"Why is the denominator of the sample variance n-1 rather than n"

Alan Berger October 9, 2020

Available at https://github.com/AlanBerger/Statistical_Inference_GitHub_files

Introduction

This is a commonly asked question, answered in most statistics textbooks and on the web, see for example https://en.wikipedia.org/wiki/Bessel%27s_correction

In this article I'm going to give a *very detailed* explanation of the "algebra" that verifies this. While this is pretty lengthy, there will be less for the reader to "figure out" on their own. I am going to assume familiarity with the basic properties of the **expectation of a random variable** X, denoted by $\mathbf{E}(\mathbf{X})$ as covered here: "A basic explanation of the Expectation of a Random Variable" available at:
https://github.com/AlanBerger/Statistical_Inference_GitHub_files/blob/master/a-basic-explanation-of-the-Expectation-of-a-Random-Variable-UTF8.pdf
(located in the same GitHub repository as this article). In particular, for random variables X_1 and X_2 (defined on the same sample space S, for example X_1 and X_2 being clinical measurements on a set S of participants in a clinical trial of a vaccine) and constants c_1 and c_2 and c it is true that $E(c_1X_1 + c_2X_2 + c) = c_1E(X_1) + c_2E(X_2) + c$. Also, if X_1 and X_2 are independent, then $E(X_1X_2) = E(X_1)E(X_2)$.

The overview is that using $n-1$ makes the sample variance an **unbiased estimator** (defined below) of the variance of the population the samples came from (assuming independent random samples). It is also usually noted that this "comes from" the fact that the sample mean is used in defining the sample variance, so there is "1 less degree of freedom" so one should use $n-1$ rather than n.

Some notation. Let $x_1, \ldots, x_i, \ldots, x_n$ be n independent random samples from some distribution that has a finite mean and variance (so, for example, not from a Cauchy distribution). We can equivalently think of this as:
each of these samples x_i being a random sample from a random variable X_i, and the random variables $\{X_i\}$ are independent and have the same distribution (are identically distributed). For convenience, let X denote one of the X_i (it will not matter which one), or equivalently, a random variable with the same distribution as each of the $\{X_i\}$. Let μ denote the true mean of the distribution of each X_i and X, which using the expectation notation is $E(X_i) = E(X) = \mu$, and σ the standard deviation of X (and each X_i). The definition of the variance V, which is the square of the standard deviation is

$$\sigma^2 = E[(X-\mu)^2] \text{ which equals } E(X^2 - 2\mu X + \mu^2) = E(X^2) - 2\mu^2 + \mu^2 = E(X^2) - [E(X)]^2$$

We will use the notation $E(X)^2$ for $[E(X)]^2$. The **sample mean**, **m**, of the $\{x_i\}$ (also often denoted by $\hat{\mu}$ or $\bar{X}$) is

$$m = \sum_{i=1}^{n} x_i/n$$

The **sample variance** denoted by $\mathbf{S^2}$ is defined by

$$S^2 = \sum_{i=1}^{n} (x_i - m)^2/(n-1)$$

If the denominator were n instead of $n-1$, this would be the average of the square of the deviation of each x_i from the sample mean m. The sample standard deviation S is defined as the square root of S^2.

One can view the sample variance as a sample value of the function, given in the equation below, of the n random variables $\{X_i\}$

$$\mathcal{V} = \sum_{i=1}^{n} [X_i - \sum_{j=1}^{n} X_j/n]^2/(n-1) \qquad \text{equation 1}$$

so $\mathcal{V}$ is itself a random variable and we can ask what is its expected value. The idea is that if we take n independent random samples from X and calculate the sample variance S^2 and do this N times, the average of these N sample variances should approach σ^2 as N gets very large. If the expected value of $\mathcal{V}$, $E(\mathcal{V})$, equals σ^2, then S^2 is said to be an **unbiased estimator** for σ^2.

Demonstration that S^2 is an unbiased estimator for σ^2 (that is, $E(\mathcal{V}) = \sigma^2$)

Given the notation and definitions we have set up above, calculating $E(\mathcal{V})$ is a matter of using the properties of the expectation, and a fair amount of algebra bookkeeping. We need to expand the sum of the squares in the definition of $\mathcal{V}$ in equation 1 above. Define

$$M = \sum_{j=1}^{n} X_j/n \qquad \text{equation 2}$$

Then using equations 1 and 2,

$$E(\mathcal{V}) = \frac{1}{n-1} E(\sum_{i=1}^{n} (X_i - M)^2) = \frac{1}{n-1} E(\sum_{i=1}^{n} X_i^2 - 2\sum_{i=1}^{n} X_i M + \sum_{i=1}^{n} M^2) \qquad \text{equation 3}$$

Handling the first of the 3 terms in the expression on the right of the last equal sign in the equation above is fairly direct, recalling that X has the same distribution as each of the X_i, so this sum consists of n identical terms and is equal to

$$\frac{n}{n-1} E(X^2) \qquad \text{equation 4}$$

For the other two terms on the right of the last equal sign in equation 3 we use what M is, given in equation 2 above. These 2 terms can be combined:

$$\frac{1}{n-1} E(-2(\sum_{i=1}^{n} X_i) M + \sum_{i=1}^{n} M^2) = \frac{1}{n-1} E(-2(nM) * M + nM^2) = \frac{1}{n-1} E(-nM^2) = \frac{-n}{n-1} E(M^2)$$

Again using the definition of M in equation 2, we have

$$\frac{-n}{n-1} E(M^2) = \frac{-1}{(n-1)n} \sum_{j=1}^{n} \sum_{k=1}^{n} E(X_j X_k) = \frac{-1}{(n-1)n} [\sum_{j \neq k=1}^{n} E(X_j X_k) + \sum_{j=1}^{n} E(X_j^2)] \qquad \text{equation 5}$$

Now since the X_i are independent and have the same distribution as X, $E(X_j X_k)$ is equal $E(X)^2$ when $j \neq k$. Note there are n^2 pairs (j, k) for j and k ranging between 1 through n, and only n of these pairs (the

"diagonal pairs") have $j = k$. Hence the sum $\sum_{j \neq k=1}^{n}$ has $n^2 - n$ terms in it, and note since $n^2 - n = (n-1)n$, the value of the expression to the right of the last equal sign in equation 5 above is

$$-E(X)^2 + \frac{-1}{n-1}E(X^2)$$

Combining this expression with equation 4, we have

$$E(\mathcal{V}) = \frac{n}{n-1}E(X^2) - E(X)^2 + \frac{-1}{n-1}E(X^2) = E(X^2) - E(X)^2 = \sigma^2$$

so indeed the sample variance is an unbiased estimator for the population variance.

If, hypothetically, we knew the exact value of μ then using n rather than $n-1$ in the denominator of the sample variance would give an unbiased estimator

Given all the algebra we have gone through above, this can be verified with a relatively short "calculation". In this special case, the sample variance is set to

$$S_\mu^2 = \sum_{i=1}^{n}(x_i - \mu)^2/n$$

where now by assumption, μ is the (known) value of the mean of the population we are sampling, i.e., a constant (not an observed value from random samples). The calculation of $E(S_\mu^2)$ is then much less complicated than what we dealt with above (recall $E(X) = \mu$):

$$E(S_\mu^2) = \frac{1}{n}\sum_{i=1}^{n} E(X_i^2 - 2\mu X_i + \mu^2) = \frac{1}{n}(nE(X^2) - 2n\mu^2 + n\mu^2) = E(X^2) - E(X)^2 = \sigma^2$$

as claimed.

License and legal notice