2.3 Sufficient Statistics and the Rao-Blackwell Theorem

Recall that we are estimating a population parameter (called target parameter) θ, by an estimator $\overline{\theta}(X_1, \ldots, X_n)$, based on a sample (i.e. sample variables $X_1, X_2, \ldots, X_n$). Our goal is to find "good" estimators, i.e. unbiased estimators with low variance, ultimately MVUE's. We saw last time that an efficient estimator is also a MVUE.

We are know exploring other means of finding such estimators, by finding "appropriate" statistics (sample functions) to use in our estimation.

Here is the main idea: So far, we have chosen estimators from a sample on the basis of intuition, like the sample mean $\overline{X}$ for the population mean μ (or the sample variance s^2 for the population variance σ^2). We did not use the *actual* sample values themselves, but rather, we summarized (reduced) them into a value. Has this process retained *all* the information about μ (or σ^2) or has some important information been lost? We want to find statistics that, in some way, summarize **all** the information in a sample about a target parameter. These are called *sufficient* statistics.

Definition 2.1. *The statistic $S = S(X_1, \ldots, X_n)$ is called* ***sufficient*** *for (the estimation of) θ, if the conditional probability distribution of $(X_1, \ldots, X_n)$, given the value of the statistic $S = s$, does not depend on the parameter θ, i.e.*

$$f(x_1, \ldots, x_n; \theta \mid S = s) = f(x_1, \ldots, x_n \mid S = s). \tag{2.1}$$

Remark 2.2.

1. Intuitively, this is saying that once the value of the statistic S is known, no other function of $X_1, \ldots, X_n$ (so no other statistic) will shed additional light on the possible values of θ. Then the conditional probability distribution of the data does not depend on the unknown parameter except through the sufficient statistic. In that sense, S contains all the information in the sample about θ.
2. If a statistic S is sufficient for a parameter θ, then the conditional distribution of *any* statistic, given the value of S, does not depend on θ, since any statistic is just a function of the sample variables $X_1, \ldots, X_n$.

$$f(A(x_1, \ldots, x_n; \theta) \mid S = s) = f(A(x_1, \ldots, x_n) | \, S = s).$$

Example 2.3. Let $X_1, \ldots, X_n$ be a random sample drawn from the Poisson $\mathcal{P}(\lambda)$ distribution of parameter $\lambda > 0$, with pdf

$$f(x; \lambda) = \frac{\lambda^x}{x!} e^{-\lambda}, \; x = 0, 1, \ldots.$$

Is $S = \sum_{k=1}^{n} X_k$ a sufficient statistic for the estimation of λ?

Solution. Since each sample variable $X_k \in \mathcal{P}(\lambda)$, their sum S also has a Poisson distribution $\mathcal{P}(n\lambda)$, with pdf

$$S \left(\begin{array}{c} s \\ \dfrac{(n\lambda)^s}{s!} e^{-n\lambda} \end{array} \right)_{s=0,1,\ldots}.$$

Then the conditional probability of $(x_1, \ldots, x_n)$, given $S = s$ is

$$\begin{aligned}
f(x_1, \ldots, x_n; \lambda \mid s) &= P(X_1 = x_1, \ldots, X_n = x_n \mid S = s) \\
&= \frac{P(X_1 = x_1, \ldots, X_n = x_n, S = s)}{P(S = s)} \\
&= \frac{P\left(X_1 = x_1, \ldots, X_{n-1} = x_{n-1}, X_n = s - \sum_{k=1}^{n-1} x_k\right)}{P(S = s)} \\
&\overset{\text{ind}}{=} \frac{P(X_1 = x_1) \ldots P(X_{n-1} = x_{n-1}) P\left(X_n = s - \sum_{k=1}^{n-1} x_k\right)}{P(S = s)} \\
&= \frac{P\left(X_n = s - \sum_{k=1}^{n-1} x_k\right) \prod_{i=1}^{n-1} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda}}{P(S = s)} \\
&= \frac{\dfrac{\lambda^{s - \sum_{k=1}^{n-1} x_k}}{\left(s - \sum_{k=1}^{n-1} x_k\right)!} e^{-\lambda} \dfrac{\lambda^{\sum_{i=1}^{n-1} x_i}}{x_1! \ldots x_{n-1}!} e^{-(n-1)\lambda}}{\dfrac{(n\lambda)^s}{s!} e^{-n\lambda}} \\
&= \frac{s!}{n^s x_1! \ldots x_n!}
\end{aligned}$$

and **does not** depend on λ. So, S is **sufficient** for the estimation of λ.

■

As we have seen, neither Definition 2.1, nor Remark 2.2 is easy to verify for a given statistic. An equivalent, but easier to use result for sufficiency was given by Fisher:

Theorem 2.4 (Fisher's Factorization Criterion)**.**
A statistic $S = S(X_1, \ldots, X_n)$ is sufficient for θ, if and only if the likelihood function

$$L(X_1, \ldots, X_n; \theta) = \prod_{i=1}^{n} f(X_i; \theta)$$

can be factored into two nonnegative functions

$$L(x_1, \ldots, x_n; \theta) \quad = \quad g(x_1, \ldots, x_n) h(s; \theta), \tag{2.2}$$

such that one factor, g, does not depend on θ and the other factor, h, which does depend on θ, depends on $(x_1, \ldots, x_n)$ only through the value of $s = S(x_1, \ldots, x_n)$.

Example 2.5. Let us consider again the previous example, where $X_1, \ldots, X_n$ is a random sample drawn from a Poisson $\mathcal{P}(\lambda), \lambda > 0$ and $S = \sum_{k=1}^{n} X_k$.

Solution. For each $i = \overline{1, n}$,

$$f(x_i; \lambda) = \frac{\lambda^{x_i}}{x_i!} e^{-\lambda}.$$

Then the likelihood function is given by

$$\begin{aligned} L(x_1, \ldots, x_n; \lambda) &= \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} \\ &= \frac{\lambda^{\left(\sum_{i=1}^{n} x_i\right)}}{x_1! \ldots x_n!} \cdot e^{-n\lambda} \\ &= \frac{1}{x_1! \ldots x_{n!}} \cdot \lambda^s e^{-n\lambda} \\ &= g(x_1, \ldots, x_n) h(s; \lambda), \end{aligned}$$

where $g(x_1, \ldots, x_n) = \dfrac{1}{x_1! \ldots x_{n!}}$ and $h(s; \lambda) = \lambda^s e^{-n\lambda}$.

Thus, by Theorem 2.4, S is a sufficient statistic for the estimation of λ.

■

Sufficient statistics play an important role in finding good estimators. We have already mentioned that for an unbiased estimator, we want the standard error to be small. Having an unbiased estimator, sufficient statistics can be used to find an unbiased estimator with a smaller variance.

Theorem 2.6 (Rao-Blackwell)**.**
Let $\hat{\theta}$ be an unbiased estimator and S be a sufficient statistic for θ. Let

$$\overline{\theta} = E(\hat{\theta} \mid S) \tag{2.3}$$

Then
a) $E(\overline{\theta}) = \theta$ *(i.e. $\overline{\theta}$ is also an unbiased estimator for θ) and*
b) $V(\overline{\theta}) \le V(\hat{\theta})$.

So, using sufficient statistics we can improve an unbiased estimator by lowering its variance. It would seem that by repeatedly applying Theorem 2.6 to an unbiased estimator, we get better and better estimators (and, eventually, a MVUE). Unfortunately, that is not so, if *the same* sufficient statistic S is used. Indeed, if $\overline{\theta} = E(\hat{\theta} \mid S) = f(S)$, then by Remark 2.2, we have

$$E(\overline{\theta} \mid S) = E\left(f(S) \mid S\right) = f(S) = \overline{\theta},$$

the same.
Stronger conditions on the statistic S need to be imposed, in order to improve the quality of unbiased estimators.

2.4 Complete Statistics and the Lehmann-Scheffé Theorem

Definition 2.7. *The statistic $S = S(X_1, \dots, X_n)$ is said to be **complete** for the family of probability distributions $f(x;\theta), \theta \in A$, if*

$$E(\varphi(S)) = 0, \forall \theta \in A \Longrightarrow \varphi \overset{a.s.}{=} 0. \tag{2.4}$$

Example 2.8. Let us consider again the statistic $S = \sum_{i=1}^{n} X_i$ for the Poisson $\mathcal{P}(\lambda), \lambda > 0$ distribution. Is it complete?

Solution. So, again, since each sample variable $X_i \in \mathcal{P}(\lambda)$, the sum S also has a Poisson distribution $\mathcal{P}(n\lambda)$, with pdf

$$S \begin{pmatrix} s \\ \dfrac{(n\lambda)^s}{s!} e^{-n\lambda} \end{pmatrix}_{s=0,1,\dots}.$$

Then the random variable $\varphi(S)$ is also discrete with pdf

$$\varphi(S) \left(\begin{array}{c} \varphi(s) \\ \dfrac{(n\lambda)^s}{s!} e^{-n\lambda} \end{array} \right)_{s=0,1,\ldots} .$$

We compute the expected value in (2.4). We have

$$\begin{aligned} E(\varphi(S)) &= \sum_{s=0}^{\infty} \varphi(s) \frac{(n\lambda)^s}{s!} e^{-n\lambda} \\ &= e^{-n\lambda} \sum_{s=0}^{\infty} \frac{\varphi(s) n^s}{s!} \lambda^s. \end{aligned}$$

So, the condition in (2.4) means

$$e^{-n\lambda} \sum_{s=0}^{\infty} \frac{\varphi(s) n^s}{s!} \lambda^s = 0, \ \forall \lambda > 0.$$

The only way that this can happen is if **all** the coefficients in the power series are equal to 0, i.e.

$$\varphi(s) = 0, \ \forall s \in \mathbb{N},$$

which means

$$\varphi \overset{a.s.}{=} 0.$$

Thus, condition (2.4) holds and S is a complete statistic for the family of Poisson $\mathcal{P}(\lambda), \lambda > 0$, distributions.

■

In the next result, we see what happens when we construct the estimator from the Rao-Blackwell Theorem using a statistic that is sufficient *and* complete.

Theorem 2.9 (Lehmann-Scheffé).
Let $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ be an unbiased estimator for θ and $S = S(X_1, \ldots, X_n)$ be a sufficient and complete statistic for θ. Then the estimator

$$\overline{\theta} = \overline{\theta}(X_1, \ldots, X_n) = E(\hat{\theta} \mid S) \qquad (2.5)$$

is a MVUE.

Proof. Let $\widetilde{\theta} = \widetilde{\theta}(X_1, \ldots, X_n)$ be *any* unbiased estimator for θ and let

$$\theta_1 = E(\widetilde{\theta} \mid S).$$

Since S is sufficient, by Theorem 2.6, θ_1 is unbiased ($E(\theta_1) = \theta$) and $V(\theta_1) \leq V(\widetilde{\theta})$.
The same things are true for $\overline{\theta}$. Then

$$E(\overline{\theta}) = E(\theta_1) = \theta,$$

i.e.

$$E\left(E(\hat{\theta} \mid S)\right) = E\left(E(\widetilde{\theta} \mid S)\right) = \theta,$$

which can be rewritten as

$$E\underbrace{\left(E(\hat{\theta} \mid S) - E(\widetilde{\theta} \mid S)\right)}_{\varphi(S)} = 0$$

and this is true for *any* unbiased estimator $\widetilde{\theta}$.
Since S is also complete, it follows that $\varphi \stackrel{a.s.}{=} 0$, which means

$$E(\hat{\theta} \mid S) \stackrel{a.s.}{=} E(\widetilde{\theta} \mid S),$$

i.e.

$$\overline{\theta} \stackrel{a.s.}{=} \theta_1$$

and

$$V(\overline{\theta}) = V(\theta_1).$$

But $V(\overline{\theta}) = V(\theta_1) \leq V(\widetilde{\theta})$, i.e.

$$V(\overline{\theta}) \leq V(\widetilde{\theta}),$$

for *any* unbiased estimator $\widetilde{\theta}$. Thus, $\overline{\theta}$ is a MVUE.

□

Example 2.10. Consider again $X_1, \ldots, X_n$, a sample drawn from the Poisson distribution $\mathcal{P}(\lambda)$, $\lambda > 0$ and the statistic $S = \sum\limits_{k=1}^{n} X_k = n\overline{X}$. Let us find a MVUE for λ.

Solution. To make computations easier, we will find an estimator for the parameter $\theta = e^{-\lambda}$ (which, then, of course, will lead to an estimator for λ). Then the pdf is

$$f(x;\theta) \;=\; \frac{\lambda^x}{x!}e^{-\lambda} \;=\; \frac{1}{x!}\theta\left(\ln\frac{1}{\theta}\right)^x, x = 0, 1, \ldots$$

and S has a Poisson pdf with parameter $n\lambda = n\left(\ln\frac{1}{\theta}\right)$.

Recall that S is a sufficient and complete statistic for λ, which means it is also a sufficient and complete statistic for θ.

In order to use Theorem 2.9, we have to start with an unbiased estimator for θ. For $i = 1, \ldots, n$, consider the variables

$$Y_i \;=\; \begin{cases} 1, & X_i = 0 \\ 0, & X_i \neq 0. \end{cases}$$

Then

$$P(Y_i = 1) \;=\; P(X_i = 0) \;=\; \frac{1}{0!}\theta\left(\ln\frac{1}{\theta}\right)^0 \;=\; \theta$$

and, obviously $P(Y_i = 0) = 1 - \theta$. Thus, each Y_i has a $Bern(\theta)$ pdf

$$Y_i \begin{pmatrix} 0 & 1 \\ 1-\theta & \theta \end{pmatrix}$$

and expectation $E(Y_i) = \theta, i = \overline{1, n}$.

Let

$$\hat{\theta} \;=\; \frac{1}{n}\sum_{i=1}^{n} Y_i \;=\; \frac{\text{the number of } X_i\text{'s that are equal to } 0}{n}.$$

Then

$$E(\hat{\theta}) \;=\; \frac{1}{n}\sum_{i=1}^{n} E(Y_i) \;=\; \frac{1}{n}\cdot n\theta \;=\; \theta,$$

so $\hat{\theta}$ is an unbiased estimator for θ. Hence, by Theorem 2.9, the estimator

$$\overline{\theta} \;=\; E(\hat{\theta} \mid S)$$

is the MVUE. Let us compute it. First off, we have

$$\overline{\theta} \;=\; E(\hat{\theta} \mid S) \;=\; E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i \mid S\right) \;=\; \frac{1}{n}\sum_{i=1}^{n} E(Y_i \mid S) \;=\; \frac{1}{n}\cdot nE(Y_1 \mid S) \;=\; E(Y_1 \mid S).$$

Now, to go further with the computation, let us recall a few things about conditional probability:

$$\begin{aligned} P(A \mid B) &= \frac{P(A \cap B)}{P(B)}, \\ P(B \mid A) &= \frac{P(A \cap B)}{P(A)}, \\ P(A \mid B) &= \frac{P(A) \cdot P(B \mid A)}{P(B)}. \end{aligned}$$

So, we have

$$\begin{aligned} E(Y_1 \mid S) &= P(Y_1 = 0 \mid S = s) \cdot 0 + P(Y_1 = 1 \mid S = s) \cdot 1 \\ &= P(X_1 = 0 \mid S = s) \\ &= \frac{P\Big((X_1 = 0) \cap (S = s)\Big)}{P(S = s)} \\ &= \frac{P(X_1 = 0) \cdot P(S = s \mid X_1 = 0)}{P(S = s)} \\ &= \frac{P(X_1 = 0) \cdot P\left(\sum_{i=1}^{n} X_i = s \mid X_1 = 0\right)}{P(S = s)} \\ &= \frac{P(X_1 = 0) \cdot P(X_2 + \ldots + X_n = s\ )}{P(S = s)} \end{aligned}$$

$P(X_1 = 0) = \dfrac{\lambda^0}{0!}e^{-\lambda} \;=\; e^{-\lambda} = \theta$ and since all X_i's have $\mathcal{P}(\lambda)$ pdf, the sum $X_2 + \ldots + X_n$ will have $\mathcal{P}((n-1)\lambda)$ pdf

$$\left(\begin{array}{c} s \\ \dfrac{((n-1)\lambda)^s}{s!}e^{-(n-1)\lambda} \end{array}\right)_{s=0,1,\ldots}.$$

So, we continue the computation

$$
\begin{aligned}
E(Y_1 \mid S) &= \frac{\theta \cdot \dfrac{1}{s!}\left((n-1)\ln\dfrac{1}{\theta}\right)^s \cdot \theta^{n-1}}{\dfrac{1}{s!}\left(n\ln\dfrac{1}{\theta}\right)^s \cdot \theta^n} \\
&= \left(\frac{n-1}{n}\right)^s \\
&= \left(1-\frac{1}{n}\right)^s.
\end{aligned}
$$

Thus, the estimator

$$
\overline{\theta} \;=\; E(\hat{\theta} \mid S) \;=\; \left(1-\frac{1}{n}\right)^S \;=\; \left(1-\frac{1}{n}\right)^{n\overline{X}}
$$

is the MVUE for θ. To find it for λ, we solve

$$
\begin{aligned}
e^{-\overline{\lambda}} &= \left(\frac{n-1}{n}\right)^{n\overline{X}} \\
-\overline{\lambda} &= n\overline{X}\;\ln\left(\frac{n-1}{n}\right) \\
\overline{\lambda} &= \left(n\;\ln\left(\frac{n}{n-1}\right)\right)\overline{X}.
\end{aligned}
$$

■

Example 2.11. Let $X_1, \ldots, X_n$ be sample variables for a random sample drawn from a Bernoulli distribution with parameter $p \in (0,1)$, unknown. Find the MVUE for p.

Solution. Recall the $Bern(p)$ pdf

$$
\begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}.
$$

So, for each $i = \overline{1,n}$, the pdf can be written

$$
f(x_i; p) = P(X_i = x_i) = p^{x_i}(1-p)^{1-x_i}, \;\; x_i \in \{0,1\},
$$

Then the likelihood function of the sample is

$$
L(x_1, \ldots, x_n; p) \;=\; \prod_{i=1}^{n} f(x_i; p)
$$

$$
\begin{aligned}
&= p^{x_1+\ldots+x_n}(1-p)^{n-(x_1+\ldots+x_n)} \\
&= p^s(1-p)^{n-s},
\end{aligned}
$$

where S is the statistic

$$S = S(X_1, \ldots, X_n) = \sum_{1=1}^{n} X_i = n\overline{X}.$$

By Theorem 2.4, with $g(x_1, \ldots, x_n) = 1$ and $h(s;p) = p^s(1-p)^{n-s}$, S is sufficient.

Now assume that $E(\varphi(S)) = 0$, for all $p \in (0,1)$. Recall that since $X_1, \ldots, X_n$ are independent and identically distributed with a Bernoulli distribution, S has a Binomial distribution with parameters n and p. Then

$$E(\varphi(S)) = \sum_{s=0}^{n} \varphi(s) C_n^s p^s (1-p)^{n-s} = (1-p)^n \sum_{s=0}^{n} \varphi(s) C_n^s \left(\frac{p}{1-p}\right)^s.$$

If $E(\varphi(S)) = 0$, for all $p \in (0,1)$, then

$$\sum_{s=0}^{n} \varphi(s) C_n^s \left(\frac{p}{1-p}\right)^s = 0,$$

for all $p \in (0,1)$, which is possible only if $\varphi(s) = 0$, for all $s = \overline{0,n}$, i.e. $\varphi \overset{a.s.}{=} 0$. Thus S is also complete.

Now let us consider the estimator $\hat{p} = \overline{X} = \frac{1}{n}S$. Since $S \in B(n,p)$, we know that $E(S) = np$ and, hence $E(\hat{p}) = p$, so $\hat{p}$ is an unbiased estimator for p. Then by Theorem 2.9, the MVUE is given by

$$\overline{p} \;=\; E(\hat{p} \mid S) \;=\; E\left(\frac{1}{n}S \mid S\right) \;=\; E\left(\frac{1}{n}S\right) \;=\; \frac{1}{n}S \;=\; \overline{X}.$$

Thus, the sample mean $\overline{X}$ is the MVUE for the population mean $p = E(X)$.

■