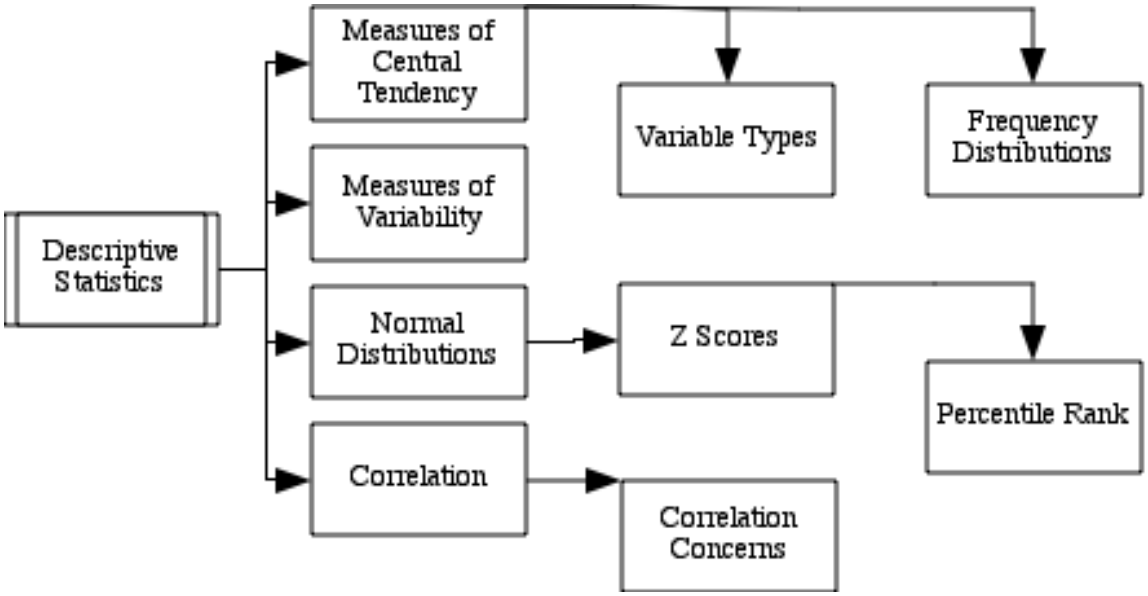
Chapter 7: Descriptive Statistics



Chapter Overview

Chapter 7 provides an introduction to basic strategies for describing groups statistically. Statistical concepts around normal distributions are discussed. The statistical procedures of z-scores and correlation are presented.

INTRODUCTION

Although qualitative studies often include numerical information, that information is meant to provide more meaning to what you have been collecting to better understand the

groups or individuals you are studying. It is another way to triangulate data gathered in qualitative research. Like all qualitative data gathering, the validity of numerical data is based on the degree to which it informs your research question. What if you are not doing a qualitative study, but have a question that requires quantitative methods and analyses?

As we discussed in Chapter 2, the usual form of a quantitative study is to gather and analyze numerical data in order to show if something you did or observed impacted one or more groups in a predictable way. Did a new curriculum raise test scores? Does age make any difference in teacher attitudes? Do intrinsic rewards decrease disruptive behavior? Once you have a quantitative research question in mind your job as a researcher is to plan a strategy for gathering and analyzing data to answer the question.

This leads to one of the biggest differences between qualitative and quantitative research. Qualitative research questions are answered by using iterative and dynamic methods. You gather data, review what has been gathered and then gather more data until you have a sufficient understanding of your topic. What data are gathered and how it is gathered may change as the study progresses. In quantitative studies this is not the case. In quantitative studies once you have designed a strategy for gathering and analyzing data you will carry out the design. The methods of your study do not change as the study progresses. You need to have planned carefully before data collection begins. The next few chapters of this book are intended to give the tools necessary to plan well and to be able to complete insightful quantitative research.

DESCRIPTIVE STATISTICS

To begin, we need to approach numerical understanding of groups in a more formal way. This will eventually lead to analysis tools designed to describe differences among groups using a kind of formal mathematical logic. The first step is to get used to the ideas around using numbers (rather than words) to describe groups.

Think about talking to a friend and describing a party you recently attended. It is likely you would be talking about others who were there. John is now in graduate school. Mary has taken a job as a graphic designer. Bob has lost a lot of weight. Carol and Steve got married. To describe the group, you describe characteristics of the individuals. We do the same thing in quantitative research. If we want to know how tall a group of seventh graders is we would measure each child's height. The resulting list would be like the description of the party. William is 52 inches tall. Sally is 47 inches tall. Juan is 50 inches tall. When you look at all of the heights you would be able to get a sense of the general height of the members of the group.

The best way to analyze a group is to know each member of the group and be able to compare some characteristic among them. The problem is that our brains are not very good at keeping lots of detailed items separate and available for use. At some point there would be so many heights to remember that you would have trouble generalizing about the group. When this happens you have to figure out a way to summarize all of the numbers. In statistics, summarizing the characteristics of a group is called **descriptive statistics.**

Descriptive statistics are a kind of shorthand to make it easier to talk about a group as a whole instead of talking about groups by describing each individual. For example, it

would not make sense to say the average age of children is 26.4 months old if there are only two of them, since you would get much more information if you were just given each of their ages. However, if you were given each child's height in a class this would get confusing. The summary numbers are much more useful when there are many members of the group.

MEASURES OF CENTRAL TENDENCY

Now it is time to begin to review those things that most of us already know about statistics. When describing groups quantitatively, we usually try to come up with some number that represents as many members of the group as possible. These "summary" numbers representing where most of the members of the group appear are called **measures of central tendency**. There are three of these. The **mode** is the number that appears most often in a list. If 6 of our seventh graders turned out to be 47 inches tall and no other specific height occurred that many times in the group, then the mode of the group would be 47. The **median** is the number where half of the group measures lower and half is higher—it is the middle, like the median of a road. If there were 23 students in the class and all of the heights were sorted into increasing order, then the 12^{th} height—the one right in the middle of the list—would be the median height of the class. If you had 24 students, the median would be the average of the 12^{th} and 13^{th} heights (since there is no one middle number). The **mean** is the arithmetic average. If you take all of the measured heights, add them all together and divide by the number of students measured, that would be the mean height of the class.

Variable Types

The problem is figuring out which of these three measures of central tendency will give you the group summary which is the best description of the group. In order to do that we first need some background about how group data are collected.

When you start to research a group you will pick specific characteristics, or **variables**, of the individuals in the group that are of interest to you. These are things that you would not expect to be the same for everyone in the group. You might be interested, as in our example above, in the height of each student. More likely you might be interested in their grade point average or how many books they each read last week. There is an infinite number of possible variables in a group and it is your job as a researcher to choose just those variables that you need to help you answer your research question.

Generally, there are three types of variables. You need to know about these so that you can choose the best measure of central tendency to describe the variable for the group. Imagine wanting to know the pet preference for the children in your class. Some would like cats most and some dogs; maybe some have snakes as a favorite pet. When you are looking at how each child answered this question you could sort the responses into dogs, cats, snakes, birds and probably not more than a few other categories—7 children like dogs best, 9 like cats most, and so on. The responses have been sorted into containers but the containers do not have a logical order. It would not make any difference if cats were put before dogs or even snakes came first. You would not get any

useful information from the order of the response categories as you were examining the results. Variables like this—responses that can only be sorted into containers that have no logical order—are called **nominal variables**. The most important thing about the response categories for these variables is the category's name, hence nominal variable.

Going back to the measures of central tendency, imagine trying to average pet preference. The question does not make any sense because the response categories do not have a fixed order. If we ask what the median pet preference is the same problem occurs. There is no way to order the responses so that you can figure out the middle point. So, with nominal variables the only measure of central tendency that is available is the mode. Which response category has the most responses in it? A statement describing a group with a mode would be something like: more students said cats were their favorite pet than any other pet type.

Most of the time we gather data from groups in ways that the response categories do have a logical order. Imagine asking your class how often they read at home. It might be very difficult for students to put an exact number to the answer of that question but they probably would be able to select from these categories: hardly ever, once a month, once a week, more than once a week. Responses to variables designed this way are called **ordinal variables**. The responses categories have a logical order, but the categories are not necessarily equivalent—a month includes a lot more possible reading days than a week. The most important characteristic of this type of variable is the relative order of the response categories.

When you got the data back from the students you would be able to sort the responses into the categories just like with nominal variables. If you wanted you could report the mode of the responses (i.e., more students said that they read once a week than students in any other category), but there is more information available to summarize the group because the response categories are in order. It is still not possible to average the responses because, as noted above, the response categories are not of equivalent size. Take all of the responses and put them in order—putting all of the "almost never" responses first, then the once a month responses, then the once a week responses followed by the more than once a week responses. Now count through the responses until you find the one in the middle of the list. This is the median response. It describes the point in the responses where half of the students responded below this point and half responded above. Since the purpose of descriptive statistics is to provide the best description of the group possible, the median gives more information about how the responses from the group are distributed than the mode does. That usually makes it the best measure of central tendency to use with ordinal variables. A sentence describing a group with a median would be something like: Reading at home once a week was the median response for the class.

Finally, whenever possible, researchers try to use variables where the response categories are ordered *and* they are of equivalent size. These are called **interval variables**. In our example of asking the students how tall they were, students responded with their height in inches. Certainly inch measurements have a natural order but it is important that each category—each inch measurement—is the same size. An inch is an

inch whether it is at the 2 inch point on the ruler or the 40 inch mark. With interval variables the most important characteristic is that each response category represents an equivalent interval.

Interval variables are the only variable type where determining the mean (averaging) is possible. When the data are gathered, the responses can be averaged and the mean becomes a much more descriptive statistic than the median or the mode. The median and mode could still be computed for interval data, but those numbers would generally not tell as much about the group as the mean would. In our case a statement describing a group with a mean might read: The mean height of the students in this class was 48.6 inches. Read another way the mean represents a point around which we would expect most of the responses from the group to cluster.

There is a special case of interval variables called **ratio variables**. Ratio variable scales always start at zero. If you are talking about height you can say that someone is twice as tall as someone else. Or, you could say that a car got one third the gas mileage as another. These are ratio statements. Think about saying that someone is twice as smart as someone else—it does not make sense because intelligence scales or standardized assessments scales do not start at zero. In most cases, the way ratio and interval variables are used in statistics is the same but it is important to remember the difference between these types of variables.

We will describe in the next chapter how to do statistical analysis with these measures of central tendency. Right now you should keep in mind that whenever possible (and it will not always be possible—how would you gather interval data on gender?) you

should gather data with interval variables because they not only provide the summary description of the group with the most information, but they also allow for the most sophisticated analyses of the data once they are gathered.

Frequency Distributions

When you are reporting descriptive statistics in a research paper you will almost always put the important numbers into a table. Sometimes it is valuable to look at your data in graphic forms other than tables to better understand what they mean. Start by determining how many responses from your group on a given variable there are for each response category—how many responded almost never, how many once a month, how many once a week and how many more than once a week. This is called a **frequency distribution**. When you make a bar chart from the frequency distribution it is called a **histogram**. Although histograms can be made for all three variable types, remember that the order of the bars in a chart for a nominal variable has no meaning. Regardless, charting frequency distributions is a good way to see how your data are spread out.

Here is an example using an interval variable. Imagine recording these scores from a 30 point test: 27, 28, 28, 27, 27, 24, 27, 26, 20, 30, 23, 23, 24. A frequency distribution and measures of central tendency of these scores would look like this.

20 — 1
21 — 0
22 — 0
23 — 2
24 — 2
25 — 0
26 — 1
27 — 4

mode = 27
median = 27
mean = 25.69

28 — 2
29 — 0
30 — 1

The histogram for the data is depicted in Figure 7-1. Review this figure to see how the scores are spread out. Can you make some generalizations about the class just from looking at the chart? The information in the frequency distribution and the histogram are the same; however, it is usually easier to get a sense of the group from the chart rather than the table of numbers.

[Insert Figure 7-1 here: Histogram A].

Stop and try this yourself. Here are scores for a second 30 point test: 22, 22, 23, 23, 25, 26, 27, 27, 27, 27, 28, 28, 29. Compute a frequency distribution, draw a histogram and then determine the values of the measures of central tendency for these scores. Does your histogram look like the one in Figure 7-2? What generalizations about the class can you make from this histogram?

[Insert Figure 7-2 here: Histogram B]

MEASURES OF VARIABILITY

The two previous examples (tests one and two) pose an interesting problem. They represent two different frequency distributions but the measures of central tendency are

identical. Since the purpose of descriptive statistics is to give the best possible summary description of the group, it appears that measures of central tendency may not always provide enough information by themselves for differentiating among different sets of data. Something else is needed to summarize differences in groups.

With interval variables (remember, the intervals between response categories are equal) the histogram represents not only how many responses are in each category but also how far apart the response categories are. Because the possible response categories are equidistant we can get a reasonable representation of whether the data are spread out across the histogram or whether the data seem to be clustered more closely together. Look at the two previous examples to get a sense of this. Another form of descriptive statistics, then, represents how far the responses are spread apart in a distribution. These are called **measures of variability**.

There are two measures of variability that are generally used. The **range** represents the difference between the highest and lowest responses in the data. Sometimes the range is listed as the actual highest and lowest scores and sometimes it is listed as the difference between those two numbers. In the first of the cases previously discussed, the range runs from a low score of 20 to a high of 30 or a difference of 10. In the second of the cases, the range would show responses as low as 22 and as high as 29 or a distance between those two points of 7. The range shows the responses being clustered into a smaller space on the response scale than the first set of data.

The second measure of variability is the **standard deviation**. In general terms this is the average distance the responses are from the mean score. To figure out what the

standard deviation is you would find the difference of each score from the mean and then average all of those differences. In reality the calculation is a bit more complicated. To statisticians it looks like this:

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

In this book we will leave that calculation to the computer. For a given set of data, higher standard deviations would mean that the responses are more spread out (greater average distance from the mean). Lower standard deviations would mean that the responses are more closely clustered around the mean (smaller average distance from the mean).

With the standard deviations computed it is possible to differentiate between our two example groups. See Table 7-1.

[Insert Table 7-1 here: Mean and Standard Deviation for Two Example Tests]

The standard deviation for the second group is a lower number indicating a smaller average distance that the scores are from the mean. We would expect the histogram to show the responses clustered more closely around the mean than they would be in the first group. Go back to the two charts (Figures 7-1 and 7-2) and see if you can identify this difference.

As a teacher looking at test scores, a smaller standard deviation would generally mean that students all performed similarly regardless of the group's overall performance. A larger standard deviation would show that the scores on the test were more spread out. Some students scored particularly high or low compared to the rest of the group. To

inform how to best improve student learning, it might be worth figuring out why that was so.

Since a standard deviation is a measure relative to a mean score, it is only available with interval data. The range is the measure of variability reported for ordinal data. Together, the measures of central tendency and the measures of variability give a powerful way to describe groups and eventually you will see that they form the foundation from which groups can be statistically compared.

You may feel that this introduction to statistical calculation inspires you to learn how to let the computer do much of this calculation for you. We have included a brief introduction into using Microsoft Excel for statistical work in Appendix F.

NORMAL DISTRIBUTIONS

If we were able to measure the height of all middle-school aged boys, we would be able to determine the mean height. And then if we plotted a histogram of the heights we would find that most of the boys' heights would be clustered around the mean height, and that a few would be considerably taller and some would be considerably shorter. We would not expect every boy to be the same height, but we would expect most of them to be relatively the same, with only a few particularly tall or short. This characteristic distribution of lots of responses being close to the mean and fewer responses appearing as you move above and below the mean is typical of many variables measured in living systems: the size of apples from a tree, the weight of sea turtles, or the cognitive ability of 14- year- olds.

Keeping the histogram of this kind of distribution in your mind, imagine connecting the tops of each bar with a line. The bar in the middle with the most responses would be very close to the mean of the group. Then a line connecting the top of each bar would begin to fall in each direction as the bars got shorter (fewer people in those response categories). Now, with a bit of mental gymnastics, imagine smoothing out the line connecting the tops of all of the bars and you would be left with a smooth curve that is high in the middle and drops down on both sides. If the group is large enough from which you were gathering data, this curve would be likely to take on a very even and symmetrical appearance (Figure 7-3). Because the curve represents "normally distributed" data (data that acts the way we expect) it is called a **normal curve**. Sometimes you will see it referred to as a characteristic curve or a standard curve.

[Insert Figure 7-3 here: Normal Curve]

This curve has certain characteristics. The mean is in the middle with data evenly distributed on both sides. The curve in its ideal form is symmetrical. The category that has the most responses (the highest bar, the mode) is also in the middle and because it is symmetrical half of the responses are on each side of the highest bar (the median). Consequently, the mean, median and mode are the same number in a normally distributed group of data. And, the curve represents this characteristic of normally distributed data in that most of the responses are close to the mean and there are fewer responses as the response category is farther from the mean.

Looking at this from the point of view of measures of central tendency and variability, the mean is in the center of the curve and the standard deviation controls how wide the curve is—with larger standard deviations the responses will be spread farther apart, with smaller standard deviations the responses would be more closely clustered around the mean.

There is one more characteristic of normal curves that is very important (see Figure 7-4). Look at the point on the curve that is one standard deviation beyond the mean. (Remember, the standard deviation is the average of the differences of the individual data points from the mean.) Having just said that the standard deviation determines the width of the curve, imagine what happens as the standard deviation gets bigger and smaller. As the standard deviation moves, the curve gets stretched out or compacted together. It is beyond the scope of this book to show you exactly why this is the case, but the area under the curve between the mean and one standard deviation (the responses on the histogram between those two points) represents the exact same percentage of the responses from the group regardless of what the mean and standard deviation of the responses are. Let's put this in mathematical terms. The percentage of the responses in a normal distribution between the mean and one standard deviation is a constant—it never changes. It turns out that 34.1 percent of the responses are in that area.

[Insert Figure 7-4 here: Area between the mean and one standard deviation in a normal curve]

Well, if the area between the mean and one standard deviation is constant then it is not a surprise that the area between one and two standard deviations is also constant. It is 13.6 percent of the responses. And, the area between two and three standard deviations represents 2.1 percent of the responses. If you add those up you will see we are getting very close to 50 percent (all those responses above the mean). In fact, in a normal distribution there will only be 0.1 percent of the responses beyond three standard deviations. Naturally, these relationships hold true if you look at standard deviations below the mean as well (see Figure 7-5). If a group mean is 24.7 and one standard deviation is 6.2 that would mean that 34.1 percent of the responses from this group would be between 24.7 and 30.9 (24.7 + 6.2). Similarly 34.1 percent of the responses would be between 24.7 and 18.5 (24.7 – 6.2).

[Insert Figure 7-5: Standard Deviation and Percent of the Normal Curve]

Z-Scores

In general, statistics are used to describe groups and relationships among groups. Seldom are they used to describe individuals. In the examples of test scores, we were looking at the class data. For heights, we were describing the entire class. There is one notable exception to this in which statistics help to compare an individual to the group. In an earlier section of this chapter, we described the relationship of standard deviation to a normal curve. It is easiest to make that description by looking at the points that are whole number standard deviations above or below the mean. Actually, we could compute the

area under the normal curve between the mean and any given standard deviation above or below the mean. For instance the area under the curve (recall, the area under the curve represents the percent of responses) between the mean and 1.23 standard deviations above the mean is about 39 percent of the total responses.

Since the normal curve is a simplified graph that really represents the frequency distribution of the group, we can look at any specific response (be it a particular test score, a certain number of books read, a specific height of child—whatever you are measuring) and figure out how many standard deviations above or below the mean that specific response is. The calculation to do this is very simple. Figure out the distance between the mean of the group and the specific response in which you are interested (subtract the mean from the specific response) and divide by the standard deviation. Using our example numbers of a mean of 24.7 and a standard deviation of 6.2, we can do this calculation for a specific individual's score of 32.

(Score-Mean)/Std. Dev. = Number of Standard Deviations Away from the Mean

or

(32-24.7)/6.2 = 1.18

If the specific response in which you are interested is to the left of the mean the result will be negative (Score-Mean will be negative) and if it is to the right it will be positive. Because ours is to the right of the mean it is a positive value. What you would

have done through this computation is determine the distance from the mean to your score of interest in units of standard deviation. This number is called a **z score**.

Rarely are specific z scores reported as a way to compare an individual's response to the larger group. Instead, what is reported is the *z* score converted into the area under the curve (the percentage of responses) that the z score represents. Since the areas between the mean and any given standard deviation are all constants, this last process is really easy. Look up the z score on a table prepared for this purpose. This table is called a z table and you can find one in Appendix D.

The number that you get from the z table needs a little explanation. What we really want to know is how to compare a person's response to the larger group. To do this the comparison that is made is to determine the percentage of the responses from the whole group that are lower than the individual's that you are examining. If the z score is positive (the score of interest is above the mean) that would mean that the percentage of responses lower than that score is the area between the mean and the score of interest plus 50 percent. The 50 percent is the area below the mean. In our example of a z score of 1.18 the area between the mean and 1.18 standard deviations is 38 percent. Add the 50 percent of the area below the mean and that shows that 88 percent of the scores of this group are below a score with a z score of 1.18. Figure 7-6 depicts this in graphic form.

[Insert Figure 7-6 here: Positive Z- Score]

If the z score is negative (it is below the mean) then you would subtract the area under the curve between the mean and score of interest from 50 percent. For instance if the z score is -1.18 you would subtract 38 (the percent of responses calculated from the z-table) from 50 with a result of 12 percent of the responses being lower than the score represented by 1.18 standard deviations lower than the mean. This is shown in Figure 7-7.

[Insert Figure 7-7 here: Negative Z-Score]

Percentile Rank

You have probably run into this number before with a slightly different name. A response which has a z score of 1.18 has a **percentile rank** of 88. This trail through z scores has led to that magic number that is on students' standardized test results. A student has a percentile rank in which his or her score on the test is compared to some group. It could be a comparison to everyone who has taken the test or it could be a comparison to some subset (i.e., school, district, gender) of the larger group. Percentile rank does not represent the percentage of the scores that a student got right or wrong on the test. It does represent the percentage of students who scored lower on the test than a specific student. Not only is the concept of computing the areas of a normal curve important for more complex statistical analysis but also, as a teacher, you may find yourself explaining percentile rank to parents.

CORRELATION

There is one more important way that you can describe a group. With percentile rank we are describing the relationship of individuals to the larger group. With a **correlation** you can describe the relationship of one characteristic of a group to another characteristic. If data were gathered from a group of teachers about their age and number of years they had been teaching we would expect that in general younger teachers would have taught fewer years and older teachers would have taught more years. This probably would not be true for every teacher you surveyed, but in general we would expect to see a propensity for this to appear over the whole group. The point of computing correlations is to see how strong that relationship is (how likely it is to be true in every case) and the direction of the relationship; that is, does one go up as the other goes up or does one go down as the other goes up.

All statistical computer applications, including Microsoft Excel, make computing correlations simple. When the responses from two variables are compared in this way, the numerical correlation, or the correlation coefficient, is a number between 1 and -1. If the relationship between two variables is absolutely predictable (in every case if you knew the score on one variable you could predict the score on another) the correlation would be +1 or -1. If the two variables change in the same way (as one is larger the other will be larger), the correlation is +1. It would be -1 if the variables behave in opposite directions (as one is larger the other gets smaller). As the measure of the relationship of two variables is less predictable, the number approaches 0 from the two extremes of 1 and -1. The closer a computed correlation is to zero the less likely that you would be able to

predict one from another. In general if a correlation appears between 1 and 0.7 or -0.7 and -1 then it would be called a strong correlation. Those between 0.3 and 0.7 or -0.3 and -0.7 are considered moderate correlations, and those between 0.3 and -0.3 are weak (Figure 7-8). These are subjective terms but they serve to help to know when to pay attention to correlations and when to assume that there really is not much of a relationship between variables.

[Insert Figure 7-8 here: Levels of Correlation Strength]

Many studies are designed so that examining the relationship among many variables in the study at the same time is important. Statistical programs will allow you to make all of the necessary computations simultaneously. If there are ten variables that you wish to check against each other, what results is a correlation table in which the correlation of each variable with every other variable of interest is listed. In these tables the variables are listed on both the horizontal and vertical axis. At the intersection of any two variables, the correlation is listed. These tables have a characteristic diagonal appearance because the results are symmetrical. Comparing variable 1 on the horizontal axis with variable 2 on the vertical is the same as variable 2 on the horizontal with variable 1 on the vertical. Consequently, those redundant comparisons are not left in the chart.

Table 7-2 shows an example **correlation matrix** from student evaluations of a course. We wanted to know how strongly correlated the responses to each question were from the students. Note that the "q" represents question.

[Insert Table 7-2 here: Question Response Correlations from a Course Evaluation (N=65)]

Sometimes researchers will include the 1 in the results table at each of the points where something is correlated with itself (the diagonal row of 1s in the table) to make the table easier to read. In Table 7-2, the highest correlation is between q1 and q2 (.79, a *strong* correlation) and the lowest is between q2 and q3 (.59, a *moderate* correlation).

Correlations are a powerful descriptive tool because they can demonstrate the degree to which two variables are measuring similar things. In the course evaluation example described earlier, we designed the survey to ask students about the quality of teaching in the course. If the questions were designed well we would expect the questions to show strong correlations. If there were some that did not show strong correlations, we might want to look more closely to see if some of the questions were really measuring something other than perceptions of teaching.

There are a few other things that you should pay attention to as you are using correlations in your studies. First, it is a good idea to do a graphic representation of the relationship of the variables. By doing a **scatter plot** (one variable represented on the x axis of a graph and the other variable on the y axis) the resulting points on the graph would ideally cluster around a line called a **regression line** or a **best fit** line that runs through the plotted data points (Figure 7-9). In Microsoft Excel this line can be generated by adding a **linear trend line** to a scatter plot. The closer all of the points are

to this line, the closer the correlation will be to 1 or -1. If the points do not seem to cluster around the line the correlation is more likely to be closer to 0. After making a scatter plot, you also need to make sure that the data appear as a single cluster no matter how tightly the cluster is around the best fit line. In some cases your responses will gather into two or more clusters. When this happens it is probable that something else is affecting these variables making the computed correlation a false representation of the relationship among the variables.

[Insert Figure 7-9 here: Correlation Scatter Plots]

Notice in the examples in Figure 7-9 that with a positive correlation the regression line goes from lower on the left to higher on the right. As one variable increases so does the other. With a negative correlation the opposite is true. The regression line goes from higher on the left to lower on the right. As one variable increases the other decreases. As the correlation approaches 0.00 the regression line becomes horizontal.

Concerns when Using Correlations

In Chapter 8 we will discuss the probability that statistical relationships you observe happened by chance. Without going into detail about that right now, statistical programs will compute, along with the correlation, the probability that the observed correlation is something that might have appeared randomly. The determination of whether the correlation is likely to have happened by chance is related to the strength of the

correlation along with the size of the group from which you have gathered the responses. Since it is possible to compute correlations without determining this probability they would appear by chance (for instance Excel does not provide probability numbers), you must be careful that you are not reporting on a relationship that is meaningless because chance is a better explanation of the relationship than any other meaning you might attribute to it. This is particularly important if you are discussing moderate correlations with group sizes of less than twenty. Use the table included in Appendix E to determine when a computed correlation is not likely to have occurred by chance.

Two other problems appear in using correlations. First, you will occasionally read reports on studies (usually but not always in non-peer reviewed sources) where the author suggests that a high correlation indicates that one variable causes the results on the other. In our example above this would be like saying that years of teaching causes age. As easy as it would be to humorously suggest that relationship, it clearly makes no sense. It may be that a causal relationship between variables exists but a correlation cannot be used to demonstrate that.

The other difficulty is that correlations need to make sense. Statistically it may be possible to demonstrate correlations between variables that don't seem to be logically connected. You might find a relationship between height and reading scores but it would be difficult to make the case of why these two should be related. Usually, there should be something in the literature that you can reference to support any curious findings. Otherwise be clear when you are reporting on unanticipated relationships that further investigation is warranted to better understand them.

There are two common ways that most correlations are computed. One is called a **Pearson's r**. It is used when both variables being compared are interval variables. When one of the variables is ordinal then a more conservative calculation is used called **Spearman's rho.** If the statistical program you are using doesn't differentiate between the two types of correlation (i.e., Excel) then you should assume it is using the Pearson calculation. Whenever possible you should try to set up correlation studies to compare interval variables.

NEXT STEPS

Descriptive statistics are often part of qualitative research studies. They can serve as important triangulation data for your results. When descriptive statistics are used in quantitative studies, they usually are the foundation for statistical analysis. Quantitative studies require substantially different designs from qualitative studies to accommodate statistical analysis. In the next chapter we will take a closer look at those analysis tools. Regardless of which type of research you are thinking of doing, understanding these basics of measures of central tendency, normal curves, and correlation will be important not only for gathering and analyzing data for you own study, but for reading and evaluating others' research as well.

CHAPTER SUMMARY

Having completed this chapter, you should be comfortable discussing the following:

- measures of central tendency: mean, mode, median

- variable types (nominal, ordinal, interval)
- frequency distributions, including histograms
- measures of variability: range and standard deviation
- description of a normal distribution
- z-scores and percentile rankings
- correlations

CHAPTER REVIEW

1. How will you know which measure of central tendency will be best for your study?

2. What are the differences among the three variable types? Why are those differences important?

3. What is a normal distribution? What does it represent?

4. What is percentile rank and how is determined?

5. When would you use a correlation in a study?

Figure 7-1: Histogram A

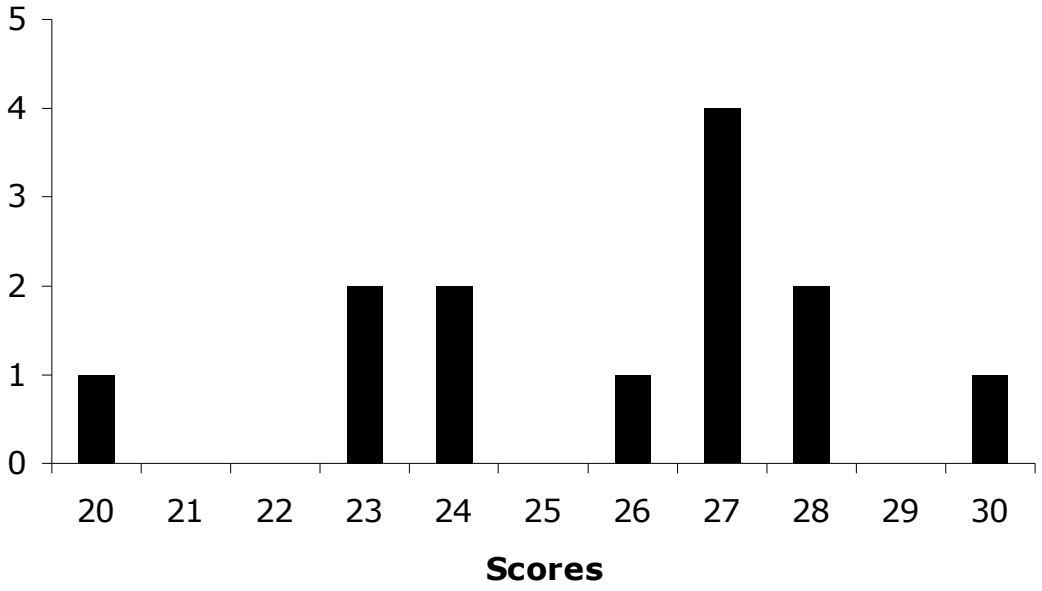


Figure 7-2: Histogram B

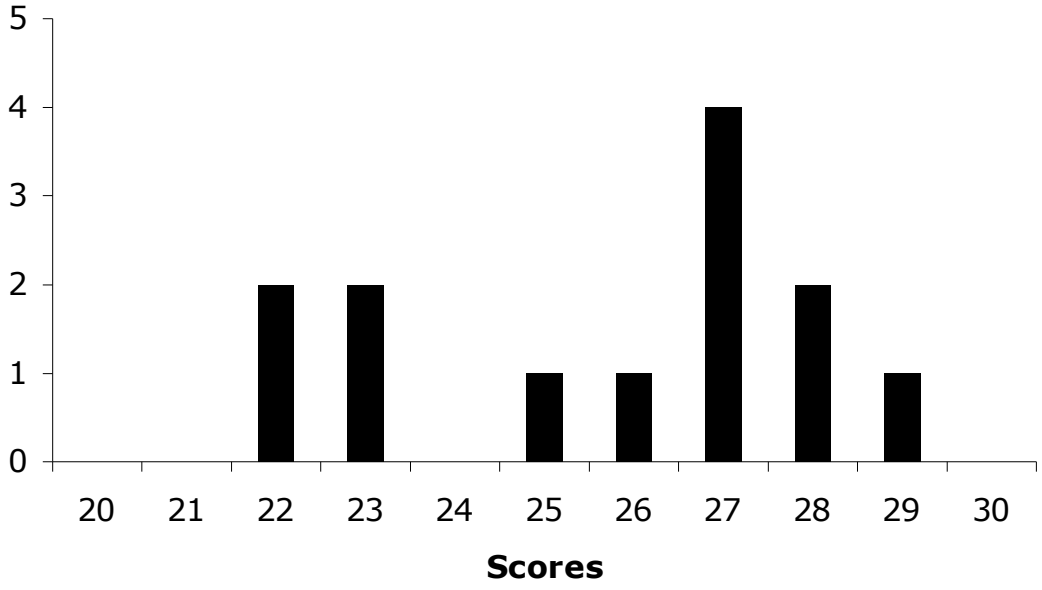


Figure 7-3: Normal Curve

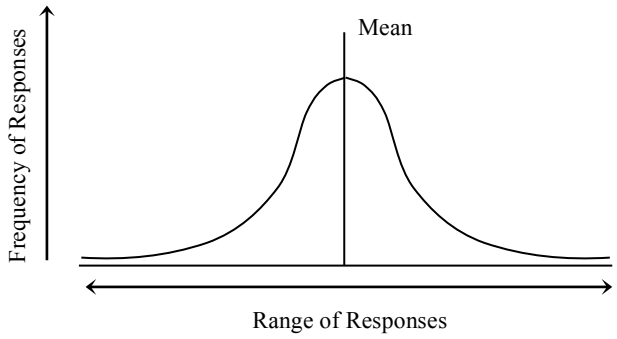


Figure 7-4: Area between the mean and one standard deviation in a normal curve

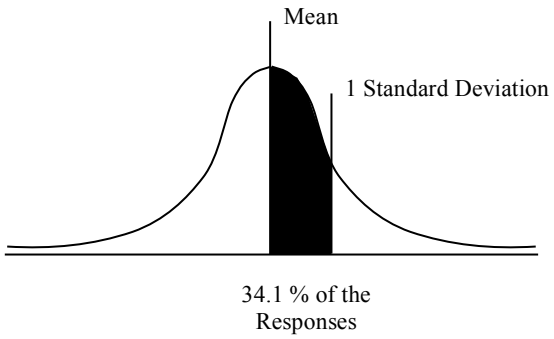


Figure 7-5: Standard Deviation and Percent of the Normal Curve

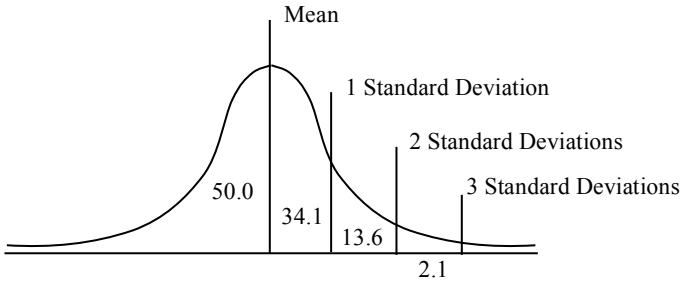


Figure 7-6: Positive Z-Score

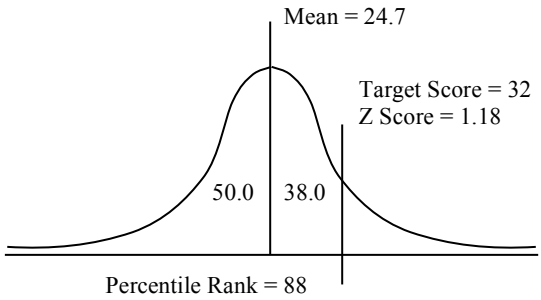


Figure 7-7: Negative Z-Score

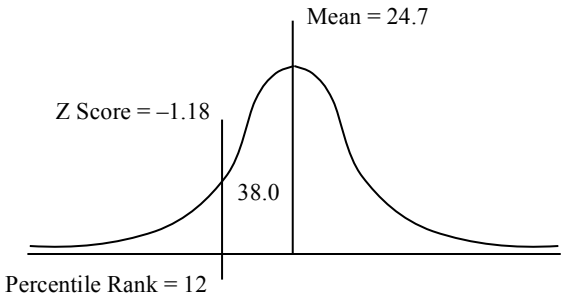


Figure 7-8: Levels of Correlation Strength

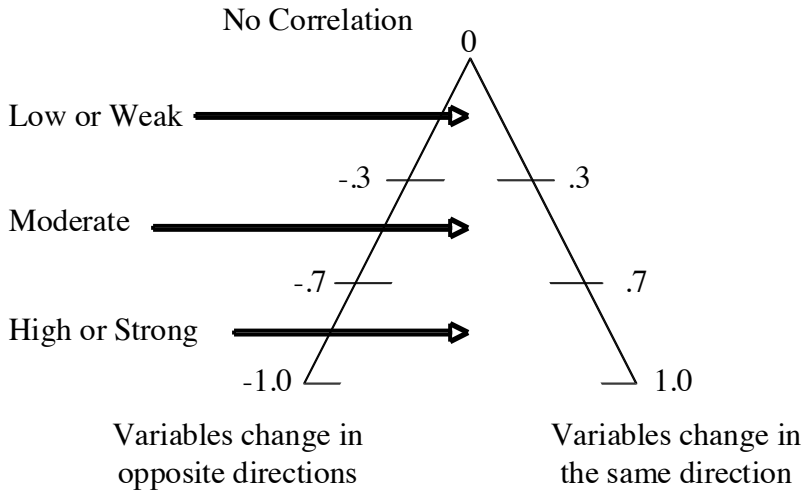
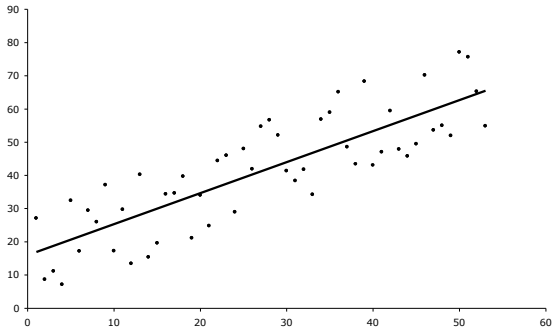
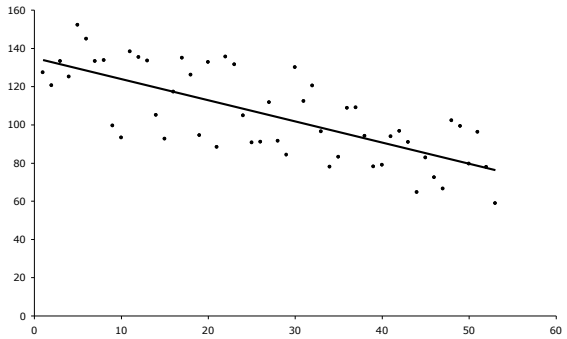


Figure 7-9: Correlation Scatter Plots

Strong positive correlation (0.83)



Strong negative correlation (-0.73)



Almost no correlation (-0.02)

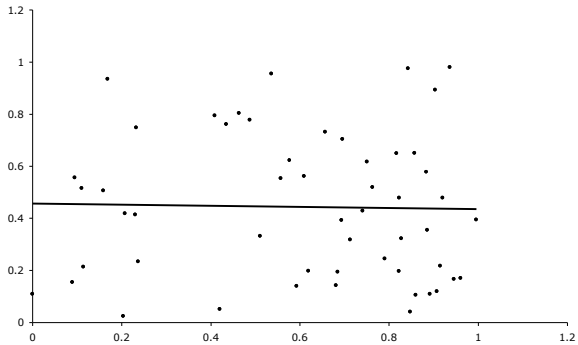


Table 7-1

Mean and Standard Deviation for Two Example Tests

	Test 1 $N = 13$	Test 2 $N = 13$
Mean	25.69	25.69
Standard Deviation (S.D.)	2.72	2.43

Table 7-2

Question Response Correlations from a Course Evaluation (N = 65)

	q1	q2	q3	Q4	q5
q1	1				
q2	.79	1			
q3	.62	.59	1		
q4	.73	.72	.77	1	
q5	.78	.73	.68	.71	1