Chapter 6

Estimation: Supplementary materials

6.5 Sufficiency

This section is about *sufficiency* — whether a statistic (or vector of statistics) contain all the information in a sample that is relevant to a parameter. For example, it turns out that $\bar{X}$ contains all the information in a sample that is relevant for the mean μ of a normal population, but the sample median does not. For a Cauchy population neither is sufficient.

When sufficient statistics exist, we can simplify the process of choosing a good estimator by focusing on estimators that use the sufficient statistics, hence ignoring irrelevant information. And if an estimator is based on irrelevant information, we can improve on it.

We'll start with a simple example, then continue with a more formal approach.

Example 1. Let's return to the weighted coin that you bought at a magic shop — the one with unknown probability of heads p (Section 6.1.1). We saw earlier that on one sequence of flips of the coin, we observed HHHTHTHT. We obtained the maximum likelihood estimate $\hat{p} = 5/8$.

Now imagine that an analyst only records the number of heads, $y = 5$. Is that the only thing that matters (for estimating p)? Or does the order also matter? Conditional on the number of heads, is the order in which the heads occur related to the parameter?

We answer these questions using conditional probabilities, $P_\theta(\text{data}|\text{number of heads})$. If this does not depend on θ then the order is irrelevant.

Note — earlier in the book we would have written that as $P(\text{data}|\text{number of heads})$, with the dependence on θ implicit. In this section it will be important to distinguish when something depends on the parameter and when it does not, so we make the dependence explicit.

Now, given $y = 5$, what is the probability of the sequence HHHTHTHT? Using $X_i = 1$ if the i^{th} flip is heads, 0 otherwise, we compute

$$\begin{aligned}
&P_p(X_1 = 1, X_2 = 1, X_3 = 1, X_4 = 0, X_5 = 1, X_6 = 0, X_7 = 1, X_8 = 0 | y = 5) \\
&= \frac{P_p(X_1 = 1, X_2 = 1, X_3 = 1, X_4 = 0, X_5 = 1, X_6 = 0, X_7 = 1, X_8 = 0, y = 5)}{P_p(y = 5)} \\
&= \frac{p^5(1-p)^3}{\binom{8}{5} p^5(1-p)^3} \\
&= \frac{1}{\binom{8}{5}}
\end{aligned}$$

For the denominator probability $P_p(y = 5)$ above, in a sequence of length 8, there are $\binom{8}{5}$ ways to choose the positions were the heads go, and those heads occur with probability p^5, and the rest of the positions are tails which occur with probability $(1-p)^3$.

Note that the final probability $1/\binom{8}{5}$ is independent of p.

Similarly if we had observed another sequence like HHHHTHTT — the maximum likelihood estimate of p would still be $5/8$ and using the same argument as above, the conditional probability of seeing this sequence given $y = 5$ would again be $1/\binom{8}{5}$. Thus, knowing the exact sequence does not provide any additional information about p than what the estimate $y = 5$ provides. We will say that the sample proportion is a sufficient estimator for p.

Similarly, the statistic (an estimator) $\hat{p} = 5/8$ is sufficient.

Before we formally define sufficiency, let's look at an example of a statistic that is not sufficient.

Example 2. We saw earlier that for $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} \text{Unif}[0, \beta]$, the method of moments estimator of β is $\hat{\beta} = 2\bar{X}$. Suppose Deepak obtains the sample $X_1 = 1, X_2 = 3, X_3 = 5, X_4 = 6$ and finds the methods of moments estimate $\hat{\beta} = 2(3.75) = 7.5$. Another researcher Claire obtains the sample $Y_1 = 1.5, Y_2 = 1.5, Y_3 = 4, Y_4 = 8$ and her method of moments estimate of β is also 7.5. Do the two samples provide the same information about the possible value of β? For Deepak, the value $\beta = 7.5$ is possible, while for Claire it is impossible because she would never observe $Y_4 = 8$. Thus, the method of moments estimator $\hat{\beta} = 2\bar{X}$ is not sufficient for β.

Definition 1. Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with pdf $f(x; \theta)$. The statistic $T = T(X_1, X_2, \ldots, X_n)$ is *sufficient* for θ if, for all t, the conditional distribution of $X_1, X_2, \ldots, X_n$ given $T = t$ does not depend on θ.

Remark If T is sufficient for the parameter θ, then any one-to-one function of T is also sufficient for θ. In particular, if T is sufficient for θ, then cT where c is a non-zero constant, is also sufficient. (See Exercise 5) ||

Example 3. Suppose that X is a random variable that takes on one of four possible values $\{a_1, a_2, a_3, a_4\}$, with probabilities that depend on a parameter θ that has two possible values. The corresponding probabilities are given in this table:

	a_1	a_2	a_3	a_4
θ_1	1/4	1/4	1/4	1/4
θ_2	1/2	1/6	1/6	1/6

We draw a single value $X = x$ and define the statistic $T(x) = 1$ if $X = a_1$ and 0 otherwise. It looks like this statistic contains all the relevant information; if $T = 1$ then θ_2 is more likely, while if $T = 0$, then θ_1 is more likely, and given that $T = 0$, it really doesn't matter whether $X = a_2, a_3,$ or a_4.

We'll check this using the definition. We need the conditional distributions of X given $T = t$, one for each value of θ.

For $\theta = \theta_1$ and $X = a_1$, we have

$$P_{\theta_1}(X = a_1|T = 1) = \frac{P_{\theta_1}(T = 1|X = a_1)P_{\theta_1}(X = a_1)}{P_{\theta_1}(X = a_1)} = \frac{1 \times 1/4}{1/4} = 1$$

and for $i = 2, 3, 4$,

$$P_{\theta_1}(X = a_i|T = 1) = \frac{P_{\theta_1}(T = 1|X = a_i)}{P_{\theta_1}(X = a_i)} = 0$$

For $\theta = \theta_2$ and $X = a_1$, we have

$$P_{\theta_2}(X = a_1|T = 1) = \frac{1 \times 1/2}{1/2} = 1$$

and for $i = 2, 3, 4$,

$$P_{\theta_2}(X = a_i|T = 1) = \frac{P_{\theta_2}(T = 1|X = a_i)}{P_{\theta_2}(X = a_i)} = 0$$

The two conditional distributions of X are the same when $T = 1$. Similar steps show that they are also the same when $T = 0$. Hence the conditional distribution does not depend on θ, so T is sufficient.

In general, computing this conditional distribution is not easy, so we rely on the following instead:

Theorem 1. *(**Factorization Criterion**) Let $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ be a random sample from a distribution with pdf $f(x; \theta)$ and $t = T(x_1, x_2, \ldots, x_n)$ a statistic. t is a sufficient statistic for θ if and only if for all values θ, the likelihood $L(\theta|x_1, x_2, \ldots, x_n) = L(\theta)$ can be written as*

$$L(\theta) = g(t, \theta)h(x_1, x_2, \ldots, x_n)$$

where g and h are real-valued functions.

Proof. We will give the proof in the case that the distribution is discrete. Let $p(x;\theta)$ denote the probability mass function of X.

First assume the likelihood factors in this way. By the definition of conditional probability,

$$\begin{aligned} &P_\theta(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n | T = t) \\ &\quad = \frac{P_\theta(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n, T = t)}{P_\theta(T = t)} \\ &\quad = \frac{P_\theta(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)}{P_\theta(T = t)} \end{aligned}$$

Now, the numerator is

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \prod_{i=1}^{n} p(x_i;\theta) = L(\theta) = g(t,\theta)h(x_1, x_2, \ldots, x_n)$$

In the denominator, we have

$$\begin{aligned} P_\theta(T = t) &= \sum_{\substack{(y_1,y_2,\ldots,y_n) \\ T(y_1,y_2,\ldots,y_n)=t}} P_\theta(X_1 = y_1, X_2 = y_2, \ldots, X_n = y_n) \\ &= \sum_{\substack{(y_1,y_2,\ldots,y_n) \\ T(y_1,y_2,\ldots,y_n)=t}} g(T(y_1, y_2, \ldots, y_n), \theta)h(y_1, y_2, \ldots, y_n) \\ &= g(t,\theta) \sum_{\substack{(y_1,y_2,\ldots,y_n) \\ T(y_1,y_2,\ldots,y_n)=t}} h(y_1, y_2, \ldots, y_n) \end{aligned}$$

Thus,

$$P_\theta(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n | T = t) = \frac{h(x_1, x_2, \ldots, x_n)}{\displaystyle\sum_{\substack{(y_1,y_2,\ldots,y_n) \\ T(y_1,y_2,\ldots,y_n)=t}} h(y_1, y_2, \ldots, y_n)},$$

which does not depend on θ.

On the other hand, suppose $T = t$ is a sufficient statistic. Then,

$$\begin{aligned} L(\theta) &= P_\theta(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \\ &= P_\theta(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n, T = t) \\ &= P_\theta(T = t)P_\theta(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n | T = t) \\ &= P_\theta(T = t)P(X_1 = x_1, X_2, \ldots, X_n = x_n | T = t) \end{aligned}$$

This satisfies the theorem, with $g(t,\theta) = P_\theta(T = t)$ and $h(x_1, x_2, \ldots, x_n) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$. □

Example 4. Let $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ be a random sample from the Poisson distribution with parameter $\lambda > 0$. We have seen that $\hat{\lambda} = \bar{x}$ is an estimate of λ.

$$\begin{aligned} L(\lambda) &= \prod_{i=1}^{n} \frac{\lambda_i^x e^{-\lambda}}{x_i!} \\ &= \frac{\lambda^{\sum_{i=1}^n x_i} e^{-n\lambda}}{\prod_{i=1}^n x_i!} \\ &= \frac{\lambda^{n\bar{x}} e^{-n\lambda}}{\prod_{i=1}^n x_i!} \end{aligned}$$

This satisfies the factorization theorem with $g(\bar{x}, \lambda) = \lambda^{n\bar{x}} e^{-n\lambda}$ and $h(x_1, x_2, \ldots, x_n) = 1/\prod_{i=1}^n x_i!$, hence $\hat{\lambda} = \bar{x}$ is a sufficient statistic.

Example 5. Let $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ be a random sample from a distribution with pdf $f(x; \lambda) = e^{-(x-\lambda)}$ for $x \geq \lambda$, 0 otherwise (a shifted exponential distribution). Let $t = x_{\min} = \min\{x_1, x_2, \ldots, x_n\}$. Show that $x_{\min}$ is a sufficient statistic for λ.

We write the density as $f(x; \lambda) = e^{-(x-\lambda)} I(x \geq \lambda)$ where $I(A)$ is the indicator function, with value 1 if A is true and 0 otherwise.

$$\begin{aligned} L(\lambda) &= \prod_{i=1}^{n} e^{-(x_i-\lambda)} I(x_i \geq \lambda) \\ &= e^{-\sum_{i=1}^n (x_i-\lambda)} I(x_i \geq \lambda \ \forall i) \\ &= e^{-\sum_{i=1}^n x_i} e^{n\lambda} I(x_{\min} \geq \lambda). \end{aligned}$$

This satisfies the factorization theorem with $g(x_{\min}, \lambda) = e^{n\lambda} I(x_{\min} \geq \lambda)$ and $h(x_1, x_2, \ldots, x_n) = e^{-\sum_{i=1}^n x_i}$.

Note that other statistics like $\bar{x}$ or $x_{\max}$ are not sufficient; it is only $x_{\min}$ that satisfies $I(x_i \geq \lambda \ \forall i) = I(x_{\min} \geq \lambda)$.

6.5.1 The Rao-Blackwell Theorem

We now turn to why sufficiency is useful. As we noted in the beginning of this chapter:

> When sufficient statistics exist, we can simplify the process of choosing a good estimator by focusing on estimators that use the sufficient statistics, hence ignoring irrelevant information. And if an estimator is based on irrelevant information, we can improve on it.

We begin with a theorem that tells us how to improve on estimators.

Theorem 2. The Rao-Blackwell Theorem *Let $\hat{\theta}$ be an unbiased estimator of θ and $\mathrm{Var}[\hat{\theta}] < \infty$. Suppose that T is a sufficient statistic for θ and let $\tilde{\theta} =$*

$\mathrm{E}[\hat{\theta} \mid T]$. *Then* $\tilde{\theta}$ *is an unbiased estimator of* θ *and* $\mathrm{Var}[\tilde{\theta}] \leq \mathrm{Var}[\hat{\theta}]$. *Furthermore,* $\mathrm{Var}[\tilde{\theta}] < \mathrm{Var}[\hat{\theta}]$ *for some* θ *unless* $\tilde{\theta} = \hat{\theta}$ *with probability one.*

Proof. We will need two results from probability: if X and Y are jointly distributed, then

$$\mathrm{E}[X] = \mathrm{E}[\mathrm{E}[X \mid Y]] \tag{6.1}$$

$$\mathrm{Var}[X] = \mathrm{Var}[\mathrm{E}[X \mid Y]] + \mathrm{E}[\mathrm{Var}[X \mid Y]]. \tag{6.2}$$

See Chapter 7 in Ross (*A First Course in Probability*).

Since T is sufficient for θ, $\tilde{\theta} = \mathrm{E}[\hat{\theta} \mid T]$ is not a function of θ. Hence, $\tilde{\theta}$ is also a statistic.

Thus, by Equation 6.1,

$$\mathrm{E}[\tilde{\theta}] = \mathrm{E}[\mathrm{E}[\hat{\theta} \mid T]] = \mathrm{E}[\hat{\theta}] = \theta.$$

We conclude that $\tilde{\theta}$ is an unbiased estimator of θ.

Using Equation 6.2, we have

$$\begin{aligned}\mathrm{Var}[\hat{\theta}] &= \mathrm{Var}[\mathrm{E}[\hat{\theta} \mid T]] + \mathrm{E}[\mathrm{Var}[\hat{\theta} \mid T]] \\ &= \mathrm{Var}[\tilde{\theta}] + \mathrm{E}[\mathrm{Var}[\hat{\theta} \mid T]]\end{aligned}$$

Thus $\mathrm{Var}[\hat{\theta}] \geq \mathrm{Var}[\tilde{\theta}]$ since $\mathrm{Var}[\hat{\theta} \mid T = t] \geq 0$ for all t. □

Thus, we see that conditioning an unbiased estimator on a sufficient statistic will lead to an unbiased estimator with smaller variance, so in searching for a "best" unbiased estimator, we will want to consider statistics that are functions of a sufficient statistic.

Example 6. Let $X_i \overset{i.i.d.}{\sim} \mathrm{Bern}(p)$, $i = 1, 2, \ldots, n$, Then $\hat{p} = X_1$ is an unbiased estimator of p with $\mathrm{Var}[\hat{p}] = p(1-p)$. In Exercise 1 of the previous section, we showed that $T = \sum_{i=1}^{n} X_i$ is sufficient for p. Let $\tilde{p} = \mathrm{E}[\hat{p} \mid T]$, and suppose $T = k$. Then

$$\begin{aligned}\tilde{p} = \mathrm{E}[X_1 \mid \sum_{i=1}^{n} X_i = k] &= 1 \times P(X_1 = 1 \mid \sum_{i=1}^{n} X_i = k) + 0 \times P(X_1 = 0 \mid \sum_{i=1}^{n} X_i = k) \\ &= \frac{P(X_1 = 1, \sum_{i=1}^{n} X_i = k)}{P(\sum_{i=1}^{n} X_i = k)} \\ &= \frac{\binom{n-1}{k-1} p^k (1-p)^{n-k}}{\binom{n}{k} p^k (1-p)^{n-k}} \\ &= \frac{k}{n}\end{aligned}$$

Thus,

$$\mathrm{Var}[\tilde{p}] = \mathrm{Var}[\frac{k}{n}] = \mathrm{Var}[\frac{\sum_{i=1}^{n} X_i}{n}] = \frac{1}{n^2} \sum_{i}^{n} \mathrm{Var}[X_i] = \frac{p(1-p)}{n}$$

which is a much smaller variance than $\mathrm{Var}[\hat{p}] = p(1-p)$.

Example 7. Consider the Deepak/Claire example above: $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim}$ Unif$[0, \beta]$, with $\hat{\beta} = 2\bar{X}$. A sufficient statistic in this case is $T = X_{max} = \max\{X_1, X_2, \ldots, X_n\}$, the largest value. The conditional expected value of the sample mean is $\mathrm{E}[\bar{X}|X_{max}] = \frac{1}{n}(X_{max} + (n-1)X_{max}/2) = \frac{n+1}{2n}X_{max}$ (there is one observation equal to X_{max}, and the remaining observations have conditional mean $X_{max}/2$) so $\tilde{\beta} = \mathrm{E}[(]\hat{\beta}|T) = \frac{n+1}{n}X_{max}$ is the improved estimator (we previously encountered this estimator in Example 6.13).

This process, of taking an unbiased estimator and obtaining a new lower-variance estimator as the expected value given a sufficient statistic, is known as *Rao-Blackwellization*.

While the Rao-Blackwell Theorem is for unbiased estimators, a more general result holds — for any estimator $\hat{\theta}$ (whether unbiased or not), taking the expected value of that estimator given a sufficient statistic $\tilde{\theta} = \mathrm{E}[\hat{\theta} \,|\, T]$ yields an estimator with the same expected value and better or equal variance.

We might be tempted to use Rao-Blackwell twice, to get an even better estimator. That doesn't work — the second round does nothing. That is, if $\hat{\theta}$ is an estimator of θ, $\tilde{\theta} = h(T) = \mathrm{E}[\hat{\theta} \,|\, T]$, then $\tilde{\tilde{\theta}} = \mathrm{E}[h(T), | \, T] = h(T)$.[1]

That doesn't mean there isn't further room for improvement; there might be a better sufficient statistic to use.

6.5.2 Minimal Sufficient Statistics

While a sufficient statistic contains all the relevant information about a parameter, it might also contain irrelevant information. For example, for data from a normal distribution with unknown mean and variance $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, then the whole sample $(X_1, \ldots, X_n)$, the order statistics $(X_{(1)}, \ldots, X_{(n)})$, and the sample mean+variance $(\bar{X}, S^2)$ are all (vector-valued) sufficient statistics, but we prefer the latter — it contains all the relevant information, without containing irrelevant information. It is a *minimal sufficient statistic*.

Definition 2. A (vector-valued) statistic is *minimal sufficient* if and only if it is a function of every other set of sufficient statistics.

For example, $(\bar{X}, S^2)$ is a function of $(X_1, \ldots, X_n)$ but the converse does not hold.

If we start with an estimator, apply Rao-Blackwell once using a statistic that is sufficient but not minimal sufficient, then we could apply Rao-Blackwell a second time using a minimal sufficient statistic for further variance reduction.

There is no universal rule for finding a minimal sufficient statistic, or even proving that a sufficient statistic is minimal. But if a problem has k parameters we find a sufficient statistic of dimension k then it is typically minimal.

An estimator that is based on (is a function of) a minimal sufficient statistic has minimum variance, and if unbiased is a *minimum variance unbiased estimator* (MVUE). (Caveat — we're not being rigorous here.)

[1] Chapter 7 of *A First Course in Probability* by Ross.

In practice, estimators with small bias may be preferred to unbiased estimators. But the idea carries over to biased estimators — estimators that are functions of minimal sufficient statistics have lower variance than other estimators with the same bias that are not, and we can improve estimators by Rao-Blackwellization using a minimal sufficient statistic.

6.5.3 Exercises

1. Let $X_i \overset{i.i.d.}{\sim} \text{Bern}(p)$, $i = 1, 2, \ldots, n$. Let $T = \sum_{i=1}^{n} X_i$. Show that T is a sufficient statistic for p.

2. Let $X_i \overset{i.i.d.}{\sim} \text{Unif}[0, \beta]$, $i = 1, 2, \ldots, n$. Show that $X_{\max}$ is a sufficient estimator of β.

3. Let $X_1, X_2, \ldots, X_n \sim \text{N}(\mu, \sigma^2)$ where σ is known. Show that $T = \bar{X}$ is sufficient for μ.

4. Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with pdf $f(x; \theta) = \frac{2x}{\theta} e^{-x^2/\theta}$ for $x > 0$ and 0 otherwise. Show that $T = \sum_{i=1}^{n} X_i^2$ is sufficient for θ.

5. If T is a sufficient statistic for a parameter θ and f is a one-to-one function, show that $f(T)$ is also a sufficient statistic.

6. If $X, \ldots, X_n$ are independent and identically distributed Gamma(α, β), show that $(\sum_{i=1}^{n} X_i, \prod_{i=1}^{n} X_i)$ is a sufficient statistic.