7.4: Information and Efficiency

Taylor

University of Virginia

Motivation

In this section we define *Fisher Information* and two of its applications. Application 1: Find the minimum possible variance for an unbiased estimator. Application 2: Show MLE estimators are asymptotically unbiased and normally distributed.

Definitions

The **Fisher information** $I(\theta)$ in a single observation from a pmf or pdf $f(x;\theta)$ is the variance of the random variable $U = \frac{\partial \log f(X;\theta)}{\partial\theta}$

$$I(\theta) = V\left[\frac{\partial \log f(X;\theta)}{\partial\theta}\right]$$

Definitions

The **Fisher information** $I(\theta)$ in a single observation from a pmf or pdf $f(x;\theta)$ is the variance of the random variable $U = \frac{\partial \log f(X;\theta)}{\partial \theta}$

$$I(\theta) = V\left[\frac{\partial \log f(X;\theta)}{\partial \theta}\right]$$

U is called the “score.”

Definitions

First thing to notice is that the mean of $U = \frac{\partial \log f(X;\theta)}{\partial\theta}$ is 0. For this slide and the next slide we'll use this property a couple of times: $(\log f)' = \frac{f'}{f}$ for $f \geq 0$. Also we're dealing with a discrete pmf here.

$$EU = E\left[\frac{\partial \log f(X;\theta)}{\partial\theta}\right] = \sum_x \frac{\partial \log f(x;\theta)}{\partial\theta} f(x;\theta) \tag{1}$$

$$= \sum_x \frac{\partial}{\partial\theta} f(x;\theta) = \frac{\partial}{\partial\theta} \sum_x f(x;\theta) \tag{2}$$

$$= \frac{\partial}{\partial\theta} 1 = 0 \tag{3}$$

This means $V\left[\frac{\partial \log f(X;\theta)}{\partial\theta}\right] = E\left[\left(\frac{\partial \log f(X;\theta)}{\partial\theta}\right)^2\right]$

Proposition

Now we find another expression for $I(\theta)$. Recall from the previous slide that $0 = \sum_x \frac{\partial \log f(x;\theta)}{\partial \theta} f(x;\theta)$. Taking derivatives again on both sides we get:

$$
\begin{aligned}
0 &= \frac{\partial}{\partial \theta} \sum_x \frac{\partial \log f(x;\theta)}{\partial \theta} f(x;\theta) \\
&= \sum_x \left\{ \left(\frac{\partial^2 \log f(x;\theta)}{\partial \theta^2} \right) f(x;\theta) + \frac{\partial \log f(x;\theta)}{\partial \theta} \left(\frac{\partial}{\partial \theta} f(x;\theta) \right) \right\} \\
&= \sum_x \left\{ \left(\frac{\partial^2 \log f(x;\theta)}{\partial \theta^2} \right) f(x;\theta) + \frac{\partial \log f(x;\theta)}{\partial \theta} \left(\frac{\partial \log f(x;\theta)}{\partial \theta} f(x;\theta) \right) \right\}
\end{aligned}
$$

Definitions

Let's rewrite that last line:

$$0 = \sum_x \left\{ \left(\frac{\partial^2 \log f(x;\theta)}{\partial \theta^2} \right) f(x;\theta) + \frac{\partial \log f(x;\theta)}{\partial \theta} \left(\frac{\partial \log f(x;\theta)}{\partial \theta} f(x;\theta) \right) \right\}$$

after we break up the sum,

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

So whenever we're allowed to do the above steps, we have

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

Example

Let X be a Bernoulli random variable. So $f(x;p) = p^x(1-p)^{1-x}$, $x = 0, 1$. Taking the log, we get $\log f(x;p) = x \log p + (1-x)\log(1-p)$.

Taking the derivative with respect to the parameter:

$$\frac{\partial \log f(X;p)}{\partial p} = \frac{X}{p} - \frac{1-X}{1-p} = \frac{X-p}{p(1-p)}$$

If we use the first way, then

$$I(\theta) = V\left[\frac{X-p}{p(1-p)}\right] = V\left[\frac{X}{p(1-p)}\right] = \frac{1}{p^2(1-p)^2} V[X] = \frac{1}{p(1-p)}$$

Example continued

Usually the second way is easier, although it doesn't really help that much in this problem. This is generally true, though, because taking expectations is easier than taking variances. So for practice let's do it again but using the second way.

Example continued

Recall that our first derivative was

$$\frac{\partial \log f(X;p)}{\partial p} = \frac{X}{p} - \frac{1-X}{1-p}$$

Taking the derivative again we get

$$\frac{\partial^2 \log f(X;p)}{\partial p^2} = -Xp^{-2} - (1-X)(1-p)^{-2}$$

So

$$\begin{aligned} I(\theta) &= -E\left[\frac{\partial^2 \log f(x;p)}{\partial p^2}\right] = \\ &= E[X]p^{-2} + E[1-X](1-p)^{-2} \\ &= 1/p + 1/(1-p) \\ &= 1/p(1-p) \end{aligned}$$

Motivation

Why are we looking at the information in a data point? Shouldn't we be looking at the information in an entire data *set*?

Motivation

Why are we looking at the information in a data point? Shouldn't we be looking at the information in an entire data *set*?

Yeah, but we start with this because it makes the math easier. Here's the information in a data set:

Because we usually look at i.i.d data, $f(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$. This means

$$\log \prod_{i=1}^{n} f(x_i; \theta) = \sum_{i=1}^{n} \log f(x_i; \theta).$$

Motivation

Why are we looking at the information in a data point? Shouldn't we be looking at the information in an entire data *set*?

Yeah, but we start with this because it makes the math easier. Here's the information in a data set:

Because we usually look at i.i.d data, $f(x_1, \ldots, x_n; \theta) = \prod_{i=1}^n f(x_i; \theta)$. This means

$$\log \prod_{i=1}^n f(x_i; \theta) = \sum_{i=1}^n \log f(x_i; \theta).$$

Finally we take the derivative and then the variance on both sides:

$$I_n(\theta) = V\left[\frac{\partial \log f(x_1, \ldots, x_n; \theta)}{\partial \theta}\right] = \sum_{i=1}^n V\left[\frac{\partial \log f(x_i; \theta)}{\partial \theta}\right] = nI(\theta)$$

Note

Be aware of which information you're using. Either you're talking about a single data point $I(\theta)$, or you're talking about the information in a dataset $I_n(\theta)$. Usually it's clear which one you should be using from the context, but be aware that confusion is a possibility.

Example

Earlier we found the information in a single Bernoulli random variable to be $1/[p(1-p)]$. By the previous slides, the information for $X_1, \ldots, X_n \overset{iid}{\sim}$ Bernoulli(p) is $I_n(\theta) = nI(\theta) = \frac{n}{p(1-p)}$.

Example

Earlier we found the information in a single Bernoulli random variable to be $1/[p(1-p)]$. By the previous slides, the information for $X_1, \ldots, X_n \overset{iid}{\sim}$ Bernoulli(p) is $I_n(\theta) = nI(\theta) = \frac{n}{p(1-p)}$.

...the information increases linearly with the sample size. And take notice that this is the reciprocal of the variance of your sample mean in this same situation.

The Cramér-Rao Inequality

Assume a random sample $X_1, \dots, X_n$ from the distribution with pmf or pdf $f(x;\theta)$ such that the set of possible values does not depend on θ. If the statistics $T = t(X_1, \dots, X_n)$ is an unbiased estimator for the parameter θ, then

$$V(T) \geq \frac{1}{I_n(\theta)}$$

Proof

The "idea" is to consider the correlation between T and the **score** (of the entire data set) $U_n = \frac{\partial}{\partial\theta} \log f(x_1, \ldots, x_n; \theta)$, keeping in mind that $-1 \leq \rho \leq 1$.

$$\begin{aligned}
\left| \frac{\mathrm{Cov}(T, U_n)}{\sqrt{\mathrm{Var}(T)}\sqrt{\mathrm{Var}(U_n)}} \right| \leq 1 &\iff \\
\frac{[\mathrm{Cov}(T, U_n)]^2}{\mathrm{Var}(U_n)} \leq \mathrm{Var}(T) &\iff \\
\frac{[\mathrm{Cov}(T, U_n)]^2}{I_n(\theta)} \leq \mathrm{Var}(T) &\iff
\end{aligned}$$

The rest follows when we show $\mathrm{Cov}(T, U) = 1$ (next slide).

Proof

Since $E[U_n] = 0$ (from a few slides ago)

$$\text{Cov}(T, U_n) = E[TU_n] - E(T)E(U_n) = E[TU_n].$$

Now

$$\begin{aligned}
E[TU_n] &= \sum_x t(x_1, \ldots x_n) \left[\frac{\partial}{\partial\theta} \log f(x_1, \ldots, x_n; \theta)\right] f(x_1, \ldots, x_n; \theta) \\
&= \sum_x t(x_1, \ldots x_n) \left[\frac{\partial}{\partial\theta} f(x_1, \ldots, x_n; \theta)\right] \\
&= \sum_x \frac{\partial}{\partial\theta} [t(x_1, \ldots x_n) f(x_1, \ldots, x_n; \theta)] \\
&= \frac{\partial}{\partial\theta} \sum_x t(x_1, \ldots x_n) f(x_1, \ldots, x_n; \theta) \\
&= \frac{\partial}{\partial\theta} E[T] = \frac{\partial}{\partial\theta} \theta = 1
\end{aligned}$$

Example

Let $X_1, \ldots, X_n \overset{iid}{\sim}$ Bernoulli(p). We showed a few days ago that $V[\bar{X}] = \frac{p(1-p)}{n}$. Then on slide 12 we showed $I_n(\theta) = \frac{n}{p(1-p)}$. So we say that $\bar{X}$ "achieves" the Cramér-Rao lower bound.

Example

Let $X_1, \ldots, X_n \overset{iid}{\sim}$ Bernoulli(p). We showed a few days ago that $V[\bar{X}] = \frac{p(1-p)}{n}$. Then on slide 12 we showed $I_n(\theta) = \frac{n}{p(1-p)}$. So we say that $\bar{X}$ "achieves" the Cramér-Rao lower bound.

This leads us to a new definition. Let T be an unbiased estimator of θ. The ratio of the lower bound to the variance of T is its **efficiency**. So the highest possible effiency is 1.

T is said to be **efficient** if it achieves this Cramér-Rao lower bound and hits this efficiency of 1 target.

Nota Bene

All efficient estimators are MVUE. However, all MVUEs are not necessarily efficient.

This is the only theorem you will learn in this class that shows and estimator T is MVUE.

Motivation

Recall the properties of MLE estimators:

1. is “close to” θ (consistency)
2. is approximately unbiased ($E[\hat{\theta}] \approx \theta$)
3. has variance as nearly as small as possible

Also recall the three different types of convergence we learned (in distribution, in mean, in probability).

Proposition

Let $\hat{\theta}$ be the MLE estimator. The first one says $\hat{\theta} \xrightarrow{p} \theta$. The book doesn't go into details of this and neither will we.

We'll deal with the last two. Together they say $\hat{\theta} \xrightarrow{D} \mathcal{N}(\theta, \frac{1}{I_n(\theta)})$. We'll outline a proof of the equivalent statement:

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{D} \mathcal{N}\left(0, \frac{1}{I(\theta)}\right)$$

(also, we're going to sweep a few details under the rug)

Proof Sketch

Let's start off with a first-order Taylor approximation of the score function $U_n(\theta) = \frac{\partial}{\partial\theta} \log f(x_1, \ldots, x_n; \theta)$ about the mle estimator $\hat{\theta}$

$$U_n(\theta) \approx U_n(\hat{\theta}) + \frac{U_n'(\hat{\theta})(\theta - \hat{\theta})}{1!} = U_n'(\hat{\theta})(\theta - \hat{\theta})$$

re-arranging (and letting $\hat{\theta} \to \theta$) we get

$$\hat{\theta} - \theta \approx \frac{U_n(\theta)}{-U_n'(\theta)}$$

so

$$\sqrt{n}(\hat{\theta} - \theta) \approx \sqrt{n}\frac{U_n(\theta)}{-U_n'(\theta)}$$

Proof Sketch

We have

$$\sqrt{n}(\hat{\theta} - \theta) = \sqrt{n}\frac{U_n(\theta)}{-U_n'(\theta)} = \frac{\frac{1}{n}U_n(\theta)/\sqrt{I(\theta)/n}}{-\frac{1}{n}U_n'(\theta)/\sqrt{I(\theta)}}$$

checking out the denominator...

$$-\frac{1}{n}U_n'(\theta) = \frac{1}{n}\left\{\left[-\frac{\partial^2}{\partial\theta^2}\log f(X_1;\theta)\right] + \cdots + \left[-\frac{\partial^2}{\partial\theta^2}\log f(X_n;\theta)\right]\right\}$$

So by the law of large numbers

$$-\frac{1}{n}U_n'(\theta) \xrightarrow{p} E\left[-\frac{\partial^2}{\partial\theta^2}\log f(X_1;\theta)\right] = I(\theta)$$

Proof Sketch

We have

$$\sqrt{n}(\hat{\theta} - \theta) = \sqrt{n}\frac{U_n(\theta)}{-U_n'(\theta)} = \frac{\frac{1}{n}U_n(\theta)/\sqrt{I(\theta)/n}}{-\frac{1}{n}U_n'(\theta)/\sqrt{I(\theta)}}$$

Looking at the numerator...

$$\frac{1}{n}U_n(\theta)/\sqrt{I(\theta)/n} = \frac{\frac{1}{n}\left[\frac{\partial}{\partial\theta}\log f(x_1;\theta) + \cdots + \frac{\partial}{\partial\theta}\log f(x_n;\theta)\right] - 0}{\sqrt{I(\theta)/n}}$$

So by the central limit theorem

$$\frac{1}{n}U_n(\theta)/\sqrt{I(\theta)/n} \xrightarrow{D} \mathcal{N}(0,1)$$

Proof Sketch

We have

$$\sqrt{n}(\hat{\theta} - \theta) = \sqrt{n}\frac{U_n(\theta)}{-U_n'(\theta)} = \frac{\frac{1}{n}U_n(\theta)/\sqrt{I(\theta)/n}}{-\frac{1}{n}U_n'(\theta)/\sqrt{I(\theta)}} = \frac{A_n}{B_n}$$

1. $A_n \overset{D}{\to} \mathcal{N}(0,1)$
2. $B_n \overset{p}{\to} \sqrt{I(\theta)}$

Finally

So

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{D} \mathcal{N}(0, I^{-1}(\theta))$$

This means we don't always need to find a sampling distribution for $\hat{\theta}$. If we have enough data, we can assume it's approximately normal.

Example 7.35 on page 377

Let $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta) = \theta x^{\theta-1}, 0 < x < 1$. Also, $\theta > 0$. Find $\hat{\theta}$ and its asymptotic distribution.

Example 7.35 on page 377

Let $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta) = \theta x^{\theta-1}, 0 < x < 1$. Also, $\theta > 0$. Find $\hat{\theta}$ and its asymptotic distribution.

$$U = \frac{\partial}{\partial\theta} \log f(x;\theta) = \frac{\partial}{\partial\theta}[\log\theta + (\theta - 1)\log x] = \frac{1}{\theta} + \log x$$

The variance of this is hard to find. Let's take another derivative and find the information that way...

$$U' = -\frac{1}{\theta^2}$$

So $I(\theta) = \frac{1}{\theta^2}$

Example 7.35 on page 377

What about the MLE? First
$\log f(x_1, \ldots, x_n; \theta) = \sum_i [\log \theta + (\theta - 1) \log x_i]$. So

$$\frac{\partial}{\partial \theta} \log f(x_1, \ldots, x_n; \theta) = \frac{n}{\theta} + \sum \log x_i$$

Setting this equal to 0 we get $\hat{\theta} = -\frac{n}{\sum_i \log(x_i)}$.

Remember $I(\theta) = \frac{1}{\theta^2}$. So, approximately,

$$\hat{\theta} \sim \mathcal{N}(\theta, \frac{\theta^2}{n})$$

if n is really large.