Bessel's Correction

The formula for calculating Population Variance is given by:

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

The formula for calculating Sample Variance is given by:

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

But why is 'n-1' in the denominator instead of 'n'?

This is known as Bessel's Correction factor. The main goal of Descriptive Statistics is to estimate the Population parameters using the Sample parameters. Population Variance can be better estimated when the Sample Variance includes 'n-1' instead of 'n'. Often, the mathematical proofs given for the same include 'Expectations', which in my opinion, deviate from straight-forward approach and rely on implicit assumptions. So, I decided to derive the same using no 'Expectations'.

Let us look at the original sample variance formula:

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

$$= \frac{1}{n} \sum_{i=1}^{n} \left((x_i - \mu) - (\bar{x} - \mu) \right)^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left((x_i - \mu)^2 - 2(x_i - \mu)(\bar{x} - \mu) + (\bar{x} - \mu)^2 \right)$$

$$= \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2 - \frac{1}{n} \sum_{i=1}^{n} 2(x_i - \mu)(\bar{x} - \mu) + \frac{1}{n} \sum_{i=1}^{n} (\bar{x} - \mu)^2$$

$$= \sigma^2 - \frac{2}{n} (\bar{x} - \mu) \sum_{i=1}^{n} (x_i - \mu) + \frac{\cancel{n}}{\cancel{n}} (\bar{x} - \mu)^2$$

$$= \sigma^2 - \frac{2}{n} (\bar{x} - \mu) \left[\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \mu \right] + (\bar{x} - \mu)^2$$

$$= \sigma^2 - \frac{2}{n} (\bar{x} - \mu)(n\bar{x} - n\mu) + (\bar{x} - \mu)^2$$

$$= \sigma^2 - \frac{2}{\cancel{n}} (\bar{x} - \mu) \cancel{n} (\bar{x} - \mu) + (\bar{x} - \mu)^2$$

$$= \sigma^2 - 2(\bar{x} - \mu)^2 + (\bar{x} - \mu)^2$$

$$= \sigma^2 - (\bar{x} - \mu)^2$$

$$= \sigma^2 - \left(\frac{1}{n} \sum_{i=1}^{n} x_i - \frac{1}{n} \sum_{i=1}^{n} \mu \right)^2$$

$$= \sigma^2 - \left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)\right)^2$$

$$= \sigma^2 - \frac{1}{n^2}\left(\sum_{i=1}^{n}(x_i - \mu)\right)^2$$

$$= \sigma^2 - \frac{1}{n^2}\left(\sum_{i=1}^{n}(x_i - \mu)^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(x_i - \mu)(x_j - \mu)\right)$$

$$\left[\because \left(\sum_{i=1}^{n} a_i\right)^2 = \sum_{i=1}^{n} a_i^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} a_i a_j\right]$$

Please refer to my previous paper for its proof.

$$s^2 = \sigma^2 - \frac{1}{n^2}\left(\sum_{i=1}^{n}(x_i - \mu)^2 + 0\right)$$ [$\because$ When the sample is unbiased, i.e., if X & Y are independent, then $Cov(X, Y) = 0$]

$$= \sigma^2 - \frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2\right)$$

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

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

$$s^2 = \sigma^2\left(\frac{n-1}{n}\right) \Rightarrow \sigma^2 = \left(\frac{n}{n-1}\right)s^2 = \left(\frac{\cancel{n}}{n-1}\right)\left(\frac{1}{\cancel{n}}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)$$

$$\therefore \sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Therefore, to estimate σ^2, s^2 is changed to $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.