Chapter 6

UNIVARIATE DESCRIPTIVE STATISTICS

"Statistics" is four things:

! an academic subject or discipline. There is a Department of Statistics in many universities. You may take a course called Statistics 101.

! a set of methods used to process and interpret quantitative data. Most of this "processing" and "interpretation" involves doing things that allow you to see patterns in the data.

! collections of data gathered with the methods described above.

! a set of figures that summarize a set of data. More precisely, statistics are figures that summarize *samples*, while parameters summarize populations.

In this book I'm going to ignore the first and third types of "statistics." I'll spend most of my time on the second type — the methods used to interpret quantitative data.

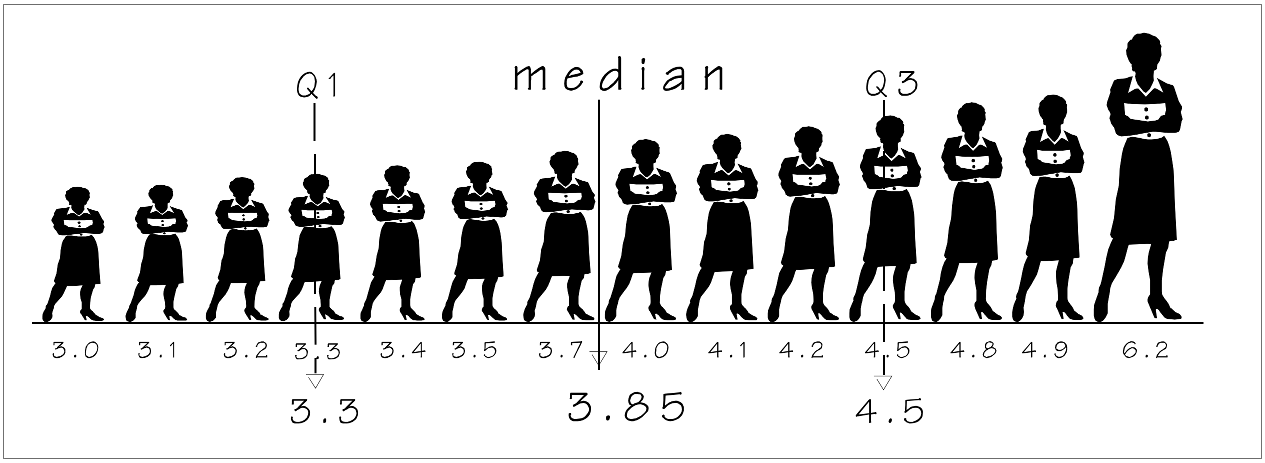
Descriptive and Inferential Statistics

On page 35 you read about sampling — what it is and why you do it. You know that samples are different from the population, and you know something about the relation between samples and their populations. It won't surprise you to learn that the statistical methods used for samples are different from the ones used for populations. The methods used for samples are known as *descriptive* statistics, while the ones used for populations are known as *inferential* statistics.

Descriptive statistics are simple. Their task is to *describe* the data in a sample. You are probably already familiar with some kinds of descriptive statistics; the quantity commonly known as "the average" is the most familiar one. When you calculate the average of the ages of a group of people, you are doing descriptive statistics: you are summarizing the data you have.

Inferential statistics are more complicated. Where you would use descriptive statistics to summarize the data from your sample, you don't have data from the population, so you have to make logical inferences about it. This is where inferential statistics are used. Their task is to help you make *inferences* about the population based on the information you have about your sample. While the math is a little bit more complicated here than it is for descriptive statistics, it is the logic involved that most people find confusing when they first learn inferential statistics.

I'll begin with the simplest kind of descriptive statistics — the ones that describe only one variable



at a time. Then I'll move to the kind that describe differences or relationships. These use two variables at a time. After I've laid the groundwork, I'll venture into the area of inferential statistics.

Descriptive Statistics

Descriptive statistics are tools you use to summarize your data in an effective and meaningful way. They are used to pull information about important aspects of your data out of the pile of numbers and to make it visible in a useful way.

On page 23 you read that variables are either discrete (also called *categorical* because they are used to sort things into categories) or continuous. The methods used for discrete variables are generally different from the ones used for continuous variables. Whether a variable is discrete or continuous turns out to have important implications for what you can do with it and how you use it. These implications will be discussed in the following pages.

There are two classes of univariate descriptive statistics: measures of **central tendency** and measures of **dispersion**.

Central tendency gets at the "typical" or "most common" value in a set of values. *Dispersion* tells how much spread or how much scattering around the central value there is. You use different measures of central tendency and dispersion for data scaled at different levels. The main reason for this is that different types of scaling produce different kinds of "numbers," and they vary in the extent to which they allow you to perform arithmetical operations (addition, multiplication, etc.) on them.

Central Tendency

Because different kinds of numbers represent different aspects of reality, you need to use different measures of central tendency for different levels of scaling. Remember that nominal data sorts cases into categories and only tells which category each case belongs to; with nominal data you can only tell whether two cases are in the same category or in different categories. Ordinal data tells you more; it orders cases from low to high. With ordinal data you can tell whether one value is higher than another, but you can't tell how much higher it may be. With interval data you can not only do everything that you do with nominal and ordinal data; you can also determine the size of the difference between two values. With interval scaling, the numbers you get behave like real numbers — you can add and subtract values. Ratio scaling goes one step further because the scale is anchored by an absolute zero value. These differences mean that you have to match your analytic approach to your data.

The **mode** is *the most common category or value* in the data. If more women are named Linda than any other name, Linda is the modal value of women's names. The mode is the only measure of central tendency that can be used with nominal data. It is a *discrete* measure, and cannot be used with continuous interval or ratio data without first grouping the data into discrete categorical ranges. For example, if you wanted to get the mode of the height of students at your university, you would have to group their heights into categories by rounding, say, to the nearest inch. This would make height a discrete variable.

! The mode is not influenced by extreme values.

! The mode is sensitive only to the most frequently occurring score; it is insensitive to all other scores.

! The mode is of little value for non-categorical (e.g., continuous) data; it is used almost exclusively for discrete variables.

The **median** is *the value at the midpoint of a rank-ordered list of all the values* in a set of data. If you have a large group of people line up in order from shortest to tallest, the height of the person in the middle of the line will be the median of the group's height.

Half the values are above the median and half are below the median. If there are an odd number of scores, the median will be the center point in the rank-ordered list of points. If there are an even number of scores, the median will be the mean of the two centermost points. *Note that the median is not the midpoint of the range* (the difference between the highest and lowest values).

If you split the list of values into the half below the median and the half above the median (a "median split"), you could find the middle value of each half. These values (marked "Q1" and "Q3" on the drawing on page 46), together with the median, divide the list of values into four parts called "*quartiles*," each of which contains 25% of all the cases in your data. The median is sometimes referred to as the "second quartile."

Because the only thing it needs from your data is the order of the values, the median can be used for all types of scaling except nominal.

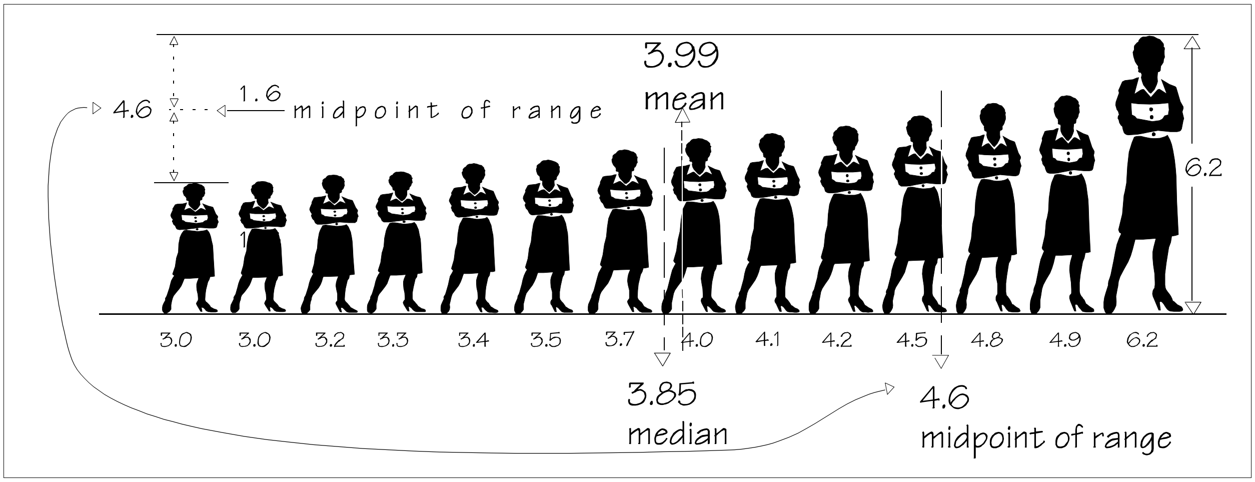
! The median can be used for discrete or continuous variables.

! The median is not influenced by extreme values.

! The median is sensitive *only* to the value of the middle point or points; it is *not* sensitive to the values of all other points.

The **mean** is *the arithmetic average of a set of values*. It is calculated by dividing the sum of all the values by the number of values. Because the calculation of the mean requires addition, it can only be used with interval or ratio data. Since every value in a set of data affects the mean, the mean uses more of the information in the data than the median does. Extreme values have a disproportionately large effect on the mean.

With the mean we see the first mathematical equation and symbols for this course. The symbol for the mean of a sample is $\overline{X}$ (pronounced "x-bar"). The symbol for the mean of a population is : (the



lowercase Greek letter “mu”). In the formula below, $\bar{x}$ is the mean of all the values for the variable x in a set of data.

The symbol $\sum$ is a Greek uppercase sigma — about as close as you can get in Greek to the letter “S” (as in “sum”). It means “add up all the things that follow.” In the equation below, it means add together the scores of all the people on the variable x :

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \ldots + x_n$$

The part to the left of the equal sign in the equation is read as “the sum of x-sub-i; i goes from 1 to n.”

To calculate the mean of a list of numbers, you add them all together and divide the result by how many there are. In other words,

$$mean = \bar{x} = \frac{\sum x_i}{n}$$

So the mean of the numbers 2, 3, 4, 5, 6 would be:

$$\frac{2+3+4+5+6}{5}$$

! The mean requires interval or ratio data.

! The mean is the preferred measure for interval or ratio data.

! The mean is generally not used for discrete variables.

! The mean is sensitive to **all** scores in a sample (every number in the data affects the mean), which makes it a more “powerful” measure than the median or mode.

! The mean's sensitivity to all scores also makes it sensitive to extreme values, which is why the median is used when there are extreme values.

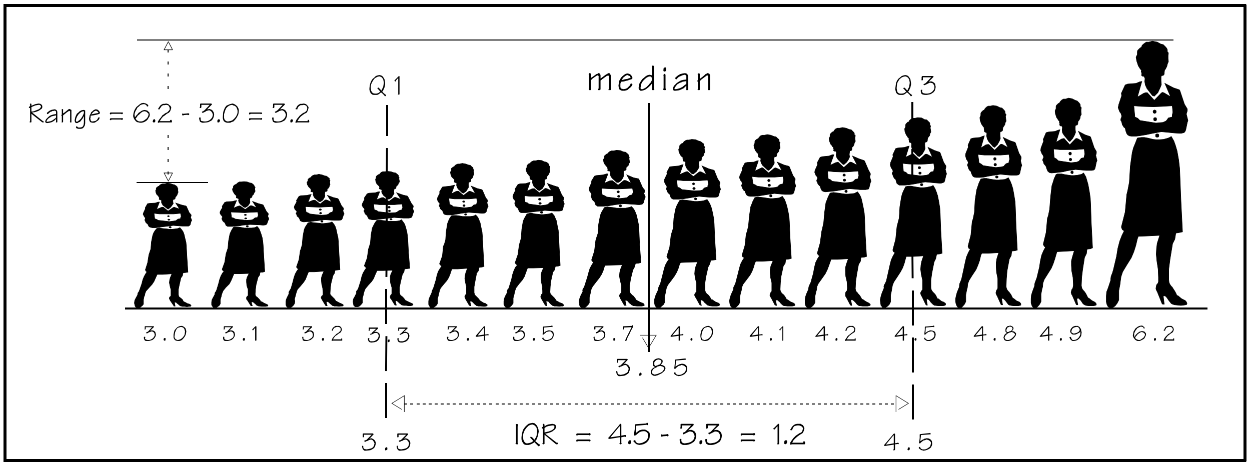
Dispersion

Something you will notice if you look at almost any set of data is that not all observations have the same score on any variable. People differ from one another in terms of their attitudes, their beliefs, or their behaviors. Magazines differ from one another in terms of their content, their format, their cost, and how often they are published. The goal of most research is to describe or explain the variability in the data.

The simplest thing you can do with your data is measure *the amount of variability* in the scores of your variables. When you do this, you will be measuring *dispersion*. Dispersion tells how scattered or spread out the values are. The less spread out the values are, the more concentrated or clustered they will be, and the more likely it is that there will be a “most common” or “typical” or “central” value. Also, the less dispersion there is, the more you can learn about the whole set of values by knowing its central value.

For *nominal* data, the only kind of comparison you can do with a pair of values is to see whether they are the same or different. You can only use a measure of dispersion that looks at the extent to which the values in the data are the same as or different from one another. The measure of dispersion that does this is the information-theoretic measure of **uncertainty**.

If all the cases in your data have the same value — if they all occupy the same nominal category — there is no uncertainty about what the typical value is. For example, if all dogs were named Spot, the most common name for a dog (the *modal* name) would be Spot. In fact, you would have little doubt (uncertainty) about what the name of the next dog you saw would be. If each case is the sole occupant of its category, there is maximum uncertainty of what the typical value is. For example, if every dog had a different, unique name, there would be no “typical” dog's name; you would be completely uncertain about what the next dog’s name would be. If half of all dogs were named Spot and the rest were named Rover, you would have more uncertainty than if they were all named Spot, but less than if they all had different names. The measure of uncertainty in your data tells you the extent to which the values in your data are different from or equal to one another. This measure is not commonly used, mainly because people don’t often calculate dispersion for nominal data.



The simplest measure of dispersion for numerical data is the **range** — the difference between the highest and lowest values, as you can see in the drawing on the next page. Because the difference is a distance, the range can only be calculated on interval or ratio data. The range is determined by only two values — the lowest and the highest — the two most extreme values in the data. As you can see from the drawing above, the range is strongly influenced by extreme values. If the tallest woman is included, the range is 3.2. If she were excluded from the data, the range would drop to 4.9 - 3.0 = 1.9. *Because of this dependence on the two most unusual values, the range doesn't tell much about the data.* It tells nothing, for example, about how far from the center typical values lie.

The **interquartile range** (**IQR**) can be used for ordinal, interval, or ratio data. It is the difference between the first and third quartiles in the data. If you remove the top 25% and the bottom 25% of all cases and then calculate the range of the remaining cases, you will get the IQR. While the IQR is more valuable than the range because it is not influenced as much by extreme values, it is more difficult to calculate, as it requires the data points to be rank ordered. Also, it doesn't work well for small samples, especially ones with an odd number of cases.

The figure on the top of the page illustrates the IQR in relation to the median. You would say either "The IQR is 1.2" or, more likely, "The middle 50% of the sample have heights between 3.3 and 4.5."

Neither the range or the interquartile range take all of the values in your data into account. The range is determined by the two most extreme values and the IQR is determined by the lowest and highest values in the middle 50% of your data. There are two measures of dispersion that take every value in your data into consideration. For nominal data, you can use the information-theoretic measure of uncertainty. For continuous (i.e. interval or ratio) data, you would use variance or its very close relative, standard deviation.

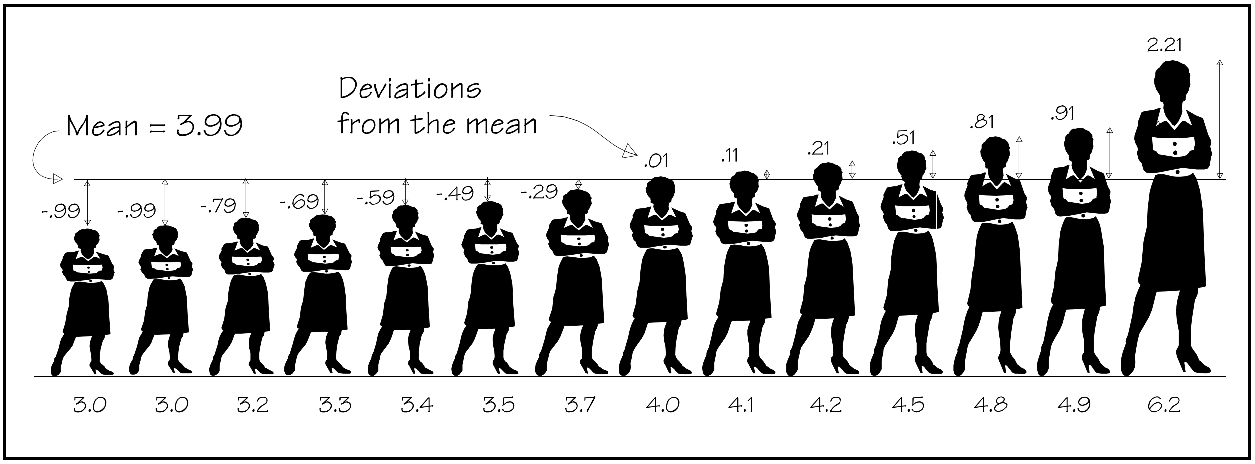
Variance

All measures of dispersion are assessments of what you might call "variability"or "variety" — the extent to which values in your data differ from one another. *Variance* is a particularly useful measure of variety or variability for interval- or ratio-scaled data. It is probably the most important statistical concept, and it is used in a very wide range of situations.

The variance of a set of numbers is based on the distance between each value and the mean of all the values. It starts with *deviation scores*. The deviation score for an individual is the difference between the individual's score and the mean. It is written like this:

$$d_i = x_i - \overline{x}$$

where "$\boldsymbol{d_i}$" is the deviation score for the i^{th} individual, and "$\boldsymbol{x}_\mathrm{i}$" is the i^{th} individual's value (i.e., the score for person number *i*).



If an individual's score is higher than the mean, the deviation score will be positive; if it is lower than the mean, the deviation score will be negative, as in the drawing above. It might seem to make sense to calculate the mean of the deviation scores. There is a problem with this, though, because the sum of the deviation scores is *always* zero (*why does this happen?*), so the mean of the deviation scores will also always be zero.

Sometimes people speak of the "*average deviation*" or the "*mean absolute deviation,*" which is the mean of the **absolute values** of the deviation scores. (The "absolute value" means that you ignore minus signs and treat all scores as if they were positive numbers.)

It turns out that there is a better way of dealing with negative deviation scores. If you square a negative number, the result becomes positive. So, if you square the deviation scores, the results will always be positive. The sum of the squared deviation scores is called the *sum of squares* (SS). **Note: The sum of squares is *not* the variance**. The SS is the first step in the calculation of variance, and it is something you will see in a variety of situations in the coming chapters.

$$SS = \sum d_i^2$$

The **variance** is the sum of squares divided by the number of scores that went into the sum. In other words, it is the *mean* of the *squared* deviation scores. This is why variance is sometimes called by the more descriptive name **mean square**. Note that this name tells you how to calculate the variance (*if you can remember what it is you have to square!*).

The symbol for a population's variance is σ^2. The letter in the symbol is a lowercase Greek sigma. Here is the equation for a *population's* variance:

$$\sigma^2 = \frac{\sum D_i^2}{N} = \frac{SS}{N}$$

In the equation, the "D_i" is deviation scores — the differences between the individual scores and the population's mean. "*N*" is the number of cases. (The "*D*" and "*N*" are uppercase to remind you that they are *population* values, not sample values.)

The symbol for the variance of a sample is s^2. When you calculate the variance for a *sample*, the size of the sample is transformed into a value called **degrees of freedom**. The degrees of freedom for a sample of size *n* is *n - 1*, so you divide by ***(n-1)*** instead of ***N***. *This change makes the variance of a sample a better estimate of the population's variance.* (This issue is discussed in more detail on page 53.) So the equation for a *sample's* variance is:

$$s^2 = \frac{\sum d_i^2}{n-1} = \frac{SS}{n-1}$$

Standard Deviation

The ***standard deviation*** is the square root of the variance. It is sometimes called the **root mean square**, because it is the square *root* of the *mean* of the *squared* deviation scores. The standard deviation is the most commonly used measure of dispersion for interval or ratio level data.

While the *variance* is a measure of the overall amount of variability or spread around the mean, the *standard deviation* is a measure of the typical deviation from the mean. Like the variance and the mean, the standard deviation is sensitive to ***all*** scores. The symbol for a sample's standard deviation is a lowercase "*s*". The symbol for population standard deviation is a lowercase Greek sigma: F. The equation for a sample's standard deviation is:

$$s = \sqrt{\frac{\sum d_i^2}{n-1}} = \sqrt{\frac{SS}{n-1}}$$

Standard Scores (or "z-scores")

If an individual person's score is converted so it tells how far from the mean the person is, it will become a relative score that will let you know how this person compares to the rest of the sample. The most common way of doing this is to calculate a *standard score*. The word "standard" in "standard score" is the same one as in "standard deviation." This is not a coincidence. To calculate standard scores, you divide the individual's deviation score (the difference between the individual's score and the mean) by the standard deviation. The equation to calculate an individual's standard score is:

$$z_i = \frac{x_i - \overline{x}}{s}$$

The subscript *i* tells which person you are doing this for. $\boldsymbol{x_i}$ is person *i*'s score on the variable. The ***s*** in the denominator is the standard deviation. If the standard deviation is 2, the mean is 5, and your score is 7, your *z*-score would be:

$$\frac{(x_i - \overline{x})}{s} = \frac{(7-5)}{2} = \frac{2}{2} = 1.0$$

If your score was 4, your *z*-score would be:

$$\frac{(4-5)}{2} = \frac{-1}{2} = -0.5$$

A positive *z*-score means that you are above the mean; a negative *z*-score means that you are below the mean.

A person's *z*-score tells how far away from the mean that person's score is, in terms of standard deviations. If your *z*-score is 1.0, you are one standard deviation above the mean. If your *z*-score is ! 3.5, you are three and a half standard deviations below the mean.

If you know the mean and standard deviation and you want to convert *z*-scores back into raw scores, you can use this equation:

$$x_i = (z_i \times s) + \overline{x}$$

Just multiply the *z*-score by the standard deviation and add the result to the mean. Here are a few examples:

Mean	Std. Dev.	Raw Score	*z*- score
10	2	6	-2.0
10	2	13	1.5
10	5	12	0.4
12	3	10	-0.667
12	3	16.4	1.466667

The table below summarizes which measures of central tendency and dispersion are used for different levels of scaling.

Level of Scaling	discrete or continuous	central tendency	dispersion
nominal	discrete	mode	uncertainty
ordinal	discrete	median	IQR
interval	continuous	median or mean	std. dev. or IQR
ratio	continuous	median or mean	std. dev. or IQR

Calculating Standard Deviation

An alternative name for standard deviation, "root mean square," tells how to calculate it: the standard deviation is the square *root* of the *mean* of the *squared* deviation scores. The deviation score (d_i) is the difference between the individual's score (x_i) and the mean $(\bar{x})$.

So the equation for standard deviation is:

$$s = \sqrt{\frac{\sum d_i^2}{n-1}}$$

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

$$= \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}}$$

This equation shows the additions, subtractions, and multiplications you have to do to calculate the standard deviation. This is the equation you first saw on page 51. The only good things that you can say about this equation are first, that it works; and second, that you can remember it if you can remember "root mean square." *(Don't forget that it is the deviation scores that get squared!)* The bad thing about it is that it involves a *lot* of work. You have to calculate the mean, then you have to subtract the mean from every value on the list. Then you have to square the differences and add them up. Finally, you divide by *n - 1* and take the square root.

Here is a different form of the equation, called the "computational form" because it is much easier to use. It is easier because you don't have to calculate the mean or the deviation scores:

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

1) First, calculate the sum of the squares of the original scores: $\sum x_i^2$

2) Then calculate the sum of the original scores, square it, and divide the result by ***n***, the sample size:

$$\left(\sum x_i\right)^2 / n$$

3) Now subtract the result of step 2 from the first sum;

4) Divide the result of step 3 by your sample size minus 1;

5) . . . and take the square root. *Voila!*

The result is exactly the same as if you calculated the mean, the deviation scores, the squares of the deviation scores, the sum of the squares . . . . With the new method, you do a *lot* less work. You only need to add up all the scores for one total and then add up the squares of the scores for the second total. The rest is easy. The table below shows a comparison of the amount of work you have to do with the two methods for samples containing 7 and 50 cases.

	original method		computational form	
	n = 7	n = 50	n = 7	n = 50
additions	14	100	14	100
subtractions	8	51	2	2
multiplications	7	50	8	50
divisions	2	2	2	2
square roots	1	1	1	1
total	32	204	27	155

The numbers in the last line of this table actually underestimate the difference in the amount of work. This is because the original method usually requires working with messy numbers — numbers involving decimals like 27.3841, 4.2938, etc. These numbers result from subtracting the mean from the original scores to calculate deviation scores. When you square the deviation scores, the numbers get even messier, increasing the amount of work you have to do with all of those additional digits, and *requiring more rounding*, which introduces rounding errors.

In comparison, almost all of the numbers you use with the computational form are whole numbers. It's easy to see that the computational form is the one you would want to use: First, because it almost completely eliminates *rounding error*, the result is more accurate. Second, it takes less work and the work you do is easier because you are using whole numbers.

Sample or Population?

The equation for a *population's* variance and standard deviation are:

$$\sigma^2 = \frac{\sum D_i^2}{N} \quad \text{and} \quad \sigma = \sqrt{\frac{\sum D_i^2}{N}}$$

In the equation, the "$\boldsymbol{D_i}$" is deviation scores — the differences between the individual scores and the population's mean. The "***N***" is the population's size. The "***D***" and "***N***" are uppercase because they are *population* values, not sample values.

To calculate F for a population, you need to calculate the sum of the squares of the deviation scores for *all* members of the population. Since most social research is interested large populations — in which getting data from all members is either impossible or impractical — you will rarely find yourself trying to *calculate* F. Instead, you will most likely be using data from a sample to *estimate* F.

Remember that a *sample* is a subset of a population, chosen in such a way that what you learn about the sample can be generalized to the population. Because sample statistics are used as the basis for estimates of population parameters, there may be some adjustments made to the equations used for the corresponding population parameters. For standard deviation and variance, you use *n-1* instead of *n* in the denominator, which produces a slightly larger result. The difference will be more significant for smaller samples than for larger ones. In general, a larger standard deviation or variance will produce a more conservative conclusion from any statistical decisions you might make. Since small samples provide results that are less stable than those from larger ones, this conservative modification is a desirable thing to do.

The equations for a *sample's* variance and standard deviation are:

$$s^2 = \frac{\sum d_i^2}{n-1} \quad \text{and} \quad s = \sqrt{\frac{\sum d_i^2}{n-1}}$$

The Uses of the Standard Deviation

Standard deviations are used for several purposes, sometimes to give information directly, and sometimes in the calculation of another statistical measure. Here are three important uses of the standard deviation:

! The standard deviation is the most common measure of dispersion when the data is scaled at the interval or ratio level. Here you are describing the extent to which the elements in a sample are spread out from one another — in particular, *how far the typical value is from the mean.* If you think about this, you will realize that knowing the standard deviation allows you to know how good an estimate of central tendency the mean is.

For example, for a course with 50 students, if the mean score on the final exam is 75 and the standard deviation is 0.5, you will know that most scores are pretty close to 75. (In fact, about 95% of all the scores will be between 74 and 76.) In this case, the mean, is a very good indicator of the "typical" score.

In contrast, if the mean is 75 and the standard deviation is 28.0, you don't really know what the "typical" score is. There might not even be any scores between 70 and 80. It turns out that you would be safe in estimating that 95% of all scores are between 25 and 125, but this is a pretty big range. Generally, you would want a more precise estimate than that.

! The sample standard deviation, when it is calculated with (*n-1*) in the denominator, is used as an estimate of the population's standard deviation, which tells you how much variety or heterogeneity there is in the population.

! When you combine the mean with the standard deviation in the calculation of *z*-scores (for a normally distributed variable), you can tell where an individual is in the distribution relative to the other members. For example, if your *z*-score is 2.52, you are above 97.5% of everyone else in the sample. If your *z*-score is 1.0, about 16% of the other people are above you.

One benefit of using *z*-scores is that it allows you to compare variables that have different means and standard deviations. Remember that the *z*-score is also known as the *standard score*. When you transform a person's raw score (the original value) into a *z*-score, you standardize it by converting it to a scale where the mean is zero and the units are marked off in standard deviations. Starting from the mean and going up, you have 0, 1, 2, and so on. These numbers mean "0, 1, 2, etc. standard deviations above the mean." Once you have standardized your scores by computing *z*-scores, you can compare a person's score on one variable to their score on another variable (or to someone else's score on another variable).

The computational formula for standard deviation

If you are curious, here is a demonstration of the basis for the computational formula for standard deviation. Compare the original and computational versions. Note that the only difference between the two equations is the numerators:

original equation	computational equation
$s = \sqrt{\frac{\sum d_i^2}{n-1}}$	$s = \sqrt{\frac{\sum x_i^2 - \left(\sum x_i\right)^2 / n}{n-1}}$

Consider the numerator of the original formula:

$$\sum d_i^2$$

Since d_i, the deviation score, is $x_i - \bar{x}$, you could rewrite the numerator like this:

$$\sum (x_i - \bar{x})^2$$

which is:

$$\sum (x_i - \bar{x})(x_i - \bar{x})$$

If you multiply the terms in the product, you get:

$$\sum (x_i x_i - x_i\bar{x} - x_i\bar{x} + \bar{x}\bar{x})$$

which you can simplify as:

$$\sum (x_i^2 - 2x_i\bar{x} + \bar{x}^2)$$

You can break it up into three sums:

$$\sum x_i^2 - \sum 2x_i\bar{x} + \sum \bar{x}^2$$

Since $\bar{x}$, the mean, is $\frac{\sum x_i}{n}$, you can rewrite the above sum as:

$$\sum x_i^2 - \sum 2x_i\left(\frac{\sum x_i}{n}\right) + \sum\left(\frac{\sum x_i}{n}\right)^2$$

which simplifies to:

$$\sum x_i^2 - 2\sum x_i\left(\frac{\sum x_i}{n}\right) + \frac{\left(\sum x_i\right)^2}{n}$$

and then to:

$$\sum x_i^2 - 2\frac{\sum x_i \sum x_i}{n} + \frac{\left(\sum x_i\right)^2}{n}$$

which gives:

$$\sum x_i^2 - 2\frac{\left(\sum x_i\right)^2}{n} + \frac{\left(\sum x_i\right)^2}{n}$$

and, at last, we have:

$$\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

. . . which is the numerator of the computational formula!

As a bonus, here is a simpler computational version:

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

How to calculate standard deviation with the original method

$$s = \sqrt{\frac{\sum d_i^2}{n-1}}$$

1. Make three columns, like the ones on the right, headed “X_i,” “d_i,” and “d_i^2”
2. Write the values of the variable in the first column, and count them to see how many there are.
   (*ex: n = 20*)
3. Add all the values in the first column to get a total.
   (*ex: total = 120*)
4. Divide the total by the number there are to get the mean. (*ex: mean = 120/20 = 6.0*)
5. Subtract the mean from each value in the first column and write the results in the second column.
6. Square each value in the second column and write the result in the third column.
7. Add all the values in the third column to get a total. This is the “sum of squares".
   (*ex: total SS = 40*)
8. Divide the sum of squares by *n - 1* to get the variance.
   (*ex: variance = 2.1052632*)
9. Take the square root of the variance to get the standard deviation.
   (*ex: std. dev. = 1.4509525*)

	(1)	X_i		d_i		d_i^2
	(2)	8	*(5)*	2	*(6)*	4
		6		0		0
		4		-2		4
		5		-1		1
		6		0		0
		5		-1		1
		4		-2		4
		8		2		4
		7		1		1
		8		2		4
		4		-2		4
		5		-1		1
		8		2		4
		7		1		1
		6		0		0
		5		-1		1
		5		-1		1
		6		0		0
		8		2		4
		5		-1		1
Sums:	*(3)*	120		0	*(7)*	40

(2) n =20

(4) Mean = 120/20 = 6.0

(8) variance = 40/19 = 2.105263158 . 2.105

(9) std. dev. = $\sqrt{\text{variance}}$ = 1.4509525 . 1.451

How to calculate standard deviation with the computational method

$$s = \sqrt{\frac{\sum x_i^2 - \left(\sum x_i\right)^2 / n}{n-1}}$$

1. Make two columns, headed “$\boldsymbol{X_i}$,” and “$\boldsymbol{X_i^2}$"
2. Write the values of the variable in the first column, and count them to see how many there are.
   (*ex: n = 20*)
3. Add all the values in the column to get a total.
   (*ex: total = 120*)
4. Square each value in the first column and write the result in the second column.
5. Add all the values in the second column to get a total.
   (*ex: total = 760*)
6. Square the first total.
   (*ex: 120 × 120 = 14,400)*
7. Divide the value you obtained in step 6 by the sample size.
   (*ex: 14,400 / 20 = 720)*
8. Subtract the result of step 7 from the result of step 5. The difference is the sum of squares (SS).
   (*ex: 760 – 720 = 40*)
9. Divide the sum of squares by (n-1) to get the variance.
   (*ex: 40 / 19 = 2.1052632)*
10. Take the square root of the variance to get the standard deviation.
    (*ex: std. dev. = 1.4509525*)

	(1)	X_i		X_i^2
	(2)	8	*(4)*	64
		6		36
		4		16
		5		25
		6		36
		5		25
		4		16
		8		64
		7		49
		8		64
		4		16
		5		25
		8		64
		7		49
		6		36
		5		25
		5		25
		6		36
		8		64
		5		25
Sums:	*(3)*	120	*(5)*	760

(6) 120 × 120 = 14400

(7) 14400 ÷ 20 = 720

(8) 760 – 720 = 40

(9) variance 40/19 = 2.105263158 (2.105)

(10) std. dev. = $\sqrt{\text{variance}}$ = 1.4509525 (1.451)

Important Terms and Concepts

average deviation
central tendency
computational formula
degrees of freedom
descriptive statistics
deviation scores
dispersion
extreme value
inferential statistics
information-theoretic uncertainty
interquartile range (IQR)
mean absolute deviation
mean
mean square
median
midpoint of range
modal value
mode
range
root mean square
standard scores
standard deviation
statistics
sum of squares (SS)
variability or spread around the mean
variance
z-score

Things to think about .....

Add one more observation with the value of 9 to the data on page 55 and see what it does to the results, using the calculation method on page 55. You will have to calculate a new sum for the first column, a new mean, new deviation scores, new squared deviation scores, etc. It's a *lot* of work! The new sum in the first column will be 120 + 9 = 129. The new *n* will be 21 instead of 20.

What is the new mean?

What is the new sum of squares?

What is the new variance?

Do the same thing with the data on page 56 and see what it does to the results, using the calculation method on page 56.

What is the new mean?

What is the new sum of squares?

What is the new variance?

Which method do you prefer?

Which result is likely to be more accurate? Why?

Examples of data and mean & standard deviation:

3 sets of data	$\bar{x}$	s	$\sum x_i$	$\sum x_i^2$
7, 28, 12, 17, 23, 13, 0, 16, 25, 7, 25, 15, 13, 9, 2, 4	13.50	8.5557	216.0	4014.0
95, 79, 83, 61, 85, 59, 52, 54, 95, 79, 74, 91, 83, 84, 94, 64	77.00	14.6924	1232.0	98102.
41, 31, 36, 42, 47, 49, 38, 45, 46, 31, 37, 48, 35, 32, 43, 39	40.00	6.05530	640.	26150.

Notes