

Descriptive Statistics

*Dr. Tom Pierce
Department of Psychology
Radford University*

Descriptive statistics comprise a collection of techniques for better understanding what the people in a group look like in terms of the measure or measures you're interested in.

In general, there are four classes of descriptive techniques. First, **frequency distributions** are used to display information about where the scores fall along the scale going from the lowest score to the highest score. Second, measures of central tendency, or **averages**, provide the best single numbers we could use to represent all of the scores on a particular measure. Third, measures of **variability** provide information about how spread out a set of scores happen to be. Fourth, the original raw scores one collects are often transformed to other types of scores in order to provide the investigator with different types of information about the research participants in a study. A *standard score* is a good example of a **transformed score** that provides much more information about an individual subject than a raw score can.

Frequency distributions

Let's say that you obtain Beck Depression Inventory scores from each of 400 research participants. The scores on this measure can range anywhere from 1 to 73. Typically, scores fall somewhere between 35 and 55. You've got 400 numbers to have to keep track of here. If someone asks you how the scores in your study came out you could say "well, subject number one had a score of 38, subject two had a 25, subject three had a 46, ...". You get the idea. This is way too many number for anyone to be able to look at and get a general idea of (1) where most of the scores fall on the scale and (2) how spread out the scores are around this central point on the scale. The job of a frequency distribution is to take a very large set of numbers and to boil it down to a much smaller set of numbers – a collection of numbers that is small enough for the pathetically limited human mind to keep track of. A good frequency distribution allows the consumer to extract the necessary information about the scores in the data set while working within our cognitive limitations.

Regular frequency distribution

The most straight-forward example of a frequency distribution goes like this. Let's say that you're given ratings of teaching effectiveness for the students in a large Introduction to Psychology class. There are 400 students in the class. The questionnaire provides students with 15 statements and the student is asked to pick a number between one and five that indicates the degree to which they agree or disagree with each statement. One of these statements is "The instructor in this course is outstanding". A response of "5" indicates that the student agrees with the statement completely. A response of "1"

indicates that the student disagrees with the statement completely. A regular frequency distribution will allow the instructor to see how many of the students rated her on every possible score ranging from one to five. In other words, how many students gave the instructor a “one”, how many gave them a “two, and so on. This information is often displayed in the form of a table, like the one below.

Table 1.1

X	f
5	150
4	200
3	40
2	10
1	0

	400

There are two columns of numbers in this table. There is a capital X at the top of the column on the left. Every possible raw score that a subject could provide is contained in this column. A capital X is used to label this column because a capital X is the symbol that is usually used to represent a **raw score**. The column on the right is labeled with a small-case letter f. The numbers in this column represent the number of times – the **frequency** – that each possible score actually showed up in the data set. The letter f at the top of the column is just short-hand for “frequency”.

Thus, this dinky little table contains everything the instructor needs in order to know every score in the data set. Instead of having to keep track of 400 numbers, the instructor only has to keep track of five – the number of times each possible score appeared in the data set. This table is said to represent a frequency distribution because the it shows us how the scores in the set are distributed as you go from the smallest possible score in the set to the highest possible score. It basically answers the question “where did the scores fall on the scale?”. This particular example is said to represent a regular frequency distribution because *every possible score is displayed in the raw score (capital X) column*.

Interval frequency distribution

A little bit different situation where a frequency distribution might come in handy is in displaying IQ scores collected from 90 people. In a random sample of people drawn from the population, what would you expect these IQ scores to look like? What is the lowest score you might expect to see in the set? What’s the highest score you might reasonably expect to see in the data set? It turns out that the lowest score in this particular set is 70 and the highest score is 139. Is it reasonable to display these data in a regular frequency distribution? No! So, why not?

What would you have to do to generate a regular frequency distribution for these data? Well, you'd start by listing the highest possible score at the top of the raw score column and then in each row below that you'd list the next lowest possible raw scores. Like this...

Table 1.2

X

139
138
137
136

You get the idea. There are SEVENTY possible raw scores between 70 and 139. That means there would be 70 rows in the table and 70 numbers that you'd have to keep track of. That's too many! The whole idea is to keep the number of values that you have to keep track of to somewhere between five and ten.

So what can you do? A regular frequency distribution isn't very efficient when the number of possible raw scores is greater than ten or twelve. So it's not a good idea to keep track of how often every possible raw score shows up in the set. A reasonable next-best thing is to keep track of the number of times scores fall within a *range of possible scores*. For example, how times did you get scores falling between 130 and 139? How many times did you get scores falling between 120 and 129? 110 and 119? You get the idea. Table 1.3 below presents a frequency distribution in which *the numbers in the frequency column represent the number of times that scores fell within a particular interval or range of possible raw scores*. This makes this version of a frequency distribution an **interval frequency distribution** or, as it is sometimes referred to, a **grouped frequency distribution**.

Table 1.3

X	f
---	---
130-139	3
120-129	7
110-119	15
100-109	28
90-99	23
80-89	9
70-79	5

	90

As you can see, when you add up all of the numbers in the frequency column you get 90. The table accounts for all 90 scores in the set and it does this by only making you keep track of seven numbers – the seven frequencies displayed in the frequencies column. From these seven numbers we can tell that most of the scores fall between 90 and 119 and that the frequency of scores in each interval drops off sharply as we go below 90 or above 119. The interval frequency distribution thus retains the advantage of allowing you to get a sense of how the scores fall as you go from the low end of the scale to the highest. BUT you retain this simplicity at a price. You have to give up some precision in knowing exactly where any one score falls on the scale of possible scores. For example, you might know that three scores fell between 130 and 139, but you have no way of knowing from the table whether there were three examples of a score of 130, or two 135s and a 137, or three examples of a 139. *An interval frequency distribution represents a tradeoff between two competing benefits – the benefit of being precise and the benefit of being concise.* When the variable you’re working with has a lot of possible scores it will be impossible to enjoy both benefits at the same time.

Cumulative frequency distribution

Sometimes the investigator is most interested in the pattern one sees in a running total of the number of scores that have been encountered as one goes from the lowest score in the set to the highest. For example, in the IQ data set, I might want to know the number of people with IQ scores at or below the interval of 70-79, then how many have scores at or below the interval of 80-89, 90-99, and so on. A ***cumulative frequency distribution*** is a table or graph that presents the number of scores within or below each interval that the investigator might be interested in, not just the number of scores within each interval.

Constructing this type of frequency distribution is easy. The frequency that corresponds to each interval is nothing more than the number of subjects that fell within that interval PLUS the number of subjects that had scores below that interval. In Table 1.4, the *cumulative frequencies* for each interval are contained in the column labeled “cf”.

Table 1.4

X	f	cf
---	---	---
130-139	3	90
120-129	7	87
110-119	15	80
100-109	28	65
90-99	23	37
80-89	9	14
70-79	5	5

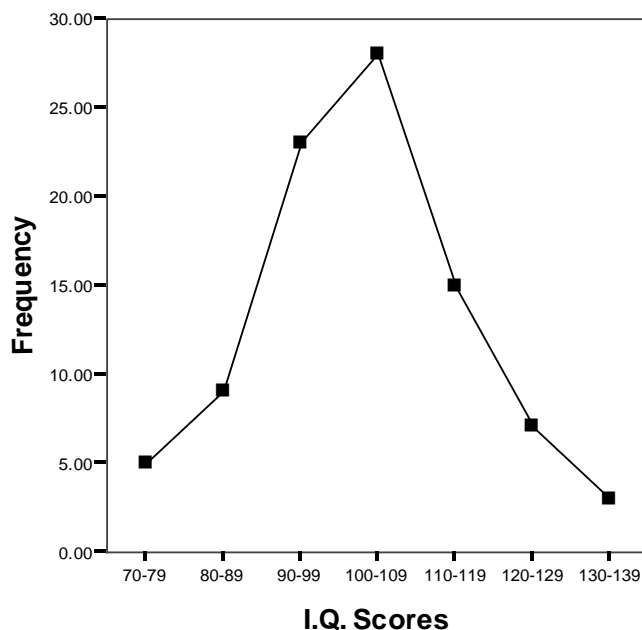
Cumulative frequencies are a standard tool for presenting data in a number of fields, including those from operant conditioning experiments and behavior modification interventions. In these studies, the cumulative frequency displays the number of correct

responses up to that point in the experiment. Psychologists have found that the shape of the distribution changes when different schedules of reinforcement are used.

Graphs of frequency distributions

Tables are one way of presenting information about where the scores in a data set are located on the scale from lowest to highest. Another common strategy for presenting the same information is to provide a graph of a frequency distribution. The idea here is that the X-axis of the graph represents the full range of scores that showed up in the data set. The Y-axis represents the number of times that scores were observed at each point on the scale of possible scores.

[Figure 1.1]

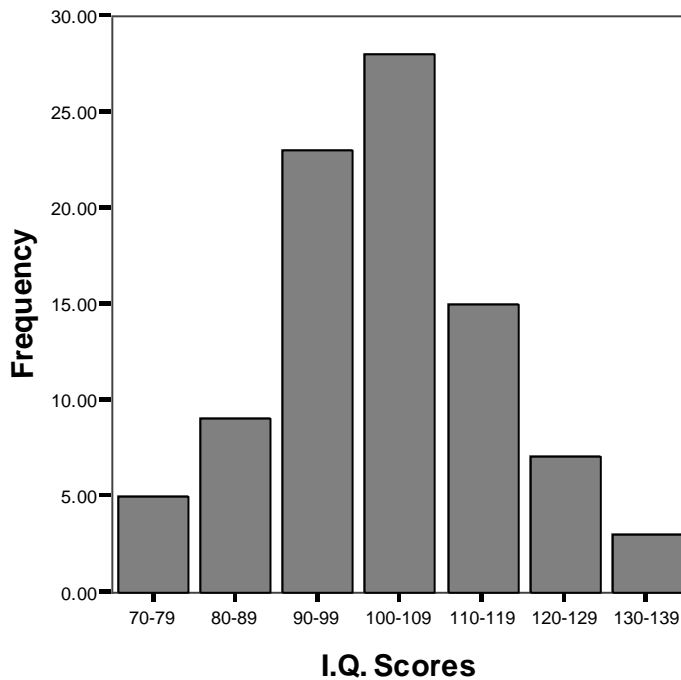


The points in the graph in figure 1.1 contain all of the information about the distribution of IQ scores that was contained in the frequency distribution table presented in Table 1.3. Obviously, the higher the point in the graph, the more scores there were at that particular location on the scale of possible scores. If we “connect the dots” in the graph, we get what’s called a **frequency polygon**.

An alternative way of displaying the same information is to generate a

bar graph. In a bar graph the frequency for each individual score (for a regular frequency distribution) or for each grouping of scores (for an interval frequency distribution) is represented by the height of a bar rising above a particular score or interval on the x-axis.

[Figure 1.2]

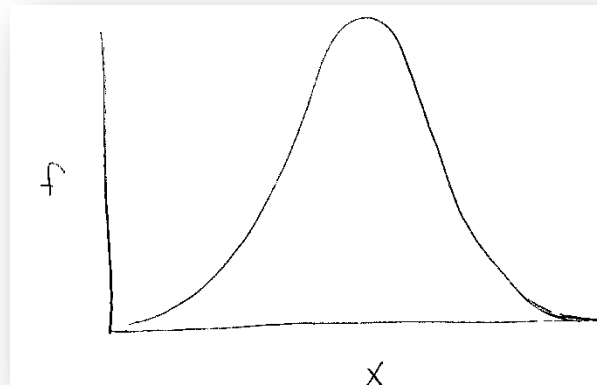


Another name you'll see for the same type of graph is a **histogram** (or histogram). The difference between a bar graph and a histogram is that a bar graph technically has a space between each individual bar, while a **histogram** has no space in between the bars. Big whoop.

Shapes of frequency distributions

There are several ways in which the shapes of frequency distributions can differ from each other. The first one we'll talk about is in terms of whether the shape of a frequency distribution is symmetrical or not. A distribution is said to be **symmetrical** when the shape on the left hand side of the curve is the same as the shape on the right hand side of the curve. For example, take a look at a graph of a **normal distribution**. See Figure 1.3.

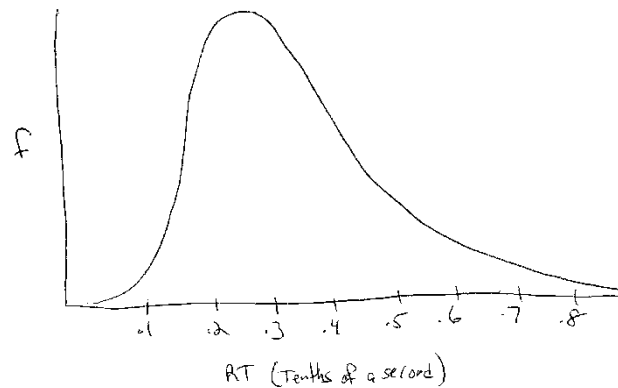
[Figure 1.3]



If you were to fold the left hand side of the curve over on top of the right hand side of the curve, the lines would overlap perfectly with each other. The distribution on the right side is a mirror image of the left side. Because of this we know that the normal curve is a symmetrical distribution.

Skewness. A distribution is said to be **asymmetrical** (i.e., without symmetry) if the two sides of the distribution are not mirror images of each other. For example, the frequency distribution below (Figure 1.4) displays reaction times on a choice reaction time task for one college-age research participant.

[Figure 1.4]

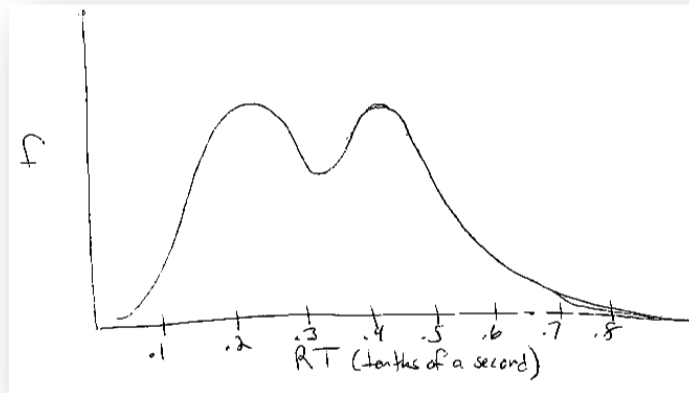


Obviously, if you were to fold the right side of the graph onto the left side, the lines wouldn't be anywhere close to overlapping. The reaction times in the set are much more bunched up at the lower end of the scale and then the curve trails off slowly towards the longer reaction times. There are many more outliers on the high end of the scale than on the low end of the scale. This particular shape for a frequency distribution is referred to as a **skewed distribution**. Distributions can be skewed either to the right or to the left. In this graph the longer tail of the distribution is pointing to the right, so we'd say that the distribution is **skewed to the right**. Other people would refer to this shape as being **positively skewed** (the tail is pointing towards the more positive numbers). If the scores in a distribution were bunched up at the higher end of the scale and the longer tail were pointing to the left (or towards the negative numbers), we'd say that the distribution was **skewed to the left** or **negatively skewed**.

The number of peaks. Another way of describing the shape of a frequency distribution is in terms of the number of noticeable peaks in the shape of the curve. The normal curve has one peak in the distribution, so we would refer to the shape of this distribution as **unimodal** (uni = one; modal = peak).

Now let's look at another data set. Below is a frequency distribution of reaction times collected from ten younger adults and ten adults over the age of sixty. See Figure 1.5

[Figure 1.5]



There are two noticeable peaks in the shape of this graph. For this reason we refer to the shape of this distribution as **bimodal**. The two peaks don't have to be exactly the same height to say that the distribution is bimodal. How would you interpret the shape of this distribution? One likely scenario is that there are two separate groups of participants represented in the study. The younger group might have provided most of the faster RTs and the older group might have provided most of the slower RTs. Inspection of the frequency distribution is often helpful in thinking about whether you've got one group of participants in terms of the variable you're using or more than one group.

Kurtosis. One additional way of describing the shape of a frequency distribution is in terms of the "peakedness" of the curve. This property is referred to as **kurtosis**. If the shape of a distribution has a sharp peak to it, the distribution is referred to as **leptokurtic**. Leptokurtic distributions are more peaked than the normal curve. If the shape of a distribution is relatively flat, it is referred to as **platykurtic** (just think of the flat bill of a platypus!). Platykurtic distributions are less peaked than the normal curve. A curve with the same degree of peakedness as the normal curve is referred to as **mesokurtic**.

Central Tendency

Often an investigator would like to have a single number to represent all of the scores in a data set. These "averages" are referred to as measures of **central tendency**. We'll discuss three measures of central tendency: the mode, the median, and the mean.

The Mode

The mode of a particular variable represents the score that showed up the most often. It's the most frequent score for that variable. In terms of a graph of a frequency distribution, the mode corresponds to the peak of the curve.

The biggest problem with the mode as a best representative is that it's based on data at only a single point on the scale of possible values. This means that it doesn't matter what's going on at any other point on the scale. The only value that counts is the one that occurred the most often. For this reason, the mode might not be a very good number to use in representing all of the scores in the set.

The Median

If you were to take all of the scores you collected on a particular variable and lined them up in increasing numerical order, the median would be the score that occurs half-way through the list of numbers. So, if there were 101 scores in the data set and all the scores were different, the median would be the 51st score. The median is the score where half the remaining scores fall below it and half the remaining scores fall above it. In terms of a graph of a frequency distribution, the median corresponds to the point on the scale where half of the area under the curve falls to the left of that point and half of the area under the curve falls to the right of that point.

A synonym for the median is the 50th percentile. If a student took their SATs and found out that they were at the 50th percentile that would mean that 50% of the other students taking the test got scores below theirs and 50% got scores above theirs. We'll say more about when to use the median in a bit.

The Mean

The mean is what most people think of as "the average". It's simply the number you get when you add up all of the raw scores in a set and then divide by the number of raw scores. The symbol for the mean is often a capital X with a bar over it. The equation for calculating the mean is...

$$\bar{X} = \frac{\Sigma X}{N}$$

Advantages of the mean. By far the most widely used measure of central tendency is the mean. So what makes the mean so much more appealing as a measure of central tendency than the other two? For starters, it's the only one the three measures of central tendency that takes all of the raw scores in the set into account. But there's something else about the mean that appeals to statisticians.

As you have probably become acutely aware, statisticians are a strange and odd group of people. And strange and odd things make strange and odd people happy. So now I'm going to tell you what makes a statistician happy about the mean as a measure of central tendency. You may think that it sounds a bit strange and odd at first, but then we've already covered why it would. Here it is: statisticians think the mean is just gosh-darn

neat because it is ***the one number that makes the sum of squared deviations around it a minimum.***

So should you feel all warm and fuzzy or something after hearing this? Probably not, but by the end of the semester this may change. I would recommend that you read this sentence again a week or so after the semester ends to see if you have begun the irreversible mutation process towards statistician.

The mean is the one number that makes the sum of squared deviations around it a minimum? Here's what I mean...

- Below we've got a set of five raw scores. Now subtract the mean from each of these scores. The equation used to generate these numbers is, of course, $X - \bar{X}$. A small case letter X ("x") is often used as the symbol for a deviation score.

X	$(X - \bar{X})$ or x
6 - 10	-4
8 - 10	-2
10 - 10	0
12 - 10	+2
14 - 10	+4

- Now we're left with a new set of numbers. Each number tells us how far one person's score fell from the mean. Positive numbers indicate that the person's score was above the mean. Negative scores indicate that the person's score was below the mean. Now, just for fun, take each of these deviation scores and square them. Humor me. The equation to generate each of these numbers is $(X - \bar{X})^2$.
- Now, last thing, take the five squared deviations and add them up. You get 40.

X	$X - \bar{X}$ or x	$(X - \bar{X})^2$
6 - 10	-4	16
8 - 10	-2	4
10 - 10	0	0
12 - 10	+2	4
14 - 10	+4	16
		40

- There is no other number you could plug in for the mean at the beginning that would give you a smaller number than 40. That's because subtracting the mean gives us deviations scores that are as small, on average, as they could be. This, in turn gives us a set of squared deviations that are as close to zero as they could be. Add them up and

you get a sum of squared deviations that's as small as it could possibly be. Essentially, this tells us that *the mean is as in the middle of the entire set of raw scores as you can get!* That's the appealing thing about the mean for a statistician.

Disadvantages of the mean. Sometimes, the characteristic that gives the mean an advantage over the mode and the median is also the thing that results in a potential downside. Because every score gets taken into account when computing the mean, scores that are very different from most of the other scores end up having a disproportionately large influence on the mean. In other words, a single or a few outliers on one side of the distribution can pull the mean towards them to the degree that the mean doesn't look very much like the vast majority of scores in the data set. When this happens the median is often a better choice as a measure of central tendency.

For example, take the following set of scores:

2,4,6,8,10

The mean of these scores is 6. The median is also 6. Now change just one number in the data set. Change the score of 10 to a score of 100. Now the mean goes up to 24! The mean is now larger than every score in the data except one. What's the median? The median stays the same at 6. In the presence of this outlier of 100 the median of 6 is much more *in the middle* of most of the scores in the data set than the mean.

Measures of Variability

Measures of central tendency provide a single number to represent all of the scores for a single variable. However, this best representative is not the only thing that one might want to know about the scores that people have provided. Measures of variability provide numbers that describe how spread out a set of scores are. In other words, are all of the scores very close to the mean or are they spread out all over the place around the mean? We'll discuss four measures of variability: the range, the sum of squares, the variance, and the standard deviation.

The Range

The range is by far the easiest measure of variability to calculate. The range is simply the difference between the highest score in the data set and the lowest score. In other words, what's the distance between the lowest score and the highest score? This is easy; just take the highest score in the set and subtract the lowest score.

Let's say we have six scores in a data set: 6, 8, 10, 12, and 14. The range would be $14 - 6$ or 8. That's it! The problem with the range is that the only two scores that influence the range are the two scores at the extremes. The other scores in the set could all be huddled up around the mean or spread out all over the place. Because the range is basically at the mercy of outliers in the data set, it's used mainly as a quick and dirty first look at the variability of a set of scores.

Some textbooks give the equation for the range as “Highest Score – Lowest Score + 1”. Because the range is almost never used as a stand-alone statistic, I don’t think it makes much difference which one gets used. The problems with the range are the same no matter which equation is used. One thing that’s never made sense to me about adding the “+ 1” to the equation is that you could run into a situation where everyone had the same score on a measure (say five people also had scores of 27) and the computation of the range turns out to “+1” even though there is absolutely no variability at all in the scores. It just seems like it makes more sense to have a statistic where a sets of scores with zero variability gets a value of zero for the measure of variability.

Sum of Squares

The biggest problem with the range is that it’s based on the values of only two scores in the data set; the extreme high and the extreme low scores. None of the other scores matter at all. It would certainly seem reasonable to want a measure of variability that takes all of the scores for a variable into account. Let’s say that, just for fun, we put all of our raw scores in a column, like this...

X
--
6
8
10
12
14

Now, if we think about it, one pretty good way to think about variability is in terms of how far the raw scores in a set differ from the center of their distribution: that is, from the mean. If the raw scores differ a lot from the mean, this would tell us that the scores are quite spread out *around the mean*. If the differences between the raw scores and the mean are very small, this would tell us that there is very little variability in the raw scores. So let’s do the same thing we did before and subtract the mean from every raw score in the set.

X	$X - \bar{X}$ or x
--	-----
6 – 10	-4
8 - 10	-2
10 – 10	0
12 – 10	+2
14 – 10	+4
-----	-----

And, of course, we end up with a new column of numbers. Each of these numbers provides the distance between a raw score and the mean of 10. For this reason, these numbers are referred to as **deviation scores**. As we said before, the symbol of a small-case letter x is often used as the symbol for a deviation score. In terms of our measure of variability the thing to look at is how close these numbers are to zero.

If there were no variability at all in the data set, what would these deviation scores be equal to? Well, obviously if there's no variability in the data set, all of the raw scores have to be the same number – which tells us that all of the raw scores would be equal to the mean – which would give us zeros for every deviation score! The further the deviations are from zero, the more variability there is among the raw scores.

So, if we want one number to represent the variability of all the scores in the set, why not compute the mean of all the deviation scores? This seems like a reasonable thing to want to do, except for one interesting little twist. Remember, the mean is the center of gravity of a frequency distribution. The definition of the mean is that it minimizes the sum of squared deviations around it. *The sum of all the negative numbers is perfectly balanced out by the sum of all the positive numbers.* This tells us that the mean was designed so that whenever we add up all of the deviation scores in a set, they'll always add up to zero. The problem with calculating the mean deviation score is that we'll always end up a big fat zero! So that won't work. Bummer.

Couldn't we just take the absolute value of all the deviation scores and then get the mean of all of these numbers? You could. You'd end up with a number called the **average deviation** – and it's a perfectly good descriptive number. However, on the mathematical-type side of things, using absolute values is apparently a bit of a dead end. For this reason, most statisticians prefer another method for getting rid of the negative numbers.

Squaring any number, positive or negative, leaves you with a positive number. The alternative to using absolute values is to take every deviation score and square them. This leaves us with a new column of numbers, squared deviations from the mean.

X	$X - \bar{X}$ or x	$(X - \bar{X})^2$
--	-----	-----
6 - 10	-4	16
8 - 10	-2	4
10 - 10	0	0
12 - 10	+2	4
14 - 10	+4	16
-----	-----	-----
		40

When we add up the numbers in the squared deviations column we get the number 40. This number forty is a perfectly good measures of variability. It takes every original raw score into account and it doesn't resort to using absolute values. This sum of all the squared deviations from the mean is referred by the shorthand term **sum of squares**. The

closer the sum of squares is to zero the less variability there is in the original raw scores. The larger the value for the sum of squares the more variability there is in the scores.

The Variance

The sum of squares is a perfectly good measure of variability. When you've got just one set of numbers to consider, the farther the sum of squares is from zero the more spread out the scores are around the mean. However, think about this situation.

Let's say that I have two sections of the same class. One section has 10 people in it. The other section has 20 people in it. They both take their first test and I want to know which class had the most variability in their scores. I find out that the sum of squares for the section with 10 students is 500. The sum of squares for the class with 20 students is 600. Obviously, the class with the sum of squares of 600 has scores that are more spread out because 600 is farther from zero than 500, right? OUCH! NO! NO! NO! You can't say this because the sum of squares is just that – a sum. It's what you get when you add up a bunch of numbers. If you add up 20 different numbers you're likely to end up with a bigger number than if you only added up 10 numbers. The sum of squares for the class with 20 students is probably a larger number because it's the sum of twice as many numbers, not because the scores are more spread out around their mean. ***You can't compare one sum to another sum when they're based on different numbers of values.***

However, you can compare the mean of one set of numbers to the mean of another set of numbers. If you take the sum of squares for the class with ten students in it (i.e., 500) and divide it by the number of students you get a value of 50. The mean of these 10 squared deviations is 50. The mean of the 20 squared deviations for the class with 20 students is 600 divided by 20, or a value of 30. The mean squared deviation for one class is 50, while the mean squared deviation for the other class is 30. The mean of 50 is directly comparable to the mean of 30. This is the advantage of computing a measure of variability known as the variance. The variance is simply ***the mean of a bunch of squared deviations.***

There are actually two equations for computing the variance that we need to discuss. When you have the raw scores for everyone in the population the equation for the variance is...

$$\sigma^2 = \frac{(X - \bar{X})^2}{N}$$

The symbol for the variance of a population is σ^2 .

When you're working with the raw scores for a sample, the equation for the variance is...

$$S^2 = \frac{(X - \bar{X})^2}{N - 1}$$

The symbol for the variance of a sample is S^2 .

So why do I have to divide by $N-1$ when the data come from a sample, but not when the data come from a population? First of all, think back to the difference between a sample and a population. A **population** is every member of the set of people, animals, of things you want to know about. A **sample** is a representative subset of a population. The only reason to obtain data from a sample is to learn something about the population the data were sampled from. Samples are supposed to give you accurate information about populations. When you calculate a descriptive statistic from a sample, that number is supposed to give you an unbiased estimate of the number you would have gotten if you'd been able to use the data from the entire population. A number calculated from all of the scores in a population is referred to as a **parameter**. A number calculated from the scores in a sample is referred to as a **statistic**. *The job of a statistic is to give us an unbiased estimate of a parameter* – that is, the estimates are no more likely to be too high than to be too low. The reason to divide by $N-1$ when calculating the variance of a sample is it results in an unbiased estimate of the variance of all the scores in the population. Dividing by N would result in a number that would, on average, be a little bit smaller than it really should be. SPSS will assume that you are working with data from a sample, so it will compute the variance by dividing by $N-1$. The number $N-1$ is referred to as the number of **degrees of freedom** for this statistic.

Degrees of Freedom

The term “degrees of freedom” will crop up a number of times in our discussions. It's not a particularly easy concept, but it turns out to be extremely useful. For example, in the context of calculating the variance, dividing by the number of degrees of freedom for the sample results enables this sample variance to be an unbiased estimate of the variance of all of the scores in the population. To give you the flavor of what the term “degrees of freedom” refers to, consider the following situation...

If you knew that the mean of a set of five scores was 8 and you know that one of the raw scores in the set is 4, are the other four numbers in the set free to vary? That is, could you move the other numbers around. Sure you could, as long as they averaged out to 8. Now, if you know that four of the numbers in the set were 4, 6, 8, and 10, is the last number in the set free to vary – could you move it around and still have the mean come out to 8. No, this last number in the set has to be 12. The number of degrees of freedom refers to the number of values that are free to vary when one set of numbers is being used to give you an estimate of another number.

The Standard Deviation

Okay, mean squared deviation, degrees of freedom, $N-1$, blah, blah, blah... What's the bottom line here. The variance represents the mean of a set of squared deviations. But saying that you know the mean of a bunch of squared deviations doesn't sound like you've got a very intuitive measure of variability. And it isn't. That's why you don't see the variance reported very often as a stand-alone measure of variability. The measure you do see reported is the standard deviation of a set of scores. The **standard deviation** is a much more intuitively appealing measure of variability because it *represents the average amount that raw scores deviate from the mean*. The variance is the average of

all the *squared deviations* from the mean. The standard deviation is an average of the original *deviations* from the mean. Nobody thinks in terms of squared deviations, but it seems pretty natural to think in terms of regular old deviations.

To obtain the standard deviation, all you have to do is to start with the number you've already gotten for the variance and then calculate the square root of that number. If you remember, the variance of our class with ten students above was 50. The square root of this number is 7.07. So, this standard deviation of 7.07 tells us that the average amount that the test scores differed from their mean was 7.07. The equation for the standard deviation for a population is..

$$\sigma = \sqrt{\frac{\sum(X - \bar{X})^2}{N}}$$

The equation for the standard deviation of a sample is...

$$S = \sqrt{\frac{\sum(X - \bar{X})^2}{N-1}}$$

Transformations

Standard Scores

It won't take long after you've started to work with your data that you'll find that the original raw scores you collected just aren't going to cut it. You might want the numbers you're working with to be able to give you a type of information they weren't designed to provide. But it may be possible to take the original numbers you collected and transform them or convert them to a different form.

For example, let's say that an instructor of a graduate level statistics class – at a medium-sized state university in southwest Virginia – has adopted a rather bizarre grading system. You get your graded first exams back and there is nothing on the exams except a number written in red ink at the top left corner of the front page. The number is 3257. This is your grade. Would you be very happy? Why not!? 3257 is a perfectly good number! It's just as much a number as 23 or 95 or 129. You're right though, this raw score all by itself doesn't give you any idea of how well you did on the exam. 3257 might be a great score or a rotten score. There's really no way to tell. The raw score by itself doesn't give you as much information as you'd like.

So let's say that I tell you that the mean score on the test is 3000. Does this help? At the very least you know that your score is above the mean and you know that your score is

257 points above the mean. Subtracting the mean from a raw score results in a new number – a transformed score – a **deviation score**. In fact we’ve already created a set of deviation scores on the way to calculating the sum of squares.

But is that good enough? Is 257 points above the mean enough to get into the “A” range or did a lot of people get deviation scores this far above the mean so that your score is only good enough for a “B”? Again, there’s no way of telling from the information you’ve got so far to know exactly where your score falls in relation to everybody else’s score. You know that your score was 257 points above the mean. But what was the average amount that the raw scores deviated from the mean? Let’s say the standard deviation of the test scores was 100. Your raw score was 257 points above the mean and the average deviation from the mean was 100. This tells you that your raw score was 2.57 of these average deviations above the mean of 3000. This number is able to show you exactly where your score falls in relation to everybody else’s. Taking one person’s deviation from the mean and dividing it by the average deviation from the mean results in a new transformed score – a **standard score** – that tells you how many standard deviations one raw score fell above or below the mean. The symbol for a standard score is Z . Thus, another term for a standard score is a **Z-score**.

The equation for a standard score is ...

$$Z = \frac{X - \bar{X}}{S}$$

Let’s say that you take every raw score in a set and convert them to standard scores.

X	X- \bar{X}	(X- \bar{X}) ²	Z
7	-2	4	-1.26
6	-1	1	-0.63
8	0	0	0.0
9	+1	1	+0.63
10	+2	4	+1.26

The standard deviation is 1.58 so dividing each deviation score by 1.58 results in the standard score for that person. One thing to notice is that the mean of the z-scores is zero. This will be true of any set of z-scores. ***The mean of any set of standard scores is always zero.*** A second thing that we know about any set of standard scores is that ***the standard deviation of any set of standard scores is always one.***

So what’s so handy about standard scores? Well, let’s say that someone asked you the following very reasonable question. You know that a person is 6.0 feet tall and weighs 150 pounds. Is this person taller or heavier? ??? Taller or heavier? “*Well, 150 is a much bigger number than 6, so the person is obviously quite a bit more heavy than they are tall*”. OUCH. I should hope no one was thinking that. Raw scores don’t carry

enough information to let you answer this kind of question. A score measured on the scale of “feet” can’t be compared to a score measured on the scale of “pounds”. However, as we’ll see in a second, standard scores don’t have this limitation because ***all standard scores are measured in the same units – standard deviations above or below the mean.***

So now let’s say you know that the mean height in the class is 5.6 feet and that the standard deviation for height is .3 feet. The mean weight in the class is 125 pounds and the standard deviation for weight is 10 pounds.

Converting the raw score for height to a standard score tells you that the person is 1.3 standard deviations above the mean for height. Converting the raw score for weight to a standard score tells you that the person is 2.5 standard deviations above the mean for weight.

The person is 1.3 standard deviations above the mean for height, but 2.5 standard deviations above the mean for weight. So, compared against the rest of the students in the class the person is more heavy than they are tall.

Standard scores have the advantage of a common unit of measurement. No matter what the original units of measurement were – feet, pounds, seconds – when you convert raw scores to standard scores the numbers will be in units of ***standard deviations above or below the mean.*** This allows you to compare scores of measures with dramatically different units of measurement. You can compare a person’s score on job performance to their score on job satisfaction or a person’s score on depression to their score for the amount of social support they’re receiving.

Percentile scores and the normal curve

Let’s go back to the example of getting a score on your first exam. A standard score of +2.5 tells you that your score was 2.5 standard deviations above the mean, which is very good. However, not every reader or listener will know how to interpret that kind of number. If you wanted to provide a number that indicated where one score was in relation to everybody else’s, it would be more intuitive to be able to say that a person’s raw score was better than 97% of the scores for other people taking the test. That value 97 is referred to as a **percentile score**.

A percentile score is a number that tells you the percentage of scores that fell below the one you were interested in. So a percentile score of 25 indicates that 25% of all the scores in the set were below that one. The median of a data set represents the 50th percentile. By definition, 50% of all the scores for a particular variable fall below the median and 50% fall above it.

Converting a standard score to a percentile score doesn’t seem like it would be a particularly easy thing to do, except that statisticians something about how the scores in a data set typically fall that makes it very easy to perform this conversion. We know that if

the distribution of scores for a variable looks like the *normal curve* then a specific percentage of scores fall between the center of the curve and any standard score we might be interested in. The shape of the normal curve describes the relationship between the possible values one might observe for a particular variable and the number of times that each score actually showed up in the data.

The normal curve table provided in Table --- gives us percentages that correspond to different standard scores we might be interested in. For example, a percentage of 34.13 is listed for the standard score of 1.00. This indicates that 34.13% of all the scores in the set fall between a standard score of zero and a standard score of 1.00. If you knew that a person had a standard score of 1.00, it's not too much of a stretch to figure out their percentile score from there. You know that the normal curve is symmetrical, so 50% of all the scores fall below a standard score of zero and 50% fall above a standard score of zero. You know that 34.13% of the scores fall between a standard score of zero and the standard score we're interested in (i.e., 1.00). So the 34.13% we just looked up in the table plus the 50% of scores below zero gives us a total of 84.13%. A person with a standard score of 1.00 thus has a percentile score of 84.13. This is, of course, assuming that the scores for that particular variable are normally distributed.

Examples:

1. What percentile score corresponds to a standard score of -1.68 ?
2. What percentage of scores fall between standard scores of $+0.45$ and $+2.52$?
3. What percentage of scores fall between standard scores of -1.57 and $+0.46$?

Transforming percentile scores to standard scores

Let's say that you work for a company and are told that the ten percent of employees with the best job performance ratings will receive a bonus for that year. Your job is to determine the lowest job performance rating that an employee have and still get the bonus. You know that the mean job performance rating is 23.0 and that the standard deviation for ratings is 5.0.

Where do you start? Well, the first thing you have to figure out is the percentile score you're dealing with. In this case, it's the 90th percentile because you're trying to find the performance score where 10% of the scores fall above it and 90% of the scores fall below it.

Okay, now what. Well, you've got the normal curve table to help you. You know that the numbers in the body of the table represent percentages – percentages of scores between a standard score of zero and the particular standard score that we're interested in. In this problem, instead of looking up a particular standard score and then getting the percentage of scores that goes with it, now we've got to do just the opposite. We've got to look up a percentage and then look for the standard score goes with it.

So what percentage do we look up in the table? This table starts with the middle of the normal curve (a standard score of 0.0) and it works its way outward from there. We're working with the 90th percentile, which means that if we take the 50% of scores on the negative side of the normal curve into account, we're left with 40% of the scores on the positive of the curve. So we need to look up 40 in the body of the normal curve table (or as close as we can get to it). The closest we can find to 40 in the body of the table is 39.97. Now, all you have to do is to see which row you're in and which column you're in. The row will give you the standard score out to one decimal place. This number turns out to be 1.2. The column tells you the number that goes in the second decimal place. This turns out to be .08. When you put these numbers together, 1.2 + .08 gives you 1.28. This means that the lowest standard score an employee could get and still be in the top 10% is +1.28.

But you're still not through. You know which standard score to use as a cutoff for awarding bonuses, but what's the raw score? To find the raw score that corresponds to our cutoff, we can use the equation for a standard score that we worked with before.

$$Z = \frac{X - \bar{X}}{S}$$

We know every number that goes in this equation except the one number we need...

$$1.28 = \frac{X - 23.0}{5.0}$$

... so all we have to do is solve for X and we'll have our answer.

$$X = (1.28)(5.0) + 23.0$$

$$X = 6.4 + 23 = \mathbf{29.4}$$

The lowest job performance score that an employee could have and still get the bonus is 29.4.

Examples:

1. What standard scores define the outer 5% of the normal curve?
2. What is the lowest standard score you could have that would still be included in the highest 5% of scores in a set?