Economics 103 – Statistics for Economists

Francis J. DiTraglia

University of Pennsylvania

Lecture 15

Sampling Distributions and Estimation – Part II

Unbiased means “Right on Average”

Bias of an Estimator

Let $\widehat{\theta}_n$ be a sample estimator of a population parameter θ_0. The *bias* of $\widehat{\theta}_n$ is $E[\widehat{\theta}_n] - \theta_0$.

Unbiased Estimator

A sample estimator $\widehat{\theta}_n$ of a population parameter θ_0 is called *unbiased* if $E[\widehat{\theta}_n] = \theta_0$

Why $(n-1)$ for sample variance?

Why $(n-1)$ for sample variance?

We will show that having $n-1$ in the denominator ensures:

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

under random sampling.

Why $(n-1)$ for sample variance?

Step # 1 – Tedious but straightforward algebra gives:

$$\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 = \left[\sum_{i=1}^{n} (X_i - \mu)^2\right] - n(\bar{X} - \mu)^2$$

You are not responsible for proving Step #1 on an exam.

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

Why $(n-1)$ for sample variance?

Step # 2 – Take Expectations of Step # 1:

$$
\begin{aligned}
E\left[\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2\right] &= E\left[\left\{\sum_{i=1}^{n}(X_i-\mu)^2\right\}-n(\bar{X}-\mu)^2\right] \\
&= E\left[\sum_{i=1}^{n}(X_i-\mu)^2\right]-E\left[n(\bar{X}-\mu)^2\right] \\
&= \sum_{i=1}^{n}E\left[(X_i-\mu)^2\right]-n\,E\left[(\bar{X}-\mu)^2\right]
\end{aligned}
$$

Where we have used the linearity of expectation.

Why $(n-1)$ for sample variance?

Step # 3 – Use assumption of random sampling:

$X_1, \ldots, X_n \sim$ iid with mean μ and variance σ^2

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

Since we showed earlier today that $E[\bar{X}] = \mu$ and $Var(\bar{X}) = \sigma^2/n$ under this random sampling assumption.

Why $(n-1)$ for sample variance?

Finally – Divide Step # 3 by $(n-1)$:

$$E[S^2] \;=\; E\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2\right]=\frac{(n-1)\sigma^2}{n-1}=\sigma^2$$

Hence, having $(n-1)$ in the denominator ensures that the sample variance is "correct on average," that is *unbiased*.

A Different Estimator of the Population Variance

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

$$E[\widehat{\sigma}^2] = E\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right] = \frac{1}{n}E\left[\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right] = \frac{(n-1)\sigma^2}{n}$$

Bias of $\widehat{\sigma}^2$

$$E[\widehat{\sigma}^2] - \sigma^2 = \frac{(n-1)\sigma^2}{n} - \sigma^2 = \frac{(n-1)\sigma^2}{n} - \frac{n\sigma^2}{n} = -\sigma^2/n$$

How Large is the Average Family?

How many brothers and sisters are in your family, including yourself?

The average number of children per family was about 2.0 twenty years ago.

What's Going On Here?

Biased Sample!

- Zero children $\Rightarrow$ didn't send any to college
- Sampling by *children* so large families oversampled