Why dividing by n – 1 yields an unbiased estimate of the population variance

Kristopher J. Preacher

2/15/12

First, prove that $E\left[x_i^2\right]=\sigma^2+\mu^2$:

$$\begin{aligned}
\sigma^2 &= E\left[\left(x_i - E[x_i]\right)^2\right] \\
&= E\left[\left(x_i - E[x_i]\right)\left(x_i - E[x_i]\right)\right] \\
&= E\left[x_i^2 - x_i E[x_i] - E[x_i]x_i + E[x_i]E[x_i]\right] \\
&= E\left[x_i^2 - 2x_i\mu + \mu^2\right] \\
&= E\left[x_i^2\right] - 2\mu E[x_i] + E\left[\mu^2\right] \\
&= E\left[x_i^2\right] - 2\mu^2 + \mu^2 \\
&= E\left[x_i^2\right] - \mu^2
\end{aligned}$$

It follows that $E\left[x_i^2\right]=\sigma^2+\mu^2$. We make use of this identity in the following.

Exhibit 1: The expectation of the sample variance (dividing by *n*) leads to a <u>biased</u> estimate of σ^2:

$$
\begin{aligned}
E\left[S^2\right] &= E\left[\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}{n}\right] \\
&= \frac{1}{n}E\left[\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(x_i-\bar{x}\right)\right] \\
&= \frac{1}{n}E\left[\sum_{i=1}^{n}\left(x_i^2-2x_i\bar{x}+\bar{x}^2\right)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(E\left[x_i^2\right]-2E\left[x_i\bar{x}\right]+E\left[\bar{x}^2\right]\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(E\left[x_i^2\right]-2E\left[x_i\frac{1}{n}\sum_{i=1}^{n}x_i\right]+E\left[\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)^2\right]\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(E\left[x_i^2\right]-\frac{2}{n}E\left[x_i\sum_{i=1}^{n}x_i\right]+\frac{1}{n^2}E\left[\left(\sum_{i=1}^{n}x_i\right)^2\right]\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(E\left[x_i^2\right]-\frac{2}{n}E\left[x_i\left(x_i+\sum_{i\neq j}x_j\right)\right]+\frac{1}{n^2}E\left[\sum_{i=1}^{n}x_i^2+\sum_{i\neq j}x_ix_j\right]\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(E\left[x_i^2\right]-\frac{2}{n}\left[\left(E\left[x_i^2\right]+E\left[x_i\sum_{i\neq j}x_j\right]\right)\right]+\frac{1}{n^2}\sum_{i=1}^{n}E\left[x_i^2\right]+\frac{1}{n^2}\sum_{i\neq j}E\left[x_i\right]E\left[x_j\right]\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(E\left[x_i^2\right]-\frac{2}{n}E\left[x_i^2\right]-\frac{2}{n}(n-1)E\left[x_ix_j\right]+\frac{1}{n^2}\sum_{i=1}^{n}E\left[x_i^2\right]+\frac{1}{n^2}\sum_{i\neq j}E\left[x_i\right]E\left[x_j\right]\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\left(\sigma^2+\mu^2\right)-\frac{2}{n}\left(\sigma^2+\mu^2\right)-\frac{2}{n}(n-1)\mu^2+\frac{1}{n^2}n\left(\sigma^2+\mu^2\right)+\frac{1}{n^2}n(n-1)\mu^2\right) \\
&= \sigma^2+\mu^2-\frac{2}{n}\sigma^2-\frac{2}{n}\mu^2-2\mu^2+\frac{2}{n}\mu^2+\frac{1}{n}\sigma^2+\frac{1}{n}\mu^2+\mu^2-\frac{1}{n}\mu^2 \\
&= \sigma^2-\frac{\sigma^2}{n}
\end{aligned}
$$

...which underestimates σ^2 by a factor of $1-\frac{1}{n}=\frac{n-1}{n}$.

Exhibit 2: The expectation of the sample variance (dividing by $n - 1$) leads to an <u>unbiased</u> estimate of σ^2:

$$
\begin{aligned}
E\left[s^2\right] &= E\left[\frac{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2}{n-1}\right] \\
&= \frac{1}{n-1}E\left[\sum_{i=1}^{n}\left(x_i-\overline{x}\right)\left(x_i-\overline{x}\right)\right] \\
&= \frac{1}{n-1}E\left[\sum_{i=1}^{n}\left(x_i^2-2x_i\overline{x}+\overline{x}^2\right)\right] \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left(E\left[x_i^2\right]-2E\left[x_i\overline{x}\right]+E\left[\overline{x}^2\right]\right) \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left(E\left[x_i^2\right]-2E\left[x_i\frac{1}{n}\sum_{i=1}^{n}x_i\right]+E\left[\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)^2\right]\right) \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left(E\left[x_i^2\right]-\frac{2}{n}E\left[x_i\sum_{i=1}^{n}x_i\right]+\frac{1}{n^2}E\left[\left(\sum_{i=1}^{n}x_i\right)^2\right]\right) \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left(E\left[x_i^2\right]-\frac{2}{n}E\left[x_i\left(x_i+\sum_{i\neq j}x_j\right)\right]+\frac{1}{n^2}E\left[\sum_{i=1}^{n}x_i^2+\sum_{i\neq j}x_ix_j\right]\right) \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left(E\left[x_i^2\right]-\frac{2}{n}\left[\left(E\left[x_i^2\right]+E\left[x_i\sum_{i\neq j}x_j\right]\right)\right]+\frac{1}{n^2}\sum_{i=1}^{n}E\left[x_i^2\right]+\frac{1}{n^2}\sum_{i\neq j}E\left[x_i\right]E\left[x_j\right]\right) \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left(E\left[x_i^2\right]-\frac{2}{n}E\left[x_i^2\right]-\frac{2}{n}(n-1)E\left[x_ix_j\right]+\frac{1}{n^2}\sum_{i=1}^{n}E\left[x_i^2\right]+\frac{1}{n^2}\sum_{i\neq j}E\left[x_i\right]E\left[x_j\right]\right) \\
&= \frac{n}{n-1}\left(\left(\sigma^2+\mu^2\right)-\frac{2}{n}\left(\sigma^2+\mu^2\right)-\frac{2}{n}(n-1)\mu^2+\frac{1}{n^2}n\left(\sigma^2+\mu^2\right)+\frac{1}{n^2}n(n-1)\mu^2\right) \\
&= \frac{n}{n-1}\left(\sigma^2-\frac{1}{n}\sigma^2\right) \\
&= \left(\frac{n}{n-1}-\frac{1}{n-1}\right)\sigma^2 \\
&= \sigma^2
\end{aligned}
$$