As seen in the preceding topic, the MLE is not necessarily even consistent, so the title of this topic is slightly misleading — however, “Asymptotic normality of the consistent root of the likelihood equation” is a bit too long! It will be necessary to review a few facts regarding Fisher information before we proceed.

As you are probably already aware, for a density (or mass) function \( f_\theta(x) \), we define the Fisher information function to be

\[
I(\theta) = E_\theta \left\{ \frac{d}{d\theta} \log f_\theta(X) \right\}^2. \tag{90}
\]

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

\[
\frac{d}{d\eta} = \frac{d\theta}{d\eta} \frac{d}{d\theta} = \frac{1}{g'(\theta)} \frac{d}{d\theta}
\]

by the chain rule, we conclude that

\[
I(\eta) = \frac{I(\theta)}{(g'(\theta))^2}. \tag{91}
\]

Loosely speaking, \( I(\theta) \) is the amount of information about \( \theta \) contained in a single observation from the density \( f_\theta(x) \). However, this interpretation doesn’t always make sense — for example, it is possible to have \( I(\theta) = 0 \) for a very informative observation (see Example 7.2.1 on page 462 of Lehmann).

Although we do not dwell on this fact in this course, expectation may be viewed as integration. Suppose that \( f_\theta(x) \) is twice differentiable with respect to \( \theta \) and that the operations of differentiation and integration may be interchanged in the following sense:

\[
\frac{d}{d\theta} E_\theta f_\theta(X) = E_\theta \frac{d}{d\theta} f_\theta(X) \tag{92}
\]

and

\[
\frac{d^2}{d\theta^2} E_\theta f_\theta(X) = E_\theta \frac{d^2}{d\theta^2} f_\theta(X). \tag{93}
\]

Equations (92) and (93) give two additional expressions for \( I(\theta) \). From Equation (92) follows

\[
I(\theta) = \text{Var}_\theta \left\{ \frac{d}{d\theta} \log f_\theta(X) \right\}, \tag{94}
\]

and Equation (93) implies

\[
I(\theta) = -E_\theta \left\{ \frac{d^2}{d\theta^2} \log f_\theta(X) \right\}. \tag{95}
\]
In many cases, Equation (95) is the easiest form of the information to work with.

Equations (94) and (95) make clear a helpful property of the information, namely that for independent random variables, the information about $\theta$ 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 $\theta$ equals $nI(\theta)$.

The reason that we need the Fisher information is that we will show that under certain regularity conditions, $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{p} N\left(0, \frac{1}{n} \frac{1}{I(\theta_0)} \right)$, where $\hat{\theta}_n$ is the consistent root of the likelihood equation.

**Example 27.1** Suppose that $X_1, \ldots, X_n$ are iid Poisson($\theta_0$) 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 $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \theta_0)$. However, the Fisher information for a single observation in this case is

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

Thus, in this example, equation (96) holds.

Rather than stating all of the regularity conditions necessary to prove Equation (96), we work backwards, figuring out the conditions as we go through the proof. The first step is to expand $\ell'(\hat{\theta}_n)$ in a power series around $\theta_0$:

$$\ell'(\hat{\theta}_n) = \ell'(\theta_0) + (\hat{\theta}_n - \theta_0)\ell''(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)^2\ell'''(\theta^*_n)$$

(97)

for some $\theta^*_n$ between $\hat{\theta}_n$ and $\theta_0$. Clearly, the validity of Equation (97) hinges on the existence of a continuous third derivative of $\ell(\theta)$. Rewriting equation (97) gives

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{\sqrt{n}\{\ell'(\hat{\theta}_n) - \ell'(\theta_0)\}}{\ell''(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)\ell'''(\theta^*_n)} = \frac{-\frac{1}{\sqrt{n}}\{\ell'(\theta_0) - \ell'(\hat{\theta}_n)\}}{-\frac{1}{n}\ell''(\theta_0) - \frac{1}{2n}(\hat{\theta}_n - \theta_0)\ell'''(\theta^*_n)}.$$  

(98)

Let’s consider the pieces of Equation (98) individually. If the conditions of Theorem 26.3 are met, then $\ell'(\hat{\theta}_n) \overset{p}{\to} 0$. If Equation (92) holds and $I(\theta_0) < \infty$, then

$$\frac{1}{\sqrt{n}}\ell'(\theta_0) = \sqrt{n} \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{d}{d\theta} \log f_{\theta_0}(X_i) \right\} \xrightarrow{p} N\{0, I(\theta_0)\}$$

by the central limit theorem and Equation (94). If Equation (93) holds, then

$$\frac{1}{n}\ell''(\theta_0) = \frac{1}{n} \sum_{i=1}^{n} \frac{d^2}{d\theta^2} \log f_{\theta_0}(X_i) \overset{p}{\to} -I(\theta_0)$$

by the weak law of large numbers and Equation (95). Finally, we would like to have the term involving $\ell'''(\theta^*_n)$ disappear, so clearly it is enough to show that $\frac{1}{\sqrt{n}}\ell'''(\theta)$ is bounded in probability in a neighborhood of $\theta_0$.

Putting all of these facts together gives a theorem.

**Theorem 27.1** Suppose that the conditions of Theorem 26.3 are satisfied, and let $\hat{\theta}_n$ denote a consistent root of the likelihood equation. Assume also that $\ell'''(\theta)$ exists and is continuous, that
equations (92) and (93) hold, and that \( \frac{1}{n} \ell''(\theta) \) is bounded in probability in a neighborhood of \( \theta_0 \). Then if \( 0 < I(\theta_0) < \infty \),

\[
\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{L} N \left( 0, \frac{1}{I(\theta_0)} \right).
\]

The theorem is proved by noting that under the stated regularity conditions, \( \ell'(\hat{\theta}_n) \xrightarrow{P} 0 \) so that the numerator in (98) converges in distribution to \( N(0, I(\theta_0)) \) by Slutsky’s theorem. Furthermore, the denominator in (98) converges to \( I(\theta_0) \), so Slutsky’s theorem gives the desired result.

Sometimes, it is not possible to find an exact zero of \( \ell'(\theta) \). One way to get a numerical approximation to a zero of \( \ell'(\theta) \) is to use Newton’s method, in which we start at a point \( \theta_0 \) and then set

\[
\theta_1 = \theta_0 - \frac{\ell'(\theta_0)}{\ell''(\theta_0)}.
\]

Ordinarily, after finding \( \theta_1 \) we would set \( \theta_0 \) equal to \( \theta_1 \) and apply Equation (99) iteratively.

However, we may show that by using a single step of Newton’s method, starting from a \( \sqrt{n} \)-consistent estimator of \( \theta_0 \), 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 27.2** Suppose that \( \hat{\theta}_n \) is any \( \sqrt{n} \)-consistent estimator of \( \theta_0 \) (i.e., \( \sqrt{n}(\hat{\theta}_n - \theta_0) \) is bounded in probability). Then under the conditions of Theorem 27.1, if we set

\[
\delta_n = \hat{\theta}_n - \frac{\ell'(\hat{\theta}_n)}{\ell''(\hat{\theta}_n)},
\]

then

\[
\sqrt{n}(\delta_n - \theta_0) \xrightarrow{L} N \left( 0, \frac{1}{I(\theta_0)} \right).
\]

**Problems for Topic 27**

**Problem 27.1** Do Problems 2.1 on p. 553 and 2.12 on p. 555.

**Problem 27.2** (a) Show that under the conditions of Theorem 27.1, including \( 0 < I(\theta_0) < \infty \), then if \( \hat{\theta}_n \) is a consistent root of the likelihood equation, \( P_{\theta_0}(\hat{\theta}_n \text{ is a local maximum}) \to 1 \).

(b) Using the result of part (a), show that for any two sequences \( \hat{\theta}_{1n} \) and \( \hat{\theta}_{2n} \) of consistent roots of the likelihood equation, \( P_{\theta_0}(\hat{\theta}_{1n} = \hat{\theta}_{2n}) \to 1 \).

**Problem 27.3** Prove Theorem 27.2.

**Hint:** Start with \( \sqrt{n}(\delta_n - \theta_0) = \sqrt{n}(\delta_n - \hat{\theta}_n) + \sqrt{n}(\hat{\theta}_n - \theta_0) \), then expand \( \ell'(\hat{\theta}_n) \) in a Taylor series about \( \theta_0 \) and rewrite \( \sqrt{n}(\hat{\theta}_n - \theta_0) \) using this expansion.

**Problem 27.4** Suppose that the following is a random sample from a logistic density with cdf \( F_\theta(x) = (1 + \exp(\theta - x))^{-1} \) (I’ll cheat and tell you that I used \( \theta = 2 \).)

| 1.0944 | 6.4723 | 3.1180 | 3.8318 | 4.1262 |
| 1.2853 | 1.0439 | 1.7472 | 4.9483 | 1.7001 |
| 1.0422 | 0.1690 | 3.6111 | 0.9970 | 2.9438 |

78
(a) Evaluate the unique root of the likelihood equation numerically. Then, taking the sample median as our known \(\sqrt{n}\)-consistent estimator \(\hat{\theta}_n\) of \(\theta\), evaluate the estimator \(\delta_n\) in equation (100) numerically.

(b) Find the asymptotic distributions of \(\sqrt{n}(\hat{\theta}_n - \theta_0)\) and \(\sqrt{n}(\delta_n - 2)\). Then, simulate 200 samples of size \(n = 15\) from the logistic distribution with \(\theta = 2\). Find the sample variances of the resulting sample medians and \(\delta_n\)-estimators. How well does the asymptotic theory match reality?

---

**Topic 28. Efficient estimation**

*Lehmann §7.4; Ferguson §21*

In Theorem 27.1, we showed that the consistent root \(\hat{\theta}_n\) of the likelihood equation satisfies

\[
\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{L} N \left( 0, \frac{1}{I(\theta_0)} \right).
\]

In Theorem 27.2, we stated that if \(\hat{\theta}_n\) is a \(\sqrt{n}\)-consistent estimator of \(\theta_0\) and \(\delta_n = \hat{\theta}_n - \ell'(\hat{\theta}_n)/\ell''(\hat{\theta}_n)\), then

\[
\sqrt{n}(\delta_n - \theta_0) \xrightarrow{L} N \left( 0, \frac{1}{I(\theta_0)} \right). \tag{101}
\]

There seems to be something special about the limiting variance \(1/I(\theta_0)\), and in fact this is true.

Much like the Cramér-Rao lower bound states that (under some regularity conditions) an unbiased estimator of \(\theta_0\) cannot have a variance smaller than \(1/I(\theta_0)\), we may state (though we don’t prove it here) the following result:

**Theorem 28.1** Given an iid sample, if \(\delta_n\) is any estimator satisfying

\[
\sqrt{n}(\delta_n - \theta_0) \xrightarrow{L} N \{0, v(\theta_0)\}
\]

for all \(\theta_0\), where \(v(\theta)\) is continuous, then \(v(\theta) \geq 1/I(\theta)\) for all \(\theta\).

In other words, \(1/I(\theta)\) is in a sense the smallest possible asymptotic variance for a \(\sqrt{n}\)-consistent estimator. For this reason, we refer to any estimator \(\delta_n\) satisfying (101) for all \(\theta_0\) an efficient estimator.

One condition in Theorem 28.1 that may be a bit puzzling is the condition that \(v(\theta)\) be continuous. If this condition is dropped, then a well-known counterexample, due to Hodges, exists:

**Example 28.1** Suppose that \(\delta_n\) is an efficient estimator of \(\theta_0\). Then if we define

\[
\delta_n^* = \begin{cases} 
0 & \text{if } n(\delta_n)^4 < 1 \\
\delta_n & \text{otherwise}
\end{cases}
\]

it is possible to show (Problem 28.1) that \(\delta_n^*\) is superefficient in the sense that

\[
\sqrt{n}(\delta_n^* - \theta_0) \xrightarrow{L} N \left( 0, \frac{1}{I(\theta_0)} \right)
\]

for all \(\theta_0 \neq 0\) but \(\sqrt{n}(\delta_n^* - \theta_0) \xrightarrow{L} 0\) if \(\theta_0 = 0\).