G}'\boldsymbol{\xi} \tag{6.19}$$

where $\boldsymbol{G}(\boldsymbol{u}) = \frac{\partial}{\partial \boldsymbol{u}}\boldsymbol{g}(\boldsymbol{u})'$ and $\boldsymbol{G} = \boldsymbol{G}(\boldsymbol{\mu})$. In particular, if $\boldsymbol{\xi} \sim \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}\right)$ then as $n \to \infty$

$$\sqrt{n}\left(\boldsymbol{g}\left(\widehat{\boldsymbol{\mu}}\right)-\boldsymbol{g}(\boldsymbol{\mu})\right) \xrightarrow{d} \mathrm{N}\left(0, \boldsymbol{G}'\boldsymbol{V}\boldsymbol{G}\right). \tag{6.20}$$

6.15 Asymptotic Distribution for Smooth Function Model

The Delta Method allows us to complete our derivation of the asymptotic distribution of the plug-in estimator in the smooth function model. By combining Theorems 6.18 and 6.23 we find the following.

Theorem 6.24 If $\boldsymbol{y}_i$ are independent and identically distributed, $\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{h}(\boldsymbol{y}))$, $\boldsymbol{\theta} = \boldsymbol{g}(\boldsymbol{\mu})$, $\mathbb{E}\|\boldsymbol{h}(\boldsymbol{y})\|^2 < \infty$, and $\boldsymbol{G}(\boldsymbol{u}) = \dfrac{\partial}{\partial \boldsymbol{u}} \boldsymbol{g}(\boldsymbol{u})'$ is continuous in a neighborhood of $\boldsymbol{\mu}$, for $\widehat{\boldsymbol{\theta}} = \boldsymbol{g}(\widehat{\boldsymbol{\mu}})$ with $\widehat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{h}(\boldsymbol{y}_i)$, then as $n \to \infty$

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{\theta}}\right)$$

where $\boldsymbol{V}_{\boldsymbol{\theta}} = \boldsymbol{G}'\boldsymbol{V}\boldsymbol{G}$, $\boldsymbol{V} = \mathbb{E}\left((\boldsymbol{h}(\boldsymbol{y}) - \boldsymbol{\mu})(\boldsymbol{h}(\boldsymbol{y}) - \boldsymbol{\mu})'\right)$ and $\boldsymbol{G} = \boldsymbol{G}(\boldsymbol{\mu})$.

Theorem 6.20 established the consistency of theta^ for theta, and Theorem 6.24 established its asymptotic normality. It is instructive to compare the conditions required for these results.

Consistency required that h(y) have a **finite mean**, while asymptotic normality requires that this variable have a **finite variance**.

Consistency required that g(u) be **continuous**, while our proof of asymptotic normality used the assumption that g(u) is **continuously differentiable**.

6.16 Covariance Matrix Estimation

To use asymptotic distribution in Theorem 6.24 we need an estimator of the asymptotic variance matrix V_theta = G'VG. The natural plug-in estimator is

$$
\begin{aligned}
\widehat{\boldsymbol{V}}_{\boldsymbol{\theta}} &= \widehat{\boldsymbol{G}}' \widehat{\boldsymbol{V}} \widehat{\boldsymbol{G}} \\
\widehat{\boldsymbol{G}} &= \boldsymbol{G}(\widehat{\boldsymbol{\mu}}) \\
\widehat{\boldsymbol{V}} &= \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{h}(\boldsymbol{y}_i) - \widehat{\boldsymbol{\mu}})(\boldsymbol{h}(\boldsymbol{y}_i) - \widehat{\boldsymbol{\mu}})'.
\end{aligned}
$$

Under the assumptions of Theorem 6.24, the WLLN implies

$$\widehat{\boldsymbol{\mu}} \xrightarrow{p} \boldsymbol{\mu}, \text{ and } \widehat{\boldsymbol{V}} \xrightarrow{p} \boldsymbol{V}.$$

CMT implies

$$\widehat{\boldsymbol{G}} \xrightarrow{p} \boldsymbol{G}$$

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\theta}} = \widehat{\boldsymbol{G}}' \widehat{\boldsymbol{V}} \widehat{\boldsymbol{G}} \xrightarrow{p} \boldsymbol{G}' \boldsymbol{V} \boldsymbol{G} = \boldsymbol{V}_{\boldsymbol{\theta}}.$$

We have established that V_theta^ is consistent for V_theta.

Theorem 6.25 Under the assumptions of Theorem 6.24, $\widehat{\boldsymbol{V}}_{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{V}_{\boldsymbol{\theta}}$ as $n \to \infty$.

6.17 t-ratios

When l = 1 we can combine Theorems 6.24 and 6.25 to obtain the asymptotic distribution of the studentized statistic

$$T = \frac{\sqrt{n}\left(\widehat{\theta} - \theta\right)}{\sqrt{\widehat{V}_\theta}} \xrightarrow{d} \frac{\mathrm{N}\left(0, V_\theta\right)}{\sqrt{V_\theta}} \sim \mathrm{N}\left(0, 1\right).$$

The final equality is by the property that affine functions of normal random variables are normally distributed (Theorem 5.3).

Theorem 6.26 Under the assumptions of Theorem 6.24, $T \xrightarrow{d} \mathrm{N}\left(0,1\right)$ as $n \to \infty$.