Bessel's Correction

When finding the variance of a population of size N, we know to compute

$$\sigma^2 = \frac{\sum(x-\mu)^2}{N}$$

where the summation is taken over all members x of the population, and μ is the population mean.

However, if we are attempting to estimate σ^2 using a sample, it turns out that simply replacing the population size N with the sample size n, replacing the population mean μ with the sample mean $\overline{x}$, and summing over all members x of the sample (instead of the population) yields a *biased* estimate of σ^2.

$$s^2_{biased} = \frac{\sum(x-\overline{x})^2}{n}$$

To intuitively see this, consider the extreme case where the sample size is $n = 1$, with the lone value in the sample being x_0. In this case, $\overline{x} = x_0$, making $s_{biased} = 0$. Unless the population consisted of N identical values (which is highly unlikely), estimating the population variance with 0 is clearly an underestimate.

In the more general case, note that the sample mean is not the same as the population mean. One's sample observations are naturally going to be closer on average to the sample mean than the population mean, resulting in the average $(x-\overline{x})^2$ value underestimating the average $(x-\mu)^2$ value. Thus, s^2_{biased} generally underestimates σ^2 -- with the difference between the two more pronounced when the sample size is small.

The good news is that this bias can be corrected!

However, the argument below to show this is a bit involved.

Before we begin this argument, let us make a couple of observations:

First, suppose that we randomly draw a sample of the form $\{x_1, x_2, \ldots, x_n\}$ from a population with mean μ. We can quickly show that $E\left[\,\overline{x}\,\right] = \mu$, using the properties of the expected value, as seen below:

$$\begin{aligned}
E\left[\,\overline{x}\,\right] &= E\left[\frac{x_1 + x_2 + \cdots + x_n}{n}\right] \\
&= \frac{1}{n} E\left[x_1 + x_2 + \cdots + x_n\right] \\
&= \frac{1}{n}\left(E\left[x_1\right] + E\left[x_2\right] + \cdots E\left[x_n\right]\right) \\
&= \frac{1}{n}(\mu + \mu + \cdots + \mu) \qquad \text{...where } \mu \text{ appears } n \text{ times} \\
&= \frac{1}{n} \cdot n\mu \\
&= \mu
\end{aligned}$$

Second, under the additional assumption that the population discussed above has variance σ^2, we can very similarly show that $Var[\,\overline{x}\,] = \sigma^2/n$.

Recall that

$$\begin{aligned}
Var[\,\overline{x}\,] &= Var\left[\frac{x_1 + x_2 + \cdots + x_n}{n}\right] \\
&= \frac{1}{n^2} Var\left[x_1 + x_2 + \cdots + x_n\right] \\
&= \frac{1}{n^2}\left(Var\left[x_1\right] + Var\left[x_2\right] + \cdots + Var\left[x_n\right]\right) \\
&= \frac{1}{n^2}(\sigma^2 + \sigma^2 + \cdots + \sigma^2) \qquad \text{...where } \sigma^2 \text{ appears } n \text{ times} \\
&= \frac{1}{n^2} \cdot n\sigma^2 \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

Having these results under our belt, we can turn our attention to the main argument...

We intend to show that

$$E[s^2_{biased}] = \left(\frac{n-1}{n}\right)\sigma^2$$

With the right side not being simply σ^2, we establish the biased nature of s^2_{biased} while simultaneously determining a factor to correct this bias.

As argument for our claim, consider the following:

$$\begin{aligned}
E[s^2_{biased}] &= E\left[\frac{1}{n}\cdot\sum_{i=1}^{n}(x_i-\overline{x})^2\right] \\
&= E\left[\frac{1}{n}\sum_{i=1}^{n}[(x_i-\mu)-(\overline{x}-\mu)]^2\right] \\
&= E\left[\frac{1}{n}\sum_{i=1}^{n}[(x_i-\mu)^2-2(\overline{x}-\mu)(x_i-\mu)+(\overline{x}-\mu)^2]\right] \\
&= E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2-\frac{2(\overline{x}-\mu)}{n}\sum_{i=1}^{n}(x_i-\mu)+\frac{1}{n}\sum_{i=1}^{n}(\overline{x}-\mu)^2\right] \\
&= E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2-2(\overline{x}-\mu)^2+\frac{1}{n}\sum_{i=1}^{n}(\overline{x}-\mu)^2\right] \\
&= E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2-2(\overline{x}-\mu)^2+\frac{1}{n}\cdot n\cdot(\overline{x}-\mu)^2\right] \\
&= E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2-2(\overline{x}-\mu)^2+(\overline{x}-\mu)^2\right] \\
&= E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2-(\overline{x}-\mu)^2\right] \\
&= E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2\right]-E\left[(\overline{x}-\mu)^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}E[(x_i-\mu)^2]-E\left[(\overline{x}-\mu)^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\sigma^2-E\left[(\overline{x}-\mu)^2\right] \\
&= \frac{1}{n}\cdot n\cdot\sigma^2-E\left[(\overline{x}-\mu)^2\right] \\
&= \sigma^2-E\left[(\overline{x}-\mu)^2\right] \\
&= \sigma^2-Var\left[\,\overline{x}\,\right] \\
&= \sigma^2-\frac{\sigma^2}{n}
\end{aligned}$$

$$= \quad \left(\frac{n-1}{n}\right)\sigma^2$$

Again, having established that

$$E[s^2_{biased}] = \left(\frac{n-1}{n}\right)\sigma^2$$

we can quickly construct an unbiased estimator, s^2, for σ^2 by multiplying s^2_{biased} by $n/(n+1)$, yielding

$$s^2 = \frac{\sum_{i=1}^n (x-\overline{x})^2}{n-1}$$

The unbiased nature of s^2 can be quickly confirmed by observing the following:

$$\begin{aligned}
E[s^2] &= E\left[\frac{\sum_{i=1}^n (x-\overline{x})^2}{n-1}\right] \\
&= \frac{1}{n-1}\cdot E\left[\sum_{i=1}^n (x-\overline{x})^2\right] \\
&= \frac{n}{n-1}\cdot E\left[\frac{1}{n}\cdot\sum_{i=1}^n (x-\overline{x})^2\right] \\
&= \frac{n}{n-1} E\left[s^2_{biased}\right] \\
&= \left(\frac{n}{n-1}\right)\left(\frac{n-1}{n}\right)\sigma^2 \\
&= \sigma^2
\end{aligned}$$