A note on standard deviation

Harley Weston,
Department of Mathematics and Statistics,
University of Regina

Suppose that you are in some course and have just received your grade on an exam. It is natural to ask how the rest of the class did on the exam so that you can put your grade in some context. Knowing the mean or median tells you the "center" or "middle" of the grades, but it would also be helpful to know some measure of the spread or variation in the grades.

Lets look at a small example. Suppose three classes of 5 students each write the same exam and the grades are:

Class 1	**Class 2**	**Class 3**
82	82	67
78	82	66
70	82	66
58	42	66
42	42	65

Each of these classes has a mean, $\bar{x}$, of 66 and yet there is great difference in the variation of the grades in each class. One measure of the variation is the range, which is the difference between the highest and lowest grades. In this example the range for the first two classes is 82 - 42 = 40 while the range for the third class is 67 - 65 = 2. The range is not a very good measure of variation here as classes 1 and 2 have the same range yet their variation seems to be quite different. One way to see this variation is to notice that in class 3 all the grades are very close to the mean, in class 1 some of the grades are close to the mean and some are far away and in class 2 all of the grades are a long way from the mean. It is this concept that leads to the definition of the standard deviation.

Lets look at class 1. For each student calculate the difference between the students grade and the mean.

Class 1	$x_i - \bar{x}$
82	16
78	12
70	4
58	-8
42	-24

The average of these differences could now be calculated as a measure of the variation, but this is zero. What is really needed is the distance from each grade to the mean not the difference. You could take the absolute value of each difference and then calculate the mean. This is called the mean deviation, i.e. mean deviation = $\frac{\sum|x_i - \bar{x}|}{n}$, where n is the number of students in the class. For class 1 this is 64/5 = 12.8. Another way to deal with the negative differences is to square each difference before adding.

Class 1	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
82	16	256
78	12	144
70	4	16
58	-8	64
42	-24	576

The sum of this column is 1056. To find what is called the standard deviation, s, divide this sum by n-1 and then, since the sum is in square units, take the square root. For class 1 this gives $s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{1056}{4}} = 16.2$

A similar calculation gives a standard deviation of 21.9 for class 2 and 0.7 for class 3. So for class 3, where the grades are all close to the mean, the standard deviation is quite small, for class 1, where the grades are spread out between 42 and 82, the standard deviation is considerably larger and for class 2, where all the grades are far from the mean, the standard deviation is larger still. The standard deviation is the quantity most commonly used by statisticians to measure the variation in a data set.

The reason that the denominator in the calculation of s is n-1 deserves a comment. To look at this lets change the example. Suppose that I am interested in the number of hours per day that high school students in North America spend doing their mathematics homework. The "population" of interest is all high school students in North America, a very large number of people. Lets call this number N. My real interest is the mean and standard deviation of this population. When talking about a population statisticians usually use Greek letters to designate these quantities, so the mean of the population is written $\mu = \frac{\sum X_i}{N}$, (μ is the Greek letter mu). Likewise the standard deviation is $\sigma = \sqrt{\frac{\sum(X_i - \mu)^2}{N}}$, ($\sigma$ is the Greek letter sigma). Notice that here the denomonator in the calculation is N.

Rather than trying to deal with this large population a statistician would usually select a "sample" of students, say n of them, and perform calculations on this smaller data set to estimate mu and sigma. Here n might be 25 or 30 or 100 or maybe even 1000, but certainly much smaller than N. To estimate mu it seems natural to use $\bar{X}$, the mean of the sample. Likewise to estimate sigma it seems reasonable to use $\sqrt{\frac{\sum(X_i - \bar{X})^2}{n}}$, but this quantity tends to underestimate sigma, particularly for small n. For this and other technical reasons the quantity $s = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}}$ is a usually preferred as the estimator to use for sigma.

If you have a calculator that computes the standard deviation it is a good exercise to see if it divides by n or n-1. Take the three number data set -1,0,1, calculate the standard deviation both ways by hand and then use your calculator to see which method it uses.

Footnote:

In the Spring of 2000 I received a request by a teacher to elaborate on the reasons why statisticians divide by n - 1 when calculating the sample variance. My reply was to describe an experiment that he could have his students perform in the classroom.

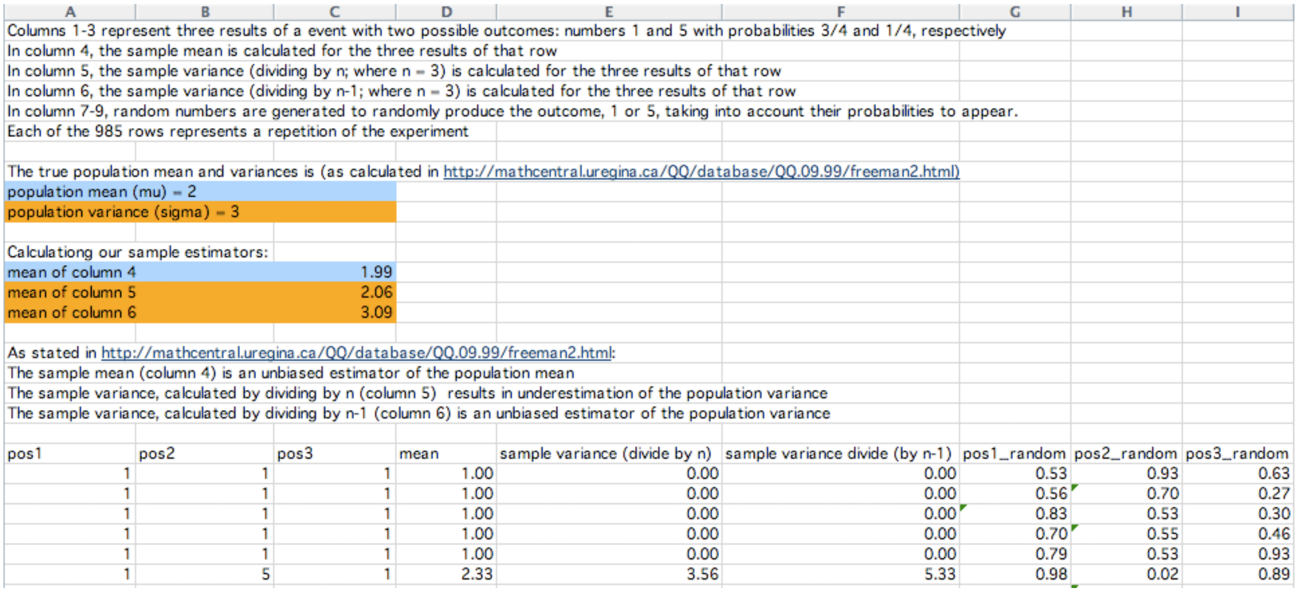
Harley

Footnote:

In December 2009 Javier Quílez Oliete, a doctoral student in Barcelona Spain, sent an Excel file where he simulated the experiment that I suggested in the previous footnote. He simulates repeatedly selecting a sample of size 3, with replacement, from a set with the two numbers 1 and 5 where the probability of selecting the 1 is 3/4 and the probability of selecting the 5 is 1/4. His simulation has 985 repetitions and for each of the 985 samples of size n = 3 he calculates the sample mean, the sample variance dividing by n, and the sample variance dividing by n - 1. He then compares these sample statistics to the population mean and variance.

Below is a snapshot of the first few lines of Janvier's spreadsheet. Click here to download Javier's spreadsheet.

Thanks Javier,
Harley



Columns 1-3 represent three results of a event with two possible outcomes: numbers 1 and 5 with probabilities 3/4 and 1/4, respectively
In column 4, the sample mean is calculated for the three results of that row
In column 5, the sample variance (dividing by n; where n = 3) is calculated for the three results of that row
In column 6, the sample variance (dividing by n-1; where n = 3) is calculated for the three results of that row
In column 7-9, random numbers are generated to randomly produce the outcome, 1 or 5, taking into account their probabilities to appear.
Each of the 985 rows represents a repetition of the experiment

The true population mean and variances is (as calculated in http://mathcentral.uregina.ca/QQ/database/QQ.09.99/freeman2.html)
population mean (mu) = 2
population variance (sigma) = 3

Calculationg our sample estimators:

mean of column 4	1.99
mean of column 5	2.06
mean of column 6	3.09

As stated in http://mathcentral.uregina.ca/QQ/database/QQ.09.99/freeman2.html:
The sample mean (column 4) is an unbiased estimator of the population mean
The sample variance, calculated by dividing by n (column 5) results in underestimation of the population variance
The sample variance, calculated by dividing by n-1 (column 6) is an unbiased estimator of the population variance

A	B	C	D	E	F	G	H	I
pos1	pos2	pos3	mean	sample variance (divide by n)	sample variance divide (by n-1)	pos1_random	pos2_random	pos3_random
1	1	1	1.00	0.00	0.00	0.53	0.93	0.63
1	1	1	1.00	0.00	0.00	0.56	0.70	0.27
1	1	1	1.00	0.00	0.00	0.83	0.53	0.30
1	1	1	1.00	0.00	0.00	0.70	0.55	0.46
1	1	1	1.00	0.00	0.00	0.79	0.53	0.93
1	5	1	2.33	3.56	5.33	0.98	0.02	0.89

Go to Math Central

To return to the previous page use your browser's back button.