Appendix C: Review of Large Sample Theory

Because large sample theory results are fundamental to modern statistical methods, for which exact results cannot be derived, we review generically and informally the basics of large sample theory.

In particular, suppose we have an estimator for a parameter of interest in a statistical model. Recall that an estimator is a function of random variables/vectors representing data, and an estimate is the numerical value that results from evaluating the estimator at a particular realization of data.

Ideally, we would like to derive the exact sampling distribution of the estimator to deduce appropriate assessments of uncertainty, such as standard errors and confidence intervals, and to develop hypothesis testing procedures. However, in complex statistical models, this is often not possible analytically. It is thus customary to appeal to theory as the sample size approaches infinity and use the theory to approximate the behavior of the estimator.

The fundamental concepts are:

- **Consistency.** Does the estimator “estimate the right stuff?” That is, for larger and larger sample sizes, does the estimator “approach” the true value of the parameter in some sense?

- **Asymptotic distribution.** Can we approximate the true, unknown sampling distribution of the estimator to use as a basis for inference and gain understanding of precision of estimation?

- **Asymptotic relative efficiency.** (There are different definitions of this concept; we consider a standard one.) Can we compare the performance of two or more competing estimators for the same quantity? If both are “consistent,” which one is “better” in terms of precision?

We review basic concepts in probability and large sample theory relevant to the above goals.

**CONSISTENCY AND ORDER IN PROBABILITY:** To evaluate whether or not an estimator “approaches” the “right stuff,” we must define precisely what we mean by this. Along with this concept is a convenient notation that summarizes behavior of relevant quantities in this sense.

**STOCHASTIC CONVERGENCE:** To discuss consistency, we need a basic understanding of convergence of random variables. The following concepts are usually introduced in a probability course, but often their practical usefulness is not elucidated.
• Estimators are functions of random variables, so that they are \textit{themselves} random variables (vectors). Thus, convergence of random variables (vectors) in a probabilistic sense is directly relevant to defining \textit{consistency}, as we now show.

For this discussion, let $Y_n$ be a \textit{generic random variable} (scalar) or vector that depends on some index $n$. Let $Y$ be another random variable or vector. Let $P$ be a relevant probability measure.

\textbf{DEFINITION C.1 Almost sure convergence.} $Y_n \xrightarrow{a.s.} Y$; i.e., $Y_n$ converges to $Y$ with probability one or almost surely, if

$$P\left(\lim_{n\to\infty} Y_n = Y\right) = 1.$$  

\textbf{DEFINITION C.2 Convergence in probability.} $Y_n \xrightarrow{p} Y$; i.e., $Y_n$ converges to $Y$ in probability if

$$\lim_{n\to\infty} P(|Y_n - Y| < \delta) = 1 \quad \text{for all } \delta > 0.$$ 

For random vectors, the definitions extend element by element.

\textbf{FACTS:} The following can be derived from the above definitions.

• $Y_n \xrightarrow{a.s.} Y$ implies that $Y_n \xrightarrow{p} Y$.

• If $h$ is a \textit{continuous} function in its argument, then if $Y_n \xrightarrow{a.s.} Y$, it follows that $h(Y_n) \xrightarrow{a.s.} h(Y)$.

  Similarly, if $Y_n \xrightarrow{p} Y$, then $h(Y_n) \xrightarrow{p} h(Y)$.

We will make routine use of the second fact in the sequel.

Taken alone, the definitions do not seem to be relevant to a study of practical issues in estimation. However, if we identify them with the generic estimation problem, their importance becomes clear.

• $n$ is the \textit{sample size}.

• $Y_n$ represents an \textit{estimator} of some parameter of interest in a \textit{statistical model}, $\eta$, say. Recall that a statistical model is a class of probability distributions assumed to have generated the data used to form the estimator. Thus, for example, a \textit{parametric} statistical model would be a probability distribution depending on a finite-dimensional parameter $\eta$. A statistical model is \textit{correct} if it includes the true distribution generating the data. In this case, the \textit{true value} of $\eta, \eta_0$, say, is then the value of $\eta$ such that the probability distribution evaluated at $\eta_0$ is that truly generating the data.
• The estimator is a **function** of \( n \) through its dependence on the \( n \) (assumed randomly sampled) observations, so we can write \( \hat{\eta}_n \). Ordinarily, we do not include a subscript “\( n \)” in standard notation for estimators, but it is important to recognize that they do depend on the sample size. If we view an estimator properly as a **function** of sample data, then the sample size serves as an **index** for a **sequence** of estimators, the functions of sample size \( n \) for each \( n \).

• \( Y \) represents the thing \( Y_n \), and thus an estimator \( \hat{\eta}_n \), “approaches.” In the estimation problem, we hope that \( \hat{\eta}_n \rightarrow \eta_0 \), where, assuming that the statistical model in which \( \eta \) appears is **correct**, \( \eta_0 \) is the **true value** generating the data.

• Thus, in **DEFINITIONS C.1** and **C.2** of modes of stochastic convergence, the estimator \( \hat{\eta}_n \) plays the role of \( Y_n \) while the value \( \eta_0 \), which in this case is a **fixed constant**, plays the role of \( Y \). So in the case of applying these definitions to **consistency** of estimators, which we define formally momentarily, the random variable or vector \( Y \) is degenerate.

**TERMINOLOGY:** Special terminology is used to describe how \( \hat{\eta}_n \) approaches \( \eta_0 \).

• **Strong consistency:** \( \hat{\eta}_n \xrightarrow{a.s.} \eta_0 \)

• **Weak consistency:** \( \hat{\eta}_n \xrightarrow{p} \eta_0 \).

**WHAT DOES THIS MEAN?**

• Both types of consistency state that the estimator **approaches** the quantity to be estimated in a probabilistic sense.

• From the definition of almost sure convergence, the interpretation of **strong** consistency is that, if the sample size \( n \) is sufficiently large, the probability that \( \hat{\eta}_n \) will assume values outside an arbitrarily small “neighborhood” of \( \eta_0 \) is zero. This follows from the fact that the limit appears inside the probability statement in **DEFINITION C.1**. Recall that, for a **deterministic sequence**, \( a_n \) has limit \( a \) if, for each \( \epsilon > 0 \), there is a value \( n_\epsilon \) such that

\[
|a_n - a| < \epsilon \quad \text{for all } n > n_\epsilon.
\]

This can be applied to the probability.

• From the definition of convergence in probability, the interpretation of **weak** consistency is that, for \( n \) large, the probability is small that \( \hat{\eta}_n \) assumes a value outside an arbitrarily small neighborhood of \( \eta_0 \). This again follows from the definition of a limit; the difference between 1 and the probability that \( \hat{\eta}_n \) is within \( \delta \) of \( \eta_0 \) is less that \( \epsilon \) if \( n \) is greater than some \( n_\epsilon \).
- The names seem to imply that strong is **better than** weak.

**PRACTICAL DIFFERENCE:** Here is a popular argument in favor of strong consistency. Suppose that one were to collect data sequentially, and, periodically, re-estimate \( \eta \) by \( \hat{\eta}_n \), where \( n \) is the number of observations collected so far. Thus, with this scheme, \( n \to \infty \). A **sequence of estimators** indexed by \( n, \hat{\eta}_n \), is thus generated.

- One would like to be assured that a value of \( n \) can be reached at which the current estimate is **sufficiently close** to the true value and will never “wander away” again after further data collection.

- Strong consistency ensures this – for \( n \) large enough, the probability that \( \hat{\eta}_n \) will stay arbitrarily close to \( \eta_0 \) is 1.

- Weak consistency does not – it states state that the probability that \( \hat{\eta}_n \) will wander away again is “small.”

This argument seems to suggest that we should **always prefer** strong consistency. However, statisticians are usually willing to settle for weak consistency; most are content that an estimator is “good” if we can make the probability of \( \hat{\eta}_n \) being “close to” the true value “large” (rather than equal to 1).

The unqualified term **consistency** in most statistical literature almost always refers to **weak** consistency. In this course, we are satisfied with weak consistency.

**TECHNICAL NOTES:**

- We have presented consistency under the conditions that there is a **statistical model** involving a parameter \( \eta \), and this model is **correctly specified**. We can thus think of this model as indexed by values of \( \eta \), and there is a true value of \( \eta, \eta_0 \), that is responsible for the data we have seen. Interest in statistical problems is of course in estimating this true value under these conditions. Usually, the term **consistency** is meant to imply this situation; i.e., that the true value of some parameter generating the data is correctly identified in the probabilistic sense.

- It is not always the case that the model is correctly specified (although we may not be aware of this). In this situation, we may still forge ahead as if it were correct, and conceive of it as being **indexed** by a parameter \( \eta \), and we can deduce estimators for \( \eta \).
However, there may not be a “true value” for $\eta$, as the model does not coincide with the true data generating mechanism. The estimator $\hat{\eta}_n$ can still be defined for each $n$ (as a function of the sample data), and it may still converge in probability (or almost surely) to some quantity, $\eta_*$, say. Such a $\hat{\eta}_n$ is sometimes referred to as being “consistent” for the value $\eta_*$, which can be confusing.

- Alternatively, even in the context of a correctly specified model, it is possible to define estimators $\hat{\eta}_n$ that do not “estimate the right stuff;” i.e., that are not consistent in that $\hat{\eta}_n \xrightarrow{P} \eta_*$, where $\eta_* \neq \eta_0$. In this case, $\hat{\eta}_n$ is said to be inconsistent.

- In general, the goal is (a) to identify a statistical model that is correct (i.e., contains the true distribution generating the data) and (b) identify a consistent estimator for the true value of a parameter $\eta$ indexing this model. If (a) is not carried out, (b) may not be possible.

- In most studies of properties of estimators, that the model is correctly specified is taken as a starting point. We take this perspective initially; however, we will also investigate what happens when certain components of models are not correctly specified.

**ORDER IN PROBABILITY:** This notation can appear confusing initially, but once mastered, is useful for streamlining presentation of large sample results. Again, let $\{Y_n\}$ denote a generic sequence of random variables/vectors indexed by $n$.

**DEFINITION C.3 “Big” $O_p$.** $Y_n$ is at most of order in probability $n^k$ if, for all $\epsilon > 0$, there exist constants $n_\epsilon, M_\epsilon > 0$ such that

$$P(n^{-k}\|Y_n\| < M_\epsilon) > 1 - \epsilon$$

for all $n > n_\epsilon$. Here, $\|\cdot\|$ is some norm measuring magnitude in the case of vector $Y_n$; if $Y_n$ is scalar, then this is just absolute value.

The notation is $Y_n = O_p(n^k)$

- The definition says that the magnitude of $n^{-k}Y_n$ stays bounded with high probability if $n$ is large enough. The cases of most interest to us are when $k$ is nonpositive.

- For example, if $k = -1/2$, then $n^{1/2}Y_n$ stays bounded as $n$ gets large with high probability. In particular, with high probability, $Y_n$ is bounded by $M_n n^{-1/2}$. This means that $Y_n$ itself is getting “small” as $n$ gets large. A practical interpretation is that $Y_n$ behaves like $n^{-1/2}$ with high probability for $n$ large enough; i.e., becomes negligible in the same “way” $n^{-1/2}$ does.
In fact, as \( n^{-1/2} \to 0 \) as \( n \to \infty \), this says that \( Y_n \) itself “approaches” (converges in probability) to zero at the same rate as \( n^{-1/2} \).

If \( k = 0 \), then \( Y_n = O_p(1) \). From Definition C.3, this says that \( Y_n \) remains bounded by the constant \( M \) for \( n \) large with high probability. In this case, \( Y_n \) is said to be bounded in probability.

Practically speaking, this says that, as \( n \) gets large, \( Y_n \) does not become negligible, nor does it “blow up.” Instead, it is “nicely behaved.”

**DEFINITION C.4** “Little” \( o_p \). \( Y_n \) is said to be of smaller order in probability than \( n^k \) if

\[
n^{-k} Y_n \xrightarrow{p} 0 \quad \text{as} \quad n \to \infty.
\]

The notation is \( Y_n = o_p(n^k) \).

The case of most interest to us is \( k = 0 \). It is shorthand for saying that \( Y_n \) converges in probability to zero; that is, for \( n \) sufficiently large, \( Y_n \) stays arbitrarily “close” to zero with high probability. If we can show an expression is \( o_p(1) \), it can be “ignored” as “negligible.”

More generally, the case \( k \leq 0 \) is the most interesting. For example, if \( k = -1/2 \), \( Y_n = o_p(n^{-1/2}) \), then \( n^{1/2} Y_n \xrightarrow{p} 0 \). This shorthand notation says that we can multiply \( Y_n \) by a factor that acts like \( n^{1/2} \), and the entire product will still be negligible. Thus, this notation is a useful way of expressing how quickly \( Y_n \xrightarrow{p} 0 \); i.e. if \( Y_n = o_p(n^{-1/2}) \), then it is “faster” than \( n^{-1/2} \).

**FACTS:** The following facts can be deduced from the definitions.

- \( Y_n = O_p(n^{-\delta}) \) for \( \delta > 0 \) implies that \( Y_n = o_p(1) \). This is intuitively obvious – if \( Y_n \) acts like \( n^{-\delta} \) when \( n \) is large with high probability, then it must “go to zero.”

  Stating that \( Y_n = O_p(n^{-\delta}) \) is more informative than just saying \( Y_n = o_p(1) \); the former not only tells us that \( Y_n \) becomes negligible with high probability, but at what rate.

- If \( Y_n = O_p(n^k) \) and \( X_n = O_p(n^j) \), then \( X_n Y_n = O_p(n^{k+j}) \). The same holds true if \( O_p \) is replaced by \( o_p \), and for combinations of \( O_p \) and \( o_p \). Thus, if we know the order in probability of each of two quantities, we can deduce how their product behaves.

- A useful special case is when \( Y_n = O_p(1) \), referred to as bounded in probability, and \( X_n = o_p(1) \) \( (X_n \xrightarrow{p} 0) \). Then \( X_n Y_n = o_p(1) \). Intuitively, this makes sense; \( Y_n \) is “well behaved,” neither getting small nor blowing up and is multiplied by something that is getting small. Thus, the product would be expected to get small. Of course, this means that \( X_n Y_n \xrightarrow{p} 0 \).
It is important to keep in mind that $Y_n = O_p(n^{-1/2})$, say, only means that the magnitude of $n^{1/2}Y_n$ is bounded by some constant. That constant could be huge, so that $n$ must be very, very large for $Y_n$ to become negligible for practical purposes. This explains in part why, sometimes, large sample approximations do not seem relevant in practice.

We now turn to concepts useful in deducing approximations to sampling distributions of estimators. Continue to regard $Y_n$ and $Y$ as generic random variables/vectors.

**DEFINITION C.5 Convergence in distribution.** Suppose $Y_n$ has cumulative distribution function (cdf) and that $Y$ has cdf $F$. $Y_n$ is said to converge in distribution (or law) to $Y$ if and only if, for each continuity point of $F$,

$$\lim_{n \to \infty} F_n(y) = F(y).$$

The standard notation is $Y_n \overset{D}{\longrightarrow} Y$ or $Y_n \overset{L}{\longrightarrow} Y$; we use the latter.

**PRACTICAL INTERPRETATION:** If $Y_n \overset{L}{\longrightarrow} Y$, this implies roughly that, for large $n$, except at a few points, the distribution of $Y_n$ (and thus probabilities associated with $Y_n$) is the same as that of $Y$. Thus, if we are interested in probability and distributional statements about $Y_n$, we can approximate these with statements about $Y$.

In the context of estimation, if we are interested in approximating the sampling distribution of an estimator, we are interested in the convergence in distribution of the estimator (or some function thereof).

**FACTS:** The following can be deduced from Definition C.5 and previous definitions.

- If $Y_n \overset{p}{\longrightarrow} Y$, then $Y_n \overset{L}{\longrightarrow} Y$. This says that if, for large $n$, the probability that $Y_n$ differs from $Y$ is small, then we would expect the probabilities with which they take on values to be “close,” and thus expect them to have distributions that are “close.”

- However, $Y_n \overset{L}{\longrightarrow} Y$ DOES NOT imply $Y_n \overset{p}{\longrightarrow} Y$ in general. For example, suppose that $Y_n$ and $Y$ have the same distribution for each $n$, but $Y_n$ and $Y$ are independent for each $n$. Then a realization of $Y_n$ is totally unrelated to a realization of $Y$!

- $Y_n \overset{L}{\longrightarrow} y$, where $y$ is a constant, does imply $Y_n \overset{p}{\longrightarrow} y$. Intuitively, because the distribution of $Y_n$ collapses to a single point, a realization of $Y_n$ must also approach that point.
Of course, if \( Y_n \xrightarrow{L} y \), a constant, then the distribution is degenerate, which is not particularly interesting if one seeks to deduce a sampling distribution to be used for constructing confidence intervals and hypothesis tests.

**SAMPLING DISTRIBUTION OF AN ESTIMATOR:** Return now to our situation of interest, where \( \hat{\eta}_n \) is an estimator for a parameter \( \eta \) in a (correct) statistical model with true value \( \eta_0 \). Showing that \( \hat{\eta}_n \xrightarrow{P} \eta_0 \) thus implies that \( \hat{\eta}_n \xrightarrow{L} \eta_0 \). However, this knowledge that the distribution of \( \hat{\eta}_n \) collapses to the single point \( \eta_0 \) is not very useful for the usual inferential goals described above. In particular, this result does not even give information on precision of estimation.

To gain insight and to provide a basis for the standard inferential objectives, we must pursue a more refined assessment of large sample behavior. Instead of considering the properties of \( \hat{\eta}_n \) itself, we instead consider a suitable function of \( \hat{\eta}_n \) whose properties are “more interesting” and relevant. For most estimators solving estimating equations, a standard approach to deriving an approximate sampling distribution that is more useful applies.

**DEFINITION C.6 Asymptotic normality.** We present this definition in the scalar case; the vector case is similar. Classically speaking, a random variable \( Y_n \) is said to be asymptotically normal if we can find sequences \( \{a_n\} \) and \( \{c_n\} \) such that

\[
c_n(Y_n - a_n) \xrightarrow{L} \mathcal{N}(0, 1).
\]

By this notation, we mean that the right hand side of this expression is a standard normal random variable.

**DEFINITION C.6** implies that, although the limit distribution of \( Y_n \) itself may be uninteresting, if we center and scale \( Y_n \) appropriately, this “standardized” version of \( Y_n \) has an interesting limit distribution.

- In particular, \( a_n \) is called the asymptotic mean and \( c_n \) is called the asymptotic variance, and **DEFINITION C.6** can be interpreted to mean that, approximately for large \( n \)

\[
Y_n \sim \mathcal{N}(a_n, c_n^{-2}).
\]

The usefulness of this result for approximating a sampling distribution is thus evident.
How is this applied in estimation situations of interest to us? As we will see, because many estimators of interest to us are *not available in a closed form*, things are not as simple as immediately identifying \( \hat{\eta}_n \) with \( Y_n \) and then determining appropriate centering and scaling constants. Instead, what is done is to find an approximation to an *appropriate centered and scaled* version of \( \hat{\eta}_n \) by applying a *Taylor series* to the *estimating equation* that defines \( \hat{\eta}_n \) implicitly. This approximation then forms the basis for deducing behavior like that in Definition C.6.

Some important tools for deducing this behavior are the following. After we state these important results, we sketch how they are used in this way.

There are numerous versions of *central limit theorems* that characterize the *convergence in distribution* of appropriately standardized *sums of independent random variables/vectors*. These can be extended to random vectors in a number of ways to allow generalization of univariate results to multivariate ones. We do not discuss the technicalities behind this. Instead, we state a particular *multivariate central limit theorem* that is useful for our purposes.

**MULTIVARIATE CENTRAL LIMIT THEOREM:** Let \( Z_i \) be independent random vectors with \( E(Z_i) = \mu_i \) and \( \text{var}(Z_i) = \Sigma_i, i = 1, \ldots, n \), such that

\[
\lim_{n \to \infty} n^{-1}(\Sigma_1 + \cdots + \Sigma_n) = \lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} \Sigma_i = \Sigma,
\]

say, letting \( F_i \) be the cdf of \( Z_i \),

\[
n^{-1} \sum_{i=1}^{n} \int_{\|z - \mu_i\| \geq \epsilon n^{1/2}} \|z - \mu_i\|^2 dF_i(z) \longrightarrow 0 \quad \text{as } n \to \infty.
\] (C.1)

Then

\[
n^{-1/2} \sum_{i=1}^{n} (Z_i - \mu_i) \overset{d}{\longrightarrow} \mathcal{N}(0, \Sigma).
\]

The *Lindeberg condition* (C.1) effectively restricts the *tail behavior* of \( Z_i \) and does not appear particularly intuitive. It turns out that (C.1) may be shown to hold if the *third moments* of \( Z_i \) exist and are finite (the so-called *Lyapunov condition*). We generally assume that higher moments of the response *exist* and are *finite*, so that the moments of relevant quantities to which this theorem will be applied can be assumed to exist and be finite. Thus, condition (C.1) is assumed without comment when we apply the multivariate central limit theorem.

Additional key results of which we will make heavy use are the following.
**SLUTSKY’S THEOREM:** Suppose that \( Y_n \overset{c}{\to} Y \) and \( V_n \overset{p}{\to} c \), where \( c \) is a constant. Then

\[
Y_n + V_n \overset{c}{\to} Y + c, \quad Y_nV_n \overset{c}{\to} cY, \quad Y_n/V_n \overset{c}{\to} Y/c,
\]

where in the last expression \( c \neq 0 \) is required. These extend readily to random vectors: If \( Y_n \overset{c}{\to} Y \) and \( \Sigma_n \overset{p}{\to} C \), where \( \Sigma_n \) and \( C \) are matrices,

\[
Y_n + \Sigma_n \overset{c}{\to} Y + C, \quad \Sigma_n Y_n \overset{c}{\to} CY, \quad \Sigma_n^{-1} Y_n \overset{c}{\to} C^{-1} Y.
\]

Thus, when \( E(Y) = \mu \) and \( \text{var}(Y) = \Sigma \), say, then we have that \( \Sigma_n Y_n \overset{c}{\to} \) to random vector with mean \( C\mu \) and covariance matrix \( C\Sigma C^T \).

Slutsky’s theorem may be invoked repeatedly so that if \( U_n \overset{p}{\to} D \) as well, then

\[
\Sigma_n Y_n + U_n \overset{c}{\to} CY + D.
\]

**WEAK LAW OF LARGE NUMBERS:** We state this in the scalar case, but it extends straightforwardly to vectors. Suppose \( Z_i \) are independent (or uncorrelated) random variables and \( a_i \) are constants. Then, if \( n^{-2} \sum_{i=1}^{n} \text{var}(Z_i)a_i^2 \to 0 \),

\[
n^{-1} \sum_{i=1}^{n} a_i Z_i - n^{-1} \sum_{i=1}^{n} a_i E(Z_i) \overset{p}{\to} 0.
\]

- The condition \( n^{-2} \sum_{i=1}^{n} \text{var}(Z_i)a_i^2 \to 0 \) is satisfied if \( n^{-1} \sum_{i=1}^{n} \text{var}(Z_i)a_i^2 \to c \) for some constant \( c \), which is often reasonable (and similar to the requirement for the central limit theorem).

- If we furthermore know that \( n^{-1} \sum_{i=1}^{n} a_i E(Z_i) \to d \), say, then we can conclude that \( n^{-1} \sum_{i=1}^{n} a_i Z_i \overset{p}{\to} d \), as

\[
n^{-1} \sum_{i=1}^{n} a_i Z_i - d = \{ n^{-1} \sum_{i=1}^{n} a_i Z_i - n^{-1} \sum_{i=1}^{n} a_i E(Z_i) \} + \{ n^{-1} \sum_{i=1}^{n} a_i E(Z_i) - d \} \overset{p}{\to} 0.
\]

We are now in a position to describe how all of this is used in more detail. We drop the \( n \) subscript on our generic estimator and treat it and the parameter of interest as vectors, writing \( \eta \) and \( \hat{\eta} \).

For estimating equations for a parameter \( \eta \) of interest with solution \( \hat{\eta} \), we can deduce using Taylor series and some additional conditions that

\[
n^{1/2}(\hat{\eta} - \eta_0) = A_n^{-1} C_n + o_p(1),
\]

where \( C_n = n^{-1/2} \sum_{i=1}^{n} \) (function of data), \( A_n = n^{-1} \sum_{i=1}^{n} \) (function of data), and \( o_p(1) \) represents terms that converge in probability to zero.
We then

- Apply the **central limit theorem** to $C_n$ to show that it *converges in distribution* to a normal random vector.

- Apply the **weak law of large numbers** to $A_n$ to show that it *converges in probability* to a constant matrix.

- Apply **Slutsky’s theorem** to $A_n^{-1}C_n$ to conclude that $n^{1/2}(\hat{\eta} - \eta_0)$ *converges in distribution* to a normal random vector with **mean zero** and some covariance matrix $\Sigma$; i.e.,

$$n^{1/2}(\hat{\eta} - \eta_0) \xrightarrow{L} \mathcal{N}(0, \Sigma), \quad \text{(C.2)}$$

say. This is often interpreted as $\hat{\eta} \sim \mathcal{N}(\eta_0, n^{-1}\Sigma)$; i.e., $\hat{\eta}$ is asymptotically normal with mean $\eta_0$ and covariance matrix $n^{-1}\Sigma$.

From these steps, we can then deduce an approximate (normal) **sampling distribution** for $\hat{\eta}$.

Suppose that we have **two competing estimators** for $\eta$ ($k \times 1$), $\hat{\eta}^{(1)}$ and $\hat{\eta}^{(2)}$, say, so that we have

$$n^{1/2}(\hat{\eta}^{(1)} - \eta_0) \xrightarrow{L} \mathcal{N}(0, \Sigma_1) \quad \text{and} \quad n^{1/2}(\hat{\eta}^{(2)} - \eta_0) \xrightarrow{L} \mathcal{N}(0, \Sigma_2)$$

for some matrices $\Sigma_1$ and $\Sigma_2$.

- Both $\hat{\eta}^{(1)}$ and $\hat{\eta}^{(2)}$ are **consistent**. It is a general fact that, if a random vector *converges in distribution*, then it is **bounded in probability**. Thus, we have

$$n^{1/2}(\hat{\eta}^{(\ell)} - \eta_0) = O_p(1)$$

for $\ell = 1, 2$. This may be expressed equivalently as

$$(\hat{\eta}^{(\ell)} - \eta_0) = O_p(n^{-1/2}).$$

- Thus, both estimators “estimate the right stuff” and approach it at the same rate. On this basis, then, they are entirely **comparable**.

- As this **does not distinguish** the two estimators from one another, consider their **precision**. In finite sample, exact theory, the estimator that is **more precise** is to be preferred. Here, we approximate the covariance matrices of the estimators by $n^{-1}\Sigma_1$ and $n^{-1}\Sigma_2$, respectively. This suggests comparing $\Sigma_1$ and $\Sigma_2$. 


• In the case \( k = 1 \), so that \( \Sigma_1 \) and \( \Sigma_2 \) are scalar variances, this suggests preferring the estimator with the smaller variance. That is, prefer \( \hat{\eta}_2 \) to \( \hat{\eta}_1 \) if \( \Sigma_2 < \Sigma_1 \). If \( \Sigma_2 = \Sigma_1 \), then the two estimators are of equal precision.

**DEFINITION C.7 Asymptotic relative efficiency.** For scalars, the asymptotic relative efficiency of \( \hat{\eta}_1 \) to \( \hat{\eta}_2 \) is defined as

\[
ARE = \frac{\Sigma_2}{\Sigma_1}.
\]

With this definition, if \( ARE = 1 \), the estimators are equally precise. If \( ARE < 1 \), then \( \hat{\eta}_1 \) is inefficient relative to \( \hat{\eta}_2 \), and if \( ARE > 1 \), then \( \hat{\eta}_1 \) offers a gain in efficiency relative to \( \hat{\eta}_2 \).

Often, one constructs the ratio with the potentially “better” estimator’s variance in the numerator, so that \( ARE < 1 \) is “good” for showing another estimator is inefficient relative to it. However, many texts and authors do this in the reverse, so that larger-than-one values are preferred.

The extension of the definition to \( k > 1 \) is that \( \hat{\eta}^{(2)} \) is preferable to \( \hat{\eta}^{(1)} \) if the covariance matrix \( \Sigma_2 \) is “smaller” than \( \Sigma_1 \) in some sense. To formalize this, if \( (\Sigma_1 - \Sigma_2) \) is nonnegative definite, then, for all \( (k \times 1) \) \( \lambda \),

\[
\lambda^T \Sigma_2 \lambda \leq \lambda^T \Sigma_1 \lambda.
\]

By choosing \( \lambda \) in turn to be the vector with a 1 in one position and zeroes elsewhere, we see that this implies that the variances on the diagonal of \( \Sigma_2 \) must be smaller than those on the diagonal of \( \Sigma_1 \), so that the (approximate) variance of each component of \( \hat{\eta}^{(2)} \) is \textit{smaller} than that of \( \hat{\eta}^{(1)} \).

If \( (\Sigma_1 - \Sigma_2) \) is nonnegative definite, it follows that

\[
|\Sigma_2| \leq |\Sigma_1|.
\]

Thus, the asymptotic relative efficiency of \( \hat{\eta}^{(1)} \) to \( \hat{\eta}^{(2)} \) is generally defined for \( k > 1 \) as

\[
ARE = \left\{ \frac{|\Sigma_2|}{|\Sigma_1|} \right\}^{1/k}.
\]

This comparison is sometimes simplified in that it turns out that \( \Sigma_1 = \alpha_1 \Sigma \) and \( \Sigma_2 = \alpha_2 \Sigma \) for some scalars \( \alpha_\ell, \ell = 1, 2 \), and common matrix \( \Sigma \). In this case, \( ARE \) reduces to \( \alpha_2/\alpha_1 \), as \( |\alpha \Sigma| = \alpha^k |\Sigma| \).

It is often argued that one estimator is efficient relative to another by examining the difference \( \Sigma_1 - \Sigma_2 \). However, simply noting that this difference is nonnegative definite does not give insight into \textit{how much} more efficient. The calculation of \( ARE \) quantifies “how much better.” In complex statistical models, \( ARE \) usually depends on the design, parameter values, and functions involved, so that a “global” statement of relative efficiency can not be made.