The "n-1"

April 21, 2003

We will attempt to understand the $n-1$ in the denominator of the sample variance. The idea will be to try and understand how to estimate the mean and variance of a random variable X from a sample. Namely we assume we **don't know** $E(X)$ **or** $V(X)$ and wish to estimate them by performing n independent samples of the random variable X. We will denote each of these trials as X_i.

First we will attempt to estimate the mean from our n independent trials. We of course hope that

$$\bar{X} = \frac{\sum X_i}{n},$$

does the job, since this is our formula from the text for the sample mean! Notice, from the 1FMP

$$E\left(\frac{\sum X_i}{n}\right) = \frac{n}{n}E(X) = E(X)$$

and from the 2FMP

$$V\left(\frac{\sum X_i}{n}\right) = \frac{n}{n^2}V(X)$$

In other words,

$$Sd(X) = \frac{Sd(X)}{\sqrt{n}}.$$

If we accept the idea that "the chance that we are MANY standard deviation from the mean is small", then we are forced to conclude that for large n that $\bar{X}$ is near the expected value with high probability.

The next question is how to we estimate the $V(X)$ from our n independent samples. We might hope that

$$w^2 = \frac{\sum (X_i - \bar{X})^2}{n}$$

would do the job, since we are trying to compute the expected value $E((X - E(X))^2)$ and we just learned that $\bar{X}$ estimates $E(X)$. Well for this to work we would need that $E(w^2)$ is in fact equal to $V(X)$. Let's perform the computation.

$$\begin{aligned} E(w^2) &= E\left(\frac{\sum(X_i - \bar{X})^2}{n}\right) \quad (1) \\ &= \frac{1}{n}\sum_{i=1}^{n} E\left((X_i - \bar{X})^2\right) \quad (2) \end{aligned}$$

Where going from (1) to (2) requires using the 1FMP. Now we let $Y_i = X_i - E(X)$. We like these Y_i since $X_i - \bar{X} = Y_i - \bar{Y}$ with Y_i **itself** satisfying that $E(Y_i) = 0$. We now plug in the Y_i, "foil", and find

$$\begin{aligned} E(w^2) &= \frac{1}{n}\sum_{i=1}^{n} E\left((Y_i - \bar{Y})^2\right) \quad (3) \\ &= \frac{1}{n}\sum_{i=1}^{n} E\left(Y_i^2 - 2Y_i\bar{Y} + \bar{Y}^2\right). \quad (4) \end{aligned}$$

At this point we recall that $\bar{Y} = \frac{\sum Y_i}{n}$ and plug this in to find

$$E(w^2) = \frac{1}{n}\sum_{i=1}^{n} E\left(\left(1 - \frac{2}{n}\right)Y_i^2 + \frac{1}{n^2}\sum_{j=1}^{n} Y_j^2 + \left(\frac{2}{n^2} - \frac{2}{n}\right)\sum_{i \neq j} Y_iY_j\right).$$

By the 3FMP and the fact that $E(Y_i) = 0$, when $i \neq j$ we have that $E(Y_iY_j) = E(Y_i)E(Y_j) = 0$. Hence, upon utilizing the 1FMP to bring the expect value through the above sum, we find that the last term in this sum disappears. We are left with the following.

$$\begin{aligned} E(w^2) &= \frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{2}{n} + \frac{n}{n^2}\right)E(Y^2) \quad (5) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{1}{n}\right)E(Y^2) \quad (6) \\ &= \frac{n-1}{n}\sum_{i=1}^{n} E((X - E(X))^2) \quad (7) \\ &= \frac{n-1}{n}\sum_{i=1}^{n} V(X) \quad (8) \end{aligned}$$

Hence to get the correct expect value we need to use

$$\sigma^2 = \frac{n}{n-1} w^2$$

which, from this computation, indeed satisfies $E(\sigma^2) = V(X)$. Hence σ^2 is the formula needed to estimate the variance of X from a sample, and indeed

$$\sigma^2 = \frac{\sum_i (X_i - \bar{X})^2}{n-1},$$

is the formula stated in the text for the sample variance.

We are left wondering why did our initial guess failed to work? Very simply we need to use $\bar{x}$ in the formula for w^2, and $\bar{x}$ is **itself** only an estimate for $E(X)$. Notice, if we happen to KNOW $\mu = E(X)$, then indeed we could indeed use

$$w^2 = \frac{\sum (X_i - \mu)^2}{n},$$

to estimate $V(X)$. Occasionally , by symmetry, we will know that $\mu = 0$. In this case, we can (and should!) estimate the variance with w^2. USUALLY, however, we do not know μ, and hence must use the $n-1$.

Comment: On a practical level, it should be noted that $\frac{1}{n}$ and $\frac{1}{n-1}$ differ by a very small amount when n is large. For example, when $n \geq 20$ we find that

$$\frac{1}{n-1} - \frac{1}{n} \leq \frac{1}{19} - \frac{1}{20} < \frac{1}{365},$$

and since we are willing to say that a year has 365 days, this difference should be considered for most practical purposes as pretty darn little!