This material builds on Hansen's *Probability and Statistics for Economists* (PSE) and Len Goff's lecture notes when he taught ECON 8070 several years ago.

# Maximum Likelihood Estimation

PSE Chapter 10

So far, the main way that we have constructed estimators is the method of moments (based on the analogy principle) where we write the parameter of interest as a function of population moments and then replace population moments with sample averages. One attractive feature of this approach is that it requires very few assumptions — particularly, we never had to say what the distribution of $X$ is.

In this set of notes, we will consider the main alternative approach, **maximum likelihood estimation**. Both method of moments and maximum likelihood have a very long history and, in some sense, can be seen as competing paradigms. Karl Pearson introduced the method of moments in the 1890s as a way to fit distributions to data, and R.A. Fisher developed maximum likelihood between about 1912 and 1922, arguing that it typically makes more efficient use of the data than the method of moments. Pearson and Fisher are two of the most famous/influential statisticians of all time.

The idea is to start by assuming that we know the distribution of $X$ up to some unknown parameter(s), and then to choose the value of the parameter that makes the data that we actually observed "most likely." We will see that, when this assumption about the distribution is correct, maximum likelihood estimators have very good properties. Maximum likelihood is also a common way to estimate nonlinear models (for example, logit and probit models for binary outcomes), which you will see in later econometrics courses.

## Parametric Models and the Likelihood Function

A **parametric model** for $X$ is a collection of pdfs (or pmfs) $\{f(x;\tilde{\theta}) : \tilde{\theta} \in \Theta\}$, indexed by a parameter $\tilde{\theta}$ that belongs to some parameter space $\Theta$. The model is **correctly specified** if the pdf of $X$ is $f(x;\theta)$ for some $\theta \in \Theta$.

> **Side-Comment:** We will use $\theta$ to denote the true value of the parameter (as we have throughout the semester) and $\tilde{\theta}$ to denote a generic value of the parameter, such as the argument of a function. Be aware that the textbook instead uses $\theta_0$ for the true value, but the notation used here should align with our notation throughout the semester.

As an example of a parametric model, the normal model is $\{f(x;\tilde{\mu},\tilde{\sigma}^2) : \tilde{\mu} \in \mathbb{R}, \tilde{\sigma}^2 > 0\}$ where $f(x;\tilde{\mu},\tilde{\sigma}^2) = \frac{1}{\sqrt{2\pi\tilde{\sigma}^2}} \exp\left(-\frac{(x-\tilde{\mu})^2}{2\tilde{\sigma}^2}\right)$. Throughout, we will maintain that the model is correctly specified and that we have an iid sample $\{X_i\}_{i=1}^n$.

The **likelihood function** is defined as the joint pdf of the data given as a function of the parameter $\tilde{\theta}$.

$$\begin{aligned} L_n(\tilde{\theta}) &= f(X_1, \ldots, X_n; \tilde{\theta}) \\ &= \prod_{i=1}^{n} f(X_i; \tilde{\theta}) \end{aligned}$$

where the second equality holds because we have an iid sample. To be clear, you should view this as a function of the parameter, and that it gives the "likelihood'' of the data in our sample across different values of the parameter. It is usually more convenient to work in terms of the **log-likelihood function**, which is defined as

$$\ell_n(\tilde{\theta}) = \log L_n(\tilde{\theta}) = \sum_{i=1}^{n} \log f(X_i; \tilde{\theta})$$

The **maximum likelihood estimator (MLE)** is the value of $\tilde{\theta}$ that maximizes the likelihood:

$$\hat{\theta} = \underset{\tilde{\theta} \in \Theta}{\operatorname{argmax}}\ L_n(\tilde{\theta}) = \underset{\tilde{\theta} \in \Theta}{\operatorname{argmax}}\ \ell_n(\tilde{\theta})$$

For discrete $X$, $L_n(\tilde{\theta})$ is literally the probability of observing the sample that we actually observed if the true parameter were $\tilde{\theta}$. For continuous $X$, the likelihood function plays the same role through the pdf. So the MLE is the value of the parameter under which the observed data would have been most likely. The likelihood and the log-likelihood are maximized at the same value of $\tilde{\theta}$ because log is a strictly increasing function. We will almost always work with the log-likelihood because sums are much easier to work with (e.g., to differentiate) than products.

### A Simple Example: Three Possible Values of $\mu$

Let's start with a very simple example. Suppose that $X_i \sim \mathcal{N}(\mu, 1)$ (so that we know $\sigma^2 = 1$), and that we know that $\mu$ is one of only three possible values: $\Theta = \{0, 1, 2\}$. Suppose that we observe the following sample of $n = 5$ observations:

```
x <- c(1.4, 0.3, 2.1, 1.0, 0.7)
mean(x)
```

```
[1] 1.1
```

For each of the three candidate values of $\tilde{\mu}$, the likelihood is the product of the normal pdf evaluated at each of the five data points: $L_n(\tilde{\mu}) = \prod_{i=1}^{5} \phi(X_i - \tilde{\mu})$ where $\phi$ is the standard normal pdf. Figure 1 shows these five pdf values (the heights of the vertical lines) under each candidate value of $\tilde{\mu}$.

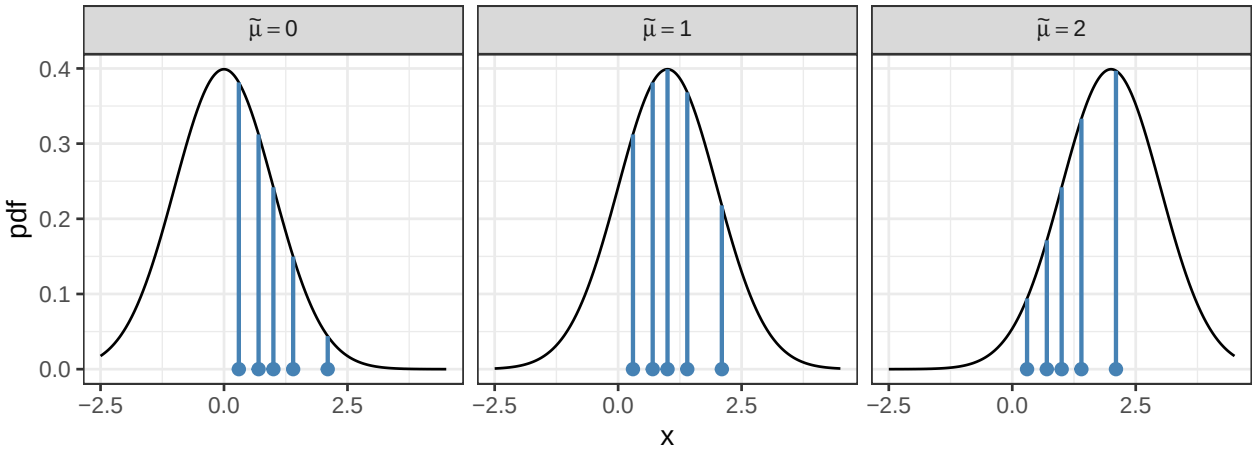



Figure 1: The data (points) along with the pdf of $\mathcal{N}(\tilde{\mu}, 1)$ for each candidate value of $\tilde{\mu}$. The likelihood is the product of the heights of the vertical lines.

When $\tilde{\mu} = 0$ or $\tilde{\mu} = 2$, several of the data points are out in the tails of the pdf, so some of the vertical lines are short, and multiplying them together gives a small number. When $\tilde{\mu} = 1$, all of the data points are near the center of the pdf. We can compute the likelihood and log-likelihood for each candidate value directly:

```
loglik <- function(mu) sum(dnorm(x, mean = mu, sd = 1, log = TRUE))
data.frame(mu = mu_vals,
  likelihood = sapply(mu_vals, function(m) exp(loglik(m))),
  loglik = sapply(mu_vals, loglik))
```

```
  mu  likelihood    loglik
1  0 0.000189771 -8.569693
2  1 0.003811652 -5.569693
3  2 0.000515851 -7.569693
```

The likelihood is largest at $\tilde{\mu} = 1$, so the MLE in this example is $\hat{\mu} = 1$.

Now suppose that, instead of only three possible values, $\mu$ could be any real number. Figure 2 plots the log-likelihood over all values of $\tilde{\mu}$, with the three candidate values from above marked as points. The log-likelihood is a smooth, concave function, and it is maximized at $\tilde{\mu} = \bar{X} = 1.1$. We will show next that this is not a coincidence.

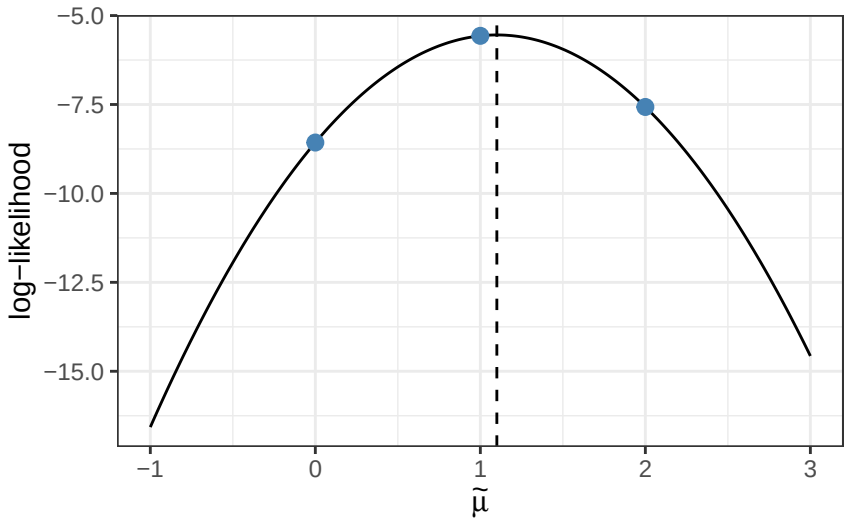



Figure 2: Log-likelihood of the $\mathcal{N}(\tilde{\mu}, 1)$ model as a function of $\tilde{\mu}$. Points are the three candidate values; the dashed line is at $\bar{X}$.

### Example: The Normal Model

Suppose that $X_i \sim \mathcal{N}(\mu, \sigma^2)$ where both $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ are unknown, so that $\theta = (\mu, \sigma^2)'$. The log-likelihood is

$$\begin{aligned}\ell_n(\tilde{\mu}, \tilde{\sigma}^2) &= \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\tilde{\sigma}^2}} \exp\left(-\frac{(X_i - \tilde{\mu})^2}{2\tilde{\sigma}^2}\right)\right) \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\tilde{\sigma}^2) - \frac{1}{2\tilde{\sigma}^2}\sum_{i=1}^{n}(X_i - \tilde{\mu})^2\end{aligned}$$

The first order conditions are

$$\begin{aligned}\frac{\partial \ell_n}{\partial \tilde{\mu}}\Big|_{(\hat{\mu}, \hat{\sigma}^2)} &= \frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}(X_i - \hat{\mu}) = 0 \\ \frac{\partial \ell_n}{\partial \tilde{\sigma}^2}\Big|_{(\hat{\mu}, \hat{\sigma}^2)} &= -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4}\sum_{i=1}^{n}(X_i - \hat{\mu})^2 = 0\end{aligned}$$

The first equation implies that $\hat{\mu} = \bar{X}$ (whatever the value of $\hat{\sigma}^2$), and plugging this into the second equation gives

$$\hat{\mu} = \bar{X} \qquad \text{and} \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Thus, the MLE of $\mu$ is the sample average — the same estimator that we obtained by the method of moments — and the MLE of $\sigma^2$ is the plug-in estimator of $\sigma^2$ from our first set of statistics notes. In the normal model, the MLE and method of moments estimators coincide. This is not true in general, but it is common for the two approaches to lead to the same (or similar) estimators in simple models.

## Consistency of the MLE

PSE Section 10.12

In the normal model, the MLE has a closed-form expression, so we can study its properties directly. In most models, though, the MLE does not have a simple closed-form expression (and is often computed numerically), and its finite-sample distribution is unknown. But, as for the method of moments estimators that we have studied, we can derive its large sample properties. We start with consistency.

It will be convenient to work with the average log-likelihood and its population counterpart, which we denote by

$$Q_n(\tilde{\theta}) = \frac{1}{n}\ell_n(\tilde{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \log f(X_i;\tilde{\theta}) \qquad \text{and} \qquad Q(\tilde{\theta}) = \mathbb{E}[\log f(X;\tilde{\theta})]$$

The MLE maximizes $Q_n(\tilde{\theta})$ (dividing by $n$ does not change the maximizer). The first step is to show that the true value $\theta$ maximizes $Q(\tilde{\theta})$. As a first step, notice that, for any $\tilde{\theta} \in \Theta$,

$$Q(\tilde{\theta}) - Q(\theta) = \mathbb{E}\left[\log \frac{f(X;\tilde{\theta})}{f(X;\theta)}\right] \leq \log \mathbb{E}\left[\frac{f(X;\tilde{\theta})}{f(X;\theta)}\right] = \log \int \frac{f(x;\tilde{\theta})}{f(x;\theta)} f(x;\theta)\, dx = \log 1 = 0$$

where the inequality is Jensen's inequality (because log is concave). This shows that $\theta$ maximizes $Q(\tilde{\theta})$ since $Q(\tilde{\theta}) - Q(\theta) = 0$ when $\tilde{\theta} = \theta$ and is otherwise less than or equal to 0 by the previous display.

> **Side-Comment:** The previous discussion is closely related to a common measure of the divergence between two pdfs. The **Kullback–Leibler divergence** between pdfs $f$ and $g$ is
>
> $$\mathrm{KL}(f,g) = -\int \log \frac{g(x)}{f(x)}\, f(x)\, dx,$$
>
> Thus, the previous display says that $Q(\tilde{\theta}) - Q(\theta) = -\mathrm{KL}(f(\cdot;\theta), f(\cdot;\tilde{\theta})) \leq 0$.
> $\mathrm{KL}(f,g)$ is called a divergence because it effectively measures the distance between two distributions (e.g., $\mathrm{KL}(f,g) \geq 0$ for all $f, g$) but is not symmetric so it is not a true distance measure.

So, the MLE maximizes $Q_n(\tilde{\theta})$, by the WLLN $Q_n(\tilde{\theta}) \xrightarrow{p} Q(\tilde{\theta})$ for each $\tilde{\theta}$, and $Q(\tilde{\theta})$ is maximized at $\theta$. It is tempting to conclude from this that $\hat{\theta} \xrightarrow{p} \theta$, but there are two gaps in this argument:

- **$\theta$ needs to be the *unique* maximizer of $Q(\tilde{\theta})$.** Otherwise, $\hat{\theta}$ could converge to some other maximizer of $Q(\tilde{\theta})$ (or bounce around between them). This is an **identification** condition. An example where this condition would fail is the model $X \sim \mathcal{N}(\mu_1 + \mu_2, 1)$ with $\theta = (\mu_1, \mu_2)'$: $(\mu_1 + c, \mu_2 - c)$ gives exactly the same distribution of $X$ for any $c$, so there is no way to learn $\mu_1$ and $\mu_2$ separately from data on $X$.

- **Uniform Convergence** In the proof of consistency below, we use the condition $\sup_{\tilde{\theta}\in\Theta} \left|Q_n(\tilde{\theta}) - Q(\tilde{\theta})\right| \xrightarrow{p} 0$. This is a stronger condition than is implied by the weak law of large numbers, which says that for any *fixed* $\tilde{\theta}$, $Q_n(\tilde{\theta})$ is close to $Q(\tilde{\theta})$ when $n$ is large. What we need instead is that $Q_n(\tilde{\theta})$ is close to $Q(\tilde{\theta})$ *uniformly* over $\Theta$.

These lead to the following assumptions.

**Assumption C (Consistency).**

C1. (Identification) For all $\tilde{\theta} \in \Theta$ with $\tilde{\theta} \neq \theta$, $Q(\tilde{\theta}) < Q(\theta)$.

C2. (Compactness and continuity) $\Theta$ is compact and $Q(\tilde{\theta})$ is continuous on $\Theta$.

C3. (Uniform convergence) $\sup_{\tilde{\theta}\in\Theta} \left|Q_n(\tilde{\theta}) - Q(\tilde{\theta})\right| \xrightarrow{p} 0$.

**Consistency of the MLE:** Under Assumption C, $\hat{\theta} \xrightarrow{p} \theta$.

*Proof:* The proof proceeds in two steps. First, we show that $Q(\hat{\theta}) \xrightarrow{p} Q(\theta)$; i.e., that the value of the population objective at $\hat{\theta}$ gets close to its maximum. Because $\theta$ maximizes $Q(\tilde{\theta})$,

$$\begin{aligned} 0 \leq Q(\theta) - Q(\hat{\theta}) &= [Q(\theta) - Q_n(\theta)] + [Q_n(\theta) - Q_n(\hat{\theta})] + [Q_n(\hat{\theta}) - Q(\hat{\theta})] \\ &\leq [Q(\theta) - Q_n(\theta)] + [Q_n(\hat{\theta}) - Q(\hat{\theta})] \\ &\leq 2 \sup_{\tilde{\theta}\in\Theta} \left|Q_n(\tilde{\theta}) - Q(\tilde{\theta})\right| \end{aligned}$$

where the second inequality holds because $\hat{\theta}$ maximizes $Q_n(\tilde{\theta})$ (so that the middle term is $\leq 0$), and the third inequality holds because each of the two remaining terms is bounded by the supremum. By C3, the last term converges in probability to 0, so $Q(\hat{\theta}) \xrightarrow{p} Q(\theta)$. Notice that uniform convergence (C3) is what allows us to handle the term $Q_n(\hat{\theta}) - Q(\hat{\theta})$, which is evaluated at the random point $\hat{\theta}$ (the WLLN alone does not tell us anything about this term).

Second, we show that this implies that $\hat{\theta}$ itself is close to $\theta$. Fix $\varepsilon > 0$, and let $A_\varepsilon = \{\tilde{\theta} \in \Theta : \|\tilde{\theta} - \theta\| \geq \varepsilon\}$ be the set of parameter values that are at least $\varepsilon$ away from $\theta$ (if $A_\varepsilon$ is empty, then $\mathrm{P}(\|\hat{\theta} - \theta\| \geq \varepsilon) = 0$ and there is nothing to show). $A_\varepsilon$ is a closed subset of the compact set $\Theta$, so it is compact, and, because $Q$ is continuous (C2), it attains its maximum over $A_\varepsilon$ at some point $\tilde{\theta}_\varepsilon \in A_\varepsilon$. Because $\tilde{\theta}_\varepsilon \neq \theta$, C1 implies that

$$\delta := Q(\theta) - Q(\tilde{\theta}_\varepsilon) = Q(\theta) - \max_{\tilde{\theta}\in A_\varepsilon} Q(\tilde{\theta}) > 0$$

That is, every parameter value at least $\varepsilon$ away from $\theta$ has a value of the population objective that is at least $\delta$ below the maximum. Therefore, if $\|\hat{\theta} - \theta\| \geq \varepsilon$, then $Q(\theta) - Q(\hat{\theta}) \geq \delta$, so that

$$\mathrm{P}(\|\hat{\theta} - \theta\| \geq \varepsilon) \leq \mathrm{P}(Q(\theta) - Q(\hat{\theta}) \geq \delta) \to 0$$

where the convergence holds by the first step. Because $\varepsilon$ was arbitrary, $\hat{\theta} \xrightarrow{p} \theta$.

**Side-Comment:** C3 is an example of a "high level" condition. It is called high level because it is "the condition we need to hold" but it is often hard to think about whether or not it is plausible in particular application. Therefore, most of the time, it is better to provide "primitive conditions" for it to hold, where these are conditions are more specific to the particular setting. The most common version of this is to provide conditions to invoke the **uniform law of large numbers (ULLN)**::

**Uniform Law of Large Numbers:** Suppose that $\{X_i\}_{i=1}^n$ are iid, that $\Theta$ is compact, that $g(x, \tilde{\theta})$ is continuous in $\tilde{\theta}$ (with probability one), and that there is a function $G(x)$ such that $|g(X, \tilde{\theta})| \leq G(X)$ for all $\tilde{\theta} \in \Theta$ and $\mathbb{E}[G(X)] < \infty$. Then, $\mathbb{E}[g(X, \tilde{\theta})]$ is continuous on $\Theta$, and

$$\sup_{\tilde{\theta} \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} g(X_i, \tilde{\theta}) - \mathbb{E}[g(X, \tilde{\theta})] \right| \xrightarrow{p} 0$$

This is PSE Theorem 18.2 (see also Lemma 2.4 in Newey and McFadden (1994), "Large sample estimation and hypothesis testing," *Handbook of Econometrics*). We will not provide a proof of this result. In the context of MLE, we can apply the ULLN (thus satisfying Assumptions C2 and C3 above) under the following conditions:

1. $\{X_i\}_{i=1}^n$ are iid,
2. $\Theta$ is compact,
3. $\log f(X; \tilde{\theta})$ is continuous in $\tilde{\theta}$ with probability one, and
4. $|\log f(X; \tilde{\theta})| \leq G(X)$ for all $\tilde{\theta} \in \Theta$, with $\mathbb{E}[G(X)] < \infty$.

Thus, it is common to see these conditions stated (together with the identification condition C1) to show consistency of MLE estimators. In particular, these are the conditions given in PSE Theorem 10.8.

## Asymptotic Normality of the MLE

PSE Sections 10.7, 10.13, and 10.15

Next, we would like to derive the asymptotic distribution of $\hat{\theta}$ so that we can construct standard errors, confidence intervals, and tests. The challenge is that, without a closed-form expression for $\hat{\theta}$, we cannot simply apply the CLT as we did for sample averages. What we do know is that the MLE

solves the sample problem of maximizing $Q_n(\tilde{\theta})$, and, thus $\hat{\theta}$ must satisfy the first order condition

$$\frac{\partial}{\partial \tilde{\theta}} Q_n(\tilde{\theta})\Big|_{\tilde{\theta}=\hat{\theta}} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial}{\partial \tilde{\theta}}\log f(X;\tilde{\theta})\right]_{\tilde{\theta}=\hat{\theta}} = 0 \tag{1}$$

The strategy will be to linearize this first order condition around $\theta$ (just like in our proof of the delta method), which will let us write $\hat{\theta}-\theta$ in terms of sample averages. Before doing so, we will introduce some more setup/terminology.

**Definition.** The **score** (for a single observation) is the $k \times 1$ vector of derivatives of the log-pdf with respect to the parameter:

$$S(\tilde{\theta}) = \frac{\partial}{\partial \tilde{\theta}}\log f(X;\tilde{\theta})$$

Notice that the score is a random vector as it depends on $X$. Its dimension is $k \times 1$, which it inherits from the dimension of $\theta$ as it is a vector of partial derivatives. The **efficient score** is the score evaluated at the true value of the parameter:

$$S = S(\theta) = \frac{\partial}{\partial \tilde{\theta}}\log f(X;\tilde{\theta})\Big|_{\tilde{\theta}=\theta}$$

For the full sample, the **likelihood score** is

$$S_n(\tilde{\theta}) = \frac{\partial}{\partial \tilde{\theta}}\ell_n(\tilde{\theta}) = \sum_{i=1}^{n} S_i(\tilde{\theta}) \text{ where } S_i(\tilde{\theta}) = \frac{\partial}{\partial \tilde{\theta}}\log f(X_i;\tilde{\theta})$$

Notice that the first order condition for the MLE in Equation 1 can be expressed as $\frac{1}{n}S_n(\hat{\theta}) = 0$, which is a system of $k$ equations.

As you will see below, we will also need to work with second derivatives of the log likelihood function, and we define related terminology next.

**Definition.** The **Hessian** (for a single observation) is the negative of the $k \times k$ matrix of second derivatives of the log-pdf:

$$H(\tilde{\theta}) = -\frac{\partial^2}{\partial \tilde{\theta}\partial \tilde{\theta}'}\log f(X;\tilde{\theta})$$

Like the score, the Hessian is random because it depends on $X$. We will write $\mathbf{H}(\tilde{\theta}) = \mathbb{E}[H(\tilde{\theta})]$ for its expectation and the **expected Hessian** is its value at the true parameter:

$$\mathbf{H}_\theta = -\mathbb{E}\left[\frac{\partial^2}{\partial \tilde{\theta}\partial \tilde{\theta}'}\log f(X;\tilde{\theta})\Big|_{\tilde{\theta}=\theta}\right]$$

For the full sample, the **likelihood Hessian** is

$$H_n(\tilde{\theta}) = -\frac{\partial^2}{\partial\tilde{\theta}\partial\tilde{\theta}'}\ell_n(\tilde{\theta}) = \sum_{i=1}^{n} H_i(\tilde{\theta})$$

where

$$H_i(\tilde{\theta}) = -\frac{\partial^2}{\partial\tilde{\theta}\partial\tilde{\theta}'}\log f(X_i;\tilde{\theta})$$

The minus sign in the definition of the Hessian is a convention (it is the one used in PSE) that makes $H_n(\hat{\theta})$ positive (semi-)definite at a maximum of the log-likelihood; larger values of $H_n(\tilde{\theta})$ correspond to a more sharply curved log-likelihood.

For the arguments below to work (and really even to define the score and Hessian above), we need some additional regularity conditions to make sure that certain derivatives are well-defined and that we can interchange differentiation and integration. These are called regularity conditions because, while you can find examples where they would not hold, they will hold in all the main MLE examples. These assumptions are stated next.

**Assumption R (Regularity).**

R1. (Interior and smoothness) $\theta$ is in the interior of $\Theta$, and there is a closed ball $\mathcal{N} \subset \Theta$ centered at $\theta$ such that $\log f(x;\tilde{\theta})$ is twice continuously differentiable in $\tilde{\theta}$ on $\mathcal{N}$ for all $x$.

R2. (Support) The support of $f(\cdot;\tilde{\theta})$, $\{x : f(x;\tilde{\theta}) > 0\}$, is the same for all $\tilde{\theta} \in \Theta$, and the first and second derivatives with respect to $\tilde{\theta}$ of integrals of the form $\int a(x)f(x;\tilde{\theta})\,dx$ (that appear below) can be calculated by differentiating inside the integral.

### Properties of the Score

Before showing asymptotic normality, we need to provide some properties of the score.

The first important property is that the efficient score has mean 0.

**Proposition:** Under Assumption R, the efficient score has mean 0: $\mathbb{E}[S] = \mathbb{E}[S(\theta)] = 0$.

To be clear, the expectation here is with respect to the true distribution of $X$, $f(x;\theta)$.

*Proof:* We will use the result from the following equation in this proof as well as several more times below.

$$\frac{\partial}{\partial\tilde{\theta}}\log f(x;\tilde{\theta}) = \frac{\partial f(x;\tilde{\theta})/\partial\tilde{\theta}}{f(x;\tilde{\theta})} \tag{2}$$

Equation 2 holds immediately by the chain rule. Next, we have that

$$\begin{aligned}\mathbb{E}[S(\theta)] &= \int \left( \frac{\partial}{\partial \tilde{\theta}} \log f(x;\tilde{\theta})\Big|_{\tilde{\theta}=\theta} \right) f(x;\theta)\, dx \\ &= \int \frac{\partial f(x;\tilde{\theta})/\partial\tilde{\theta}\big|_{\tilde{\theta}=\theta}}{f(x;\theta)} f(x;\theta)\, dx = \int \frac{\partial f(x;\tilde{\theta})}{\partial \tilde{\theta}}\Big|_{\tilde{\theta}=\theta} dx = \frac{\partial}{\partial \tilde{\theta}} \underbrace{\int f(x;\tilde{\theta})\, dx}_{=1}\Big|_{\tilde{\theta}=\theta} = 0\end{aligned}$$

where the first equality holds by the definition of $S(\theta)$, the second equality uses Equation 2, the third equality cancels terms, the fourth equality interchanges differentiation and integration, and the last equality holds because a pdf integrates to 1 for every value of $\tilde{\theta}$ (so its derivative with respect to each element of $\tilde{\theta}$ is 0).

This result should not come as a big surprise. In particular, $\mathbb{E}[S] = 0$ is the population first order condition for $\theta$ maximizing $Q(\theta)$ (the expected log-likelihood).

Next, let's consider the variance of the score (called the Fisher information).

**Definition.** The **Fisher information** is the $k \times k$ variance matrix of the efficient score:

$$\mathbf{J}_\theta = \mathbb{E}[SS'] = \text{var}(S)$$

where the second equality uses that the efficient score has mean 0.

The Fisher information is named after R.A. Fisher, who introduced it in the 1920s as a measure of how much "information" a single observation carries about $\theta$. We will see below that, when $\mathbf{J}_\theta$ is larger, $\theta$ can be estimated more precisely.

### Asymptotic Normality

We need a few more conditions for the arguments for asymptotic normality below. For a matrix $\mathbf{A}$, define its norm $\|\mathbf{A}\|$ as the square root of the sum of its squared elements, which is a measure of its size.

**Assumption AN (Asymptotic Normality).**

AN1. (Finite information) $\mathbb{E}[\|S\|^2] < \infty$.

AN2. (Uniform convergence of the Hessian) $\sup_{\tilde{\theta}\in\mathcal{N}} \left\| \frac{1}{n} H_n(\tilde{\theta}) - \mathbf{H}(\tilde{\theta}) \right\| \xrightarrow{p} 0$, and $\mathbf{H}(\tilde{\theta})$ is continuous in a neighborhood $\mathcal{N}$ of $\theta$.

AN3. (Invertibility) $\mathbf{H}_\theta$ is invertible.

AN1 is a finite variance condition, used for applying the CLT. AN2 is a uniform convergence condition, similar to C3 above. AN3 is a specific version of an identification condition that requires the log-likelihood to be curved at $\theta$.

**Proposition:** Under Assumptions C, R, and AN,

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}\left(0, \mathbf{H}_\theta^{-1}\mathbf{J}_\theta\mathbf{H}_\theta^{-1}\right)$$

*Proof:* From a mean value theorem argument, we have that

$$\begin{aligned} 0 = \frac{1}{\sqrt{n}}S_n(\hat{\theta}) &= \frac{1}{\sqrt{n}}S_n(\theta) + \frac{1}{n}H_n(\bar{\theta})\sqrt{n}(\hat{\theta} - \theta) \\ &= \frac{1}{\sqrt{n}}S_n(\theta) + \mathbf{H}(\bar{\theta})(\bar{\theta})\sqrt{n}(\hat{\theta} - \theta) + o_p(1) \\ &= \frac{1}{\sqrt{n}}S_n(\theta) + \mathbf{H}_\theta\sqrt{n}(\hat{\theta} - \theta) + o_p(1) \end{aligned}$$

where the second equality holds by AN2, and the third equality holds because $\bar{\theta}$ is between $\hat{\theta}$ and $\theta$ and because $\hat{\theta}$ is consistent for $\theta$. After rearranging, this implies that

$$\sqrt{n}(\hat{\theta} - \theta) = \mathbf{H}_\theta^{-1}\frac{1}{\sqrt{n}}S_n(\theta) + o_p(1) = \mathbf{H}_\theta^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} S_i(\theta) + o_p(1)$$

Since $S_i(\theta)$ is iid and has mean 0, by the CLT, we have that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} S_i(\theta) \xrightarrow{d} \mathcal{N}(0, \mathbf{J}_\theta)$$

and this further implies that

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \mathbf{H}_\theta^{-1}\mathbf{J}_\theta\mathbf{H}_\theta^{-1})$$

**Side-Comment:** The asymptotic variance in this proposition is often called a **sandwich** form where the variance of the score, $\mathbf{J}_\theta$, is the "meat," and the inverse of the expected Hessian, $\mathbf{H}_\theta^{-1}$, is the "bread."

Next, we will see a very famous result in statistics that implies that the asymptotic variance of maximum likelihood estimators can be simplified. Before stating this result, it is useful to have notation for expectations under distributions in the model other than the true one. For a function $a(X)$, let

$$\mathbb{E}_{\tilde{\theta}}[a(X)] := \int a(x)f(x;\tilde{\theta})\,dx$$

denote the mean that $a(X)$ would have if $X$ were drawn from $f(\cdot;\tilde{\theta})$. Plain $\mathbb{E}[\cdot] = \mathbb{E}_\theta[\cdot]$ continues to denote the expectation under the true distribution.

**Information Matrix Equality:** Under Assumption R, $\mathbf{J}_\theta = \mathbf{H}_\theta$.

*Proof:* Define the function

$$r(\tilde{\theta}) := \mathbb{E}_{\tilde{\theta}}[S(\tilde{\theta})] = \int \left(\frac{\partial}{\partial \tilde{\theta}} \log f(x;\tilde{\theta})\right) f(x;\tilde{\theta})\, dx$$

(a $k \times 1$ vector). You can use essentially the same argument as for the proof of $\mathbb{E}[S] = 0$ above to show that $r(\tilde{\theta}) = 0$ for every value of $\tilde{\theta}$. Because of this, it follows that

$$\mathbf{0} = \frac{\partial}{\partial \tilde{\theta}'} r(\tilde{\theta})$$

i.e., the derivative of $r$ is also equal to 0. And, to be clear about notation, both sides of the previous equation are $k \times k$ matrices. Therefore,

$$\begin{aligned}
\mathbf{0} = \frac{\partial r(\tilde{\theta})}{\partial \tilde{\theta}'} &= \frac{\partial}{\partial \tilde{\theta}'} \int \left(\frac{\partial}{\partial \tilde{\theta}} \log f(x;\tilde{\theta})\right) f(x;\tilde{\theta})\, dx \\
&= \int \frac{\partial}{\partial \tilde{\theta}'} \left[\left(\frac{\partial}{\partial \tilde{\theta}} \log f(x;\tilde{\theta})\right) f(x;\tilde{\theta})\right] dx \\
&= \int \left(\frac{\partial^2}{\partial \tilde{\theta} \partial \tilde{\theta}'} \log f(x;\tilde{\theta})\right) f(x;\tilde{\theta})\, dx + \int \left(\frac{\partial}{\partial \tilde{\theta}} \log f(x;\tilde{\theta})\right) \frac{\partial f(x;\tilde{\theta})}{\partial \tilde{\theta}'}\, dx \\
&= \int \left(\frac{\partial^2}{\partial \tilde{\theta} \partial \tilde{\theta}'} \log f(x;\tilde{\theta})\right) f(x;\tilde{\theta})\, dx \\
&\qquad + \int \left(\frac{\partial}{\partial \tilde{\theta}} \log f(x;\tilde{\theta})\right) \left(\frac{\partial}{\partial \tilde{\theta}} \log f(x;\tilde{\theta})\right)' f(x;\tilde{\theta})\, dx
\end{aligned}$$

where the first line holds by the definition of $r(\tilde{\theta})$, the second line interchanges differentiation and integration, the third line uses the product rule for derivatives, and the last equality uses

$$\frac{\partial f(x;\tilde{\theta})}{\partial \tilde{\theta}'} = \left(\frac{\partial}{\partial \tilde{\theta}} \log f(x;\tilde{\theta})\right)' f(x;\tilde{\theta})$$

which holds by rearranging Equation 2 above. Evaluating the previous expression at $\tilde{\theta} = \theta$, notice that the first term is $-\mathbf{H}_\theta$ and the second term is $\mathbb{E}[SS'] = \mathbf{J}_\theta$. Thus, $\mathbf{0} = -\mathbf{H}_\theta + \mathbf{J}_\theta \implies \mathbf{J}_\theta = \mathbf{H}_\theta$.

The information matrix equality immediately gives the following simplification to the asymptotic normality of maximum likelihood estimators.

**Asymptotic Normality of the MLE:** Under Assumptions C, R, and AN,

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}\left(0, \mathbf{J}_\theta^{-1}\right)$$

Besides this simplification, the information matrix equality gives a nice interpretation of the Fisher information as the (expected) curvature of the log-likelihood at the true value of the parameter. If the log-likelihood is sharply curved, the data strongly distinguish the true value of $\theta$ from nearby values, so the data are very "informative" about $\theta$. If the log-likelihood is flat, many values of $\tilde{\theta}$ fit the data about equally well.

**Example:** In the normal model with both parameters unknown, $\theta = (\mu, \sigma^2)'$ and $\log f(X; \tilde{\mu}, \tilde{\sigma}^2) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\tilde{\sigma}^2) - \frac{(X-\tilde{\mu})^2}{2\tilde{\sigma}^2}$. Taking derivatives with respect to $\tilde{\mu}$ and $\tilde{\sigma}^2$,

$$S(\tilde{\theta}) = \begin{pmatrix} \dfrac{X - \tilde{\mu}}{\tilde{\sigma}^2} \\ -\dfrac{1}{2\tilde{\sigma}^2} + \dfrac{(X-\tilde{\mu})^2}{2\tilde{\sigma}^4} \end{pmatrix}$$

$$H(\tilde{\theta}) = \begin{pmatrix} \dfrac{1}{\tilde{\sigma}^2} & \dfrac{X - \tilde{\mu}}{\tilde{\sigma}^4} \\ \dfrac{X - \tilde{\mu}}{\tilde{\sigma}^4} & -\dfrac{1}{2\tilde{\sigma}^4} + \dfrac{(X-\tilde{\mu})^2}{\tilde{\sigma}^6} \end{pmatrix}$$

Evaluating $H(\tilde{\theta})$ at the true values and using $\mathbb{E}[X - \mu] = 0$ and $\mathbb{E}[(X-\mu)^2] = \sigma^2$, it follows that

$$\mathbf{H}_\theta = \begin{pmatrix} \dfrac{1}{\sigma^2} & 0 \\ 0 & \dfrac{1}{2\sigma^4} \end{pmatrix}$$

and that $\mathbf{J}_\theta = \mathbf{H}_\theta$ from the information matrix equality. It is possible to calculate $\mathbf{J}_\theta$ directly by calculating $\mathbb{E}[SS']$. It will give you the same answer, but it is more tedious, and you will need $\mathbb{E}[(X-\mu)^3] = 0$ and $\mathbb{E}[(X-\mu)^4] = 3\sigma^4$ for a normal random variable.
Because $\mathbf{J}_\theta^{-1}$ is diagonal with diagonal elements $\sigma^2$ and $2\sigma^4$, the theorem says that $\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$ (as before) and, additionally, that $\sqrt{n}(\hat{\sigma}^2 - \sigma^2) \xrightarrow{d} \mathcal{N}(0, 2\sigma^4)$.

### Standard errors

The asymptotic variance can be estimated by $\hat{\mathbf{J}}_\theta^{-1}$ where $\hat{\mathbf{J}}_\theta = \frac{1}{n}\sum_{i=1}^n S_i(\hat{\theta})S_i(\hat{\theta})'$ or $\hat{\mathbf{J}} = \frac{1}{n}\sum_{i=1}^n H_i(\hat{\theta})$. To be clear, in finite samples, these will give you different estimates of the asymptotic variance, though the information matrix equality implies that they are asymptotically equal. Whatever estimator is used for the asymptotic variance, $\text{se}(\hat{\theta}_j) = \frac{\sqrt{\left[\hat{\mathbf{J}}^{-1}\right]_{jj}}}{\sqrt{n}}$, where $\left[\hat{\mathbf{J}}^{-1}\right]_{jj}$ denotes the $j$-th diagonal element of $\hat{\mathbf{J}}^{-1}$. The results from the previous set of notes on t-statistics, confidence intervals, and Wald tests then apply directly.

## Efficiency

This section shows that maximum likelihood estimators typically have good properties with respect to being efficient.

### The Cramér–Rao Lower Bound

PSE Sections 10.9, 10.11, and 10.14

The previous section showed that the asymptotic variance of the MLE is $\mathbf{J}_\theta^{-1}$. A natural question is whether this is "small" — could some other estimator do better? Earlier in the semester, we showed that $\bar{X}$ is the best (minimum variance) estimator of $\mathbb{E}[X]$ among *linear* unbiased estimators. The Cramér–Rao lower bound provides a stronger notion of efficiency. In a parametric model, it gives a lower bound on the variance of *any* unbiased estimator.

**Cramér–Rao Lower Bound:** Under Assumption R, and if $\mathbf{J}_\theta$ is invertible, if $\check{\theta}$ is any unbiased estimator of $\theta$, then

$$\text{var}(\check{\theta}) \geq \frac{\mathbf{J}_\theta^{-1}}{n}$$

*Proof:* Write the estimator as a function of the data, $\check{\theta} = \check{\theta}(X_1, \ldots, X_n)$, and let $f_n(x_1, \ldots, x_n; \tilde{\theta}) = \prod_{i=1}^n f(x_i; \tilde{\theta})$ denote the joint pdf of the sample. Define the function

$$m(\tilde{\theta}) := \mathbb{E}_{\tilde{\theta}}[\check{\theta}] = \int \check{\theta}(x_1, \ldots, x_n) f_n(x_1, \ldots, x_n; \tilde{\theta})\, dx_1 \cdots dx_n$$

which is the mean that $\check{\theta}$ would have if the data were drawn from $f(\cdot; \tilde{\theta})$ (using the notation from the proof of the information matrix equality). Unbiasedness means that, whatever the value of $\tilde{\theta}$, it must be the case that

$$m(\tilde{\theta}) = \tilde{\theta} \quad \text{for all } \tilde{\theta} \in \Theta,$$

or, equivalently, that

$$\tilde{\theta} = \int \breve{\theta}(x_1, \dots, x_n) f_n(x_1, \dots, x_n; \tilde{\theta})\, dx_1 \cdots dx_n$$

The first step in the proof is to show that

$$\text{cov}(\breve{\theta}, S_n(\theta)) = \mathbf{I}_k \tag{3}$$

To show this, start by taking the derivative of each side of the previous equation for $\tilde{\theta}$ with respect to $\tilde{\theta}'$, from which we have that

$$\begin{aligned}
\mathbf{I}_k &= \frac{\partial}{\partial \tilde{\theta}'} \int \breve{\theta}(x_1, \dots, x_n) f_n(x_1, \dots, x_n; \tilde{\theta})\, dx_1 \cdots dx_n \\
&= \int \breve{\theta}(x_1, \dots, x_n) \frac{\partial f_n(x_1, \dots, x_n; \tilde{\theta})}{\partial \tilde{\theta}'}\, dx_1 \cdots dx_n \\
&= \int \breve{\theta}(x_1, \dots, x_n) \frac{\partial \log f_n(x_1, \dots, x_n; \tilde{\theta})}{\partial \tilde{\theta}'}\, f_n(x_1, \dots, x_n; \tilde{\theta}) dx_1 \cdots dx_n
\end{aligned} \tag{4}$$

where the second equality holds by changing the order of the derivative and integral, and the third equality holds by the same reasoning as for Equation 2 above.

Next, notice that part of the term in the last line is given by $\log f_n(x_1, \dots, x_n; \tilde{\theta}) = \sum_{i=1}^{n} \log f(x_i; \tilde{\theta})$, which implies that

$$\begin{aligned}
\frac{\partial \log f_n(x_1, \dots, x_n; \tilde{\theta})}{\partial \tilde{\theta}'} &= \frac{\partial}{\partial \tilde{\theta}'} \sum_{i=1}^{n} \log f(x_i; \tilde{\theta}) \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial \tilde{\theta}'} \log f(x_i; \tilde{\theta}) \\
&= \sum_{i=1}^{n} S_i(\tilde{\theta})' = S_n(\tilde{\theta})
\end{aligned}$$

Plugging this expression back into Equation 4 implies $\mathbb{E}_{\tilde{\theta}}[\breve{\theta} S_n(\tilde{\theta})'] = \mathbf{I}_k$. Since the argument above holds for any $\tilde{\theta}$, it also holds at $\tilde{\theta} = \theta$, implying

$$\mathbb{E}[\breve{\theta} S_n(\theta)'] = 0$$

Because $\mathbb{E}[S_n(\theta)] = 0$ (which is easy to show based on our previous result that $\mathbb{E}[S] = 0$), this means that $\text{cov}(\breve{\theta}, S_n(\theta)) = \underbrace{\mathbb{E}[\breve{\theta} S_n(\theta)']}_{=\mathbf{I}_k} - \mathbb{E}[\breve{\theta}] \underbrace{\mathbb{E}[S_n(\theta)]}_{=0} = \mathbf{I}_k$. This establishes the claim in Equation 3.

For the next step, consider the variance of $\breve{\theta} - (n\mathbf{J}_\theta)^{-1} S_n(\theta)$. For the argument below, we will use the matrix version of $\text{var}(W - aZ) = \text{var}(W) + \text{var}(aZ) - 2\text{cov}(W, aZ) = \text{var}(W) + a^2\text{var}(Z) - 2a\text{cov}(W, Z)$, noting that the random terms in this covariance are $\breve{\theta}$ and $S_n(\theta)$, while $\mathbf{J}_\theta$ is nonrandom

since it is an expectation.

$$\begin{aligned}\operatorname{var}\Big(\check{\theta} - (n\mathbf{J}_\theta)^{-1} S_n(\theta)\Big) &= \operatorname{var}(\check{\theta}) - \operatorname{cov}(\check{\theta}, S_n(\theta))(n\mathbf{J}_\theta)^{-1} - (n\mathbf{J}_\theta)^{-1}\operatorname{cov}(S_n(\theta), \check{\theta}) \\ &\qquad + (n\mathbf{J}_\theta)^{-1}\operatorname{var}(S_n(\theta))(n\mathbf{J}_\theta)^{-1} \\ &= \operatorname{var}(\check{\theta}) - 2(n\mathbf{J}_\theta)^{-1} + (n\mathbf{J}_\theta)^{-1} \\ &= \operatorname{var}(\check{\theta}) - (n\mathbf{J}_\theta)^{-1}\end{aligned}$$

where the second equality uses $\operatorname{cov}(\check{\theta}, S_n(\theta)) = \mathbf{I}_k$ and $\operatorname{var}(S_n(\theta)) = n\mathbf{J}_\theta$.

$\operatorname{var}\Big(\check{\theta} - (n\mathbf{J}_\theta)^{-1} S_n(\theta)\Big)$ is a variance matrix, so $\operatorname{var}\Big(\check{\theta} - (n\mathbf{J}_\theta)^{-1} S_n(\theta)\Big) \geq \mathbf{0}$. This implies $\operatorname{var}(\check{\theta}) - (n\mathbf{J}_\theta)^{-1}$, which is the result.

The main implication of the Cramér–Rao bound is that maximum likelihood estimators are typically efficient. An unbiased estimator whose variance equals the Cramér–Rao lower bound is called **efficient**. No other unbiased estimator can have a smaller variance.

The Cramér–Rao bound can be extended to functions of parameters. In particular, under the same conditions as above, let $g : \mathbb{R}^k \to \mathbb{R}$ be differentiable with $k \times 1$ gradient $\nabla g(\theta)$. If $\check{g}$ is an unbiased estimator of $g(\theta)$, then

$$\operatorname{var}(\check{g}) \geq \nabla g(\theta)' (n\mathbf{J}_\theta)^{-1} \nabla g(\theta)$$

I won't provide the proof of this result, but it follows the same kind of argument we just used for the Cramér–Rao bound itself.

**Example:** In the normal model with both parameters unknown, we showed above that $\mathbf{J}_\theta$ is diagonal, so that

$$(n\mathbf{J}_\theta)^{-1} = \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{pmatrix}$$

For $\mu$, we can apply the Cramér–Rao lower bound for functions of parameters with $g(\theta) = \mu$, so that $\nabla g(\theta) = (1, 0)'$ and the bound is

$$\nabla g(\theta)'(n\mathbf{J}_\theta)^{-1}\nabla g(\theta) = \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \frac{\sigma^2}{n}$$

We also know that $\bar{X}$ is unbiased for $\mu$ and that $\text{var}(\bar{X}) = \sigma^2/n$. Therefore, $\bar{X}$ attains the lower bound; that is, *if $X$ is normally distributed*, then $\bar{X}$ has the smallest variance among all unbiased estimators of $\mu$, not just among linear ones.
For $\sigma^2$, taking $g(\theta) = \sigma^2$ so that $\nabla g(\theta) = (0, 1)'$, the same calculation implies that any unbiased estimator has variance at least $2\sigma^4/n$. Consider the unbiased estimator $s^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$. From the previous set of notes, $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$, which has variance $2(n-1)$, so that

$$\text{var}(s^2) = \frac{\sigma^4}{(n-1)^2} \cdot 2(n-1) = \frac{2\sigma^4}{n-1} > \frac{2\sigma^4}{n}$$

Thus, $s^2$ does *not* attain the Cramér–Rao lower bound. It turns out that, in this case, *no* unbiased estimator attains the bound. The Cramér–Rao lower bound is a lower bound, but it need not be achievable.

It is worth emphasizing the trade-off between the Cramér–Rao result and the BLUE result. The BLUE result only compared $\bar{X}$ to linear estimators, but it did not require knowing the distribution of $X$. The Cramér–Rao result compares $\bar{X}$ to all unbiased estimators, but it relies on knowing that $X$ is normally distributed. There is a sort of trade-off between how strong each efficiency result is and how strong the assumptions behind it are.

### Asymptotic Efficiency

The Cramér–Rao lower bound only applies to unbiased estimators, and many estimators that we care about (e.g., MLEs, or nonlinear functions of sample averages) are biased in finite samples. For these, it is more natural to compare asymptotic variances. A natural conjecture is that, if $\sqrt{n}(\check{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \mathbf{V})$, then $\mathbf{V} \geq \mathbf{J}_\theta^{-1}$. This conjecture turns out to be false without an additional

restriction on the estimator. We need to rule out certain `irregular'' estimators (you can look up Hodges' estimator for an example of an irregular estimator). However, within the class of`regular'' estimators, the claim is true and implies that maximum likelihood estimators are asymptotically efficient within this class.