

SUMMARY STATISTICS

Documents prepared for use in course B01.1305,
New York University, Stern School of Business

The standard deviation page 3

Computing a covariance and correlation from data page 6

© Gary Simon, 2007
revision date 8JAN 2007

Cover photo: Fire Island Deer, Long Island, 2005

≈≈≈ SUMMARY STATISTICS ≈≈≈

The empirical rules regarding standard deviations are these:

About $\frac{2}{3}$ of the values in a list are between $\bar{x} - s$ and $\bar{x} + s$.

About 95% of the values in a list are between $\bar{x} - 2s$ and $\bar{x} + 2s$.

Here \bar{x} represents the list mean and s is the list standard deviation. The “list” is usually a sample.

These can give us a good feel for what kind of answers we should get in assessing standard deviation, but certainly serious calculation will need a serious method. Here are four methods for getting a standard deviation.

We’ll talk first about the *sample* standard deviation, used when you think your data values represent some subset of a population. There is a distinction between *sample* standard deviation and *population* standard deviation. It’s not usually a big deal.

Method 1: This is the definition. The formula for the sample standard deviation s is

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

We often define $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$, so that $s = \sqrt{\frac{S_{xx}}{n-1}}$. The symbol S_{xx} is the “sample corrected sum of squares.” It’s a computational intermediary and has no direct interpretation of its own.

Example: Consider this list of 5 values: 28 32 31 29 39

Start by finding the total 159 and hence the average $\frac{159}{5} = 31.8$. Now note the deviations from average and their squares.

Value	Deviation	Deviation ²
28	-3.8	14.44
32	0.2	0.04
31	-0.8	0.64
29	-2.8	7.84
39	7.2	51.84
TOTAL	0.0	74.80

The value 74.80 is S_{xx} .

Complete the arithmetic as $s = \sqrt{\frac{S_{xx}}{n-1}} = \sqrt{\frac{74.80}{5-1}} = \sqrt{\frac{74.80}{4}} = \sqrt{18.7} \approx 4.32$

This arithmetic in this example is unrealistically easy. The computation for Method 1 is usually very messy and error-prone.

Method 2: The short-cut method. This uses the fact that

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

In our case, $n = 5$, $\sum_{i=1}^n x_i = 159$ and $\sum_{i=1}^n x_i^2 = 5,131$, giving

$$S_{xx} = 5,131 - \frac{(159)^2}{5} = 5,131 - 5,056.2 = 74.80$$

The rest of calculation proceeds as above. This is the method used by calculators and by most computer routines.

Method 3: Use a handheld calculator. This is generally *not* recommended unless n is small. By the way, calculators can also use the *population* form $\frac{S_{xx}}{N}$ (rather than $\frac{S_{xx}}{n-1}$).

We prefer the $n - 1$ form in nearly all cases. (It's conventional to use the upper case N for the size of a population, while lower case n is used for the size of a sample.)

Calculators that find standard deviations often have keys with s_{n-1} and s_N (or some variation). Usually with real data you want the $n - 1$ version, since you rarely have access to the whole population.

The difference between dividing by “count” versus “count minus 1” is numerically small.

If we had defined population S with divisor $N - 1$ (and there are several good reasons to do so) then the world would have been simpler.

Method 4: Use a computer program. This is the best technique.

The clerical task of entering numbers into a data file is never more difficult than the work for methods 1, 2, or 3. Moreover, numbers in a data file can be easily checked and edited.

If your data are worthy or serious attention, then the standard deviation is not the only thing you are going to be asking about them!

The quantity $\frac{S_{xx}}{n-1} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}}{n-1} = \frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2}{n-1}$ is called the *sample variance*. Thus

$$[\text{standard deviation}]^2 = \text{variance}$$

$$\text{standard deviation} = \sqrt{\text{variance}}$$

Finally, let's state that a *population* consists of the entire set of possible objects of interest. If we've got the numbers for the whole population, then the calculation of the population standard deviation would divide by the population size (usually denoted as N) and call the result σ (rather than s). That is,

$$\sigma = \sqrt{\frac{S_{xx}}{N}} = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}}$$

When working with the entire population, the mean $\frac{\sum_{i=1}^N x_i}{N}$ is usually denoted as μ (rather than \bar{x}).

If you can't decide whether you're looking at a population or a sample, then almost certainly you've got the sample. Use s , the one with divisor $n - 1$.

◆◆◆COMPUTING A COVARIANCE AND CORRELATION FROM DATA◆◆◆

Suppose that x_1, x_2, \dots, x_n is a list of values with mean \bar{x} . The sample variance of the x 's is defined as

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

The calculation of s_x^2 by hand or by computer is usually done through the formula

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

The square root is s_x , the sample standard deviation of the x 's.

If we had matching random variables y_1, y_2, \dots, y_n with mean \bar{y} , then we could make parallel calculations. By matching, we mean that x_1 and y_1 are collected from the same data point, x_2 and y_2 are collected from the same data point, and so on. Thus, the sample variance of the y 's is

$$s_y^2 = \frac{S_{yy}}{n-1} = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$$

In parallel with the above, the calculation of s_y^2 by hand or by computer is usually done through the formula

$$s_y^2 = \frac{S_{yy}}{n-1} = \frac{1}{n-1} \left[\sum_{i=1}^n y_i^2 - \frac{\left(\sum_{i=1}^n y_i \right)^2}{n} \right]$$

The square root is s_y , the sample standard deviation of the y 's.

◆◆◆COMPUTING A COVARIANCE AND CORRELATION FROM DATA◆◆◆

We can compute additional quantities which tells how the x 's and y 's tend to behave relative to each other. The first quantity is the *sample covariance*, defined as

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

The calculation of s_{xy} is usually done through the formula

$$s_{xy} = \frac{1}{n-1} \left[\sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{n} \right]$$

Covariances are not that easy to interpret. They are calculated in product units; for example, if the x 's are in dollars and the y 's in tons, it happens that s_{xy} is in units of dollar-tons. Statisticians routinely present this in the form of correlations. We define the *sample correlation* between the x 's and y 's to be

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

It is always true that $-1 \leq r_{xy} \leq +1$. Here are some quick interpretations for the correlation:

If $r_{xy} = +1$, then for some numbers a and b (with $b > 0$), it happens that $y_i = a + b x_i$. That is, a plot of the data would show that all points lie on a straight line of positive slope.

If $r_{xy} = -1$, then for some numbers a and b (with $b < 0$), it happens that $y_i = a + b x_i$. That is, a plot of the data would show that all points lie on a straight line of negative slope.

If $r_{xy} = 0$, then there is a complete absence of a straight-line relationship between the x 's and the y 's. A plot of the data would show aimless scatter, though it occasionally happens that a non-linear (curved) relationship corresponds to a correlation of zero. In practice, sample correlations are rarely exactly zero, though they can be very close to zero.

Some statistics constructed from a sample of data will have correlation exactly zero. Regression residuals and regression fitted values have this property.

In-between values of r_{xy} are taken as measures of the strength of the relationship. Thus a value $r_{xy} = 0.31$ would indicate a weak positive relationship between the x 's and the y 's, while a value $r_{xy} = -0.82$ would indicate a fairly strong negative relationship.