Why the $N-1$ (Bessel's correction in sample variance calculations)?

January 21, 2022

Let us look at some commonly used formulae for mean and covariance. Suppose that we have a PDF, $p(x)$ with true mean μ and variance ν. Suppose that we have N samples from that PDF, x_i where $1 \leq i \leq N$. The sample mean is found from the simple formula:

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{1}$$

This is an example of use of the estimator (denoted with an over-line) that means average over a finite sample (N members) generalised as follows

$$\overline{\bullet} = \frac{1}{N}\sum_{i=1}^{N} \bullet_i. \tag{2}$$

In addition, the sample variance is the mean of the 'squared deviations from the mean' calculated from the same sample. As its name suggests, the mean squared deviation involves a 'mean of squares', so we again use formula (2). The deviation from the mean of any member of the population is $x_i - \mu$, so the mean squared deviation is

$$\overline{(x-\mu)^2} = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2. \tag{3}$$

In most real situations though, we don't know the true mean, μ, only our statistical estimate of it, $\overline{x}$ from (1). We could simply substitute $\mu \to \overline{x}$ in (3), but this gives a slightly biased answer. A more accurate result is gained by also substituting the normalisation factor $1/N$ with $1/(N-1)$ (or equivalently, multiplying by $N/(N-1)$):

$$\overline{(x-\mu)^2} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\overline{x})^2. \tag{4}$$

This is known as Bessel's correction. Many people know that this correction of the normalisation factor needs to be done, but not necessarily why.

The origin of the problem obviously lies in the fact that μ and $\overline{x}$ differ. Let us relate these via an error, ϵ_μ:

$$\overline{x} = \mu + \epsilon_\mu, \tag{5}$$

and substitute this into the formula that we obviously know to be true (3), and develop it:

$$\begin{aligned}
\overline{(x-\mu)^2} &= \frac{1}{N}\sum_{i=1}^{N}(x_i-\overline{x}+\epsilon_\mu)^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}\left[(x_i-\overline{x})^2+\epsilon_\mu^2+2\epsilon_\mu(x_i-\overline{x})\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}(x_i-\overline{x})^2+\epsilon_\mu^2+\frac{2\epsilon_\mu}{N}\sum_{i=1}^{N}(x_i-\overline{x}). \qquad (6)
\end{aligned}$$

The first summation has a form of the sample variance that we are interested in, but there are two more terms. The second term needs a bit more analysis (below) and the the last summation is automatically zero.

The second term, ϵ_μ^2, is equivalent to the following using (5):

$$\epsilon_\mu^2 = (\overline{x}-\mu)^2. \qquad (7)$$

We do not know the value that this term has for any particular sample, but what we can do is estimate its expected value based on a hypothetical large number of estimates of the mean. Let us define a new estimator that is based on such a large number, M, of estimates of repeated samples. The overline notation is already reserved for the estimator used over a single sample, so let us define our new estimator over multiple samples as $E(\bullet)$:

$$E(\bullet) = \lim_{M\to\infty}\frac{1}{M}\sum_{j=1}^{M}\bullet^{(j)}, \qquad (8)$$

where the bullet mark here indicates a quantity that is an output of (2). This estimator is only hypothetical though, invoked theoretically to understand the expected value of ϵ_μ^2 in (7). Let us then apply this estimator to ϵ_μ^2 (where we have M hypothetical estimates of ϵ_μ^2, each one labeled ${\epsilon_\mu^{(j)}}^2$):

$$\begin{aligned}
E(\epsilon_\mu^2) &= \lim_{M\to\infty}\frac{1}{M}\sum_{j=1}^{M}{\epsilon_\mu^{(j)}}^2 \\
&= \lim_{M\to\infty}\frac{1}{M}\sum_{j=1}^{M}\left(\overline{x}^{(j)}-\mu\right)^2 \\
&= \lim_{M\to\infty}\frac{1}{M}\sum_{j=1}^{M}\left(\frac{1}{N}\sum_{i=1}^{N}x_i^{(j)}-\frac{1}{N}\sum_{i=1}^{N}\mu\right)^2 \\
&\quad \lim_{M\to\infty}\frac{1}{M}\sum_{j=1}^{M}\left(\frac{1}{N}\sum_{i=1}^{N}\left[x_i^{(j)}-\mu\right]\right)^2, \qquad (9)
\end{aligned}$$

where $\overline{x}^{(j)}$ is the hypothetical sample mean based on the jth sample comprising the N hypothetical

members $x_i^{(j)}$. This formula can be developed further:

$$\begin{aligned} E(\epsilon_\mu^2) &= \lim_{M\to\infty} \frac{1}{M}\sum_{j=1}^{M}\left(\frac{1}{N^2}\sum_{i=1}^{N}\left[x_i^{(j)}-\mu\right]\sum_{i'=1}^{N}\left[x_{i'}^{(j)}-\mu\right]\right) \\ &= \lim_{M\to\infty}\left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{i'=1}^{N}\frac{1}{M}\sum_{j=1}^{M}\left[x_i^{(j)}-\mu\right]\left[x_{i'}^{(j)}-\mu\right]\right) \\ &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{i'=1}^{N} E\left(\left[x_i-\mu\right]\left[x_{i'}-\mu\right]\right). \end{aligned} \tag{10}$$

Here we have legally changed the positions of the summations and reintroduced our new estimator notation. This estimator is essentially the covariance between the ith and i'th members of any sample. Since our members are random draws from $p(x)$, these are uncorrelated, unless $i = i'$ in which case the sample member is the same, in which case the covariance is ν, the variance of the distribution. Hence

$$E\left(\left[x_i-\mu\right]\left[x_{i'}-\mu\right]\right) = \nu\delta_{ii'}, \tag{11}$$

where $\delta_{ii'}$ is the Kronecker delta function short for

$$\delta_{ii'} = \begin{cases} 1 & \text{if } i = i' \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

Feeding this back into (10) gives

$$\begin{aligned} E(\epsilon_\mu^2) &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{i'=1}^{N}\nu\delta_{ii'} \\ &= \frac{\nu}{N^2}\sum_{i=1}^{N} 1 \\ &= \frac{\nu}{N}. \end{aligned} \tag{13}$$

Equation (13) is an important and well-known result that the expected squared error in the sample mean is the variance of the PDF divided by N. Note that in the limit of a large sample, this error decays to zero.

We can now return to (6) with this result by replacing ϵ_μ^2 with its estimator, and noting that $\nu = \overline{(x-\mu)^2}$:

$$\begin{aligned} \overline{(x-\mu)^2} &\approx \frac{1}{N}\sum_{i=1}^{N}(x_i-\overline{x})^2 + \frac{\nu}{N} \\ &\approx \frac{1}{N}\sum_{i=1}^{N}(x_i-\overline{x})^2 + \frac{\overline{(x-\mu)^2}}{N}, \end{aligned}$$

so

$$\begin{aligned} \overline{(x-\mu)^2}\left(1-\frac{1}{N}\right) &\approx \frac{1}{N}\sum_{i=1}^{N}(x_i-\overline{x})^2 \\ \overline{(x-\mu)^2}\frac{N-1}{N} &\approx \frac{1}{N}\sum_{i=1}^{N}(x_i-\overline{x})^2 \\ \text{so } \overline{(x-\mu)^2} &\approx \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\overline{x})^2, \end{aligned} \tag{14}$$

thus demonstrating Bessel's correction.

Ross Bannister, Data Assimilation Research Centre and National Centre for Earth Observation, University of Reading, UK.