36-705: Intermediate Statistics **Fall 2017**

Lecture 12: September 27

Lecturer: Siva Balakrishnan

Today we will discuss sufficiency in more detail and then begin to discuss some general strategies for constructing estimators.

12.1 Minimal sufficiency

As we have seen previously sufficient statistics are not unique. Furthermore, it seems at least intuitively that some sufficient statistics present much more reduction than others (for instance in the Poisson model both the mean and the entire sample are sufficient).

This motivates the following definition of minimal sufficient statistics:

Minimal Sufficiency: A statistic $T(x_1, \ldots, x_n)$ is minimal sufficient if it is sufficient, and furthermore for any other sufficient statistic $S(x_1, \ldots, x_n)$ we can write $T(x_1, \ldots, x_n) = g(S(x_1, \ldots, x_n))$, i.e. T is a function of S.

There is unfortunately no straightforward way to verify this condition. Analogous to the factorization theorem we have a condition that we can check.

Theorem 12.1 *Define*

$$R(x_1, \ldots, x_n, y_1, \ldots, y_n; \theta) = \frac{p(y_1, \ldots, y_n; \theta)}{p(x_1, \ldots, x_n; \theta)}.$$

Suppose that a statistic T has the following property:

$R(x_1, \ldots, x_n, y_1, \ldots, y_n; \theta)$ **does not depend on** θ **if and only if** $T(y_1, \ldots, y_n) = T(x_1, \ldots, x_n)$**.**

Then T is a MSS.

Before we prove the theorem let us consider some examples.

Example 12.2 *Suppose that $Y_1, \ldots, Y_n$ are i.i.d Poisson (θ).*

$$p(y_1, \ldots, y_n; \theta) = \frac{e^{-n\theta}\theta^{\sum y_i}}{\prod y_i}, \quad \frac{p(y_1, \ldots, y_n; \theta)}{p(x_1, \ldots, x_n; \theta)} = \frac{\theta^{\sum y_i - \sum x_i}}{\prod y_i! / \prod x_i!}$$

which is independent of θ iff $\sum y_i = \sum x_i$. This implies that $T(X_1, \ldots, X_n) = \sum_{i=1}^n X_i$ is a minimal sufficient statistic for θ.

The minimal sufficient statistic is not unique. But, the minimal sufficient partition is unique.

Example 12.3 *Cauchy.*

$$p(x;\theta) = \frac{1}{\pi(1+(x-\theta)^2)}.$$

Then

$$\frac{p(y_1,\ldots,y_n;\theta)}{p(x_1,\ldots,x_n;\theta)} = \frac{\prod_{i=1}^{n}\{1+(x_i-\theta)^2\}}{\prod_{j=1}^{n}\{1+(y_j-\theta)^2\}}.$$

The ratio is a constant function of θ if

$$T(X_1,\ldots,X_n) = (X_{(1)},\cdots,X_{(n)}).$$

It is technically harder to show that the ratio is independent of θ only if T is the order statistics, but it could be done using theorems about polynomials. Having shown this, one can conclude that the order statistics are the minimal sufficient statistics for θ.

Proof: This proof is a bit technical so feel free to skip it.

We prove this in two steps. We first show that T is a sufficient statistic and then we check that it is minimal. We define the partition induced by T, as $\{A_t : t \in \text{Range}(T)\}$ and for each set in the partition A_t we associate a representative $(x_{t1},\ldots,x_{tn}) \in A_t$.

T **is sufficient:** We look at the joint distribution at any $(x_1,\ldots,x_n)$. Suppose that $T(x_1,\ldots,x_n) = u$, then consider $(y_1,\ldots,y_n) := (x_{u1},\ldots,x_{un})$. Observe that, $(y_1,\ldots,y_n)$ depends only on $T(x_1,\ldots,x_n)$, i.e. the point y is a function of the statistic T only. Now we have that,

$$p(x_1,\ldots,x_n;\theta) = p(y_1,\ldots,y_n;\theta)R(y_1,\ldots,y_n,x_1,\ldots,x_n;\theta),$$

and since $T(x_1,\ldots,x_n) = T(y_1,\ldots,y_n)$, R does not depend on θ. Recalling that $(y_1,\ldots,y_n)$ is only a function of $T(x_1,\ldots,x_n)$ we have that,

$$p(x_1,\ldots,x_n;\theta) = g(T(x_1,\ldots,x_n);\theta)h(x_1,\ldots,x_n),$$

where g corresponds to the first term and h corresponds to the R term. We conclude that T is sufficient.

T **is minimal:** As a preliminary we note that the definition of a minimal sufficient statistic could be equivalently written as: T is a MSS if for any other sufficient statistic S, if we have that $S(x_1,\ldots,x_n) = S(y_1,\ldots,y_n)$ then we also have that $T(x_1,\ldots,x_n) = T(y_1,\ldots,y_n)$. This is equivalent to the statement that T is a function of S.

Consider, any other sufficient statistic S. Suppose that, $S(x_1, \ldots, x_n) = S(y_1, \ldots, y_n)$, then by the factorization theorem we have that,

$$\begin{aligned} p(x_1, \ldots, x_n; \theta) &= g(S(x_1, \ldots, x_n); \theta) h(x_1, \ldots, x_n) \\ &= g(S(y_1, \ldots, y_n); \theta) h(y_1, \ldots, y_n) \frac{h(x_1, \ldots, x_n)}{h(y_1, \ldots, y_n)} \\ &= p(y_1, \ldots, y_n; \theta) \frac{h(x_1, \ldots, x_n)}{h(y_1, \ldots, y_n)}, \end{aligned}$$

so we have that $R(x_1, \ldots, x_n, y_1, \ldots, y_n; \theta)$ does not depend on θ. So we conclude that $T(x_1, \ldots, x_n) = T(y_1, \ldots, y_n)$ and so T is minimal. ■

12.2 Minimal sufficiency and the likelihood

Although minimal sufficient statistics are not unique they induce a unique partition on the possible datasets. This partition is also induced by the likelihood, i.e.

Suppose we have a partition such that $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$ are placed in the same set of the partition iff $L(\theta; x_1, \ldots, x_n) \propto L(\theta; y_1, \ldots, y_n)$, then the partition is the minimal sufficient partition.

You will prove this on your homework but it is a simple consequence of the characterization we have seen in the previous section.

12.3 Sufficiency - the risk reduction viewpoint

We will return to the concept of risk more formally in the next few lectures, but for now let us try to understand the main ideas.

Setting: Suppose we observe $X_1, \ldots, X_n \sim p(X; \theta)$ and we would like to estimate θ, i.e. we want to construct some function of the data that is close in some sense to θ. We construct an estimator $\widehat{\theta}(X_1, \ldots, X_n)$. In order to evaluate our estimator we might consider how far our estimate is from θ on average, i.e. we can define

$$R(\widehat{\theta}, \theta) = \mathbb{E}(\widehat{\theta} - \theta)^2.$$

We will see this again later on but the risk of an estimator can be decomposed into its bias and variance, i.e.

$$\mathbb{E}(\widehat{\theta} - \theta)^2 = (\mathbb{E}\widehat{\theta} - \theta)^2 + \mathbb{E}(\widehat{\theta} - \mathbb{E}\widehat{\theta})^2,$$

where the first term is referred to as the bias and the second is the variance.

There is a strong sense in which estimators which do not depend only on sufficient statistics can be improved. This is known as the Rao-Blackwell theorem.

Let $\widehat{\theta}$ be an estimator. Let T be any sufficient statistic and define $\widetilde{\theta} = \mathbb{E}[\widehat{\theta}|T]$.

Rao-Blackwell theorem:

$$R(\widetilde{\theta}, \theta) \leq R(\widehat{\theta}, \theta).$$

We will not spend too much time on this but lets see a quick example and then prove the result.

Example: Suppose we toss a coin n times, i.e. $X_1, \ldots, X_n \sim \text{Ber}(\theta)$. We consider the estimator:

$$\widehat{\theta} = X_1,$$

and the sufficient statistic $T = \sum_{i=1}^n X_i$, then

$$\widetilde{\theta} = \mathbb{E}[X_1|T] = \mathbb{E}[X_1|\sum_i X_i].$$

I claim that the conditional expectation is simply the average, i.e.

$$\widetilde{\theta} = \frac{1}{n}\sum_{i=1}^n X_i.$$

First, let us check this in the case when $n = 2$. If $X_1 + X_2 = 2$ then $X_1 = 1$, and if $X_1 + X_2 = 0$, $X_1 = 0$. In the case, when $X_1 + X_2 = 1$, we have $X_1 = 1$ with probability $1/2$ and 0 with probability $1/2$. So we conclude the conditional expectation is $(X_1 + X_2)/2$.

More generally, if we have $\sum X_i = k$, then of the $\binom{k}{n}$ equally likely possibilities we have that $X_1 = 1$ for $\binom{k-1}{n-1}$ of them so that the conditional expectation is simply:

$$\mathbb{E}[X_1|\sum_i X_i = k] = \frac{\binom{k}{n}}{\binom{k-1}{n-1}} = \frac{k}{n},$$

as desired.

We observe that both estimators are unbiased but the variance of the Rao-Blackwellized estimator is $\theta(1-\theta)/n$ as opposed to the original estimator which has variance $\theta(1-\theta)$.

Proof of Rao-Blackwell: Observe that,

$$\begin{aligned} R(\widetilde{\theta}, \theta) &= \mathbb{E}[(\mathbb{E}[\widehat{\theta}|T] - \theta)^2] \\ &= \mathbb{E}[(\mathbb{E}[\widehat{\theta} - \theta|T])^2] \\ &\leq \mathbb{E}[\mathbb{E}[(\widehat{\theta} - \theta)^2|T]] \\ &= R(\widehat{\theta}, \theta). \end{aligned}$$

The inequality is Jensen's inequality (equivalently just $\text{Var(X)} = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \geq 0$).

A question worth pondering is: why does it matter for Rao-Blackwellization that T is a sufficient statistic?

12.4 More examples with the likelihood

Example 12.4 *Suppose that* $X = (X_1, X_2, X_3) \sim \text{Multinomial}(n, p)$ *where*

$$p = (p_1, p_2, p_3) = (\theta, \theta, 1 - 2\theta).$$

So

$$p(x; \theta) = \binom{n}{x_1\ x_2\ x_3} p_1^{x_1} p_2^{x_2} p_3^{x_3} = \theta^{x_1 + x_2}(1 - 2\theta)^{x_3}.$$

Suppose that $X = (1, 3, 2)$. *Then*

$$L(\theta) = \frac{6!}{1!\ 3!\ 2!}\theta^1\theta^3(1 - 2\theta)^2 \propto \theta^4(1 - 2\theta)^2.$$

Now suppose that $X = (2, 2, 2)$. *Then*

$$L(\theta) = \frac{6!}{2!\ 2!\ 2!}\theta^2\theta^2(1 - 2\theta)^2 \propto \theta^4(1 - 2\theta)^2.$$

Hence, the likelihood function is the same for these two datasets.

Example 12.5 $X_1, \cdots, X_n \sim N(\mu, 1)$. *Then,*

$$L(\mu) = \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right\} \propto \exp\left\{-\frac{n}{2}(\overline{x} - \mu)^2\right\}.$$

Example 12.6 *Let* $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. *Then*

$$L(p) \propto p^X(1 - p)^{n-X}$$

for $p \in [0, 1]$ *where* $X = \sum_i X_i$.

12.5 Estimation

Now we begin discussing more formally the estimation problem.

$X_1, \ldots, X_n \sim p(x;\theta)$. Want to estimate $\theta = (\theta_1, \ldots, \theta_k)$. An *estimator*

$$\widehat{\theta} = \widehat{\theta}_n = w(X_1, \ldots, X_n)$$

is a function of the data. Keep in mind that the parameter is a fixed, unknown constant. The estimator is a random variable.

For now, we will discuss three methods of constructing estimators:

1. The Method of Moments (MOM)
2. Maximum likelihood (MLE)
3. Bayesian estimators.

Some Terminology. Throughout these notes, we will use the following terminology:

1. $\mathbb{E}_\theta(\widehat{\theta}) = \int \cdots \int \widehat{\theta}(x_1, \ldots, x_n) p(x_1;\theta) \cdots p(x_n;\theta) dx_1 \cdots dx_n$.
2. Bias: $\mathbb{E}_\theta(\widehat{\theta}) - \theta$.
3. The distribution of $\widehat{\theta}$ is called its *sampling distribution*.
4. The standard deviation of $\widehat{\theta}$ is called the *standard error* denoted by $\text{se}(\widehat{\theta})$.
5. $\widehat{\theta}$ is *consistent* if $\widehat{\theta} \xrightarrow{p} \theta$ as $n \to \infty$.
6. Later we will see that if bias $\to 0$ and $\text{Var}(\widehat{\theta}) \to 0$ as $n \to \infty$ then $\widehat{\theta}$ is consistent.

12.6 The Method of Moments

Suppose that $\theta = (\theta_1, \ldots, \theta_k)$. Define

$$\begin{aligned}
m_1 &= \frac{1}{n}\sum_{i=1}^n X_i, & \mu_1(\theta) &= \mathbb{E}(X_i) \\
m_2 &= \frac{1}{n}\sum_{i=1}^n X_i^2, & \mu_2(\theta) &= \mathbb{E}(X_i^2) \\
&\vdots & &\vdots \\
m_k &= \frac{1}{n}\sum_{i=1}^n X_i^k, & \mu_k(\theta) &= \mathbb{E}(X_i^k).
\end{aligned}$$

Let $\widehat{\theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_k)$ solve:

$$m_j = \mu_j(\widehat{\theta}), \quad j = 1, \ldots, k.$$

In other words, we equate the first k sample moments with the first k theoretical moments. This defines k equations with k unknowns.

Example 12.7 *$N(\beta, \sigma^2)$ with $\theta = (\beta, \sigma^2)$. Then $\mu_1 = \beta$ and $\mu_2 = \sigma^2 + \beta^2$. Equate:*

$$\frac{1}{n}\sum_{i=1}^n X_i = \widehat{\beta}, \quad \frac{1}{n}\sum_{i=1}^n X_i^2 = \widehat{\sigma}^2 + \widehat{\beta}^2$$

to get

$$\widehat{\beta} = \overline{X}_, \quad \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2.$$

Example 12.8 *Suppose*

$$X_1, \ldots, X_n \sim \text{Binomial}(k, p)$$

where both k and p are unknown. We get

$$kp = \overline{X}_n, \quad \frac{1}{n}\sum_{i=1}^n X_i^2 = kp(1-p) + k^2p^2$$

giving

$$\widehat{p} = \frac{\overline{X}}{k}, \quad \widehat{k} = \frac{\overline{X}^2}{\overline{X} - \frac{1}{n}\sum_i (X_i - \overline{X})^2}.$$

The method of moments was popular many years ago because it is often easy to compute. Lately, it has attracted attention again. For example, there is a large literature on estimating "mixtures of Gaussians" using the method of moments.