Lecture 16: Estimation

Relevant textbook passages:

Larsen–Marx [2]: Sections 5.2–5.7

Note: I have rearranged the material this year, and these notes have been hastily put together. They should be edited, but I think they are still useable.

16.1 What makes an estimator a good estimator?

Recall:

- A random experiment has a set \( X \) of possible outcomes.

- \( \Theta \) is the set of parameters of the set of possible data generating processes for the model of the random experiment, or effectively the set of dgp.

- \( P_\theta \) is the probability measure on \( X \) corresponding to \( \theta \).

- \( f(x; \theta) \) is the pdf or pmf of the outcome \( X \) for the dgp \( \theta \).

- An estimator \( T: X \rightarrow \Theta \).

\[ T \text{ cannot depend on } \theta. \]

- So \( T \) is a random variable.

- But we want it to be related to \( \theta \), where \( \theta \) is the “true” dgp.

- An estimator \( T \) is unbiased if \( E T = \theta \).

Unbiasedness

An estimator is a function \( T: X \rightarrow \Theta \). What do we mean by \( E T \)?

Since the datum \( X \) is a random variable with pmf or pdf \( f(x; \theta) \), the expected value of \( T(X) \) depends on \( \theta \), which is unknown.

\[
E_\theta T(X) = \theta, \quad \text{where of course, } E_\theta T(X) = \int T(x) f(x; \theta) \, dx.
\]

Unbiased estimators need not exist.

16.1.1 Example (cf. Lehmann and Hodges [1, p. 247]) There is no unbiased estimator for the binomial odds ratio.

Suppose \( T_n \) is an estimator of \( p/(1 - p) \).

For \( n = 2 \),

\[
E T = T(0)(1 - p)^2 + T(1)(1 - p)p + T(2)p^2.
\]

Now unbiased would require \( E T = p/(1 - p) \rightarrow \infty \) as \( p \rightarrow 1 \), but \( E T \) is bounded.
The same idea works for \( n > 2 \).

\[
E T = \sum_{k=0}^{n} T(k) \binom{n}{k} p^k (1 - p)^{n-k}
\]

which is bounded above by \( \max\{T(k) : k = 1, \ldots, n\} \), and so \( \neq p/(1 - p) \) for \( p \) close to one. \( \square \)

Consistency

- Imagine independent replications of the experiment, and let \( T_n \) be the estimator of \( \theta \) based on \( n \) replications.
  - \( T \) (more properly the sequence of \( T_n \)'s) is **consistent** if
    \[
    \text{plim}_{n \to \infty} T_n = \theta.
    \]
  - That is for every \( \theta \in \Theta \) and \( \varepsilon > 0 \),
    \[
    P_\theta (|T_n - \theta| > \varepsilon) \to 0 \quad \text{as} \quad n \to \infty.
    \]
  - \( T \) is **strongly consistent** if
    \[
    P_\theta (T_n \to \theta) = 1.
    \]

Even if an estimator is biased, it may still be consistent. For example, we shall soon see that the MLE of the variance of a Normal is biased (by a factor of \( (n - 1)/n \), but is still consistent, as the bias disappears in the limit.

Efficiency

\( T \) is **efficient** if for \( \theta \in \Theta \), \( T \) has the minimum variance of any unbiased estimator,

\[
\text{Var}_\theta T = \min \{ \text{Var}_\theta T' : E_\theta T' = \theta \}
\]

Asymptotic normality

When \( X = R \), it would be nice if an appropriately normalized \( \tilde{T}_n \) satisfied

\[
\tilde{T}_n \xrightarrow{D} N(0, 1).
\]

This property is often used to (feebly) justify treating the estimator as a Normal random variable for moderate sample sizes.

### 16.2 Maximum Likelihood Estimators

The main reason we are interested in Maximum Likelihood Estimators is not that R. A. Fisher thought they were a good idea, but because of the following claim.

Claim: For a wide variety of models, MLEs are consistent, efficient, asymptotically normal, and often unbiased.

I will discuss the efficiency claim in some detail next time, and then give you some references for the consistency claim. For now, just trust me that MLES are worth investigating.
16.3 First order conditions for an extremum

In order to find MLEs, we first need to know how to find maximizers of a function.

If \( f: \mathbb{R}^n \rightarrow \mathbb{R} \) is differentiable, \( \hat{x} \) is interior to the domain of \( f \), \( \hat{x} \) (locally) maximizes \( f \), then
\[
\frac{\partial f(\hat{x}_1, \ldots, \hat{x}_n)}{\partial x_i} = 0 \quad (i = 1, \ldots, n).
\]

Figure 16.1. A nicely behaved maximum: \( f' = 0 \) and \( f'' < 0 \).

Unfortunately, these are also the first order conditions for a minimizer.

If \( f \) is concave, then these conditions are also sufficient for \( \hat{x} \) to be a maximizer of \( f \). One way to tell if \( f \) is concave is to check that the matrix of second partials
\[
\begin{bmatrix}
\vdots \\
\frac{\partial^2 f(x)}{\partial x_i \partial x_j} & \ldots \\
\vdots 
\end{bmatrix}
\]
is negative semidefinite. If this is new to you, you may want to look at Section 3 of my on-line notes on maximization.

One of the ways that a lot of numerical optimization is done is to numerically find places where the partial derivatives are all zero. That is, reduce the problem to finding zeros of a function. Newton’s method and various modifications of it are frequently used for this purpose.

16.4 The likelihood function for independent experiments

Often a random experiment is actually a sequence of \( n \) independent random experiments with the same likelihood, or a set of \( n \) independent observations of identically distributed random variables \( X_1, \ldots, X_n \). If \( R \) denotes the range of each \( X_i \), then the set \( S \) of experimental outcomes is \( R^n \), or better yet \( \bigcup_{n=1}^{\infty} R^n \).

Let \( X_1, X_2, \ldots, X_n \) be independent and identically distributed with common pmf or pdf \( f(x; \theta) \).

Given observations \( X_1 = x_1, \ldots, X_n = x_n \), the (joint) likelihood function is
\[
L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i; \theta).
\]
Taking logarithms gives
\[ \ln L(\theta; x_1, \ldots, x_n) = \sum_{i=1}^{n} \ln f(x_i; \theta). \]

Larsen–Marx [2, Comment, p. 284] say you should not treat the likelihood as a function of the data, but that is clearly nonsense, since it is a function of the data. The utility of the likelihood function is justified by what it tells you and how you can use, not by any a priori metaphysical principle.

16.4.1 Example (A single Bernoulli trial) Admittedly, there is not much to learn from a single Bernoulli trial. The likelihood function is
\[ L(p; x) = px + (1 - p)(1 - x) \quad (x = 0, 1). \]
When \( x = 1 \) this is maximized (subject to the constraint that \( 0 \leq p \leq 1 \)) at \( p = 1 \), and when \( x = 0 \) it is maximized at \( p = 0 \). Thus the maximum likelihood estimator is
\[ \hat{p}(x) = \begin{cases} 1 & x = 1 \\ 0 & x = 0. \end{cases} \]
The MLE has the virtue of being an unbiased estimator since
\[ E \hat{p}(X) = p \bar{p}(1) + (1 - p)\bar{p}(0) = p. \]
The question of consistency makes no sense here, since by definition, we are considering only one observation. If we had \( n \) observations, we would be in the realm of the Binomial distribution. The variance of \( \hat{p}(X) \) is \( p(1 - p) \). It is trivial to come up with a lower variance estimator—just choose a constant—but then the estimator would not be unbiased.

16.4.2 Example (Binomial \((n, p)\)) We saw last time that the MLE of \( p \) for a Binomial \((n, p)\) random variable \( X \) is just \( X/n \). This is unbiased and consistent (by the Law of Large Numbers).

But there is one more point I want to make. The likelihood function is
\[ L(p; k) = \binom{n}{k} p^k (1 - p)^{n-k}. \]
The leading term \( \binom{n}{k} \) is positive and independent of \( p \), and so it has no relevance to MLE, and it is often convenient to omit it, and just write
\[ L(p; k) \propto p^k (1 - p)^{n-k}. \]
where the symbol \( \propto \) is read “is proportional to.”

16.4.3 Example (Independent and identically distributed normals) Let \( X_1, \ldots, X_n \) be independent and identically distributed \( N(\mu, \sigma^2) \) random variables. Given the sample \( x_1, \ldots, x_n \), the likelihood function is
\[
L(\mu, \sigma^2; x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2} \left( \frac{x_i - \mu}{\sigma} \right)^2}
\]
Again, we may ignore constants and write
\[
L(\mu, \sigma^2; x_1, \ldots, x_n) \propto \sigma^{-n} \prod_{i=1}^{n} e^{-\frac{1}{2} \left( \frac{x_i - \mu}{\sigma} \right)^2}
\]
or, by taking logs we would get (up to a constant)

$$\ln L(\mu, \sigma^2; x_1, \ldots, x_n) = \frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^2.$$ 

To find the maximizer of the log-likelihood we set both partials $\frac{\partial}{\partial \mu}$ and $\frac{\partial}{\partial \sigma^2}$ to zero. Now

$$\frac{\partial}{\partial \mu} \ln L(\hat{\mu}, \hat{\sigma}^2) = \frac{1}{\hat{\sigma}^2} \sum_{i=1}^{n} (x_i - \hat{\mu})$$

and (treating $\sigma^2$ as a single symbol),

$$\frac{\partial}{\partial \sigma^2} \ln L(\hat{\mu}, \hat{\sigma}^2) = -\frac{n}{2} \frac{1}{\hat{\sigma}^2} + \frac{1}{2} \left( \frac{1}{\hat{\sigma}^2} \right)^2 \sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

Setting (1) to zero implies

$$\hat{\mu}_{\text{MLE}} = \frac{\sum_{i=1}^{n} x_i}{n},$$

That is, the MLE of $\mu$ is the sample average. Multiplying (2) by $2(\hat{\sigma}^2)^2$ and setting it to zero gives:

$$-n\hat{\sigma}^2 + \sum_{i=1}^{n} (x_i - \hat{\mu})^2 = 0,$$

or letting

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} (= \hat{\mu}),$$

we get

$$-n\hat{\sigma}^2 + \sum_{i=1}^{n} (x_i - \bar{x})^2 = 0,$$

so

$$\hat{\sigma}^2_{\text{MLE}} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}.$$

Now $\hat{\mu}_{\text{MLE}}$ is unbiased and consistent, but $\hat{\sigma}^2_{\text{MLE}}$ is biased. To see this, let’s compute its expectation. We start with the expectation of $(X_i - \bar{X})^2$. First, let

$$Z = \sum_{j \neq i} X_j.$$

Then $Z$ has mean $(n - 1)\mu$ and variance $(n - 1)\sigma^2$ as the sum of $n - 1$ independent $N(\mu, \sigma^2)$ rvs. Moreover

$$\mathbb{E} Z^2 = (n - 1)\mu^2 + (n - 1)\sigma^2,$$

since for any rv $\text{Var} Y = \mathbb{E}(Y^2) - (\mathbb{E} Y)^2$. Also note that $X_i$ and $Z$ are independent, so

$$\mathbb{E} X_i Z = (\mathbb{E} X_i) (\mathbb{E} Z) = (n - 1)\mu^2.$$
Finally observe that

\[ \bar{X} = \frac{X_i + Z}{n}. \]

Thus

\[
E((X_i - \bar{X})^2) = E \left( X_i - \frac{X_i + Z}{n} \right)^2 \\
= E \left( \frac{n-1}{n} X_i - \frac{1}{n} Z \right)^2 \\
= \frac{1}{n^2} E \left( (n-1)^2 X_i^2 - 2(n-1)X_iZ + Z^2 \right) \\
= \frac{1}{n^2} \left( (n-1)^2(\mu^2 + \sigma^2) - 2(n-1)^2\mu^2 + (n-1)^2\mu^2 + (n-1)\sigma^2 \right) \\
= \frac{1}{n^2} \left( [(n-1)^2 - 2(n-1)^2 + (n-1)^2] \mu^2 + [(n-1)^2 + (n-1)] \sigma^2 \right) \\
= \frac{1}{n^2} n(n-1)\sigma^2 \\
= \frac{n-1}{n} \sigma^2.
\]

It follows from (4) that

\[ E \sigma^2_{\text{MLE}} = \frac{n-1}{n} \sigma^2. \]

Thus \( \sigma^2_{\text{MLE}} \) is biased, but the bias tends to zero as \( n \to \infty \), so the estimator is consistent.

An unbiased estimate of \( \sigma^2 \) is given by

\[ S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = \frac{n}{n-1} \widehat{\sigma^2}_{\text{MLE}}. \]

Now go back and realize that the computation of the expectations depends only on the fact that the \( X_i \) are independent and identically distributed with mean \( \mu \) and variance \( \sigma^2 \), not that they are normal.

16.5 Estimating functions of parameters

Suppose I don’t care about \( \theta \) per se, but rather some function \( g(\theta) \). E.g., suppose I want to estimate the standard deviation of a normal and not its variance.

(This only makes sense if \( g \) is a one-to-one function of \( \theta \)—suppose \( g(\theta) = g(\theta') = \gamma \). What likelihood should I assign to \( \gamma \), \( f(x; \theta) \) or \( f(x; \theta') \)?)

If \( g \) is a one-to-one function of \( \theta \) define the likelihood of \( g(\theta) \) by

\[ L(g(\theta); x) = f(x; \theta). \]

Or

\[ L(\gamma; x) = f(x; g^{-1}(\gamma)). \]

Then the maximum likelihood estimator of \( g(\theta) \) is just \( \hat{g}_{\text{MLE}} \).
This property is sometimes referred to invariance.

So if $\hat{\theta}_{MLE}$ is the MLE of $\theta$, then $\frac{1}{\hat{\theta}_{MLE}}$ is the MLE of $\frac{1}{\theta}$. But be warned! If $\hat{\theta}_{MLE}$ is an unbiased estimator of $\theta$, then $\frac{1}{\hat{\theta}_{MLE}}$ is not an unbiased estimate of $\frac{1}{\theta}$. Why? Jensen’s Inequality.

Unless $\hat{\theta}_{MLE}$ is degenerate,

$$E \left( \frac{1}{\hat{\theta}_{MLE}} \right) \neq \frac{1}{E \hat{\theta}_{MLE}} = \frac{1}{\theta}.$$

Likewise if $\hat{\sigma}_{MLE}^2$ is an unbiased estimator of the variance, then $\hat{\sigma}_{MLE}$ is not an unbiased estimator of the standard deviation!

### 16.6 Sufficient statistics

I’ve already done things like write the likelihood function for a binomial in terms of $k$, the number of successes instead of the entire sequence $x_1, \ldots, x_n$ of successes and failures. That’s because $k$ is all that matters for the likelihood function. We’ll formalize and generalize this idea.

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed with common pdf $f(x; \theta)$. The likelihood function is

$$L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i; \theta).$$

Let $T = \psi(X_1, \ldots, X_n)$ be a statistic. It has a density $f_T(t; \theta)$. If the likelihood function factors as

$$L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i; \theta) = f_T(\psi(x_1, \ldots, x_n); \theta) b(x_1, \ldots, x_n),$$

that is if $\theta$ enters the likelihood function only through the distribution of $T$, then $T$ is called a **sufficient statistic** for $\theta$.

In terms of the log-likelihood, the condition for sufficiency is

$$\ln L(\theta; x_1, \ldots, x_n) = \sum_{i=1}^{n} \ln f(x_i; \theta) = \ln f_T(\psi(x_1, \ldots, x_n); \theta) + \ln b(x_1, \ldots, x_n),$$

Note that in order to maximize the likelihood function with respect to $\theta$, it suffices to maximize $f_T(\psi(x_1, \ldots, x_n); \theta)$.

#### 16.6.1 Example

In the normal case, the sample mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

and the unbiased estimate of the variance

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1},$$

are sufficient for the pair $(\mu, \sigma^2)$. To see this, write the log-likelihood function as

$$\ln L(\mu, \sigma^2; x_1, \ldots, x_n) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2. \quad (5)$$

Now

$$\sum_{i=1}^{n} (x_i - \mu)^2 = \sum_{i=1}^{n} (x_i^2 - 2\mu x_i + \mu^2) = \sum_{i=1}^{n} x_i^2 - 2\mu \sum_{i=1}^{n} x_i + n\mu^2$$

$$= n\bar{x}^2$$

[108x742]KC Border Estimation 16–7

[123x700]This property is sometimes referred to invariance.

So if $\hat{\theta}_{MLE}$ is the MLE of $\theta$, then $\frac{1}{\hat{\theta}_{MLE}}$ is the MLE of $\frac{1}{\theta}$. But be warned! If $\hat{\theta}_{MLE}$ is an unbiased estimator of $\theta$, then $\frac{1}{\hat{\theta}_{MLE}}$ is not an unbiased estimate of $\frac{1}{\theta}$. Why? Jensen’s Inequality.

Unless $\hat{\theta}_{MLE}$ is degenerate,

$$E \left( \frac{1}{\hat{\theta}_{MLE}} \right) \neq \frac{1}{E \hat{\theta}_{MLE}} = \frac{1}{\theta}.$$

Likewise if $\hat{\sigma}_{MLE}^2$ is an unbiased estimator of the variance, then $\hat{\sigma}_{MLE}$ is not an unbiased estimator of the standard deviation!

### 16.6 Sufficient statistics

I’ve already done things like write the likelihood function for a binomial in terms of $k$, the number of successes instead of the entire sequence $x_1, \ldots, x_n$ of successes and failures. That’s because $k$ is all that matters for the likelihood function. We’ll formalize and generalize this idea.

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed with common pdf $f(x; \theta)$. The likelihood function is

$$L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i; \theta).$$

Let $T = \psi(X_1, \ldots, X_n)$ be a statistic. It has a density $f_T(t; \theta)$. If the likelihood function factors as

$$L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i; \theta) = f_T(\psi(x_1, \ldots, x_n); \theta) b(x_1, \ldots, x_n),$$

that is if $\theta$ enters the likelihood function only through the distribution of $T$, then $T$ is called a **sufficient statistic** for $\theta$.

In terms of the log-likelihood, the condition for sufficiency is

$$\ln L(\theta; x_1, \ldots, x_n) = \sum_{i=1}^{n} \ln f(x_i; \theta) = \ln f_T(\psi(x_1, \ldots, x_n); \theta) + \ln b(x_1, \ldots, x_n),$$

Note that in order to maximize the likelihood function with respect to $\theta$, it suffices to maximize $f_T(\psi(x_1, \ldots, x_n); \theta)$.

#### 16.6.1 Example

In the normal case, the sample mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

and the unbiased estimate of the variance

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1},$$

are sufficient for the pair $(\mu, \sigma^2)$. To see this, write the log-likelihood function as

$$\ln L(\mu, \sigma^2; x_1, \ldots, x_n) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2. \quad (5)$$

Now

$$\sum_{i=1}^{n} (x_i - \mu)^2 = \sum_{i=1}^{n} (x_i^2 - 2\mu x_i + \mu^2) = \sum_{i=1}^{n} x_i^2 - 2\mu \sum_{i=1}^{n} x_i + n\mu^2$$

$$= n\bar{x}^2$$

[108x742]KC Border Estimation 16–7

[123x700]This property is sometimes referred to invariance.

So if $\hat{\theta}_{MLE}$ is the MLE of $\theta$, then $\frac{1}{\hat{\theta}_{MLE}}$ is the MLE of $\frac{1}{\theta}$. But be warned! If $\hat{\theta}_{MLE}$ is an unbiased estimator of $\theta$, then $\frac{1}{\hat{\theta}_{MLE}}$ is not an unbiased estimate of $\frac{1}{\theta}$. Why? Jensen’s Inequality.

Unless $\hat{\theta}_{MLE}$ is degenerate,

$$E \left( \frac{1}{\hat{\theta}_{MLE}} \right) \neq \frac{1}{E \hat{\theta}_{MLE}} = \frac{1}{\theta}.$$

Likewise if $\hat{\sigma}_{MLE}^2$ is an unbiased estimator of the variance, then $\hat{\sigma}_{MLE}$ is not an unbiased estimator of the standard deviation!
and
\[(n - 1)S^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - 2\bar{x} \sum_{i=1}^{n} x_i + n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2,\]

so
\[\sum_{i=1}^{n} x_i^2 = (n - 1)S^2 + n\bar{x}^2. \quad (7)\]

Substituting (7) into (6), we get
\[\sum_{i=1}^{n} (x_i - \mu)^2 = (n - 1)S^2 + n\bar{x}^2 - 2n\mu\bar{x} + n\mu^2,\]

so (5) becomes
\[
\ln L(\mu, \sigma^2; \bar{x}, S^2) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2} \frac{1}{\sigma^2} (n - 1)S^2 + n\bar{x}^2 - 2n\mu\bar{x} + n\mu^2
\]
\[= -\frac{n}{2} \left[ \ln(2\pi) + \ln(\sigma^2) + \frac{1}{\sigma^2} \left( \frac{n - 1}{n} S^2 - 2\mu\bar{x} + \bar{x}^2 + \mu^2 \right) \right]. \]

Not that for the purposes of MLE, the coefficient \(n/2\) and the constant \(\ln(2\pi)\) do not affect the location of the maximizer, so if we wish, we can discard them and simply work with
\[-\ln(\sigma^2) - \frac{1}{\sigma^2} \left( \frac{n - 1}{n} S^2 - 2\mu\bar{x} + \bar{x}^2 + \mu^2 \right)\]

From this expression, we can re-derive the maximum likelihood estimators of \(\mu\) and \(\sigma^2\). The first order conditions for a maximum are that the partial derivatives with respect to \(\mu\) and \(\sigma^2\) are zero. So at the point \((\mu, \sigma^2) = (\hat{\mu}, \hat{\sigma}^2)\)
\[
\frac{\partial}{\partial \mu} = -\frac{1}{\sigma^2} (-2\bar{x} + 2\hat{\mu}) = 0,
\]
which implies
\[\hat{\mu} = \bar{x},\]
and
\[
\frac{\partial}{\partial \sigma^2} = -\frac{1}{\sigma^2} + \frac{1}{(\sigma^2)^2} \left( \frac{n - 1}{n} S^2 - 2\mu\bar{x} + \bar{x}^2 + \mu^2 \right) = 0,
\]
which, after multiplying by \((\sigma^2)^2\), implies
\[\hat{\sigma}^2 = \frac{n - 1}{n} S^2.\]

Thankfully this agrees with our previous derivation. \(\square\)

**Bibliography**
