Today we will continue our discussion on sufficiency.

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 example, suppose that $X_1, \ldots, X_n \sim N(\mu, 1)$. Then $\overline{X}_n$ is a sufficient statistic. But so is the whole data set. 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.

Theorem 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 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 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: 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,

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

$$= g(S(y_1, \ldots, y_n); \theta)h(y_1, \ldots, y_n)\frac{h(x_1, \ldots, x_n)}{h(y_1, \ldots, y_n)}$$

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

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. ∎

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.

Lemma 4 *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.

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

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

We 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 $\binom{n}{k}$ equally likely possibilities we have that $X_1 = 1$ for $\binom{n-1}{k-1}$ of them so that the conditional expectation is simply:

$$\mathbb{E}\left[X_1 \ \Big| \ \sum_i X_i = k\right] = \frac{\binom{n-1}{k-1}}{\binom{n}{k}} = \frac{k}{n} = \overline{X}_n.$$

We observe that both estimators are unbiased (have mean equal to θ) 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,

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

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?