January, 2004

MEI Newsletter extra
Standard deviation

Notation for standard deviation

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, $msd = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$	Variance, $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$
Root mean square deviation, $rmsd = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}$	Standard deviation, $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$

These changes will be reflected in the examination papers for the new specification and in the new (3rd edition) textbooks.

How does the divisor $n-1$ arise ?

When teaching you try to build upon students' existing concepts. That is how you hope to avoid them developing a black box mentality in which they apply formulae or procedures blindly, without any understanding. The idea of an "average", found by dividing by *n*, is firmly established in all students' minds. They have been doing it for years. Now, at the start of AS Level, they encounter a new measure for spread which involves an unexpected procedure, working with the squares of the deviations. Dividing the *n* squared quantities involved by the number *n*, at the appropriate stage, will seem a (reasonably) natural thing to do. How can you justify the use of $n-1$?

One way is to use the idea of *independent variables*. Think of two numbers, say 4 and 8. To find their mean you need both of the numbers, $\bar{x} = \frac{4+8}{2} = 6$. However in going on to find the spread, you might go on to consider the deviations from the mean. Since some of these are positive and some negative it is also quite natural to think of their absolute values. However, the absolute values of $(8-6)$ and $(6-8)$ are the same. At this stage there is only one independent variable. A similar argument can be extended to more than two numbers; there is always a loss of one independent variable in using deviation as the basis for a measure of spread.

There is an argument that the measure with divisor *n* should never have surfaced in the first place. It can be stated as follows.

> *"Calculating the mean of a set of n numbers involves the use of n independent values. If you then go on to calculate the standard deviation you have only n-1 independent values left. Thus standard deviation is a measure on n-1 independent variables, and so n-1 is the appropriate divisor."*

Many teachers, however, feel that it is better to overcome one barrier at a time and that there is already enough for students to worry about in coming to terms with the rest of the calculation algorithm. Perhaps this is why the divisor *n* is now widely taught.

The new MEI position is to accept that the use of divisor *n* is widespread but to give the results of those calculations their own names: *mean square deviation* (in place of variance) and *root mean square deviation* (in place of standard deviation). The equivalent notations are *msd* and *rmsd* respectively. These names are merely summaries of the calculation algorithms. A calculation leading to *variance* or *standard deviation* must have involved divisor $n-1$.

These terms appear in the specification for the new *Statistics 1* and students are expected both to be able to do the calculations long-hand and also to interpret the answers that their calculators give them.

The fact that using a divisor of $n-1$ produces an unbiased estimate of the population variance is a *consequence* of the calculation having been carried out appropriately but it is *not* the reason for doing it that way in the first place. Looking more widely, in all but the most elementary of cases theory usually requires the use of the measure obtained with divisor $n-1$, and this can be taken as confirmation that this is the appropriate measure for general use.

New formulae

As has already been mentioned, the name root mean square deviation is a summary of the algorithm underlying the formula

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

In this case the formula for standard deviation is obtained simply by replacing *n* by $n-1$ in the denominator.

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

However, this is not the only way to calculate *rmsd*. An alternative, and frequently used, formula is

$$rmsd = \sqrt{\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2\right)}$$

but in this case replacing *n* by $n-1$ does not yield the correct formula since the $\overline{x}^2$ is not part of the fraction. One possible way to overcome this difficulty is to convert root mean square deviation to standard deviation, *s*, using

$$s = rmsd\sqrt{\frac{n}{n-1}}$$

but this would not be good practice. It would not help students to understand the true nature of standard deviation, the key measure of spread, but would rather encourage them to apply a calculation algorithm blindly.

A much better way, however, is to use a different set of formulae throughout, constructed around the sum of squares, S_{xx}. This is given by

$$S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2 \ .$$

The various formulae are then written as follows.

Mean square deviation, $msd = \frac{S_{xx}}{n}$	Variance, $s^2 = \frac{S_{xx}}{n-1}$
Root mean square deviation, $rmsd = \sqrt{\frac{S_{xx}}{n}}$	Standard deviation, $s = \sqrt{\frac{S_{xx}}{n-1}}$

The value of S_{xx} is calculated using the more convenient of the two forms. Both forms require the use of the mean, $\overline{x}$, and large errors can arise from rounding it. This danger can be reduced by writing the second form of the formula as

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 \ .$$

When used in this form the mean is not calculated at all.

The new Student's Handbook and the new edition of the Statistics 1 textbook will both present the formulae in the sum of squares form. Their use is recommended.

Teaching standard deviation

Teachers may find it helpful to introduce root mean square deviation through the mean square deviation or *msd*, the equivalent of variance. It is less of a mouthful to say and, by being one step shorter, more readily understood. Proceeding to *rmsd* will then seem quite a natural step.

A full explanation of the next step, going to variance and standard deviation, requires the new concept of a number of independent variables. It is possible to look at the inclusion of this idea in *Statistics 1* rather negatively as another obstacle to be overcome, but it can also be seen as a good opportunity to give an informal introduction to a very important idea in statistics. Students will meet degrees of freedom doing the χ^2 test on contingency tables in the new *Statistics 2*, in the *t* test in *Statistics 3* and in analysis of variance in *Statistics 4*.

A simple way to introduce this idea is to discuss the marks that a class of *n* students achieve in a test. The most interesting feature for an individual student is almost certainly the *n*-1 comparisons between his/her own score and those of the other students. A more sophisticated example involves the use a spreadsheet to take many samples from a (manageably small) population and then demonstrate that division by *n* produces estimates which are, on average, too small whereas those from division by $n-1$ are, on average, just about right.

Knock-on effects to later units

In *Statistics 2* students meet correlation and regression and we recommend that the sum of squares notation is used here too. The possible formulae for S_{xx} have already been given; those for S_{yy} are equivalent. In addition the sum of products, S_{xy}, is now needed and this is given by $S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ with alternative forms $S_{xy} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$ and $S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)$.
Using these allows the formulae to be written simply, as follows.

Product moment correlation $r = \dfrac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$

Regression $y = a + bx$ *where* $b = \dfrac{S_{xy}}{S_{xx}}$ *and* $a = \bar{y} - b\bar{x}$

In more precise notation, $\hat{\alpha}$ and $\hat{\beta}$ are used rather than *a* and *b* to emphasise that these are estimates of underlying parameters.

These formulae are unaffected by the more precise definitions of variance and standard deviation now in *Statistics 1*. This is because the formulae do not involve division by either *n* or $n-1$; the quantities S_{xx}, S_{yy} and S_{xy} are used directly.

By contrast, the old-fashioned form of the correlation formula, $r = \dfrac{\text{cov}(x, y)}{\text{sd}(x)\text{sd}(y)}$, is no longer right. It should be $r = \dfrac{\text{cov}(x, y)}{\text{rmsd}(x)\text{rmsd}(y)}$ (where cov(*x*,*y*) is sometimes referred to as the sample covariance). Similarly, the old-fashioned form of the expression for the coefficient of *x* in the formula for the least squares regression line, $\dfrac{\text{cov}(x, y)}{\text{var}(x)}$, is no longer right (assuming, as is conventional, that $\text{cov}(x, y)$ is itself defined with divisor *n*); the denominator should now be msd(*x*). However, the continued use of these formulae is definitely not recommended.

Another situation which students meet in *Statistics 2* is where they compare the mean and variance of sample data to investigate whether a Poisson distribution might be appropriate. In this case it is indeed the sample variance, s^2, which should be used. In *Statistics 3*, the standard deviation, *s*, must be used in the *t* test, and not *rmsd*.

Other users

Standard deviation is no longer part of the GCSE Mathematics syllabus and consequently many students will meet it for the first time in *Statistics 1* but unfortunately this is not true for everyone. There is a totally mistaken view that standard deviation is needed for top marks in GCSE coursework and so some students are taught it there, and it also features in Statistics GCSE. In such cases it may be necessary to wipe the slate clean and start again.

Several other subjects use standard deviation. In some of them the convention that the term standard deviation is only used for the $n-1$ case is already established but others may require some missionary work on your part ! You will also have to contend with the various “notations” used on calculators many of which are highly misleading.

MEI Statistics Development Group *January 2004.*