Social Studies 201

October 8, 2004

Standard Deviation and **Variance**

See text, section 5.9, pp. 225-258.

Note: The examples in these notes may be different than used in class. However, the examples are similar and the methods used are identical to what was presented in class.

The variance and the standard deviation are presented together in these notes. Each originates from the same set of calculations, the difference of individual values from the mean of the distribution.

In these notes, and in many statistical texts, the variance is given the symbol s^2 and the standard deviation is given the symbol s. As might be recognized, the variance is the square of the standard deviation, that is, the standard deviation multiplied by itself. Alternatively, beginning with the variance, the standard deviation is the square root of the variance. That is, the standard deviation is the number, which when multiplied by itself, produces the variance.

Unlike the previous two measures of variation (range and interquartile range), there is no easily understood or intuitive explanation for the standard deviation and variance. Hopefully you will develop an understanding and appreciation of what the standard deviation and variance represent as you work your way through these notes.

In you have difficulty understanding these measures, you can always think of them as simply measures of variation where a larger value indicates that a distribution is more varied and a smaller value indicates that a distribution is less varied. For example, suppose two classes have grade distributions where the standard deviations are $s = 5$ in class A and $s = 10$ in class B. This indicates that the grades for class B are twice as varied as those for class A. While the value of 5 or 10 for a standard deviation may be difficult to interpret, by comparing standard deviations for different distributions, you should develop some appreciation of how to use this measure.

Each of the standard deviation and variance is constructed using the value of each case in a distribution. Recall that each of the range and interquartile range used only two values. In contrast, the value of each case is employed

when obtaining the standard deviation and variance. From these values, the first step is to calculate the mean of the distribution. The next step is to determine the difference in value of each case from the mean – these differences are often referred to as the deviation about the mean. The standard deviation is a sort of "average" or "standard" of these differences from the mean; the variance is an average of the squares of these differences from the mean. The exact formula for each measure is provided in the notes that follow.

Scale of measurement. The standard deviation and variance can be calculated for variables that have an interval or ratio scale of measurement (such as height, age, income). This is the same condition that was used to obtain the mean. The reason for requiring this level of measurement is that differences of values of a variable must be meaningful in order to calculate the mean and differences about the mean. For variables with no more than an ordinal level of measurement, such as attitude scales, the standard deviation and variance are often calculated and used. In this case, caution should be exercised. What the statistical analysis is assuming in this situation is that equal units on the ordinal scale represent equal units of the phenomenon. In the case of variable with no more than a nominal scale (ethnicity, religion, political party preferred), obtaining the mean, standard deviation, or variance would make not sense.

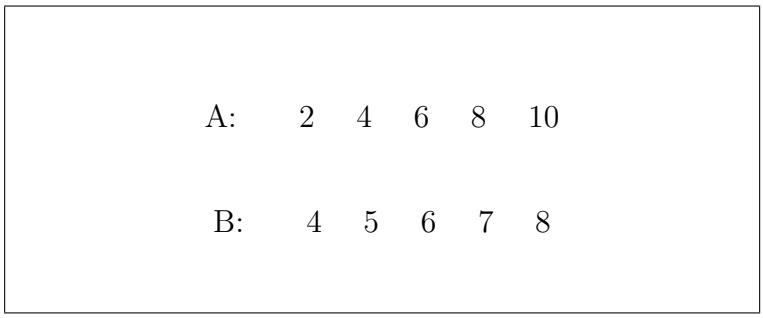
Hypothetical data sets. Each of the standard deviation and variance is constructed from differences of individual values from the mean. A data set with diverse values of a variable will have large differences of each value from the mean, producing a large standard deviation and variance. In contrast, a data set with less diverse values will have smaller differences of each value from the mean, producing a smaller standard deviation and variance.

Consider two data sets, A and B, each with five cases, as illustrated in Figure 1. After a quick examination of data sets A and B, it is apparent that A is twice as varied as is B. Both data sets have the same mean, a value of 6. But A is more varied than B.

For data set A, the first two value is 2, four units less than the mean of 6. For B, the first value is 4, two units below the mean of 6. Similarly, the second value in A is 4, two units below the mean; for B, the second value is 5, one unit below the mean. The middle value of each data set, 6, is right at the mean. For the values above the mean, the situation is the same as

for below the mean. That is, the values for A are twice as distant from the mean as the corresponding values for B. This demonstrates that A can be considered to be twice as varied as A. This is supported by a calculation of the range – for A the range is $10 - 2 = 8$ and for B it is $8 - 4 = 4$.

Figure 1: Variability of two small data sets A and B



While the formula for the standard deviation and variance have not yet been discussed, their values for data sets A and B are given in Table 1. While the exact value of the standard deviation is likely to be a mystery at this point, observe that the standard deviation for A (3.16) is double the standard deviation for B (1.58). This should make sense, given the discussion above.

Table 1: Summary statistics for hypothetical data sets A and B

	Value of statistic for data set:	
Measure	A	B
Mean	6	6
Standard deviation	3.16	1.58
Variance	10.00	2.50
Range	8	4

The variance is much larger for A than for B, again consistent with the fact that the values in A are more spread out than in B. But the variance for A is four times that for B, making the relative variances seem too different from each other. As a result, for much statistical analysis, the standard deviation is the preferable measure of variation. The variance has many applications in statistical analysis, but is less used when comparing two distributions. As will be seen in the following notes, the variance is often calculated first, and then the standard deviation, the square root of the variance, is obtained from the variance.

Definitions and formulae

There are a number of formulae for the standard deviation and variance, depending on how the data are organized. The first formula is for ungrouped data – a list of values of a variable. Later in these notes there are different formulae for data which are organized or grouped into categories or intervals, or what are referred to as grouped data. Formulae, examples, and explanations for each of these follow.

Symbols. In Module 4, the following symbols are used

$$\text{Variance} = s^2$$

$$\text{Standard deviation} = s$$

That is, the variance is the square of the standard deviation and the standard deviation is the square root of the variance.

Square root. The square root of a value X is written $\sqrt{X}$ and is another number, which when multiplied by itself, produces the original value. That is $\sqrt{X} \times \sqrt{X} = X$. Examples are:

$$\sqrt{25} = 5 \quad \text{since} \quad 5 \times 5 = 25.$$

$$\sqrt{10} = 3.162 \quad \text{since} \quad 3.162 \times 3.162 = 10.$$

$$\sqrt{0.58} = 0.762 \quad \text{since} \quad 0.762 \times 0.762 = 0.58.$$

Most calculators have a button with a symbol such as $\sqrt{x}$ or $\sqrt{}$ on it. In order to obtain the square root of a value, enter the value into the calculator, so the value appears in the display, and press the square root button. You can always check to make sure you have the correct square root by multiplying the square root by itself. This should produce the original value.

Ungrouped data

See text, pp. 225-237.

Definition. A variable X that takes on values $X_1, X_2, X_3, ...X_n$ has mean

$$\bar{X} = \Sigma X/n.$$

The variance is

$$s^2 = \frac{\Sigma(X - \bar{X})^2}{n-1}$$

and the standard deviation is the square root of the variance or

$$s = \sqrt{s^2} = \sqrt{\frac{\Sigma(X - \bar{X})^2}{n-1}}$$

These formulae are intimidating if you are not familiar with algebraic notation. However, if you follow the steps listed here, this will help guide you through the calculation of the variance and standard deviation. These steps will be used in the examples that follow. The tabular format introduced in the examples should also help you with the calculations.

The procedure for calculating the variance and standard deviation for ungrouped data is as follows.

- First sum up all the values of the variable X, divide this by n and obtain the mean of $\bar{X} = \Sigma X/n$. From Module 3, you should be familiar with this first step.

- Next subtract each individual value of X from the mean to obtain the differences about the mean. These are the values $X_1 - \bar{X}$, $X_2 - \bar{X}$, $X_2 - \bar{X}$, ... $X_n - \bar{X}$. These differences are often referred to as deviations about the mean. Some of these will be negative values, others will be positive values, and some might be zero.

- Multiply each of these differences about the mean by themselves, that is, obtain the squares of each of these differences. This produces the values $(X_1 - \bar{X})^2$, $(X_2 - \bar{X})^2$, $(X_3 - \bar{X})^2$, ... $(X_n - \bar{X})^2$.

- Add all the squares of the differences about the mean to obtain

$$\Sigma(X - \bar{X})^2.$$

- Divide this sum of the square of the differences about the mean by $n-1$ to obtain the variance s^2, that is,

$$\text{Variance} = s^2 = \frac{\Sigma(X - \bar{X})^2}{n-1}.$$

  This is amean of these squares of the differences about the mean, although $n-1$, rather than n, is used in the denominator.

- The standard deviation, s, is the square root of the variance of the last step, that is,

$$\text{Standard deviation} = s = \sqrt{s^2} = \sqrt{\frac{\Sigma(X - \bar{X})^2}{n-1}}$$

Work through some of the following examples and hopefully you will find them helpful in learning to obtain the variance and standard deviation.

Example 4.5 – Standard deviation for data sets A and B. For each of data sets A and B, calculate the variance and standard deviation.

Answer. Using the steps outlined above, the variance and standard deviation of data set A is obtained as follows.

- The sum of all the values of the variable is $\Sigma X = 2 + 4 + 6 + 8 + 10 = 30$. There are $n = 5$ values and the mean is $\bar{X} = \Sigma X/n = 30/5 = 6$.
- The differences of each individual value of X from the mean are

$$X_1 - \bar{X} = 2 = 6 = -4$$
$$X_2 - \bar{X} = 4 - 6 = -2$$
$$X_3 - \bar{X} = 6 - 6 = 0$$
$$X_4 - \bar{X} = 8 - 6 = 2$$
$$X_5 - \bar{X} = 10 - 6 = 4$$

- Multiply each of these differences about the mean by themselves, that is, obtain the squares of each of these differences. This produces the values

$$-4^2 = 16$$

$$-2^2 = 4$$

$$0^2 = 0$$

$$2^2 = 4$$

$$4^2 = 16$$

- Sum the squares of the differences about the mean to obtain

$$\Sigma(X - \bar{X})^2 = 16 + 4 + 0 + 4 + 16 = 40.$$

- Divide this sum of the square of the differences about the mean by $n - 1 = 5 - 1 = 4$ to obtain the variance, s^2. The variance is

$$s^2 = \frac{\Sigma(X - \bar{X})^2}{n - 1} = \frac{40}{4} = 10.$$

- The standard deviation, s, is the square root of the variance of the last step, that is,

$$s = \sqrt{s^2} = \sqrt{\frac{\Sigma(X - \bar{X})^2}{n - 1}} = \sqrt{10} = 3.162.$$

Tabular format. These calculations are usually performed using a table such as Table 2. Following is a description of how to construct and use the tabular format to obtain the variance and standard deviation for data set A.

In tabular format, you will need three columns, a column for X, a column for $X - \bar{X}$, and a column for the square of these deviations about the mean, $(X - \bar{X})^2$.

Table 2: Calculations for standard deviation, data set A

X	$X - \bar{X}$	$(X - \bar{X})^2$
2	-4	16
4	-2	4
6	0	0
8	2	4
10	4	16
30	0.0	40

Begin by placing all the values of the variable vertically in the first column. Since you must obtain the mean prior to calculating the deviations about the mean, sum the values in the first column to obtain ΣX. Divide this by the number of cases to obtain the mean.

In Table 2, the sum of the entries in the first column is 30 and the mean is

$$\bar{X} = \frac{\Sigma X}{n} = 30/5 = 6.$$

Then the column $X - \bar{X}$ is obtained by subtracting the mean from each of the individual values. These differences are squared and placed in the proper row in the third column.

For example in the first row, $X = 2$ and this is $X - \bar{X} = 2 - 6 = -4$ less than the mean of 6. This deviation of -4 is placed in the second column of the first row; its square is $-4 \times -4 = 16$ and this is placed in the third column of the first row.

For the second row, $X = 4$ and the deviation about the mean is $4 = 6 = -2$. This value is place in the second column and its square $(-2 \times -2 = 4)$ is placed in the third column.

Perform the same operations for each row of the table and enter each value in the appropriate row and column.

The sum of the third column is $\Sigma(X - \bar{X})^2 = 40$. This sum is divided by $n - 1 = 5 - 1 = 4$ to produce the variance $s^2 =$

$40/4 = 10$. Finally, the standard deviation is the square root of the variance, that is, $s = \sqrt{10} = 3.162$. The variance is 10 units of the original variable and the standard deviation is 3.16 units (rounded to two decimal places).

Exercise: Derive the variance and standard deviation for data set B of Figure 1. Show they are equal to the values listed in Table 1. That is obtain $s^2 = 2.50$ and $s = 1.58$.

Additional notes on variance and standard deviation

- **Deviations about the mean**. Both the variance and standard deviation are constructed from differences of individual values of X from the mean, the values $X - \bar{X}$. Statisticians often refer to these as **deviations about the mean**. Large values for these deviations produce a large variance and standard deviation; small values for these deviations produce a small variance and standard deviation.

- **Standard deviation and variance as averages**. At the beginning of this section, it was indicated that the standard deviation and variance are types of averages. While the variance might not initially be recognizable as an average, it is a mean since it is a sum divided by $n - 1$. This differs slightly from the mean of Module 3, in that the denominator is $n-1$ rather than n. This is for technical and statistical reasons – always use $n - 1$ but think of the variance as the mean of squares of deviations about the mean. While it may be a little difficult to interpret this mean, larger values indicate greater variation and smaller values indicate smaller variation.

  The standard deviation is also an average of sorts. The standard deviation is a square root of the variance – the square root of the average of squares of deviations about the mean. While this is not considered a normal type of mean, in some ways it is an "average" in representing an average deviation about the mean. Rather than referring to it as the average, the term "standard" is used. In the example of data set A, the individual deviations were $-4, -2, 0, 2, 4$ and the standard deviation was 3.162. This latter value can be thought of as a rough "average" or "standard" of these individual deviations.

- **Unit**. The unit for the standard deviation is the same as the unit for the variable X. For example, if the values 2, 4, 6, 8, 10 of data set A and 4, 5, 6, 7, 8 of data set B represented the ages of children in years, the standard deviation would also be in units of years.

  The variance is the square of the standard deviation so is in the original unit squared. If the case of data sets A and B, if these were ages in years, then the unit for the variance would be years squared. Because "years squared" is an unfamiliar unit with which to work, it is generally preferable to work with the standard deviation – at least it is in a familiar unit. That is, if variance is in units of "years squared" the square root of this is in "years," the familiar unit we use to measure age.

- **Tabular format**. Rather than proceeding directly through the formula or the bulleted items above, most analysts consider it more efficient to obtain the standard deviation using a table or tabular format. In the examples that follow, a tabular format is used.

- **Using a calculator**. Some calculators have procedures permitting you to calculate the standard deviation entirely on the calculator. If you use a calculator for this, check the standard deviation for a few examples or exercises in these notes, to ensure that you are using the proper procedure on the calculator. Some calculators use a slightly different formula than presented here. For this course, the formulae you are to use are those presented in these notes.

  Example 4.6 – Variation in homicides by province. Calculate the variance and standard deviation for each of the two groups of provinces in Table 3. Write a short note contrasting the variability in the two sets of provinces. Speculate why the variability might be so different for the two data sets.

  Answer. The variable is number of homicides, a ratio scale variable, since each homicide is equal in number to any other homicide and zero homicides represents none at all. As a result, the variance and standard deviation can be meaningfully obtained.

  A tabular format is used for the answer in this example. The calculations for determining the standard deviation of the num-

Table 3: Number of homicides in two areas of Canada, 2001

Provinces east of Manitoba		Western Provinces	
Province	No. of homicides	Province	No. of homicides
Nfld	1	Manitoba	34
PEI	2	Sask.	27
NS	9	Alberta	70
NB	8	B.C.	85
Quebec	140		
Ontario	170		

ber of homicides for provinces east of Manitoba are contained in Table 4 and for the four western provinces in Table 5. In each table, a column is provided for the X value, then a column for the differences of each X value from the mean, $X - \bar{X}$. Finally, the third column contains the squares of these differences from the mean, $(X - \bar{X})^2$. From the entries into these columns and the sum of these columns, the variance and standard deviation can be relatively easily calculated.

Table 4: Calculations for Mean and Standard Deviation, Homicides in provinces east of Manitoba

X	$X - \bar{X}$	$(X - \bar{X})^2$
1	-54	2,916
2	-53	2,809
9	-46	2,116
8	-47	2,209
140	85	7,225
170	115	13,225
330	0.0	30,500

Provinces east of Manitoba. The first step is to calculate the

mean. For the provinces east of Manitoba in Table 4, the sum of the number of homicides is 330 (first column) and there are $n = 6$ provinces. The mean number of homicides across these provinces is

$$\Sigma X/n = 330/6 = 55$$

.

The next step is to subtract the mean from each value of the variable. These values are placed in the second column of Table 4. For Newfoundland, there was only one homicide and the deviation about the mean number of homicides is $1 - 55 = -54$, that is 54 homicides below the mean. For Quebec, there were 140 homicides, or $140 - 55 = 85$ homicides above the mean. The other values of the deviations about the mean are calculated in a similar manner.

The next step is use each of the differences about the mean in the second column and square them, that is, multiply them by themselves. These squares are placed in the appropriate row of the third column. This results in the squares of the differences from the mean, the $(X - \bar{X})^2$ values, of the third column of Table 4. For example, for Newfoundland, with 54 homicides less than the mean, the squared difference is $-54 \times -54 = 2,916$. Quebec has 85 homicides above the mean and this squared difference is $85 \times 85 = 7,225$. Other entries in the third column are calculated in a similar manner. The entries in this column are then added together and, in the total of the third column is 30,500.

The sums in the last row of the table are now used to determe the variance. For the six provinces east of Manitoba, the variance is the sum of the third column divided by $n - 1$, that is,

$$s^2 = \frac{\Sigma(X - \bar{X})^2}{n - 1} = \frac{30,500}{6 - 1} = \frac{30,500}{5} = 6,100.$$

The standard deviation is the square root of the variance, that is,

$$s = \sqrt{s^2} = \sqrt{6,100} = 78.102$$

or 78.1 homicides.

Western provinces. The calculations for the four western provinces are contained in Table 5. The procedure is the same as fore the eastern provinces, with the mean number of homicides across the four western provinces being

$$\Sigma X/n = 216/4 = 54.$$

Note that this is almost exactly equal to the mean for the eastern provinces so there is little difference in the average homicide experience for the two sets of provinces.

Table 5: Calculations for mean and standard Deviation, homicides in western Canada

X	$X - \bar{X}$	$(X - \bar{X})^2$
34	-20	400
27	-27	729
70	16	256
85	31	961
216	0.0	2,346

For this table, again obtain the deviations about the mean, enter each in the second column, square this value, and place the square in the third column. Summing the third column gives a value of 2,346 for the sum of the squares of the deviations about the mean. Entering the sums in the formula for the variance gives

$$s^2 = \frac{\Sigma(X - \bar{X})^2}{n-1} = \frac{2,346}{4-1} = \frac{2,346}{3} = 782.$$

The standard deviation is the square root of the variance, that is,

$$s = \sqrt{s^2} = \sqrt{782} = 27.964$$

or 28.0 homicides.

Comparison. A quick glance at the number of homicides by province in the original Table 3 demonstrates that the variation

Table 6: Summary statistics for homicides in two areas of Canada

Measure	Eastern Provinces	Western Provinces
$\bar{X}$	55	54
s^2	6,100	782
s	78.1	28.0
Range	169	58

in homicides by province is greater in the six eastern provinces than in the four western provinces. The range of homicides in the provinces east of Manitoba is from 1 to 170 or $170 - 1 = 169$. In contrast, the range of homicides in the four western provinces is from 27 to 85 or $85 - 27 = 58$. The four numbers for Western Canada are closer to each other than are the six numbers for provinces east of Manitoba. We thus expect to find a larger variance and standard deviation for the six central and eastern provinces than for the four western provinces.

The measures of variation are summarized in Table 6. From this summary, the much greater variability for the eastern, as compared with western, provinces is apparent. The standard deviation and range are each about two and one-half to three times larger for the eastern than the western provinces. The mean number of homicides is very similar, but variation among the provinces is much greater in the east than the west, at least in terms of number of homicides.

Speculating about possible reasons for this result, differences in the homicide rate or differences in the population would likely be the reason for the different variability. The homicide rates for these provinces was given in Exercise 3.xxxx and there is some difference in these across provinces. The main reason for the different number of homicides though is likely to be the different populations for the different provinces. Ontario and Quebec each have large populations and the four Atlantic provinces each have

small populations. It is no surprise that the number of homicides differs greatly across these six provinces.

In contrast, the four western provinces do not differ so much in population, so the number of homicides would not be expected to differ so greately in the four western provinces. British Columbia and Alberta have larger populations than Saskatchewan and Manitoba, but the difference in population size among these four provinces is not so great as across the six eastern provinces.

Finally, note that the unit for each of the standard deviation and range is the number of homicides. The unit for the variance is homicides squared, a difficult unit to comprehend.

Alternative formula for ungrouped data

See text, p. 231 and following.

For ungrouped data (a list of numbers) the following formula is often considered a more efficient way to obtain the variance and standard deviation. At first glance, this formula may appear more intimidating than the earlier formula. Work through the example following the definition and you will find the use of this formula can probably save you time.

Definition. For a variable X that takes on values $X_1, X_2, X_3, ...X_n$ the variance is

$$s^2 = \frac{1}{n-1}\left[\Sigma X^2 - \frac{(\Sigma X)^2}{n}\right].$$

The standard deviation is the square root of the variance or

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\left[\Sigma X^2 - \frac{(\Sigma X)^2}{n}\right]}$$

These formulas produce the same variance and standard deviation as earlier (see text, pp. 235-6 for a proof).

Using these formulae, the procedure is to list all the n values of the variable X and then square all these values to produce X^2 for each of the original values. The list of X and X^2 values are then summed to produce

ΣX and ΣX^2. The latter sum is the first entry in the large bracket and ΣX is the entry in the numerator of the last part of the expression in the large bracket of

$$s^2 = \frac{1}{n-1}\left[\Sigma X^2 - \frac{(\Sigma X)^2}{n}\right].$$

Using this approach with a tabular format, there need only be two columns, an X column and a square of the Xs column, X^2. The sums of these columns are calculated and entered into the above expression. An example follows.

Example 4.7 – Variation in homicides using alternative formula. Use the data in Table 3 to obtain the variance and standard deviation for the number of homicides across the four western provinces.

Answer. Calculations, using the alternative formula, are contained in Table 7.

Table 7: Alternative calculations for variance in number of homicides – 4 western provinces

X	X^2
34	1,156
27	729
70	4,900
85	7,225
216	14,010

The values of X are listed in the first column of Table 7, and the corresponding squares of these values are in the second column. That is, the first value of X is 34 and $34 \times 34 = 1,156$. For the second value of $X = 27$, its square is $X^2 = 27 \times 27 = 729$. Other numbers in Table 7 are obtained in a similar manner. The sum of the X values is $\Sigma X = 216$ and the sum of the squares of X is $\Sigma X^2 = 14,010$. There are $n = 4$ provinces.

Entering these into the formulae for the mean and variance gives a mean of

$$\bar{X} = \frac{\Sigma X}{n} = \frac{216}{4} = 54.$$

and a variance of

$$\begin{aligned} s^2 &= \frac{1}{n-1}\left[\Sigma X^2 - \frac{(\Sigma X)^2}{n}\right] \\ &= \frac{1}{4-1}\left[14,010 - \frac{216^2}{4}\right] \\ &= \frac{1}{4-1}\left[14,010 - \frac{46,656^2}{4}\right] \\ &= \frac{14,010 - 11,664}{3} \\ &= \frac{2,346}{3} \\ &= 782. \end{aligned}$$

The standard deviation is the square root of the variance, that is,

$$s = \sqrt{s^2} = \sqrt{782} = 27.964$$

or 28.0 homicides per province. The alternative formula yields the same results as the earlier formula.