Topic 14

Unbiased Estimation

14.1 Introduction

In creating a parameter estimator, a fundamental question is whether or not the estimator differs from the parameter in a systematic manner. Let's examine this by looking a the computation of the mean and the variance of 16 flips of a fair coin.

Give this task to 10 individuals and ask them report the number of heads. We can simulate this in R as follows

```
> (x<-rbinom(10,16,0.5))
 [1]   8   5   9   7   7   9   7   8   8 10
```

Our estimate is obtained by taking these 10 answers and averaging them. Intuitively we anticipate an answer around 8. For these 10 observations, we find, in this case, that

```
> sum(x)/10
[1] 7.8
```

The result is a bit below 8. Is this systematic? To assess this, we appeal to the ideas behind Monte Carlo to perform a 1000 simulations of the example above.

```
> meanx<-rep(0,1000)
> for (i in 1:1000){meanx[i]<-mean(rbinom(10,16,0.5))}
> mean(meanx)
[1] 8.0049
```

From this, we surmise that we the estimate of the sample mean $\bar{x}$ neither systematically overestimates or underestimates the distributional mean. From our knowledge of the binomial distribution, we know that the mean $\mu = np = 16 \cdot 0.5 = 8$. In addition, the sample mean $\bar{X}$ also has mean

$$E\bar{X} = \frac{1}{10}(8+8+8+8+8+8+8+8+8+8) = \frac{80}{10} = 8$$

verifying that we have no systematic error.

The phrase that we use is that the sample mean $\bar{X}$ is an **unbiased** estimator of the distributional mean μ. Here is the precise definition.

Definition 14.1. *For observations $X = (X_1, X_2, \ldots, X_n)$ based on a distribution having parameter value θ, and for $d(X)$ an estimator for $h(\theta)$, the* **bias** *is the mean of the difference $d(X) - h(\theta)$, i.e.,*

$$b_d(\theta) = E_\theta d(X) - h(\theta). \tag{14.1}$$

If $b_d(\theta) = 0$ for all values of the parameter, then $d(X)$ is called an **unbiased estimator**. *Any estimator that is not unbiased is called* **biased**.

Example 14.2. *Let $X_1, X_2, \ldots, X_n$ be Bernoulli trials with success parameter p and set the estimator for p to be $d(X) = \bar{X}$, the sample mean. Then,*

$$E_p\bar{X} = \frac{1}{n}(EX_1 + EX_2 + \cdots + EX_n) = \frac{1}{n}(p + p + \cdots + p) = p$$

Thus, $\bar{X}$ is an unbiased estimator for p. In this circumstance, we generally write $\hat{p}$ instead of $\bar{X}$. In addition, we can use the fact that for independent random variables, the variance of the sum is the sum of the variances to see that

$$\begin{aligned}\text{Var}(\hat{p}) &= \frac{1}{n^2}(\text{Var}(X_1) + \text{Var}(X_2) + \cdots + \text{Var}(X_n)) \\ &= \frac{1}{n^2}(p(1-p) + p(1-p) + \cdots + p(1-p)) = \frac{1}{n}p(1-p).\end{aligned}$$

Example 14.3. *If $X_1, \ldots, X_n$ form a simple random sample with unknown finite mean μ, then $\bar{X}$ is an unbiased estimator of μ. If the X_i have variance σ^2, then*

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}. \tag{14.2}$$

We can assess the quality of an estimator by computing its **mean square error**, defined by

$$E_\theta[(d(X) - h(\theta))^2]. \tag{14.3}$$

Estimators with smaller mean square error are generally preferred to those with larger. Next we derive a simple relationship between mean square error and variance. We begin by substituting (14.1) into (14.3), rearranging terms, and expanding the square.

$$\begin{aligned}E_\theta[(d(X) - h(\theta))^2] &= E_\theta[(d(X) - (E_\theta d(X) - b_d(\theta)))^2] = E_\theta[((d(X) - E_\theta d(X)) + b_d(\theta))^2] \\ &= E_\theta[(d(X) - E_\theta d(X))^2] + 2b_d(\theta)E_\theta[d(X) - E_\theta d(X)] + b_d(\theta)^2 \\ &= \text{Var}_\theta(d(X)) + b_d(\theta)^2\end{aligned}$$

Thus, the representation of the mean square error as equal to the variance of the estimator plus the square of the bias is called the **bias-variance decomposition**. In particular:

- The mean square error for an unbiased estimator is its variance.
- Bias always increases the mean square error.

14.2 Computing Bias

For the variance σ^2, we have been presented with two choices:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \text{and} \quad \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{14.4}$$

Using bias as our criterion, we can now resolve between the two choices for the estimators for the variance σ^2. Again, we use simulations to make a conjecture, we then follow up with a computation to verify our guess. For 16 tosses of a fair coin, we know that the variance is $np(1-p) = 16 \cdot 1/2 \cdot 1/2 = 4$

For the example above, we begin by simulating the coin tosses and compute the sum of squares $\sum_{i=1}^{10}(x_i - \bar{x})^2$,

```
> ssx<-rep(0,1000)
> for (i in 1:1000){x<-rbinom(10,16,0.5);ssx[i]<-sum((x-mean(x))^2)}
> mean(ssx)
[1] 35.8511
```

The choice is to divide either by 10, for the first choice, or 9, for the second.

```
> mean(ssx)/10;mean(ssx)/9
[1] 3.58511
[1] 3.983456
```

Exercise 14.4. *Repeat the simulation above, compute the sum of squares $\sum_{i=1}^{10}(x_i-8)^2$. Show that these simulations support dividing by 10 rather than 9. verify that $\sum_{i=1}^{n}(X_i-\mu)^2/n$ is an unbiased estimator for σ^2 for independent random variable $X_1,\dots,X_n$ whose common distribution has mean μ and variance σ^2.*

In this case, because we know all the aspects of the simulation, and thus we know that the answer ought to be near 4. Consequently, division by 9 appears to be the appropriate choice. Let's check this out, beginning with what seems to be the *inappropriate choice* to see what goes wrong..

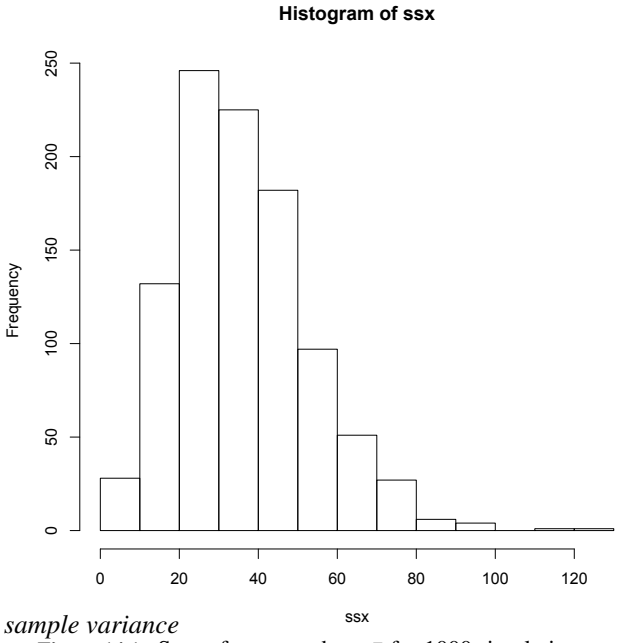

Figure 14.1: Sum of squares about $\bar{x}$ for 1000 simulations.

Example 14.5. *If a simple random sample $X_1, X_2, \dots,$ has unknown finite variance σ^2, then, we can consider the sample variance*

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

To find the mean of S^2, we divide the difference between an observation X_i and the distributional mean into two steps - the first from X_i to the sample mean $\bar{x}$ and and then from the sample mean to the distributional mean, i.e.,

$$X_i - \mu = (X_i - \bar{X}) + (\bar{X} - \mu).$$

We shall soon see that the lack of knowledge of μ is the source of the bias. Make this substitution and expand the square to obtain

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\mu)^2 &= \sum_{i=1}^{n}((X_i-\bar{X})+(\bar{X}-\mu))^2\\ &= \sum_{i=1}^{n}(X_i-\bar{X})^2 + 2\sum_{i=1}^{n}(X_i-\bar{X})(\bar{X}-\mu) + \sum_{i=1}^{n}(\bar{X}-\mu)^2\\ &= \sum_{i=1}^{n}(X_i-\bar{X})^2 + 2(\bar{X}-\mu)\sum_{i=1}^{n}(X_i-\bar{X}) + n(\bar{X}-\mu)^2\\ &= \sum_{i=1}^{n}(X_i-\bar{X})^2 + n(\bar{X}-\mu)^2\end{aligned}$$

(Check for yourself that the middle term in the third line equals 0.) Subtract the term $n(\bar{X}-\mu)^2$ from both sides and divide by n to obtain the identity

$$\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - (\bar{X}-\mu)^2.$$

Using the identity above and the linearity property of expectation we find that

$$\begin{aligned} ES^2 &= E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2\right] \\ &= E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2-(\bar{X}-\mu)^2\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}E[(X_i-\mu)^2]-E[(\bar{X}-\mu)^2] \\ &= \frac{1}{n}\sum_{i=1}^{n}\text{Var}(X_i)-\text{Var}(\bar{X}) \\ &= \frac{1}{n}n\sigma^2-\frac{1}{n}\sigma^2=\frac{n-1}{n}\sigma^2\neq\sigma^2. \end{aligned}$$

The last line uses (14.2). This shows that S^2 is a biased estimator for σ^2. Using the definition in (14.1), we can see that it is biased downwards.

$$b(\sigma^2)=\frac{n-1}{n}\sigma^2-\sigma^2=-\frac{1}{n}\sigma^2.$$

Note that the bias is equal to $-\text{Var}(\bar{X})$. In addition, because

$$E\left[\frac{n}{n-1}S^2\right]=\frac{n}{n-1}E\left[S^2\right]=\frac{n}{n-1}\left(\frac{n-1}{n}\sigma^2\right)=\sigma^2$$

and

$$S_u^2=\frac{n}{n-1}S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

is an unbiased estimator for σ^2. As we shall learn in the next section, because the square root is concave downward, $S_u=\sqrt{S_u^2}$ as an estimator for σ is **downwardly biased**.

Example 14.6. *We have seen, in the case of n Bernoulli trials having x successes, that $\hat{p}=x/n$ is an unbiased estimator for the parameter p. This is the case, for example, in taking a simple random sample of genetic markers at a particular biallelic locus. Let one allele denote the wildtype and the second a variant. If the circumstances in which variant is recessive, then an individual expresses the variant phenotype only in the case that both chromosomes contain this marker. In the case of independent alleles from each parent, the probability of the variant phenotype is p^2. Naïvely, we could use the estimator $\hat{p}^2$. (Later, we will see that this is the maximum likelihood estimator.) To determine the bias of this estimator, note that*

$$E\hat{p}^2=(E\hat{p})^2+\text{Var}(\hat{p})=p^2+\frac{1}{n}p(1-p). \tag{14.5}$$

Thus, the bias $b(p)=p(1-p)/n$ and the estimator $\hat{p}^2$ is biased upward.

Exercise 14.7. *For Bernoulli trials $X_1,\ldots,X_n$,*

$$\frac{1}{n}\sum_{i=1}^{n}(X_i-\hat{p})^2=\hat{p}(1-\hat{p}).$$

Based on this exercise, and the computation above yielding an unbiased estimator, S_u^2, for the variance,

$$E\left[\frac{1}{n-1}\hat{p}(1-\hat{p})\right]=\frac{1}{n}E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\hat{p})^2\right]=\frac{1}{n}E[S_u^2]=\frac{1}{n}Var(X_1)=\frac{1}{n}p(1-p).$$

In other words,

$$\frac{1}{n-1}\hat{p}(1-\hat{p})$$

is an unbiased estimator of $p(1-p)/n$. Returning to (14.5),

$$E\left[\hat{p}^2 - \frac{1}{n-1}\hat{p}(1-\hat{p})\right] = \left(p^2 + \frac{1}{n}p(1-p)\right) - \frac{1}{n}p(1-p) = p^2.$$

Thus,

$$\widehat{p^2}_u = \hat{p}^2 - \frac{1}{n-1}\hat{p}(1-\hat{p})$$

is an unbiased estimator of p^2.

To compare the two estimators for p^2, assume that we find 13 variant alleles in a sample of 30, then $\hat{p} = 13/30 = 0.4333$,

$$\hat{p}^2 = \left(\frac{13}{30}\right)^2 = 0.1878, \quad \textit{and} \quad \widehat{p^2}_u = \left(\frac{13}{30}\right)^2 - \frac{1}{29}\left(\frac{13}{30}\right)\left(\frac{17}{30}\right) = 0.1878 - 0.0085 = 0.1793.$$

The bias for the estimate $\hat{p}^2$, in this case 0.0085, is subtracted to give the unbiased estimate $\widehat{p^2}_u$.

The **heterozygosity** *of a biallelic locus is $h = 2p(1-p)$. From the discussion above, we see that h has the unbiased estimator*

$$\hat{h} = \frac{2n}{n-1}\hat{p}(1-\hat{p}) = \frac{2n}{n-1}\left(\frac{x}{n}\right)\left(\frac{n-x}{n}\right) = \frac{2x(n-x)}{n(n-1)}.$$

14.3 Compensating for Bias

In the methods of moments estimation, we have used $g(\bar{X})$ as an estimator for $g(\mu)$. If g is a **convex function**, we can say something about the bias of this estimator. In Figure 14.2, we see the method of moments estimator for the estimator $g(\bar{X})$ for a parameter β in the Pareto distribution. The choice of $\beta = 3$ corresponds to a mean of $\mu = 3/2$ for the Pareto random variables. The central limit theorem states that the sample mean $\bar{X}$ is nearly normally distributed with mean 3/2. Thus, the distribution of $\bar{X}$ is nearly symmetric around 3/2. From the figure, we can see that the interval from 1.4 to 1.5 under the function g maps into a longer interval above $\beta = 3$ than the interval from 1.5 to 1.6 maps below $\beta = 3$. Thus, the function g spreads the values of $\bar{X}$ above $\beta = 3$ more than below. Consequently, we anticipate that the estimator $\hat{\beta}$ will be **upwardly biased**.

To address this phenomena in more general terms, we use the characterization of a convex function as a differentiable function whose graph lies above any tangent line. If we look at the value μ for the convex function g, then this statement becomes

$$g(x) - g(\mu) \geq g'(\mu)(x-\mu).$$

Now replace x with the random variable $\bar{X}$ and take expectations.

$$E_\mu[g(\bar{X}) - g(\mu)] \geq E_\mu[g'(\mu)(\bar{X}-\mu)] = g'(\mu)E_\mu[\bar{X}-\mu] = 0.$$

Consequently,

$$E_\mu g(\bar{X}) \geq g(\mu) \tag{14.6}$$

and $g(\bar{X})$ is **biased upwards**. The expression in (14.6) is known as **Jensen's inequality**.

Exercise 14.8. *Show that the estimator S_u is a downwardly biased estimator for σ.*

To estimate the size of the bias, we look at a quadratic approximation for g centered at the value μ

$$g(x) - g(\mu) \approx g'(\mu)(x-\mu) + \frac{1}{2}g''(\mu)(x-\mu)^2.$$

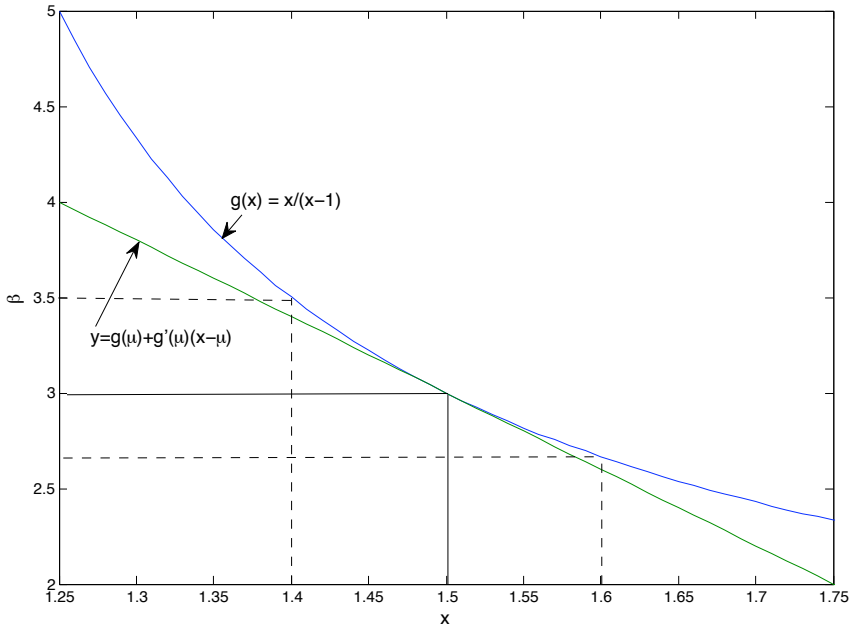

Figure 14.2: Graph of a convex function. Note that the tangent line is below the graph of g. Here we show the case in which $\mu = 1.5$ and $\beta = g(\mu) = 3$. Notice that the interval from $x = 1.4$ to $x = 1.5$ has a longer range than the interval from $x = 1.5$ to $x = 1.6$ Because g spreads the values of $\bar{X}$ above $\beta = 3$ more than below, the estimator $\hat{\beta}$ for β is biased upward. We can use a second order Taylor series expansion to correct most of this bias.

Again, replace x in this expression with the random variable $\bar{X}$ and then take expectations. Then, the bias

$$b_g(\mu) = E_\mu[g(\bar{X})] - g(\mu) \approx E_\mu[g'(\mu)(\bar{X} - \mu)] + \frac{1}{2}E[g''(\mu)(\bar{X} - \mu)^2] = \frac{1}{2}g''(\mu)\text{Var}(\bar{X}) = \frac{1}{2}g''(\mu)\frac{\sigma^2}{n}. \quad (14.7)$$

(Remember that $E_\mu[g'(\mu)(\bar{X} - \mu)] = 0$.) Thus, the bias has the intuitive properties of being

- large for strongly convex functions, i.e., ones with a large value for the second derivative evaluated at the mean μ,
- large for observations having high variance σ^2, and
- small when the number of observations n is large.

Exercise 14.9. *Use (14.7) to estimate the bias in using $\hat{p}^2$ as an estimate of p^2 is a sequence of n Bernoulli trials and note that it matches the value (14.5).*

Example 14.10. *For the method of moments estimator for the Pareto random variable, we determined that*

$$g(\mu) = \frac{\mu}{\mu - 1}.$$

and that $\bar{X}$ has

$$\textit{mean} \quad \mu = \tfrac{\beta}{\beta-1} \quad \textit{and} \quad \textit{variance} \quad \tfrac{\sigma^2}{n} = \tfrac{\beta}{n(\beta-1)^2(\beta-2)}$$

By taking the second derivative, we see that $g''(\mu) = 2(\mu - 1)^{-3} > 0$ and, because $\mu > 1$, g is a convex function. Next, we have

$$g''\left(\frac{\beta}{\beta - 1}\right) = \frac{2}{\left(\frac{\beta}{\beta-1} - 1\right)^3} = 2(\beta - 1)^3.$$

Thus, the bias

$$b_g(\beta) \approx \frac{1}{2} g''(\mu) \frac{\sigma^2}{n} = \frac{1}{2} 2(\beta-1)^3 \frac{\beta}{n(\beta-1)^2(\beta-2)} = \frac{\beta(\beta-1)}{n(\beta-2)}.$$

So, for $\beta = 3$ and $n = 100$, the bias is approximately 0.06. Compare this to the estimated value of 0.053 from the simulation in the previous section.

Example 14.11. *For estimating the population in mark and recapture, we used the estimate*

$$N = g(\mu) = \frac{kt}{\mu}$$

for the total population. Here μ is the mean number recaptured, k is the number captured in the second capture event and t is the number tagged. The second derivative

$$g''(\mu) = \frac{2kt}{\mu^3} > 0$$

and hence the method of moments estimate is biased upwards. In this siutation, $n = 1$ and the number recaptured is a hypergeometric random variable. Hence its variance

$$\sigma^2 = \frac{kt}{N} \frac{(N-t)(N-k)}{N(N-1)}.$$

Thus, the bias

$$b_g(N) = \frac{1}{2} \frac{2kt}{\mu^3} \frac{kt}{N} \frac{(N-t)(N-k)}{N(N-1)} = \frac{(N-t)(N-k)}{\mu(N-1)} = \frac{(kt/\mu - t)(kt/\mu - k)}{\mu(kt/\mu - 1)} = \frac{kt(k-\mu)(t-\mu)}{\mu^2(kt-\mu)}.$$

In the simulation example, $N = 2000, t = 200, k = 400$ and $\mu = 40$. This gives an estimate for the bias of 36.02. We can compare this to the bias of 2031.03-2000 = 31.03 based on the simulation in Example 13.2.

This suggests a new estimator by taking the method of moments estimator and subtracting the approximation of the bias.

$$\hat{N} = \frac{kt}{r} - \frac{kt(k-r)(t-r)}{r^2(kt-r)} = \frac{kt}{r}\left(1 - \frac{(k-r)(t-r)}{r(kt-r)}\right).$$

The delta method gives us that the standard deviation of the estimator is $|g'(\mu)|\sigma/\sqrt{n}$. Thus the ratio of the bias of an estimator to its standard deviation as determined by the delta method is approximately

$$\frac{g''(\mu)\sigma^2/(2n)}{|g'(\mu)|\sigma/\sqrt{n}} = \frac{1}{2} \frac{g''(\mu)}{|g'(\mu)|} \frac{\sigma}{\sqrt{n}}.$$

If this ratio is $\ll 1$, then the bias correction is not very important. In the case of the example above, this ratio is

$$\frac{36.02}{268.40} = 0.134$$

and its usefulness in correcting bias is small.

14.4 Consistency

Despite the desirability of using an unbiased estimator, sometimes such an estimator is hard to find and at other times impossible. However, note that in the examples above both the size of the bias and the variance in the estimator decrease inversely proportional to n, the number of observations. Thus, these estimators improve, under both of these criteria, with more observations. A concept that describes properties such as these is called **consistency**.

Definition 14.12. *Given data $X_1, X_2, \ldots$ and a real valued function h of the parameter space, a sequence of estimators d_n, based on the first n observations, is called* **consistent** *if for every choice of θ*

$$\lim_{n\to\infty} d_n(X_1, X_2, \ldots, X_n) = h(\theta)$$

whenever θ is the true state of nature.

Thus, the bias of the estimator disappears in the limit of a large number of observations. In addition, the distribution of the estimators $d_n(X_1, X_2, \ldots, X_n)$ become more and more concentrated near $h(\theta)$.

For the next example, we need to recall the sequence definition of continuity: A function g is continuous at a real number x provided that for every sequence $\{x_n; n \geq 1\}$ with

$$x_n \to x, \text{ then, we have that } g(x_n) \to g(x).$$

A function is called continuous if it is continuous at every value of x in the domain of g. Thus, we can write the expression above more succinctly by saying that for every convergent sequence $\{x_n; n \geq 1\}$,

$$\lim_{n\to\infty} g(x_n) = g(\lim_{n\to\infty} x_n).$$

Example 14.13. *For a method of moment estimator, let's focus on the case of a single parameter ($d = 1$). For independent observations, $X_1, X_2, \ldots$, having mean $\mu = k(\theta)$, we have that*

$$E\bar{X}_n = \mu,$$

i. e. $\bar{X}_n$, the sample mean for the first n observations, is an unbiased estimator for $\mu = k(\theta)$. Also, by the law of large numbers, we have that

$$\lim_{n\to\infty} \bar{X}_n = \mu.$$

Assume that k has a continuous inverse $g = k^{-1}$. In particular, because $\mu = k(\theta)$, we have that $g(\mu) = \theta$. Next, using the methods of moments procedure, define, for n observations, the estimators

$$\hat{\theta}_n(X_1, X_2, \ldots, X_n) = g\left(\frac{1}{n}(X_1 + \cdots + X_n)\right) = g(\bar{X}_n).$$

for the parameter θ. Using the continuity of g, we find that

$$\lim_{n\to\infty} \hat{\theta}_n(X_1, X_2, \ldots, X_n) = \lim_{n\to\infty} g(\bar{X}_n) = g(\lim_{n\to\infty} \bar{X}_n) = g(\mu) = \theta$$

and so we have that $g(\bar{X}_n)$ is a consistent sequence of estimators for θ.

14.5 Cramér-Rao Bound

This topic is somewhat more advanced and can be skipped for the first reading. This section gives us an introduction to the log-likelihood and its derivative, the **score** functions. We shall encounter these functions again when we introduce maximum likelihood estimation. In addition, the Cramér Rao bound, which is based on the variance of the score function, known as the Fisher information, gives a lower bound for the variance of an unbiased estimator. These concepts will be necessary to describe the variance for maximum likelihood estimators.

Among unbiased estimators, one important goal is to find an estimator that has as small a variance as possible, A more precise goal would be to find an unbiased estimator d that has **uniform minimum variance**. In other words, $d(X)$ has has a smaller variance than for any other unbiased estimator $\tilde{d}$ for every value θ of the parameter.

$$\text{Var}_\theta d(X) \le \text{Var}_\theta \tilde{d}(X) \quad \text{for all } \theta \in \Theta.$$

The **efficiency** $e(\tilde{d})$ of unbiased estimator $\tilde{d}$ is the minimum value of the ratio

$$\frac{\text{Var}_\theta d(X)}{\text{Var}_\theta \tilde{d}(X)}$$

over all values of θ. Thus, the efficiency is between 0 and 1 with a goal of finding estimators with efficiency as near to one as possible.

For unbiased estimators, the Cramér-Rao bound tells us how small a variance is ever possible. The formula is a bit mysterious at first. However, we shall soon learn that this bound is a consequence of the bound on correlation that we have previously learned

Recall that for two random variables Y and Z, the correlation

$$\rho(Y,Z) = \frac{\text{Cov}(Y,Z)}{\sqrt{\text{Var}(Y)\text{Var}(Z)}}. \tag{14.8}$$

takes values between -1 and 1. Thus, $\rho(Y,Z)^2 \le 1$ and so

$$\text{Cov}(Y,Z)^2 \le \text{Var}(Y)\text{Var}(Z). \tag{14.9}$$

Exercise 14.14. *If* $EZ = 0$*, the* $\text{Cov}(Y,Z) = EYZ$

We begin with data $X = (X_1, \dots, X_n)$ drawn from an unknown probability P_θ. The parameter space $\Theta \subset \mathbb{R}$. Denote the joint density of these random variables

$$\mathbf{f}(\mathbf{x}|\theta), \quad \text{where } \mathbf{x} = (x_1 \dots, x_n).$$

In the case that the data comes from a simple random sample then the joint density is the product of the marginal densities.

$$\mathbf{f}(\mathbf{x}|\theta) = f(x_1|\theta) \cdots f(x_n|\theta) \tag{14.10}$$

For continuous random variables, the two basic properties of the density are that $\mathbf{f}(\mathbf{x}|\theta) \ge 0$ for all $\mathbf{x}$ and that

$$1 = \int_{\mathbb{R}^n} \mathbf{f}(\mathbf{x}|\theta)\, d\mathbf{x}. \tag{14.11}$$

Now, let d be the unbiased estimator of $h(\theta)$, then by the basic formula for computing expectation, we have for continuous random variables

$$h(\theta) = E_\theta d(X) = \int_{\mathbb{R}^n} d(\mathbf{x})\mathbf{f}(\mathbf{x}|\theta)\, d\mathbf{x}. \tag{14.12}$$

If the functions in (14.11) and (14.12) are differentiable with respect to the parameter θ and we can pass the derivative through the integral, then we first differentiate both sides of equation (14.11), and then use the logarithm function to write this derivate as the expectation of a random variable,

$$0 = \int_{\mathbb{R}^n} \frac{\partial \mathbf{f}(\mathbf{x}|\theta)}{\partial\theta}\, d\mathbf{x} = \int_{\mathbb{R}^n} \frac{\partial \mathbf{f}(\mathbf{x}|\theta)/\partial\theta}{\mathbf{f}(\mathbf{x}|\theta)} \mathbf{f}(\mathbf{x}|\theta)\, d\mathbf{x} = \int_{\mathbb{R}^n} \frac{\partial \ln \mathbf{f}(\mathbf{x}|\theta)}{\partial\theta} \mathbf{f}(\mathbf{x}|\theta)\, d\mathbf{x} = E_\theta\left[\frac{\partial \ln \mathbf{f}(X|\theta)}{\partial\theta}\right]. \tag{14.13}$$

From a similar calculation using (14.12),

$$h'(\theta) = E_\theta\left[d(X)\frac{\partial \ln \mathbf{f}(X|\theta)}{\partial\theta}\right]. \tag{14.14}$$

Now, return to the review on correlation with $Y = d(X)$, the unbiased estimator for $h(\theta)$ and the **score function** $Z = \partial \ln \mathbf{f}(X|\theta)/\partial\theta$. From equations (14.14) and then (14.9), we find that

$$h'(\theta)^2 = E_\theta\left[d(X)\frac{\partial \ln \mathbf{f}(X|\theta)}{\partial\theta}\right]^2 = \text{Cov}_\theta\left(d(X), \frac{\partial \ln \mathbf{f}(X|\theta)}{\partial\theta}\right) \le \text{Var}_\theta(d(X))\text{Var}_\theta\left(\frac{\partial \ln \mathbf{f}(X|\theta)}{\partial\theta}\right),$$

or,

$$\text{Var}_\theta(d(X)) \ge \frac{h'(\theta)^2}{I(\theta)}. \tag{14.15}$$

where

$$I(\theta) = \text{Var}_\theta\left(\frac{\partial \ln \mathbf{f}(X|\theta)}{\partial\theta}\right) = E_\theta\left[\left(\frac{\partial \ln \mathbf{f}(X|\theta)}{\partial\theta}\right)^2\right]$$

is called the **Fisher information**. For the equality, recall that the variance $\text{Var}(Z) = EZ^2 - (EZ)^2$ and recall from equation (14.13) that the random variable $Z = \partial \ln \mathbf{f}(X|\theta)/\partial\theta$ has mean $EZ = 0$.

Equation (14.15), called the **Cramér-Rao lower bound** or the **information inequality**, states that the lower bound for the variance of an unbiased estimator is the reciprocal of the Fisher information. In other words, the *higher* the information, the *lower* is the possible value of the variance of an unbiased estimator.

If we return to the case of a simple random sample, then take the logarithm of both sides of equation (14.10)

$$\ln \mathbf{f}(\mathbf{x}|\theta) = \ln f(x_1|\theta) + \cdots + \ln f(x_n|\theta)$$

and then differentiate with respect to the parameter θ,

$$\frac{\partial \ln \mathbf{f}(\mathbf{x}|\theta)}{\partial\theta} = \frac{\partial \ln f(x_1|\theta)}{\partial\theta} + \cdots + \frac{\partial \ln f(x_n|\theta)}{\partial\theta}.$$

The random variables $\{\partial \ln f(X_k|\theta)/\partial\theta; 1 \le k \le n\}$ are independent and have the same distribution. Using the fact that the variance of the sum is the sum of the variances for independent random variables, we see that I_n, the Fisher information for n observations is n times the Fisher information of a single observation.

$$I_n(\theta) = \text{Var}\left(\frac{\partial \ln f(X_1|\theta)}{\partial\theta} + \cdots + \frac{\partial \ln f(X_n|\theta)}{\partial\theta}\right) = n\text{Var}(\frac{\partial \ln f(X_1|\theta)}{\partial\theta}) = nE[(\frac{\partial \ln f(X_1|\theta)}{\partial\theta})^2].$$

Notice the correspondence. *Information* is *linearly* proportional to the number of observations. If our estimator is a sample mean or a function of the sample mean, then the *variance* is *inversely* proportional to the number of observations.

Example 14.15. *For independent Bernoulli random variables with unknown success probability θ, the density is*

$$f(x|\theta) = \theta^x(1-\theta)^{(1-x)}.$$

The mean is θ and the variance is $\theta(1-\theta)$. Taking logarithms, we find that

$$\ln f(x|\theta) = x\ln\theta + (1-x)\ln(1-\theta),$$

$$\frac{\partial}{\partial\theta}\ln f(x|\theta) = \frac{x}{\theta} - \frac{1-x}{1-\theta} = \frac{x-\theta}{\theta(1-\theta)}.$$

The Fisher information associated to a single observation

$$I(\theta) = E\left[\left(\frac{\partial}{\partial\theta}\ln f(X|\theta)\right)^2\right] = \frac{1}{\theta^2(1-\theta)^2}E[(X-\theta)^2] = \frac{1}{\theta^2(1-\theta)^2}\text{Var}(X)$$
$$= \frac{1}{\theta^2(1-\theta)^2}\theta(1-\theta) = \frac{1}{\theta(1-\theta)}.$$

Thus, the information for n observations $I_n(\theta) = n/(\theta(1-\theta))$. Thus, by the Cramér-Rao lower bound, any unbiased estimator of θ based on n observations must have variance al least $\theta(1-\theta)/n$. Now, notice that if we take $d(\mathbf{x}) = \bar{x}$, then

$$E_\theta \bar{X} = \theta, \quad \textit{and} \quad \mathrm{Var}_\theta d(X) = \mathrm{Var}(\bar{X}) = \frac{\theta(1-\theta)}{n}.$$

These two equations show that $\bar{X}$ is a unbiased estimator having uniformly minimum variance.

Exercise 14.16. *For independent normal random variables with known variance σ_0^2 and unknown mean μ, $\bar{X}$ is a uniformly minimum variance unbiased estimator.*

Exercise 14.17. *Take two derivatives of $\ln \mathbf{f}(x|\theta)$ to show that*

$$I(\theta) = E_\theta \left[\left(\frac{\partial \ln \mathbf{f}(X|\theta)}{\partial \theta} \right)^2 \right] = -E_\theta \left[\frac{\partial^2 \ln \mathbf{f}(X|\theta)}{\partial \theta^2} \right]. \tag{14.16}$$

This identity is often a useful alternative to compute the Fisher Information.

Example 14.18. *For an exponential random variable,*

$$\ln f(x|\lambda) = \ln \lambda - \lambda x, \quad \frac{\partial^2 f(x|\lambda)}{\partial \lambda^2} = -\frac{1}{\lambda^2}.$$

Thus, by (14.16),

$$I(\lambda) = \frac{1}{\lambda^2}.$$

Now, $\bar{X}$ is an unbiased estimator for $h(\lambda) = 1/\lambda$ with variance

$$\frac{1}{n\lambda^2}.$$

By the Cramér-Rao lower bound, we have that

$$\frac{g'(\lambda)^2}{nI(\lambda)} = \frac{1/\lambda^4}{n\lambda^2} = \frac{1}{n\lambda^2}.$$

Because $\bar{X}$ has this variance, it is a uniformly minimum variance unbiased estimator.

Example 14.19. *To give an estimator that does not achieve the Cramér-Rao bound, let $X_1, X_2, \ldots, X_n$ be a simple random sample of Pareto random variables with density*

$$f_X(x|\beta) = \frac{\beta}{x^{\beta+1}}, \quad x > 1.$$

The mean and the variance

$$\mu = \frac{\beta}{\beta - 1}, \qquad \sigma^2 = \frac{\beta}{(\beta-1)^2(\beta-2)}.$$

Thus, $\bar{X}$ is an unbiased estimator of $\mu = \beta/(\beta-1)$

$$\mathrm{Var}(\bar{X}) = \frac{\beta}{n(\beta-1)^2(\beta-2)}.$$

To compute the Fisher information, note that

$$\ln f(x|\beta) = \ln \beta - (\beta+1)\ln x \quad \textit{and thus} \quad \frac{\partial^2 \ln f(x|\beta)}{\partial \beta^2} = -\frac{1}{\beta^2}.$$

Using (14.16), we have that

$$I(\beta) = \frac{1}{\beta^2}.$$

Next, for

$$\mu = g(\beta) = \frac{\beta}{\beta - 1}, \quad g'(\beta) = -\frac{1}{(\beta-1)^2}, \quad \textit{and} \quad g'(\beta)^2 = \frac{1}{(\beta-1)^4}.$$

Thus, the Cramér-Rao bound for the estimator is

$$\frac{g'(\beta)^2}{I_n(\beta)} = \frac{\beta^2}{n(\beta-1)^4}.$$

and the efficiency compared to the Cramér-Rao bound is

$$\frac{g'(\beta)^2/I_n(\beta)}{\mathrm{Var}(\bar{X})} = \frac{\beta^2}{n(\beta-1)^4} \cdot \frac{n(\beta-1)^2(\beta-2)}{\beta} = \frac{\beta(\beta-2)}{(\beta-1)^2} = 1 - \frac{1}{(\beta-1)^2}.$$

The Pareto distribution does not have a variance unless $\beta > 2$. For β just above 2, the efficiency compared to its Cramér-Rao bound is low but improves with larger β.

14.6 A Note on Efficient Estimators

For an efficient estimator, we need find the cases that lead to equality in the correlation inequality (14.8). Recall that equality occurs precisely when the correlation is ± 1. This occurs when the estimator $d(X)$ and the score function $\partial \ln f_X(X|\theta)/\partial\theta$ are linearly related with probability 1.

$$\frac{\partial}{\partial\theta} \ln f_X(X|\theta) = a(\theta)d(X) + b(\theta).$$

After integrating, we obtain,

$$\ln f_X(X|\theta) = \int a(\theta)d\theta d(X) + \int b(\theta)d\theta + j(X) = \pi(\theta)d(X) + B(\theta) + j(X)$$

Note that the constant of integration of integration is a function of X. Now exponentiate both sides of this equation

$$f_X(X|\theta) = c(\theta)h(X)\exp(\pi(\theta)d(X)). \tag{14.17}$$

Here $c(\theta) = \exp B(\theta)$ and $h(X) = \exp j(X)$.

We shall call density functions satisfying equation (14.17) an **exponential family** with **natural parameter** $\pi(\theta)$. Thus, if we have independent random variables $X_1, X_2, \ldots X_n$, then the joint density is the product of the densities, namely,

$$\mathbf{f}(X|\theta) = c(\theta)^n h(X_1)\cdots h(X_n)\exp(\pi(\theta)(d(X_1) + \cdots + d(X_n)). \tag{14.18}$$

In addition, as a consequence of this linear relation in (14.18),

$$\overline{d(X)} = \frac{1}{n}(d(X_1) + \cdots + d(X_n))$$

is an efficient estimator for $h(\theta)$.

Example 14.20 (Poisson random variables)**.**

$$f(x|\lambda) = \frac{\lambda^x}{x!}e^{-\lambda} = e^{-\lambda}\frac{1}{x!}\exp(x\ln\lambda).$$

Thus, Poisson random variables are an exponential family with $c(\lambda) = \exp(-\lambda)$, $h(x) = 1/x!$, *and natural parameter* $\pi(\lambda) = \ln \lambda$. *Because*

$$\lambda = E_\lambda \bar{X},$$

$\bar{X}$ *is an unbiased estimator of the parameter* λ.

The score function

$$\frac{\partial}{\partial \lambda} \ln f(x|\lambda) = \frac{\partial}{\partial \lambda}(x \ln \lambda - \ln x! - \lambda) = \frac{x}{\lambda} - 1.$$

The Fisher information for one observation is

$$I(\lambda) = E_\lambda \left[\left(\frac{X}{\lambda} - 1 \right)^2 \right] = \frac{1}{\lambda^2} E_\lambda[(X - \lambda)^2] = \frac{1}{\lambda}.$$

Thus, $I_n(\lambda) = n/\lambda$ *is the Fisher information for* n *observations. In addition,*

$$\mathit{Var}_\lambda(\bar{X}) = \frac{\lambda}{n}$$

and $d(x) = \bar{x}$ *has efficiency*

$$\frac{\mathit{Var}(\bar{X})}{1/I_n(\lambda)} = 1.$$

This could have been predicted. The density of n *independent observations is*

$$\mathbf{f}(\mathbf{x}|\lambda) = \frac{e^{-\lambda}}{x_1!}\lambda^{x_1} \cdots \frac{e^{-\lambda}}{x_n!}\lambda^{x_n} = \frac{e^{-n\lambda}\lambda^{x_1 \cdots + x_n}}{x_1! \cdots x_n!} = \frac{e^{-n\lambda}\lambda^{n\bar{x}}}{x_1! \cdots x_n!}$$

and so the score function

$$\frac{\partial}{\partial \lambda} \ln \mathbf{f}(\mathbf{x}|\lambda) = \frac{\partial}{\partial \lambda}(-n\lambda + n\bar{x} \ln \lambda) = -n + \frac{n\bar{x}}{\lambda}$$

showing that the estimate $\bar{x}$ *and the score function are linearly related.*

Exercise 14.21. *Show that a Bernoulli random variable with parameter* p *is an exponential family.*

Exercise 14.22. *Show that a normal random variable with known variance* σ_0^2 *and unknown mean* μ *is an exponential family.*

14.7 Answers to Selected Exercises

14.4. Repeat the simulation, replacing `mean(x)` by 8.

```
> ssx<-rep(0,1000)
> for (i in 1:1000){x<-rbinom(10,16,0.5);ssx[i]<-sum((x-8)^2)}
> mean(ssx)/10;mean(ssx)/9
[1] 3.9918
[1] 4.435333
```

Note that division by 10 gives an answer very close to the correct value of 4. To verify that the estimator is unbiased, we write

$$E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2\right] = \frac{1}{n}\sum_{i=1}^{n} E[(X_i - \mu)^2] = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}(X_i) = \frac{1}{n}\sum_{i=1}^{n} \sigma^2 = \sigma^2.$$

14.7. For a Bernoulli trial note that $X_i^2 = X_i$. Expand the square to obtain

$$\sum_{i=1}^{n}(X_i - \hat{p})^2 = \sum_{i=1}^{n} X_i^2 - \hat{p}\sum_{i=1}^{n} X_i + n\hat{p}^2 = n\hat{p} - 2n\hat{p}^2 + n\hat{p}^2 = n(\hat{p} - \hat{p}^2) = n\hat{p}(1-\hat{p}).$$

Divide by n to obtain the result.

14.8. Recall that $ES_u^2 = \sigma^2$. Check the second derivative to see that $g(t) = \sqrt{t}$ is concave down for all t. For concave down functions, the direction of the inequality in Jensen's inequality is reversed. Setting $t = S_u^2$, we have that

$$ES_u = Eg(S_u^2) \le g(ES_u^2) = g(\sigma^2) = \sigma$$

and S_u is a downwardly biased estimator of σ.

14.9. Set $g(p) = p^2$. Then, $g''(p) = 2$. Recall that the variance of a Bernoulli random variable $\sigma^2 = p(1-p)$ and the bias

$$b_g(p) \approx \frac{1}{2}g''(p)\frac{\sigma^2}{n} = \frac{1}{2}2\frac{p(1-p)}{n} = \frac{p(1-p)}{n}.$$

14.14. $\mathrm{Cov}(Y, Z) = EYZ - EY \cdot EZ = EYZ$ whenever $EZ = 0$.

14.16. For independent normal random variables with known variance σ_0^2 and unknown mean μ, the density

$$f(x|\mu) = \frac{1}{\sigma_0\sqrt{2\pi}} \exp -\frac{(x-\mu)^2}{2\sigma_0^2},$$

$$\ln f(x|\mu) = -\ln(\sigma_0\sqrt{2\pi}) - \frac{(x-\mu)^2}{2\sigma_0^2}.$$

Thus, the score function

$$\frac{\partial}{\partial\mu}\ln f(x|\mu) = \frac{1}{\sigma_0^2}(x-\mu).$$

and the Fisher information associated to a single observation

$$I(\mu) = E\left[\left(\frac{\partial}{\partial\mu}\ln f(X|\mu)\right)^2\right] = \frac{1}{\sigma_0^4}E[(X-\mu)^2] = \frac{1}{\sigma_0^4}\mathrm{Var}(X) = \frac{1}{\sigma_0^2}.$$

Again, the information is the reciprocal of the variance. Thus, by the Cramér-Rao lower bound, any unbiased estimator based on n observations must have variance al least σ_0^2/n. However, if we take $d(\mathbf{x}) = \bar{x}$, then

$$\mathrm{Var}_\mu d(X) = \frac{\sigma_0^2}{n}.$$

and $\bar{x}$ is a uniformly minimum variance unbiased estimator.

14.17. First, we take two derivatives of $\ln \mathbf{f}(x|\theta)$.

$$\frac{\partial \ln \mathbf{f}(x|\theta)}{\partial\theta} = \frac{\partial \mathbf{f}(x|\theta)/\partial\theta}{\mathbf{f}(x|\theta)} \tag{14.19}$$

and

$$\begin{aligned}\frac{\partial^2 \ln \mathbf{f}(x|\theta)}{\partial\theta^2} &= \frac{\partial^2\mathbf{f}(x|\theta)/\partial\theta^2}{\mathbf{f}(x|\theta)} - \frac{(\partial\mathbf{f}(x|\theta)/\partial\theta)^2}{\mathbf{f}(x|\theta)^2} = \frac{\partial^2\mathbf{f}(x|\theta)/\partial\theta^2}{\mathbf{f}(x|\theta)} - \left(\frac{\partial\mathbf{f}(x|\theta)/\partial\theta)}{\mathbf{f}(x|\theta)}\right)^2 \\ &= \frac{\partial^2\mathbf{f}(x|\theta)/\partial\theta^2}{\mathbf{f}(x|\theta)} - \left(\frac{\partial \ln \mathbf{f}(x|\theta)}{\partial\theta}\right)^2\end{aligned}$$

upon substitution from identity (14.19). Thus, the expected values satisfy

$$E_\theta\left[\frac{\partial^2 \ln \mathbf{f}(X|\theta)}{\partial\theta^2}\right] = E_\theta\left[\frac{\partial^2 \mathbf{f}(X|\theta)/\partial\theta^2}{\mathbf{f}(X|\theta)}\right] - E_\theta\left[\left(\frac{\partial \ln \mathbf{f}(X|\theta)}{\partial\theta}\right)^2\right].$$

Consquently, the exercise is complete if we show that $E_\theta\left[\frac{\partial^2 \mathbf{f}(X|\theta)/\partial\theta^2}{\mathbf{f}(X|\theta)}\right] = 0$. However, for a continuous random variable,

$$E_\theta\left[\frac{\partial^2 \mathbf{f}(X|\theta)/\partial\theta^2}{\mathbf{f}(X|\theta)}\right] = \int \frac{\partial^2 \mathbf{f}(x|\theta)/\partial\theta^2}{\mathbf{f}(x|\theta)}\,\mathbf{f}(x|\theta)\,dx = \int \frac{\partial^2 \mathbf{f}(x|\theta)}{\partial\theta^2}\,dx = \frac{\partial^2}{\partial\theta^2}\int \mathbf{f}(x|\theta)\,dx = \frac{\partial^2}{\partial\theta^2}1 = 0.$$

Note that the computation require that we be able to pass two derivatives with respect to θ through the integral sign.

14.21. The Bernoulli density

$$f(x|p) = p^x(1-p)^{1-x} = (1-p)\left(\frac{p}{1-p}\right)^x = (1-p)\exp\left(x\ln\left(\frac{p}{1-p}\right)\right).$$

Thus, $c(p) = 1-p, h(x) = 1$ and the natural parameter $\pi(p) = \ln\left(\frac{p}{1-p}\right)$, the **log-odds**.

14.22. The normal density

$$f(x|\mu) = \frac{1}{\sigma_0\sqrt{2\pi}}\exp -\frac{(x-\mu)^2}{2\sigma_0^2} = \frac{1}{\sigma_0\sqrt{2\pi}}e^{-\mu^2/2\sigma_0}e^{-x^2/2\sigma_0}\exp\frac{x\mu}{\sigma_0^2}$$

Thus, $c(\mu) = \frac{1}{\sigma_0\sqrt{2\pi}}e^{-\mu^2/2\sigma_0}, h(x) = e^{-x^2/2\sigma_0}$ and the natural parameter $\pi(\mu) = \mu/\sigma_0^2$.