7.3: Sufficiency

Taylor

University of Virginia

Motivation

When we talked about MLE, we looked at $f(x_1, \ldots, x_n; \theta)$ not as a function of the xs, but as a function of θ.

Taking this idea a bit further, if $f(x_1, \ldots, x_n; \theta)$ doesn't really have a θ in it, then our data would be useless if we were trying to infer things about θ, right?

Now consider this: assume $f(x_1, \ldots, x_n; \theta)$ has a θ in it still. But say we had some test statistic $T_n = T_n(X_1, \ldots, X_n)$. What if we took the conditional distribution

$$f(x_1, \ldots, x_n | T = t)?$$

What if this was free of θ. Does that mean T gives us all the information about θ that we could possibly have?

Definitions

If this were the case, we would call T a **sufficient statistic**.

Definition

A statistic $T = t(X_1, \ldots, X_n)$ is said to be **sufficient** for making inferences about a parameter θ if the conditional distribution of $X_1, \ldots, X_n$ given that $T = t$ does not depend upon θ for every possible value t of the test statistic T.

Example 7.25 on page 361

Say $X_1, X_2, X_3 \overset{iid}{\sim} \text{Poisson}(\lambda)$. Let's define $T = \sum_{i=1}^{3} X_i$. We know that $T \sim \text{Poisson}(3\lambda)$ (show with mgfs).

$$
\begin{aligned}
P(X_1 = 2, X_2 = 1, X_3 = 1 | T = 4) &= \frac{P(X_1 = 2, X_2 = 1, X_3 = 1 \cap T = 4)}{P(T = 4)} \\
&= \frac{P(X_1 = 2, X_2 = 1, X_3 = 1)}{P(T = 4)} \\
&= \frac{\frac{e^{-\lambda}\lambda^2}{2!}\frac{e^{-\lambda}\lambda^1}{1!}\frac{e^{-\lambda}\lambda^1}{1!}}{\frac{e^{-3\lambda}(3\lambda)^4}{4!}} \\
&= \frac{4!}{2!3^4} = 4/27
\end{aligned}
$$

Example

For that value of T, $f(X|T)$ didn't depend on θ. We didn't quite show sufficiency yet, though. It has to be true for any possible T. Fortunately nothing we really did depended on our choice of T. It just had to make sense; that is, the data had to be consistent with it.

$$
\begin{aligned}
&P(X_1 = x_1, X_2 = x_2, X_3 = x_3 | T = t) \\
&= \frac{P(X_1 = x_1, X_2 = x_2, X_3 = x_3 \cap T = t)}{P(T = t)} \\
&= \frac{P(X_1 = x_1, X_2 = x_2, X_3 = x_3)}{P(T = t)} \\
&= \frac{\frac{e^{-\lambda}\lambda^{x_1}}{x_1!}\frac{e^{-\lambda}\lambda^{x_2}}{x_2!}\frac{e^{-\lambda}\lambda^{x_3}}{x_3!}}{\frac{e^{-3\lambda}(3\lambda)^t}{t!}} \\
&= \frac{t!}{x_1!x_2!x_3!3^t}
\end{aligned}
$$

Motivation

How can a sufficient statistic be identified? Using the definition, we would just have to try different Ts until it worked. That would probably take a while.

The Neyman Factorization Theorem

The Neyman Factorization Theorem

Let $f(x_1, \ldots, x_n; \theta)$ denote the joint pmf or pdf of $X_1, \ldots, X_n$. Then $T = t(X_1, \ldots, X_n)$ is a sufficient statistic for θ if and only if the joint pmf/pdf can be represented as a product of two factors in which the first factor involves θ and the data **only through** $t(x_1, \ldots, x_n)$ whereas the second factor involves $x_1, \ldots, x_n$ but **does not depend on** θ:

$$f(x_1, \ldots, x_n; \theta) = g(t; \theta) \cdot h(x_1, \ldots, x_n)$$

A sketch of the proof (discrete case)

1.) Assume T is sufficient. Want to show
$P(\mathbf{X} = \mathbf{x}; \theta) = f(\mathbf{x}; \theta) = g(t; \theta) \cdot h(x_1, \ldots, x_n)$.

$$\begin{aligned} P(\mathbf{X} = \mathbf{x}; \theta) &= P(\mathbf{X} = \mathbf{x}, T = t; \theta) \\ &= P(T = t; \theta) P(\mathbf{X} = \mathbf{x} \mid T = t) \end{aligned}$$

and we're done. $P(T = t; \theta) = g(t; \theta)$ and
$P(\mathbf{X} = \mathbf{x} \mid T = t) = h(x_1, \ldots, x_n)$.

A sketch of the proof (part 2)

2.) Now assume $P(\mathbf{X} = \mathbf{x}; \theta) = f(\mathbf{x}; \theta) = g(t; \theta) \cdot h(x_1, \ldots, x_n)$. We want to show T is sufficient.

$$
\begin{aligned}
P(\mathbf{X} = \mathbf{x} | T = t; \theta) &= \frac{P(\mathbf{X} = \mathbf{x}, T = t; \theta)}{P(T = t; \theta)} \\
&= \frac{P(\mathbf{X} = \mathbf{x}, \theta)}{P(T = t; \theta)} \\
&= \frac{g(t; \theta) h(x_1, \ldots, x_n)}{\sum_{\mathbf{u}: t(\mathbf{u}) = t} P(\mathbf{X} = \mathbf{u}; \theta)} \\
&= \frac{g(t; \theta) h(x_1, \ldots, x_n)}{\sum_{\mathbf{u}: t(\mathbf{u}) = t} g[t(\mathbf{u}); \theta] \cdot h(\mathbf{u})} \\
&= \frac{h(x_1, \ldots, x_n)}{\sum_{\mathbf{u}: t(\mathbf{u}) = t} h(\mathbf{u})}
\end{aligned}
$$

Example 7.27 on page 364

Assume $X_1, \ldots, X_n \overset{iid}{\sim} \text{Uniform}(0, \theta)$. The joint pdf is then

$$f(x_1, \ldots, x_n; \theta) = \theta^{-n} \cdot 1(0 \leq x_{(1)}) \cdot 1(x_{(n)} \leq \theta)$$

And we're done. The Neyman Factorization makes recognizing the sufficient statistic very easy.

1. $\theta^{-n} 1(x_{(n)} \leq \theta)$ is the part that relies the parameter and the sufficient statistic.
2. $1(0 \leq x_{(1)})$ is the part that is free of θ, and relies on the data.

Definitions

What happens if we have more than one parameter? We need *joint sufficiency*.

Definitions

What happens if we have more than one parameter? We need *joint sufficiency*.

Suppose the joint pmf/pdf involves k unknown parameters, call them $\theta_1, \ldots, \theta_k$.

Joint Sufficiency Definition

The m statistics $T_1 = t_1(X_1, \ldots, X_n), \ldots, T_m = t_m(X_1, \ldots, X_n)$ are said to be **jointly sufficient** for the parameters if the conditional distribution of the X_is given $T_1 = t_1, \ldots, T_m = t_m$ does not depend on *any* of the unknown parameters, and this is true for all possible values of $t_1, \ldots, t_m$.

A few things

1. sometimes $m = k$, sometimes $m > k$, sometimes $m < k$
2. 99/100 we'll use Neyman Factorization theorem to verify/identify sufficient statistics

Example 7.29 on page 366

Let $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. The joint pdf is

$$f(\mathbf{x}; \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x_i - \mu)^2}{2\sigma^2}\right]$$

With a bit of algebra,

$$f(\mathbf{x}; \mu, \sigma^2) = \underbrace{\left[\frac{1}{\sigma^n} \exp\left\{-\left(\sum x_i^2 - 2\mu \sum x_i + n\mu^2\right)/2\sigma^2\right\}\right]}_{g(t_1, t_2; \mu, \sigma^2)} \underbrace{\left(\frac{1}{2\pi}\right)^{n/2}}_{h(x_1, \ldots, x_n)}$$

and we're done ($t_1 = \sum_i x_i$ and $t_2 = \sum_i x_i^2$).

Example 7.29 on page 366

Show these:

1. Sufficient statistic for Poisson(λ) random sample is $\sum_i X_i$,
2. Sufficient statistic for Binomial(1,p) random sample is $\sum_i X_i$,
3. Sufficient statistic for Gamma(α, β) random sample is $(\prod_i X_i, \sum_i X_i)$,
4. Sufficient statistic for Uniform($-\theta, \theta$) is $\max_i |X_i|$.

Motivation

We saw $\sum_i X_i$ and $\sum_i X_i^2$ were our sufficient statistics. But we were expecting $\bar{X}$ and $\frac{\sum_i (X_i - \bar{X})^2}{n}$. How can we reconcile this?

Motivation

We saw $\sum_i X_i$ and $\sum_i X_i^2$ were our sufficient statistics. But we were expecting $\bar{X}$ and $\frac{\sum_i (X_i - \bar{X})^2}{n}$. How can we reconcile this?

It turns out that there is a bijective transformation from the first two to the second two. If this is true, can't we just plug in functions of $\bar{X}$ and $\frac{\sum_i (X_i - \bar{X})^2}{n}$ for the original terms $\sum_i X_i$ and $\sum_i X_i^2$?

Motivation

We saw $\sum_i X_i$ and $\sum_i X_i^2$ were our sufficient statistics. But we were expecting $\bar{X}$ and $\frac{\sum_i (X_i - \bar{X})^2}{n}$. How can we reconcile this?

It turns out that there is a bijective transformation from the first two to the second two. If this is true, can't we just plug in functions of $\bar{X}$ and $\frac{\sum_i (X_i - \bar{X})^2}{n}$ for the original terms $\sum_i X_i$ and $\sum_i X_i^2$?

Yes, it's no problem. This is why sufficient statistics aren't *unique*. Or they're unique up to a bijective transformation, at least.

Definition

In that last example, we had n data points, and a sufficient statistic of length 2. What if we just took the sufficient statistic $T_1 = X_1, \ldots, T_n = X_n$? Then we'd have a sufficient statistic of length n, according to our definition.

Definition

In that last example, we had n data points, and a sufficient statistic of length 2. What if we just took the sufficient statistic $T_1 = X_1, \ldots, T_n = X_n$? Then we'd have a sufficient statistic of length n, according to our definition.

A **minimal sufficient statistic** (possibly of length greater than 1) is a function of every other sufficient statistic. It is the maximum possible reduction of the data we can do. That is, it has the minimal dimension.

Definition

In that last example, we had n data points, and a sufficient statistic of length 2. What if we just took the sufficient statistic $T_1 = X_1, \ldots, T_n = X_n$? Then we'd have a sufficient statistic of length n, according to our definition.

A **minimal sufficient statistic** (possibly of length greater than 1) is a function of every other sufficient statistic. It is the maximum possible reduction of the data we can do. That is, it has the minimal dimension.

We only mention this. We're not going to talk too much about it, really.

Motivation

We've seen that sufficient statistics can be used to "reduce" a data set. For example, if we have a trillion points of normal data, we can reduce it down into two numbers, without losing any "information."

Another reason that we learn about these is that they "suggest" what our estimators should be. This is what we'll talk about next.

The Rao-Blackwell Theorem

Rao-Blackwell Theorem

Suppose the joint distribution of $X_1, \ldots, X_n$ depends on some unknown parameter θ and that T is sufficient for θ. Consider estimating $h(\theta)$, a specified function of θ. If U is unbiased for estimating $h(\theta)$, and it does not involve T, then $U^* = E[U|T]$ is also unbiased and has variance no greater than the original unbiased estimator.

Summary: taking the conditional expectation of an estimator can't hurt. The new estimator is still unbiased, and the variance might be lower!

Example

Suppose that the number of major defects on a car has a Poisson distribution with parameter λ. Consider estimating $e^{-\lambda} = P(X = 0)$ Let's use the estimator $U = 1(X_1 = 0)$. Clearly an awful estimator...it will either be 1 or 0, and it is supposed to give us some idea about a probability.

This is unbiased. $EU = 1P(X_1 = 0) + 0P(X_1 \neq 0) = e^{-\lambda}\lambda^0/0! = e^{-\lambda}$. The sufficient statistic is $T = \sum_i X_i$. Our new and improved estimator is $U^* = E[U|T = t]$

Example continued

$$
\begin{aligned}
E[U|T=t] &= 1P(U=1|T=t) + 0P(U=0|T=t) \\
&= P(X_1 = 0 \Big| \sum_i X_i = t) \\
&= \frac{P\left(X_1 = 0 \cap \sum_{i=1}^n X_i = t\right)}{P(\sum_i X_i = t)} \\
&= \frac{P(X_1 = 0 \cap \sum_{i=2}^n X_i = t - 0)}{P(\sum_i X_i = t)} \\
&= \frac{\left[\frac{e^{-\lambda}\lambda^0}{0!}\right]\left[\frac{e^{-(n-1)\lambda}((n-1)\lambda)^t}{t!}\right]}{\left[\frac{e^{-(n)\lambda}((n)\lambda)^t}{t!}\right]} \\
&= \left(\frac{n-1}{n}\right)^t
\end{aligned}
$$

Voila: $\widehat{e^{-\lambda}} = \left(\frac{n-1}{n}\right)^{\sum_i X_i}$

Another Example

Assume $X_1, \ldots, X_n \overset{iid}{\sim} \text{Normal}(\mu, \sigma^2)$. You showed previously that $T_1 = \sum_i X_i$ and $T_2 = \sum_i X_i^2$ are jointly sufficient for the parameters μ and σ^2.

Rao-blackwellize the estimator $\hat{\mu} = \bar{X}$.

Another Example

Assume $X_1, \ldots, X_n \overset{iid}{\sim} \text{Normal}(\mu, \sigma^2)$. You showed previously that $T_1 = \sum_i X_i$ and $T_2 = \sum_i X_i^2$ are jointly sufficient for the parameters μ and σ^2.

Rao-blackwellize the estimator $\hat{\mu} = \bar{X}$.

$$U^* = E[\hat{\mu} | T_1 = t_1, T_2 = t_2] = \hat{\mu} E[1 | T_1 = t_1, T_2 = t_2] = \hat{\mu}$$

Another Example

Assume $X_1, \ldots, X_n \overset{iid}{\sim} \text{Normal}(\mu, \sigma^2)$. You showed previously that $T_1 = \sum_i X_i$ and $T_2 = \sum_i X_i^2$ are jointly sufficient for the parameters μ and σ^2.

Rao-blackwellize the estimator $\hat{\mu} = \bar{X}$.

$$U^* = E[\hat{\mu}|T_1 = t_1, T_2 = t_2] = \hat{\mu}E[1|T_1 = t_1, T_2 = t_2] = \hat{\mu}$$

It's already Rao-Blackwellized because it's a function of the sufficient statistics. Remember we treat this as a "constant" with conditional expectations.

Still Unbiased

The new estimator is still unbiased:

$$\begin{aligned} E[U^*] &= E[E(U|T)] && \text{(definition of } U^*\text{)} \\ &= E[U] && \text{(LTE)} \end{aligned}$$

Less Variance

The new estimator has less variance

$$\begin{aligned}
\text{Var}[U] &= \text{Var}[E(U|T)] + E[\text{Var}(U|T)] && \text{(LTV)} \\
&\geq \text{Var}[E(U|T)] && \text{(expectation of nonnegative r.v.)} \\
&= \text{Var}[U^*] && \text{(definition of } U^*\text{)}
\end{aligned}$$

Nota Bene

Perhaps the inequality in the previous slide is not strict.

Rao-Blackwellized an unbiased estimator only improves its variance. It doesn't guarantee that it's UMVUE/MVUE, but it can't hurt and it's generally easy to do.