Likelihood and Maximum Likelihood Estimation[*†]

1 Introduction

Previously we have discussed various properties of estimator—unbiasedness, consistency, etc—but with very little mention of where such an estimator comes from. In this part, we shall investigate one particularly important process by which an estimator can be constructed, namely, *maximum likelihood*. This is a method which, by and large, can be applied in any problem, provided that one knows and can write down the joint PMF/PDF of the data. These ideas will surely appear in any upper-level statistics course.

Let's first set some notation and terminology. Observable data $X_1, \ldots, X_n$ has a specified model, say, a collection of distribution functions $\{F_\theta : \theta \in \Theta\}$ indexed by the parameter space Θ. Data is observed, but we don't know which of the models F_θ it came from. Here, we shall assume that the model is correct, i.e., that there is a θ value such that $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F_\theta$.[1] The goal, then, is to identify the "best" model—the one that explain the data the best. This amounts to identifying the true but unknown θ value. Hence, our goal is to estimate the unknown θ.

In the sections that follow, I shall describe this so-called likelihood function and how it is used to construct point estimators. The rest of the notes will develop general properties of these estimators; these are important classical results in statistical theory. In these notes, focus is primarily on the single parameter case; Section 7 extends the ideas to the multi-parameter case.

2 Likelihood

Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F_\theta$, where θ is unknown. For the time being, we assume that θ resides in a subset Θ of $\mathbb{R}$. We further suppose that, for each θ, $F_\theta(x)$ admits a

---

[*]These notes are meant to supplement in-class lectures. The author makes no guarantees that these notes are free of typos or other, more serious errors.

[†]HMC refers to Hogg, McKean, and Craig, *Introduction to Mathematical Statistics*, 7th ed., 2012.

[1]This is a *huge* assumption. It can be relaxed, but then the details get much more complicated—there's some notion of geometry on the collection of probability distributions, and we can think about projections onto the model. We won't bother with this here.

PMF/PDF $f_\theta(x)$. By the assumed independence, the joint distribution of $(X_1, \ldots, X_n)$ is characterized by

$$f_\theta(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_\theta(x_i),$$

i.e., "independence means multiply." We understand the above expression to be a function of $(x_1, \ldots, x_n)$ for fixed θ. That is, we will fix $(x_1, \ldots, x_n)$ at the observed $(X_1, \ldots, X_n)$, and imagine the above expression as a function of θ only.

Definition 1. If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f_\theta$, then the *likelihood function* is

$$L(\theta) = f_\theta(X_1, \ldots, X_n) = \prod_{i=1}^{n} f_\theta(X_i), \tag{1}$$

treated as a function of θ. In what follows, we may occasionally add subscripts, i.e., $L_X(\theta)$ or $L_n(\theta)$, to indicate the dependence of the likelihood on data $X = (X_1, \ldots, X_n)$ or on sample size n. Also write

$$\ell(\theta) = \log L(\theta), \tag{2}$$

for the log-likelihood; the same subscript rules apply to $\ell(\theta)$.

So clearly $L(\theta)$ and $\ell(\theta)$ depend on data $X = (X_1, \ldots, X_n)$, but they're treated as functions of θ only. How can we interpret this function? The first thing to mention is a warning—*the likelihood function is NOT a PMF/PDF for θ!* So it doesn't make sense to integrate over θ values like you would a PDF. We're mostly interested in the shape of the likelihood curve or, equivalently, the relative comparisons of the $L(\theta)$ for different θ's. This is made more precise below:

> If $L(\theta_1) > L(\theta_2)$ (equivalently, if $\ell(\theta_1) > \ell(\theta_2)$), then θ_1 is more likely to have been responsible for producing the observed $X_1, \ldots, X_n$. In other words, F_{θ_1} is a better model than F_{θ_2} in terms of how well it fits the observed data.

So, we can understand likelihood (and log-likelihood) of providing a sort of *ranking* of the θ values in terms of how well they match with the observations.

Exercise 1. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Ber}(\theta)$, with $\theta \in (0, 1)$. Write down an expression for the likelihood $L(\theta)$ and log-likelihood $\ell(\theta)$. On what function of $(X_1, \ldots, X_n)$ does $\ell(\theta)$ depend. Suppose that $n = 7$ and T equals 3, where T is that function of $(X_1, \ldots, X_n)$ previously identified; sketch a graph of $\ell(\theta)$.

Exercise 2. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\theta, 1)$. Find an expression for the log-likelihood $\ell(\theta)$.

3 Maximum likelihood estimators (MLEs)

In light of our interpretation of likelihood as providing a ranking of the possible θ values in terms of how well the corresponding models fit the data, it makes sense to estimate the unknown θ by the "highest ranked" value. Since larger likelihood means higher rank, the idea is to estimate θ by the maximizer of the likelihood function, if possible.

Definition 2. Given $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f_\theta$, let $L(\theta)$ and $\ell(\theta)$ be the likelihood and log-likelihood functions, respectively. Then the maximum likelihood estimator (MLE) of θ is defined as

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta) = \arg\max_{\theta \in \Theta} \ell(\theta), \tag{3}$$

where "arg" says to return the argument at which the maximum is attained. Note that $\hat{\theta}$ implicitly depends on $(X_1, \ldots, X_n)$ because the (log-)likelihood does.

Thus, we have defined a process by which an estimator of the unknown parameter can be constructed. I call this a "process" because it can be done in the same way for (essentially) any problem: write down the likelihood function and then maximize it. In addition to the simplicity of the process, the estimator also has the nice interpretation as being the "highest ranked" of all possible θ values, given the observed data. There are also some deeper motivations for such considerations (e.g., the *Likelihood Principle*) which we won't discuss here.

I should mention that while I've called the construction of the MLE "simple," I mean that only at a fundamental level. Actually doing the maximization step can be tricky, and sometimes requires sophisticated numerical methods (see supplement). In the nicest of cases, the estimation problem reduces to solving the *likelihood equation*,

$$(\partial/\partial\theta)\ell(\theta) = 0.$$

This, of course, only makes sense if $\ell(\theta)$ is differentiable, as in the next two examples.

Exercise 3. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Ber}(\theta)$, for $\theta \in (0, 1)$. Find the MLE of θ.

Exercise 4. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\theta, 1)$, for $\theta \in (0, 1)$. Find the MLE of θ.

It can happen that extra considerations can make an ordinarily nice problem not so nice. These extra considerations are typically in the form of constraints on the parameter space Θ. The next example gives a couple illustrations.

Exercise 5. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Pois}(\theta)$, where $\theta > 0$.

(a) Find the MLE of θ.

(b) Suppose that we know $\theta \geq b$, where b is a known positive number. Using this additional information, find the MLE of θ.

(c) Suppose now that θ is known to be an integer. Find the MLE of θ.

It may also happen the the (log-)likelihood is not differentiable at one or more points. In such cases, the likelihood equation itself doesn't make sense. This doesn't mean the problem can't be solved; it just means that we need to be careful. Here's an example.

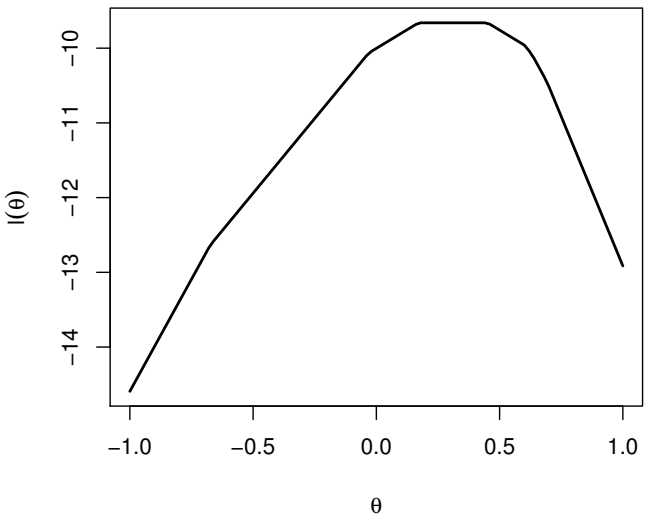


Figure 1: Graph of the Laplace log-likelihood function for a sample of size $n = 10$.

Exercise 6. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Unif}(0, \theta)$ find the MLE of θ.

I should also mention that, even if the likelihood equation is valid, it may be that the necessary work to solve it cannot be done by hand. In such cases, numerical methods are needed. Some examples are given in the supplementary notes.

Finally, in some cases, the MLE is not unique (more than one solution to the likelihood equation) and in others no MLE exists (the likelihood function is unbounded). Example 1 demonstrates the former. The simplest example of the latter is in cases where the likelihood is continuous and there is an open set constraint on θ. An important practical example is in mixture models.

Example 1. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f_\theta(x) = e^{-|x-\theta|}/2$; this distribution is often called the shifted Laplace or double-exponential distribution. For illustration, I consider a sample of size $n = 10$ from the Laplace distribution with $\theta = 0$. In Figure 1 we see that the log-likelihood flattens out, so there is an entire interval where the likelihood equation is satisfied; therefore, there the MLE is not unique. (You should try to write R code to recreate this example.)

4 Basic properties

4.1 Invariance

In the context of unbiasedness, recall the claim that, if $\hat{\theta}$ is an unbiased estimator of θ, then $\hat{\eta} = g(\hat{\theta})$ is not necessarily and unbiased estimator of $\eta = g(\theta)$; in fact, unbiasedness holds if and only if g is a linear function. That is, unbiasedness is not invariant with respect to transformations. However, MLEs are invariant in this sense—if $\hat{\theta}$ is the MLE of θ, then $\hat{\eta} = g(\hat{\theta})$ is the MLE of $\eta = g(\theta)$.

Theorem 1 (HMC, Theorem 6.1.2)**.** *Suppose $\hat{\theta}$ is the MLE of θ. Then, for specified function g, $\hat{\eta} = g(\hat{\theta})$ is the MLE of $\eta = g(\theta)$.*

Proof. The result holds for any function g, but to see the main idea, suppose that g is one-to-one. Then our familiar likelihood, written as a function of η, is simply $L(g^{-1}(\eta))$. The largest this function can be is $L(\hat{\theta})$. Therefore, to maximize, choose $\hat{\eta}$ such that $g^{-1}(\hat{\eta}) = \hat{\theta}$, i.e., take $\hat{\eta} = g(\hat{\theta})$. □

This is a very useful result, for it allows us to estimate lots of different characteristics of a distribution. Think about it: since f_θ depends on θ, any interesting quantity (expected values, probabilities, etc) will be a function of θ. Therefore, if we can find the MLE of θ, then we can easily produce the MLE for any of these quantities.

Exercise 7. If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Ber}(\theta)$, find the MLE of $\eta = \sqrt{\theta(1-\theta)}$. What quantity does η represent for the $\mathsf{Ber}(\theta)$ distribution?

Exercise 8. Let $X \sim \mathsf{Pois}(\theta)$. Find the MLE of $\eta = e^{-2\theta}$.

This invariance property is nice, but there is a somewhat undesirable consequence: *MLEs are generally NOT unbiased*. Both of the exercises above demonstrate this. For a simpler example, consider $X \sim \mathsf{N}(\theta, 1)$. The MLE of θ is $\hat{\theta} = X$ and, according to Theorem 1, the MLE of $\eta = \theta^2$ is $\hat{\eta} = \hat{\theta}^2 = X^2$. However, $\mathsf{E}_\theta(X^2) = \theta^2 + 1 \neq \theta^2$, so the MLE is NOT unbiased.

Before you get too discouraged about this, note that unbiasedness is not such an important property. In fact, we will show below that MLEs are, at least for large n, the best one can do.

4.2 Consistency

In certain examples, it can be verified directly that the MLE is consistent, e.g., this follows from the law of large numbers if the distribution is $\mathsf{N}(\theta, 1)$, $\mathsf{Pois}(\theta)$, etc. It would be better, though, if we could say something about the behavior of MLEs in general. It turns out that this is, indeed, possible—it is a consequence of the process of maximizing the likelihood function, not of the particular distributional form.

We need a bit more notation. Throughout, θ denotes a generic parameter value, while $\theta^\star$ is the "true" but unknown value; HMC use the notation θ_0 instead of $\theta^\star$.[3] The goal is to demonstrate that the MLE, denoted now by $\hat{\theta}_n$ to indicate its dependence on n, will be close to $\theta^\star$ in the following sense:

For any $\theta^\star$, the MLE $\hat{\theta}_n$ converges to $\theta^\star$ in $\mathsf{P}_{\theta^\star}$-probability as $n \to \infty$, i.e.,

$$\lim_{n\to\infty} \mathsf{P}_{\theta^\star}\{|\hat{\theta}_n - \theta^\star| > \varepsilon\} = 0, \quad \forall\, \varepsilon > 0.$$

[3]Note that there is nothing special about any particular $\theta^\star$ value—the results to be presented hold for any such value. It's simply for convenience that we distinguish this value in the notation and keep it fixed throughout the discussion.

We shall also need to put forth some general assumptions about the model, etc. These are generally referred to as *regularity conditions*, and we will list this as R0, R1, etc. Several of these regularity conditions will appear in our development below, but we add new ones to the list only when they're needed. Here's the first three:

R0. If $\theta \neq \theta'$, then f_θ and $f_{\theta'}$ are different distributions.

R1. The support of f_θ, i.e., $\text{supp}(f_\theta) := \{x : f_\theta(x) > 0\}$, is the same for all θ.

R2. $\theta^\star$ is an interior point of Θ.

R0 is a condition called "identifiability," and it simply means that it is possible to estimate θ based on only a sample from f_θ. R1 ensures that ratios $f_\theta(X)/f_{\theta'}(X)$ cannot equal ∞ with positive probability. R2 ensures that there is an open subset of Θ that contains $\theta^\star$; R2 will also help later when we need a Taylor approximation of log-likelihood.

Exercise 9. Can you think of any familiar distributions that *do not* satisfy R1?

The first result provides a taste of why $\hat{\theta}$ should be close to $\theta^\star$ when n is large. It falls short of establishing the required consistency, but it does give some nice intuition.

Proposition 1 (HMC, Theorem 6.1.1)**.** *If R0 and R1 hold, then, for any $\theta \neq \theta^\star$,*

$$\lim_{n\to\infty} \mathsf{P}_{\theta^\star}\{L_X(\theta^\star) > L_X(\theta)\} = 1.$$

Sketch of the proof. Note the equivalence of the events:

$$\begin{aligned} L_X(\theta^\star) > L_X(\theta) &\iff L_X(\theta^\star)/L_X(\theta) > 1 \\ &\iff K_n(\theta^\star, \theta) := \frac{1}{n}\sum_{i=1}^n \log \frac{f_{\theta^\star}(X_i)}{f_\theta(X_i)} > 0. \end{aligned}$$

Define the quantity[4]

$$K(\theta^\star, \theta) = \mathsf{E}_{\theta^\star}\Big\{\log \frac{f_{\theta^\star}(X)}{f_\theta(X)}\Big\},$$

From Jensen's inequality (HMC, Theorem 1.10.5), it follows that $K(\theta^\star, \theta) \geq 0$ with equality iff $\theta = \theta^\star$; in our case, $K(\theta^\star, \theta)$ is strictly positive. From the LLN:

$$K_n(\theta^\star, \theta) \to K(\theta^\star, \theta) \quad \text{in } \mathsf{P}_{\theta^\star}\text{-probability.}$$

That is, $K_n(\theta^\star, \theta)$ is near $K(\theta^\star, \theta)$, a positive number, with probability approaching 1. The claim follows since the event of interest is equivalent to $K_n(\theta^\star, \theta) > 0$. □

The intuition is that the likelihood function at the "true" $\theta^\star$ tends to be larger than any other likelihood value. So, if we estimate θ by maximizing the likelihood, that maximizer ought to be close to $\theta^\star$. To get the desired consistency, there are some technical hurdles to overcome—the key issue is that we're maximizing a random function, so some kind of uniform convergence of likelihood is required.

If we add R2 and some smoothness, we can do a little better than Proposition 1.

[4]This is known as the *Kullback–Leibler divergence*, a sort of measure of the distance between two distributions $f_{\theta^\star}$ and f_θ.

Theorem 2 (HMC, Theorem 6.1.3)**.** *In addition to R0–R2, assume that $f_\theta(x)$ is differentiable in θ for each x. Then there exists a consistent sequence of solutions of the likelihood equation.*

The proof is a bit involved so it's omitted here; but see p. 325 in HMC. This is very interesting fact but, being an existence result alone, it's not immediately clear how useful it is. For example, as we know, the likelihood equation could have many solutions for a given n. For the question "which sequence of solutions is consistent?" the theorem provides no guidance. But it does suggest that the process of solving the likelihood equation is a reasonable approach. There is one special case in which Theorem 2 gives a fully satisfactory answer.

Corollary 1 (HMC, Corollary 6.1.1)**.** *In addition to the assumptions of Theorem 2, suppose the likelihood equation admits a unique solution $\hat{\theta}_n$ for each n. Then $\hat{\theta}_n$ is consistent.*

I shall end this section with a short historical commentary. Much of the ideas (though not the proofs) were developed by Sir Ronald A. Fisher, arguably the most influential statistician in history. At the time (1920s), the field of statistics was very new and without a formal mathematical framework. Fisher's ideas on likelihood and maximum likelihood estimation set the stage for all the theoretical work that has been done since then. He is also responsible for the ideas of information and efficiency in the coming sections, as well as the notion of sufficiency to be discussed later in the course. The p-value in hypothesis testing is his idea, as well as the notion of randomization in designed experiments. Two of Fisher's other big ideas, which are less understood, are conditional inference (conditioning on ancillary statistics) and fiducial inference. Besides being one of the fathers of statistics, Fisher was also an extraordinary geneticist and mathematician. Personally, Fisher was a bit of a fiery character—there are well-documented heated arguments between Fisher, Neyman, and others about the philosophy of statistics. This "hot-headedness" was likely a result of Fisher's passion for the subject, as I have heard from people who knew him that he was a kind and thoughtful man.

5 Fisher information and the Cramer–Rao bound

To further study properties of MLEs, we introduce a concept of *information*. Before we can do this, however, we need two more regularity conditions.

R3. $f_\theta(x)$ is twice differentiable in θ for each x;

R4. $\int f_\theta(x)\,dx$ in the continuous case, or $\sum_x f_\theta(x)$ in the discrete case, is twice differentiable in θ, and the derivative can be evaluated by interchanging the order of differentiation and integration/summation.

The first condition is to guarantee that the problem is sufficiently smooth. R4 is the first condition that's really technical. It holds for most problems, but it really has nothing to do with statistics or probability. For completeness, please see the appendix with some details about interchange of derivatives and integrals/sums.

In what follows I will work with the case of continuous distributions with PDF $f_\theta(x)$. The discrete case is exactly the same, but with summation over x where integration over

x appears below. For moment, consider a single $X \sim f_\theta(x)$. Here is a simple calculus identity that will help simplify some notation, etc:

$$\frac{\partial}{\partial\theta} f_\theta(x) = \frac{\frac{\partial}{\partial\theta} f_\theta(x)}{f_\theta(x)} \cdot f_\theta(x) = \frac{\partial}{\partial\theta} \log f_\theta(x) \cdot f_\theta(x).$$

Using the fact that $1 = \int f_\theta(x)\,dx$ for all θ, if we differentiate both sides with respect to θ and apply R4 we get

$$0 = \int \frac{\partial}{\partial\theta} f_\theta(x)\,dx = \int \frac{\partial}{\partial\theta} \log f_\theta(x) \cdot f_\theta(x)\,dx = \mathsf{E}_\theta\Big\{ \frac{\partial}{\partial\theta} \log f_\theta(X) \Big\}.$$

The random variable $U_\theta(X) := \frac{\partial}{\partial\theta} \log f_\theta(X)$ is called the *score function*, and depends on both X and θ. We have shown that the score function has mean zero.

Differentiate the fundamental identity $1 = \int f_\theta(x)\,dx$ a second time and apply R4 once more we get

$$\begin{aligned} 0 &= \int \frac{\partial}{\partial\theta} \Big[\frac{\partial}{\partial\theta} \log f_\theta(x) \cdot f_\theta(x) \Big]\,dx \\ &= \cdots \\ &= \mathsf{E}_\theta\Big\{ \frac{\partial^2}{\partial\theta^2} \log f_\theta(X) \Big\} + \mathsf{E}_\theta\Big\{ \Big(\frac{\partial}{\partial\theta} \log f_\theta(X) \Big)^2 \Big\}. \end{aligned}$$

It follows that the latter two expectations are equal in magnitude—one negative, the other positive. This magnitude is called the *Fisher information*; that is,

$$I(\theta) = \mathsf{E}_\theta\Big\{ \Big(\frac{\partial}{\partial\theta} \log f_\theta(X) \Big)^2 \Big\} = -\mathsf{E}_\theta\Big\{ \frac{\partial^2}{\partial\theta^2} \log f_\theta(X) \Big\}. \tag{4}$$

This definition is understood that the Fisher information $I(\theta)$ can be evaluated with either of the two expressions on the right-hand side. You may use whichever is most convenient. It is clear that the first expression for $I(\theta)$ in (4) is positive (why?) and, therefore, defines the magnitude mentioned above. So the second expectation is negative and multiplication by -1 makes it positive.

If you recall the score function $U_\theta(X)$ defined above, then you'll notice that $I(\theta) = \mathsf{E}_\theta\{U_\theta(X)^2\}$. If you also recall that $U_\theta(X)$ has mean zero, then you'll see that the Fisher information is simply the variance $\mathsf{V}_\theta\{U_\theta(X)\}$. But despite this simple expression for $I(\theta)$ in terms of a variance of the score, it turns out that it's usually easier to evaluate $I(\theta)$ using the version with second derivatives.

Exercise 10. Find $I(\theta)$ when X is $\mathsf{Ber}(\theta)$, $\mathsf{Pois}(\theta)$, and $\mathsf{Exp}(\theta)$.

Exercise 11. Let $X \sim \mathsf{N}(\theta, \sigma^2)$ where $\sigma > 0$ is a known number. Find $I(\theta)$.

Exercise 12. Let $X \sim f_\theta(x)$, where the PDF is of the form $f_\theta(x) = g(x - \theta)$, with g an arbitrary PDF. Show that $I(\theta)$ is a constant, independent of θ. (Hint: In the integration, make a change of variable $z = x - \theta$.)

Exercise 13. Let $I(\theta)$ be the Fisher information defined above. Let $\eta = g(\theta)$ be a reparametrization, where g is a one-to-one differentiable function. If $\tilde{I}(\eta)$ is the Fisher information for the new parameter η, show that $\tilde{I}(\eta) = I(\theta) \cdot [g^{-1}(\eta)]^2$.

So far, we have considered only a single observation $X \sim f_\theta(x)$ What happens when we have an independent sample $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f_\theta(x)$? We simply replace $f_\theta(x)$ in the calculations above with the likelihood function $L(\theta) = L_X(\theta)$. Fortunately, since the model is iid, we don't have to redo all the calculations. For the score function, $U_\theta(X) = U_\theta(X_1, \ldots, X_n)$, we have

$$\begin{aligned} U_\theta(X_1, \ldots, X_n) &= \frac{\partial}{\partial\theta} \log L_X(\theta) \\ &= \frac{\partial}{\partial\theta} \sum_{i=1}^n \log f_\theta(X_i) \quad \text{(by independence)} \\ &= \sum_{i=1}^n \frac{\partial}{\partial\theta} \log f_\theta(X_i) \quad \text{(linearity of derivative)} \\ &= \sum_{i=1}^n U_\theta(X_i), \quad \text{(definition of } U_\theta(X_i)\text{)} \end{aligned}$$

the sum of the individual score functions. The Fisher information in the sample of size n is still defined as the variance of the score function. However, since we have a nice representation of the score as a sum of individual scores, we have

$$\begin{aligned} \mathsf{V}_\theta\{U_\theta(X_1, \ldots, X_n)\} &= \mathsf{V}_\theta\{U_\theta(X_1) + \cdots + U_\theta(X_n)\} \\ &= [\text{missing details}] \\ &= nI(\theta). \end{aligned}$$

Exercise 14. Fill in the missing details in the expression above.

We have, therefore, shown that the information in a sample of size n is simply n times the information in a single sample. This derivation depends critically on the iid assumption, but that's the only case in applications; but know that in dependent or non-iid data problems the Fisher information would be different.

I have so far deferred the explanation of why $I(\theta)$ is called "information." A complete understanding cannot be given yet—wait earning *sufficient statistics*—but the derivation above gives us some guidance. Intuitively, we expect that, as n increases (i.e., more data is collected), we should have more "information" about what distribution data was sample from and, therefore, we should be able to estimate θ better, in some sense. Our derivation shows that, since $I(\theta)$ is non-negative, as sample size n increases, the information about θ in that sample $nI(\theta)$ increases (linearly). So our intuition is satisfied in this case. For dependent-data problems, for example, information in the sample will still increase, but slower than linear. The Cramer–Rao lower bound result that follows should also help solidify this intuition.

In ECON 816 or other statistics or econometrics class, you could already learn point estimation and the idea of making mean-square error small. Of course, mean-square error is closely related to the variance of the estimator. The result that follows helps relate the variance of an estimator to the Fisher information. The message is that, if information is large, then better estimation should be possible.

Theorem 3 (Cramer–Rao; Theorem 6.2.1 in HMC)**.** *Let* $X_1, \ldots, X_n \overset{iid}{\sim} f_\theta(x)$*, and assume R0–R4 hold. Let* $T_n = T_n(X_1, \ldots, X_n)$ *be a statistic, with* $\mathsf{E}_\theta(T_n) = \tau(\theta)$*. Then*

$$\mathsf{V}_\theta(T_n) \geq \frac{[\tau'(\theta)]^2}{nI(\theta)}, \quad \forall\, \theta,$$

where $\tau'(\theta)$ *denotes the derivative of* $\tau(\theta)$*.*

Proof. See Appendix B. □

The following corollary helps us better understand the message of the Cramer–Rao inequality. Here we focus on the case where T_n is an unbiased estimator of θ.

Corollary 2 (HMC, Corollary 6.2.1)**.** *Let* T_n *be an unbiased estimator of* θ*. Then under the assumptions of Theorem 3,* $\mathsf{V}_\theta(T_n) \geq [nI(\theta)]^{-1}$*.*

Therefore, in this special case, the Cramer–Rao inequality can be understood as giving a lower bound on the variance of an unbiased estimator of θ. From a practical point of view, this provides us a gauge for measuring the quality of unbiased estimators. For example, if we find an unbiased estimator whose variance is exactly equal to the Cramer–Rao bound, then we know that no other unbiased estimator can do better than this one. We follow up on this idea in Section 6.

Exercise 15. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Pois}(\theta)$. Find the Cramer–Rao lower bound for unbiased estimators of θ. Find the variance of $\bar{X}$ and compare to this lower bound. We've seen before that S^2 is also an unbiased estimator of θ. What does your comparison of the Cramer–Rao lower bound and $\mathsf{V}_\theta(\bar{X})$ say about the relative performance of of $\bar{X}$ and S^2? You don't have to evaluate the variance of S^2, just explain how Corollary 2 helps with your comparison.

6 Efficiency and asymptotic normality

To follow up, more formally, on the notion of measuring performance of estimators by comparing their variance to the Cramer–Rao lower bound, we define a notion of efficiency. If $\hat{\theta}_n$ is an unbiased estimator of θ, then the *efficiency* (Pittman Effcicincy) of $\hat{\theta}_n$ is

$$\mathsf{eff}_\theta(\hat{\theta}_n) = \text{LB}/\mathsf{V}_\theta(\hat{\theta}_n), \quad \text{where} \quad \text{LB} = 1/nI(\theta).$$

An estimator is *efficient* if $\mathsf{eff}_\theta(\hat{\theta}_n) = 1$.

Exercise 16. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(0, \theta)$, where $\theta > 0$ denotes the variance.

(a) Let $\hat{\theta}_n^{(1)}$ be the sample variance. Find $\mathsf{eff}_\theta(\hat{\theta}_n^{(1)})$.

(b) Find the MLE of θ, and write this as $\hat{\theta}_n^{(2)}$. Find $\mathsf{eff}_\theta(\hat{\theta}_n^{(2)})$.

(c) Compare $\hat{\theta}_n^{(1)}$ and $\hat{\theta}_n^{(2)}$ based on their efficiencies.

We are particularly interested in the efficiency of MLEs, but there's not so many problems where the MLE has a nice expression, and even fewer of these cases can we write down a formula for its variance. So it would be nice to have some idea about the efficiency of MLEs without having to write down its variance. The next theorem, a fundamental result in statistics, gives us such a result. Indeed, a consequence of this theorem is that the MLE *asymptotically efficient* in the sense that, as $n \to \infty$, the efficiency of the MLE approaches 1. We need one more regularity condition:

R5. $f_\theta(x)$ is thrice differentiable in θ for each x, and there exists a constant $c > 0$ and a function $M(x) > 0$ such that $\mathsf{E}_\theta[M(X)] < \infty$ and, for "true value" $\theta^\star$, $|\frac{\partial^3}{\partial \theta^3} \log f_\theta(x)| \leq M(x)$ for all x and for all $\theta \in (\theta^\star - c, \theta^\star + c)$.

This assumption allows us to write a two-term Taylor approximation for $\ell(\theta)$, which is the driving part of the proof, sketched in Appendix C.

Theorem 4 (HMC, Theorem 6.2.2)**.** *Let $X_1, \ldots, X_n \overset{iid}{\sim} f_\theta(x)$, with "true value" $\theta^\star$. If R0–R5 hold, and $I(\theta^\star) \in (0, \infty)$, then for any consistent sequence of solutions $\hat{\theta}_n$ of the likelihood equation $\sqrt{n}(\hat{\theta}_n - \theta^\star) \to \mathsf{N}(0, I(\theta^\star)^{-1})$ in distribution as $n \to \infty$.*

For simplicity, let's drop the $\star$ superscript. Then we can understand the result as saying that, when n is large, the MLE $\hat{\theta}_n$ is approximately normal with mean θ and variance $[nI(\theta)]^{-1}$. So the claim about asymptotic efficiency of the MLE is clear.

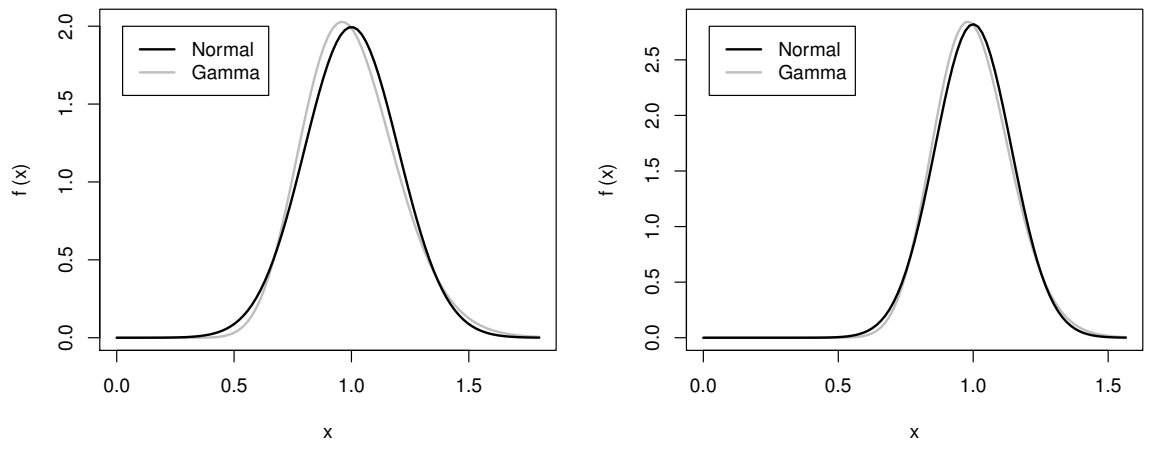
Given the importance of MLEs in applied statistics, Theorem 4 is fundamental. It says that no matter how the MLE is obtained—closed form expression, complicated numerical algorithms, etc—the sampling distribution is approximately normal when n is large. Many statistical computing packages report hypothesis tests and confidence intervals in relatively complex problems, such as logistic regression, and these are based on the sampling distribution result in Theorem 4.

Example 2. An interesting question is: how accurate is the normal approximation for finite n? Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Exp}(\theta)$. If $\theta = 1$, then Theorem 4 says the MLE $\bar{X}_n$ is approximately normal with mean 1 and variance n^{-1}. However, it can be shown that $\bar{X}_n \sim \mathsf{Gamma}(n, n^{-1})$. Figure 2 shows the exact distribution of $\bar{X}_n$ and the normal approximation for two relatively small values of n. At $n = 25$ there's some noticeable differences between the two distributions, but for $n = 50$ there's hardly any difference.

Exercise 17. Show that if $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Exp}(\theta)$, then $\bar{X}_n \sim \mathsf{Gamma}(n, n^{-1}\theta)$.

Theorem 4 is much more broad than it looks initially. As it's stated, it applies only to the MLE of θ (specifically, consistent solutions of the likelihood equation). But in light of the invariance of MLE (Theorem 1) and the Delta Theorem which can be found in ECON 817, we can develop a similar asymptotic normality result for any function of the MLE.

Exercise 18. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Exp}(\theta)$. The MLE is $\hat{\theta}_n = \bar{X}_n$. Use Theorem 4 and the Delta Theorem to find the limiting distribution of $\log \bar{X}_n$.



(a) $n = 25$ (b) $n = 50$

Figure 2: Exact and approximate sampling distributions for the MLE in Example 2.

Asymptotic normality of MLEs, in combination with the Delta Theorem, is very useful in the construction of confidence intervals. Unfortunately, we do not have sufficient time to cover this important application in detail. But some supplementary material on maximum likelihood confidence intervals is provided in a separate document.

This consideration of the asymptotic efficiency of MLEs is effectively a comparison of the *asymptotic variance* of the MLE, which according to Theorem 4, is $I(\theta)^{-1}$. This is just like the "v_θ" in the Delta Theorem statement in ECON 817. So, a way to compare two estimators is look at the ratio of their respective asymptotic variances. That is, the *asymptotic relative efficiency* of $\hat{\theta}_n^{(1)}$ and $\hat{\theta}_n^{(2)}$ is

$$\mathsf{are}_\theta(\hat{\theta}_n^{(1)}, \hat{\theta}_n^{(2)}) = \frac{\mathsf{aV}_\theta(\hat{\theta}_n^{(1)})}{\mathsf{aV}_\theta(\hat{\theta}_n^{(2)})},$$

where aV denotes the asymptotic variance. If this ratio is bigger (resp. smaller) than 1, then $\hat{\theta}_n^{(2)}$ is "better" (resp. "worse") than $\hat{\theta}_n^{(1)}$.

Example 3. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\theta, \sigma^2)$, with σ known. The MLE is $\hat{\theta}_n^{(1)} = \bar{X}_n$, and it's easy to check that the MLE is efficient. An alternative estimator is $\hat{\theta}_n^{(2)} = M_n$, the sample median. The exact variance of M_n is difficult to get, so we shall compare these two estimators based on asymptotic relative efficiency. For this, we need a sort of CLT for M_n (the 50th percentile):

> (CLT for percentiles) *Let* $X_1, \ldots, X_n \overset{iid}{\sim} f(x) = F'(x)$. *For any* $p \in (0, 1)$, *let* η_p *be the* $100p$*th percentile, i.e.,* $F(\eta_p) = p$. *Likewise, let* $\hat{\eta}_p$ *be the* $100p$*th sample percentile. If* $f(\eta_p) > 0$, *then*
>
> $$\sqrt{n}(\hat{\eta}_p - \eta_p) \to \mathsf{N}\big(0, p(1-p)/f(\eta_p)^2\big), \quad \textit{in distribution.}$$

In this case, the asymptotic variance of M_n is

$$\mathsf{aV}_\theta(\hat{\theta}_n^{(2)}) = \frac{0.5 \cdot 0.5}{\left(\sqrt{1/2\pi\sigma^2}\right)^2} = \frac{\pi\sigma^2}{2}.$$

Since $\mathsf{aV}_\theta(\hat{\theta}_n^{(1)}) = \sigma^2$, the asymptotic relative efficiency is

$$\mathsf{are}_\theta(\hat{\theta}_n^{(1)}, \hat{\theta}_n^{(2)}) = \frac{\sigma^2}{\pi\sigma^2/2} = \frac{2}{\pi} < 1.$$

This ratio is less than 1, so we conclude that $\hat{\theta}_n^{(1)}$ is "better" asymptotically.

7 Multi-parameter cases

Now suppose that $\theta \in \Theta \subseteq \mathbb{R}^d$, for integer $d \geq 1$. An important example is $\Theta = \{\theta = (\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+\}$ for the normal distribution where both mean μ and variance σ^2 are unknown. In general, let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f_\theta(x)$. Then we may define the likelihood, and log-likelihood functions just as before:

$$L(\theta) = \prod_{i=1}^{n} f_\theta(X_i) \quad \text{and} \quad \ell(\theta) = \log L(\theta).$$

Likelihood still can be understood as providing a ranking of the possible parameter values and, therefore, maximizing the likelihood function to estimate the unknown θ still makes sense. That is, the MLE $\hat{\theta}$ is still defined as

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta) = \arg\max_{\theta \in \Theta} \ell(\theta).$$

Conceptually, everything is the same as in the one-dimensional parameter case. Technically, however, things are messier, e.g., we need vectors, matrices, etc. We can immediately see how things get more technically involved, by considering the analogue of the likelihood equation: $\hat{\theta}$ is the solution to

$$\nabla \ell(\theta) = 0.$$

Here ∇ is the gradient operator, producing a vector of component wise partial derivatives,

$$\nabla \ell(\theta) = \Big(\frac{\partial \ell(\theta)}{\partial \theta_1}, \ldots, \frac{\partial \ell(\theta)}{\partial \theta_d} \Big)^\top,$$

and superscript $\top$ being the transpose operator.

Exercise 19. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\mu, \sigma^2)$, with $\theta = (\mu, \sigma^2)$ unknown. Find the MLE.

For multiple parameters, it is less likely that a closed-form solution to the likelihood equation is available. Typically, some kind of numerical methods will be needed to find the MLE. Next is a simple example of this scenario.

Exercise 20. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Gamma}(\alpha, \beta)$, with $\theta = (\alpha, \beta)$ unknown. Write down the likelihood equation and confirm that no closed-form solution is available.

For a single observation $X \sim f_\theta(x)$, the score vector is

$$U_\theta(X) = \nabla \log f_\theta(X) = \Big(\frac{\partial}{\partial \theta_1} \log f_\theta(X), \ldots, \frac{\partial}{\partial \theta_d} \log f_\theta(X) \Big)^\top.$$

In this case, the score is a $d \times 1$ (column) random vector. Recall that, for random vectors, there are notions of a mean vector and a covariance matrix. In particular, if Z is a d-dimensional random vector, then

$$\begin{aligned} \mathsf{E}(Z) &= \big(\mathsf{E}(Z_1), \ldots, \mathsf{E}(Z_d) \big)^\top \\ \mathsf{C}(Z) &= \mathsf{E}(ZZ^\top) - \mathsf{E}(Z)\mathsf{E}(Z)^\top. \end{aligned}$$

So the mean of a random vector is a $d \times 1$ vector and the covariance is a $d \times d$ matrix (provided these quantities exist). Under versions of the regularity conditions in the one-parameter case, it can be shown that

$$\mathsf{E}_\theta[U_\theta(X)] = 0 \quad \text{(a } d\text{-vector of zeros)}.$$

Just like in the one-parameter case, we define the Fisher information as the (co)variance of the score, i.e., $I(\theta) = \mathsf{C}_\theta[U_\theta(X)]$, which is a $d \times d$ matrix, rather than a number. Under regularity conditions, each component of this matrix looks like a one-dimensional information; in particular, its (j,k)th element satisfies

$$\begin{aligned} I(\theta)_{jk} &= \mathsf{E}_\theta \Big\{ \frac{\partial}{\partial \theta_j} \log f_\theta(X) \cdot \frac{\partial}{\partial \theta_k} \log f_\theta(X) \Big\} \\ &= -\mathsf{E}_\theta \Big\{ \frac{\partial^2}{\partial \theta_j \partial \theta_k} \log f_\theta(X) \Big\}. \end{aligned}$$

Typically, $I(\theta)$ is a symmetric matrix (i.e., $I(\theta) = I(\theta)^\top$); for us, this will always be true. This means you only need to evaluate $d(d+1)/2$ of the d^2 total matrix entries.

Exercise 21. For $X \sim \mathsf{N}(\mu, \sigma^2)$, with $\theta = (\mu, \sigma^2)$, find $I(\theta)$.

What about if we have an iid sample $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f_\theta(x)$ of size n? Everything goes just as before, except we're working with vectors/matrices. In particular, we replace the density $f_\theta(X)$ in the definition of the score vector with the likelihood function $L_X(\theta)$, and just as before, the Fisher information matrix for a sample of size n is just n times the information matrix $I(\theta)$ for a single observation.

For brevity, I shall summarize the d-dimensional analogues of the large-sample results derived above with care for one-dimensional problems. Here I will not explicitly state the regularity conditions, but know that they are essentially just higher-dimensional versions of R0–R5 listed above.

- Under regularity conditions, there exists a consistent sequence $\hat{\theta}_n$ (a d-vector) of solutions of the likelihood equation.
- Under regularity conditions, for any consistent sequence of solutions $\hat{\theta}_n$,

  $$\sqrt{n}(\hat{\theta}_n - \theta) \to \mathsf{N}_d(0, I(\theta)^{-1}) \quad \text{in distribution (for all } \theta),$$

  where $\mathsf{N}_d(0, I(\theta)^{-1})$ denotes a d-dimensional normal distribution with mean vector 0 and covariance matrix $I(\theta)^{-1}$, the $d \times d$ inverse of the Fisher information matrix.

- (Delta Theorem) Let $g : \mathbb{R}^d \to \mathbb{R}^k$ have continuous partial derivatives, and define the $k \times d$ matrix
$$D = \big(\partial g(\theta)_i / \partial \theta_j\big)_{i=1,\ldots,k; j=1,\ldots,d}.$$
Then, under regularity conditions,
$$\sqrt{n}[g(\hat{\theta}_n) - g(\theta)] \to \mathsf{N}_k(0, DI(\theta)^{-1}D^\top).$$
For example, take $g : \mathbb{R}^d \to \mathbb{R}$ so that $g(\theta) = \theta_j$. Then D is a $1 \times d$ matrix of all zeros except a 1 appearing in the $(1, j)$ position. With this choice,
$$\sqrt{n}(\hat{\theta}_{n,j} - \theta_j) \to \mathsf{N}(0, I(\theta)^{-1}_{jj}),$$
which is the one-dimensional counterpart, like Theorem 4.

Exercise 22. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Gamma}(\alpha, \beta)$, where $\theta = (\alpha, \beta)^\top$ is unknown. Denote the MLE by $\hat{\theta}_n = (\hat{\alpha}_n, \hat{\beta}_n)^\top$; there's no closed-form expression for the the MLE, but it can be readily evaluated numerically. State the limiting distribution of $\hat{\theta}_n$.

A Interchanging derivatives and integrals/sums

Condition R4 requires that derivatives of integrals/sums can be evaluated by differentiating the integrand/summand. The question of whether these operations can be interchanged has nothing to do with statistics, these are calculus/analysis issues. But, for completeness, I wanted to give a brief explanation of what's going on.

Let $f(x, \theta)$ be a function of two variables, assumed to be differentiable with respect to θ for each x. Here f need not be a PDF/PMF, just a function like in calculus. Let's consider the simplest case: suppose x ranges over a finite set, say, $\{1, 2, \ldots, r\}$. Then it's a trivial result from calculus that
$$\frac{d}{d\theta} \sum_{x=1}^{r} f(x, \theta) = \sum_{x=1}^{r} \frac{\partial}{\partial \theta} f(x, \theta).$$
This is referred to as the linearity property of differentiation. Similarly, suppose x ranges over a bounded interval $[a, b]$, where neither a nor b depends on θ (this last assumption can easily be relaxed). Then the famous Leibnitz formula gives
$$\frac{d}{d\theta} \int_a^b f(x, \theta)\, dx = \int_a^b \frac{\partial}{\partial \theta} f(x, \theta)\, dx.$$
So, in these two cases, differentiation and summation/integration can be interchanged with essentially no conditions. The common feature of these two situations is that summation/integration is over a finite/bounded range. Things are not so simple when "infinities" are involved.

Both the summation and integration problems over bounded and unbounded ranges can be lumped together under one umbrella in a measure-theoretic context, and the question of interchange with differentiation can be answered with the Lebesgue Dominated

Convergence Theorem. The general details are too much, so here I'll work on the two cases separately.

Start with the summation problem. That is, we want to know when

$$\frac{d}{d\theta}\sum_{x=1}^{\infty} f(x,\theta) = \sum_{x=1}^{\infty}\frac{\partial}{\partial\theta} f(x,\theta). \tag{5}$$

The three sufficient conditions are

S1. $\sum_{x=1}^{\infty} f(x,\theta)$ converges for all θ in an interval (a,b);

S2. $\frac{\partial}{\partial\theta} f(x,\theta)$ is continuous in θ for all x;

S3. $\sum_{x=1}^{\infty} \frac{\partial}{\partial\theta} f(x,\theta)$ converges uniformly on every compact subset of (a,b).

That is, if S1–S3 hold, then (5) is valid.

In the integration problem, we want to know when

$$\frac{d}{d\theta}\int_{-\infty}^{\infty} f(x,\theta)\,dx = \int_{-\infty}^{\infty}\frac{\partial}{\partial\theta} f(x,\theta)\,dx. \tag{6}$$

In this case, there is just one sufficient condition, with two parts. Suppose that there exists a function $g(x,\theta)$ and a number $\delta > 0$ such that

$$\left|\frac{f(x,\theta+\delta') - f(x,\theta)}{\delta'}\right| \le g(x,\theta) \quad \text{for all } x \text{ and all } |\delta'| \le \delta,$$

and

$$\int_{-\infty}^{\infty} g(x,\theta)\,dx < \infty.$$

Then statement (6) is valid.

B Proof of Theorem 3

For two random variables X and Y, the covariance (if it exists) is defined as $\mathsf{C}(X,Y) = \mathsf{E}(XY) - \mathsf{E}(X)\mathsf{E}(Y)$. The Cauchy–Schwartz inequality (you may have seen this in a linear algebra course) that says $|\mathsf{C}(X,Y)| \le \sqrt{\mathsf{V}(X)\mathsf{V}(Y)}$.

Here I will work with the case $n = 1$; write $X = X_1$, $T = T(X)$ for the statistic in question, and $U = U_\theta(X)$ for the score function. The first goal is to evaluate the covariance $\mathsf{C}_\theta(T,U)$. For this, recall that U has zero mean, so $\mathsf{C}_\theta(T,U) = \mathsf{E}_\theta(TU)$. Recall that $\frac{\partial}{\partial\theta}\log f_\theta(x)\cdot f_\theta(x) = \frac{\partial}{\partial\theta} f_\theta(x)$; then the expectation of TU can be written as

$$\begin{aligned}
\mathsf{E}_\theta(TU) &= \int T(x)U_\theta(x)f_\theta(x)\,dx \\
&= \int T(x)\frac{\partial}{\partial\theta}\log f_\theta(x) f_\theta(x)\,dx \\
&= \int T(x)\frac{\partial}{\partial\theta} f_\theta(x)\,dx \\
&= \frac{\partial}{\partial\theta}\int T(x) f_\theta(x)\,dx \quad \text{(by R4)} \\
&= \tau'(\theta).
\end{aligned}$$

Now we know that $\mathsf{V}_\theta(U) = I(\theta)$, so the Cauchy–Schwartz inequality above gives

$$|\tau'(\theta)| \le \sqrt{\mathsf{V}_\theta(T) I(\theta)}.$$

Squaring both sides and solving for $\mathsf{V}_\theta(T)$ gives the desired result.

C Proof of Theorem 4

The basic idea of the proof is fairly simple, although carrying out the precise details is a bit tedious. So here I'll just give a sketch to communicate the ideas.

First, do a Taylor approximation of $\ell_n'(\hat{\theta}_n)$ in a neighborhood of $\hat{\theta}_n = \theta^\star$. Since $\hat{\theta}_n$ is a solution to the likelihood equation, we know that $\ell_n'(\hat{\theta}_n) = 0$. Therefore, this Taylor approximation looks like

$$0 = \ell_n'(\theta) = \ell_n'(\theta^\star) + \ell_n''(\theta_n)(\hat{\theta}_n - \theta^\star) + \text{error},$$

where θ_n is some value between $\hat{\theta}_n$ and $\theta^\star$. Since $\hat{\theta}_n$ is consistent, it follows that θ_n is too. Ignoring the error and rearranging the terms in the Taylor approximation gives

$$\sqrt{n}(\hat{\theta}_n - \theta^\star) = -\frac{n^{1/2}\ell_n'(\theta^\star)}{\ell_n''(\theta_n)} = -\frac{n^{-1/2}\ell_n'(\theta^\star)}{n^{-1}\ell_n''(\theta_n)}.$$

Now we'll look at the numerator and denominator separately.

We can apply the usual CLT to study the numerator. Indeed, note that

$$\bar{U}_n := \frac{1}{n}\ell_n(\theta^\star) = \frac{1}{n}\sum_{i=1}^n U_{\theta^\star}(X_i)$$

is an average of iid mean-zero, variance-$I(\theta^\star)$ random variables. So the usual CLT says $n^{-1/2}\ell_n(\theta^\star) = \sqrt{n}(\bar{U}_n - 0) \to \mathsf{N}(0, I(\theta^\star))$ in distribution.

For the denominator, we'll do a bit of fudging. Recall that θ_n is close to $\theta^\star$ for large n. So we'll just replace $n^{-1}\ell_n''(\theta_n)$ in the denominator with $n^{-1}\ell_n''(\theta^\star)$. A careful argument using the regularity conditions can make this step rigorous. Now $n^{-1}\ell_n''(\theta^\star)$ is an average of iid mean-$I(\theta^\star)$ random variables, so the usual LLN says $n^{-1}\ell_n''(\theta^\star)$ converges in probability to $I(\theta^\star)$.

If we use Slutsky's theorem, we get

$$\sqrt{n}(\hat{\theta}_n - \theta^\star) = -\frac{n^{1/2}\ell_n'(\theta^\star)}{\ell_n''(\theta_n)} = -\frac{n^{-1/2}\ell_n'(\theta^\star)}{n^{-1}\ell_n''(\theta_n)} \to I(\theta^\star)^{-1} \cdot \mathsf{N}(0, I(\theta^\star)), \quad \text{in distribution.}$$

But multiplying a normal random variable by a number changes the variance by the square of that number. That is,

$$I(\theta^\star)^{-1} \cdot \mathsf{N}(0, I(\theta^\star)) \equiv \mathsf{N}(0, I(\theta^\star)^{-1}).$$

This completes the (sketch of the) proof.