Sample standard deviation, why divide by $(n-1)$?

Rémi Viné

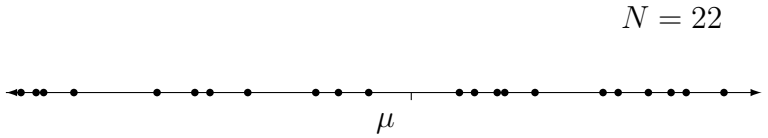
2022

1 The conundrum

Population standard deviation is defined as $\sigma = \frac{1}{N}\sum_i^N (x_i - \mu)^{0.5}$ but sample standard deviation is defined differently. Unlike for the mean, the difference is not limited to the mere fact that statistics replace parameters. There is a difference in the denominator: it is smaller for the unbiased sample standard deviation, so that the sample standard deviation is larger: $s = \frac{1}{n-1}\sum_i^n (x_i - \bar{x})^{0.5}$. Why is that so? When estimating the sample standard deviation, the difference $x_i - \bar{x}$ is missing the distance $\bar{x} - \mu$. This correction is named the Bessel's correction.

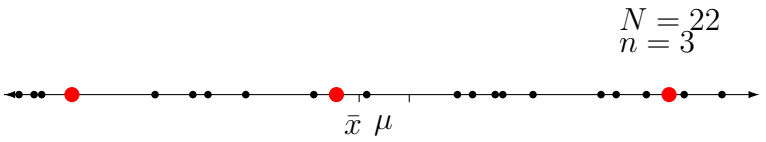
2 Intuition

Assume we have a population of 22 observations that are displayed on the following dot plot.



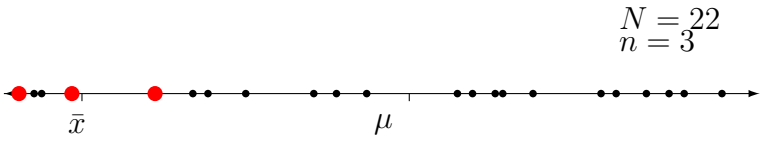
Dot plot (unstacked)

Assume now that a first sample of three observations is picked. In this case, μ is close to $\bar{x}$ and the distances between observations and the sample average are close to the ones in the entire population, so that $\sum_i^N (x_i - \mu)^2 \approx \sum_i^n (x_i - \bar{x})^2$



Dot plot (unstacked)

However, the sample of three can also easily be represented as such:



Dot plot (unstacked)

Clearly, in this case $\sum_i^N (x_i - \mu)^2 \neq \sum_i^n (x_i - \bar{x})^2$. The sample standard deviation would be much smaller than the population standard deviation. This has dramatic impact for inferences as it will lead to build confidence interval with unrealistically small margins of error and it will lead to wrong test statistics affecting hypothesis tests.

Dividing by $(n-1)$ inflates the sample standard deviation so that it is more likely to be closer to the population standard deviation. Remember that $lim_{n \to +\infty}(n-1) = n$, so that the difference is virtually non-existent for large samples.

3 Derivation

When estimating the sample standard deviation, the difference $x_i - \bar{x}$ is missing the distance $\bar{x} - \mu$. As long as variables are independent, which is assumed in most cases, the variance of a sum is equal to the sum of the variance (Bienaymé's identity), so that the bias introduced between the sample standard deviation and the population standard deviation shrinks to estimating the expected value of $(\bar{x} - \mu)^2$

$$E[\sigma^2 - s_n^2] = E[\frac{1}{n}\sum_i^n (x_i - \mu)^2 - \frac{1}{n}\sum_i^n (x_i - \bar{x})^2]$$

$$E[\sigma^2 - s_n^2] = E[\frac{1}{n}\sum_i^n (x_i^2 + \mu^2 - 2x_i\mu - x_i^2 - \bar{x}^2 + 2x_i\bar{x})]$$

$$E[\sigma^2 - s_n^2] = E[\mu^2 - \bar{x}^2 + 2(\bar{x} - \mu)\frac{1}{n}\sum_i^n x_i]$$

$$E[\sigma^2 - s_n^2] = E[\mu^2 - \bar{x}^2 + 2(\bar{x} - \mu)\bar{x}]$$

$$E[\sigma^2 - s_n^2] = E[\mu^2 - \bar{x}^2 + 2\bar{x}^2 - 2\mu\bar{x}]$$

$$E[\sigma^2 - s_n^2] = E[\mu^2 + \bar{x}^2 - 2\mu\bar{x}]$$

$$E[\sigma^2 - s_n^2] = E[(\bar{x} - \mu)^2]$$

$$E[\sigma^2 - s_n^2] = Var(\bar{x})$$

From Bienaymé's identity, we have $Var(\bar{x}) = Var(\frac{1}{n}\sum_i^n x_i)$. If variables are independent (or at least uncorrelated),this is equivalent to:

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

Therefore, the expected value of the gap of the (squared) deviations between the sample standard deviation and the population standard deviation is $\frac{\sigma}{n}$. Here again, it is easy to see that the bias shrinks with n, as, for any finite σ , $lim_{n\to+\infty}\frac{\sigma}{n} = 0^+$.

$$E[\sigma^2 - s_n^2] = \frac{\sigma^2}{n}$$

$$E[\sigma^2] - E[s_n^2] = \frac{\sigma^2}{n}$$

$$E[s_n^2] = \sigma^2 - \frac{\sigma^2}{n} = \frac{(n-1)}{n}\sigma^2$$

Therefore, the biased estimator s^2 needs to be "corrected" by $\frac{n}{(n-1)}$. Hence, the unbiased estimator, denoted s^2 is:

$$s^2 = \frac{n}{(n-1)} s_n^2$$

Since $s_n^2 = \frac{1}{n}\sum_i^n (x_i - \bar{x})^2 \Rightarrow s^2 = \frac{1}{(n-1)}\sum_i^n (x_i - \bar{x})^2$

□