4. Sufficiency

4.1. **Sufficient statistics.**

Definition 4.1. A statistic $T = T(X_1, \ldots, X_n)$ is called *sufficient* if the conditional probabilities

$$P_\theta(X_1 = x_1, \ldots, X_n = x_n | T = t) \tag{4.1}$$

are independent of θ for all $x_1, \ldots, x_n$ and t.

This assumes that the distributions are discrete. I'll give the continuous version of this definition in a moment, but let us first try to understand it intuitively. Let's take one more look at the coin flip example, so $P(X_1 = 1) = \theta$, $P(X_1 = 0) = 1 - \theta$, and let's take $T = n\overline{X} = X_1 + \ldots + X_n$.

I claim that T is sufficient. To check this, recall that $T \sim B(n, \theta)$, so

$$P(T = t) = \binom{n}{t} \theta^t (1-\theta)^{n-t}.$$

We can focus on the case $x_1 + \ldots + x_n = t$; if this fails, then the conditional probability from (4.1) simply equals zero for all θ. With this extra assumption in place, the condition $T = t$ follows from $X_1 = x_1, \ldots, X_n = x_n$, so the probability of both events occurring equals

$$\begin{aligned} P(X_1 = x_1, \ldots, X_n = x_n) &= \theta^{x_1 + \ldots + x_n} (1-\theta)^{n - x_1 - \ldots - x_n} \\ &= \theta^t (1-\theta)^{n-t}. \end{aligned} \tag{4.2}$$

Putting things together, we see that

$$P(X_1 = x_1, \ldots, X_n = x_n | T = t) = \frac{P(X_1 = x_1, \ldots, X_n = x_n)}{P(T = t)} = \frac{1}{\binom{n}{t}},$$

which is independent of θ, as claimed. (This final result could have been written down right away with no calculation: each of the $\binom{n}{t}$ sequences $x_1, \ldots, x_n$ that contain exactly t ones is equally likely, given that $T = t$.)

Now let's try to interpret this. One possible view of (4.1) is to say that once you have been informed about the value of T, further details about the values that your random sample took are of no use if you want to estimate θ. This is so because the conditional probabilities of such events, given the value of T, are independent of θ, and thus the occurrence of such an event does not support any kind of inference on θ, on top of what T already told you. In this sense, just knowing T is *sufficient.* In the example we just did, this claim is intuitively obvious: say you flip a coin ten times. If you are now told how many times heads occurred (= the sufficient statistic), you know everything

about the random sample that is worth knowing; it would not be of any additional help if you were now also told in which order exactly those outcomes occurred. (I should also point out that we are not trying to find the best formalization of certain intuitive ideas that we associate with the word *sufficient;* rather, sufficiency in the technical sense is an important notion because of various mathematical properties that are enjoyed by sufficient statistics.)

Some of our observations from the above calculation are valid in general. Clearly, the probability from (4.1) equals zero unless $t = T(x_1, \ldots, x_n)$. Moreover, in this case, we have that

$$\begin{aligned} P(X_1 = x_1, \ldots, X_n = x_n | T = t) &= \frac{P(X_1 = x_1, \ldots, X_n = x_n)}{P(T = t)} \\ &= \frac{L(x_1, \ldots, x_n)}{P(T = t)}, \end{aligned} \tag{4.3}$$

where we again used the notation L for the likelihood function, as introduced in Chapter 3. This suggests the following version of Definition 4.1 for continuous distributions: we call a statistic T *sufficient* if

$$\frac{L(x_1, \ldots, x_n; \theta)}{f_T(t; \theta)}, \qquad t = T(x_1, \ldots, x_n) \tag{4.4}$$

is independent of θ. In this setting, the likelihood function is the product of the individual densities: $L = f(x_1) \cdots f(x_n)$

Example 4.1. Consider again the uniform distribution $f = 1/\theta$ on $0 \leq x \leq \theta$. Is $T = \max(X_1, \ldots, X_n)$ a sufficient statistic? Or how about $\overline{X}$? (Perhaps think for a moment about what the answers should be before you read on.)

Recall from Example 3.9 that T has density $f(t) = nt^{n-1}/\theta^n$. Moreover, $L = \theta^{-n}$ if $t = \max x_j \leq \theta$, and $L = 0$ otherwise. So the quotient from (4.4) equals $1/(nt^{n-1})$, which is independent of θ, and thus T is sufficient.

As for $\overline{X}$, I actually don't want to discuss this in general as the formulae get out of hand quickly. Instead, I want to focus on $n = 2$ and the statistic $U = X_1 + X_2$. As usual, we find its distribution by convolving the density $f = (1/\theta)\chi_{(0,\theta)}$ of X_j with itself. For $0 \leq t \leq \theta$, the product $f(s)f(t-s)$ is non-zero (and thus $= 1/\theta^2$) for $0 \leq s \leq t$, and thus $f_U(t) = t/\theta^2$ for these t. The case $\theta \leq t \leq 2\theta$ is similar, and we find that

$$f_U(t) = \begin{cases} t/\theta^2 & 0 < t < \theta \\ (2\theta - t)/\theta^2 & \theta < t < 2\theta \end{cases}.$$

So if $\theta < x_1 + x_2 < 2\theta$, $x_1, x_2 < \theta$, and $t = x_1 + x_2$, then we obtain that the quotient from (4.4) equals $1/(2\theta - t)$. This depends on θ, so U is not sufficient.

Again, everything makes perfect sense on an intuitive level also: if you are told what the largest observed datum is equal to, then, given this, any set of values of the random sample consistent with this information is equally likely, and extra information on θ isn't going to change anything. On the other hand, if I tell you that $\overline{X} = 1$, say, then θ could be 1, or it could be that $\theta = 100$, or anything else ≥ 1, and this dramatically affects how likely given configurations are.

Exercise 4.1. Now consider the discrete analog, the urn with an unknown number $N \geq 1$ of balls in it. More formally, consider the distribution $P(X_1 = j) = 1/N$ for $j = 1, 2, \ldots, N$. Show that $T = \max X_j$ is again sufficient.

Exercise 4.2. Consider a random sample of size $n = 2$ for the coin flip distribution $P(X_1 = x) = \theta^x(1-\theta)^{1-x}$, $x = 0, 1$.
(a) It seems intuitively clear that $T = X_1$ is not sufficient (the loss of information suffered by dropping X_2 should affect our ability to draw inferences on θ). Can you show this more formally?
(b) Show that $T = X_1 + 2X_2$ is sufficient.
(c) However, show that $T = X_1 + X_2 + 2X_3$ is not sufficient for a random sample of size $n = 3$.

Sufficiency can often be more conveniently checked with the following criterion:

Theorem 4.2 (Neyman)**.** *$T = T(X_1, \ldots, X_n)$ is a sufficient statistic if and only if there are functions $k_1, k_2 \geq 0$ such that*

$$L(x_1, \ldots, x_n; \theta) = k_1(T(x_1, \ldots, x_n); \theta)k_2(x_1, \ldots, x_n). \tag{4.5}$$

In other words, the θ dependence can be isolated in a factor that depends on the values of the random sample only through the statistic T.

Let's quickly revisit our examples from above. In the coin flip example, we see from (4.2) that L indeed has such a factorization, with $k_1 = L = \theta^t(1-\theta)^{n-t}$ (writing $t = T(x_1, \ldots, x_n) = x_1 + \ldots x_n$, as before) and $k_2 = 1$.

For a random sample drawn from a uniform distribution, as in Example 4.1, we have that $L = \chi_{(t,\infty)}(\theta)\theta^{-n}$, and again, this can be our k_1, and $k_2 = 1$.

Proof of Theorem 4.2. I'll discuss the case of a discrete distribution; the continuous case is similar. If T is sufficient, then, as we saw above,

in (4.3), the quotient

$$k_2(x_1,\dots,x_n) := \frac{L(x_1,\dots,x_n;\theta)}{P(T=T(x_1,\dots,x_n))}$$

is independent of θ, and we obtain the required factorization if we put $k_1 = P(T=t)$, with $t = T(x_1,\dots,x_n)$, as usual.

Conversely, if (4.5) holds, then

$$P(T=t) = \sum_{T(x)=t} L(x) = k_1(t;\theta) \sum_{T(x)=t} k_2(x) = k_1(t;\theta)F(t).$$

Here, the sums are over those $x_1,\dots,x_n$ for which $T(x_1,\dots,x_n)=t$, and I've used the convenient abbreviation $x=(x_1,\dots,x_n)$. Let again $t=T(x)$. Then, referring to (4.3) and (4.5) one more time, we obtain that

$$P(X_1=x_1,\dots,X_n=x_n|T=t) = \frac{L(x;\theta)}{P(T=t)} = \frac{k_1(t;\theta)k_2(x)}{k_1(t;\theta)F(t)} = \frac{k_2(x)}{F(t)},$$

and this is independent of θ, as required. □

With the help of Theorem 4.2, we can often extract sufficient statistics from the form of the likelihood function. Let's take a look at some examples.

Example 4.2. The Poisson distribution $P(X=x) = (\theta^x/x!)e^{-\theta}$ has the likelihood function

$$L = \frac{\theta^{x_1+\dots+x_n}}{x_1!\cdots x_n!}e^{-n\theta}.$$

From this, we can now read off that $T = X_1+\dots+X_n$ is a sufficient statistic: indeed, this gives a factorization as in (4.5), with $k_1 = \theta^t e^{-n\theta}$ and $k_2 = 1/(x_1!\cdots x_n!)$.

Example 4.3. Recall that the $\chi^2(\theta)$ distribution has the density $f(x) = c_\theta x^{\theta/2-1}e^{-x/2}$ $(x\geq 0)$, for $\theta = 1,2,\dots$. Thus the likelihood function equals

$$L = c_\theta^n (x_1\cdots x_n)^{\theta/2-1} e^{-(1/2)(x_1+\dots+x_n)},$$

and now Neyman's Theorem shows that $T = X_1X_2\cdots X_n$ is a sufficient statistic.

Exercise 4.3. Consider the $N(0,\sigma)$ distribution, and let $\theta=\sigma^2$. Find a sufficient statistic.

Exercise 4.4. Consider the $N(\theta,1)$ distribution. Find a sufficient statistic.

Exercise 4.5. Consider the exponential distribution $f(x;\theta) = \theta e^{-\theta x}$ $(x>0)$. Find a sufficient statistic.

Exercise 4.6. Suppose the parameters θ, η and the statistics S, T are related by $\theta = g(\eta)$ and $S = h(T)$, respectively. Suppose that S is sufficient for θ. Show that then T is sufficient for η. (So, roughly speaking, sufficiency is not affected by taking functions of the parameter and/or the statistic.)
Suggestion: Use the criterion from Neyman's Theorem.

4.2. **Conditional expectation.** Our first central result on sufficient statistics will depend on the notion of *conditional expectation,* so we'll discuss this first. Let X, Y be random variables. We want to define $E(X|Y)$, the conditional expectation of X, given Y. This will be a random variable (!), and we are trying to formalize the idea of a partial averaging process over those parts of the sample space on which Y is constant. Alternatively, you can think of $E(X|Y)$ at a point $\omega \in \Omega$ as your best guess on X, given that $Y(\omega) = y$.

An example will be useful to make these ideas more concrete. Let's roll a die, so $\Omega = \{1, 2, \dots, 6\}$, and we (of course) use Laplace probability on this sample space. Let $X(\omega) = \omega$ and

$$Y(\omega) = \begin{cases} 1 & \omega = 2, 4, 6 \\ -1 & \omega = 1, 3, 5 \end{cases}.$$

So X just records the outcome, and Y tells us whether this number is even or odd. If we are now told that $Y = 1$ occurred, then the average value of X consistent with this information is $(1/3)(2 + 4 + 6) = 4$. Similarly, if $Y = -1$, then the new average value of X, given this information, equals $(1/3)(1+3+5) = 3$. Now notice that these partial averages can be viewed as weighted averages, taken with respect to the conditional probabilities:

$$4 = \sum_{k=1}^{6} kP(X = k|Y = 1), \qquad 3 = \sum_{k=1}^{6} kP(X = k|Y = -1)$$

since $P(X = k|Y = 1) = 1/3$ if k is even and $= 0$ if k is odd, and similarly for $P(X = k|Y = -1)$. This can be our definition of conditional expectation in the discrete case:

$$E(X|Y)(\omega) = \sum_{k} x_k P(X = x_k|Y = y), \qquad y = Y(\omega) \tag{4.6}$$

Exercise 4.7. Show that conditional expectation is linear (just like the plain expectation), that is, $E(aX + bY|Z) = aE(X|Z) + bE(Y|Z)$.

In the following exercise, we rewrite (4.6) in a way that makes the *partial averaging* view of the conditional expectation even more explicit:

Exercise 4.8. Partition $\Omega = \bigcup A_j$, where A_j is the event that $Y = y_j$. Show that then

$$E(X|Y)(\omega) = \frac{\sum_{\omega' \in A_j} X(\omega')P(\{\omega'\})}{P(A_j)}$$

for $\omega \in A_j$.

Proposition 4.3. *$E(X|Y) = f(Y)$ for some function f, and*

$$E(E(X|Y)) = EX. \tag{4.7}$$

It is in fact also possible to characterize conditional expectation by a similar (but somewhat more general) pair of properties, and this is the path usually taken in more advanced (= based on measure theory) treatments. For our purposes, the more direct definition given above is more convenient and accessible.

Proof. The first property is an immediate consequence of the definition (4.6) because this in particular says that $E(X|Y)(\omega)$ really only depends on the value of $Y(\omega)$, not on ω itself, so f may be obtained by sending $y \in \mathbb{R}$ with $Y(\omega) = y$ to $f(y) = E(X|Y)(\omega)$.

As for the second property, we compute

$$\begin{aligned} E(E(X|Y)) &= \sum_{\omega\in\Omega} E(X|Y)(\omega)P(\{\omega\}) \\ &= \sum_{\omega\in\Omega}\sum_k x_k P(X = x_k|Y = Y(\omega))P(\{\omega\}) \\ &= \sum_k x_k \sum_j P(Y = y_j)P(X = x_k|Y = y_j) \\ &= \sum_k x_k P(X = x_k) = EX. \end{aligned}$$

Here, we passed to the third line by partitioning Ω into the events $Y = y_j$, for all possible values of Y. We then recognize the resulting sum over j as the total probability formula for $P(X = x_k)$ and so arrive at the formula from the fourth line. □

Exercise 4.9. Show that if Y is a constant random variable, say $Y = 0$ with probability 1, then $E(X|Y) = EX$.

Theorem 4.4. *If X, Y are independent, then $E(X|Y) = EX$. If $X = f(Y)$ is a function of Y, then $E(X|Y) = X$. More generally, $E(Xf(Y)|Y) = f(Y)E(X|Y)$.*

These properties are immediately plausible: if X, Y are independent, then knowledge about Y is useless for making predictions on X, so all

we can do is take the normal expectation. Note also that the first claim really says that the random variable $E(X|Y)$ is constant and this constant equals the number EX. Similarly, a function of Y can be predicted with certainty if Y is given, so no averaging is necessary here.

Proof. If X, Y are independent, then $P(X = x_k|Y = y) = P(X = x_k)$, so the first claim follows from (4.6). Similarly, if $X = f(Y)$, then $P(X = x_k|Y = y) = 1$ if $x_k = f(y)$ and $P(X = x_k|Y = y) = 0$ otherwise, and $P(Xf(Y) = x_k|Y = y) = P(X = x_k/f(y)|Y = y)$ if $f(y) \neq 0$. If $f(y) = 0$, then $P(Xf(Y) = 0|Y = y) = 1$. These observations combined with (4.6) give the last two statements. □

Exercise 4.10. Work out this last part of the argument in more detail.

The following property will be especially important for us. Again, everything makes immediate intuitive sense. The theorem says that a partial averaging process will not increase the variance.

Theorem 4.5. $\mathrm{Var}(E(X|Y)) \leq \mathrm{Var}(X)$.

Proof.

$$\begin{aligned} \mathrm{Var}(X) &= E(X - E(X|Y) + E(X|Y) - EX)^2 \\ &= E(X - E(X|Y))^2 + E(E(X|Y) - EX)^2 \\ &\quad + 2E(X - E(X|Y))(E(X|Y) - EX) \end{aligned} \tag{4.8}$$

I now claim that this last expectation equals zero. To see this, we condition on Y. Recall from Proposition 4.1 that $E(\ldots|Y)$ is a function of Y. Thus, by Theorem 4.4,

$$\begin{aligned} &E\left[(X - E(X|Y))(E(X|Y) - EX)|Y\right] \\ &\qquad = (E(X|Y) - EX)E(X - E(X|Y)|Y) \\ &\qquad = (E(X|Y) - EX)(E(X|Y) - E(X|Y)) = 0, \end{aligned}$$

and since (4.7) implies that $EU = 0$ as soon as $E(U|V) = 0$ for some V, my claim from above follows.

Moreover, by the same identity, $E(E(X|Y)) = EX$, so the second term from (4.8) equals $\mathrm{Var}(E(X|Y))$ and since the third term has just been identified as zero, the claim of the theorem now follows. □

Now let's discuss the continuous case. We introduce the *conditional density*

$$f_{X|Y}(x, y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}; \tag{4.9}$$

here $f_{X,Y}$ denotes the joint density of X, Y. Then, in complete analogy to (4.6), we define $E(X|Y)$ as the random variable that is given by

$$E(X|Y)(\omega) = \int x f_{X|Y}(x, y)\, dx, \quad y = Y(\omega). \tag{4.10}$$

All the general properties of $E(X|Y)$ that were discussed above also hold in the continuous setting.

Example 4.4. Let X, Y be two random variables with joint density

$$f(x, y) = \begin{cases} 6y & 0 < y < x < 1 \\ 0 & \text{otherwise} \end{cases}. \tag{4.11}$$

Let's confirm that this is indeed a density:

$$\int_0^1 dy \int_y^1 dx\, 6y = 6 \int_0^1 y(1-y)\, dy = 6\left(\frac{1}{2} - \frac{1}{3}\right) = 1,$$

as required.

Next, let's try to find $Z = E(Y|X)$. For this, we will need the density of X, which we obtain as one of the *marginal densities* of f:

$$f_X(x) = \int f(x, y)\, dy = 6 \int_0^x y\, dy = 3x^2, \qquad 0 < x < 1,$$

and $f_X(x) = 0$ otherwise.

Exercise 4.11. Show more explicitly that this method of finding the individual densities is correct. *Suggestion:* Express $P(X \le x)$ in terms of the joint density and then take the derivative to find the density; it might be a good idea to do this in a general setting, not for the concrete density from above.

Now by (4.10), when $X = x$, then

$$Z = E(Y|X) = \int_0^x y\, \frac{6y}{3x^2}\, dy = \frac{2x}{3}, \qquad 0 < x < 1.$$

More succinctly (and precisely), $Z = 2X/3$.

Finally, let's take a look at $\mathrm{Var}(Y)$ and $\mathrm{Var}(Z)$. Let's start out with this latter random variable, since we already have the density of X. We compute

$$EX = \int_0^1 x \cdot 3x^2\, dx = \frac{3}{4}, \qquad EX^2 = 3 \int_0^1 x^4\, dx = \frac{3}{5},$$

so $\mathrm{Var}(X) = (3/5) - (3/4)^2 = 3/80$ and thus $\mathrm{Var}(Z) = 1/60$.

To find the density of Y, we go back to (4.11):

$$f_Y(y) = \int_y^1 6y\,dx = 6y(1-y), \qquad 0 < y < 1$$

This gives

$$EY = 6\int_0^1 y^2(1-y)\,dy = \frac{1}{2}, \qquad EY^2 = 6\int_0^1 y^3(1-y)\,dy = \frac{3}{10},$$

thus $\mathrm{Var}(Y) = 1/20$. Notice that this is larger than $\mathrm{Var}(E(Y|X))$, as predicted by Theorem 4.5.

Exercise 4.12. Let X, Y be random variables with joint density $f(x,y) = x+y$ for $0 < x, y < 1$ and $f(x,y) = 0$ otherwise. Proceed as in Example 4.4 to find $E(Y|X)$. Then again compare $\mathrm{Var}(Y)$ with $\mathrm{Var}(E(Y|X))$.

4.3. **The Rao-Blackwell Theorem.** Recall that among all unbiased estimators, we prefer those with small variance about the correct expected value. Now Theorem 4.5 may be interpreted as saying that the variance can be improved by conditioning; moreover, (4.7) guarantees that this will not introduce a bias. These simple, but extremely useful observations already give us a good part of the somewhat grandiosely named *Rao-Blackwell Theorem* (in this order, for some reason, at least if we go along with the 24,100 hits for "R-B" on a popular search engine as opposed to just 1,500 for "B-R").

Theorem 4.6 (Rao-Blackwell)**.** *Let T_1 be an unbiased estimator, and let Y be a sufficient statistic. Then $T_2 = E(T_1|Y) = \varphi(Y)$ defines a new unbiased estimator with* $\mathrm{Var}(T_2) \le \mathrm{Var}(T_1)$.

In short: unbiased estimators can be improved by conditioning on sufficient statistics. Also, given an arbitrary unbiased estimator, there is a function of the sufficient statistic that performs at least as well, perhaps better.

Exercise 4.13. Why do we want Y to be a sufficient statistic? Why don't we get the same conclusions for an *arbitrary* statistic Y, by the argument based on Theorem 4.5 and (4.7) that we just outlined? *Warning:* The answer to this is simple, but the notation we are currently using somewhat deceptively hides the crucial point.

Proof. As already observed, the inequality on the variances is Theorem 4.5, applied to the case at hand. Also, (4.7) gives that $ET_2 = E(E(T_1|Y)) = ET_1 = \theta$, as required. Finally, T_2 is indeed a statistic because (in the discrete case) the sufficiency of Y implies that the conditional probabilities $P(T_1 = t|Y = y)$ are independent of θ and thus

T_2 is independent of θ as well, by (4.6). (Please go back to Exercise 4.13 now if you haven't solved it already.) □

We said in the first section that we can think of a sufficient statistic as containing all the information about the random sample that is worth knowing. If this intuition is correct, then it would seem foolish to let an estimator take distinct values on a part $Y = y$ of the sample space where the sufficient statistic is constant, and indeed the Rao-Blackwell Theorem says that we can improve estimators by ironing out such unnecessary variations.

Definition 4.7. We call a statistic T a *minimum variance unbiased estimator* (MVUE) if T is unbiased and $\mathrm{Var}(T) \leq \mathrm{Var}(U)$ for all unbiased estimators U (and for all θ).

So an efficient statistic is an MVUE, but there could be other situations (which we will actually encounter rather soon) where the CR bound isn't achieved, but (non-efficient) MVUEs still exist.

The Rao-Blackwell Theorem shows that if we can come up with a sufficient statistic Y, then the search for an MVUE can be restricted to functions $T = \varphi(Y)$ of Y. This is so because if U is an MVUE, then so is $T = E(U|Y)$, by the Rao-Blackwell Theorem, and this statistic is a function of Y, by the same result (and Proposition 4.3).

Example 4.5. Back to the coin flip example. In fact, to keep things simple, I now only want to consider random samples of size $n = 2$. We know that $Y = X_1 + X_2$ is sufficient. (We also know that $T = \overline{X} = Y/2$ is efficient, and thus the MVUE, but let's ignore this for now.)

Let's consider the estimator

$$T = \frac{1}{4}X_1 + \frac{3}{4}X_2.$$

This is unbiased, but

$$\mathrm{Var}(T) = \left(\left(\frac{1}{4}\right)^2 + \left(\frac{3}{4}\right)^2\right)\mathrm{Var}(X_1) = \frac{5}{8}\,\mathrm{Var}(X_1),$$

which is larger than the optimal value $(1/2)\mathrm{Var}(X_1)$. By Theorem 4.6, T can be improved by conditioning on Y, so what is $U = E(T|Y)$ equal to? If $Y = 0$, then $X_1 = X_2 = 0$, so $T = 0$, and no averaging is necessary in this case, and $U = 0$ as well. Similarly, $Y = 2$ implies that $X_1 = X_2 = 1$, and thus $U = 1$ in this case. If $Y = 1$, then $X_1 = 0, X_2 = 1$ or the other way around, and the two conditional probabilities satisfy

$$P(X_1 = 0, X_2 = 1|Y = 1) = P(X_1 = 1, X_2 = 0|Y = 1) = 1/2.$$

Thus $U = 1/2 \cdot 3/4 + 1/2 \cdot 1/4 = 1/2$.

This can be summed up by noticing that $U = Y/2 = \overline{X}$; the process of conditioning T on the sufficient statistic Y recovers the MVUE.

Exercise 4.14. Show by a similar calculation that $E(T|Y) = \overline{X}$ also if $T = X_1$, and $Y = X_1 + X_2$ is as above. In fact, can you also show that $E(X_1|Y) = \overline{X}$ for a random sample of arbitrary size (and $Y = X_1 + \ldots + X_n$)?

It is no coincidence that we are invariably led back to $\overline{X}$ in these examples. Later we will see that this is the only unbiased function of Y.

Example 4.6. Let's take another look at the urn with an unknown number of balls in it: $P(X_1 = x) = 1/N$ for $x = 1, 2, \ldots, N$ and $P(X_1 = x) = 0$ otherwise, and $N = 1, 2, \ldots$ takes the role of the parameter θ. You showed in Exercise 4.1 that $Y = \max X_j$ is a sufficient statistic; recall also that $T = 2\overline{X} - 1$ is unbiased (which seemed pretty much the only thing that could be said in favor of this estimator). Backed up by the Rao-Blackwell Theorem, let's now try to improve T by conditioning on Y. As a warm-up, let's first do this for a random sample of size $n = 2$.

If $Y = 1$, then $X_1 = X_2 = 1$, so $T = 1$ as well and thus $U = E(T|Y) = 1$ on this part of the sample space. If $Y = 2$, then the three outcomes $(1, 2)$, $(2, 1)$, $(2, 2)$ for (X_1, X_2) are consistent with this value of Y, and $P(X_1 = x_1, X_2 = x_2|Y = y) = 1/3$ for each of these.

Exercise 4.15. Show this in more detail. In fact, can you right away show the general claim that $P(X_1 = x_1, \ldots, X_n = x_n|Y = y)$ has a constant value for all outcomes $x_1, \ldots, x_n$ with $Y(x_1, \ldots, x_n) = y$?

It follows that $U = (1/3)(2 + 2 + 3) = 7/3$ if $Y = 2$.

We can continue in this style to work out U on the set where $Y = 3, 4, \ldots$, but we should in fact be ready for the general case now. So let's consider a random sample of size n, and suppose that $Y = y$. Then the possible values of the random sample are those n tuples x_j with $1 \leq x_j \leq y$ for $j = 1, 2, \ldots, n$ and $x_k = y$ for at least one index k. By Exercise 4.15, all these outcomes have the same conditional probability, so we can find probabilities by counting.

More precisely, we have that (when $Y = y$)

$$E(T|Y) = \frac{1}{M(y)} \sum_{x:Y(x)=y} T(x), \tag{4.12}$$

where we again abbreviated $x = (x_1, \ldots, x_n)$ and $M(y)$ denotes the number of x with $Y(x) = y$. Since this is all sequences x drawn from

$1, 2, \ldots, y$ except those that avoid y altogether, we have that

$$M(y) = y^n - (y-1)^n.$$

Next, let's take a look at the sum from (4.12). Recall that $T = 2\overline{X} - 1$ and focus on $\overline{X}$. Consider for the moment only those x that take the value y exactly k times, for a fixed $k = 1, \ldots, n$. There are $\binom{n}{k}$ choices for the slots where those y's occur. If we fix such a choice, say $x_1 = \ldots = x_k = y$ (for notational convenience), then

$$\overline{X}(x) = \frac{1}{n}\left(ky + x_{k+1} + \ldots + x_n\right).$$

Now take the sum over $x_j = 1, 2, \ldots, y-1$ for $j > k$. Since

$$\sum_{x=1}^{y-1} x = \frac{(y-1)y}{2},$$

we obtain that

$$\sum_{x_{k+1},\ldots,x_n=1}^{y-1} \overline{X}(x) = \frac{1}{n}(y-1)^{n-k}\left(ky + (n-k)\,\frac{y}{2}\right),$$

and there are $\binom{n}{k}$ such (partial) sums for fixed k, hence

$$\sum_{Y(x)=y} T(x) = \frac{1}{n}\sum_{k=1}^{n}\binom{n}{k}(y-1)^{n-k}(n+k)y - M(y).$$

We can simplify this sum. Notice first of all that

$$\sum_{k=1}^{n}\binom{n}{k}1^k(y-1)^{n-k} = y^n - (y-1)^n = M(y).$$

Next,

$$\binom{n}{k}k = \frac{n!}{(k-1)!(n-1-(k-1))!} = n\binom{n-1}{k-1},$$

so

$$\sum_{k=1}^{n}\binom{n}{k}(y-1)^{n-k}k = n\sum_{k=0}^{n-1}\binom{n-1}{k}(y-1)^{n-1-k} = ny^{n-1}.$$

Putting things together, we find that

$$\sum_{Y(x)=y} T(x) = yM(y) + y^n - M(y)$$

and thus, by (4.12),

$$E(T|Y) = Y - 1 + \frac{Y^n}{Y^n - (Y-1)^n} = Y + \frac{(Y-1)^n}{Y^n - (Y-1)^n}.$$

The last fraction can be viewed as a small correction, which makes our estimator unbiased, and essentially this is our favorite estimator $Y = \max X_j$.

Example 4.7. This worked reasonably well, so perhaps we can now also do the continuous version of this example. We consider the uniform distribution on $(0, \theta)$, that is, $f(x) = 1/\theta$ for $0 < x < \theta$. We know that the analogous statistic $Y = \max X_j$ is again sufficient, and $T = 2\overline{X}$ is an unbiased estimator. Let's try to work out $U = E(T|Y)$. I'll do this somewhat informally, which will simplify the calculation tremendously. (A rather different, more rigorous approach will be discussed later, when we have additional theory available.)

If $Y = y$, then $X_j = y$ for at least one j, and in fact I will ignore the possibility of having two or more such indices j. This is justified (we hope) because we are dealing with continuous random variables now, which take a given precise value with probability zero. Then, if we fix a j and set $X_j = y$, the remaining X_k are uniformly distributed over $(0, y)$, and thus the partial average of $S = X_1 + \ldots X_n$ becomes $y + (n-1)y/2$. Since this is independent of j, it follows that $E(S|Y) = Y + (n-1)Y/2$ as well (as in the previous example, we build up the partial average defining the conditional expectation as an average of partial averages over smaller parts). Thus

$$E(T|Y) = \frac{2}{n}\, E(S|Y) = \frac{n+1}{n}\, Y,$$

which is reassuring because this is exactly the high performance unbiased estimator that we found earlier, in Chapter 3.

Example 4.8. Consider the $N(0, \sigma)$ distribution; we are trying to estimate the variance $\theta = \sigma^2$. You showed in Exercise 4.3 that $Y = \sum X_j^2$ is a sufficient statistic, and we also know that the sample variance $S^2 = (1/(n-1)) \sum (X_j - \overline{X})^2$ is an unbiased estimator for θ. The Rao-Blackwell Theorem suggests to improve matters by using the statistic $E(S^2|Y)$ instead. By a calculation that we already did a few times on other occasions, we can write

$$(n-1)S^2 = \sum_{j=1}^{n} X_j^2 - n\overline{X}^2 = Y - n\overline{X}^2.$$

Now $E(Y|Y) = Y$, so the first term poses no difficulties. As for the second term, we multiply out and obtain

$$n^2 E(\overline{X}^2|Y) = E\left(\sum_{j=1}^{n} X_j^2 \big| Y\right) + \sum_{j \neq k} E(X_j X_k|Y) = Y + \sum_{j \neq k} E(X_j X_k|Y).$$

This last conditional expectation equals zero because Y is insensitive to signs, so the conditional distribution of X_jX_k, given Y, is still symmetric about zero. Somewhat more formally, we can observe that if we replace X_j by $-X_j$, then all (conditional and joint) distributions involved here remain unchanged, so $E(X_jX_k|Y) = -E(X_jX_k|Y)$. We conclude that

$$E(S^2|Y) = \frac{1}{n-1}\left(Y - \frac{Y}{n}\right) = \frac{Y}{n}.$$

The Rao-Blackwell Theorem tells us that Y/n performs at least as well as S^2, and in fact we know from our earlier discussion in Example 3.11 that Y/n is efficient while S^2 isn't.

Example 4.9. Let's return one more time to the uniform distribution $f(x) = 1/\theta$, $0 < x < \theta$, but this time I'll proceed in a formally fully correct way. Consider a random sample of size $n = 3$, and let $M = M(X_1, X_2, X_3)$ be its *median:* we order the random sample $X_{j_1} < X_{j_2} < X_{j_3}$, and M is defined as the middle value X_{j_2}. To find the distribution of M, observe that

$$P(M \le x) = 3\,\frac{x^2(\theta - x)}{\theta^3} + \frac{x^3}{\theta^3};$$

indeed, the median will be $\le x$ if either all three data are $\le x$ (the last term) or exactly two are $\le x$ (the first term on the right-hand side). By differentiating, we find that M has density

$$f(x) = 6\,\frac{x(\theta - x)}{\theta^3}.$$

In particular, this implies that

$$EM = \frac{6}{\theta^3}\int_0^\theta x^2(\theta - x)\,dx = \frac{\theta}{2},$$

as could have been predicted from symmetry considerations. So $2M$ is an unbiased estimator.

As usual, the Rao-Blackwell Theorem suggests to condition on $Y = \max(X_1, X_2, X_3)$, a temptation which I will give in to. Let's first think about the joint distribution of M, Y. We have that

$$P(M \le x, Y \le y) = 3\,\frac{x^2(y - x)}{\theta^3} + \frac{x^3}{\theta^3}, \qquad 0 < x \le y < \theta,$$

by essentially the same argument as above: either all three data points are $\le x$, or exactly two of them satisfy this condition and the third one lies between x and y. Of course, if $x > y$, then the conditions become

contradictory and the probability of the (now empty) event equals zero. From this, we find the joint density as $f_{M,Y} = \partial^2 P/\partial x \partial y$:

$$f_{M,Y}(x,y) = \frac{6x}{\theta^3}, \qquad 0 < x \le y < \theta. \tag{4.13}$$

Exercise 4.16. Check our work by (re-)deriving the densities of M and Y from (4.13).

Recall that (or derive it again) $f_Y(y) = 3y^2/\theta^3$, and thus by (4.9) $f_{M|Y}(x,y) = 2x/y^2$ if $0 < x \le y < \theta$.

Exercise 4.17. Observe that θ has dropped out of this formula. Are you surprised at this? (Don't just answer *yes* or *no;* find reasons for your surprise or lack thereof.)

It follows that when $Y = y$, then

$$E(M|Y) = \int_0^y x\,\frac{2x}{y^2}\,dx = \frac{2y}{3}.$$

In other words, $E(M|Y) = 2Y/3$.

Exercise 4.18. In the absence of any extra information, the X_j are restricted to $[0,\theta]$, and thus by symmetry $EM = \theta/2$, as we saw above. If $Y = y$, then the X_j are now restricted to $[0,y]$, so reasoning by analogy, we expect that $E(M|Y) = Y/2$, right? What is your answer to this?

We wanted to improve the unbiased estimator $2M$, and we have now obtained $E(2M|Y) = 4Y/3$. This is satisfying; just as in Example 4.7, we recover (in the special case $n = 3$) the estimator $(n+1)Y/n$ one more time.

Exercise 4.19. However, is $T = 4Y/3$ really better than $2M$ (the Rao-Blackwell Theorem leaves open the possibility that both estimators could have the same variance)? To answer this, recall from our discussion in Chapter 3 that $\mathrm{Var}(T) = \theta^2/(3 \cdot 5) = \theta^2/15$; see especially (3.11). Now compute $\mathrm{Var}(2M)$ and compare.

Exercise 4.20. Consider the density $f(x) = \theta^{-1}e^{-x/\theta}$ $(x > 0)$, and draw a random sample of size $n = 2$.
(a) Show again that $Y = X_1 + X_2$ is sufficient (this also follows by combining Exercises 4.5, 4.6).
(b) Find the joint density $f_{X_1,Y}(x,y)$ and then the conditional density $f_{X_1|Y}(x,y)$; for this, you will need $f_Y(y)$, which we discussed in Chapter 2, see pg. 25.
(c) Use your answer to part (b) to find $E(X_1|Y)$.

Another general observation about sufficient statistics worth making is the following:

Theorem 4.8. *Let Y be a sufficient statistic. If a unique MLE $\widehat{\theta}$ exists, then $\widehat{\theta} = \varphi(Y)$.*

We said earlier that MVUEs, if they exist, are functions of a given sufficient statistic, so this gives some additional confidence that MLEs are reasonable. Of course, Theorem 4.8 just says that $\widehat{\theta}$ is *some* function of Y, and it may not be the right one. In fact, we know that MLEs can easily fail to be unbiased.

Proof. This is an immediate consequence of Neyman's theorem, which says that the likelihood function may be factorized as follows:

$$L(x_1, \ldots, x_n; \theta) = k_1(Y(x_1, \ldots, x_n); \theta) k_2(x_1, \ldots, x_n)$$

The MLE is defined by the condition that $\widehat{\theta}(x_1, \ldots, x_n)$ maximizes L for fixed values $x_1, \ldots, x_n$, but since k_2 is independent of θ, this is the same as maximizing k_1, and thus the solution to this maximization problem only depends on $Y(x_1, \ldots, x_n)$. □

Exercise 4.21. Find at least four different examples (from our earlier discussion of MLEs in Chapter 3) that confirm the claim of Theorem 4.8.