STA261: Lecture 3

Sufficiency

Alex Stringer

July 11th, 2018

Disclaimer

The materials in these slides are intended to be a companion to the course textbook, *Mathematical Statistics and Data Analysis, Third Edition*, by John A Rice. Material in the slides may or may not be taken directly from this source. These slides were organized and typeset by Alex Stringer.

A big thanks to Jerry Brunner as well for providing inspiration for assignment questions.

License

Recap

Last week, we learned about **consistency**, which is Property 1 of an estimator: as we get more and more data, estimates from this estimator should get closer and closer to the parameter they are estimating.

We also learned the Method of Moments, which gives consistent estimators algorithmically.

This Week

This week we are going to talk about Property **2**: We should base our estimator off of all the information in the sample. Our estimator should be a *summary* of the full sample, or rather should contain all the information present in the sample about the parameter θ. Whether we know the entire dataset or just our estimate $\hat{\theta}$, we should make the same conclusions regarding θ.

Let's look at an example.

Use all of the data

Suppose we have $X_1, X_2 \sim N(\mu, 1)$ independently, and we wish to estimate μ. I propose the following three estimators:

1. $\hat{\mu}_1 : X_1$
2. $\hat{\mu}_2 : X_2$
3. $\hat{\mu}_3 : \frac{X_1+X_2}{2}$

Which do you prefer?

Use all of the data

It may seem like a silly question; of course we want to use $\bar{X}$ instead of a single data point X_1. But answering *why* is the hard part.

Let's formulate the question a different way: given that we have observed a particular value of X_1, does observing X_2 change the amount of information in the sample about μ?

Conditioning on an estimator

Let's look at what happens if we take $\hat{\mu} = X_1$. I told you X_1 and X_2 were independent, so the conditional distribution of X_2 given X_1 is just

$$f_{X_2|X_1}(x_2|x_1) = f_{X_2}(x_2) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_2 - \mu)^2\right)$$

This conditional distribution is a function of μ. So even knowing the value of our estimator $\hat{\mu} = X_1$, the values that X_2 are likely to take on will depend on μ.

Therefore, knowing X_2 gives us *information* about μ.

Example

Suppose we observe $(X_1, X_2) = (4, 6)$. We know $\sigma = 1$ (it was given in the problem statement).

If $X_1 = 4$, what is the range of plausible values for the parameter μ?

- $\mu = 3$ wouldn't be weird; X_1 would be 1 SD from the mean
- $\mu = 2$ wouldn't be weird; X_1 would be 2 SD from the mean
- $\mu = 5$ wouldn't be weird; X_1 would be 1 SD from the mean
- $\mu = 6$ wouldn't be weird; X_1 would be 2 SD from the mean

Obviously $\mu = 4$ is the least weird value- our "best guess" with the information we have.

Example

If $X_2 = 6$, what are the range of plausible values for the parameter μ?

- $\mu = 5$ wouldn't be weird; X_2 would be 1 SD from the mean
- $\mu = 4$ wouldn't be weird; X_2 would be 2 SD from the mean
- $\mu = 7$ wouldn't be weird; X_2 would be 1 SD from the mean
- $\mu = 8$ wouldn't be weird; X_2 would be 2 SD from the mean

Obviously $\mu = 6$ is the least weird value- our "best guess" with the information we have.

Example

But now the question: having observed $X_1 = 4$, does observing $X_2 = 6$ *change* what we think the plausible values of μ could be?

- $\mu = 2$
- $\mu = 3$
- $\mu = 4$
- $\mu = 5$
- $\mu = 6$
- $\mu = 7$
- $\mu = 8$

Conditioning on an estimator

Now consider $\hat{\mu} = \bar{X}$, so in our example $\hat{\mu} = \frac{X_1+X_2}{2} = 5$. We may consider the conditonal distribution of $X_2|\bar{X} = t$:

$$X_2|\bar{X} = t \sim N(t, 1/2)$$

This does not depend on μ.

Conditioning on an estimator

In particular, once we have observed $\bar{X} = t$, knowing the specific value of X_2 does not change what values of μ we might think are plausible. That is, $\bar{X}$ contains all the *information* in the sample about μ, such that *knowing* $\bar{X}$ *renders the rest of the sample uninformative about* μ.

Keep in mind that when we say "information", we are talking about information with respect to a particular parameter. Knowing the rest of the sample is still necessary if we want to estimate other things, or check model assumptions.

Conditioning on an estimator

For example, the following two samples give the same information about μ:

$$\mathbf{x}_1 = (1, 1.1, 0.9)$$
$$\mathbf{x}_2 = (-9999, 3, 9999)$$

but we wouldn't think those data came from the same distribution. Specifically, they contain different information about σ.

Check out Anscombe's Quartet:
https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Definition: Sufficiency (Textbook, section 8.7)

We can formalize this notion as follows.

Defintion: A statistic $T(X_1, \ldots, X_n)$ is said to be **sufficient** for the parameter θ if the conditional distribution of the sample $X_1, \ldots, X_n$ given $T = t$ does not depend on θ, for any t.

Notes

Technically, any function that takes in data and returns a (possibly lower-dimensional) summary is called a **statistic**, and we often use the term *sufficient statistic*.

Remember how we defined an estimator last week: as a *function* $\hat{\theta} : \mathbb{R}^n \to \mathbb{R}^m$ that takes in realizations of a set of random variables ("data") and returns an estimate of a parameter.

Estmators are statistics; it's just another way of saying it.

Sufficiency says that knowing the sufficient statistic and knowing the whole sample gives us the same information about which values of θ were likely to have generated the observed data.

Example (textbook, page 306)

Let $X_1, \ldots, X_n$ be a random sample from a Bernoulli distribution with parameter θ. Show $\hat{\theta} = \sum_{i=1}^{n} X_i$ is sufficient for θ.

We need to evaluate

$$P(X_1 = x_1, \ldots, X_n = x_n | \hat{\theta} = t) = \frac{P(X_1 = x_1, \ldots, X_n = x_n, \hat{\theta} = t)}{P(\hat{\theta} = t)}$$

and then we can just look at whether or not the result is a function of θ.

Example (textbook, page 306)

In this case we can just work out all the probabilities we need from scratch.

Notice that $\hat{\theta} \sim Bin(n, \theta)$ so $P(\hat{\theta} = t) = \binom{n}{t}\theta^t(1-\theta)^{n-t}$.

Example (textbook, page 306)

For the numerator, first notice that the value of $\hat{\theta}$ is completely determined by the values of $X_1, \ldots, X_n$, so really the joint probability of the sample and the estimator is just the probability of the sample itself. This is the probability of seeing any particular combination of t 1's and $n - t$ 0's, which is equal to the binomial probability function *without* the first factor (why?)

$$\begin{aligned} P(X_1 = x_1, \ldots, X_n = x_n, \hat{\theta} = t) &= P(X_1 = x_1, \ldots, X_n = x_n) \\ &= \theta^t (1 - \theta)^{n-t} \end{aligned}$$

Example (textbook, page 306)

We can evaluate directly:

$$
\begin{aligned}
P(X_1 = x_1, \ldots, X_n = x_n | \hat{\theta} = t) &= \frac{P(X_1 = x_1, \ldots, X_n = x_n, \hat{\theta} = t)}{P(\hat{\theta} = t)} \\
&= \frac{\theta^t (1-\theta)^{n-t}}{\binom{n}{t} \theta^t (1-\theta)^{n-t}} \\
&= \frac{1}{\binom{n}{t}}
\end{aligned}
$$

which does not depend on θ. Hence $\hat{\theta} = \sum_{i=1}^{n} X_i$ is sufficient for θ.

Factorization

You saw that computing this conditional distribution is annoying, and in some cases, not possible analytically. Even in the simple case of the normal conditional on $\overline{X}$, I had to use specifc theorems about the normal distribution that we don't have time to get in to. And even that would only have worked for that example, not in general.

We need a better way to find sufficient estimators, and test if an estimator is sufficient for a parameter.

Factorization Theorem

This is sometimes called "Neyman Factorization", or just "The Factorization Theorem".

Let $\hat{\theta}$ be an estimator of θ. Then $\hat{\theta}$ is sufficient for θ **if and only if** the joint density of $X_1, \ldots, X_n$ can be factored as follows:

$$f_{\mathbf{X}}(x_1, \ldots, x_n) = g(\hat{\theta}, \theta) \times h(\mathbf{x})$$

Example

This looks still to be pretty abstract, but it happens remarkably often.

E.g. if $X_1, \ldots, X_n \sim N(\mu, 1)$ independently then the joint density is

$$
\begin{aligned}
f(\mathbf{x}) &= \prod_{i=1}^{n} f_{x_i}(x_i) \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^n \times \prod_{i=1}^{n} \exp\left(-\frac{1}{2}(x_i - \mu)^2\right) \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^n \times \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right) \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^n \times \exp\left(-\frac{1}{2}\sum_{i=1}^{n} x_i^2\right) \times \exp\left(-\frac{1}{2}\left(-2n\bar{x}\mu + n\mu^2\right)\right) \\
&= h(\mathbf{x}) \times g(\bar{x}, \mu)
\end{aligned}
$$

Example

In the Bernoulli example from before, the joint density of the sample is

$$P(X_1 = x_1, \ldots, X_n = x_n|\theta) = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}$$
$$= h(\mathbf{x}) \times g\left(\sum_{i=1}^n X_i, \theta\right)$$

It's not cheating to take $h(\mathbf{x}) = 1$.

Proof (discrete case only; textbook, page 307)

($\Leftarrow$): Suppose $P(\mathbf{X} = \mathbf{x}) = g(\hat{\theta}, \theta) \times h(\mathbf{x})$. Then

$$\begin{aligned} P(\hat{\theta} = t) &= \sum_{\mathbf{x}|\hat{\theta}(\mathbf{x})=t} P(\mathbf{X} = \mathbf{x}) \\ &= \sum_{\mathbf{x}|\hat{\theta}(\mathbf{x})=t} g(\hat{\theta}, \theta) \times h(\mathbf{x}) \\ &= g(t, \theta) \times c(\mathbf{x}) \end{aligned}$$

where $c(\mathbf{x})$ is a constant that does not depend on θ.

Proof (Textbook, page 307)

The joint distribution $P(\mathbf{X} = \mathbf{x}, \hat{\theta} = t)$ is actually equal to the marginal distribution of the sample, $P(\mathbf{X} = \mathbf{x})$. This is because the value of $\mathbf{X}$ (which remember, we are denoting $\mathbf{x}$) completely determines the value of $\hat{\theta}(\mathbf{x})$.

Another way of putting this is to say that $P(\hat{\theta} = t|\mathbf{X} = \mathbf{x}) = 1$, which implies

$$P(\mathbf{X} = \mathbf{x}, \hat{\theta} = t) = P(\mathbf{X} = \mathbf{x})$$

Proof (Textbook, page 307)

We can then evaluate the conditional distribution

$$\begin{aligned} P(\mathbf{X}=\mathbf{x}|\hat{\theta}=t) &= \frac{P(\mathbf{X}=\mathbf{x},\hat{\theta}=t)}{P(\hat{\theta}=t)} \\ &= \frac{P(\mathbf{X}=\mathbf{x})}{P(\hat{\theta}=t)} \\ &= \frac{g(t,\theta)\times h(\mathbf{x})}{g(t,\theta)\times c(\mathbf{x})} \\ &= \frac{h(\mathbf{x})}{c(\mathbf{x})} \end{aligned}$$

which doesn't depend on θ. Hence $\hat{\theta}$ is sufficient for θ.

Proof (Textbook, page 307)

($\Rightarrow$): now suppose $\hat{\theta}$ is sufficient for θ. Then $P(\mathbf{X} = \mathbf{x}|\hat{\theta})$ doesn't depend on θ. We can write

$$P(\mathbf{X} = \mathbf{x}, \hat{\theta} = t) = P(\mathbf{X} = \mathbf{x}) = P(\hat{\theta} = t) \times P(\mathbf{X} = \mathbf{x}|\hat{\theta} = t)$$

But in this case, $P(\mathbf{X} = \mathbf{x}|\hat{\theta})$ doesn't depend on θ, so we have

$$P(\mathbf{X} = \mathbf{x}) = g(\hat{\theta}, \theta) \times h(\mathbf{x})$$

as was to be shown.

Example

We can look at more examples.

Let $X_i \sim Gamma(\alpha, 1), i = 1 \ldots n$, so that each X_i has density

$$f_{X_i}(x_i) = \frac{1}{\Gamma(\alpha)} x_i^{\alpha-1} e^{-x_i}$$

Show $\hat{\alpha} = \prod_{i=1}^{n} x_i$ is sufficient for α.

Example

Compute the joint density

$$f_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} f_{X_i}(x_i)$$

$$= \left(\frac{1}{\Gamma(\alpha)}\right)^n \left(\prod_{i=1}^{n} x_i\right)^{\alpha-1} \exp\left(-\sum_{i=1}^{n} x_i\right)$$

which factors with

$$g\left(\prod_{i=1}^{n} x_i, \alpha\right) = \left(\frac{1}{\Gamma(\alpha)}\right)^n \left(\prod_{i=1}^{n} x_i\right)^{\alpha-1}$$

$$h(\mathbf{x}) = \exp\left(-\sum_{i=1}^{n} x_i\right)$$

Note

Like with consistency, just because an estimator happens to be sufficient for a parameter doesn't mean that it's a good estimator overall.

For example, $\hat{\alpha} = \frac{\prod_{i=1}^{n} x_i}{1{,}000{,}000}$ is also sufficient for α in the previous example.

One-to-one Functions of Sufficient Statistics are Sufficient

In fact, any one-to-one function of a sufficient statistic is sufficient.

Proof: ...