Stats 135, Fall 2019
Lecture 12, Wednesday, 9/25/2019

1 Review

Last time, we went over §8.8 and sufficiency. Recall our definition that T is a sufficient statistic for θ if the conditional distribution of $X_1, \dots, X_n \mid T$ does not depend on θ.
Essentially, it allows data reduction (we need not store all the data).
We worked through the example of $X_1, X_2 \sim iid\text{Bernoulli}(\theta)$, and the sufficient statistic of the sample data $T := \overline{X}$. We computed the conditional densitites $f(X_1, X_2 \mid T = t, \theta)$ and we saw that these are not dependent on θ anymore. We are specifying (fixing) θ.
Lucas notes that obviously setting $S := X_1$ (only the first data point) does not give a sufficient statistic. Suppose $(X_1 = 0, X_2 = 0)$, so that $S = 0$. Then

$$f(X_1, X_2 \mid S = 0, \theta) = \frac{f(X_1, X_2 \mid \theta)}{f(S)} = \frac{(1-\theta)(1-\theta)}{(1-\theta)(1-\theta) + (1-\theta)\theta},$$

which certainly is a function of θ.

2 Inventing Shorthand Notation

We'll write $X^n := (X_1, \dots, X_n)$ and $Y^n := (Y_1, \dots, Y_n)$.

Theorem 2.1. (Factorization Theorem)
T is sufficient if and only if

$$f(X^n \mid \theta) = g\left(T(X^n), \theta\right) h(X^n).$$

See Thm A in the book. From this we see that an MLE estimate of θ optimizes $g(T(X^n), \theta)$ and hence is a function of T. In other words, we don't need to know all of our data; only $T(X^n)$. Recall that as a function of θ, the MLE is the θ that maximizes $f(X^n \mid \theta)$, and this is the same as maximizing $g(T(X^n), \theta)$. Notice that this function depends on $T(X^n)$ (not necessarily all the data).
This is the Corollary A (p. 309) of Rice, which states:

If T is sufficient for θ, the MLE is a function of T.

Example: Let $X^n \overset{iid}{\sim} \text{Bernoulli}(\theta)$.
Then

$$g(X^n|\theta) = \underbrace{\theta^{\sum X_i}(1-\theta)^{n-\sum X_i}}_{g(T(X^n,\theta))} \cdot \underbrace{1}_{h(X^n)},$$

where $T(X^n) = \sum X_i$. This proves $T := \sum X_i$ is sufficient for this example. If, for example, we want to show that $\sum X_i$ is a sufficient statistic, we could use this factorization theorem and say that $\sum X_i$ is only a part of this function g. Lucas notes that we may have a T in mind, otherwise it's an argument.

Remark: If we define $T := (X_1, \dots, X_n)$, taking T to be all our data, this would trivially be sufficient by factorization. But $T := \sum X_i$ is better as it leads to a data reduction.

We will eventually get to a minimal sufficient statistic.

2.1 A more complicated example

Let $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Then

$$\begin{aligned} f(X^n|\mu, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(X_i-\mu)^2} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum(X_i-\mu)^2}. \end{aligned}$$

Lucas says we need to do some massaging. The trick here is that we can write

$$\sum(X_i - \mu)^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2 + n(\overline{X} - \mu)^2,$$

where this follows from some algebra. Now given that we can do this, we see:

$$\begin{aligned} f(X^n|\mu) &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum(X_i-\mu)^2} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(X_i-\overline{X})^2 + n(\overline{X}-\mu)^2\right]} \\ &= \underbrace{e^{-\frac{n(\overline{X}-\mu)^2}{2\sigma^2}}}_{g(T(X^n),\mu)} \underbrace{\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} e^{-\frac{\sum(X_i-\overline{X})^2}{2\sigma^2}}}_{h(X^n)} \end{aligned}$$

This implies that $T := \overline{X}$ is a sufficient statistic.

Remark: Note that we CANNOT factor $f(X^n|\mu)$ as:

$$f(X^n|\mu) = \underbrace{e^{-\frac{n(\overline{X}-\mu)^2}{2\sigma^2}} e^{-\frac{\sum(X_i-\overline{X})^2}{2\sigma^2}}}_{g(\overline{X},\mu)} \underbrace{\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}}}_{h(X^n)}.$$

In other words, because this left factor isn't a function of only $\overline{X}$ and μ (it is also a function of $S^2 = \frac{1}{n-1}\sum(X_i - \overline{X})^2$), then $T := (\overline{X}, S^2)$ is a sufficient statistic for $\theta = \mu$ (with σ^2 known and constant), but it is a higher-dimensional statistic than $T = \overline{X}$, which is known as a minimal statistic (lowest-dimensional possible).

2.2 Yet Another Example

Suppose σ^2 is NOT known. Let $X_1, \dots, X_n \sim N(\mu, \sigma^2)$. Then

$$f(X_1, \dots, X_n \mid \mu, \sigma^2) = \underbrace{\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{(n-1)S^2}{2\sigma^2}\right) \cdot \exp\left(\frac{n(\overline{X}-\mu)^2}{2\sigma^2}\right)}_{g(\overline{X},S^2,\mu,\sigma^2)} \cdot \underbrace{1}_{h(X^n)}$$

Recall that given our data, h has to be a number. This factorization shows that

$$T := \left(\overline{X}, S^2\right)$$

is a sufficient statistic for $\theta = (\mu, \sigma^2)$.
Here we need $T = (\overline{X}, S^2)$ since σ^2 is not a constatnt.

2.3 And more examples

Example: Consider $\hat{\theta}_{ML} = \overline{X}$ for $X^n \sim$ Bernoulli(θ) with $T := \overline{X}$. Then

$$\hat{\theta}_{ML} = \overline{X},$$

and our ML estimator is our sufficient statistic.

Example: Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$.
$T = (\overline{X}, S^2)$, $\hat{\theta}_{ML} = \left(\overline{X}, \frac{n-1}{n} S^2\right)$.
The question arises as to why we do not set the sufficient statistic simply to the ML estimator.
Lucas notes that we have an entire family of estimators that are sufficient and this is to introduce that concept.

3 Partitions

Definition: Likelihood Parition (LP) -

We say that likelihoods $f(X^n|\theta)$ and $f(y^n|\theta)$ are equivalent if

$$f(X^n|\theta) = f(Y^n|\theta)$$

for some constant c that may depend on $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$, but not θ. This is called a likelihood partition (LP).

Let's take a look at our familiar example.

Example: Take $X_1, X_2 \sim iid$Bernoulli(θ).
$(1-\theta)\theta$ and itself for $(0,1)$ and $(1,0)$ form a partition because they are multiples (equal) to one another. However, $(1-\theta)\theta$ is not a multiple of $(1-\theta)^2$ or θ^2, so those are in separate partitions.

Definition: Statistic Partition (SP) -
Statistics $T(X^n)$, $T(Y^n)$ are equivalent if

$$T(X^n) = Y(y^n).$$

This is called a statistic partition (SP).

The likelihood partition is the "coarsest" partition (the minimal sufficient statistic), whereas taking all our data gives the "finest" partition (every data entry is in a separate cell).
One way to think about it is to start with our LP as our minimal sufficient statistic, and we get other sufficient statistics by 'adding lines' into our partition.

Definition: -

We say that a statistic $T(X^n)$ is sufficient for θ if and only if SP $\subseteq$ LP, which is to say that SP is contained in (finer than) LP, and that we get SP from LP by adding lines.

Now we can define:

Definition: Minimal Sufficiency -

A statistic is a minimal sufficient statistic (MSS) if SP = LP (in that we get the same partition).