Sample Variance: Why divide by $n-1$?

Pramook Khungurn

November 1, 2014

As someone with little training in statistics, it often puzzles me why, when estimatiing the sample/empirical variance of a random varianble, we divide the sum by $n-1$ instead of n.

Let X be a random variable. Suppose we have n i.i.d. samples of X, which we denote by X_1, X_2, ..., X_n. Let $\bar{X}_n$ denote the empirical mean:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Of course, we know that $\bar{X}_n$ is an unbiased estimator of $E[X]$.

We want to construct an unbiased estimate of $Var(X)$. Now, we might be tempted to use the estimator:

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

However, this estimator is biased. One thing to notice is that $\bar{X}_n$ is not exactly $E[X]$. So, what is the expectation of this estimator? We have

$$\begin{aligned}
E[S_n^2] &= E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right] = \frac{1}{n}E\left[\sum_{i=1}^{n}(X_i^2 - 2X_i\bar{X}_n + \bar{X}_n^2)\right] = \frac{1}{n}\left(\sum_{i=1}^{n}E[X_i^2] - 2\sum_{i=1}^{n}E[X_i\bar{X}_n] + \sum_{i=1}^{n}E[\bar{X}_n^2]\right) \\
&= \frac{1}{n}\left(nE[X^2] - 2\sum_{i=1}^{n}E\left[X_i\left(\frac{1}{n}\sum_{j=1}^{n}X_j\right)\right] + nE\left[\left(\frac{1}{n}\sum_{j=1}^{n}X_j\right)^2\right]\right) \\
&= \frac{1}{n}\left(nE[X^2] - \frac{2}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}E[X_iX_j] + nE\left[\frac{1}{n^2}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}X_iX_j\right)\right]\right) \\
&= \frac{1}{n}\left(nE[X^2] - \frac{2}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}E[X_iX_j] + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}E[X_iX_j]\right) \\
&= \frac{1}{n}\left(nE[X^2] - \frac{1}{n}\left(\sum_{i=1}^{n}E[X_i^2] + \sum_{i\neq j}E[X_iX_j]\right)\right) \\
&= \frac{1}{n}\left(nE[X^2] - E[X^2] - (n-1)(E[X])^2\right) \\
&= \frac{n-1}{n}\left(E[X^2] - (E[X])^2\right) = \frac{n-1}{n}Var(X).
\end{aligned}$$

Therefore, the unbiased estimator is given by:

$$S_{n-1}^2 := \frac{n}{n-1}S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$