Proof that Sample Variance is Unbiased Plus Lots of Other Cool Stuff

Scott D. Anderson

http://www.spelman.edu/~anderson/teaching/437/unbiased/unbiased.html

Fall 1999

Expected Value of S^2

The following is a proof that the formula for the sample variance, S^2, is unbiased. Recall that it seemed like we should divide by n, but instead we divide by n-1. Here's why.

First, recall the formula for the sample variance:

$$\text{var}(x) = S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Now, we want to compute the expected value of this:

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

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

Now, let's multiply both sides of the equation by n-1, just so we don't have to keep carrying that around, and square out the right side, just like we did with that shortcut formula for SSX, above.

$$(n-1)E\left[S^2\right] = E\left[\sum_{i=1}^{n} x_i^2 - 2\bar{x}x_i + \bar{x}^2\right]$$

$$(n-1)E[S^2] = E\left[\sum x_i^2\right] - E\left[\sum 2\bar{x}x_i\right] + E\left[\sum \bar{x}^2\right]$$

$$(n-1)E[S^2] = E\left[\sum x_i^2\right] - E\left[2\bar{x}\sum x_i\right] + E\left[\bar{x}^2\sum 1\right]$$

Now, if you think about it, it's clear that $\sum x_i = n\bar{x}$, so we can rewrite the middle term on the RHS in terms of $\bar{x}$:

$$(n-1)E[S^2] = E\left[\sum x_i^2\right] - E\left[2\bar{x}(n\bar{x})\right] + E\left[\bar{x}^2\sum 1\right]$$

$$(n-1)E[S^2] = E\left[\sum x_i^2\right] - nE\left[\bar{x}^2\right]$$

$$(n-1)E[S^2] = nE\left[x_i^2\right] - nE\left[\bar{x}^2\right]$$

$$\frac{n-1}{n}E[S^2] = E\left[x_i^2\right] - E\left[\bar{x}^2\right]$$

Let's write that again as a numbered equation:

$$\frac{n-1}{n}E[S^2] = E\left[x_i^2\right] - E\left[\bar{x}^2\right] \tag{1}$$

Unfortunately, the expected value of the square of something is not equal to the square of the expected value, so we seem to have hit an impasse with both terms on the RHS. But, we're not out of tricks yet. Each of those terms is an expected value of something squared: a second moment. Let's use the trick about moments that we saw above.

First, let Y be the random variable defined by the sample mean, $\bar{x}$. We're trying to figure out the expected value of its square.

$$E[Y^2] = E[\bar{x}^2] = \text{var}[Y] + E[Y]^2$$

$$E[Y^2] = E[\bar{x}^2] = \text{var}\left[\frac{1}{n}\sum x_i\right] + \mu^2$$

$$E[Y^2] = E[\bar{x}^2] = \frac{1}{n^2}\text{var}\left[\sum x_i\right] + \mu^2$$

$$E[Y^2] = E[\bar{x}^2] = \frac{1}{n^2} \sum \text{var}[x_i] + \mu^2$$

$$E[Y^2] = E[\bar{x}^2] = \frac{1}{n^2} \sum \sigma^2 + \mu^2$$

$$E[Y^2] = E[\bar{x}^2] = \frac{1}{n^2}(n\sigma^2) + \mu^2$$

$$E[Y^2] = E[\bar{x}^2] = \frac{1}{n}\sigma^2 + \mu^2$$

We can substitute this stuff for the second term on the RHS of equation 1. Also, note that the first term on the RHS of eqation 1 is the second moment of X, so that can also be re-written. Doing both substitutions gives us:

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

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

$$E[S^2] = \sigma^2$$

Whew! That was hard, but solvable. This is why S^2 with the n-1 denominator is an unbiased estimator.