Topic 14
Unbiased Estimation
Defining and Computing Bias

Outline

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.

```
> (x<-rbinom(10,16,0.5))
 [1] 11 11  8  8  9 10  8  9  5  8
```

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] 8.7
```

The result is a bit above 8. Is this systematic?

Introduction

To assess this, we appeal to the ideas behind Monte Carlo to perform a 5000 simulations of the example above.

```
> meanx<-rep(0,5000)
> for (i in 1:5000){meanx[i]<-mean(rbinom(10,16,0.5))}
> mean(meanx)
[1] 7.99818
```

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 $\mu = np = 16 \cdot 0.5 = 8$. Thus, 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.

Definition of Bias

Definition. 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.,

$$b_d(\theta) = E_\theta d(X) - h(\theta).$$

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.

Exercise. 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

$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

Mean Square Error

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

$$E_\theta[(d(X) - h(\theta))^2].$$

To derive a simple relationship between mean square error and variance, we begin by substituting the equation for bias into the question above, rearranging terms, and expanding the square.

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

NB. $E_\theta[d(X) - E_\theta d(X)] = 0$. So, bias *increases* mean square error.

Computing Bias

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

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

Using bias as our criterion to resolve between the two choices, we use simulations to make a conjecture. For 16 tosses of a fair coin, we know that the variance is $np(1-p) = 16 \cdot 1/2 \cdot 1/2 = 4$.

```
> ssx<-rep(0,5000)
> for (i in 1:5000){x<-rbinom(10,16,0.5);ssx[i]<-sum((x-mean(x))^2)}
> mean(ssx)
[1] 35.87472
```

Computing Bias

Because we know all the aspects of the simulation, we know that the answer ought to be near 4.

```
> mean(ssx)/10;mean(ssx)/9
[1] 3.587472
[1] 3.98608
```

Consequently, division by 9 appears to be the appropriate choice. Let's check this out, beginning with

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

seemingly the *inappropriate choice* to see what goes wrong..

Computing Bias

We divide the difference between an observation X_i and μ - the first from X_i to the sample mean $\bar{X}$ and then from the sample mean to the distributional mean, μ i.e.,

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

Make this substitution and expand the square to obtain

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

Computing Bias

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

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

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

Computing Bias

This shows that S^2 is a biased estimator for σ^2. Using the definition of bias, we can see that it is biased *downwards*.

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

Exercise. Show that the alternative choice

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

is an unbiased estimator of σ^2.

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