The Correlation Coefficient

In chapter 1 you learned that the term “correlation” refers to a process for establishing whether or not relationships exist between two variables. You learned that one way to get a general idea about whether or not two variables are related is to plot them on a “scatterplot”. If the dots on the scatterplot tend to go from the lower left to the upper right it means that as one variable goes up the other variable tends to go up also. This is a called a “*positive relationship*”.

On the other hand, if the dots on the scatterplot tend to go from the upper left corner to the lower right corner of the scatterplot, it means that as values on one variable go up values on the other variable go down. This is called a “*negative relationship*”. If you are unclear about this, ***please*** return to Chapter 11 and make sure that you understand what is written there before you continue!

While using a scatterplot is an appropriate way to get a *general* idea about whether or not two variables are related, there are problems with this approach. These include:

- Creating scatterplots can be tedious and time consuming (unless you use a computer)
- A scatterplot does not really tell you exactly how strong any given relationship may be
- If a relationship is weak—as most relationships in the social sciences are—it may not even be possible to tell from looking at a scatterplot whether or not it exists
- It is difficult to make accurate predictions about variables based solely on looking at scatterplots—unless the relationship is very strong

And so let’s add a new tool to add to our statistical tool box. What we need is a single summary number that answers the following questions:

a) Does a relationship exist?
b) If so, is it a positive or a negative relationship? and
c) Is it a strong or a weak relationship?

Additionally, it would be great if that same summary number would allow us to make accurate predictions about one variable when we have knowledge about the other variable. For example, we would like to be able to predict whether or not a convicted criminal would be likely to commit another crime after he or she was released from prison.

We are not asking for much, are we?

Well, there is such a number. It is called the *correlation coefficient*.

Correlation Coefficient: A single summary number that gives you a good idea about *how closely one variable is related to another variable.*

In order for you to be able to understand this new statistical tool, we will need to start with a scatterplot and then work our way into a formula that will take the information provided in that scatterplot and translate it into the correlation coefficient. As with most applied statistics, the math is not difficult. It is the concept that is important. I typically refer to formulae as recipes and all the data as ingredients. The same is true with the formula for the Correlation Coefficient. It is simply a recipe. You are about to learn how to cook up a pie—a nice and tasty Correlation Pie!

Let's begin with an example. Suppose we are trying to determine whether a the length of time a person has been employed with a company (a proxy for experience) is related to how much the person is paid (compensation). We could start by trying to find out if there is any kind of relationship between "time with company" and "compensation" using a scatterplot.

In order to answer the question "Is compensation related to the length of time a person has worked for the company?" we could do something like the following:

- STEP 1 – Create a data file that contains all individuals employed by the company during a specific period of time.
- STEP 2 – Calculate how long each person has been employed with the company..
- STEP 3 – Record how much each person is compensated in, say, hourly pay (in the real world you would probably use annual total compensation).
- STEP 4 – Create a scatterplot to see if there seems to be a relationship.

Suppose that our study resulted in the data found in table 12-1, below.

TABLE 2-1
Example Data File Containing Fictitious Data

Employee's Initials	Compensation (In dollars per hour)	Number of months employed with the company
J.K.	5	45
S.T.	15	32
K.L.	18	37
J.C	20	33
R.W.	25	24
Z.H.	25	29
K.Q.	30	26
W.D.	34	22
D.Q.	38	24
J.B.	50	15

Once we have collected these data, we could create the scatterplot found in Figure 2-1, below. Notice that the dots tend to lay in a path that goes from the upper left area of the scatterplot to the lower right portion of the scatterplot. What type of relationship does this seem to indicate? How strong does the relationship seem to be?

The scatterplot in Figure 2-1 indicates that there is a negative relationship between "Time With Company" and "Hourly Pay". This means that the longer an individual has been employed with the company, the less they tend to be paid—a very strange finding!

Note that this does not mean that Time With Company actually causes lower compensation (correlation does not equal causation) it only shows that there is a relationship between the two variables and that the variable tends to be negative in nature.

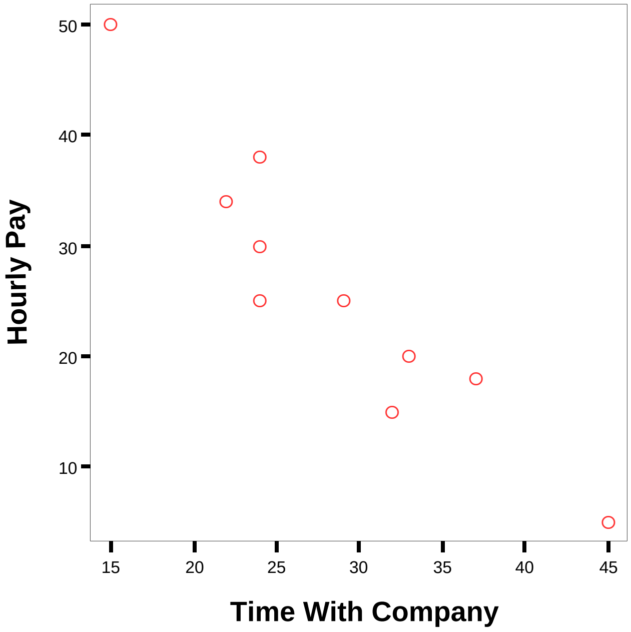
Important Note:

"Correlation does not equal causation". To be correlated only means that two variables are related. You cannot say that one of them "causes" the other. Correlation tells you that as one variable changes, the other seems to change in a predictable way. If you want to show that one variable actually causes changes in another variable, you need to use a different kind of statistic which you will learn about later in this book.

You should also be able to see that the negative relationship between Time With Company and Comensation seems to be pretty strong. But how strong is it? This is our main problem. We really can't say anything more than direction of the relationship (negative) and that it is strong. We are not able to say just how strong that relationship is.

A *really* smart guy named Karl Pearson figured out how to calculate a summary number that allows you to answer the question "How strong is the relationship?" In honor of his genius, the statistic was named after him. It is called Pearson's Correlation Coefficient. Since the symbol used to identify Pearson's Correlation Coefficient is a lower case "r", it is often called "*Pearson's r*".

FIGURE 2-1
Scatterplot of minutes of exercise by post-partum depression symptoms (Fictitious data)



The Formula for Pearson's Correlation Coefficient

$$r_{xy} = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n}}{\sqrt{(SS_x)(SS_y)}} \quad \text{OR} \quad r_{xy} = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n}}{\sqrt{\left[\left(\Sigma X^2 - \frac{(\Sigma X)^2}{n_x}\right)\left(\Sigma Y^2 - \frac{(\Sigma Y)^2}{n_y}\right)\right]}}$$

Gosh! Is that scary looking or what? Are you feeling intimidated? Is your heart starting to pound and your blood pressure starting to rise? Are your palms getting sweaty and are you starting to feel a little faint?

If so (and I am sure this describes pretty accurately how some who are reading this feel right about now!), take a deep breath and relax. Then take a close look at the formula.

Can you tell me how many separate things you really need to calculate in order to work this beast out? Think it through.

Think of it like a loaf of bread. Just as a loaf of bread is made up of nothing more than a series of ingredients that have been mixed together and then worked through a process (mixing, kneading, and baking), so it is with this formula. Look for the ingredients. They are listed below:

- $\sum X$ This simply tells you to add up all the X scores
- $\sum Y$ This tells you to add up all the Y scores
- $\sum X^2$ This tells you to square each X score and then add them up
- $\sum X^2$ This tells you to square each Y score and then add them up
- $\sum XY$ This tells you to multiply each X score by its associated Y score and then add the resulting products together (this is called a "cross-products")
- n This refers to the number of "pairs" of data you have.

These are the ingredients you need. The rest is simply a matter of adding them, subtracting them, dividing them, multiplying them, and finally taking a square root. All of this is easy stuff with your calculator.

Let's work through an example. I am going to use the same data we used in Table 2-1 when we were interested in seeing if there was a relationship between an employee's Time With Company and his or her compensation. However, even though we are going to use the same data, the table I am going to set up to make our calculations easier will look a lot different.

Take a look at table 2-2, below. Notice that I have created a place in this table for each piece of information I need to calculate r_{xy} using the computational formula ($\sum X$, $\sum Y$, $\sum X^2$, $\sum Y^2$, $\sum XY$)

TABLE 2-2
Example of a way to set up data to make sure you don't make mistakes when using the computational formula to calculate Pearson's r

X	**X^2**	**Y**	**Y^2**	**XY**
5		45		
15		32		
18		37		
20		33		
25		24		
25		29		
30		26		
34		22		
38		24		
50		15		
$\sum X$=	$\sum X^2$=	$\sum Y$=	$\sum Y^2$=	$\sum XY$=

Notice that I listed my X values under “X” and right next to that I have a column where I will put my squared X values (X^2). Then I have a column where I list my Y values under “Y” with a column right next to it where I can put my squared Y values (Y^2). I also have a column where I can put the results of multiplying each X value by its associated Y value. Finally, notice that I even have a row along the bottom where I can put all of the “Sum of” values.

Filling In The Table

In order to use the table, start by filling in the X^2 and Y^2 values. See Table 2-3, below.

TABLE 2-3
Entering the squared values of the X and Y scores

X	**X^2**	**Y**	**Y^2**	**XY**
5	25	45	2,025	
15	225	32	1,024	
18	324	37	1,369	
20	400	33	1,089	
25	625	24	576	
25	625	29	841	
30	900	26	676	
34	1,156	22	484	
38	1,444	24	576	
50	2,500	15	225	
$\sum X=$	$\sum X^2=$	$\sum Y=$	$\sum Y^2=$	$\sum XY=$

Next, multiply each X score by its paired Y score which will give you the cross-products of X and Y. See Table 2-4, below, for an example. Notice that I have bolded the scores you need to multiply in order to get the cross-products.

TABLE 2-4
Calculating the cross-products by multiplying each X score by its corresponding Y score

X	X^2	Y	Y^2	XY
5	25	**45**	2,025	225
15	225	**32**	1,024	480
18	324	**37**	1,369	666
20	400	**33**	1,089	660
25	625	**24**	576	600
25	625	**29**	841	725
30	900	**26**	676	780
34	1,156	**22**	484	748
38	1,444	**24**	576	912
50	2,500	**15**	225	750
$\sum X=$	$\sum X^2=$	$\sum Y=$	$\sum Y^2=$	$\sum XY=$

After you have filled in the last column which contains the cross-products of X and Y, all you have to do is fill in the last row of the table which contains all of you “Sum Of” statements. In other words, just add up all of the X scores to get the $\sum X$, etc. See Table 2-5, below to see an example based on our current data.

TABLE 2-5
Calculating the sums of each of the columns

X	X^2	Y	Y^2	XY
5	25	**45**	2,025	225
15	225	**32**	1,024	480
18	324	**37**	1,369	666
20	400	**33**	1,089	660
25	625	**24**	576	600
25	625	**29**	841	725
30	900	**26**	676	780
34	1,156	**22**	484	748
38	1,444	**24**	576	912
50	2,500	**15**	225	750
$\sum X=$ 260	$\sum X^2=$ 8,224	$\sum Y=$ 287	$\sum Y^2=$ 8,885	$\sum XY=$ 6,546

Now, you have everything you need to fill out that nasty looking computational formula for Pearson's Correlation Coefficient! That's not too hard, now is it?
Let's just begin by plugging the numbers in where they go.

- Wherever you see a "∑X" just enter the number you calculated in your table, which is **260**.
- Wherever you see a "∑X^2" enter the number you calculated for that in your table, which is **8,224**.
- Wherever you see a "∑Y" enter the number you calculated for that which is **287**.
- Wherever you see a "∑Y^2" enter the number you calculated which is **8,885**.
- Wherever you see a "∑XY" enter the number you calculated with is **6,546**.
- Finally, wherever you see an "n" enter the number of pairs of data you have which in this example is **10**.

Look at the computational formula again.

$$r_{xy} = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n}}{\sqrt{\left[\left(\Sigma X^2 - \frac{(\Sigma X)^2}{n_x}\right)\left(\Sigma Y^2 - \frac{(\Sigma Y)^2}{n_y}\right)\right]}}$$

In order to use the formula to calculate a correlation coefficient by hand, all you have to do is carefully go through the following steps:

STEP 1 – Enter the numbers you have calculated in the spaces where they should go. See below. Make sure you understand where to put each of the numbers (e.g., ∑XY , ∑X, etc.).

$$r_{xy} = \frac{6{,}546 - \frac{(260)(287)}{10}}{\sqrt{\left[\left(8{,}224 - \frac{(260)^2}{10_x}\right)\left(8{,}885 - \frac{(287)^2}{10_y}\right)\right]}}$$

STEP 2 – Multiply the (∑X)(∑Y) in the numerator (the top part of the formula) and do the squaring to $(\sum X)^2$ and $(\sum Y)^2$ in the denominator (the bottom part of the formula). Make sure that you clearly understand what I have done below.

$$r_{xy} = \frac{6{,}546 - \dfrac{74{,}620}{10}}{\sqrt{\left[\left(8{,}224 - \dfrac{67{,}600}{10_x}\right)\left(8{,}885 - \dfrac{82{,}369}{10_y}\right)\right]}}$$

STEP 3 – Do the division parts in the formula.

$$r_{xy} = \frac{6{,}546 - 7{,}462}{\sqrt{[(8{,}224 - 6{,}760)(8{,}885 - 8{,}236.9)]}}$$

STEP 4 – Do the subtraction parts of the formula.

$$r_{xy} = \frac{-916}{\sqrt{[(1{,}464)(648.1)]}}$$

STEP 5 – Multiply the numbers in the denominator.

$$r_{xy} = \frac{-916}{\sqrt{948818.4}}$$

STEP 6 – Take the square root of the denominator.

$$r_{xy} = \frac{-916}{974.073}$$

STEP 7 – Take the last step and divide the numerator by the denominator and you will get…

r_{xy} = -.940

…the Correlation Coefficient! Trust me. Once you get used to setting up your data into a table like I showed you in this example, you can compute a correlation coefficient easily in less than 10 minutes (as long as you are not doing it with too many numbers).

What Good Is A Correlation Coefficient?

As can see above, we just did a whole lot of calculating just to end up with a single number: **-0.94038**. How ridiculous is that? Seems kind of like a waste of time, huh? Well, guess again! It is actually very cool! ("Yeah, right!" you say, but let me explain.)

Important Things Correlation Coefficients Tell You

They Tell You The Direction Of A Relationship

If your correlation coefficient is a negative number you can tell, just by looking at it, that there is a negative relationship between the two variables. As you may recall from the last chapter, a negative relationship means that as values on one variable increase (go up) the values on the other variable tend to decrease (go down) in a predictable manner.

If your correlation coefficient is a positive number, then you know that you have a positive relationship. This means that as one variable increases (or decreases) the values of the other variable tend to go in the same direction. If one increases, so does the other. If one decreases, so does the other in a predictable manner.

Correlation Coefficients Always Fall Between -1.00 and +1.00

All correlation coefficients range from -1.00 to +1.00. A correlation coefficient of -1.00 tells you that there is a *perfect negative relationship* between the two variables. This means that as values on one variable *increase* there is a perfectly predictable *decrease* in values on the other variable. In other words, as one variable goes up, the other goes in the opposite direction (it goes down).

A correlation coefficient of +1.00 tells you that there is a *perfect positive relationship* between the two variables. This means that as values on one variable *increase* there is a perfectly predictable *increase* in values on the other variable. In other words, as one variable goes up so does the other.

A correlation coefficient of 0.00 tells you that there is a zero correlation, or no relationship, between the two variables. In other words, as one variable changes (goes up or down) you can't really say anything about what happens to the other variable.

Sometimes the other variable goes up and sometimes it goes down. However, these changes are not predictable.

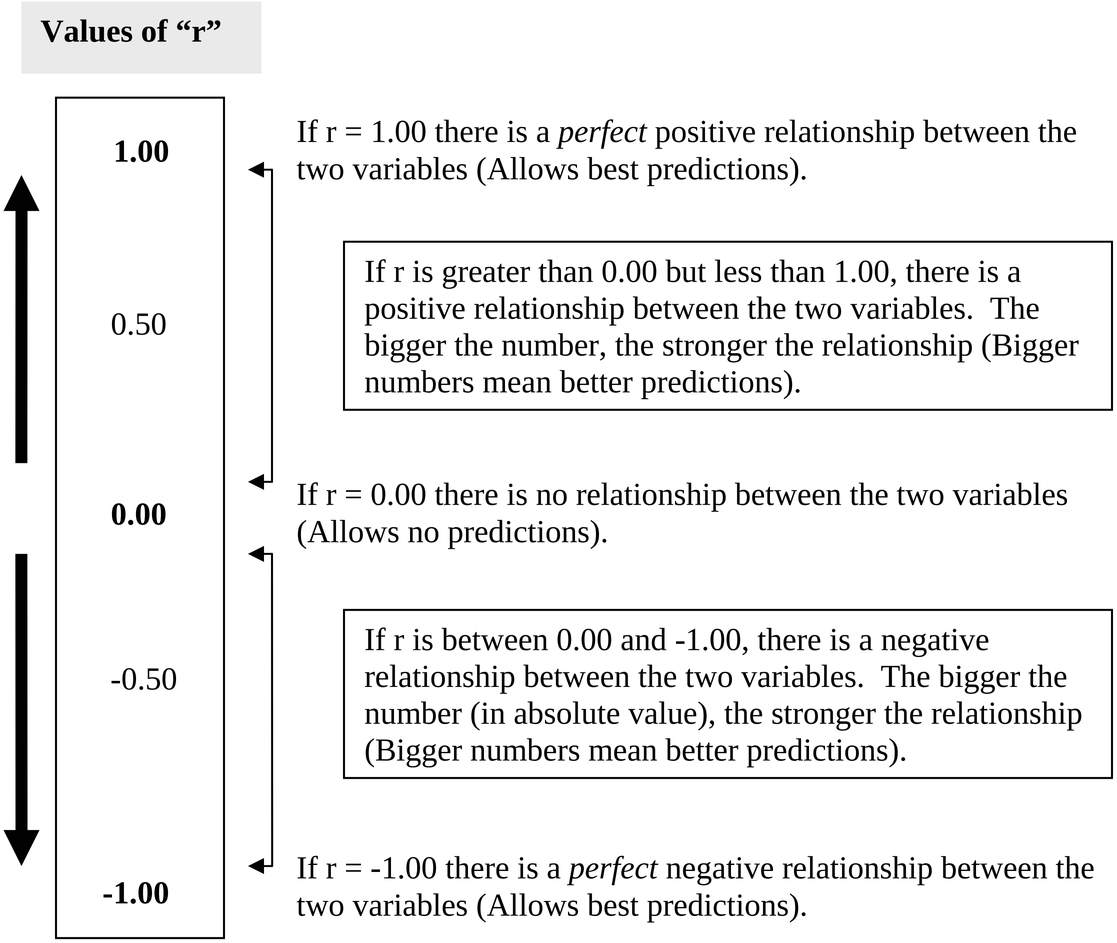
Correlation coefficients are always between - 1.00 and +1.00. If you ever get a correlation coefficient that is larger than + or – 1.00 then you have made a calculation error. Always pay attention to this if you are calculating a correlation coefficient!

Larger Correlation Coefficients Mean Stronger Relationships

Most correlation coefficients (assuming there really is a relationship between the two variables you are examining) tend to be somewhat lower than plus or minus 1.00 (meaning that they are not perfect relationships) but are somewhat above 0.00. Remember that a correlation coefficient of 0.00 means that there is no relationship between your two variables based on the data you are looking at.

The closer a correlation coefficient is to 0.00, the weaker the relationship is and the less able you are to tell exactly what happens to one variable based on knowledge of the other variable. The closer a correlation coefficient approaches plus or minus 1.00 the stronger the relationship is and the more accurately you are able to predict what happens to one variable based on the knowledge you have of the other variable. See Figure 2-4, below, for a graphical representation of what I am talking about.

FIGURE 2-2
What Does A Correlation Coefficient Mean?



Values of "r"

1.00

0.50

0.00

-0.50

-1.00

If r = 1.00 there is a *perfect* positive relationship between the two variables (Allows best predictions).

If r is greater than 0.00 but less than 1.00, there is a positive relationship between the two variables. The bigger the number, the stronger the relationship (Bigger numbers mean better predictions).

If r = 0.00 there is no relationship between the two variables (Allows no predictions).

If r is between 0.00 and -1.00, there is a negative relationship between the two variables. The bigger the number (in absolute value), the stronger the relationship (Bigger numbers mean better predictions).

If r = -1.00 there is a *perfect* negative relationship between the two variables (Allows best predictions).

Suppose that you read a research study that reported a correlation of r = .75 between compensation and number of years of education. What could you say about the relationship between these two variables and what could you do with the information?

- You could tell, just by looking at the correlation coefficient that there is a positive relationship between level of education and compensation. This means that people with more education tend to earn higher salaries. Similarly, people with low levels of education tend to have correspondingly lower salaries.
- The relationship between years of education and compensation seems to be pretty strong. You can tell this because the correlation coefficient is much closer to 1.00 than it is to 0.00. Notice, however, that the correlation coefficient is not 1.00. Therefore, it is not a perfect relationship.
- As a result, if you were examining an employer's compensation practices, you could use the knowledge that education is related to compensation to "model" the company's compensation practices. This means that you can generate predictions of how much each employee should be compensated based on his or her level of education. Your predictions will not be perfectly accurate because the

relationship between education and compensation is not perfect. However, you will be much more accurate than if you simply guessed!

Let me share another example. Let's say that you are interested in finding out whether gender is related to (or predictive of) compensation. You conduct create a data set that identifies the gender for each employee (0 = female and 1 = male, for example) and correlate gender with the employee's compensation.

You compute a correlation coefficient to evaluate whether gender is related to compensation and you get an r = -.45. What could you say based on this correlation coefficient?

- You could tell from looking at the correlation coefficient that, since it is a negative number, there is a negative relationship between gender and compensation (as gender goes up compensation goes down). In this case, we would say that people with "1s" tend to have lower compensation than people with "0s". In other words, males tend to have lower compensation than females. This kind of finding with a variable like gender (or alternatively with minority/non-minority status where 0 = minority and 1 = non-minority) is a BAD thing! Ideally, we want the correlation between gender and/or minority status to be r = 0.00—meaning that the variable has nothing to do with compensation. Anything else could indicate potential discrimination or bias.
- In this example, the relationship between gender and compensation seems to be moderately strong (it is about halfway between 0.00 and -1.00).
- Okay, so how would you use this information? Since you have established that a relationship exists between these two variables, you can use knowledge about how gender is generate how much people would be compensated just considering gender. Then, what you would do is look at those females who had high compensation rates to see if there is some valid reason (e.g., higher-level positions, more education, etc.) Finally, you could use this information to help identify how much the pay of males needs to be increased to eliminate the relationship.

There's More!

In addition to telling you:

A. Whether two variables are related to one another,
B. Whether the relationship is positive or negative and
C. How large the relationship is,

the correlation coefficient tells you one more important bit of information--it tells you exactly how much variation in one variable is related to changes in the other variable.

Many students who are new to the concept of correlation coefficients make the mistake of thinking that a correlation coefficient is a percentage. They tend to think that when r = .90 it means that 90% of the changes in one variable are accounted for or related to the other variable. Even worse, some think that this means that any predictions you make will be 90% accurate. *This is **not** correct!* A correlation coefficient is a "ratio" not a percent.

However it is very easy to translate the correlation coefficient into a percentage. All you have to do is "square the correlation coefficient" which means that you multiply it by itself. So, if the symbol for a correlation coefficient is "r", then the symbol for this new statistic is simply "r^2" which can be called "r squared".

There is a name for this new statistic—the Coefficient of Determination. The coefficient of determination tells you the percentage of variability in one variable that is directly related to variability in the other variable.

The Coefficient of Determination

r^2, also called the "Coefficient of Determination", tells you how much variation in one variable is directly related to (or accounted for) by the variation in the other variable.

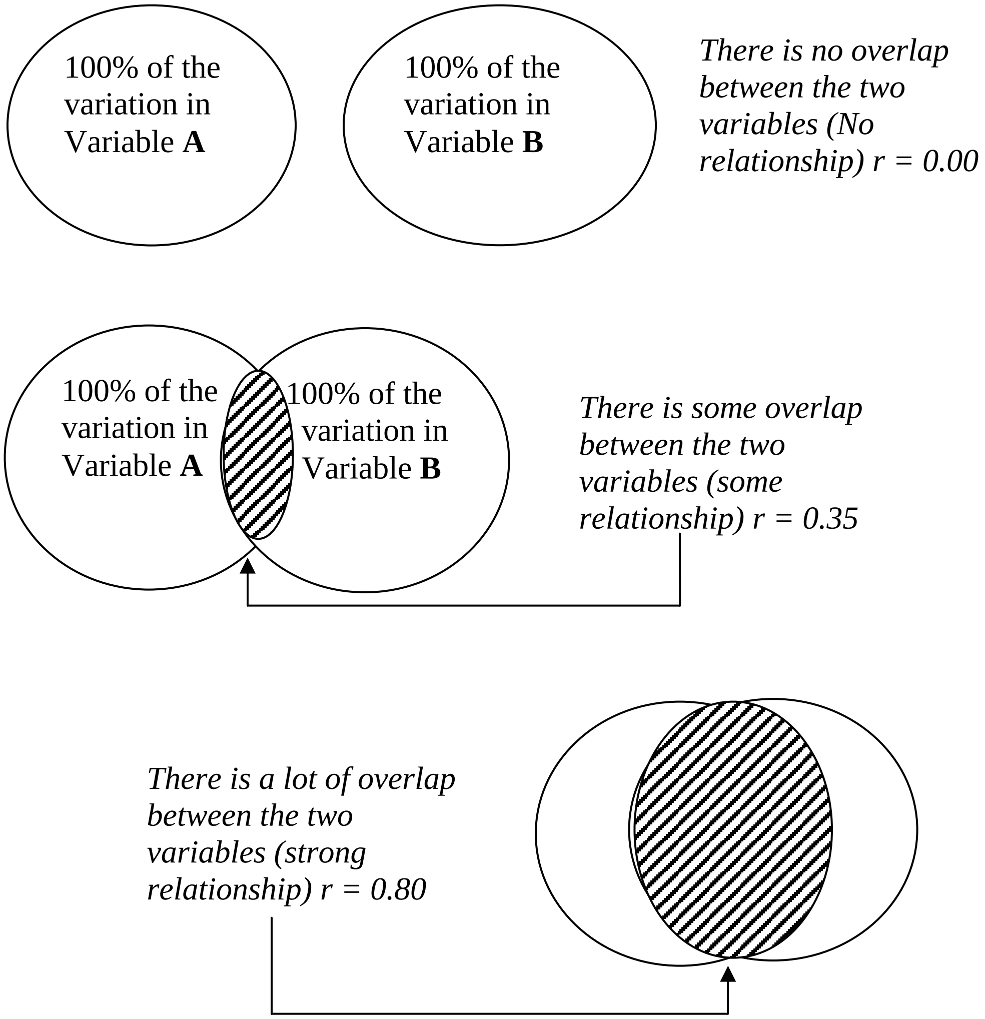
As you will learn in the next chapter, the Coefficient of Determination helps you get an idea about how accurate any predictions you make about one variable from your knowledge of the other variable will be. The Coefficient of Determination is very important. Look at Figure 2-3, below.

From looking at Figure 12-5 several things should be fairly clear.

A. In the first example, there is no overlap between the two variables. This means that there is no relationship and so what we might no about Variable A tells us nothing at all about Variable B.
B. In the second example, there is some overlap. The correlation coefficient is r = 0.25. If you square that to get the coefficient of determination (r^2) you would get 12.25%. This tells you that 12.25% of how a person scored on Variable B is directly related to how they scored on Variable A (and vice-versa). In other words, if you know a person's score on Variable A you really know about 12.25% of what there is to know about how they scored on Variable B! That is quite an improvement over the first example. You could actually make an enhanced prediction based on your knowledge.
C. In the third example, there is even more overlap between the two variables. The correlation coefficient is r = 0.80. By squaring r to get r^2, you can see that fully 64% of the variation in scores on Variable B is directly related to how they scored on Variable A. In other words, if you know a person's score on Variable A, you know about 64% of what there is to know about how they scored on Variable B. This means that in the third example you could make much more accurate predictions about how people scored on Variable B just from knowing their score on Variable A than you could in either of the first two examples.

To summarize, larger correlation coefficients mean stronger relationships and therefore mean that you would get higher r^2 values. Higher r^2 values mean more variance accounted for and allow better, more accurate, predictions about one variable based on knowledge of the other. The next chapter focuses solely on the notion of how to make predictions using the correlation coefficient. It is called linear regression.

FIGURE 2-3
What Is Meant By Saying One Variable Is Associated With Another Variable


100% of the variation in Variable **A**
100% of the variation in Variable **B**
100% of the variation in Variable **A**
100% of the variation in Variable **B**

There is no overlap between the two variables (No relationship) r = 0.00

There is some overlap between the two variables (some relationship) r = 0.35

There is a lot of overlap between the two variables (strong relationship) r = 0.80

How To Interpret Correlation Coefficients

Hopefully, you now understand the following:

- That the Correlation Coefficient is a measure of relationship.
- That the valid values of the Correlation Coefficient are -1.00 to +1.00.
- Positive Correlation Coefficients mean positive relationships while negative Correlation Coefficients mean negative relationships.

- Large Correlation Coefficients (closer to +/- 1.00) mean stronger relationships whereas smaller Correlation Coefficients (close to 0.00) mean weaker relationships.
- A Correlation Coefficient of 0.00 means no relationship while +/- 1.00 means a perfect relationship.
- The Correlation Coefficient works by providing you with a single summary number telling you the average amount that a person's score on one variable is related to another variable.
- How to calculate the Correlation Coefficient using the z-score formula (also called the definitional formula).
- That while the Correlation Coefficient is not a measure of the percent of one variable that is accounted for by the other variable, r^2 (also called the Coefficient of Determination) is a measure of the percent of variation in one variable that is accounted for by the other variable.
- Large r^2 values mean more shared variation which means more accurate predictions are possible about one variable based on nothing more than knowledge of the other variable.
- This ability to make predictions is very powerful and can be extremely helpful in business, government, education and in psychology. It can save millions of dollars and improve the lives of many people.

Making Statistical Inferences

Okay, you have calculated a correlation coefficient. The next step you take is critical. You need to determine whether or not your results are "real" or if you got them purely by chance. Remember, rare events happen. Therefore, it is always possible that your analysis is based on a group of really strange people, or "outliers" and that your results are simply due to a chance event. You need to determine whether or not your findings can be extrapolated to the general population.

Remember, one of the chief goals of statistical analysis is to conduct a study with a small sample and then "generalize" your findings to the larger population. This process is called making a "statistical inference."

Before you can say that your correlation coefficient—which tells you very accurately about what is happening in your sample—would be equally true for the general population, you need to ask yourself the important question: "Could I have gotten a correlation coefficient as large as I found simply by chance?" In other words, is it possible that there really is no relationship between these two variables but that somehow I just happened to select a sample of people whose scores made it seem like there was a relationship?

IMPORTANT!

Before you can claim that the relationship you found between the two variables based on your sample also exists in the general population, you need to determine how likely it is that you would have gotten your result just by chance.

How do you determine whether or not your correlation is simply a chance occurrence or if it really is true of the population at large? All you have to do is look your correlation up in a table and compare it to a number that you will find there and you will get your answer. This concept will be discussed in greater detail later in this textbook, but for now I am just going to tell you what you need to do to answer this important question.

You will need three things in order to determine whether you can infer that the relationship you found in your sample also is true (in other words, "is generalizable" to) in the larger population (NOTE: Most statistical packages will compute this for you. However, the steps below tell you how to do it by hand):

1. The Correlation Coefficient that you calculated (for example, r = .65)
2. Something called the "degrees of freedom" which is simply the number of pairs of data in your sample minus 2 (Number of pairs – 2). For example, if you had 10 people in the data set you used to calculate the correlation coefficient that means you have 10 pairs of data (each person has two scores, one on the X variable and one on the Y variable). Therefore:
    a. n = 10
    b. The number of pairs of data would be 10
    c. Number of pairs – 2 would be equal to 8
3. The table you will find in Appendix B which lists "The Critical Values of the Correlation Coefficient"

Let me give you an example that should make clear what you need to do.

1. Suppose you have calculated a correlation coefficient of r = .65
2. Your correlation coefficient was based on 10 people therefore your degrees of freedom is 8.
3. If you look at the table in the back on nearly every statistics book which should be called something like "Critical Values of r" or "Critical Values of the Correlation Coefficient" you will see something like the following:

TABLE 2-6
Critical Values of the Correlation Coefficient

	Two Tailed Test	
	Type 1 Error Rate	
Degrees of Freedom	**.05**	**.01**
1	0.997	0.999
2	0.950	0.990
3	0.878	0.959
4	0.811	0.917
5	0.754	0.874
6	0.707	0.834
7	0.666	0.798
8	**0.632**	**0.765**
9	0.602	0.735
10	0.576	0.708
11	0.553	0.684
Etc…	Etc…	Etc…

Notice that I have increased the font size and I have bolded and underlined the row that has 8 degrees of freedom in it? I am trying to make it clear that the first thing you need to do is look down the degrees of freedom column until you see the row with the number of degrees of freedom that matches your sample degrees of freedom. In this case, since you had 10 pairs of data, your degrees of freedom is 8.

The next thing you need to do is look at the two columns to the right that are listed under "Type 1 Error Rate". A Type 1 Error refers to the chance you would have had of finding a correlation between two variables as large as you did—purely by chance—when no correlation really exists. This is called an "error" because you found a relationship that really does not exist in the general population.

IMPORTANT!

A **Type 1 Error** refers to the chance you would have of finding a correlation as large as you did when, in reality, there is no real relationship in the general population. In other words, you found a relationship that does not really exist! A bad thing!

Under the heading of “Type 1 Error Rate” you will see two columns. One of the columns has a “.05” at the top and the other column has a “.01”. These numbers refer to the chance you would have of making a Type 1 Error. In other words, the “.05” tells you that you would have a 5% chance of making a Type 1 Error. Still another way of saying it is “If you conducted this same study 100 times using different people in your sample, you would only find a correlation coefficient as large as you did about 5 times purely by chance *even if a relationship does not exist*.”

Once you have looked down the column for Degrees of Freedom in the table named “Critical Values of the Correlation Coefficient” (Table 12-11, above), look across to the number listed under .05. Do you see that the number listed there is **0.632**? This number is called “*the critical value of r*”.

Why is 0.632 called the “critical value of r”? Because when you compare the absolute value (ignoring the negative or positive sign) of the correlation coefficient that you calculated from your sample of data with this number, if your correlation coefficient is *equal to or bigger* than this critical value, you can say with reasonable confidence that you would have found a relationship this strong no more that 5% of the time if it really did not exist in the population. In other words, the evidence suggests that there really is a relationship between your two variables.

In other words, if based on your research you claim that there is in fact a relationship between whatever two variables you have been studying, you only have about a 5% chance of being wrong. Please note that you could still be wrong—but you can actually say something about how big of a chance that is! This explanation is overly simplistic but it is good enough to give you the general idea. We will discuss decision errors and statistical significance in greater detail in the chapter on Hypothesis Testing.

Now look at the number listed under the heading of **.01**. Notice that the number that is listed for 8 degrees of freedom is **0.765**. This tells you that if your correlation coefficient is equal to or larger than this critical value, then you could expect to find a correlation that large when no relationship really exists—only 1% of the time. In other words, your risk of making a Type 1 Error is less than when you used the number listed under the heading of **.05**.

4. The last step, therefore is to compare your correlation coefficient with the critical value listed in the table. For the example we have been using, r = .65, since the critical value listed under the heading of .05 with 8 degrees of freedom is 0.632, you can tell that *our correlation coefficient is larger than the critical value listed in the table*. This means that if we accept our findings as being "true" meaning that we believe there is in fact a relationship between these two variables in the general population, we run only about a 5% chance of being wrong.
5. When your correlation coefficient is equal to or larger than the critical value, you can say it is "*statistically significant*". Whenever you hear that a research finding is statistically significant, it tells you how much confidence you can have in those findings and you can tell just how much chance there is that they may be wrong!

Just to make sure that you are getting the idea here, try a few examples. Cover up the shaded box below (which has the answers) and try to determine if each of the correlations below are statistically significant at the .05 and .01 levels. As you consider your answers, try and think about what it means if your correlation coefficient is equal to or larger than the critical value.

r = .43 n = 9 degrees of freedom?______ Significant? ( ) .05 ( ) .01

r = .87 n = 4 degrees of freedom?______ Significant? ( ) .05 ( ) .01

r = .83 n = 6 degrees of freedom?______ Significant? ( ) .05 ( ) .01

r = .10 n = 11 degrees of freedom?______ Significant? ( ) .05 ( ) .01

r = .72 n = 8 degrees of freedom?______ Significant? ( ) .05 ( ) .01

The Answers:

r = .43 n = 9 degrees of freedom? **7** Significant? ( ) .05 ( ) .01 **Not Significant**

r = .87 n = 4 degrees of freedom? **2** Significant? ( ) .05 ( ) .01 **Not Significant**

r = .83 n = 6 degrees of freedom? **4** Significant? (**X**) .05 ( ) .01

r = .10 n = 11 degrees of freedom? **9** Significant? ( ) .05 ( ) .01 **Not Significant**

r = .72 n = 8 degrees of freedom? **6** Significant? (**X**) .05 ( ) .01

Researchers in the Social Sciences usually are willing to accept a 5% chance of Type 1 Errors. So, you can probably be safe just getting used to looking under the .05 column.

If you want to have more confidence in your findings, you can decide that you will only accept findings that meet the requirements for .01.

It's All In The Ingredients

Many years ago my sister decided to make a lemon meringue pie. Into the kitchen she went and began her work. At the last minute, however, she had a moment of inspiration. She noticed that there, next to the kitchen sink was a bottle of Peach Brandy. For some reason she thought "I bet that a little brandy would really spice up this pie!"

Taking the bottle in her hands with that little feeling of "I am doing something that I am not supposed to and I am *really* liking it!" she poured some of the liquid into her pie filling. The deed was done.

Later, as we were enjoying the pie, my sister looked around and proclaimed with some satisfaction, "You know why it tastes so good? It has a secret ingredient!" We prodded her with questions until she finally told us what the "secret ingredient" was.

"Peach Brandy", she said.

After a moment my father, who was the guy who drank the Peach Brandy said, "But we don't have any Peach Brandy."

My sister said, "Sure we do! It was next to the kitchen sink!"

At this point I took a close look at the pie. You see, I found the empty Peach Brandy bottle when my dad threw it away and, since I thought the bottle looked "cool', I decided to save it. To clean it out, I had filled it with soapy water. The secret was out. My sister's secret pie ingredient was dish soap!

Believe it or not, this story makes a very important point. Even though my sister accidentally put soap into her pie, it still looked like a pie. It tasted like a pie—kind of (although there was something funny about the taste). You could eat the pie and it wouldn't kill you (at least it didn't kill any of us!). However the actually pie was, in fact defective in a way that you just could not detect simply by looking at it.

The same is true with the correlation coefficient (and statistics in general). The formula for a correlation coefficient is like a recipe. The information that goes into it ($\sum X$, $\sum Y$, etc.) are the ingredients. If you put the wrong ingredients into the formula, you will get a number that may very well look like a correlation coefficient. The only problem is that it is a defective correlation coefficient. Therefore, it will not mean what it should mean.

If you try to interpret a correlation coefficient that has been "cooked" using the wrong ingredients, you will come to incorrect conclusions! This can be a very bad thing!

For example, suppose you conducted a study to see how likely a convicted sexually violent predator was to commit another sexual crime once they are released from prison. As a result of your study, you find a correlation coefficient of $r = .97$ between a predator's personality characteristics and their likelihood of re-offending.

Based on your results, you make decisions about whether or not you are going to release a person from prison. So far, so good.

What if your correlation coefficient was calculated using "ingredients" that were somehow faulty? That would mean your correlation coefficient was faulty. This, in turn, would mean that any conclusions you draw about a person's likelihood of committing another crime would be faulty also.

As a result, you would fail to let some people out who should have been let out and you might release others who should never have been released. The latter result could seriously endanger the community into which the potential sexually violent predator will be released.

Therefore, if you "bake" your correlation coefficient with faulty "ingredients" then you are likely to draw conclusions or make decisions that are not correct.

The moral of the story, therefore, is to make sure you have the correct "ingredients" when you are calculating correlation coefficients. These ingredients are listed below—statisticians call these ingredients "assumptions".

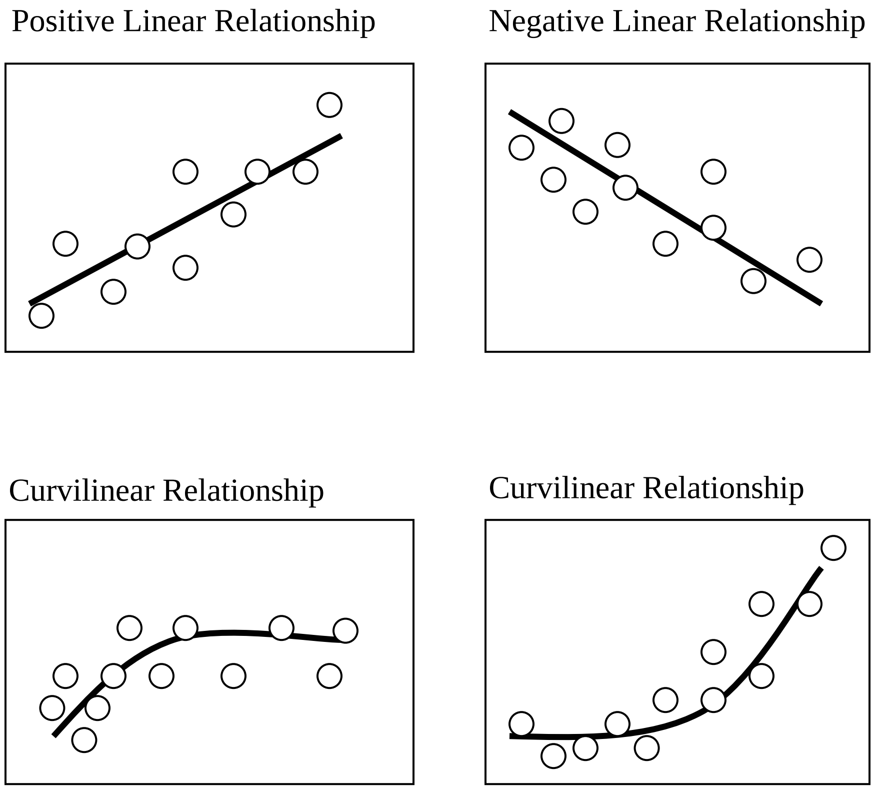
Assumptions for Using Pearson's Correlation Coefficient

1. Your data on **both variables** is measured on *either* an **Interval Scale** or a **Ratio Scale**.

   Interval Scales have *equal intervals* between points on your scale but they *do not have a true zero point*.
   Ratio Scales have both *equal intervals* between points on their scale and they *do have a true zero point*.

2. The **traits you are measuring are normally distributed in the population**. In other words, even though the data in your sample may not be normally distributed (if you plot them in a histogram they do not form a bell-shaped curve) you are pretty sure that if you could collect data from the entire population the results would be normally distributed.
3. The relationship, if there is any, between the two variables is best characterized by a straight line. This is called a "**linear relationship**". The best way to check this is to plot the variables on a scatterplot and see if there is a clear trend from lower left to upper right (a positive relationship) or from the upper left to the lower right (a negative relationship). If the relationship seems to change directions somewhere in the scatterplot, this means that you do not have a linear relationship. Instead, it would be curvilinear and Pearson's r is not the best type of correlation coefficient to use. There are others, however, that are beyond the scope of this book so they will not be discussed (See Figure 12-6, below). It is ok if this assumption is violated as long as its not too bad (sounds really specific, huh?)
4. **Homoscedasticity** – A fancy term that says scores on the Y variable are "normally distributed across each value of the X variable. Figure 12-7, below, is probably more easily understood than a verbal description. Again, one of the easiest ways to assess homoscedasticity is to plot the variables on a scatterplot and make sure the "spread" of the dots is approximately equal along the entire length of the distribution.

FIGURE 2-4
Examples of Linear and Non-Linear Relationships



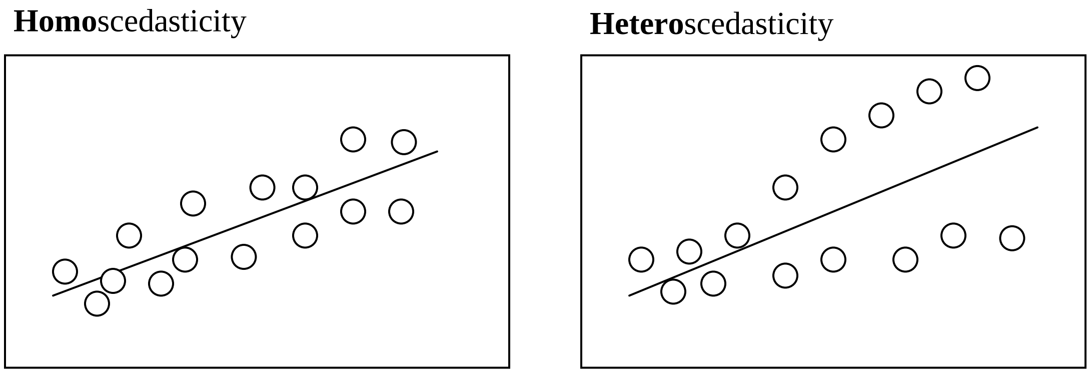
Positive Linear Relationship

Negative Linear Relationship

Curvilinear Relationship

Curvilinear Relationship

FIGURE 2-5
Homoscedasticity vs. Heteroscedasticity



Homoscedasticity

Heteroscedasticity

Notice that in homoscedasticity the dots are evenly distributed around the line while in Heteroscedasticity the dots get farther from the line in some parts of the distribution while close in other parts. *Heteroscedasticity is bad*!

Whenever you are going to calculate a Pearson Correlation Coefficient, make sure that your data meet the assumptions! Otherwise your correlation coefficient—and your conclusions—will be half-baked!

What If the Data Do Not Meet the Assumptions for Pearson's r?

My dad always told me that I should always use the correct tool for the job. I have tried to follow his advice. Once, however, I went camping and forgot all of my cooking utensils and was forced to try and open a can of chili with an axe and eat it with a stick. While my approach was not optimal, I did not starve.

Similarly, if your data do not meet the assumptions for using a Pearson Correlation Coefficient, you are not out of luck. There are other tools available to you for assessing whether or not relationships exist between two variables. These tools have different assumptions. I will not go into the details of how to compute these other kinds of correlation coefficients, but you should be aware that they exist and know that they are there to help you in those situations where you cannot use Pearson's r.

Spearman's Rank Order Correlation Coefficient (r_s)

When your data meet all of the assumptions for Pearson's r *except* that one or both of the variables are measured on an ordinal scale rather than on an interval or ratio scale, you can use Spearman's Rank Order Correlation Coefficient. Sometimes this coefficient is referred to as "Spearman's Rho". The symbol for it is r_s.

Biserial Correlation Coefficient (r_{bis})

Sometimes your data meets all of the assumptions for Pearson's r except that one of the variables has been "**artificially**" converted to a dichotomous variable. For example, if you take a multiple choice test item with four possible alternatives and "recode" it so that a person has either answered it correctly (1) or incorrectly (0). You have taken a variable that could conceivably be considered an "interval scale" with four points to a scale with only 2 points.

Another example would occur if you took the variable "total income" and recoded it into two groups such as "below the poverty line" (0) and "above the poverty line" (1). In this case, you have taken a ratio level variable and converted it into a dichotomous variable which only has two possible values.

In cases such as those just described, you would use a correlation coefficient called the Biserial Correlation Coefficient. The symbol for this correlation coefficient is r_{bis}.

Point Biseriel Correlation Coefficient (r_{pbis})

The Point Biserial Correlation Coefficient is very similar to the Biserial Correlation Coefficient. The big difference is related to the "dichotomous" variable.

Whereas in the Biserial Correlation Coefficient one of the variables is continuous while the other one *would be* if it had not been artificially made into a dichotomous variable, in the Point Biserial Correlation Coefficient, the dichotomous variable is "naturally" dichotomous. For example, gender is (for all practical purposes) truly dichotomous—there are only two choices.

When you have one variable that is continuous on an interval or ratio scale and the other is naturally dichotomous, the Point Biserial Correlation Coefficient is the best choice to use when it comes to measuring whether or not a relationship exists between the variables.

All Correlation Coefficients Are Interpreted In the Same Way

Whatever correlation coefficient you use, it is interpreted in generally the same way. The value of the correlation coefficient must be between -1.00 and +1.00; larger correlation coefficients mean stronger relationships; squaring the correlation coefficient tells you the amount of variation in one variable that is accounted for by the other variable.

Chapter Summary

1. A correlation coefficient is a single summary number that tells you:
   a. Whether there is a relationship between two variables.
   b. Whether the relationship positive or negative.
   c. How strong or weak the relationship is.
2. When calculating correlation coefficients by hand, it is best to lay the problem out in the form of a table that clearly reminds you what you need to calculate. This will help you avoid making mistakes.
3. The correlation coefficient is not a percentage. In other words it does not tell you what percent of one variable is accounted for by the other variable.
4. The Coefficient of Determination (r^2) tells you the percent of variance in one variable that is accounted for by the other variable.
5. Large correlation coefficients mean there is a strong relationship. The result is that r^2 will be large meaning that a larger proportion of variation in one variable is accounted for by variation in the other variable.
6. Stronger relationships will allow you to make more accurate predictions than weaker relationships.
7. In order to use Pearson's Correlation Coefficient to establish whether relationships exist between two variables, your data should meet certain assumptions. These are:
   a. Your variables are measured on an Interval or Ratio Scale.
   b. The traits you are measuring with your variables are normally distributed in the population (although Pearson's Correlation Coefficient will work well even if your data differ somewhat from normality).

   c. The relationship you are measuring is "linear" in nature meaning that it is best characterized by a straight line on a scatterplot.
   d. The relationship between your two variables is homoscedastic meaning that for each value on the X variable the values of the Y variable are roughly similar distance from the middle of the "cloud of dots".

8. If you calculate a Correlation Coefficient with data that do not meet the assumptions, you will get a number that looks like a correlation coefficient but it may not be accurate and so you cannot accurately interpret it.

Terms to Learn

You should be able to define the following terms based on what you have learned in this chapter.

Bisearial Correlation Coefficient
Coefficient of Determination
Correlation Coefficient
Curvilinear Relationship
Heteroscedasticity
Homoscedasticity
Linear Relationship
Point Bisearial Correlation Coefficient
Spearman's Rank Ordered Correlation Coefficient
The symbol "r"
Type I Error
