Course Notes for Math 162: Mathematical Statistics
The Cramér-Rao Inequality

Adam Merberg and Steven J. Miller

May 8, 2008

Abstract

The Cramér-Rao Inequality provides a lower bound for the variance of an unbiased estimator of a parameter. It allows us to conclude that an unbiased estimator is a minimum variance unbiased estimator for a parameter. In these notes we prove the Cramér-Rao inequality and examine some applications. We conclude with a discussion of a probability distribution for which the Cramér-Rao inequality provides no useful information.

Contents

1 Description of the Problem

Point estimation is the use of a statistic to estimate the value of some parameter of a population having a particular type of density. The statistic we use is called the **point estimator** and its value is the **point estimate**. A desirable property for a point estimator $\hat{\Theta}$ for a parameter θ is that the expected value of $\hat{\Theta}$ is θ. If $\hat{\Theta}$ is a random variable with density f and values $\hat{\theta}$, this is equivalent to saying

$$\mathbb{E}[\hat{\Theta}] \;=\; \int_{-\infty}^{\infty} \hat{\theta} f(\hat{\theta})\, d\hat{\theta} \;=\; \theta.$$

An estimator having this property is said to be **unbiased**.

Often in the process of making a point estimate, we must choose among several unbiased estimators for a given parameter. Thus we need to consider additional criteria to select one of the estimators for use. For example, suppose that $X_1, X_2, \ldots, X_m$ are a random sample from a normal population of mean μ and variance σ^2 with n an odd integer, $m = 2n + 1$. Let the density of this function be given by $f(x; \mu, \sigma^2)$. Suppose we wish to estimate the mean, μ, of this population. It is well-known that both the sample mean and the sample median are unbiased estimators of the mean (c.f. [MM]).

Often, we will take the unbiased estimator having the smallest variance. The variance of $\hat{\Theta}$ is, as for any random variable, the second moment about the mean:

$$\mathrm{var}(\hat{\Theta}) \;=\; \int_{-\infty}^{\infty} (\hat{\theta} - \mu_{\hat{\Theta}})^2 f(\hat{\theta})\, d\hat{\theta}.$$

Here, $\mu_{\hat{\Theta}}$ is the mean of the random variable $\hat{\Theta}$, which is θ in the case of an unbiased estimator. Choosing the estimator with the smaller variance is a natural thing to do, but by no means is it the only possible choice. If two estimators have

the same expected value, then while their average values will be equal the estimator with greater variance will have larger fluctuations about this common value.

An estimator with a smaller variance is said to be **relatively more efficient** because it will tend to have values that are concentrated more closely about the correct value of the parameter, thus it allows us to be more confident that our estimate will be as close to the actual value as we would like. Furthermore, the quantity

$$\frac{\operatorname{var}\hat{\Theta}_1}{\operatorname{var}\hat{\Theta}_2}$$

is used as a measure of the efficiency of $\hat{\Theta}_2$ relative to $\hat{\Theta}_1$ [MM]. We hope to maximize efficiency by minimizing variance.

In our example, the mean of the population has variance $\sigma^2/m = \sigma^2/(2n+1)$. If the **population median** is $\tilde{\mu}$, that is $\tilde{\mu}$ is such that

$$\int_{-\infty}^{\tilde{\mu}} f(x;\mu,\sigma^2)\,dx \;=\; \frac{1}{2},$$

then, according to [MM], the sampling distribution of the median is approximately normal with mean $\tilde{\mu}$ and variance

$$\frac{1}{8n\cdot f(\tilde{\mu})^2}.$$

Since the normal distribution of our example is symmetric, we must have $\tilde{\mu} = \mu$, which makes it easy to show that $f(\tilde{\mu}) = 1/\sqrt{2\pi\sigma^2}$. The variance of the sample median is therefore $\pi\sigma^2/4n$.

Certainly, in our example, the mean has the smaller variance of the two estimators, but we would like to know whether an estimator with smaller variance exists. More precisely, it would be very useful to have a lower bound on the variance of an unbiased estimator. Clearly, the variance must be non-negative[1], but it would be useful to have a less trivial lower bound. The Cramér-Rao Inequality is a theorem that provides such a bound under very general conditions. It does not, however, provide any assurance that any estimator exists that has the minimum variance allowed by this bound.

2 The Cramér-Rao Inequality

The Cramér-Rao Inequality provides us with a lower bound on the variance of an unbiased estimator for a parameter.

Cramér-Rao Inequality. *Let $f(x;\theta)$ be a probability density with continuous parameter θ. Let $X_1,\ldots,X_n$ be independent random variables with density $f(x;\theta)$, and let $\hat{\Theta}(X_1,\ldots,X_n)$ be an unbiased estimator of θ. Assume that $f(x;\theta)$ satisfies two conditions:*

1. *We have*

$$\frac{\partial}{\partial\theta}\left[\int\cdots\int \hat{\Theta}(x_1,\ldots,x_n)\prod_{i=1}^{n} f(x_i;\theta)\,dx_i\right] \;=\; \int\cdots\int \hat{\Theta}(x_1,\ldots,x_n)\frac{\partial\prod_{i=1}^{n} f(x_i;\theta)}{\partial\theta}\,dx_1\cdots dx_n, \qquad (2.1)$$

   Conditions under which this holds are reproduced from [HH] in Appendix A.

2. *For each θ, the variance of $\hat{\Theta}(X_1,\ldots,X_n)$ is finite.*

Then

$$\operatorname{var}(\hat{\Theta}) \;\geq\; \frac{1}{n\,\mathbb{E}\left[\left(\frac{\partial\log f(x;\theta)}{\partial\theta}\right)^2\right]}, \qquad (2.2)$$

where $\mathbb{E}$ denotes the expected value with respect to the probability density function $f(x;\theta)$.

Proof. We prove the theorem as in [CaBe]. Let $\widehat{\Theta}(\vec{X}) = \widehat{\Theta}(X_1,\ldots,X_n)$. We assume that our estimator depends only on the sample values $X_1,\ldots,X_n$ and is independent of θ. Since $\widehat{\Theta}(\vec{X})$ is unbiased as an estimator for θ, we have $\mathbb{E}[\widehat{\Theta}] = \theta$. From this we have:

$$\begin{aligned} 0 &= \mathbb{E}[\hat{\Theta}-\theta] \\ &= \int\cdots\int \Big(\widehat{\Theta}(x_1,\ldots,x_n)-\theta\Big)\,f(x_1;\theta)\cdots f(x_n;\theta)\,dx_1\cdots dx_n. \end{aligned}$$

[1] It is possible for the variance of an estimator to be zero. Consider the following case: we always estimate the mean to be 0, no matter what sample values we observe. This is a terrific estimate if the mean happens to be 0, and is a poor estimate otherwise. Note, however, that the variance of our estimator is zero!

We denote $f(x_1;\theta)\cdots f(x_n;\theta)$ by $\phi(\vec{x};\theta)$ and $dx_1\cdots dx_n$ by $d\vec{x}$. We now have

$$0 \;=\; \int\cdots\int \left(\widehat{\Theta}(\vec{x})-\theta\right)\phi(\vec{x};\theta)\,d\vec{x}.$$

Differentiating both sides of this equation and using the the assumption of equation (2.1).

$$\begin{aligned}
0 &= \frac{\partial}{\partial\theta}\int\cdots\int \left(\widehat{\Theta}(\vec{x})-\theta\right)\phi(\vec{x};\theta)\,d\vec{x} \\
&= \int\cdots\int \frac{\partial}{\partial\theta}\left[\left(\widehat{\Theta}(\vec{x})-\theta\right)\phi(\vec{x};\theta)\right]\,d\vec{x} \\
&= \int\cdots\int \left[\frac{\partial}{\partial\theta}\left(\widehat{\Theta}(\vec{x})-\theta\right)\right]\phi(\vec{x};\theta)\,d\vec{x} + \int\cdots\int (\widehat{\Theta}(\vec{x})-\theta)\frac{\partial\phi(\vec{x};\theta)}{\partial\theta}\,d\vec{x}. \qquad (2.3)
\end{aligned}$$

Since $\widehat{\Theta}(\vec{X})$ is independent of θ, we have $\frac{\partial}{\partial\theta}(\widehat{\Theta}(\vec{x})-\theta) = 0-1=-1$, whence

$$\begin{aligned}
0 &= \int\cdots\int -\phi(\vec{x};\theta)\,d\vec{x} + \int\cdots\int (\widehat{\Theta}(\vec{x})-\theta)\frac{\partial\phi(\vec{x};\theta)}{\partial\theta}\,d\vec{x} \\
0 &= -1 + \int\cdots\int (\widehat{\Theta}(\vec{x})-\theta)\frac{\partial\phi(\vec{x};\theta)}{\partial\theta}\,d\vec{x}. \qquad (2.4)
\end{aligned}$$

We now expand $\frac{\partial\phi(\vec{x};\theta)}{\partial\theta}$:

$$\begin{aligned}
\frac{\partial\phi(\vec{x};\theta)}{\partial\theta} &= \frac{\partial}{\partial\theta}\left(f(x_1;\theta)\cdots f(x_n;\theta)\right) \\
&= \frac{\partial f(x_1;\theta)}{\partial\theta}\cdot f(x_2;\theta)\cdots f(x_n;\theta) + \frac{\partial f(x_2;\theta)}{\partial\theta}\cdot f(x_1;\theta)f(x_3;\theta)\cdots f(x_n;\theta)+\cdots \\
&\quad + \frac{\partial f(x_n;\theta)}{\partial\theta}\cdot f(x_1;\theta)\cdots f(x_{n-1}). \qquad (2.5)
\end{aligned}$$

We can simplify the above by using the identity $\frac{\partial\log g}{\partial\theta} = \frac{1}{g}\frac{\partial g}{\partial\theta}$. We multiply the i^{th} summand by $f(x_i;\theta)/f(x_i;\theta)$, and find

$$\begin{aligned}
\frac{\partial\phi(\vec{x};\theta)}{\partial\theta} &= \sum_{i=1}^{n}\frac{\partial\log f(x_i;\theta)}{\partial\theta}f(x_1;\theta)\cdots f(x_n;\theta) \\
&= \phi(\vec{x};\theta)\sum_{i=1}^{n}\frac{\partial\log f(x_i;\theta)}{\partial\theta}. \qquad (2.6)
\end{aligned}$$

Combining our expansion for $\partial\phi(\vec{x};\theta)/\partial\theta$ with (2.4) yields

$$\begin{aligned}
1 &= \int\cdots\int (\widehat{\Theta}(\vec{x})-\theta)\left[\phi(\vec{x};\theta)\sum_{i=1}^{n}\frac{\partial\log f(x_i;\theta)}{\partial\theta}\right]\,d\vec{x} \\
&= \int\cdots\int \left[(\widehat{\Theta}(\vec{x})-\theta)\cdot\phi(\vec{x};\theta)^{1/2}\right]\left[\phi(\vec{x};\theta)^{1/2}\cdot\sum_{i=1}^{n}\frac{\partial\log f(x_i;\theta)}{\partial\theta}\right]\,d\vec{x}. \qquad (2.7)
\end{aligned}$$

The above argument looks strange: we have taken the nice factor $\phi(\vec{x};\theta)$ and written it as $\phi(\vec{x};\theta)^{1/2}\cdot\phi(\vec{x};\theta)^{1/2}$. The reason we do this is so that we may apply the Cauchy-Schwarz Inequality (the statement of this inequality, as well as a proof, are provided in Appendix B). There are other reasons[2] why it is natural to split $\phi(\vec{x};\theta)$ in two.

[2]Note that the random variable $(\widehat{\Theta}(\vec{x})-\theta)^2$ is almost surely not integrable with respect to $d\vec{x}$. For example, let X be a random variable with probability distribution $f(x)$; further, assume all moments of f are finite. Then

$$\int_{-\infty}^{\infty} x^2 f(x)dx \;<\; \infty;$$

however,

$$\int_{-\infty}^{\infty} x^2 dx \;=\; \infty.$$

While we don't expect x^2 to be integrable, we do expect $x^2 f(x)$ to be integrable; we need the density $f(x)$ to give decay and ensure convergence. It is the same situation here; $(\widehat{\Theta}(\vec{x})-\theta)^2$ is not integrable, but multiplying by $\phi(\vec{x};\theta)$ will give something that (in general) is integrable.

We square both sides of (2.7), obtaining

$$1 \;=\; \left(\int \cdots \int \left[(\widehat{\Theta}(\vec{x}) - \theta) \cdot \phi(\vec{x};\theta)^{1/2} \right] \left[\phi(\vec{x};\theta)^{1/2} \cdot \sum_{i=1}^{n} \frac{\partial \log f(x_i;\theta)}{\partial \theta} \right] d\vec{x} \right)^2. \tag{2.8}$$

We now apply the Cauchy-Schwarz Inequality to (2.8). Thus

$$1 \quad \leq \quad \int \cdots \int \left(\widehat{\Theta}(\vec{x}) - \theta \right)^2 \;\cdot\; \phi(\vec{x};\theta) d\vec{x} \cdot \int \cdots \int \left(\sum_{i=1}^{n} \frac{\partial \log f(x_i;\theta)}{\partial \theta} \right)^2 \phi(\vec{x};\theta) d\vec{x}. \tag{2.9}$$

There are two multiple integrals to evaluate on the right hand side. The first multiple integral is just the definition of the variance of the estimator $\widehat{\Theta}$, which we denote by $\mathrm{var}(\widehat{\Theta})$. Thus (2.8) becomes

$$1 \;\leq\; \mathrm{var}(\widehat{\Theta}) \cdot \int \cdots \int \left(\sum_{i=1}^{n} \frac{\partial \log f(x_i;\theta)}{\partial \theta} \right)^2 \phi(\vec{x};\theta) d\vec{x}. \tag{2.10}$$

To finish the proof of the Cramér-Rao Inequality, it suffices to show

$$\int \cdots \int \left(\sum_{i=1}^{n} \frac{\partial \log f(x_i;\theta)}{\partial \theta} \right)^2 \phi(\vec{x};\theta) d\vec{x} \;=\; n\mathbb{E}\left[\left(\frac{\partial \log f(x;\theta)}{\partial \theta} \right)^2 \right]. \tag{2.11}$$

This is because if we can prove (2.11), simple division will yield the Cramér-Rao Inequality from (2.10). We now prove (2.11).

We have

$$\begin{aligned} \int \cdots \int \left(\sum_{i=1}^{n} \frac{\partial \log f(x_i;\theta)}{\partial \theta} \right)^2 \phi(\vec{x};\theta) d\vec{x} &= \int \cdots \int \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial \log f(x_i;\theta)}{\partial \theta} \frac{\partial \log f(x_j;\theta)}{\partial \theta} \phi(\vec{x};\theta) d\vec{x} \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} \int \cdots \int \frac{\partial \log f(x_i;\theta)}{\partial \theta} \frac{\partial \log f(x_j;\theta)}{\partial \theta} \phi(\vec{x};\theta) d\vec{x} \\ &= I_1 + I_2, \end{aligned} \tag{2.12}$$

where

$$\begin{aligned} I_1 &= \int \cdots \int \sum_{i=1}^{n} \left(\frac{\partial \log f(x_i;\theta)}{\partial \theta} \right)^2 \phi(\vec{x};\theta) d\vec{x} \\ I_2 &= \sum_{\substack{1 \leq i,j \leq n \\ i \neq j}} \int \cdots \int \frac{\partial \log f(x_i;\theta)}{\partial \theta} \frac{\partial \log f(x_j;\theta)}{\partial \theta} \phi(\vec{x};\theta) d\vec{x}. \end{aligned} \tag{2.13}$$

The proof is completed by showing $I_1 = n\mathbb{E}\left[\left(\frac{\partial \log f(x;\theta)}{\partial \theta} \right)^2 \right]$ and $I_2 = 0$.

We have

$$\begin{aligned} I_1 &= \int \cdots \int \sum_{i=1}^{n} \left(\frac{\partial \log f(x_i;\theta)}{\partial \theta} \right)^2 \phi(\vec{x};\theta) d\vec{x} \\ &= \sum_{i=1}^{n} \int \left(\frac{\partial \log f(x_i;\theta)}{\partial \theta} \right)^2 f(x_i;\theta) dx_i \cdot \int \cdots \int \prod_{\substack{\ell=1 \\ \ell \neq i}}^{n} f(x_\ell;\theta) dx_\ell \\ &= \sum_{i=1}^{n} \int \left(\frac{\partial \log f(x_i;\theta)}{\partial \theta} \right)^2 f(x_i;\theta) dx_i \cdot 1^{n-1} \\ &= \sum_{i=1}^{n} \mathbb{E}\left[\left(\frac{\partial \log f(x_i;\theta)}{\partial \theta} \right)^2 \right] \\ &= n\mathbb{E}\left[\left(\frac{\partial \log f(x_i;\theta)}{\partial \theta} \right)^2 \right]. \end{aligned} \tag{2.14}$$

In the above calculation, we used the fact that $f(x_i;\theta)$ is a probability density, and therefore integrates to one. In the final expected values, x_i is a dummy variable, and we may denote these n expected values with a common symbol.

We now turn to the analysis of I_2. In obvious notation, we may write

$$I_2 \;=\; \sum_{\substack{1\le i,j\le n\\ i\neq j}} I_2(i,j). \tag{2.15}$$

To show $I_2 = 0$ it suffices to show each $I_2(i,j) = 0$, which we now proceed to do. Note

$$\begin{aligned} I_2(i,j) &= \int\cdots\int \frac{\partial \log f(x_i;\theta)}{\partial\theta}\frac{\partial \log f(x_j;\theta)}{\partial\theta}\phi(\vec{x};\theta)d\vec{x} \\ &= \int \frac{\partial \log f(x_i;\theta)}{\partial\theta} f(x_i;\theta)dx_i \cdot \int \frac{\partial \log f(x_j;\theta)}{\partial\theta}dx_j \cdot \int\cdots\int \prod_{\substack{\ell=1\\ \ell\neq i,j}}^{n} f(x_\ell;\theta)dx_\ell \\ &= \int \frac{\partial \log f(x_i;\theta)}{\partial\theta} f(x_i;\theta)dx_i \cdot \int \frac{\partial \log f(x_j;\theta)}{\partial\theta}dx_j \cdot 1^{n-2} \\ &= \mathbb{E}\left[\frac{\partial \log f(x_i;\theta)}{\partial\theta}\right]\cdot\mathbb{E}\left[\frac{\partial \log f(x_j;\theta)}{\partial\theta}\right]; \end{aligned} \tag{2.16}$$

however, each of these two expected values is zero! To see this, note

$$1 \;=\; \int f(x;\theta)dx. \tag{2.17}$$

If we differentiate both sides of (2.17) with respect to θ, we find

$$\begin{aligned} 0 &= \int \frac{\partial f(x;\theta)}{\partial\theta}dx \\ &= \int \frac{1}{f(x;\theta)}\frac{\partial f(x;\theta)}{\partial\theta} f(x;\theta)dx \\ &= \int \frac{\partial \log f(x;\theta)}{\partial\theta} f(x;\theta)dx \;=\; \mathbb{E}\left[\frac{\partial \log f(x;\theta)}{\partial\theta}\right]. \end{aligned} \tag{2.18}$$

This shows $I_2(i,j) = 0$, which completes the proof. □

An estimator for which equality holds in (2.2) is called a **minimum variance unbiased estimator** or simply a **best unbiased estimator**. The expected value in the Cramér-Rao Inequality is called the **information number** or the **Fisher information** of the sample.

We notice that the theorem makes no statement about whether equality holds for any particular estimator $\hat{\Theta}$. Indeed, in Appendix D, we give an example in which the information is infinite, and the bound provided is therefore $\mathrm{var}(\hat{\Theta}) \ge 0$, which is trivial.

3 Examples and Exercises

Example 3.1. We first consider estimating the parameter of an exponential population based on a sample of size $m = 2n + 1$. This population has density

$$f(x;\theta) \;=\; \begin{cases} \frac{1}{\theta}e^{-x/\theta} & \text{if } x \ge 0 \\ 0 & \text{if } x < 0. \end{cases} \tag{3.19}$$

We consider two estimators, one based on the sample mean and the other on the sample median. We know from the Central Limit Theorem that for large m, the sample mean will have a normal distribution whose mean is θ, the population mean, and whose variance is $\theta^2/m = \theta^2/(2n+1)$, where θ^2 is the variance computed from the exponential density (the mean and variance are computed in Appendix C).

For large n, the sample median Y_{n+1} has approximately a normal distribution with mean equal to $\tilde{\mu}$, the population median, and variance $1/(8n\cdot f(\tilde{\mu})^2)$ [MM]. By definition, the population median satisfies

$$\int_{-\infty}^{\tilde{\mu}} f(x)\,dx \;=\; \int_0^{\tilde{\mu}} \frac{1}{\theta}e^{-x/\theta}\,dx \;=\; \frac{1}{2}. \tag{3.20}$$

Evaluating the integral, we find

$$\begin{aligned}\frac{1}{2} &= \int_0^{\tilde{\mu}} \frac{1}{\theta} e^{-x/\theta}\, dx \\ &= \left[-e^{-x/\theta}\right]_0^{\tilde{\mu}} \\ &= -e^{-\tilde{\mu}/\theta} + 1 \end{aligned} \tag{3.21}$$

so that $\tilde{\mu} = \theta \log 2$. It follows that $\hat{\Theta} = \frac{Y_{m+1}}{\log 2}$ is an unbiased estimator for θ. The variance of the sample median is

$$\begin{aligned}\left(8n \cdot f(\tilde{\mu})^2\right)^{-1} &= \left(8 \cdot n \left[\frac{1}{\theta} e^{-\theta \log 2/\theta}\right]^2\right)^{-1} \\ &= \frac{\theta^2}{2n}. \end{aligned} \tag{3.22}$$

The variance of $\hat{\Theta} = \frac{Y_{m+1}}{\log 2}$ is therefore

$$\frac{\theta^2}{2n(\log 2)^2}, \tag{3.23}$$

which for $n \geq\ 0$ is larger than the variance $\theta^2/(2n+1)$ of the sample mean.

Noting that the conditions for applying the Cramér-Rao Inequality are satisfied, we now find the bound provided by the theorem. We begin by computing the information of the sample:

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\partial \log f(x;\theta)}{\partial \theta}\right)^2\right] &= \mathbb{E}\left[\left(\frac{\partial \log\left(\frac{1}{\theta} e^{-x/\theta}\right)}{\partial \theta}\right)^2\right] \\ &= \mathbb{E}\left[\left(\frac{\partial}{\partial \theta}\left(-\log\theta - \frac{x}{\theta}\right)\right)^2\right] \\ &= \mathbb{E}\left[\left(-\frac{1}{\theta} + \frac{x}{\theta^2}\right)^2\right] \\ &= \frac{1}{\theta^4}\mathbb{E}\left[(x-\theta)^2\right]. \end{aligned} \tag{3.24}$$

Since θ is the mean of the population, the expected value in the last line is the variance of the population, which is θ^2, so the information is $1/\theta^2$. The bound yielded by the Cramér-Rao Inequality is therefore $\frac{\theta^2}{m}$. We see that this is equal to the variance of the sample mean; the sample mean is a minimum-variance unbiased estimator.

Example 3.2. We also consider a case in which condition (2.1) does not hold, so that the Cramér-Rao Inequality cannot be used. We look at a sampling of size $n = 1$ from a population with a uniform distribution on $[0, \theta]$. This population has density

$$f(x;\theta) = \begin{cases} 1/\theta & \text{if } 0 \leq x \leq \theta \\ 0 & \text{otherwise.} \end{cases} \tag{3.25}$$

It is easily shown that an unbiased estimator for θ is $\hat{\Theta}(X) = 2X$, where X is the single observation from the sample.

We compute the left and right sides of (2.1) for this special case. On the left side, we have

$$\begin{aligned}\frac{\partial}{\partial \theta}\left[\int \hat{\Theta}(x) f(x;\theta)\, dx\right] &= \frac{\partial}{\partial \theta}\left[\int_0^{\theta} 2x \frac{1}{\theta}\, dx\right] \\ &= \frac{\partial}{\partial \theta}\left[\frac{1}{\theta}\int_0^{\theta} 2x\, dx\right] \\ &= \frac{\partial}{\partial \theta}(\theta) \\ &= 1. \end{aligned} \tag{3.26}$$

On the right side, we have

$$\begin{aligned}\int \hat{\Theta}(x)\frac{\partial f(x;\theta)}{\partial\theta}\,dx &= \int_0^\theta 2x\frac{\partial}{\partial\theta}\left(\frac{1}{\theta}\right)dx \\ &= -\frac{1}{\theta^2}\int_0^\theta 2x\,dx \\ &= -1. \end{aligned} \tag{3.27}$$

It is therefore clear that condition (2.1) does not hold, so we cannot assume that the Cramér-Rao Inequality holds. Indeed, we will show that it does not. We first compute the information of the sample:

$$\begin{aligned} n\,\mathbb{E}\left[\left(\frac{\partial \log f(x;\theta)}{\partial\theta}\right)^2\right] &= \mathbb{E}\left[\left(\frac{\partial \log(1/\theta)}{\partial\theta}\right)^2\right] \\ &= \mathbb{E}\left[\left(\frac{\partial(-\log\theta)}{\partial\theta}\right)^2\right] \\ &= \mathbb{E}\left[\left(\frac{-1}{\theta}\right)^2\right] \\ &= \frac{1}{\theta^2}. \end{aligned} \tag{3.28}$$

Therefore, if applicable, the Cramér-Rao Inequality would tell us that $\mathrm{var}(\hat{\Theta}) \geq \theta^2$. We now compute the variance of $\hat{\Theta} = 2X$:

$$\begin{aligned}\mathrm{var}(2X) &= \mathbb{E}\left[(2X)^2\right] - \mathbb{E}\left[2X\right]^2 \\ &= \int_0^\theta (2x)^2\cdot\frac{1}{\theta}\,dx - \theta^2 \\ &= \frac{4\theta^2}{3} - \theta^2 \\ &= \frac{\theta^2}{3}. \end{aligned} \tag{3.29}$$

We therefore see that the Cramér-Rao Inequality need not be satisfied when condition (2.1) is not satisfied. We note that this example has the property that the region in which the density function is nonzero depends on the parameter that we are estimating. In such cases we must be particularly careful as condition (2.1) will often not be satisfied.

Exercise 3.3. *Show that the sample mean is a minimum variance unbiased estimator for the mean of a normal population.*

Exercise 3.4. *Let X be a random variable with a binomial distribution with parameters n and θ. Is $n\cdot\frac{X}{n}\cdot\left(1-\frac{X}{n}\right)$ a minimum variance unbiased estimator for the variance of X?*

A Interchanging Integration and Differentiation

Theorem A.1 (Differentiating under the integral sign)**.** *Let $f(x,t):\mathbb{R}^{n+1}\to\mathbb{R}$ be a function such that for each fixed t the integral*

$$F(t) \;=\; \int_{\mathbb{R}^n} f(t,x)dx_1\cdots dx_n \tag{A.30}$$

exists. For all x, suppose that $\partial f/\partial t$ exists[3], and that there is a continuous Riemann integrable function[4] $g(x)$ such that

$$\left|\frac{f(s,x)-f(t,x)}{s-t}\right| \;\leq\; g(x) \tag{A.31}$$

for all $s\neq t$. Then F is differentiable, and

$$\frac{dF}{dt} \;=\; \int_{\mathbb{R}^n}\frac{\partial f}{\partial t}(t,x)dx_1\cdots dx_n. \tag{A.32}$$

[3]Technically, all we need is that $\partial f/\partial t$ exists for almost all x, i.e., except for a set of measure zero.
[4]This condition can be weakened; it suffices for $g(x)$ to be a Lebesgue integrable function.

The above statement is modified from that of Theorem 4.11.22 of [HH]. See page 518 of [HH] for a proof. We have stated a slightly weaker version (and commented in the footnotes on the most general statement) because these weaker cases often suffice for our applications.

Exercise A.2. *It is not always the case that one can interchange orders of operations. We saw in Example 3.2 a case where we cannot interchange the integration and differentiation. We give an example which shows that we cannot always interchange orders of integration. For simplicity, we give a sequence a_{mn} such that $\sum_m(\sum_n a_{m,n}) \neq \sum_n(\sum_m a_{m,n})$. For $m, n \geq 0$ let*

$$a_{m,n} \;=\; \begin{cases} 1 & \text{if } n = m \\ -1 & \text{if } n = m+1 \\ 0 & \text{otherwise.} \end{cases} \tag{A.33}$$

Show that the two different orders of summation yield different answers. The reason the Fubini Theorem is not applicable here is that $\sum_n \sum_m |a_{mn}| = \infty$.

B The Cauchy-Schwarz Inequality

The Cauchy-Schwarz Inequality is a general result from linear algebra pertaining to inner product spaces. Here we will consider only an application to Riemann integrable functions. For a more thorough treatment of the general form of the inequality, we refer the reader to Chapter 8 of [HK].

Cauchy-Schwarz Inequality. *Let f, g be Riemann integrable real-valued functions of $\mathbb{R}^n$. Then*

$$\left(\int \cdots \int f(x_1, \ldots, x_n) g(x_1, \ldots, x_n)\, dx_1 \cdots dx_n\right)^2 \leq \int \cdots \int f(x_1, \ldots, x_n)^2\, dx_1, \ldots, dx_n \cdot \int \cdots \int g(x_1, \ldots, x_n)^2\, dx_1 \cdots dx_n.$$

Proof. The proof given here is a special case of that given in [HK] (page 377). For notational convenience, we define

$$I(f, g) \;=\; \int \cdots \int f(x_1, \ldots, x_n) g(x_1, \ldots, x_n)\, dx_1 \cdots dx_n.$$

The statement of the theorem is then

$$I(f, g)^2 \leq I(f, f) I(g, g).$$

The following are results of basic properties of integrals, and we leave it as an exercise for the reader to show that they hold:

1. $I(f + g, h) = I(f, h) + I(g, h)$
2. $I(f, g) = I(g, f)$
3. $I(c \cdot f, g) = c \cdot I(f, g)$
4. $I(f, f) \geq 0$ for all f.

In the case that $I(f, f) = 0$ we must also have $I(f, g) = 0$, so the inequality holds in this case. Otherwise, we let

$$h \;=\; g - \frac{I(g, f)}{I(f, f)} \cdot f.$$

We consider $I(h, h)$, noting by property 4 above that this number must be nonnegative. Using the properties verified by the reader, we gave

$$\begin{aligned} 0 \leq I(h, h) &= I\left(g - \frac{I(g, f)}{I(f, f)} \cdot f, g - \frac{I(g, f)}{I(f, f)} \cdot f\right) \\ &= I(g, g) - \frac{I(g, f)}{I(f, f)} \cdot I(f, g) - \frac{I(g, f)}{I(f, f)} \cdot I(g, f) + \frac{I(g, f)^2}{I(f, f)^2} \cdot I(f, f) \\ &= I(g, g) - \frac{I(g, f)^2}{I(f, f)}. \end{aligned} \tag{B.34}$$

It thus follows that

$$I(f, g)^2 \leq I(f, f) I(g, g). \tag{B.35}$$

□

C The Exponential Density

An exponential random variable X with parameter θ and values x has density function

$$f(x;\theta) \;=\; \begin{cases} \frac{1}{\theta}e^{-x/\theta} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases} \tag{C.36}$$

We will compute the mean and variance of this random variable.

The mean of the random variable X is

$$\begin{aligned} \mathbb{E}[x] &= \int_{-\infty}^{\infty} x f(x)\, dx \\ &= \int_{0}^{\infty} x \left(\frac{1}{\theta}e^{-x/\theta}\right) dx. \end{aligned} \tag{C.37}$$

We evaluate the integral by parts to find

$$\begin{aligned} \int_{0}^{\infty} x \left(\frac{1}{\theta}e^{-x/\theta}\right) dx &= \lim_{l\to\infty} \int_{0}^{l} \left(\frac{x}{\theta}e^{-x/\theta}\right) dx \\ &= \lim_{l\to\infty} \left[-xe^{-x/\theta} - \theta e^{-x/\theta}\right]_{0}^{l} \\ &= \lim_{l\to\infty} \left[\left(-le^{-l} - \theta e^{-l/\theta}\right) - \left(-0e^{-0/\theta} - \theta e^{-x/\theta}\right)\right] \\ &= \theta. \end{aligned} \tag{C.38}$$

The mean of X is therefore θ.

To compute the variance, we use the fact that

$$\mathrm{var}(X) = \mathbb{E}[x^2] - \mathbb{E}[x]^2, \tag{C.39}$$

which holds for any random variable having finite variance. We have just found the quantity $\mathbb{E}[x]$, so we need only compute $\mathbb{E}[x^2]$. From the definition of expected value,

$$\begin{aligned} \mathbb{E}[x^2] &= \int_{-\infty}^{\infty} x^2 f(x)\, dx \\ &= \int_{0}^{\infty} x^2 \left(\frac{1}{\theta}e^{-x/\theta}\right) dx. \end{aligned} \tag{C.40}$$

We integrate by parts two times to obtain

$$\begin{aligned} \int_{0}^{\infty} x^2 \left(\frac{1}{\theta}e^{-x/\theta}\right) dx &= \lim_{l\to\infty} \int_{0}^{l} \left(\frac{x^2}{\theta}e^{-x/\theta}\right) dx \\ &= \lim_{l\to\infty} \left[-(x^2 + 2\theta x + 2\theta^2)e^{-x/\theta}\right]_{0}^{l} \\ &= \lim_{l\to\infty} \left[-\left(l^2 + 2\theta l + 2\theta^2\right)e^{-l/\theta} + \left(0^2 + 2\theta\cdot 0 + 2\theta^2\right)e^{-0/\theta}\right] \\ &= 2\theta^2 \end{aligned} \tag{C.41}$$

It therefore follows from equation (C.39) that

$$\mathrm{var}(X) \;=\; \mathbb{E}[x^2] - \mathbb{E}[x]^2 \;=\; 2\theta^2 - \theta^2 \;=\; \theta^2. \tag{C.42}$$

D When the Cramér-Rao Inequality Provides No Information

In this appendix we analyze a probability density where the Cramér-Rao inequality yields the bound that the variance of an unbiased estimator is non-negative; this is a useless bound, as variances must be non-negative.

D.1 An Almost Pareto Density

Consider

$$f(x;\theta) \;=\; \begin{cases} a_\theta \frac{1}{x^\theta \log^3 x} & \text{if } x \ge e \\ 0 & \text{otherwise,} \end{cases} \tag{D.43}$$

where a_θ is chosen so that $f(x;\theta)$ is a probability density function. Thus

$$\int_e^\infty a_\theta \frac{\mathrm{d}x}{x^\theta \log^3 x} \;=\; 1. \tag{D.44}$$

We chose to have $\log^3 x$ in the denominator to ensure that the above integral converges, as does $\log x$ times the integrand; however, the expected value (in the expectation in (2.2)) will not converge.

For example, $1/x \log x$ diverges (its integral looks like $\log\log x$) but $1/x\log^2 x$ converges (its integral looks like $1/\log x$); see pages 62–63 of [Rud] for more on close sequences where one converges but the other does not. This distribution is close to the Pareto distribution (or a power law). Pareto distributions are very useful in describing many natural phenomena; see for example [DM, Ne, NM]. The inclusion of the factor of $\log^{-3} x$ allows us to have the exponent of x in the density function equal 1 *and have the density function defined for arbitrarily large* x; it is also needed in order to apply the Dominated Convergence Theorem to justify some of the arguments below. If we remove the logarithmic factors, then we obtain a probability distribution only if the density vanishes for large x. As $\log^3 x$ is a very slowly varying function, our distribution $f(x;\theta)$ may be of use in modeling data from an unbounded distribution where one wants to allow a power law with exponent 1, but cannot as the resulting probability integral would diverge. Such a situation occurs frequently in the Benford Law literature; see [Hi, Rai] for more details.

We study the variance bounds for unbiased estimators $\widehat{\Theta}$ of θ, and in particular we show that when $\theta = 1$ then the Cramér-Rao inequality yields a useless bound.

Note that it is not uncommon for the variance of an unbiased estimator to depend on the value of the parameter being estimated. For example, consider the uniform distribution on $[0,\theta]$. Let $\overline{X}$ denote the sample mean of n independent observations, and $Y_n = \max_{1\le i\le n} X_i$ be the largest observation. The expected value of $2\overline{X}$ and $\frac{n+1}{n}Y_n$ are both θ (implying each is an unbiased estimator for θ); however, $\mathrm{Var}(2\overline{X}) = \theta^2/3n$ and $\mathrm{Var}(\frac{n+1}{n}Y_n) = \theta^2/n(n+1)$ both depend on θ, the parameter being estimated (see, for example, page 324 of [MM] for these calculations).

Lemma D.1. *As a function of $\theta \in [1,\infty)$, a_θ is a strictly increasing function and $a_1 = 2$. It has a one-sided derivative at $\theta = 1$, and $\frac{\mathrm{d}a_\theta}{\mathrm{d}\theta} \in (0,\infty)$.*

Proof. We have

$$a_\theta \int_e^\infty \frac{\mathrm{d}x}{x^\theta \log^3 x} \;=\; 1. \tag{D.45}$$

When $\theta = 1$ we have

$$a_1 \;=\; \left[\int_e^\infty \frac{\mathrm{d}x}{x\log^3 x}\right]^{-1}, \tag{D.46}$$

which is clearly positive and finite. In fact, $a_1 = 2$ because the integral is

$$\int_e^\infty \frac{\mathrm{d}x}{x\log^3 x} \;=\; \int_e^\infty \log^{-3} x \, \frac{\mathrm{d}\log x}{\mathrm{d}x} \;=\; \frac{-1}{2\log^2 x}\Big|_e^\infty \;=\; \frac{1}{2}; \tag{D.47}$$

though all we need below is that a_1 is finite and non-zero, we have chosen to start integrating at e to make a_1 easy to compute.

It is clear that a_θ is strictly increasing with θ, as the integral in (D.46) is strictly decreasing with increasing θ (because the integrand is decreasing with increasing θ).

We are left with determining the one-sided derivative of a_θ at $\theta = 1$, as the derivative at any other point is handled similarly (but with easier convergence arguments). It is technically easier to study the derivative of $1/a_\theta$, as

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\frac{1}{a_\theta} \;=\; -\frac{1}{a_\theta^2}\frac{\mathrm{d}a_\theta}{\mathrm{d}\theta} \tag{D.48}$$

and

$$\frac{1}{a_\theta} \;=\; \int_e^\infty \frac{\mathrm{d}x}{x^\theta \log^3 x}. \tag{D.49}$$

The reason we consider the derivative of $1/a_\theta$ is that this avoids having to take the derivative of the reciprocals of integrals. As a_1 is finite and non-zero, it is easy to pass to $\frac{\mathrm{d}a_\theta}{\mathrm{d}\theta}|_{\theta=1}$. Thus we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\theta}\frac{1}{a_\theta}\Big|_{\theta=1} &= \lim_{h\to 0^+}\frac{1}{h}\left[\int_e^\infty \frac{\mathrm{d}x}{x^{1+h}\log^3 x} - \int_e^\infty \frac{\mathrm{d}x}{x\log^3 x}\right] \\ &= \lim_{h\to 0^+}\int_e^\infty \frac{1-x^h}{h}\frac{1}{x^h}\frac{\mathrm{d}x}{x\log^3 x}. \end{aligned} \tag{D.50}$$

We want to interchange the integration with respect to x and the limit with respect to h above. This interchange is permissible by the Dominated Convergence Theorem (see Appendix D.3 for details of the justification).

Note

$$\lim_{h\to 0^+}\frac{1-x^h}{h}\frac{1}{x^h} = -\log x; \tag{D.51}$$

one way to see this is to use the limit of a product is the product of the limits, and then use L'Hospital's rule, writing x^h as $e^{h\log x}$. Therefore

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\frac{1}{a_\theta}\Big|_{\theta=1} = -\int_e^\infty \frac{\mathrm{d}x}{x\log^2 x}; \tag{D.52}$$

as this is finite and non-zero, this completes the proof and shows $\frac{\mathrm{d}a_\theta}{\mathrm{d}\theta}|_{\theta=1}\in(0,\infty)$. □

Remark D.2. *We see now why we chose $f(x;\theta) = a_\theta/x^\theta\log^3 x$ instead of $f(x;\theta) = a_\theta/x^\theta\log^2 x$. If we only had two factors of $\log x$ in the denominator, then the one-sided derivative of a_θ at $\theta = 1$ would be infinite.*

Remark D.3. *Though the actual value of $\frac{\mathrm{d}a_\theta}{\mathrm{d}\theta}|_{\theta=1}$ does not matter, we can compute it quite easily. By* (D.52) *we have*

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\theta}\frac{1}{a_\theta}\Big|_{\theta=1} &= -\int_e^\infty \frac{\mathrm{d}x}{x\log^2 x} \\ &= -\int_e^\infty \log^{-2} x\ \frac{\mathrm{d}\log x}{\mathrm{d}x} \\ &= \frac{1}{\log x}\Big|_e^\infty = -1. \end{aligned} \tag{D.53}$$

Thus by (D.48)*, and the fact that $a_1 = 2$ (Lemma D.1), we have*

$$\frac{\mathrm{d}a_\theta}{\mathrm{d}\theta}\Big|_{\theta=1} = -a_1^2\cdot\frac{\mathrm{d}}{\mathrm{d}\theta}\frac{1}{a_\theta}\Big|_{\theta=1} = 4. \tag{D.54}$$

D.2 Computing the Information

We now compute the expected value, $\mathbb{E}\left[\left(\frac{\partial\log f(x;\theta)}{\partial\theta}\right)^2\right]$; showing it is infinite when $\theta = 1$ completes the proof of our main result. Note

$$\begin{aligned}\log f(x;\theta) &= \log a_\theta - \theta\log x + \log\log^{-3} x \\ \frac{\partial\log f(x;\theta)}{\partial\theta} &= \frac{1}{a_\theta}\frac{\mathrm{d}a_\theta}{\mathrm{d}\theta} - \log x. \end{aligned} \tag{D.55}$$

By Lemma D.1 we know that $\frac{\mathrm{d}a_\theta}{\mathrm{d}\theta}$ is finite for each $\theta\geq 1$. Thus

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\partial\log f(x;\theta)}{\partial\theta}\right)^2\right] &= \mathbb{E}\left[\left(\frac{1}{a_\theta}\frac{\mathrm{d}a_\theta}{\mathrm{d}\theta} - \log x\right)^2\right] \\ &= \int_e^\infty\left(\frac{1}{a_\theta}\frac{\mathrm{d}a_\theta}{\mathrm{d}\theta} - \log x\right)^2\cdot a_\theta\frac{\mathrm{d}x}{x^\theta\log^3 x}. \end{aligned} \tag{D.56}$$

If $\theta > 1$ then the expectation is finite and non-zero. We are left with the interesting case when $\theta = 1$. As $\frac{\mathrm{d}a_\theta}{\mathrm{d}\theta}|_{\theta=1}$ is finite and non-zero, for x sufficiently large (say $x\geq x_1$ for some x_1, though by Remark D.3 we see that we may take any $x_1\geq e^4$) we have

$$\left|\frac{1}{a_1}\frac{\mathrm{d}a_\theta}{\mathrm{d}\theta}\Big|_{\theta=1}\right| \leq \frac{\log x}{2}. \tag{D.57}$$

As $a_1 = 2$, we have

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{\partial \log f(x;\theta)}{\partial \theta}\right)^2\right]\Bigg|_{\theta=1} &\geq \int_{x_1}^{\infty} \left(\frac{\log x}{2}\right)^2 a_1 \frac{\mathrm{d}x}{x \log^3 x} \\
&= \int_{x_1}^{\infty} \frac{\mathrm{d}x}{2x \log x} \\
&= \frac{1}{2}\int_{x_1}^{\infty} \log^{-1} x \ \frac{\mathrm{d}\log x}{\mathrm{d}x} \\
&= \frac{1}{2}\log\log x\Big|_{x_1}^{\infty} \\
&= \infty. \qquad \text{(D.58)}
\end{aligned}$$

Thus the expectation is infinite. Let $\widehat{\Theta}$ be *any* unbiased estimator of θ. If $\theta = 1$ then the Cramér-Rao Inequality gives

$$\mathrm{var}(\widehat{\Theta}) \ \geq \ 0, \qquad \text{(D.59)}$$

which provides no information as variances are always non-negative.

D.3 Applying the Dominated Convergence Theorem

We justify applying the Dominated Convergence Theorem in the proof of Lemma D.1. See, for example, [SS] for the conditions and a proof of the Dominated Convergence Theorem.

Lemma D.4. *For each fixed $h > 0$ and any $x \geq e$, we have*

$$\left|\frac{1-x^h}{h}\frac{1}{x^h}\right| \ \leq \ e\log x, \qquad \text{(D.60)}$$

and $\frac{e\log x}{x\log^3 x}$ is positive and integrable, and dominates each $\frac{1-x^h}{h}\frac{1}{x^h}\frac{1}{x\log^3 x}$.

Proof. We first prove (D.60). As $x \geq e$ and $h > 0$, note $x^h \geq 1$. Consider the case of $1/h \leq \log x$. Since $|1-x^h| < 1 + x^h \leq 2x^h$, we have

$$\frac{|1-x^h|}{hx^h} \ < \ \frac{2x^h}{hx^h} \ \leq \ \frac{2}{h} \ \leq \ 2\log x. \qquad \text{(D.61)}$$

We are left with the case of $1/h > \log x$, or $h\log x < 1$. We have

$$\begin{aligned}
|1-x^h| &= |1-e^{h\log x}| \\
&= \left|1 - \sum_{n=0}^{\infty} \frac{(h\log x)^n}{n!}\right| \\
&= h\log x \sum_{n=1}^{\infty} \frac{(h\log x)^{n-1}}{n!} \\
&< h\log x \sum_{n=1}^{\infty} \frac{(h\log x)^{n-1}}{(n-1)!} \ = \ h\log x \cdot e^{h\log x}. \qquad \text{(D.62)}
\end{aligned}$$

This, combined with $h\log x < 1$ and $x^h \geq 1$ yields

$$\frac{|1-x^h|}{hx^h} \ < \ \frac{eh\log x}{h} \ = \ e\log x. \qquad \text{(D.63)}$$

It is clear that $\frac{\log x}{x\log^3 x}$ is positive and integrable, and by L'Hospital's rule (see (D.51)) we have that

$$\lim_{h\to 0^+} \frac{1-x^h}{h}\frac{1}{x^h}\frac{1}{x\log^3 x} \ = \ -\frac{1}{x\log^2 x}. \qquad \text{(D.64)}$$

Thus the Dominated Convergence Theorem implies that

$$\lim_{h\to 0^+} \int_e^{\infty} \frac{1-x^h}{h}\frac{1}{x^h}\frac{\mathrm{d}x}{x\log^3 x} \ = \ -\int_e^{\infty} \frac{\mathrm{d}x}{x\log^2 x} \ = \ -1 \qquad \text{(D.65)}$$

(the last equality is derived in Remark D.3). □

References

[CaBe] G. Casella and R. Berger, *Statistical Inference*, 2nd edition, Duxbury Advanced Series, Pacific Grove, CA, 2002.

[DM] D. Devoto and S. Martinez, *Truncated Pareto Law and oresize distribution of ground rocks*, Mathematical Geology **30** (1998), no. 6, 661–673.

[Hi] T. Hill, *A statistical derivation of the significant-digit law*, Statistical Science **10** (1996), 354–363.

[HK] Kenneth Hoffman and Ray Kunze. *Linear algebra*. Second edition. Prentice-Hall Inc., Englewood Cliffs, N.J., 1971.

[HH] J. H. Hubbard and B. B. Hubbard, *Vector Calculus, Linear Algebra, and Differential Forms*, second edition, Prentice Hall, Upper Saddle River, NJ, 2002.

[MM] I. Miller and M. Miller, *John E. Freund's Mathematical Statistics with Applications*, seventh edition, Prentice Hall, 2004.

[Ne] M. E. J. Newman, *Power laws, Pareto distributions and Zipfs law*, Contemporary Physics **46** (2005), no. 5, 323-351.

[NM] M. Nigrini and S. J. Miller, *Benford's Law applied to hydrology data – results and relevance to other geophysical data*, preprint.

[Rai] R. A. Raimi, *The first digit problem*, Amer. Math. Monthly **83** (1976), no. 7, 521–538.

[Rud] W. Rudin, *Principles of Mathematical Analysis*, third edition, International Series in Pure and Applied Mathematics, McGraw-Hill Inc., New York, 1976.

[SS] E. Stein and R. Shakarchi, *Real Analysis: Measure Theory, Integration, and Hilbert Spaces*, Princeton University Press, Princeton, NJ, 2005.