Mr G's Little Booklet on

The difference between S_x and σ_x?

Robert Goodhand

Preface

If I enter data into my calculator the value for the standard deviation of Sx the sample is higher than the best estimate for the parent population $\sigma_{\mathbf{x}}$ when I would expect this to be the other way round. Have Texas got it wrong or am I misunderstanding something?

Standard Deviations

Suppose I enter { 1,1,2,3,4,4,5,6,7,7 } into list L1 in my calculator and then check out STAT/CALC/1-Var Stats. I get two standard deviations Sx = 2.261 and σx = 2.145. Now I know the standard deviation of the sample will be less than the standard deviation of the whole population the sample was drawn from - because in a bigger set there are likely to be more extremes - so my first observation is that these seem to be labelled back-to-front because I'm thinking σx the standard deviation of the parent population should be the higher value not the lower. In fact I've pondered this since the first stats calculators became commercially available around 1976 – and have commented on this to numerous students. A check soon confirms that σx results from division by n and Sx results from division by n–1.

Key to understanding this is that the calculator first has to determine its best estimate of the population mean μ, and it does this by calculating the actual mean of the sample $\bar{x}$.

Suppose we step back and just look at the mean first. In our example, the mean of is 4. Now if I deduct 4 from each value and, keeping track of positives and negatives, sum these to get 0. That then gives me an insight into a more subtle definition of the mean – it is that value at which the sum of the differences is a minimum.

The same idea applies to the standard deviation. I could for instance guess any value for the mean and from that calculate the standard deviation - don't worry whether we divide by n or n – 1. The fact is we get the lowest value for the standard deviation when we choose the mean of the actual numbers we are working with.

Now back to our sample. It is unlikely that those few values actually give the exact value of the mean of the parent population – in all likelihood I'm going to be a bit adrift. So my calculated value is too "neat" because I have used the most optimum value - the actual mean of the sample - and so my calculated value for the standard deviation is too low. The real mean will be a bit different from the $\bar{x}$ value

and so the real standard deviation – that is the standard deviation of the parent population - is a bit higher. And now I have the golden rule

If I am using a calculated estimate of the mean from the sample data then I divide by n – 1 because my calculated value is using an artificially optimum value of the mean.

So back to our TI - are those labels really back to front? I suppose I ought to have had more faith in Texas Instruments because they <u>are</u> correct – though you need to take careful note of what you have actually typed in before selecting σx or Sx.

Let me summarise in a chart.

Using TI	**I enter sample data**	**I enter the entire population**
σx		This gives me the standard deviation of the entire population because now $\bar{x}$ really is μ so I already have the optimum value and hence divide by n.
Sx	This is the sample standard deviation because I am estimating μ with $\bar{x}$ so I divide by n–1	

Sx is also the unbiased estimate of the standard deviation of the parent population. If the sample is from a parent population of <u>known</u> mean – say in this case 4. I mustn't be tempted to

use the simpler "mean of squares minus square of means" formula to calculate Sx. That formula is dependent on me using $\bar{x}$. I simply have to go back to first principles and sum the square of differences but I do now divide by n and not n – 1. That gives me a tighter value for standard deviation which is the reward from knowing the exact mean.

This is my intuitive explanation of why we sometimes divide by n – 1. The "correct" explanation is that with $\bar{x}$ the sum of differences is 0 so if I didn't happen to have one of the values I could calculate it from this fact alone. Therefore I have one fewer degree of freedom than I had when calculating the mean – because I could not then have deduced a missing value from the others and therefore divide by n, as expected.

For completeness we have two unidentified sections in the chart above. The top left MEI define as the root mean square deviation – rmsd - though an internet search doesn't give this term much universal usage. MEI take a strictly correct approach to the subject but it does lead to the more laboured explanation when using the "mean of squares less the square of means" to calculate standard deviation. Personally I think on first introduction to this topic a more simplified approach is better. I had to wait 40 years to understand it after all.

Finally let me summarise everything in a simplified chart

Using TI	**I enter sample data**	**I enter the entire population**
σx	root mean square deviation	population standard deviation
Sx	sample standard deviation	(nothing at all)

Which standard deviation you use depends on what you entered

And that's about it! ☺ *rg*

Appendix : Theory of Moments

The symbol E […] denotes the expected value of whatever appears in the brackets.

The expected value of a function $\theta(X)$ is denoted by $E[\ \theta(X)\ ]$ such that

$E[\ \theta(X)\ ] = {}_{\text{all } x}\Sigma\ \theta(x)\ f(x)$ for x discrete and ${}_{-\infty}\int^{+\infty}\theta\ (x)\ f(x)\ dx$ for x continuous

Population Moments

The k^{th} moment with respect to the origin of X is denoted by

$\mu'_k = E[\ X^k\ ]$

The full theory of moments leads to the proof of the central limit theorem which is the key theorem in statistics. Specifically

$\mu'_0 = E[\ X^0\ ]$

$\mu'_1 = E[\ X^1\ ]$ is commonly called the mean and denoted just by μ.

$\mu'_2 = E[\ X^2\ ]$

$\mu'_3 = E[\ X^3\ ]$

$\mu'_4 = E[\ X^4\ ]$

Central Moments

The k^{th} moment about the mean termed the *central moment* is denoted by

$\mu_k = E[\ (X-\mu)^k\ ]$

Specifically

$\mu_0 = E[\ (X-\mu)^0\ ] = 1$

$\mu_1 = E[\ (X-\mu)\ ] = 0$

$\mu_2 = E[\ (X-\mu)^2\ ]$ is commonly called the variance and denoted by σ^2

$\mu_3 = E[\ (X-\mu)^3\ ]$ and is related to the skew

$\mu_4 = E[\ (X-\mu)^4\ ]$ and is related to the kurtosis

Standardised (Central) Moments

The standardised moment is defined as μ^k/σ^k and is dimensionless

The first is 0, the second is 1, the third is skew and the fourth is kurtosis

Sample Moments

$m'_k = {}_{i=1}\Sigma^n X_i^k/n$. Specifically

$m'_0 = {}_{i=1}\Sigma^n X_i^0/n$

$m'_1 = {}_{i=1}\Sigma^n X_i/n$ is commonly written $\bar{X}$

$m'_2 = {}_{i=1}\Sigma^n X_i^2/n$

$m'_3 = {}_{i=1}\Sigma^n X_i^3/n$

$m'_4 = {}_{i=1}\Sigma^n X_i^4/n$

$m_k = {}_{i=1}\Sigma^n (X_i-\bar{X})^k/n$. Specifically

$m_0 = {}_{i=1}\Sigma^n(X_i - \bar{X})^0/n$

$m_1 = {}_{i=1}\Sigma^n(X_i - \bar{X})/n = 0$

$m_2 = {}_{i=1}\Sigma^n(X_i - \bar{X})^2/n$ is the mean square deviation referred to by MEI.

$m_3 = {}_{i=1}\Sigma^n(X_i - \bar{X})^3/n$

$m_4 = {}_{i=1}\Sigma^n(X_i - \bar{X})^4/n$

It can be rigorously shown that

$E[m'_k] = \mu'_k$ for all k (including k = 0)

and m'_k is an unbiased estimator of μ'_k for all k

Specifically $\bar{X}$ is an unbiased estimate for μ

But in general $E[m_k] \neq \mu_k$

(except for k = 0 and k = 1 which are trivial)

But we need to use s^2 rather than m_2 as an unbiased estimator of σ^2

Notation for standard deviation

MEI Newsletter Jan 2004

There have been lists of approved notation for many years, and when approving recent AS/A Level specifications (both for 2000 and for 2004) the QCA decided to insist on their use and no longer countenance any departures. This brings this country into line with internationally agreed notation. As a result, attention has been focused on the appropriate divisor when using sample data to calculate standard deviation, n – 1 or n.

Internationally, the symbol in standard use for the random variable Sample Variance, defined with divisor n –1 is S^2. A particular value of this random variable is consequently denoted by s^2, and calculated with a divisor n–1. This usage is written into British Standards (BS 3534-1,1993) and International Standards (ISO 3534).

These conventions give S^2 and S unambiguous meanings, together with their values, s^2 and s, in any instance. However they leave us with something of a vacuum in notation because there is no longer any notation for the quantities calculated with divisor n. At GCSE, and until recently at A

Level, it had been common to use the letter s to denote "standard deviation" calculated with divisor n, but this notation is no longer tenable. The letter σ in common with other Greek letters, is reserved for the population parameter and so it cannot be used either.

The same problem arises with the nomenclature. The terms variance and standard deviation refer to the outcomes of calculations using divisor n–1 but there are no names for the outcomes of the equivalent calculations with divisor n. Sometimes people try to overcome the difficulty by using the term "sample standard deviation" but this does not really help; it means different things to different people and so it is still unclear whether division by n –1 or n is implied.

So we really need new names and symbols for the measures obtained with divisor n. It is not just that without it statistics will continue to suffer from a lack of precision in its nomenclature and notation, but that it will make teaching very much easier.

In the 2nd edition of the MEI Statistics 1 textbook, accompanying the 2000 specification, we made a first tentative move in this direction by introducing a new notation, sd, to mean "standard deviation calculated with divisor n". In the 2004 specification, after a great deal of thought, we are going further and introducing new names and notation.

Mean square deviation,

$\text{msd} = {}^{1}/_{n} \Sigma (x - \bar{x})^2$

Variance

$s^2 = {}^{1}/_{n-1} \Sigma (x - \bar{x})^2$

Root mean square deviation

$\text{rmsd} = \sqrt{({}^{1}/_{n} \Sigma (x - \bar{x})^2)}$

Standard deviation

$s = \sqrt{s^2} = \sqrt{\{ {}^{1}/_{n-1} \Sigma (x - \bar{x})^2\}}$

Reprinted without their permission

s^2 is a single symbol. s^2 is the square of s.