# Lecture 20: Theoretical Propositions about Estimation

**Relevant textbook passages:**

**Larsen–Marx [12]:** Sections 5.2–5.7

## 20.1 What makes an estimator a good estimator?

Last time we discussed the problem of estimating the probability of success in a Binomial data model, and found the maximum likelihood estimator of the probability $p$ of success is just the fraction of successes in the sample. This is certainly an intuitive estimator, and makes common sense. But there are other estimators we could consider. One example is to always estimate $p = 3/4$. This has the virtue that it is precise (has variance 0) and is computationally quite tractable. It is also clearly nonsense. But can we come up with criteria that we can use to choose among estimators when the answer is not so obvious. In this lecture, we will try to find desiderata for estimators, and investigate when maximum likelihood satisfies these criteria.

Recall:

- A random experiment has a set $\mathcal{X}$ 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 dgps.
- $P_\theta$ is the probability measure on $\mathcal{X}$ corresponding to $\theta$.

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

- An estimator $T \colon \mathcal{X} \to \Theta$.

[$T$ 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.

### 20.1.1 Unbiasedness

An estimator $T$ is **unbiased** if $\boldsymbol{E}\, T = \theta$. But what do we mean by $\boldsymbol{E}\, T$? **Larsen–Marx [12]:** § 5.4

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.

> The estimator $T$ is an unbiased estimator of $\theta$ if for every $\theta \in \Theta$
>
> $$\boldsymbol{E}_\theta\, T(X) = \theta, \qquad \text{where of course, } \boldsymbol{E}_\theta\, T(X) = \int T(x) f(x,\theta)\, dx.$$

Unfortunately, unbiased estimators need not exist.

**20.1.1 Example (cf. Lehmann and Hodges [10, p. 247])** There is no unbiased estimator for the Binomial odds ratio.

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

For $n = 2$,

$$\boldsymbol{E}\,T = T(0)(1-p)^2 + T(1)(1-p)p + T(2)p^2.$$

Now unbiased would require $\boldsymbol{E}\,T = p/(1-p) \to \infty$ as $p \to 1$, but $\boldsymbol{E}\,T$ is bounded.

The same idea works for $n > 2$.

$$\boldsymbol{E}\,T = \sum_{k=0}^{n} T(k)\binom{n}{k}p^k(1-p)^{n-k}$$

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

### 20.1.2 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

$$\operatorname*{plim}_{n\to\infty} T_n = \theta.$$

That is, for every $\theta \in \Theta$ and $\varepsilon > 0$,

$$P_\theta(\,|T_n - \theta| > \varepsilon) \to 0 \text{ as } n \to \infty.$$

- $T$ 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.

### 20.1.3 Efficiency

Since $T$ is a random variable, it has a variance. It would be desirable to keep that variance small. We say that an unbiased estimator $T$ is **efficient** if for $\theta \in \Theta$, $T$ has the minimum variance of any unbiased estimator,

$$\boldsymbol{Var}_\theta\,T = \min\{\,\boldsymbol{Var}_\theta\,T' : \boldsymbol{E}_\theta\,T' = \theta\}$$

### 20.1.4 Asymptotic normality

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

$$\tilde{T}_n \xrightarrow{\mathcal{D}} N(0,1).$$

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

## 20.2 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 data models $f(x, \theta)$, MLEs are consistent, efficient, asymptotically normal, and often unbiased.

I will discuss the efficiency claim in a moment, and then give you some references for the consistency claim. For now, just trust me that MLES are worth investigating.

## 20.3 ⋆ First order conditions for an extremum

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

If $f\colon \boldsymbol{R}^{\mathrm{n}} \to \boldsymbol{R}$ is differentiable, $\hat{x}$ is interior to the domain of $f$, $\hat{x}$ (locally) maximizes $f$, then

$$\frac{\partial f(\hat{x}_1, \dots, \hat{x}_n)}{\partial x_i} = 0 \qquad (i = 1, \dots, n).$$

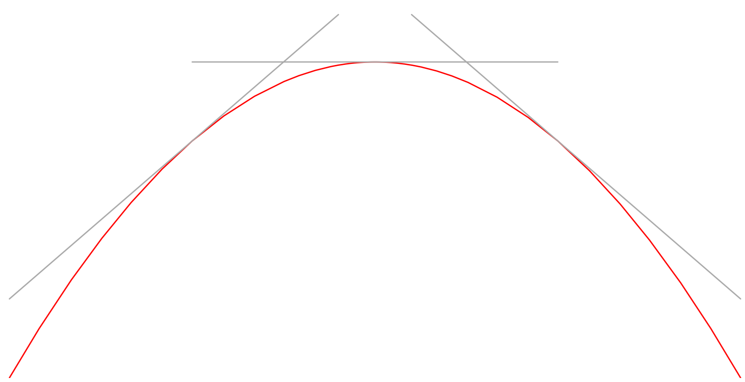

Figure 20.1. A nicely behaved maximum: $f' = 0$ and $f'' < 0$.

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

$$\begin{bmatrix} & \vdots & \\ \cdots & \frac{\partial^2 f(x)}{\partial x_i \partial x_j} & \cdots \\ & \vdots & \end{bmatrix}$$

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.

## 20.4 Estimating functions of parameters

Not in Larsen and Marx.

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, not its variance.

We first have to figure the likelihood as a function of $\gamma = g(\theta)$, not as a function of $\theta$. Unless $g$ is one-to-one, this doesn't make sense. For suppose $g(\theta) = g(\theta') = \gamma$. What likelihood should I assign to $\gamma$, $f(\boldsymbol{x};\theta)$ or $f(\boldsymbol{x};\theta')$?

But when $g$ is a one-to-one function of $\theta$, we can define the likelihood function for $\gamma = g(\theta)$. Observe that

$$f(\boldsymbol{x};\gamma) = f(\boldsymbol{x}; g^{-1}(\gamma)),$$

so the likelihood function $\tilde{L}$ for $\gamma$ is

$$\tilde{L}\big(\gamma;\boldsymbol{x}\big) = f(\boldsymbol{x};\gamma) = f(\boldsymbol{x}; g^{-1}(\gamma)) = L\big(g^{-1}(\gamma);\boldsymbol{x}\big).$$

By definition, $\hat{\theta}_{\text{MLE}}$ maximizes $L(\cdot;\boldsymbol{x})$ so that $\tilde{L}\big(\gamma;\boldsymbol{x}\big)$ is maximized when

$$g^{-1}(\gamma) = \hat{\theta}_{\text{MLE}},$$

so applying $g$ to both sides we get the following.

**20.4.1 Proposition** *For a one-to-one function $g$, the maximum likelihood estimate of $\gamma = g(\theta)$, that is, the value $\hat{\gamma}_{\text{MLE}}$ that maximizes $\tilde{L}(\gamma;\boldsymbol{x})$, is just $g(\hat{\theta}_{\text{MLE}})$,*

$$\hat{\gamma}_{\text{MLE}} = g(\hat{\theta}_{\text{MLE}}).$$

This property is referred to as **invariance**.

But be warned! If $\hat{\theta}_{\text{MLE}}$ is an unbiased estimator of $\theta$, then $g(\hat{\theta}_{\text{MLE}})$ is *not*, in general, an unbiased estimate of $g(\theta)$. Why? Usually, it's Jensen's Inequality. For instance, if $\hat{\sigma}^2_{\text{MLE}}$ is an unbiased estimator of the variance, then $\hat{\sigma}_{\text{MLE}} = \sqrt{\hat{\sigma}^2_{\text{MLE}}}$ is *not* an unbiased estimator of the standard deviation! Or if the random variable $\hat{\lambda}_{\text{MLE}}$ is an unbiased estimator of the rate $\lambda$ of an exponential distribution, then the random variable $\hat{\mu}_{\text{MLE}} = 1/\hat{\lambda}_{\text{MLE}}$ is the MLE of its mean $\mu = 1/\lambda$, but by Jensen's Inequality $\boldsymbol{E}\,\hat{\mu}_{\text{MLE}} = \boldsymbol{E}(1/\hat{\lambda}_{\text{MLE}}) > 1/\,\boldsymbol{E}(1/\hat{\lambda}_{\text{MLE}}) = 1/\lambda = \mu$.

## 20.5 Sufficient statistics

**Larsen–Marx [12]:** § 5.6

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 $x_1,\dots,x_n$ of successes and failures. That's because $k$ is all that matters for the likelihood function. We'll formalize and generalize this idea.

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

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

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

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

*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

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

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

**20.5.2 Example (Sufficient statistic for the Binomial)** The Binomial$(n, p)$ likelihood is

$$L(p; k) = \binom{n}{k} \cdot p^k (1-p)^{n-k} = \binom{n}{k} \cdot \left(p^{\frac{k}{n}} (1-p)^{1-\frac{k}{n}}\right)^n.$$

(Here $k$ plays the role of $x$ and $p$ is the abstract $\theta$.) Thus $T(k) = k/n$ is a sufficient statistic for $p$ since

$$L(p; k) = b(k) f_T\big(T(k); p\big),$$

where $b(k) = \binom{n}{k}$ and $f_T(t; p) = \big(p^t (1-p)^{1-t}\big)^n$. □

**20.5.3 Example (Sufficient statistic for the Normal)**

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

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

Now

$$\sum_{i=1}^{n} (x_i - \mu)^2 = \sum_{i=1}^{n} (x_i^2 - 2\mu x_i + \mu^2) = \sum_{i=1}^{n} x_i^2 - 2\mu \underbrace{\sum_{i=1}^{n} x_i}_{= n\bar{x}} + n\mu^2 \tag{3}$$

and

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

so

$$\sum_{i=1}^{n} x_i^2 = (n-1)S^2 + n\bar{x}^2. \tag{4}$$

Substituting (4) into (3), 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 (2) becomes

$$\begin{aligned}\ln L(\mu,\sigma^2;\bar{x},S^2) &= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2}\frac{1}{\sigma^2}\big((n-1)S^2 + n\bar{x}^2 - 2n\mu\bar{x} + n\mu^2\big)\\ &= -\frac{n}{2}\left[\ln(2\pi) + \ln(\sigma^2) + \frac{1}{\sigma^2}\left(\frac{n-1}{n}S^2 - 2\mu\bar{x} + \bar{x}^2 + \mu^2\right)\right]. \qquad (5)\end{aligned}$$

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 $(\mu,\sigma^2) = (\hat{\mu},\widehat{\sigma^2})$

$$\frac{\partial}{\partial\mu} = -\frac{1}{\widehat{\sigma^2}}(-2\bar{x} + 2\hat{\mu}) = 0,$$

which implies

$$\hat{\mu} = \bar{x},$$

and

$$\frac{\partial}{\partial\sigma^2} = -\frac{1}{\widehat{\sigma^2}} + \frac{1}{(\widehat{\sigma^2})^2}\left(\frac{n-1}{n}S^2 \underbrace{-2\hat{\mu}\bar{x} + \bar{x}^2 + \hat{\mu}^2}_{=0}\right) = 0,$$

which, after multiplying by $(\widehat{\sigma^2})^2$, implies

$$\widehat{\sigma^2} = \frac{n-1}{n}S^2.$$

Thankfully this agrees with our previous derivation. □

A family of densities $f(x;\theta)$ of the form

$$f(x;\theta) = a(\theta)b(x)\exp\left[\sum_{j=1}^{d} g_j(\theta)h_j(x)\right] \qquad (6)$$

is called an **exponential family of distributions** ([5, p. 161], [15, p. 195]). Some authors may rewrite this as

$$f(x;\theta) = \exp\left[\beta(x) + \alpha(\theta) + \sum_{j=1}^{d} g_j(\theta)h_j(x)\right] \qquad (6')$$

were $\alpha(\theta) = \ln a(\theta)$ and $\beta(x) = \ln b(x)$. Larsen–Marx [12, Exercise 5.6.9, p. 330] use the term **exponential form** for families of this sort.

Suppose $f$ has the form given by (6). Then

$$\begin{aligned} f(x_1;\theta)f(x_2;\theta) &= a(\theta)^2 b(x_1)b(x_2)\exp\left[\sum_{j=1}^d g_j(\theta)h_j(x_1)\right]\exp\left[\sum_{j=1}^d g_j(\theta)h_j(x_2)\right] \\ &= a(\theta)^2 b(x_1)b(x_2)\exp\left[\sum_{j=1}^d g_j(\theta)\big[h_j(x_1)+h_j(x_2)\big]\right] \end{aligned}$$

More generally, for a random experiment repeated independently $n$ times, with outcome $\boldsymbol{x} = (x_1,\dots,x_n)$ can likelihood be written

$$L(\theta;\boldsymbol{x}) = \prod_{i=1}^n f(x_i;\theta) = a(\theta)^n \prod_{i=1}^n b(x_i)\exp\left[\sum_{j=1}^d g_j(\theta)\sum_{i=1}^n h_j(x_i)\right].$$

Letting $H_j(x_1,\dots,x_n) = \sum_{i=1}^n h_j(x_i)$, $(j=1,\dots,d)$, and setting $\boldsymbol{H}(\boldsymbol{x}) = \big(H_1(\boldsymbol{x}),\dots,H_d(\boldsymbol{x})\big)$, we may write

$$L(\theta;\boldsymbol{x}) = f(\boldsymbol{x};\theta) = a(\theta)^n\prod_{i=1}^n b(x_i)\exp[\boldsymbol{g}(\theta)\cdot\boldsymbol{H}(\boldsymbol{x})],$$

which shows that $\boldsymbol{H}$ is a sufficient statistic for $\theta$.

The key point here is that even as the sample size gets arbitrarily large, the sufficient statistic remains $d$-dimensional.

It can be shown, see, e.g., Darmois [4], Koopman [11, Theorem 1], or Pitman [13],[1] that for the case of absolutely continuous distributions, having an exponential family is *necessary* to have a sufficient statistic of fixed dimension, subject to some smoothness conditions on the likelihood function. For this reason exponential families have played a key role in statistical theory. The key regularity properties are smoothness of $f$ (that is, $f$ has derivatives of all orders), and that the support of $f$ (that is, $\{x : f(x;\theta) > 0\}$) be independent of $\theta$. Anderson [1] has extended the original analysis to the case of discrete distributions. Perhaps the clearest exposition is Pedersen and Barndorff-Nielsen [2].

Note that the uniform distribution on $[\theta_1,\theta_2]$ is not an exponential family, but it still has a two dimensional sufficient statistic: $(\min_i x_i, \max_i x_i)$. This does not violate the statement above, since the support of the uniform does depend crucially on $(\theta_1,\theta_2)$.

## 20.6 Mean-square error of an estimator

Suppose we want to estimate some function $g$ of the parameter $\theta$ of an underlying probability model $f(x;\theta)$, using the estimator $T\colon \mathcal{X}\to\Theta$.

Define the **Mean Square Error** of the estimator $T$ of $g(\theta)$ to be the function $\mathrm{MSE}_T$ of $\theta$ given by

Not in Larsen and Marx.

$$\mathrm{MSE}_T(\theta) = \boldsymbol{E}_\theta\left[\Big(T-g(\theta)\Big)^2\right] = \int \Big(T(\boldsymbol{x})-g(\theta)\Big)^2 f(\boldsymbol{x};\theta)\,d\boldsymbol{x}.$$

If $T$ is an unbiased estimator of $\theta$, then since $\boldsymbol{E}_\theta T = \theta$, the mean square error of $T$ is just the variance of $T$.

Otherwise, define the **bias** of $T$ by

$$b_T(\theta) = \boldsymbol{E}_\theta T - g(\theta).$$

[1] This Pitman is not your textbook author, but rather his father.

Note that this depends on the unknown $\theta$.

Now we may decompose the mean square error into two terms:

$$\begin{aligned}
\mathrm{MSE}_T(\theta) &= \boldsymbol{E}_\theta\left[\Big(T - g(\theta)\Big)^2\right] \\
&= \boldsymbol{E}_\theta\left[\Big(T \underbrace{-E_\theta T + E_\theta T}_{0} - g(\theta)\Big)^2\right] \\
&= \boldsymbol{E}_\theta\left[\Big((T - E_\theta T) + b_T(\theta)\Big)^2\right] \\
&= \boldsymbol{E}_\theta\left[(T - E_\theta T)^2 + 2\underbrace{(T - E_\theta T)}_{\boldsymbol{E}_\theta = 0} b_T(\theta) + b_T(\theta)^2\right] \\
&= \boldsymbol{Var}_\theta\, T + \big(b_T(\theta)\big)^2.
\end{aligned}$$

The mean-square error of $T$ depends on the unknown parameter $\theta$.

$$\mathrm{MSE}_T(\theta) = \boldsymbol{Var}\,T + \big(b_T(\theta)\big)^2.$$

There is always a tradeoff between variance and bias. A constant estimator $\bar{T} = \bar{\theta}$ has variance zero, and $\mathrm{MSE}_{\bar{T}}(\bar{\theta}) = 0$, but it has potentially very large bias $b_{\bar{T}}(\theta)$ for $\theta \neq \bar{\theta}$, so $\mathrm{MSE}_{\bar{T}}(\theta)$ can be quite large for $\theta \neq \bar{\theta}$.

## 20.7 ⋆ A property of log-likelihood

I've already argued that taking the log of the likelihood function is a numerically reasonable thing to do. Here is another fact that illustrates the usefulness of the log-likelihood.

If $f(x;\theta)$ is a density for $x$, it gives rise to the likelihood function

$$L(\theta; x) = f(x; \theta).$$

Then since $f$ is a density we have

$$h(\theta) := \int L(\theta; x)\,dx = \int f(x;\theta)\,dx = 1.$$

Since the right-hand side does not depend on $\theta$, we must have $h'(\theta) = 0$ for every $\theta$. Often we can compute another expression for $h'$ by "differentiating under the integral." (See the on-line note for details on when this is valid.) In this case,

$$h'(\theta) := \int D_1 L(\theta; x)\,dx = \int \frac{\partial f(x;\theta)}{\partial \theta}\,dx = 0,$$

where $D_1$ denotes the partial derivative with respect to the first argument. To simplify notation, let $f'(x;\theta)$ denote $\frac{\partial f(x;\theta)}{\partial \theta}$, and let

$$\mathcal{L}(\theta; x) = \ln L(\theta; x).$$

Multiplying both the numerator and denominator of the last term by $f(x;\theta)$ gives

$$h'(\theta) = \int \frac{f'(x;\theta)}{f(x;\theta)} f(x;\theta)\,dx = \boldsymbol{E}_\theta \frac{\partial \mathcal{L}}{\partial \theta} = 0,$$

where $\boldsymbol{E}_\theta$ means that the expectation is taken with respect to the density $f(\cdot,\theta)$. To repeat:

$$\boldsymbol{E}_\theta \frac{\partial \mathcal{L}}{\partial \theta} = 0.$$

## 20.8 The Cramér–Rao Lower Bound

The following result is known the Cramér–Rao Lower Bound, even though it may have first been proven by Maurice Fréchet. It is sometimes known as the **information inequality**.

**20.8.1 The Fréchet–Cramér–Rao Lower Bound** *Assume $f$ is continuously differentiable with respect to $\theta$ and assume that the support $\{x : f(x;\theta) > 0\}$ does not depend on $\theta$. Let $T$ be an estimator of $\theta$, with differentiable bias function $b(\theta)$. Then $\boldsymbol{Var}_\theta\, T$ is bounded below, and:*

$$\boldsymbol{Var}_\theta\, T \geqslant \frac{\big[1 + b'(\theta)\big]^2}{n\,\boldsymbol{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2\right]}.$$

When $\theta$ is a parameter vector, there is a matrix interpretation of the latter.

For a proof of the Fréchet–Cramér–Rao Lower Bound see Supplement 9. For unbiased estimators this reduces to the following.

**20.8.2 Corollary (Unbiased Fréchet–Cramér–Rao Lower Bound)** *Assume $f$ is continuously differentiable with respect to $\theta$ and assume that the support $\{x : f(x;\theta) > 0\}$ does not depend on $\theta$. Let $T$ be an unbiased estimator of $\theta$. Then $\boldsymbol{Var}_\theta\, T$ is bounded below, and:*

**Larsen–Marx [12]:** § 5.5

$$\boldsymbol{Var}_\theta\, T \geqslant \frac{1}{n\,\boldsymbol{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2\right]}.$$

### 20.8.1 Special cases of the lower bound

The quantity

$$I = \boldsymbol{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2\right]$$

that appears in the denominator of the Fréchet–Cramér–Rao lower bound has an interesting interpretation. The log-likelihood function $\mathcal{L}(\theta;x) = \ln L(\theta;x)$ regarded as a function of the random variable $X$ is a random variable, and so is its derivative (with respect to $\theta$) $\mathcal{L}'(\theta;X)$. We saw in section 20.7$\star$ that $\boldsymbol{E}_\theta\,\mathcal{L}'(\theta;X) = 0$ for each $\theta$. Thus

$$I = \boldsymbol{E}_\theta\big(\mathcal{L}'(\theta;X)\big)^2 = \boldsymbol{Var}\,\mathcal{L}'(\theta;X).$$

R. A. Fisher [6, pp. 338ff], [7, pp. 709–710], [8, pp. 305–306] interpreted this as the “intrinsic accuracy” of the estimator. The quantity $I$ has since become known as the **Fisher information**. Distributions with low Fisher information or intrinsic accuracy must have high variance unbiased estimators of their parameters, but the lower bound theorem was not proven until over a decade after Fisher focused attention on $I$.

**20.8.3 Example (The lower bound and the Normal case)** The Normal density with variance $\sigma^2$ is

$$f(x;\mu) = (2\pi\sigma^2)^{-1/2} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2},$$

so we can write the log likelihood for $\mu$ as

$$\mathcal{L}(\mu;x) = -\frac{1}{2}\left[\ln(2\pi\sigma^2) + (x-\mu)^2/\sigma^2\right].$$

so

$$\mathcal{L}'(\mu;x) = \frac{1}{\sigma^2}(x-\mu).$$

You can see directly that $\boldsymbol{E}_\mu\, \mathcal{L}'(\mu;X) = \frac{1}{\sigma^2}\, \boldsymbol{E}(X-\mu) = 0$, and

$$I = \boldsymbol{E}(\mathcal{L}'(\mu;X))^2 = \boldsymbol{E}\left(\frac{X-\mu}{\sigma^2}\right)^2 = \frac{\boldsymbol{E}(X-\mu)^2}{(\sigma^2)^2} = \frac{1}{\sigma^2}.$$

So for an unbiased estimator $\hat{\mu}$ of $\mu$ the lower bound reduces to

$$\boldsymbol{Var}_\mu\, \hat{\mu} \geqslant \frac{1}{n/\sigma^2} = \frac{\sigma^2}{n} = \boldsymbol{Var}\, \bar{X}.$$

That is, any unbiased estimator has variance at least as large as the variance of the sufficient statistic $\bar{X}$. □

**20.8.4 Example (The lower bound and the Binomial case)** The probability mass function of a Bernoulli($p$) random variable $X$ is

$$f(k;p) = \binom{n}{k} p^k (1-p)^{n-k},$$

so the log likelihood is

$$\mathcal{L}(p;k) = \ln\binom{n}{k} k \ln p + (n-k)\ln(1-p).$$

Thus the first partial is

$$\mathcal{L}'(p;k) = \frac{k}{p} - \frac{n-k}{1-p} = \frac{k-np}{p(1-p)}.$$

Again it is easy to see why $\boldsymbol{E}_p\, \mathcal{L}'(p;X) = \boldsymbol{E}\, \frac{X-np}{p(1-p)} = 0$, and that

$$I = \boldsymbol{E}_p\big(\mathcal{L}'(p;X)\big)^2 = \boldsymbol{E}\left(\frac{X-np}{p(1-p)}\right)^2 = \frac{1}{p(1-p)}.$$

So for a sample of $n$ independent Bernoulli random variables, the bound on the variance of an unbiased estimator $\hat{p}$ reduces to

$$\boldsymbol{Var}_p\, \hat{p} \geqslant \frac{1}{n/p(1-p)} = \frac{p(1-p)}{n} = \boldsymbol{Var}\, \bar{X}.$$

Again, any unbiased estimator has variance at least as large as the variance of the sufficient statistic $\bar{X}$. □

In these examples, the bound is hardly mysterious. And it is not surprising that the sample mean (the maximum likelihood estimator) achieves that minimum variance.

## 20.9 MLE and Lower Bound

For the next result see, e.g., van der Waerden [16, §38, pp.162–165]. It shows that for exponential families, the Maximum Likelihood Estimator achieves the Cramér–Rao lower bound.

**20.9.1 Theorem** *Let $\theta$ be one-dimensional, and let $T(x)$ be the MLE estimator of $\theta$. Assume the likelihood function factors as*

$$L(\theta; x) = f(x; \theta) = b(x) f_T(T(x); \theta),$$

*so that $T$ is a sufficient statistic. If $f_T$ is of the* ***exponential form***

$$f_T(t; \theta) = e^{g(\theta)t + b(\theta)},$$

*and if $T$ is unbiased, then its variance achieves the Cramér–Rao lower bound, so $T$ is the minimum variance unbiased estimator of $\theta$.*

**20.9.2 Example** We've already seen that for the Normal case,

$$L(\mu; \sigma^2, x_1, \dots, x_n) \propto \sigma^{-n} e^{-\frac{1}{2\sigma^2}\sum_i x_i^2} \cdot e^{\frac{n}{\sigma^2}(\mu\bar{x} - \frac{1}{2}\mu^2)}$$

which is of the desired form for $T = \bar{x}$ as an estimator of $\mu$. In other words, $\hat{\mu}_{\text{MLE}} = \bar{x}$ is the minimum variance unbiased estimator of $\mu$. □

## 20.10 ⋆ Consistency of MLE

The classic papers on the consistency of Maximum Likelihood Estimators are by Abraham Wald [17], who proves strong consistency, and his colleague Jacob Wolfowitz [18], who simplifies Wald's arguments to show convergence in probability.

**20.10.1 Proposition (Maximum Likelihood Estimators are consistent)** *Under mild technical conditions described in Supplement 10, Maximum Likelihood Estimators are consistent and strongly consistent.*

The intuition of why this happens is straightforward. Given a sample $x_1, \dots, x_n$, for each $\theta$ the likelihood is

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

so the log-likelihood is

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

If we divide this by $n$, we get the sample average log-likelihood, which by the Law of Large Numbers, should converge to its expected value,

$$\frac{\sum_{i=1}^{n} \ln f(x_i; \theta)}{n} \xrightarrow[n\to\infty]{\text{plim}} \boldsymbol{E}_{\theta_0} \ln f(X; \theta),$$

where $\theta_0$ is the "true" value of $\theta$. Of course, we need to make enough assumptions to guarantee that this expectation exists.

Now we make an additional **identification** assumption, namely that for different $\theta$'s, we get different densities with positive probability. Or in, other words, for each $\theta \neq \theta_0$,

$$P_{\theta_0}\big(f(X; \theta) \neq f(X; \theta_0)\big) > 0. \tag{7}$$

This enables us to show that if $\theta_0$ is the parameter governing the data generating process, then $\theta_0$ uniquely maximizes the expected log-likelihood. That is the next lemma.

**20.10.2 Lemma** *For $\theta \neq \theta_0$,*

$$\boldsymbol{E}_{\theta_0} \ln f(X;\theta) < \boldsymbol{E}_{\theta_0} \ln f(X;\theta_0),$$

*assuming these expectations exist.*

*Proof*: Since $f(x;\theta)$ is a pdf for each $\theta$, we have $\int f(x;\theta)\,dx = 1$. Define $\mathbf{1}_0$ to be the indicator function of the support of $\theta_0$. That is,

$$\mathbf{1}_0(x) = \begin{cases} 1 & \text{if } f(x;\theta_0) > 0 \\ 0 & \text{if } f(x;\theta_0) = 0. \end{cases}$$

Then

$$1 = \int f(x;\theta)\,dx \geqslant \int f(x;\theta)\mathbf{1}_0(x)\,dx = \int_{\{x : f(x;\theta_0)>0\}} \frac{f(x;\theta)}{f(x;\theta_0)} f(x;\theta_0)\,dx = \boldsymbol{E}_{\theta_0} \frac{f(X;\theta)}{f(X;\theta_0)}. \tag{8}$$

By Jensen's Inequality, since ln is a strictly concave function, for any nondegenerate random variable $Y$,

$$\boldsymbol{E} \ln(Y) < \ln(\boldsymbol{E}\, Y).$$

So for $Y = e^U$, where $U$ is nondegenerate, we have

$$\boldsymbol{E}\, U < \ln(\boldsymbol{E}\, e^U).$$

Letting $U = \ln f(X;\theta) - \ln f(X;\theta_0)$. By Assumption (7), $U$ is nondegenerate, so

$$\boldsymbol{E}_{\theta_0}\Big(\ln f(X;\theta) - \ln f(X;\theta_0)\Big) < \ln\left(\boldsymbol{E}_{\theta_0} \frac{f(X;\theta)}{f(X;\theta_0)}\right) \leqslant \ln 1 = 0.$$

This proves the lemma. ∎

So the idea is the sample-average likelihood converges for each $\theta$ to its expected value by the Law of Large Numbers. By Lemma 20.10.2, the true $\theta_0$ maximizes the expected log-likelihood, which is continuous in $\theta$, so for every $\theta \neq \theta_0$ and every $\varepsilon > 0$, there is a large enough $n$ (depending on $\theta$ and $\varepsilon$) so that with probability $\geqslant 1 - \varepsilon$

$$\sum_{i=1}^n \ln f(x_i;\theta) < \sum_{i=1}^n \ln f(x_i;\theta_0).$$

This means $\theta$ cannot be the MLE. Now we need a few technical conditions to show that as $n \to \infty$ that the MLE actually does converge to something as opposed to drifting off to infinity. I hope this gives you enough guidance to understand the roles of the assumptions in [17].

## Bibliography

[1] E. B. Andersen. 1970. Sufficiency and exponential families for discrete sample spaces. *Journal of the American Statistical Association* 65(331):1248–1255. DOI: 10.2307/2284291

[2] K. Barndorff-Nielsen and O. Pedersen. 1968. Sufficient data reduction and exponential families. *Mathematica Scandinavica* 22:197–202. http://www.mscand.dk/article/view/10883

[3] H. Cramér. 1946. A contribution to the theory of statistical estimation. *Skandinavisk Aktuarietidskrift* 29:85–94.

[4] G. Darmois. 1935. Sur le lois de probabilité à estimation exhaustive. *Comptes Rendus des Séances de l'Académie des Sciences (Paris)* 260:1265–1266.

[5] M. H. DeGroot. 1970. *Optimal statistical decisions*. New York: McGraw-Hill.

[6] R. A. Fisher. 1922. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences* 222:309–368. DOI: 10.1098/rsta.1922.0009

[7] ——— . 1933. Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society* 144(5):700–725. DOI: 10.1017/S0305004100009580

[8] ——— . 1934. Two new properties of mathematical likelihood. *Proceedings of the Royal Society of London. Series A* 144(852):285–307. DOI: 10.1098/rspa.1934.0050

[9] M. Fréchet. 1943. Sur l'extension de certaines evaluations statistiques au cas de petits echantillons. *Revue de l'Institut International de Statistique / Review of the International Statistical Institute* 11(3/4):182–205. http://www.jstor.org/stable/1401114

[10] J. L. Hodges, Jr. and E. L. Lehmann. 2005. *Basic concepts of probability and statistics*, 2d. ed. Number 48 in Classics in Applied Mathematics. Philadelphia: SIAM.

[11] B. O. Koopman. 1936. On distributions admitting a sufficient statistic. *Transactions of the American Mathematical Society* 39(3):399–409. http://www.jstor.org/stable/1989758.pdf

[12] R. J. Larsen and M. L. Marx. 2012. *An introduction to mathematical statistics and its applications*, fifth ed. Boston: Prentice Hall.

[13] E. J. G. Pitman. 1936. Sufficient statistics and intrinsic accuracy. *Mathematical Proceedings of the Cambridge Philosophical Society* 32(4):567–579. DOI: 10.1017/S0305004100019307

[14] C. R. Rao. 1945. Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society* 37(3):81–91. http://bulletin.calmathsoc.org/article.php?ID=B.1945.37.14

[15] ——— . 1973. *Linear statistical inference and its applications*, 2d. ed. Wiley Series in Probability and Mathematical Statistics. New York: Wiley.

[16] B. L. van der Waerden. 1969. *Mathematical statistics*. Number 156 in Grundlehren der mathematischen Wissenschaften in Einzeldarstellungen mit besonderer Berücksichtigung der Anwendungsgebiete. New York, Berlin, and Heidelberg: Springer–Verlag. Translated by Virginia Thompson and Ellen Sherman from *Mathematische Statistik*, published by Springer-Verlag in 1965, as volume 87 in the series Grundlerhen der mathematischen Wissenschaften.

[17] A. Wald. 1949. Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics* 20(4):595–601. http://www.jstor.org/stable/2236315

[18] J. Wolfowitz. 1949. On Wald's proof of the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics* 20(4):601–602. http://www.jstor.org/stable/2236316