ELE8088: Control & Estimation Theory **QUB, 2021**

Handout 16: Good and Better Estimators

Lecturer: Pantelis Sopasakis *Date:* _____

Topics: Statistics and Estimators ∘ MSE, bias and variance ∘ UMVUE ∘ Fisher information ∘ Cramér-Rao lower bound ∘ Sufficient statistics ∘ Rao-Blackwellisation.

Last update: February 1, 2022 at 00:21:41.

16.1 Statistics and Estimators: Definitions

Here we focus on the problem of estimating a *deterministic* parameter, θ (can be scalar or vector-valued), given a collection of observations. Note that θ is assumed to be a *deterministic* parameter, so we will not assume a prior distribution on θ. In other words, we shall follow a *frequentist* rather than a *Bayesian* approach here.

We assume that we measure the values of N random variables, $X_1, \ldots, X_N$, that take values in a set E (e.g., can be $E \subseteq \mathbb{R}$ or $E \subseteq \mathbb{R}^n$) which are typically independently identically distributed according to a parametric distribution with pdf p_θ. This collection of random variables is referred to as our (random) **sample**.

The parameter θ is assumed to be drawn from a set Θ called the **parameter set**. The collection $\{p_\theta; \theta \in \Theta\}$ is known as the **statistical model**. A desirable property of the statistical model is that different parameter values correspond to different distributions, that is, $p_\theta = p_{\theta'} \Rightarrow \theta = \theta'$. If this holds we say that the statistical model is **identifiable**.

As an example, consider the case of $\mathbb{R}$-valued random variables, $X_1, \ldots, X_N \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, where $\theta = (\mu, \sigma^2)$ and $\Theta = \mathbb{R} \times (0, \sigma^2)$. It can be easily checked that this statistical model is identifiable.

We define a **statistic** as a (measurable) function of our sample, that is, $T = r(X_1, \ldots, X_N)$. Note that since $X_1, \ldots, X_N$ are random variables, T is a random variable as well. A well known and widely used statistic is the *sample mean*

$$\bar{X}_N = \tfrac{1}{N} \sum_{i=1}^{N} X_i. \tag{16.1}$$

Another popular statistic is the *sample variance* defined as

$$s^2 = \tfrac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X}_N)^2. \tag{16.2}$$

Other statistics can be $T = \max\{X_1, \ldots, X_N\}$, $T = \min\{X_1, \ldots, X_N\}$, $T = X_1$ or even $T = 100$ — there is an infinite choice of statistics given a sample.

A statistic that is used to estimate a parameter θ is referred to as an **estimator** of θ; an estimator is a statistic that is "close to" θ in some sense.

We define the **bias** of an estimator $\widehat{\theta} = \widehat{\theta}(X_1, \ldots, X_N)$ to be the expectation of the estimation error, $\widehat{\theta} - \theta$, given θ, that is[1]

$$\text{bias}(\widehat{\theta}) = \mathbb{E}[\widehat{\theta} - \theta \mid \theta], \tag{16.3}$$

if the expectation exists. Note that the bias of $\widehat{\theta}$ can depend on θ — our estimator can have a different bias for different values of the (unknown) parameter θ. If $\text{bias}(\widehat{\theta}) = 0$ for all θ we say that the estimator is **unbiased**.

The **variance** of an estimator $\widehat{\theta}$ is defined as

$$\text{var}(\widehat{\theta}) = \text{Var}\left[\widehat{\theta} \mid \theta\right] = \mathbb{E}\left[(\widehat{\theta} - \mathbb{E}[\widehat{\theta} \mid \theta])^2 \mid \theta\right], \tag{16.4}$$

if the expectation exists. We also define the **mean square error** (MSE) of $\widehat{\theta}$ as

$$\text{mse}(\widehat{\theta}) = \mathbb{E}[(\widehat{\theta} - \theta)^2 \mid \theta], \tag{16.5}$$

if the above expectation exists.

Theorem 16.1 (MSE, Variance and Bias) *The mean square error of an estimator is given by*

$$\text{mse}(\widehat{\theta}) = \text{var}(\widehat{\theta}) + \text{bias}(\widehat{\theta})^2. \tag{16.6}$$

Proof: The proof relies on the definition of the MSE in Equation (16.5) and the formula $\text{Var}[X \mid Y] = \mathbb{E}[X^2 \mid Y] - \mathbb{E}[X \mid Y]^2$. This proof is left to the reader as an exercise (☕☕). ■

Note that according to Equation (16.6), if an estimator $\widehat{\theta}$ is unbiased, its MSE is equal to its variance.

[1] Strictly speaking, in this context, this is not a conditional expectation since θ is not a random variable. We write it like this to underline the that that the bias is a function of θ.

In practice it is typically desirable for an estimator to exhibit minimum MSE; at the same time, we want it to be unbiased. These requirements are sometimes impossible to reconcile. In some cases, no unbiased estimators exist. Let us give a few examples to understand the above concepts.

Example 1 (Sample mean is unbiased). Suppose that $X_1, \dots, X_N$ are iid with mean μ. The sample mean is given in Equation (16.1) and is an estimtor of μ. We can easily see that the bias of $\bar{X}_N$ is

$$\text{bias}(\bar{X}_N) = \mathbb{E}[\bar{X}_N - \mu \mid \mu] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} X_i \middle| \mu\right] = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[X_i \mid \mu] = \mu. \tag{16.7}$$

This proves that $\bar{X}_N$ is an unbiased estimator of μ. ♡

Example 2 (MSE of sample mean). Assume that the variance of X_i is known and equal to σ^2 for all $i \in \mathbb{N}_{[1,N]}$. According to Equation (16.4), the variance of the sample mean is

$$\begin{aligned}\text{var}(\bar{X}_N) = \text{Var}[\bar{X}_N \mid \mu] &= \text{Var}\left[\frac{1}{N}\sum_{i=1}^{N} X_i\right] \\ &= \frac{1}{N^2}\text{Var}\left[\sum_{i=1}^{N} X_i\right] \overset{\text{indep.}}{=} \frac{1}{N^2}\sum_{i=1}^{N}\text{Var}[X_i] = \frac{\sigma^2}{N}.\end{aligned} \tag{16.8}$$

Since the estimator is unbiased, according to Theorem 16.1, its MSE is equal to its variance. We see that as $N \to \infty$, the MSE of this estimator converges to zero (at a rate of $\mathcal{O}(1/N)$). ♡

Example 3 (Sample variance is biased). The sample variance given in Equation (16.2) is a biased estimator of σ^2. Indeed,

$$\begin{aligned}\mathbb{E}[s^2 \mid \sigma^2] &= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X}_N)^2\right] \\ &= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}((X_i - \mu)^2 - (\bar{X}_N - \mu)^2)^2\right] \\ &= \frac{1}{N}\sum_{i=1}^{N}\underbrace{\mathbb{E}[(X_i - \mu)^2]}_{\text{Var}[X_i]} - \frac{1}{N}\sum_{i=1}^{N}\underbrace{\mathbb{E}[(\bar{X}_N - \mu)^2]}_{\text{Var}[\bar{X}_N]} = \sigma^2 - \frac{\sigma^2}{N},\end{aligned} \tag{16.9}$$

therefore, the bias of s^2 is

$$\text{bias}(s^2) = \mathbb{E}[s^2 \mid \sigma^2] - \sigma^2 = -\frac{\sigma^2}{N}. \tag{16.10}$$

We see that s^2 is a biased estimator, but the bias converges to 0 as $N \to \infty$. We say that s^2 is an **asymptotically unbiased** estimator of σ^2. ♡

Example 4 (Unbiased estimator of the variance). Following the same procedure as above we can show that the following estimate of σ^2 is an unbiased estimator

$$s^2_{\text{corr}} = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X}_N)^2. \tag{16.11}$$

This estimator is known as *Bessel's correction*. ♡

Example 5 (Lack of unbiased estimator). There are cases where there are no unbiased estimators. For example, suppose that $X \sim \text{Exp}(\lambda)$, $\lambda > 0$. Recall that the pdf of the exponential distribution, $\text{Exp}(\lambda)$, is

$$p_X(x;\lambda) = \lambda e^{-\lambda x}, x \geq 0. \tag{16.12}$$

Suppose that $\widehat{\lambda} = \widehat{\lambda}(X)$ is an unbiased estimator of λ. Then, by definition the following needs to hold for all $\lambda > 0$

$$\begin{aligned}
\text{bias}(\widehat{\lambda}) = 0 &\Leftrightarrow \mathbb{E}[\widehat{\lambda} \mid \lambda] = \lambda \\
&\Leftrightarrow \int_0^\infty \widehat{\lambda}(x) p_X(x) \mathrm{d}x = \lambda \\
&\Leftrightarrow \int_0^\infty \widehat{\lambda}(x) \lambda e^{-\lambda x} \mathrm{d}x = \lambda \\
&\Leftrightarrow \int_0^\infty \widehat{\lambda}(x) e^{-\lambda x} \mathrm{d}x = 1 \\
&\Leftrightarrow \{\mathcal{L}\widehat{\lambda}(x)\}(\lambda) = 1,
\end{aligned} \tag{16.13}$$

where the last equality is due to the fact that $\int_0^\infty \widehat{\lambda}(x)\lambda e^{-\lambda x}\mathrm{d}x$ is the Laplace transform of $\widehat{\lambda}$ evaluated at λ. However such a function, $\widehat{\lambda}(x)$ cannot exist[2]. ♡

[2]The reason is that the property $\lim_{\lambda\to\infty}\{\mathcal{L}\widehat{\lambda}(x)\}(\lambda) = 0$ should be satisfied.

Lastly, we define a uniformly minimum-variance unbiased estimator (**UMVUE**) to be an unbiased estimator such that there is no other unbiased estimator with a lower variance. Formally, let $X_1, \ldots, X_N$ be a random sample and $\widehat{\theta}(X_1, \ldots, X_N)$ is an *unbiased* estimator of a parameter θ. Then, $\widehat{\theta}$ is UMVUE if

$$\text{var}[\widehat{\theta}] \leq \text{var}[\widetilde{\theta}], \tag{16.14}$$

for all θ and for all unbiased estimators $\widetilde{\theta}$.

Bias-Variance Tradeoff: UMVUEs are considered good estimators. But sometimes, we may decide to choose an estimator that has some small nonzero bias, but comes with a lower variance. This is known as the bias-variance tradeoff problem or dilemma.

16.2 Fisher Information for one-parameter models

We shall introduce the concept of the Fisher information via an example. Let $X \sim \mathcal{N}(\mu, \sigma^2)$ where σ^2 is known and μ is unknown. Let us consider two cases:

- Case I: $X \sim \mathcal{N}(\mu, 20)$ (large variance),
- Case II: $X \sim \mathcal{N}(\mu, 0.1)$ (small variance),

where the value of μ is the same in both cases. Suppose we obtain one measurement, $X = x$, so the log-likelihood of μ is

$$\ell(\mu; x) = \log p_X(x; \mu) = -\log(\sqrt{2\pi}\sigma) - \frac{(x-\mu)^2}{2\sigma^2}. \tag{16.15}$$

Figure 16.1 shows the log-likelihood functions, $\ell(\mu; x)$, for a few samples $X \sim \mathcal{N}(\mu, 0.1)$ (Case I). We can say that each observation is quite *informative* about the parameter μ.

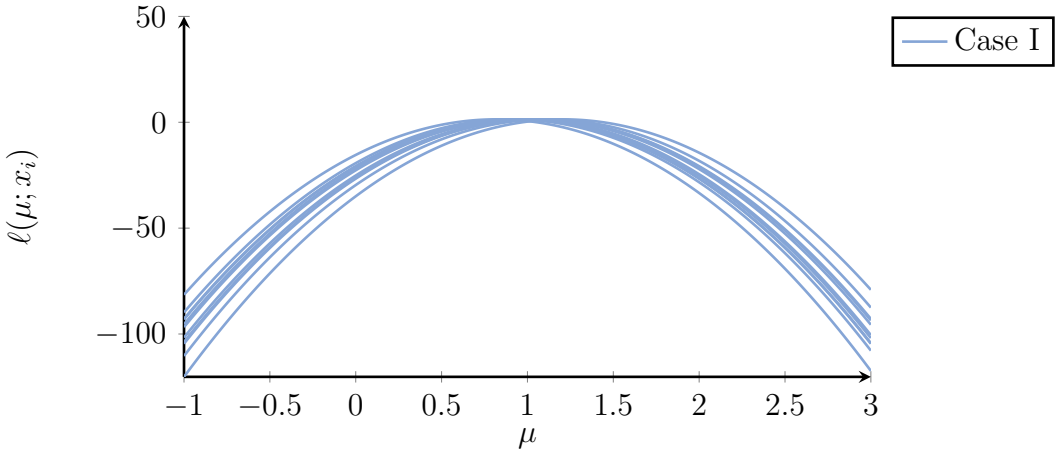


Figure 16.1: (Case I) Plot of $\ell(\mu; x_i)$ for $N = 20$ measurements. The likelihood function has, on average, a healthy curvature (it is not too flat).

It is is not difficult to guess the true value of μ from the above plot (in fact it is $\mu = 1$). On the other hand, in Case II we have the following plot (Figure 16.2) of the log-likelihood function for each observation

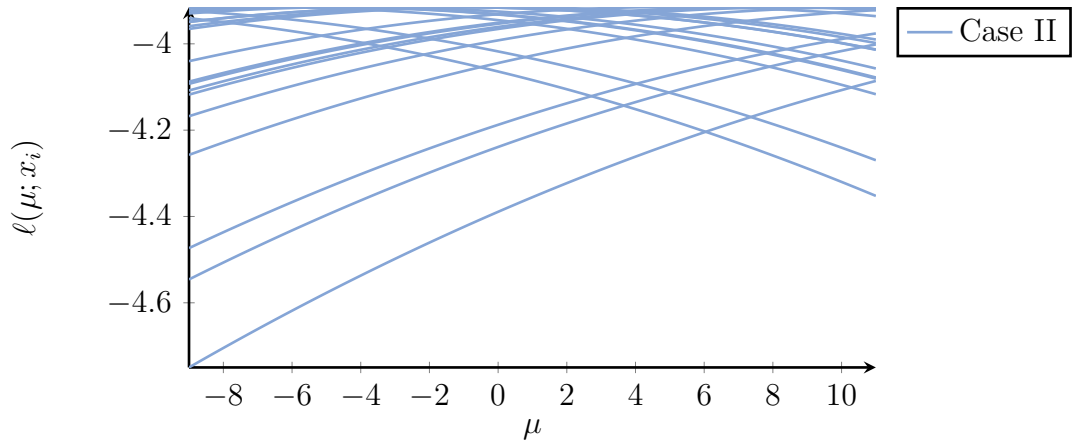


Figure 16.2: (Case II) Plot of $\ell(\mu; x_i)$ for $N = 20$ measurements. The likelihood function is, on average, quite flat.

Here it is more difficult to guess what the true value of μ is, or, to put it differently, the data are *less informative* about μ, the main reason being that

Fisher proposed the following quantity to quantify the *information* that an observation X[3] offers about the parameter θ:

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\ell(\theta; X)\middle|\theta\right], \tag{16.16}$$

where the expectation is taken with respect to X and ℓ denotes the log-likelihood. In other words, using the pdf of X,

$$I(\theta) = \int_E \frac{\partial^2}{\partial\theta^2}\ell(\theta; x)p_X(x; \theta)\mathrm{d}x. \tag{16.17}$$

This quantity is known as the **Fisher information** of the statistical model. The Fisher information quantifies how much information the data gives us for the unknown parameter.

There is an alternative expression for the determination of $I(\theta)$ — we will prove it in a moment — which uses the first derivative of ℓ:

$$I(\theta) = \mathbb{E}\left[\left(\frac{\partial\ell(\theta; X)}{\partial\theta}\right)^2\middle|\theta\right]. \tag{16.18}$$

[3] or several observations, $X_1, \ldots, X_N$

Again, as we did in Equation (16.17), we can use the pdf of X to compute $I(\theta)$, that is,

$$I(\theta) = \int_E \left(\frac{\partial \ell(\theta; x)}{\partial \theta} \right)^2 p_X(x; \theta) \mathrm{d}x. \tag{16.19}$$

The Fisher information is defined if the following **basic regularity conditions** are satisfied: (i) $p_X(x;\theta)$ is differentiable wrt θ almost everywhere, (ii) the support of $p_X(x;\theta)$ does not depend on θ. In other words, the set $\{x : p_X(x;\theta) > 0\}$ does not depend on θ, and (iii) it holds that $\frac{\partial}{\partial\theta} \int p_X(x;\theta)\mathrm{d}x = \int \frac{\partial}{\partial\theta} p_X(x;\theta)\mathrm{d}x$.

Note that the first assumption is *not satisfied for the uniform distribution,* $U(0,\theta)$, and the Fisher information is not defined for $U(0,\theta)$.

Theorem 16.2 (Fisher Information Equivalent Formulas) *Suppose $\ell(\theta; X)$ is the likelihood of a parameter θ given some observation(s) X and the above regularity assumptions are satisfied. Then,*

$$\mathbb{E}\left[\frac{\partial \ell(\theta; X)}{\partial \theta} \middle| \theta \right] = 0. \tag{16.20}$$

Additionally, assume that it is twice differentiable and (iv) assumption (iii) holds for the second order derivative wrt θ. Then, the Fisher information is given by

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2} \ell(\theta; X) \middle| \theta \right] = \mathbb{E}\left[\left(\frac{\partial \ell(\theta; X)}{\partial \theta} \right)^2 \middle| \theta \right] = \mathrm{Var}\left[\frac{\partial \ell(\theta; X)}{\partial \theta} \middle| \theta \right]. \tag{16.21}$$

Proof: We have

$$\begin{aligned} \mathbb{E}\left[\frac{\partial \ell(\theta; X)}{\partial \theta} \middle| \theta \right] &= \mathbb{E}\left[\frac{\partial \log p(X; \theta)}{\partial \theta} \middle| \theta \right] = \int_E \frac{\frac{\partial p_X(x;\theta)}{\partial\theta}}{\cancel{p_X(x;\theta)}} \cancel{p_X(x;\theta)} \mathrm{d}x \\ &= \int_E \frac{\partial p_X(x;\theta)}{\partial \theta} \mathrm{d}x \overset{\text{Ass. (iii)}}{=} \frac{\partial}{\partial\theta} \underbrace{\int_E p_X(x;\theta)\mathrm{d}x}_{=1} = 0, \end{aligned} \tag{16.22}$$

so the last equality in Equation (16.21) follows.

To prove that the two expectations are equal we use the fact that

$$\frac{\partial^2}{\partial\theta^2}\log p_X(X;\theta) = \frac{\frac{\partial^2 p_X(X;\theta)}{\partial\theta^2}}{p_X(X;\theta)} - \left(\frac{\frac{\partial^2 p_X(X;\theta)}{\partial\theta^2}}{p_X(X;\theta)}\right)^2 = \frac{\frac{\partial^2 p_X(X;\theta)}{\partial\theta^2}}{p_X(X;\theta)} - \left(\frac{\partial}{\partial\theta}\log p_X(X;\theta)\right)^2. \quad (16.23)$$

Note that

$$\mathbb{E}\left[\frac{\frac{\partial^2 p_X(X;\theta)}{\partial\theta^2}}{p_X(X;\theta)}\right] = \int_E \frac{\frac{\partial^2 p_X(x;\theta)}{\partial\theta^2}}{\cancel{p_X(x;\theta)}}\cancel{p_X(x;\theta)}\mathrm{d}x \overset{\text{(iv)}}{=} \frac{\partial^2 p_X(x;\theta)}{\partial\theta^2}\underbrace{\int_E p_X(x;\theta)\mathrm{d}x}_{=1} = 0. \quad (16.24)$$

Now taking the expectation on Equation (16.23) and using Equation (16.24), Equation (16.21) follows. ∎

Example 6 (Fisher information of Exponential). Suppose $X \overset{\text{iid}}{\sim} \text{Exp}(\lambda)$ where $\lambda > 0$ is an unknown value. We want to quantify how informative a measurement of X. The likelihood function is $L(\lambda, x) = \lambda e^{-\lambda x}$, where $x \geq 0$, so the log-likelihood is $\ell(\lambda, x) = \log\lambda - x\lambda$. Suppose that the true value of λ is 3. Let us samples some values $x_1, \ldots, x_10$ using Python (`numpy.random.Generator.exponential`) and plot $\ell(\lambda, x_i)$ which is shown in Figure 16.3.

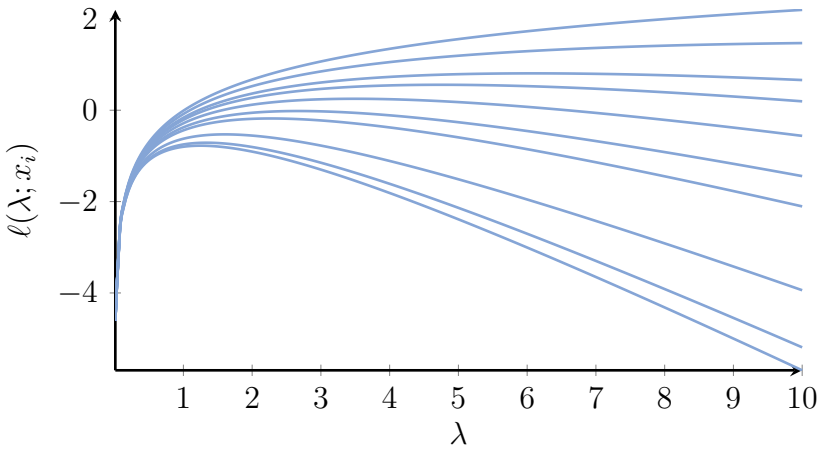


Figure 16.3: Log-likelihood, $\ell(\lambda; x_i)$, for different values sampled from Exp(3).

In this example we see that the log-likelihood is not quadratic in λ, so its curvature depends on λ. In particular, the curvature is

$$\frac{\partial^2}{\partial\lambda^2}\ell(\lambda; X) = -\frac{1}{\lambda^2}, \quad (16.25)$$

therefore, the Fisher information is $I(\lambda) = -\frac{1}{\lambda^2}$. ♡

Example 7 (Fisher information of Bernoulli). The likelihood function of a Bernoulli trial is $L(p; X) = p^X(1-p)^{1-X}$, so the log-likelihood is

$$\ell(p; X) = X \log p + (1 - X) \log(1 - p). \tag{16.26}$$

The second derivative of ℓ with respect to p is

$$\frac{\partial^2 \ell(p; X)}{\partial p^2} = \frac{X}{p^2} + \frac{1 - X}{(1 - p)^2}. \tag{16.27}$$

The Fisher information is

$$I(p) = \mathbb{E}\left[\frac{\partial^2 \ell(p; X)}{\partial p^2}\middle| p\right] = \mathbb{E}\left[\frac{X}{p^2} + \frac{1 - X}{(1 - p)^2}\middle| p\right] = \ldots = \frac{1}{p(1 - p)}. \tag{16.28}$$

The plot of the Fisher information, $I(p)$, is shown in Figure 16.4.

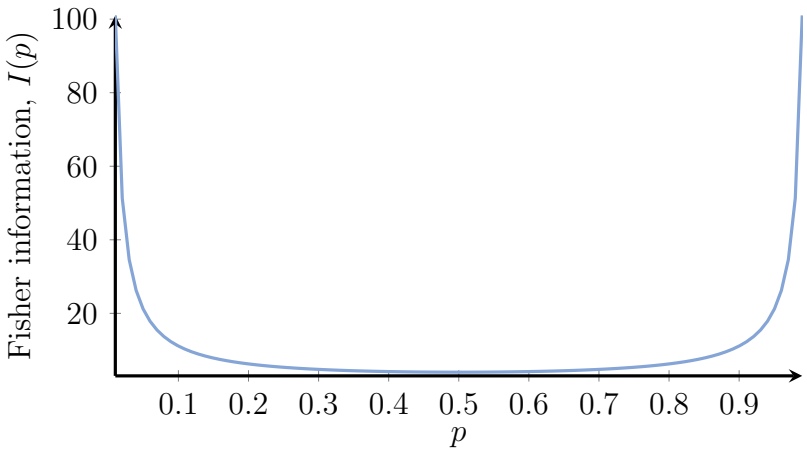


Figure 16.4: Fisher information for Ber(p) as a function of p. The outcomes of blatantly unfair coins are more informative about the parameter p.

We see that the Fisher information incrases for values of p close to 0 or 1 (for blatantly unfair coins). ♡

Exercise 1 (☕). Determine the Fisher information in each of the following cases: (i) Poisson(λ), (ii) Exp(λ), (iii) $\mathcal{N}(\mu, \sigma^2)$ with known σ^2, (iv) $\mathcal{N}(0, \sigma^2)$ with unknown σ^2. Then fill in the folloing table:

Statistical model	θ	$I(\theta)$
Ber(p)	p	$\frac{1}{p(1-p)}$
Poisson(λ)	λ	
Exp(λ)	λ	
$\mathcal{N}(\mu, \sigma^2)$	μ	
$\mathcal{N}(\mu, \sigma^2)$	σ^2	
$U(0, \theta)$	θ	

For each of the above distributions, plot $I(\theta)$ and produce figures akin to Figure 16.1 or Figure 16.3 for different values of θ.

Exercise 2 (☕). Suppose that the Fisher information associated with a likelihood function $\ell(\theta; X)$ is $I(\theta)$. Suppose we have a sample of N observations, $X_1, \ldots, X_N$. Show that the Fisher information for $\ell(\theta; X_1, \ldots, X_N)$ is $NI(\theta)$.

Exercise 3 (Fisher information of transformation) (☕☕). Suppose the log-likelihood function, $\ell(\theta; X)$, has Fisher information $I(\theta)$. Now suppose that the statistical model is parametrised with a parameter λ such that $\theta = g(\lambda)$, for some continuously differentiable function g. Let $I_1(\lambda)$ be the Fisher information of the new model. Show that

$$I_1(\lambda) = g'(\lambda)^2 I(g(\lambda)). \tag{16.29}$$

Exercise 4 (Application of Exercise 3) (☕). Suppose $X \sim \text{Exp}(\lambda)$. We know that the Fisher information is $I(\lambda) = 1/\lambda^2$. We reparametrise this model using $\lambda = \xi^2$, that is $X \sim \text{Exp}(\xi^2)$. What is the Fisher information of the parameter ξ for the new model?

16.3 Cramér-Rao Bound for one-parameter models

The Cramér-Rao Bound is a lower bound on the variance of an unbiased estimator. The result we are about to state relies on the following weak regularity assumptions on the likelihood function, $\ell(\theta)$, and the estimator, $\widehat{\theta}$.

Regularity assumptions. In addition to the basic regularity assumptions we stated on page 16-9, suppose that for all x for which $p_X(x;\theta) > 0$, the derivative $\partial/\partial\theta\, p_X(x;\theta)$ exists, and the following holds

$$\frac{\partial}{\partial\theta}\int_E \widehat{\theta}(x)p_X(x;\theta)\mathrm{d}x = \int_E \widehat{\theta}(x)\frac{\partial}{\partial\theta}p_X,(x;\theta)\mathrm{d}x \tag{16.30}$$

whenever the right hand side exists and is finite.

Theorem 16.3 (Cramér-Rao Bound) *Let X be a sample from with pdf $p_X(x;\theta)$ and $\widehat{\theta}$ is an unbiased estimator of θ. Suppose that the regularity assumptions hold. Then,*

$$\mathrm{var}[\widehat{\theta}] \geq \frac{1}{I(\theta)}. \tag{16.31}$$

It follows from Theorem 16.3 that if an unbiased estimator attains the lower bound, i.e., if $\mathrm{var}[\widehat{\theta}] = \frac{1}{I(\theta)}$, then it is a UMVUE. However, this lower bound is not tight in the sense that not all UMVUEs attain this lower bound. Let us give an example where this happens.

Example 8 (UMVUE via Cramér-Rao bound). In Example 1 we showed that the sample mean, $\bar{X}_N$, is an unbiased estimator of the mean μ and in Example 2 we showed that its variance is $\mathrm{var}[\bar{X}_N] = \sigma^2/N$ (σ^2 is assumed to be known). Now assume that $X_1, \ldots, X_N \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ for which the regularity assumptions hold and the Fisher information for μ is

$$I(\mu) = \frac{N}{\sigma^2}. \tag{16.32}$$

We see that $\mathrm{var}[\bar{X}_N] = 1/I(\mu)$, therefore, $\bar{X}_N$ is a UMVUE for μ. ♡

Example 9 (Estimator of Bernoulli parameter is UMVUE). Let $X_1, \ldots, X_N \overset{\text{iid}}{\sim} \text{Ber}(p)$. The estimator

$$\widehat{p}(X_1, \ldots, X_N) = \tfrac{1}{N} \sum_{i=1}^{N} X_i,$$

is unbiased (why?) and its variance is (see Equation (16.4))

$$\text{var}[\widehat{p}] = \text{Var}\left[\tfrac{1}{N} \sum_{i=1}^{N} X_i \middle| p \right] \overset{\text{indep.}}{=} \tfrac{1}{N^2} \sum_{i=1}^{N} \text{Var}[X_i] = \frac{p(1-p)}{N}. \tag{16.33}$$

Now if $X \sim \text{Ber}(p)$, according to Equation (16.28) the Fisher information of p is $I(p) = \frac{1}{p(1-p)}$ and following Exercise 2, the Fisher information for the case of N independent observations becomes

$$I(p) = \frac{N}{p(1-p)}. \tag{16.34}$$

The reader can verify that the regularity assumptions hold of the Bernoulli distribution. We see that $\text{var}[\widehat{p}] = I(p)^{-1}$, therefore, $\widehat{p}$ is a UMVUE of p. ♡

An unbiased estimator that attains the Cramér-Rao lower bound is called an **efficient** estimator. All efficient estimators are UMVUE, however not all UMVUEs are efficient.

We shall now state a very useful result that can be used to determine an efficient estimator (if it exists). The proof is a bit technical, so we will skip it.

Theorem 16.4 (Factorisation for efficient estimators) *An efficient estimator, $\widehat{\theta}$, exists if and only if*

$$\frac{\partial \ell(\theta; x)}{\partial \theta} = I(\theta)[\widehat{\theta}(x) - \theta]. \tag{16.35}$$

If follows from the theorem that $\text{var}[\widehat{\theta}] = I(\theta)^{-1}$. Let us give an example where we "factorise" $\frac{\partial \ell(\theta;x)}{\partial \theta}$ as in Equation (16.35) to determine an efficient estimator.

Example 10 (Estimator of σ^2 with known mean). Suppose $X_1, \ldots, X_N \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, where μ is known and σ^2 is an unknown parameter. The likelihood of σ^2 given a measurements $X = x$ is

$$\ell(\sigma^2) = \log p_X(x; \mu, \sigma^2) = -\tfrac{1}{2}\log(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}. \tag{16.36}$$

We leave it to the reader to verify that the Fischer information for one observation is $1/2\sigma^4$, so for N observations the Fisher information is (see Exercise 2)

$$I(\sigma^2) = \frac{N}{2\sigma^4}. \tag{16.37}$$

In Example 4 we showed that $s^2_{\text{corr}} = \frac{1}{N-1}\sum_{i=1}^N (X_i - \bar{X}_N)^2$ is an unbiased estimator of σ^2. It can be shown that the variance of this estimator is

$$\text{var}[s^2_{\text{corr}}] = \frac{2\sigma^4}{n-1} > \frac{\sigma^4}{n}, \tag{16.38}$$

so s^2_{corr} does not attain the lower bound of the Cramér-Rao inequality. Let us now try to factorise $\frac{\partial \ell(\sigma^2; x)}{\partial \sigma^2}$ as in Equation (16.35):

$$\begin{aligned}
\frac{\partial \ell(\theta; x)}{\partial \sigma^2} &= \frac{\partial}{\partial \sigma^2} \log p(x_1, x_2, \ldots, x_N; \mu, \sigma^2) \\
&= \frac{\partial}{\partial \sigma^2} \log \prod_{i=1}^N p(x_i; \mu, \sigma^2) = \frac{\partial}{\partial \sigma^2} \sum_{i=1}^N \log p(x_i; \mu, \sigma^2) \\
&= \frac{\partial}{\partial \sigma^2} \sum_{i=1}^N \log \left[\tfrac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2} \right) \right] \\
&= \frac{\partial}{\partial \sigma^2} \sum_{i=1}^N \left[-\tfrac{1}{2}\log(2\pi\sigma^2) - \frac{(x_i-\mu)^2}{2\sigma^2} \right] \\
&= \frac{\partial}{\partial \sigma^2} \left[-\frac{N}{2}\log(2\pi\sigma^2) - \sum_{i=1}^N \frac{(x_i-\mu)^2}{2\sigma^2} \right] \\
&= -\frac{N}{2\sigma^2} + \sum_{i=1}^N \frac{(x_i-\mu)^2}{2\sigma^4}.
\end{aligned} \tag{16.39}$$

If we now take out as a common factor $N/2\sigma^4$ — i.e., the Fisher information according to Equation (16.37) — we have

$$\frac{\partial \ell(\theta; x)}{\partial \sigma^2} = \underbrace{\frac{N}{2\sigma^4}}_{I(\sigma^2)} \left(\underbrace{\frac{\sum_{i=1}^N (x_i-\mu)^2}{N}}_{\text{efficient estimator}} - \sigma^2 \right). \tag{16.40}$$

The regularity assumptions are satisfied in this case so from Theorem 16.4 we conclude that the estimator

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}, \tag{16.41}$$

is an efficient estimator of σ^2 (thus UMVUE) and, by definition, its variance is $1/I(\sigma^2) = \frac{\sigma^4}{N}$. There is no other unbiased estimator of σ^2 with a lower variance. Note that this holds only under the assumption that μ is known. ♡

16.4 Sufficient statistics

If we are given a sample $X_1, \ldots, X_N \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, where σ^2 is known, and we need to determine μ we saw in Example 8 that we can use the statistic (estimator)

$$\bar{X}_N(X_1, \ldots, X_N) = \frac{\sum_{i=1}^{N} X_i}{N}, \tag{16.42}$$

which is a UMVUE. So, do we need to know the entire sample to estimate μ? We see that we just need to know the sum of the observations. Would the knowledge of all observations be of any use? Can we do anything better if we have all the observations instead of their sum? (spoiler: no) Such statistics that summarise the data in such a way that the knownledge of the data themselves is not needed are called sufficient.

A statistic is **sufficient** for a statistical model and a certain parameter if, roughly speaking, the knownledge of the sample offers no additional information than the statistic itself. To express this mathematically we say that a statistic $T(X_1, \ldots, X_N)$ is sufficient if the conditional distribution of the sample, $X_1, \ldots, X_N$, given $T = t$ and θ does not depend on θ.

However, this definition is difficult to use in practice. Instead, to show that a given statistic is sufficient we can use the following theorem that we state without a proof.

Theorem 16.5 (Fisher-Neymann Factorisation Theorem) *Suppose $X_1, \ldots, X_N$ are iid samples. A statistic $T(X_1, \ldots, X_N)$ is sufficient if and only if the likelihood of θ given the observations can be written as*

$$L(\theta; x_1, \ldots, x_N) = u(x_1, \ldots, x_N) \cdot v(t, \theta), \tag{16.43}$$

where $t = T(x_1, \ldots, x_N)$ and where u and v are nonnegative functions.

Let us give a few examples of sufficient statistics for various statistical models.

Example 11 (Normal data, unknown mean/known variance). Let $X_1, \ldots, X_N \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ with known variance, σ^2, and unknown mean, μ. We will show that the statistic $T(X_1, \ldots, X_N) = \sum_{i=1}^N X_i$ is sufficient for this statistical model for μ. The likelihood of μ given the observations is

$$L(\mu; x_1, \ldots, x_N) = c \exp\left(-\frac{\sum_{i=1}^N (x_i - \mu)^2}{2\sigma^2} \right) \tag{16.44}$$

for a constant c, which does not depend on μ

$$= c \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^N x_i^2 + \frac{\mu}{2\sigma^2} \underbrace{\sum_{i=1}^N x_i}_{T} - \frac{N\mu^2}{2\sigma^2} \right), \tag{16.45}$$

so we can define

$$u(x_1, \ldots, x_N) = c \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^N x_i^2 \right), \tag{16.46a}$$

$$v(T, \mu) = \exp\left(\frac{N\mu^2}{2\sigma^2} + \frac{\mu}{\sigma^2} T \right). \tag{16.46b}$$

According to Theorem 16.5, T is a sufficient statistic. ♡

Example 12 (Bernoulli data). Let $X_1, \ldots, X_N \overset{\text{iid}}{\sim} \text{Ber}(\theta)$, where p is unknown. The joint pdf of the sample is

$$p(x_1, \ldots, x_N; \theta) = \prod_{i=1}^N \theta^{x_i} (1-\theta)^{1-x_i} = \underbrace{\theta^{\sum_{i=1}^N x_i}}_{u(x_1, \ldots, x_N)} \underbrace{(1-\theta)^{N - \sum_{i=1}^N X_i}}_{v(\sum_{i=1}^N X_i, \theta)}, \tag{16.47}$$

from which we see that $T = \sum_{i=1}^N X_i$ is a sufficient statistic for θ. ♡

Exercise 5 (Gamma data, unknown α, known β) (☕). Let $X_1, \ldots, X_N \overset{\text{iid}}{\sim} \Gamma(\alpha, \beta)$, where α is an unknown parameter and β is known. (i) Show that the statistic

$$T_1(X_1, \ldots, X_N) = \sum_{i=1}^N \log X_i,$$

is a sufficient statistic for α. (ii) Show that the statistic

$$T_2(X_1, \ldots, X_N) = \prod_{i=1}^N X_i,$$

is also a sufficient statistic.

Exercise 6 (Sufficient statistic for $U(0,\theta)$) (☕☕☕). Let $\theta > 0$ be an unknown parameter and $X_1, \ldots, X_N \overset{\text{iid}}{\sim} U(0,\theta)$ (uniform distribution on $[0,\theta]$). Determine a sufficient statistic for θ for this statistical model[4].

Exercise 7 (Poisson data) (☕). Suppose $X_1, \ldots, X_N \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$ with unknown parameter $\lambda > 0$. Show that $T(X_1, \ldots, X_N) = \sum_{i=1}^N X_i$ is a sufficient statistic.

Exercise 8 (Sufficient statistics) (☕☕). Fill in the following table with sufficient statistics:

Statistical model	θ	Sufficient Statistic	Statistical model	θ	Sufficient Statistic
$U(0,\theta)$	θ	$\max_i X_i$	Poisson(λ)	λ	
Ber(p)	p	$\sum_{i=1}^N X_i$	Exp(λ)	λ	
$\Gamma(\alpha,\beta)$	α	$\sum_{i=1}^N \log X_i$	$\mathcal{N}(\mu,\sigma^2)$	μ	
$\Gamma(\alpha,\beta)$	β		$\mathcal{N}(\mu,\sigma^2)$	σ^2	

Exercise 9 (Exponential families) (☕☕). A parametric distribution, $p_X(x;\theta)$, is said to belong to the *exponential family* if it can be written as

$$p_X(x;\theta) = h(x)\exp\left[\eta(\theta)t(x) - a(\theta)\right]. \tag{16.48}$$

(i) show tha the normal, exponential, Poisson, Bernoulli, and gamma (either with $\theta = \alpha$ or $\theta = \beta$) distributions belong to the exponential family, (ii) $T(X_1, \ldots, X_N) = \sum_{i=1}^N t(X_i)$ is a sufficient statistic, (iii) show that the normal distribution, $\mathcal{N}(\mu, 1)$, and the exponential distribution, Exp(λ), belong to the exponential family.

[4] Spoiler: show that $T = \max_i X_i$ is a sufficient statistic.

16.5 Better Estimators with Rao-Blackwellisation

16.5.1 Rao-Blackwell Theorem

Let $X_1, \ldots, X_N$ be a sample of independent random variables with pdf $p(x;\theta)$. Suppose we have a rough estimator $\widehat{\theta}(X_1, \ldots, X_N)$ of an unknown parameter θ and let $T(X_1, \ldots, X_N)$ be a *sufficient* statistic for θ. We define the **Rao-Balckwell estimator** which is given by

$$\widetilde{\theta}_{\text{rb}} = \mathbb{E}[\widehat{\theta} \mid T]. \tag{16.49}$$

The Rao-Blackwell estimator is very often better than $\widehat{\theta}$ in terms of its MSE. Additionally, if the original estimator, $\widehat{\theta}$, is unbiased, its "Rao-Balckwellisation" in Equation (16.49) will also be unbiased. In practice, the Rao-Blackwellisation of a crude estimator is often significantly better in terms of MSE and under certain conditions, it yields a UMVUE (see Section 16.5.2).

Important remark: The Rao-Blackwell estimator is a function of T! This results from T being a sufficient statistic. In order to use $\widetilde{\theta}_{\text{rb}}$ we only need to know the value of the sufficient statistic T — not the entire sample, $X_1, \ldots, X_N$. In other words, we have an estimator $\widetilde{\theta}_{\text{rb}}(T(X_1, \ldots, X_N))$. For example, if we have the observations $X_1 = x_1, \ldots, X_N = x_N$, the Rao-Blackwell estimate of θ is

$$\widetilde{\theta}_{\text{rb}}(t) = \mathbb{E}[\widehat{\theta} \mid T = t], \tag{16.50}$$

where $t = T(x_1, \ldots, x_N)$.

Theorem 16.6 (Rao-Blackwell Theorem) *Let $\widehat{\theta}(X_1, \ldots, X_N)$ be an estimator of a parameter θ, let T be a complete statistic and $\widetilde{\theta}_{\text{rb}}$ be the estimator given in Equation (16.49). Assume that the variance of $\widehat{\theta}$ is finite for all $\theta \in \Theta$. Then,*

$$\text{mse}(\widetilde{\theta}_{\text{rb}}) \leq \text{mse}(\widehat{\theta}), \tag{16.51}$$

for all $\theta \in \Theta$. Additionally, if $\widehat{\theta}$ unbiased if and only if $\widetilde{\theta}_{\text{rb}}$ is unbiased.

Proof: The MSE of $\widetilde{\theta}_{\rm rb}$ is

$$\begin{aligned}\mathrm{mse}(\widetilde{\theta}_{\rm rb}) &= \mathbb{E}[(\widetilde{\theta}_{\rm rb}-\theta)^2] = \mathbb{E}[(\mathbb{E}[\widehat{\theta}\mid T]-\theta)^2] = \mathbb{E}[\mathbb{E}[\widehat{\theta}-\theta\mid T]^2] \\ &\leq \mathbb{E}[\mathbb{E}[\widehat{\theta}-\theta\mid T]]^2 = \mathbb{E}[\widehat{\theta}-\theta]^2 = \mathrm{mse}(\widehat{\theta}). \end{aligned} \tag{16.52}$$

Note that the expectation $\mathbb{E}[(\widetilde{\theta}_{\rm rb}-\theta)^2]$ is with respect to t. The inequality follows from Jensen's inequality according to which $\mathbb{E}[Z^2] \leq \mathbb{E}[Z]^2$.

The fact that $\widetilde{\theta}_{\rm rb}$ is unbiased whenever the original estimator is unbiased is left to the reader as an exercise. ■

We will now give a few examples, but first we need to do a quick brush up... Suppose $X \sim \mathrm{Ber}(\theta)$. Then $\mathsf{P}[X=1] =$ _____ and

$$\mathsf{P}[X=k] = \text{\_\_\_\_\_}, \tag{16.53}$$

for $k \in \{0,1\}$. Moreover, if $X_1,\ldots,X_N \overset{\text{iid}}{\sim} \mathrm{Ber}(\theta)$ then $T = \sum_{i=1}^N X_i \sim$ _____ and

$$\mathsf{P}[T=t] = \text{\_\_\_\_\_}, \tag{16.54}$$

for $t \in$ _____.

If X is a *discrete* random variable, and Y is another random variable with $\mathsf{P}[Y=y] > 0$, then

$$\mathbb{E}[X \mid Y=y] = \text{\_\_\_\_\_}. \tag{16.55}$$

On the other hand, if X and Y are continuous random variables with joint pdf $p_{X,Y}$, then

$$\mathbb{E}[X \mid Y=y] = \text{\_\_\_\_\_}. \tag{16.56}$$

Let us start with an example involving the Bernoulli distribution, which is a discrete distribution.

Example 13 (Rao-Blackwellisation, Bernoulli distribution). For this example we need to recall that when $X_1, \dots, X_N \overset{\text{iid}}{\sim} \text{Ber}(\theta)$, then $\sum_{i=1}^N X_i \sim \text{Binom}(N, \theta)$[5].

Let $X_1, \dots, X_N \overset{\text{iid}}{\sim} \text{Ber}(\theta)$, where θ is unknown. The estimator $\widehat{\theta} = X_1$ is unbiased ($\mathbb{E}[\widehat{\theta}] = \mathbb{E}[X_1 \mid \theta] = \theta$) and has variance $\text{var}[\widehat{\theta}] = \text{Var}[X_1 \mid \theta] = \theta(1-\theta)$, but it is clearly a very crude estimator. Indeed, it can only return 0 or 1.

In Example 12 we showed that $T = \sum_{i=1}^N X_i$ is a sufficient statistic. We can use it to Rao-Blackwellise the above estimator. We have

$$\widetilde{\theta}_{\text{rb}}(t) = \mathbb{E}[X_1 \mid T = t] = \mathbb{E}\left[X_1 \mid \sum_{i=1}^N X_i = t\right] \tag{16.57}$$

and since X_1 is a discrete random variance we can use the formula

$$= \frac{\mathsf{P}\left[X_i = 1, \sum_{i=1}^N X_i = t\right]}{\mathsf{P}\left[\sum_{i=1}^N X_i = t\right]} \tag{16.58}$$

$$= \frac{\mathsf{P}\left[X_i = 1, \sum_{i=2}^N X_i = t-1\right]}{\mathsf{P}\left[\sum_{i=1}^N X_i = t\right]} \tag{16.59}$$

since X_1 and $\sum_{i=2}^N$ are independent

$$= \frac{\cancelto{\theta}{\mathsf{P}[X_i = 1]} \cdot \mathsf{P}\left[\sum_{i=2}^N X_i = t-1\right]}{\mathsf{P}\left[\sum_{i=1}^N X_i = t\right]} \tag{16.60}$$

and using the fact that $\sum_{i=1}^N X_i \sim \text{Binom}(N, \theta)$

$$= \frac{\theta\binom{N-1}{t-1}\theta^{t-1}(1-\theta)^{N-t}}{\binom{N}{t}\theta^t(1-\theta)^{N-t}} = \frac{t}{N}, \tag{16.61}$$

therefore, the Rao-Blackwellised estimator is T/N, it is unbiased, and its variance is $\text{var}[\widetilde{\theta}_{\text{rb}}] = \frac{1}{N}\theta(1-\theta)$. In fact, $\widetilde{\theta}_{\text{rb}}$ is the maximum likelihood estimator of θ and as we can see from Example 7 it attains the Cramér-Rao lower bound, therefore it is a UMVUE. ♡

[5] Take a moment to revise the Bernoulli and binomial distributions in Handout 13.

Example 14 (Rao-Blackwellisation, Uniform Distribution). Suppose $X_1, \ldots, X_N \overset{\text{iid}}{\sim} U(0, \theta)$, where $\theta > 0$ is an unknown parameter. Using the factorisation theorem (Theorem 16.5) we can show that $T = \max_{i=1,\ldots,N} X_i$ is a sufficient statistic (Exercise 6).

Define the statistic $S = \sum_{i=1}^{N} X_i$. Then $\mathbb{E}[S] = N\theta/2$, so a naive approach can be to use the following estimator of θ

$$\widehat{\theta} = \frac{2}{N} \sum_{i=1}^{N} X_i, \tag{16.62}$$

which is unbiased (but not UMVUE). Its Rao-Blackwellisation is

$$\widetilde{\theta}_{\text{rb}}(t) = \mathbb{E}[\widehat{\theta} \mid T = t] = \mathbb{E}\left[\frac{2}{N} \sum_{i=1}^{N} X_i \middle| \max_i X_i = t \right] \tag{16.63}$$

and because of the independence of X_i this is

$$= \frac{2}{N} \sum_{i=1}^{N} \mathbb{E}\left[X_i | \max_i X_i = t \right] = 2\mathbb{E}\left[X_1 | \max_i X_i = t \right]. \tag{16.64}$$

We now do this trick

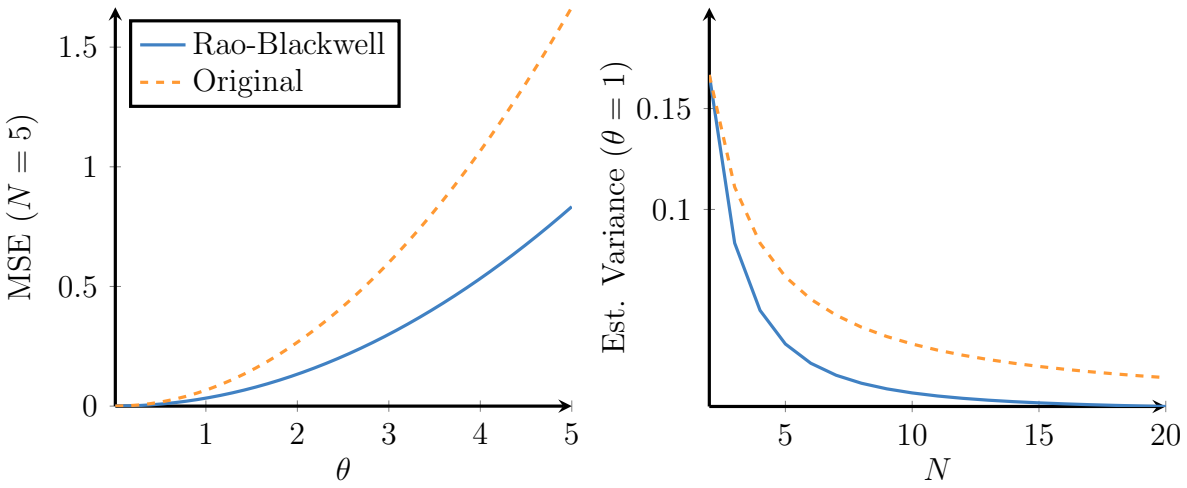
$$\begin{aligned} \mathbb{E}\left[X_1 | \max_i X_i = t \right] = & \underbrace{\mathbb{E}\left[X_1 | \max_i X_i = t, X_1 = \max_i X_i \right]}_{t} \underbrace{\mathsf{P}[X_1 = \max_i X_i]}_{\frac{1}{N}} \\ & + \underbrace{\mathbb{E}\left[X_1 | \max_i X_i = t, X_1 < \max_i X_i \right]}_{\frac{t}{2}} \underbrace{\mathsf{P}[X_1 < \max_i X_i]}_{1-\frac{1}{N}}. \end{aligned} \tag{16.65}$$

As a result

$$\widetilde{\theta}_{\text{rb}}(t) = \frac{N+1}{N} t. \tag{16.66}$$

It is easy to verify that the variance of $\widehat{\theta}$ is $\theta^2/3N$, whereas we will state without a proof that the variance of the Rao-Blackwell estimator is $\frac{\theta^2}{N(N+1)}$.

Next we see the variance of the original estimator, $\widehat{\theta}$, and its Rao-Blackwellisation, $\widetilde{\theta}_{\text{rb}}$, with respect to θ for $N = 5$ (left) and for different values of N with $\theta = 1$ (right).



We see that Rao-Blackwellisation lead to a significant improvement. ♡

Remark. Is the Rao-Blackwell estimator of Equation (16.66) a UMVUE? We could check whether it is efficient. However, for the uniform distribution, $U(0, \theta)$, **the Fisher information is not defined**, so we cannot use the Cramér-Rao bound.

16.5.2 Can we do better?*

The Rao-Blackwell estimator is a (possibly/hopefully) better estimator than the original one and if the original estimator is unbiased, so is its Rao-Blackwellisation. But can we do better? Is it possible that the Rao-Blackwell estimator is a UMVUE? Gaili and Meilijson showed that in some cases, the Rao-Blackwell estimator is not a UMVUE and one can find a better estimator[6]. Under an additional condition known as *completeness* the Rao-Blackwell estimator is a UMVUE — this is the celebrated Lehmann-Scheffé theorem that we will state in this section.

Earlier we defined various statistics, $T = T(X_1, \ldots, X_N)$, which of course are random variables and often we known their distributions (which depend on θ). Before we proceed, let us list some statistics and their distributions.

Sample dist.	Statistic	Statistic dist.
$X_i \overset{\text{iid}}{\sim} \text{Ber}(\theta)$	$T = \sum_{i=1}^N X_i$	$T \sim \text{Binom}(N, \theta)$
$X_i \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$	$T = \sum_{i=1}^N X_i$	$T \sim \text{Poisson}(N\lambda)$
$X_i \overset{\text{iid}}{\sim} \text{Exp}(\lambda)$	$T = \sum_{i=1}^N X_i$	$T \sim \Gamma(N, \lambda)$
$X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$	$T = \sum_{i=1}^N X_i$	$T \sim \mathcal{N}(N\mu, N\sigma^2)$
$X_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$	$T = \sum_{i=1}^N X_i^2$	$T \sim \sigma^2 \chi^2_N$ (Note*)
$X_i \overset{\text{iid}}{\sim} U(0, \theta)$	$T = \max_{i=1\ldots,N} X_i$	$p_T(t) = N\theta^{-N} t^{N-1}$
	$T = \sum_{i=1}^N X_i$	Irwin-Hall distribution**
$X_i \overset{\text{iid}}{\sim} \Gamma(\alpha, \beta)$	$T = \sum_{i=1}^N \log X_i$	Ref.***

* Chi-squared distribution with N degrees of freedom.

** It is a common mistake to assume that $T \sim U(0, N\theta)$

*** See S. A. Saberali and N. C. Beaulieu, "Calculating the distribution of sums of log-gamma random variables," 2012 IEEE ICC, 2012, pp. 2416-2421.

A word of warning before we state the definition of completeness:

[6] Tal Galili and Isaac Meilijson (2016) An Example of an Improvable Rao–Blackwell Improvement, Inefficient Maximum Likelihood Estimator, and Unbiased Generalized Bayes Estimator, The American Statistician, 70:1, 108-113, DOI: 10.1080/00031305.2015.1100683.

> Many definitions in statistics are intuitive, but unfortunately "complete statistic" is not one of them.
>
> From https://www.statisticshowto.com/complete-statistic/.

Indeed, completeness is a technical condition, which is understood through its use in the Lehmann-Scheffé theorem, Basu's theorem and other results.

Let us now give the definition of completeness. Let $T(X_1, \ldots, X_N)$ be a statistic with a pdf $p_T(t;\theta)$[7] and p_T belongs to a family of parametric distributions $\mathcal{P}_\theta = \{p_T(\,\cdot\,;\theta), \theta \in \Theta\}$. We say that $\mathcal{P}_\theta$ is a **complete family** if $\mathbb{E}[g(T)] = 0$ for all (measurable) functions g and for all $\theta \in \Theta$ implies that $g(T) = 0$ almost surely[8] for all θ. Equivalently, we can say that T is a **complete statistic**[9].

Example 15 (Complete statistic for Bernoulli). Suppose $X_1, \ldots, X_N \overset{\text{iid}}{\sim} \text{Ber}(\theta)$, where $\theta \in (0,1)$ is an unknown parameter. Consider the statistic $T = \sum_{i=1}^N X_i$, which follows the binomial distribution, $\text{Binom}(N, \theta)$ (see Handout 13). Suppose that $\mathbb{E}[g(T)] = 0$ for all $\theta \in [0,1]$, that is

$$\begin{aligned}
&\sum_{t=0}^{N} g(t)\binom{N}{t}\theta^t(1-\theta)^{N-t} = 0, \ \forall\theta \in (0,1) \\
\Leftrightarrow &(1-\theta)\sum_{t=0}^{N} g(t)\binom{N}{t}\theta^t(1-\theta)^{-t} = 0, \ \forall\theta \in (0,1)
\end{aligned}$$

[7] For example, if we have $X_1, \ldots, X_N \overset{\text{iid}}{\sim} \mathcal{N}(\mu, 1)$ and we define $T = \frac{1}{N}\sum_{i=1}^N X_i$, then we know that $T \sim \mathcal{N}(\mu, 1/N)$ so $p_T(t;\mu) = (\sqrt{2\pi\sigma^2})^{-1/2}\exp\left(-\frac{N(t-\mu)^2}{2}\right)$. Note that the distribution of T depends parametrically on the parameter μ.

[8] Detail: here "almost surely" is meant with respect to P_θ, where P_θ is the probability measure that corresponds to $p(\,\cdot\,;\theta)$, for all $\theta \in \Theta$.

[9] An alternative statement of the definition of completeness is: "T is complete if there are no nontrivial unbiased estimators of 0 based on T."

$$\Leftrightarrow \sum_{t=0}^{N} g(t)\binom{N}{t}\left(\frac{\theta}{1-\theta}\right)^t = 0,\ \forall\theta \in (0,1) \tag{16.67}$$

Define $z = \frac{\theta}{1-\theta} > 0$; then we have

$$\underbrace{\sum_{t=0}^{N} g(t)\binom{N}{t} z^t}_{\text{polynomial in } z} = 0,\ \forall z > 0, \tag{16.68}$$

and the only way that a polynomial is zero for all $z > 0$ is that it is the zero polynomial, i.e., $g(t) = 0$ for all t, which shows that T is complete. ♡

Example 16 (Complete statistic for Poisson). Suppose $X_1, \ldots, X_N \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$ and define $T = \sum_{i=1}^{N} X_i$. We know that $T \sim \text{Poisson}(N\lambda)$[10]. Recall that T takes values in $\mathbb{N}$, so a function g of T is essentially a sequence, $(g_t)_{t\in\mathbb{N}}$. Now suppose g is such that $\mathbb{E}[g(T)] = 0$ for all λ; equivalently

$$\sum_{t=0}^{\infty} g_t \frac{(N\lambda)^t e^{-N\lambda}}{t!} = 0,\ \forall\lambda > 0 \Leftrightarrow \underbrace{\sum_{t=0}^{\infty} g_t \frac{(N\lambda)^t}{t!}}_{s(\lambda)} = 0,\ \forall\lambda > 0. \tag{16.69}$$

By taking $\lambda \to 0^+$ we observe that $g_0 = 0$. Then, by differentiating s with respect to λ and taking $\lambda \to 0^+$ we observe that $g_0 = 0$. Recursively, $g_t = 0$ for all $t \in \mathbb{N}$, which shows that T is a complete statistic. ♡

Example 17 (Complete statistic for Normal). Suppose $X_1, \ldots, X_N \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, where σ^2 is known and μ is an unknown parameter. Define $T = \sum_{i=1}^{N} X_i$, and $T \sim \mathcal{N}(N\mu, N\sigma^2)$. Suppose g is a measurable function such that $\mathbb{E}[g(T)] = 0$ for all $\mu \in \mathbb{R}$. From LotUS

$$\begin{aligned} &\int_{-\infty}^{\infty} g(t) p_T(t; \mu, \sigma^2)\mathrm{d}t = 0,\ \forall\mu \in \mathbb{R} \\ \Leftrightarrow\ &\int_{-\infty}^{\infty} g(t) \exp\left(-\frac{(t - N\mu)^2}{2N\sigma^2}\right)\mathrm{d}t = 0,\ \forall\mu \in \mathbb{R} \end{aligned}$$

[10] **Exercise 10 (☕☕).** Prove that if $X \sim \text{Poisson}(\lambda)$ and $X' \sim \text{Poisson}(\lambda')$, for some $\lambda, \lambda' > 0$, then $X + X' \sim \text{Poisson}(\lambda + \lambda')$. Hint: Use the fact that the pdf of $X + X'$ is the convolution of the pdfs of X and X'.

$$\Leftrightarrow \int_{-\infty}^{\infty} g(t) \exp\left(-\frac{t^2}{2N\sigma^2}\right) \cancel{\exp\left(-\frac{N^2\mu^2}{2N\sigma^2}\right)} \exp\left(\frac{2N\mu t}{2N\sigma^2}\right) \mathrm{d}t = 0, \ \forall \mu \in \mathbb{R}$$

$$\Leftrightarrow \int_{-\infty}^{\infty} g(t) \exp\left(-\frac{t^2}{2N\sigma^2}\right) \exp\left(\frac{\mu}{\sigma^2} t\right) \mathrm{d}t = 0, \ \forall \mu \in \mathbb{R} \tag{16.70}$$

Recall that the two-sided or *bilateral* Laplace transform of a function $f : \mathbb{R} \to \mathbb{R}$ is $\mathcal{B}\{f(t)\}(s) = \int_{-\infty}^{\infty} f(t) e^{-st} \mathrm{d}t$, for $s \in \mathbb{C}$, if the integral converges. That said, Equation (16.70) can be written equivalently as

$$\mathcal{B}\left\{ g(t) \exp\left(-\frac{t^2}{2N\sigma^2}\right) \right\} \left(\frac{\mu}{\sigma^2}\right) = 0, \forall \mu \in \mathbb{R}, \tag{16.71}$$

that is,

$$\mathcal{B}\left\{ g(t) \exp\left(-\frac{t^2}{2N\sigma^2}\right) \right\} (s) = 0, \forall s \in \mathbb{R}. \tag{16.72}$$

In turn this means that $g(t) \exp\left(-\frac{t^2}{2N\sigma^2}\right) = 0$ for almost all t, so $g = 0$ almost surely, which shows that T is complete. ♡

Let us have a look at the following simple implication of completeness, which motivates its definition and brings us closer to the Lehmann-Scheffé Theorem.

Proposition 16.7 (Uniqueness) *If $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are two unbiased estimators that depend on a complete and sufficient statistic, they are P_θ-almost surely equal to one another for some $\theta \in \Theta$, i.e., $\mathsf{P}_\theta[\widehat{\theta}_1 = \widehat{\theta}_2] = 1$, for some $\theta \in \Theta$.*

Proof: Suppose we have two sufficient, complete, unibased estimators, $\widehat{\theta}_1$ and $\widehat{\theta}_2$ that depend on a complete sufficient statistic $T = T(X_1, \ldots, X_N)$. It is $\mathbb{E}[\widehat{\theta}_1] = \theta$ and $\mathbb{E}[\widehat{\theta}_2] = \theta$. Suppose that the two estimators are essentially different in the sense

$$\mathsf{P}_\theta[\widehat{\theta}_1(T) = \widehat{\theta}_2(T)] \neq 0, \ \forall \theta \in \Theta. \tag{16.73}$$

Now define $g(T) = \widehat{\theta}_1(T) - \widehat{\theta}_2(T)$ for which $\mathbb{E}[g(T)] = 0$, but $\mathsf{P}_\theta[g(T) = 0] \neq 0$, meaning that T is not complete. ■

Theorem 16.8 (Lehmann-Sheffé) *Let $X_1, \ldots, X_N$ be a (iid) sample with distribution $p(\cdot;\theta)$, where $\theta \in \Theta$ is an unknown parameter. Let $T = T(X_1, \ldots, X_N)$ is a sufficient and complete statistic for θ. Let $\widetilde{\theta}(T)$ be an unbiased estimator of θ — i.e., $\mathbb{E}[\widetilde{\theta}(T) \mid \theta] = \theta$. Then, $\widetilde{\theta}(T)$ is the unique UMVUE of θ.*

Proof: Suppose $\widehat{\theta}$ be an unbiased estimator of θ. The Rao-Blackwell theorem implies that the estimator $\widetilde{\theta}_{\rm rb}(T) = \mathbb{E}[\widehat{\theta} \mid T]$ with $\mathrm{var}(\widetilde{\theta}_{\rm rb}) \leq \mathrm{var}(\widehat{\theta})$. Moreover, since $\widetilde{\theta}$ and $\widetilde{\theta}_{\rm rb}$ are unbiased, then $\mathbb{E}[\widetilde{\theta}(T) - \widetilde{\theta}_{\rm rb}(T)] = 0$. Since T is complete, $\widetilde{\theta}(T) = \widetilde{\theta}_{\rm rb}(T)$ almost surely, so they have the same variance, that is,

$$\mathrm{var}[\widetilde{\theta}(T)] = \mathrm{var}[\widetilde{\theta}_{\rm rb}(T)] \leq \mathrm{var}[\widehat{\theta}], \tag{16.74}$$

and this holds for any unbiased estimator, $\widehat{\theta}$. This proves that $\widetilde{\theta}$ is a UMVUE. The uniqueness follows from Proposition 16.7. ■

Remark. In many cases, it is easy to determine a complete and sufficient statistic. The Lehmann-Scheffé suggests that if we can use a sufficient statistic to define an unbiased estimator, it will be a UMVUE.

Example 18. In Exercise 7 and Example 16 we showed that $T = \sum_{i=1}^{N} X_i$ is a sufficient and complete statistic for $X_1, \ldots, X_N \overset{\rm iid}{\sim} \mathrm{Poisson}(\lambda)$. The estimator $\widehat{\lambda} = \frac{1}{N}T$ is unbiased (why?), therefore, by the Lehmann-Scheffé theorem, it is the unique UMVUE. ♡

16.6 References

1. E.L. Lehmann and G. Casella, Theory of Point Estimation, Springer, Second Ed., 1998

2.