5 Large Sample (Asymptotic) Properties of Estimators

[CB10.1, BD5.2-5.3]
Asymptotics in statistics is usually thought of as the study of the limiting behavior of statistics or, more specifically, of distributions of statistics, based on observing n i.i.d. observations $X_1, \cdots, X_n$ as $n \to \infty$. Asymptotics, in this context, always refers to a sequence of statistics $\{T_n(X_1, \cdots, X_n)\}_{n\geq 1}$, for instance the sequence of means $\{\bar{X}_n\}_{n\geq 1}$, or the sequence of medians, or it refers to the sequence of their distributions $\{\mathrm{L}_F(T_n(X_1, \cdots, X_n))\}_{n\geq 1}$. Asymptotic statements are always statements about the sequence.

The strong law of large numbers (Kolmogorov) tells us that if $X_1, X_2, \cdots, X_n$ are independent and identically distributed, the existence of a finite constant c for which $\bar{X} \stackrel{a.s.}{\to} c$ holds iff $E(X_1)$ is finite and equals c. [Serfling pg 27]

We interpret this as saying that, for n sufficiently large, X_n is approximately equal to its expectation. The trouble is that for any specified degree of approximation, say, $\epsilon = .01$, this does not tell us how large n has to be for the approximation not holding to this degree, that is $|\bar{X}(\omega) - c| > \epsilon$. Is $n > 100$ enough or does it have to be $n > 100,000$?

Central Limit Theorem(Lindeberg-Levy) If $X_1, X_2, \cdots, X_n$ are independent and identically distributed (distribution F) with finite mean ($E_F(X_1) = \mu$) and variance ($Var_F(X_1) = \sigma^2$), then $\sqrt{n}(\bar{X}_n - \mu) \Rightarrow \mathcal{N}(0, \sigma^2)$.

Relaxation of assumptions lead to other CLT's like Lindeberg Levy where independence is assumed but different means and variances are allowed satisfying suitable conditions.

As an approximation, this reads $P(\bar{X} \leq x) \approx \Phi(\sqrt{n}(x - \mu)/\sigma)$. Again we are faced with the questions of how good the approximation is for given n, x and F. What we in principle prefer are bounds, which are available in the classical situations of WLLN and CLT above.

By Chebychev's inequality, if $E_F(X_l^2) < \infty$, then $P_F[|\bar{X}_n - \mu| \geq \epsilon] \leq \sigma^2/n\epsilon^2$.As a bound this is typically far too conservative. If $|X_1| \leq 1$, the much more delicate Hoeffding bound gives $P_F[|\bar{X}_n - \mu| \geq \epsilon] \leq 2exp(-n\epsilon^2/2)$. Because $|X_1 \leq 1$ implies that $\sigma^2 \leq 1$ when σ^2 is unknown the RHS of Chebychev becomes $1/n\epsilon^2$. For $\epsilon = .1, n = 400$ Chebychev is 0.25 whereas Hoeffding is 0.14. Of course $|X_1 \leq 1$ can be replaced with $|X_1 \leq M$.

The celebrated Berry-Esseen bound states that if $E_F|X_1|^3 <^{\infty}$, $sup_x|P(\sqrt{n}(\bar{X}_n - \mu)/\sigma \leq x) - \Phi(x)| \leq CE_F|X_1|^3/\sigma^3\sqrt{n}$ where C is a universal constant known to be $< 33/4$.

5.1 Consistency

Consistency refers to convergence in probability (weak) or almost surely (strong). In the iid case, by LLN, $\bar{X}_n$ is consistent.

Ex: $X_1, \cdots X_n$ iid Bernoulli(p). $\hat{p}_n = \bar{X}_n$ is consistent for p by LLN. Consider $\theta = p(1 - p)$ which is the variance of X_1 and its method of moments

estimator $\hat{p}_n(1-\hat{p}_n)$. This is consistent.

Result: If X_n converges to X in probability and g is a continuous function, then $g(X_n)$ converges to $g(X)$ in probability.

Theorem 1 *(Consistency of MoM estimators) $X_1, \cdots, X_n$ iid. $X_i \in \mathcal{X}$. Let $g = (g_1, \cdots, g_d)$ map $\mathcal{X}$ onto $\mathcal{Y} \subseteq \mathbb{R}^d$. Suppose $E_\theta g_j(X_1) < \infty, 1 < j < d, \forall \theta$. Let $m_j(\theta) = E_\theta g_j(X_l), 1 < j < d$ and $q(\theta) = h(m(\theta))$, where $h : \mathcal{Y} \to \mathbb{R}^p$. Then, if h is continuous $\hat{q} = h(\bar{g})$ is consistent for $q(\theta)$.*

Theorem 2 *(Consistency of MLE in exponential family) Suppose $\mathcal{P}$ is a canonical exponential family of rank d generated by T. Suppose $\mathcal{E}$ the support of the canonical parameter η, is open. Then, if $X_l, \cdots, X_n$ are a sample from $P_\eta \in \mathcal{P}$*

1. *P(The MLE $\hat{\eta}$ exists) $\to$ 1*
2. *$\hat{\eta}$ is consistent.*

Pf: Pg 304 of BD. Not to be done in class.

We begin the discussion of the consistency of the MLE by defining the so-called Kullback-Liebler information.

Definition 1 *If $f_{\theta_0}(x)$ and $f_{\theta_1}(x)$ are two densities, the Kullback-Leibler information number equals $K(f_{\theta_0}, f_{\theta_1}) = E_{\theta_0} \log \frac{f_{\theta_0}(X)}{f_{\theta_1}(X)}$. If $P_{\theta_0}\left(f_{\theta_0}(X) > 0 \quad \text{and} \quad f_{\theta_1}(X) = 0\right) > 0$, then $K(f_{\theta_0}, f_{\theta_1})$ is defined to be 1.*

We may show that the Kullback-Leibler information must be nonnegative using Jensen's inequality.

Theorem 3 (Jensen's inequality) *If $g(t)$ is a convex function, then for any random variable $X, g(EX) \leq Eg(X)$. Furthermore, if $g(t)$ is strictly convex, then $Eg(X) = g(EX)$ only if $P(X = c) = 1$ for some constant c.*

Considering the Kullback-Leibler information once again, we first note that

$$E_{\theta_0} \frac{f_{\theta_1}(X)}{f_{\theta_0}(X)} = E_{\theta_1}\left(I_{f_{\theta_0}(X)>0}\right) \leq 1.$$

Therefore, by the strict convexity of the function $-\log x$,

$$K(f_{\theta_0}, f_{\theta_1}) = E_{\theta_0} - \log \frac{f_{\theta_1}(X)}{f_{\theta_0}(X)} \geq -\log E_{\theta_0} \frac{f_{\theta_1}(X)}{f_{\theta_0}(X)} \geq 0, \tag{1}$$

with equality if and only if $P_{\theta_0} f_{\theta_0}(X) = f_{\theta_1}(X) = 1$. Inequality (1) is sometimes called the Shannon- Kolmogorov information inequality.

If $X_1, \cdots, X_n$ are iid with density $f_{\theta_0}(x)$, then $l(\theta) = \sum_{i=1}^n \log f_{\theta_0}(x_i)$. Thus, the Shannon-Kolmogorov information inequality may be used to prove the consistency of the maximum likelihood estimator in the case of a finite parameter space.

Theorem 4 (Consistency of MLE) *Suppose Ω is finite and that $X_1, \cdots, X_n$ are iid with density $f_{\theta_0}(x)$. Furthermore, suppose that the model is identifiable, which is to say that different values of θ lead to different distributions. Then if $\hat{\theta}_n$ denotes the maximum likelihood estimator, $\hat{\theta}_n \xrightarrow{P} \theta_0$.*

Proof: Notice that

$$\frac{1}{n}\sum_{i=1}^{n} \log \frac{f_\theta(X_i)}{f_{\theta_0}(X_i)} \xrightarrow{P} E_{\theta_0} \log \frac{f_\theta(X_i)}{f_{\theta_0}(X_i)} = -K(f_{\theta_0}, f_\theta) \tag{2}$$

The value of $-K(f_{\theta_0}, f_\theta)$ is strictly negative for $\theta \neq \theta_0$ by the identifiability of the model. Therefore, since $\hat{\theta}_n$ is the maximizer of the left hand side of Equation (2),

$$P(\hat{\theta}_n \neq \theta_0) = P\left(\max_{\theta \neq \theta_0}(\frac{1}{n}\sum_{i=1}^{n} \log \frac{f_\theta(X_i)}{f_{\theta_0}(X_i)}) > 0\right) \leq \sum_{\theta \neq \theta_0} P\left(\frac{1}{n}\sum_{i=1}^{n} \log \frac{f_\theta(X_i)}{f_{\theta_0}(X_i)} > 0\right) \to 0. \tag{3}$$

5.2 Asymptotic normality

In the simplest form of the central limit theorem, we consider a sequence $X_1, X_2, ...X_n$ of independent and identically distributed (univariate) random variables with mean μ and finite variance σ^2. In this case, the central limit theorem states that

$$\sqrt{n}(\bar{X_n} - \mu) \Rightarrow \mathcal{N}(0, \sigma^2)$$

5.2.1 Delta Method

In this section, we wish to consider the asymptotic distribution of, say, some function of $\bar{X_n}$. In the simplest case, the answer depends on results already known: Consider a linear function $h(t) = at + b$ for some known constants a and b. Clearly $\mathrm{E}(h(\bar{X_n})) = a\mu + b = h(\mu)$ by the linearity of the expectation operator. Therefore, it is reasonable to ask whether $\sqrt{n}(h(\bar{X_n}) - h(\mu))$ tends to some distribution as $n \to \infty$. But the linearity of $h(t)$ allows one to write

$$\sqrt{n}(h(\bar{X_n}) - h(\mu)) = a\sqrt{n}(\bar{X_n} - \mu)$$

We conclude that

$$\sqrt{n}(h(\bar{X_n}) - h(\mu)) \Rightarrow \mathcal{N}(0, a^2\sigma^2)$$

None of the preceding development is especially deep; one might even say that it is obvious that a linear transformation of the random variable T_n alters its asymptotic distribution by a constant multiple. Yet what if the function $h(t)$ is nonlinear? It is in this nonlinear case that a strong understanding of the argument above, as simple as it may be, pays real dividends. For if T_n is consistent for θ (say), then we know that, roughly speaking, T_n will be very close to θ for large n. Therefore, the only meaningful aspect of the behavior of

$h(t)$ is its behavior in a small neighborhood of θ. And in a small neighborhood of θ, $h(\theta)$ may be considered to be roughly a linear function. Formally we use the Taylor expansion to obtain the following result:

Theorem 5 (First Order Delta Method) *If*

$$\sqrt{n}(T_n - \theta) \Rightarrow \mathcal{N}(0, \tau^2) \tag{4}$$

then

$$\sqrt{n}(h(T_n) - h(\theta)) \Rightarrow \mathcal{N}(0, \tau^2(h'(\theta))^2)$$

provided $h'(\theta)$ exists and is not zero.

Proof: Step 1: It follows from equation (4) that $T_n \to \theta$ in probability.
Step 2: Consider the Taylor expansion of h around θ.

$$h(x) = h(\theta) + (x - \theta)(h'(\theta) + r)$$

where $r \to 0$ as $x \to \theta$.
Define R_n as the remainder in

$$h(T_n) = h(\theta) + (T_n - \theta)(h'(\theta) + R_n)$$

By step 1, $T_n \to \theta$ in probability.
Hence $R_n \to 0$ in probability.
This implies $h'(\theta) + R_n \to h'(\theta)$ in probability.

Step 3: The result follows by applying Slutsky's theorem to $\sqrt{n}(h(T_n) - h(\theta))$.
$\sqrt{n}(h(T_n) - h(\theta)) = \sqrt{n}(T_n - \theta) \times (h'(\theta) + R_n)$.
Let $Y_n = (h'(\theta) + R_n)$ and $X_n = \sqrt{n}(T_n - \theta)$ as above.
$X_n \Rightarrow X$ and $Y_n \to c$ in probability where $c = h'(\theta), X \sim \mathcal{N}(0, \tau^2)$.
By Slutsky's theorem, $\sqrt{n}(h(T_n) - h(\theta)) = Y_n X_n \Rightarrow cX$.
The distribution of cX is $\mathcal{N}(0, \tau^2(h'(\theta))^2)$.

Example 1(Exponential Rate) Let $X_i, i = 1, 2, \cdots, n$ be independent Exponential(λ) random variables and let $T_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then by CLT,

$$\sqrt{n}(T_n - \lambda) \Rightarrow \mathcal{N}(0, \lambda^2)$$

Suppose we are now interested in the large sample behavior of the estimate $\frac{1}{T_n}$ of the rate $h(\lambda) = \frac{1}{\lambda}$.
Since $h'(\lambda) = -\frac{1}{\lambda^2}$, it follows from Theorem 5 that

$$\sqrt{n}(\frac{1}{T_n} - \frac{1}{\lambda}) \Rightarrow \mathcal{N}(0, \left(-\frac{1}{\lambda^2}\right)\lambda^2 = \frac{1}{\lambda^2})$$

Example 2 (Binomial Variance) Let $X_i, i = 1, 2, \cdots, n$ be independent Bernoulli random variables and let $T_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then by CLT,

$$\sqrt{n}(T_n - p) \Rightarrow \mathcal{N}(0, p(1-p))$$

Suppose we are now interested in the large sample behavior of the estimate $T_n(1-T_n)$ of the variance $h(p) = p(1-p)$.
Since $h'(p) = 1-2p$, it follows from Theorem 5, when $p \neq 1/2$, that

$$\sqrt{n}(T_n(1-T_n) - p(1-p)) \Rightarrow \mathcal{N}(0, (1-2p)^2 p(1-p))$$

What happens when $h'(\theta) = 0$?

Theorem 6 (Second Order Delta Method) *If*

$$\sqrt{n}(T_n - \theta) \Rightarrow \mathcal{N}(0, \tau^2) \quad \text{and} \quad h'(\theta) = 0 \tag{5}$$

then

$$n(h(T_n) - h(\theta)) \Rightarrow \frac{1}{2}\tau^2 h''(\theta)\chi_1^2$$

Proof Consider the Taylor expansion of $h(T_n)$ around $h(\theta)$ upto the second term.

$$h(T_n) = h(\theta) + (T_n - \theta)h'(\theta) + \frac{1}{2}(T_n - \theta)^2(h''(\theta) + R_n)$$

where $R_n \to 0$ as $T_n \to \theta$.

Step 1: It follows from equation (5) that $T_n \to \theta$ in probability.
Hence $R_n \to 0$ in probability. This implies $h''(\theta) + R_n \to h''(\theta)$ in probability.

Step 2: $\frac{1}{\tau^2} n(T_n - \theta)^2 \Rightarrow \chi_1^2$.
This follows from equation (5) after dividing by τ and squaring a standard normal random variable.

Step 3: The result follows by applying Slutsky's theorem to $n(h(T_n) - h(\theta))$.
$n(h(T_n) - h(\theta)) = n(T_n - \theta)^2 \times (h''(\theta) + R_n)$ since $h'(\theta) = 0$.
Let $Y_n = \tau^2(h''(\theta) + R_n)$ and $X_n = \frac{1}{\tau^2} n(T_n - \theta)^2$.
$X_n \Rightarrow X$ and $Y_n \to c$ in probability where $c = \tau^2 h''(\theta), X \sim \chi_1^2$.
By Slutsky's theorem, $n(h(T_n) - h(\theta)) = Y_n X_n \Rightarrow cX$.
The distribution of cX is $\tau^2 h''(\theta)\chi_1^2$.

Example 3'(Binomial Variance at $p = 1/2$) For $h(p) = p(1-p)$, we have at $p = 1/2$, $h'(1/2) = 0$ and $h''(1/2) = -2$. Hence from theorem 6, at $p = 1/2$,

$$n\left[T_n(1-T_n) - \frac{1}{4}\right] \Rightarrow -\frac{1}{4}\chi_1^2 \tag{6}$$

Although the equation (6) might appear strange, note that $T_n(1-T_n) \leq 1/4$, so the left side is always negative. An equivalent form is

$$4n\left[\frac{1}{4} - T_n(1-T_n)\right] \Rightarrow \chi_1^2$$

We now present a result on multivariate Delta method without proof.

Theorem 7 (Multivariate Delta Method) *Let* $(X_{1\nu}, \cdots, X_{s\nu}), \quad , \nu = 1, \cdots, n$ *be* n *independent* s*-tuples of random variables with* $\mathrm{E}(X_{i\nu}) = \xi_i$ *and* $\mathrm{Cov}(X_{i\nu}, X_{j\nu}) =$

σ_{ij}. Let $\bar{X}_i = \sum_{\nu=1}^{n} X_{i\nu}/n$, and suppose that h is a real valued function of s arguments with continuous first partial derivatives. Then

$$\sqrt{n}\left[h(\bar{X}_1, \cdots, \bar{X}_s) - h(\xi_1, \cdots, \xi_s)\right] \Rightarrow \mathcal{N}(0, \upsilon^2), \quad \text{where} \quad \upsilon^2 = \sum_{i=1}^{s}\sum_{j=1}^{s} \sigma_{ij} \frac{\partial h}{\partial \xi_i}\frac{\partial h}{\partial \xi_j}$$

Example 4 (Variance of Variance estimator) Suppose $X_1, \cdots, X_n$ are iid random variables with mean μ and variance σ^2. We are interested in the joint distribution of $s^2 = \frac{1}{n}\sum(X_i - \bar{X})^2$, the estimator of σ^2. Denoting $\mathrm{E}(X^k)$ by m_k, we have

$$\begin{aligned}
\mathrm{E}(\bar{X}) &= m_1 \\
\mathrm{E}(\bar{X}^2) &= m_2 \\
\mathrm{Cov}(\bar{X}, \bar{X}^2) &= (m_3 - m_1 m_2)/n \\
\mathrm{Var}(\bar{X}) &= (m_2 - m_1^2)/n \\
\mathrm{Var}(\bar{X}^2) &= (m_4 - m_2^2)/n
\end{aligned}$$

The parameter of interest is $\sigma^2 = h(m_1, m_2) = m_2 - m_1^2$. The derivatives of h are $\frac{\partial h}{\partial m_1} = -2m_1$ and $\frac{\partial h}{\partial m_2} = 1$.

$$\sqrt{n}\left[h(\bar{X}, \bar{X}^2) - h(m_1, m_2)\right] \Rightarrow \mathcal{N}(0, \upsilon^2), \quad \text{where}$$

$$\begin{aligned}
\upsilon^2 = \mathbf{D}h\Sigma\mathbf{D}h^T &= \begin{pmatrix} -2m_1 & 1 \end{pmatrix} \begin{pmatrix} m_2 - m_1^2 & m_3 - m_1 m_2 \\ m_3 - m_1 m_2 & m_4 - m_2^2 \end{pmatrix} \begin{pmatrix} -2m_1 \\ 1 \end{pmatrix} \\
&= -4m_1^4 + 8m_1^2 m_2 + m_4 - m_2^2 - 4m_1 m_3
\end{aligned}$$

The central limit theorem and the delta method will prove very useful in deriving asymptotic distribution results about functions of sample moments.

Example 9 (Distribution of sample T statistic) Suppose $X_1, X_2, \cdots, X_n$ are iid with $\mathrm{E}(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2 < \infty$. Define $s_n^2 = \frac{1}{n}\sum(X_i - \bar{X})^2$, and let

$$T_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{s_n}.$$

Letting $A_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$ and $B_n = \sigma/s_n$, we obtain $Tn = A_n B_n$. Therefore, since $A_n \Rightarrow \mathrm{N}(0,1)$ by the central limit theorem and $B_n \xrightarrow{P} 1$ by the weak law of large numbers, Slutsky's theorem implies that $T_n \Rightarrow \mathcal{N}(0,1)$. In other words, T statistics are asymptotically normal under the null hypothesis.

5.2.2 Asymptotic Normality of MLE

It will be necessary to review a few facts regarding Fisher information before we proceed. For a density (or mass) function $f_\theta(x)$, we define the Fisher information function to be

$$I(\theta) = E_\theta \left\{ \frac{d}{d\theta} \log f_\theta(X) \right\}^2 \tag{7}$$

If $\eta = g(\theta)$ for some invertible and differentiable function $g(\Delta)$, then since

$$\frac{d}{d\eta} = \frac{d\theta}{d\eta}\frac{d}{d\theta} = \frac{1}{g'(\theta)}\frac{d}{d\theta} \tag{8}$$

by the chain rule, we conclude that

$$I(\eta) = \frac{I(\theta)}{\{g'(\theta)\}^2} \tag{9}$$

Loosely speaking, $I(\theta)$ is the amount of information about θ contained in a single observation from the density $f_\theta(x)$.
Suppose that $f_\theta(x)$ is twice differentiable with respect to θ and that the operations of differentiation and integration may be interchanged in the following sense:

$$E_\theta \left\{ \frac{d}{d\theta} \log f_\theta(X) \right\} = E_\theta \left\{ \frac{\frac{d}{d\theta} f_\theta(X)}{f_\theta(X)} \right\} = \int \frac{d}{d\theta} f_\theta(X) dx = \frac{d}{d\theta} \int f_\theta(X) dx = \frac{d}{d\theta} 1 = 0 \tag{10}$$

$$E_\theta \left\{ \frac{d}{d\theta} \frac{\frac{d}{d\theta} f_\theta(X)}{f_\theta(X)} \right\} = E_\theta \left\{ \frac{\frac{d^2}{d\theta^2} f_\theta(X)}{f_\theta(X)} \right\} - I(\theta) = \frac{d^2}{d\theta^2} \int f_\theta(X) dx - I(\theta) = -I(\theta) \tag{11}$$

Equations (10) and (11) give two additional expressions for $I(\theta)$. From Equation (10) follows

$$I(\theta) = \mathrm{Var}_\theta \left\{ \frac{d}{d\theta} \log f_\theta(X) \right\} \tag{12}$$

and Equation (11) implies

$$I(\theta) = -E_\theta \left\{ \frac{d^2}{d\theta^2} \log f_\theta(X) \right\}. \tag{13}$$

In many cases, Equation (13) is the easiest form of the information to work with. Equations (12) and (13) make clear a helpful property of the information, namely that for independent random variables, the information about θ contained in the joint sample is simply the sum of the individual information components. In particular, if we have an iid sample from $f_\theta(x)$, then the information about θ equals $nI(\theta)$. The reason that we need the Fisher information is that we will show that under certain regularity conditions,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \Rightarrow \mathcal{N}\left\{ 0, \frac{1}{I(\theta_0)} \right\}, \tag{14}$$

where $\hat{\theta}_n$ is the MLE.

Example 1 (Poisson case) Suppose that $X_1, \cdots, X_n$ are iid Poisson(θ_0) random variables. Then the likelihood equation has a unique root, namely $\hat{\theta}_n = \bar{X}_n$, and we know that by the central limit theorem $\sqrt{n}(\hat{\theta}_n - \theta_0) \Rightarrow \mathcal{N}(0, \theta_0)$. However, the Fisher information for a single observation in this case is

$$-E_\theta \left\{ \frac{d^2}{d\theta^2} \log f_\theta(X) \right\} = E_\theta \frac{X}{\theta^2} = \frac{1}{\theta} \tag{15}$$

Thus, in this example, equation (14) holds.

Rather than stating all of the regularity conditions necessary to prove Equation (12), we work backwards, figuring out the conditions as we go through the proof. The first step is to expand $l'(\hat{\theta}_n)$ in a power series around θ_0:

$$l'(\hat{\theta}_n) = l'(\theta_0) + (\hat{\theta}_n - \theta_0) l''(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)^2 l'''(\theta_n^*) \tag{16}$$

for some θ_n^* between $\hat{\theta}_n$ and θ_0. Clearly, the validity of Equation (16) hinges on the existence of a continuous third derivative of $l(\theta)$. Rewriting equation (16) gives

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{\sqrt{n}\{l'(\hat{\theta}_n) - l'(\theta_0)\}}{l''(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0) l'''(\theta_n^*)} = \frac{\frac{1}{\sqrt{n}}\{l'(\theta_0) - l'(\hat{\theta}_n)\}}{-\frac{1}{n} l''(\theta_0) - \frac{1}{2n}(\hat{\theta}_n - \theta_0) l'''(\theta_n^*)} \tag{17}$$

Let's consider the pieces of Equation (17) individually. If the MLE is consistent, then $l'(\hat{\theta}_n) \xrightarrow{P} 0$. If Equation (10) holds and $I(\theta_0) < \infty$, then

$$\frac{1}{\sqrt{n}} l'(\theta_0) = \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^{n} \frac{d}{d\theta} \log f_{\theta_0}(X_i) \right) \Rightarrow \mathcal{N}(0, I(\theta_0)) \tag{18}$$

by the central limit theorem and Equation (12). If Equation (11) holds, then

$$\frac{1}{n} l''(\theta_0) = \frac{1}{n} \sum_{i=1}^{n} \frac{d^2}{d\theta^2} \log f_{\theta_0}(X_i) \xrightarrow{P} -I(\theta_0) \tag{19}$$

by the weak law of large numbers and Equation (13). Finally, we would like to have the term involving $l'''(\theta_n^*)$ disappear, so clearly it is enough to show that $\frac{1}{n} l'''(\theta)$ is bounded in probability in a neighborhood of θ_0. Putting all of these facts together gives a theorem.

Theorem 8 *Let $\hat{\theta}_n$ denote a consistent root of the likelihood equation. Assume also that $\mathrm{l}'''(\theta)$ exists and is continuous, that equations (10) and (11) hold, and that $\frac{1}{n}\mathrm{l}'''(\theta)$ is bounded in probability in a neighborhood of θ_0. Then if $0 < I(\theta_0) < \infty$, (14) holds.*

The theorem is proved by noting that under the stated regularity conditions, $l'(\hat{\theta}_n) \xrightarrow{P} 0$ so that the numerator in (17) converges in distribution to $\mathcal{N}\{0, I(\theta_0)\}$ by Slutsky's theorem. Furthermore, the denominator in (17) converges to $I(\theta_0)$, so another application of Slutsky's theorem gives the desired result.

Sometimes, it is not possible to find an exact zero of $l'(\theta)$. One way to get a numerical approximation to a zero of $l'(\theta)$ is to use Newton's method, in which we start at a point θ_0 and then set

$$\theta_1 = \theta_0 - \frac{l'(\theta_0)}{l''(\theta_0)}. \tag{20}$$

Ordinarily, after finding θ_1 we would set θ_0 equal to θ_1 and apply Equation (20) iteratively. However, we may show that by using a single step of Newton's method, starting from a $\sqrt{n}$-consistent estimator of θ_0, we may obtain an estimator with the same asymptotic distribution as $\hat{\theta}_n$. The proof of the following theorem is left as an exercise:

Theorem 9 *Suppose that $\tilde{\theta}_n$ is any $\sqrt{n}$-consistent estimator of θ_0 (i.e., $\sqrt{n}(\tilde{\theta}_n - \theta_0)$ is bounded in probability). Then under the conditions of Theorem 7, if we set*

$$\delta_n = \tilde{\theta}_n - \frac{\mathrm{l}'(\tilde{\theta}_n)}{\mathrm{l}''(\tilde{\theta}_n)} \tag{21}$$

then

$$\sqrt{n}(\delta_n - \theta_0) \Rightarrow \mathcal{N}(0, \frac{1}{I(\theta_0)}) \tag{22}$$

5.3 Relative efficiency

we have considered various cases where the distribution of estimators converged at rate $\sqrt{n}$ to the normal distribution. If there are multiple estimators of the same parameter with this property, then all of them are $\sqrt{n}$ consistent. We can use the asymptotic variance as a means of comparing such estimators. This is the idea of asymptotic relative efficiency.

Definition 2 *If two estimators W_n and V_n satisfy*

$$\sqrt{n}[V_n - \theta] \Rightarrow \mathcal{N}(0, \sigma_V^2)$$
$$\sqrt{n}[W_n - \theta] \Rightarrow \mathcal{N}(0, \sigma_W^2)$$

The **asymptotic relative efficiency(ARE)** *of V_n with respect to W_n is*

$$\mathrm{ARE}(V_n, W_n) = \frac{\sigma_W^2}{\sigma_V^2} \tag{23}$$

Example 1(ARE of Poisson Estimators) Suppose $X_1, \cdots, X_n$ are iid Poisson(λ), and we are interested in estimating $\tau = P_\lambda(X_1 = 0) = \exp(-\lambda)$. For example number of customers who come into a bank in a given time period is modeled as

a Poisson random variable and we are interested in the probability that no one will enter the bank in one time period. A natural (but somewhat naive) estimator comes from defining $Y_i = I(X_i = 0)$. The Y_is are iid Bernoulli($\exp(-\lambda)$) and hence it follows that

$$\sqrt{n}(\bar{Y}_n - \exp(-\lambda)) \Rightarrow \mathcal{N}(0, \exp(-\lambda)(1 - \exp(-\lambda)))$$

Additionally,the MLE of $\exp(-\lambda)$ is $\hat{\tau} = \exp(-\hat{\lambda})$ where $\hat{\lambda} = \bar{X}_n$ is the MLE of λ. Using the Delta method, we have

$$\sqrt{n}(\hat{\tau} - \tau) \Rightarrow \mathcal{N}(0, \lambda \exp(-2\lambda))$$

The ARE of $\bar{Y}_n$ wrt the MLE is

$$\text{ARE}(\bar{Y}, \exp(-\bar{X})) = \frac{\lambda \exp(-2\lambda)}{\exp(-\lambda)(1 - \exp(-\lambda))} = \frac{\lambda \exp(-\lambda)}{(1 - \exp(-\lambda))}$$

Examination of this function shows that it is strictly decreasing with a maximum of 1 at $\lambda = 0$ and tailing off rapidly (< 0.1 when $\lambda = 4$) to 0 as $\lambda \to \infty$. So in this case the MLE is better in terms of ARE.

5.4 Asymptotic Bias and Efficiency

(CB 470-471)

There are two ways in which we can look at the bias as sample size goes to infinity. We can look at the finite sample bias $\text{Bias}(T_n)$ and take the limit as $n \to \infty$. This is called the limiting bias. We can also look for a suitably scaled version of the estimator converges in distribution to a non-degenerate random variable and look at the bias of that limiting distribution. This is the asymptotic bias. Here are the precise definitions:

Definition 3 *An estimator T_n of $\tau(\theta)$ is unbiased in the limit, if* $\lim_{n\to\infty} \text{E}(T_n) = \tau(\theta)$.

Definition 4 *For an estimator T_n, suppose that $k_n(T_n - \tau(\theta)) \Rightarrow \mathcal{H}$. The estimator T_n is asymptotically unbiased if the expectation of $\mathcal{H}$ is zero.*

Example 1 (Asymptotically biased estimator) Let $X_1, \cdots, X_n$ are iid $U(0, \theta)$.

$$\text{The MLE of} \quad \theta \quad \text{is} \quad X_{(n)} \tag{24}$$

$$P(X_{(n)} \leq a) = (a/\theta)^n \quad \text{and} \quad \text{E}(X_{(n)}) = \theta \tag{25}$$

Hence $P(n(\theta - X_{(n)}) \leq a) = P(X_{(n)} \geq \theta - a/n) = 1 - (1 - a/n\theta)^n \to 1 - e^{-a/\theta}$. Thus $n(\theta - X_{(n)}) \Rightarrow \text{Exp}(\frac{1}{\theta})$. The expectation of the limiting random variable is not zero. So $X_{(n)}$ is not asymptotically unbiased. From (25) $X_{(n)}$ is unbiased in the limit.

Similar concepts exist for efficiency, which concerned with the asymptotic variance of the estimator.

Definition 5 *For an estimator T_n, if $\lim_{n\to\infty} k_n \mathrm{Var}(T_n) = \tau^2 < \infty$, where k_n is a sequence of constants, then τ^2 is called the limiting variance.*

Definition 6 *For an estimator T_n, suppose that $k_n(T_n - \tau(\theta)) \Rightarrow \mathcal{H}$. Then $Var(H)$ is called the asymptotic variance of T_n.*

In most cases these two are the same. But in complicated cases, this may not hold. It is always the case that the asymptotic variance is smaller than the limiting variance (Lehmann and Casella Sec 6.1).

Example 2 Let us consider the mean $\bar{X}_n$ of n iid normal observations with mean μ and variance σ^2. Suppose we are interested in estimating $\frac{1}{\mu}$ and we use the estimator $T_n = \frac{1}{\bar{X}_n}$. For each finite n the distribution of $\sqrt{n}\bar{X}_n$ is $\mathcal{N}(0, \sigma^2)$.

$$\mathrm{Var}(\sqrt{n}T_n) = \infty, \quad \text{by direct integral of} \quad \frac{1}{x^2} \quad \text{with respect to the normal pdf.} \tag{26}$$

So, the limiting variance of T_n is infinity. On the other hand, by Delta method,

$$\sqrt{n}(T_n - \frac{1}{\mu}) \Rightarrow \mathcal{N}(0, \frac{\sigma^2}{\mu^4})$$

So the asymptotic variance of T_n is $\frac{\sigma^2}{\mu^4}$.

In the spirit of the Cramer Rao lower bound, there is an optimal asymptotic variance.

Definition 7 *A sequence of estimators W_n is asymptotically efficient for a parameter $\tau\theta$ if $\sqrt{n}(W_n - \tau(\theta) \Rightarrow \mathcal{N}(0, \nu(\theta)$ and*

$$\nu(\theta) = \frac{(\tau'(\theta)^2)}{\mathrm{E}_\theta((\frac{\partial}{\partial\theta}\log f(X \mid \theta))^2)} = \frac{(\tau'(\theta)^2)}{I(\theta)}, \tag{27}$$

that is the asymptotic variance of W_n achieves the Cramer-Rao lower bound.

For a long time it was believed that if

$$\sqrt{n}(W_n - \tau(\theta) \Rightarrow \mathcal{N}(0, \nu(\theta), \tag{28}$$

then

$$\nu(\theta) \geq \frac{(\tau'(\theta)^2)}{I(\theta)} \tag{29}$$

under regularity conditions on the densities. This belief was shattered by the example (due to hodges; see LaCam 1953) below:

Example 3 (Superefficient Estimator): Let $X_1, \cdots, X_n$ be iid $\mathcal{N}(\theta, 1)$ and the parameter of interest is θ. In this case, $h(\theta) = \theta$, and

$$\begin{aligned} I(\theta) &= \mathrm{E}_\theta((\frac{\partial}{\partial\theta}\log f(X \mid \theta))^2) \\ &= \mathrm{E}_\theta((\frac{\partial}{\partial\theta}\frac{1}{2}(X-\theta)^2)^2) \\ &= \mathrm{E}_\theta(X-\theta)^2 \\ &= 1 \end{aligned}$$

Thus equation(29) reduces $\nu(\theta) \geq 1$. Now consider the sequence of estimators

$$T_n = \begin{cases} \bar{X} & \text{if} \quad \mid \bar{X} \mid \geq 1/n^{1/4} \\ a\bar{X} & \text{if} \quad \mid \bar{X} \mid < 1/n^{1/4} \end{cases}$$

$$\text{Then,} \quad \sqrt{n}(T_n - \theta) \Rightarrow \mathcal{N}(0, \nu(\theta)), \tag{30}$$

$$\text{where} \quad \nu(\theta) = 1 \quad \text{when} \quad \theta \neq 0 \quad \text{and} \quad \nu(\theta) = a^2 \quad \text{when} \quad \theta = 0. \tag{31}$$

If $a < 1$, inequality (29) is violated at $\theta = 0$.

This phenomenon is quite common and is called superefficiency. There will typically exist estimators satisfying (28) but with $\nu(\theta)$ violating (29) at least for some values of θ. However, it was shown by LaCam(1953) that for any sequence of estimators satisfying (28), the set S of points of super-efficiency has Lebesgue measure zero.

5.5 Results and concepts from probability

1. Convergence almost surely(a.s), convergence in probability(P), convergence in distribution(d).
2. a.s.$\Rightarrow$ P$\Rightarrow$ d. But not the other way around.
3. Strong and weak laws of large numbers.
4. Central Limit Theorem
5. Chebyshev and Jensen inequalities
6. Continuous mapping Theorem: (pg 24 od Serfling) g is a continuous function. Then,
   (a) $X_n \Rightarrow X$ implies $g(X_n) \Rightarrow g(X)$.
   (b) $X_n \stackrel{P}{\rightarrow} X$ implies $g(X_n) \stackrel{P}{\rightarrow} g(X)$
   (c) $X_n \stackrel{a.s.}{\rightarrow} X$ implies $g(X_n) \stackrel{a.s.}{\rightarrow} g(X)$
7. Slutsky's Theorem:(pg 19 of Serfling) $X_n \Rightarrow X$ and $Y_n \stackrel{P}{\rightarrow} c$, where c is a constant. Then
   (a) $X_n + Y_n \Rightarrow X + c$
   (b) $X_n Yn \Rightarrow cX$
   (c) $X_n / Yn \Rightarrow X/c$ provided $c \neq 0$.