CHAPTER 6 GOALS

- Learn about the Pearson Product-Moment Correlation Coefficient (r)
- Learn about the uses and abuses of correlational designs
- Learn the essential elements of simple regression analysis
- Learn how to interpret the results of multiple regression
- Learn how to calculate and interpret Spearman’s r, Point-Biserial r, and the Phi correlation

Are stock prices related to the price of gold? Is unemployment related to inflation? Is the amount of money spent on research and development related to a company’s net worth? Correlation can answer these questions, and there is no statistical technique more useful or more abused than correlation.

Correlation is a statistical method that determines the degree of relationship between two different variables. It is also known as a “bivariate” statistic, with bi- meaning two and variate indicating variable or variance. The two variables are usually a pair of scores for a person or object. The relationship between any two variables are can vary from strong to weak or none. When a relationship is strong, this means that knowing a person's or object’s score on one variable helps to predict their score on the second variable. In other words, if a person has a high score of variable A (compared to all the other peoples’ scores on A, then they are likely to have a high score on variable B (compared to the other peoples’ scores on B). The latter would be considered a strong positive correlation. If the correlation or relationship between variable A and B is a weak one, then knowing a person's score on variable A does not help to predict their score on variable B. One very nice feature of the correlation coefficient is that it can only range from \(-1.00\) to \(+1.00\). Any values outside this range are invalid. Here is a graphic
representation of correlation’s range. Note that the correlation coefficient is represented in a sample by the value “r.”

\[\begin{array}{cccc}
1.00 & .50 & 0 & +.50 & +1.00 \\
\text{strong negative relationship} & \text{weak or none} & \text{strong positive relationship}
\end{array}\]

When the correlation coefficient approaches \( r = +1.00 \) (or greater than \( r = +.50 \)) it means there is a strong positive relationship or high degree of relationship between the two variables. This also means that the higher the score of a participant on one variable, the higher the score will be on the other variable. Also, if a participant scores very low on one variable then their score will also be low on the other variable. For example, there is a positive correlation between years of education and wealth. Overall, the greater the number of years of education a person has, the greater their wealth. A strong correlation between these two variables also means the lower the number of years of education, the lower the wealth of that person. If the correlation was perfect one (\( r = +1.00 \)), then there would be not a single exception in the entire sample to increasing years of education and increasing wealth. It would mean that there would be a perfect linear relationship between the two variables. However, perfect relationships do not exist between two variables in the real world of statistical sampling. Thus, a strong but not perfect relationship between education and wealth in the real world would mean that the relationship holds for most people in the sample but there are some exceptions. In other words, some highly educated people are not wealthy, and some uneducated people are wealthy.

When the correlation coefficient approaches \( r = -1.00 \) (or less than \( r = -.50 \)), it means that there is a
**strong negative relationship.** This means that the higher the score of a person on one variable, the lower the score will be on the other variable. For example, there might be a strong negative relationship between the value of gold and the Dow Jones Industrial Average. In other words, when the value of gold is high, the stock market will be lower and when the stock market is doing well, the value of gold will be lower.

A correlation coefficient that is close to $r = 0.00$ (note that the typical correlation coefficient is reported to two decimal places) means knowing a person's score on one variable tells you nothing about their score on the other variable. For example, there might be a zero correlation between the number of letters in a person's last name and the number of miles they drive per day. If you know the number of letters in a last name, it tells you nothing about how many miles they drive per day. There is no relationship between the two variables; therefore, there is a zero correlation.

It is also important to note that there are no hard rules about labeling the size of a correlation coefficient. Statisticians generally do not get excited about a correlation until it is greater than $r = 0.30$ or less than $r = -0.30$.

The correlational statistical technique usually accompanies correlational designs. In a correlational design, the experimenter typically has little or no control over the variables to be studied. The variables may be statistically analyzed long after they were initially produced or measured. Such data is called **archival.** The experimenter no longer has any experimental power to control the gathering of the data. The data has already been gathered, and the experimenter now has only statistical power in his or her control. Cronbach (1967), an American statistician, stated well the difference between the experimental and correlational techniques, “… the experimentalist [is] an expert puppeteer, able to keep untangled the strands to half-a-dozen independent variables. The correlational psychologist is a mere observer of a play where Nature pulls a thousand strings.”
One of the potential benefits of a correlational analysis is that *sometimes* a strong correlation between two variables may provide clues about possible cause-effect relationships. However, some statisticians claim a strong correlation *never* implies a cause-effect relationship. As much as correlational designs and statistical techniques are abused in this regard, I can understand the conservative statisticians’ concerns. I think that correlational designs and techniques may allow a researcher to develop ideas about potential cause-effect relationships between variables. At that point, the researcher may conduct a controlled experiment and determine whether their cause-effect hunch between two variables has some support. Indeed, after a controlled experiment, a researcher may claim a cause-effect relationship between two variables. Because correlational designs and techniques *may* yield clues for future controlled experimental investigations of cause-effect relationships, correlational designs and correlational statistical analyses are probably the most ubiquitous in all of statistics. Their mere frequency, therefore, may help to contribute to their continued abuse yet it is also something about their very nature that contributes to their misinterpretation.

**Correlation: Use and Abuse**

The crux of the nature and the problem with correlation is that, just because two variables are correlated, it does not mean that one variable *caused* the other. We mentioned earlier of a governor who wanted to supply every parent of a newborn child in his state with a classical CD or tape in order to boost the child’s IQ. The governor supported his decision by citing studies, which have shown a positive relationship between listening to classical music and intelligence. In fact, the controversy has grown to the point where it is referred to as the Mozart Effect. The governor is making at least two false assumptions. First, he is assuming a *causal relationship* between classical music and intelligence, that is, classical music causes intelligence to rise. However, the technique of correlation does not allow the implication of causation.
If $x$ and $y$ are correlated, then $x$ is related to $y$, and $y$ is related to $x$. Therefore, it may not be that classical music increases intelligence ($x$ is related to $y$) but maybe more intelligent people listen to classical music ($y$ is related to $x$). Correlation does not distinguish nor give us any guidance whatsoever about when $x$ is correlated with $y$ whether it is $x$ is related to $y$ or whether $y$ is related to $x$. Second, the governor is making the mistake not only of basing his decision on a few correlational studies (when he should wait for evidence from experiments) but also he has not waited for scientific replication. It is far too early to assume that classical music raises people’s intelligence based on a few correlational studies. There actually has been an experimental study of the effect, conducted by the same experimenter who claimed the effect in the first place. Let’s apply some of the principles of Sagan’s Baloney Detection Kit. Have the claims been verified by another source? At this point, the effect has received little support by researchers other than those who first claimed it. How does this claim fit in the world, as we know it? It does not fit very well. It would require new and undiscovered brain mechanisms. Does it seem too good to be true? Yes! Wouldn’t it be wonderful if just playing a Mozart CD boosted every babies’ IQ? Of course it would, but with such a wonderful claim as is, we must be very skeptical. We must ask for even higher standards of research excellence. Scientists must always be cautious. Findings must be replicated through experiments in a variety of settings with a variety of people by a variety of different researchers. Interestingly, the Mozart Effect may be another uncommon example of the benign result of rejecting a true null hypothesis. What are the consequences of a Type One error in this case? Parents are exposing their children to classical music when it really doesn’t boost their children IQ’s. I also know some highly educated and highly skeptical parents and grandparents who buy their children and grandchildren classical music toys that are marketed directly because of the probably unreal Mozart Effect. These parents and grandparents are fully aware there is little probability that the Mozart Effect is real but it is a high-risk but low cost and benign consequence situation. If there’s even a 1 in 10,000
chance that the Mozart Effect is real, the musical toys do not cost that much (because some kind of toy will be purchased anyway) and exposure to classical music is at the very very worst, harmless. A psychologist, McBurney (1996), suggests one way to counter believing in things we like to believe is to ask ourselves what the consequences would be if it were really true. For example, shouldn’t everyone be listening to classical music? Wouldn’t there be laws against playing any other kind of music in nurseries, kindergartens, and schools? Wouldn’t we force our own children to listen to classical music? Or do we want our children to end up dumber than the kids next door?

**A Warning: Correlation Does Not Imply Causation**

A major caution must be reiterated. Correlation does not imply causation. Because there is a strong positive or strong negative correlation between two variables, this *does not* mean that one variable is caused by the other variable. As noted previously, many statisticians claim that a strong correlation never implies a cause-effect relationship between two variables. Yet, there are daily published abuses of the correlational design and statistical technique, not only in newspapers but major scientific journals!

A sampling of these misinterpretations follows:

1. **Marijuana use and heroin use are positively correlated.** Some drug opponents note that heroin use is frequently correlated with marijuana use. Therefore, they reason that stopping marijuana use will stop the use of heroin. Clear-thinking statisticians note that even a higher correlation is obtained between the drinking of milk in childhood and later adult heroin use. Thus, it is just as absurd to think that if early milk use is banned, subsequent heroin use will be curbed, as it is to suppose that banning marijuana will stop heroin abuse.

2. **Milk use is positively correlated to cancer rates.** While this is not a popular finding within the milk industry, there is a moderately positive correlation with drinking milk and getting cancer (Paulos, 1990). Could drinking milk cause cancer? Probably not. However, milk consumption
is greater in wealthier countries. In wealthier countries people live longer. Greater longevity means people live long enough to eventually get some type of cancer. Thus, milk and cancer are correlated but drinking milk does not cause cancer (nor does getting cancer cause one to drink more milk).

3. **Weekly church attendance is negatively correlated with drug abuse.** A recent study demonstrated that adolescents who attended church weekly were much less likely to abuse illegal drugs or alcohol. Does weekly church attendance cause a decrease in drug abuse? If the Federal Government passed a law for mandatory church attendance, would there be a decrease in drug abuse? Probably not, and there might even be an increase in drug abuse. This study is another example of the abuse of the interpretation of the correlational design and statistical analysis.

4. **Lead Levels are positively correlated to antisocial behavior.** A 1996 correlational study (Needleman et al.) examined the relationship of lead levels in the bones of 301 children and found that higher levels of lead were associated with higher rates of antisocial behavior. “This is the first rigorous study to demonstrate a significant association between lead and antisocial behavior,” said one environmental health professor about the study. While the studies authors may have been very excited, the study is still correlational in design and analysis, thus, implications of causation should have been avoided. Perhaps antisocial children have a unique metabolic chemistry such that their bodies do not metabolize lead like normal children. Perhaps lead is not a cause of antisocial behavior but the result of being antisocial. Therefore, the reduction of lead exposure in early childhood may not reduce antisocial behavior at all. Also, note that there was a statistically “significant” relationship. As you will learn later in this chapter, with large samples (like 301 children in this study) even very weak relationships can be statistically significant with correlational techniques.
5. **The risk of getting Alzheimer’s Dementia is negatively correlated with smoking cigarettes.**

In studies funded by the tobacco industry, it was found that the risk of getting Alzheimer’s dementia was negatively correlated with smoking cigarettes (yes, the risk went down with an increased use of cigarettes!). The implication of these findings for the tobacco industry was that increases in smoking (probably from increases in nicotine) would prevent the onset of Alzheimer’s Dementia. This serves as a good example of the error in assuming a causal relationship because a correlation exists between two variables, and a demonstration of how a third variable may be controlling the other two variables. The risk of getting Alzheimer’s Dementia increases with longevity. While about 10% of people over 65 are diagnosed with Alzheimer’s, nearly 50% of those over age 90 are diagnosed with Alzheimer’s. Heavy smokers die at a rate of about 500,000 a year. Smokers do not get the chance to get Alzheimer’s Dementia because they do not live long enough.

6. **Sexual activity is negatively correlated with increases in education.** A 1997 report based on 10,000 interviews found that those with less education reported more sexual contacts per year than those who had been to post-graduate schools (PhD programs, law school, medical school, etc.). Again, this is another example of a correlational design with a correlational analysis, therefore, causation cannot be inferred in these findings. And once again, age may be the mitigating factor in this study. People with less than a baccalaureate degree tend to be younger than those with post-graduate degrees. Younger people tend to be more sexually active than older people. The study did not control for the age of the participants.

7. **An active sex life is positively correlated with longevity.** The 1997 newspaper headline for this study published in a British medical journal read, “Study suggests frequent sex equals long life.” The study was conducted on a sample of 918 Welsh men divided into three groups: sex
twice or more a week, an intermediate group, and those who had sex less than once a month. In a ten-year follow-up, the sexually inactive group had the highest death rate. The authors said the results could not be attributed to age or health. The authors, however, did not control for psychological variables like depression, which could have been the precursor of physical disease. Thus, depression may have lowered sexual interest and subsequent physical disease may have accounted for the increase in death rates. Imagine doctors prescribing or ordering their patients to have frequent sex in order to increase longevity. It is quite possible we might witness a sudden increase in death rates. In order to change this correlational design to an experiment, the authors should have randomly assigned the men to one of the three sexually frequency groups, ordered the men to have sex according to the group they were assigned, and then assessed their longevity ten years later. Of course, this experimental design is not feasible but any causal inferences from the original correlational design are equally unfeasible.

8. **Coffee drinking is negatively correlated with suicidal risk.** In this 1996 study published in an American medical journal, 86,626 registered nurses were evaluated over a 10-year period. Increased coffee drinking was associated with lower rates of suicide. The authors concluded that the caffeine in the coffee might enhance mood and well being resulting in lowered suicide risk. While the number of nurses studied is impressive, once again this study is a correlational design and causation cannot be inferred. A stronger argument might have been made for the positive mood hypothesis had the authors randomly assigned depressed patients to groups, then prescribed various levels of caffeine. If the groups receiving the highest levels of caffeine had a subsequently higher sense of well being and lowered suicide rates, then the hypothesis might be more plausible.

Another Warning: Chance is Lumpy
The contemporary Yale statistician Robert Abelson (1995) postulated Abelson’s First Law of Statistics: **Chance is lumpy**. By this, he means that if 86,626 nurses are measured for hundreds of variables over a ten year period, it would be highly surprising if dozens of significant relationships were not found! Or as my brother would say as he defends another one of his get-rich-quick schemes, “Throw enough mud up on a wall, and some of it’s bound to stick.” What Abelson is concerned with is that people fail to appreciate that long runs of occurrences can often be attributed to pure chance or random processes. Often people will attribute some unusual finding or run of luck to a mysterious or a systematic process when, in fact, only chance is operating. As Abelson notes, “attributing a data set to mere chance is deflating.”

Abelson also notes that there is also the psychological tendency to minimize the great variability that exists in small samples. Thus, we tend to overestimate the generalizability of small samples, when in reality, our conclusions may have varied widely across other samples.

One solution to the chance-is-lumpy problem is to try to replicate our findings. If this study is our first, then we should keep in mind that we should be somewhat conservative in our initial conclusions and be mindful of the repercussions of our findings and equally mindful of the tricks that can occur when we measure hundreds of variables in huge samples or a few variables in small samples.

**Correlation and Prediction**

The correlation coefficient may also be used as an indicator of prediction. If a strong positive or negative correlation is obtained, then the relationship between the two variables may be likened to a predictive relationship. For example, in the previous strong negative relationship between gold prices and the stock market, it would be suspected that stronger gold prices would mean a weakening stock market. It could be said that the price of gold predicts the stock market and vice versa. A strong stock market would predict weakening gold prices. If there was no relationship between the variables, or the
correlation coefficient is close to or equal to zero, then no predictions can be made with any reliability.

The Four Common Types of Correlation

1. **Pearson's r**: A measure of the strength of a relationship between two continuous variables.

2. **Spearman’s r**: A measure of the similarity between two ordinal rankings of a single set of data.

3. **Point-Biserial r**: A measure of the strength of a relationship between one continuous variable and one dichotomous variable (a two-level-only variable like gender).

4. **Phi (φ) Correlation**: A measure of the strength of a relationship between two dichotomous variables.

The Pearson Product-Moment Correlation Coefficient

The single most common type of correlation is the **Pearson Product-Moment Correlation Coefficient**, which measures the degree of relationship between two continuous variables. A continuous variable is a variable, which can be measured along a line scale. For example, gold prices are continuous because they can range along a line scale from about $300 to about $500 an ounce. Gender (male or female) is not considered a continuous variable because if numbers (like 1 or 2) were assigned to the two categories, a person could not be a 1.3 or a 1.7. The story of the naming of the correlation coefficient appears at the end of this chapter in the History Trivia section.

**Pearson’s coefficient r** is obtained for a sample drawn from a population. The **population value of Pearson’s coefficient is called rho (ρ)**, and thus, r is an estimate of ρ.

The formula for r is as follows:

\[
 r = \frac{N \sum xy - (\sum x)(\sum y)}{\sqrt{N \sum x^2 - (\sum x)^2} \sqrt{N \sum y^2 - (\sum y)^2}}
\]

Note: In this formula, N is equal to the number of pairs of scores and \( \sum xy \) is called the sum of the cross
products. Let’s see how the formula works in the following example.

A tobacco company statistician wishes to know whether heavy smoking is related to longevity. From a sample of recently deceased smokers, the number of cigarettes (estimated on a per day for their last five years after visits with their surviving relatives) is paired with the number of years that they lived.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Cigarettes</th>
<th>Years Lived</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25</td>
<td>63</td>
</tr>
<tr>
<td>2</td>
<td>35</td>
<td>68</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>72</td>
</tr>
<tr>
<td>4</td>
<td>40</td>
<td>62</td>
</tr>
<tr>
<td>5</td>
<td>85</td>
<td>65</td>
</tr>
<tr>
<td>6</td>
<td>75</td>
<td>46</td>
</tr>
<tr>
<td>7</td>
<td>60</td>
<td>51</td>
</tr>
<tr>
<td>8</td>
<td>45</td>
<td>60</td>
</tr>
<tr>
<td>9</td>
<td>50</td>
<td>55</td>
</tr>
</tbody>
</table>

We will arbitrarily name one variable \( x \) and the other variable \( y \). The results of Pearson’s \( r \) will be exactly same no matter which variable is labeled \( x \) or \( y \).

**Step 1.** First, obtain \( \Sigma x \), \( \Sigma x^2 \), \( (\Sigma x)^2 \), and \( (\Sigma y)^2 \)

\[
\Sigma x = 25 + 35 + 10 + 40 + 85 + 75 + 60 + 45 + 50 = 425
\]

\[
\Sigma x^2 = 25^2 + 35^2 + 10^2 + 40^2 + 85^2 + 75^2 + 60^2 + 45^2 + 50^2 = 24,525
\]

\[
\Sigma y = 63 + 68 + 72 + 62 + 65 + 46 + 51 + 60 + 55 = 518
\]

\[
\Sigma y^2 = 63^2 + 68^2 + 72^2 + 62^2 + 65^2 + 46^2 + 51^2 + 60^2 + 55^2 = 33,188
\]

\[
(\Sigma x)^2 = HINT: \text{ square } \Sigma x \text{ or } (25 + 35 + 10 + 40 + 85 + 75 + 60 + 45 + 50)^2
\]

\[
(\Sigma x)^2 = 425^2 = 180,625
\]

\[
(\Sigma y)^2 = HINT: \text{ square } \Sigma y \text{ or } (63 + 68 + 72 + 62 + 65 + 46 + 51 + 60 + 55)^2
\]

\[
(\Sigma y)^2 = 542^2 = 293,764
\]
Step 2. Obtain the sum of the cross products ($\Sigma xy$) by multiplying each x score by its paired y score.

$$\Sigma xy = (25 \times 63) + (35 \times 68) + (10 \times 72) + (40 \times 62) + (85 \times 65) + (75 \times 51) + (45 \times 60) + (50 \times 55)$$

$$\Sigma xy = 1575 + 2380 + \ldots + 2750$$

$$\Sigma xy = 24,640$$

Step 3. Obtain the value of the numerator in the $r$ formula. Remember, $N$ is equal to the number of pairs of scores. In this example, there are 9 pairs of scores.

$$N\Sigma xy - (\Sigma x)(\Sigma y)$$

$$= (9) (24,640) - (425) (518)$$

$$= 221,760 - 230,350 = -8,590$$

Step 4. Obtain the value of the denominator in the $r$ formula. Remember that the square root is obtained after all other denominator values have been computed, simplified, and reduced to a single number.

$$\sqrt{\left[N\Sigma x^2 - (\Sigma x)^2\right] \left[N\Sigma y^2 - (\Sigma y)^2\right]}$$

$$\sqrt{\left[(9)(24,525) - (425)^2\right] \left[(9)(33,188) - (542)^2\right]}$$

$$= \sqrt{220,725 - 180,625 \left[298,692 - 293,764\right]}$$

$$= \sqrt{40,100 \left[4,928\right]}$$

$$= 14,057.482$$

Step 5. Divide the result of Step 3 by the result of Step 4.
Note that the Pearson $r$ is usually rounded off to two decimal places. Thus, $r = -.61$ means that there is a strong negative correlation between smoking and longevity. This indicates that the higher the number of cigarettes smoked in the past five years, the lower the number of years lived. And the lower the number of cigarettes, the higher the number of years lived. Remember, this relationship between these two variables DOES NOT mean that heavy smoking causes one to live less. It may, however, give clues as to further research ideas for experiments. In this case, an experiment might be set up (perhaps with animals) with an experimental group and a control group to determine whether cigarette smoking actually has a causal relationship with early morbidity.

**Testing for the Significance of a Correlation Coefficient**

A correlation coefficient may be tested to determine whether the coefficient significantly differs from zero. The value $r$ is obtained on a sample. The value rho ($\rho$) is the population's correlation coefficient. It is hoped that $r$ closely approximates rho. The null and alternative hypotheses are as follows:

- Ho: $\rho = 0$
- Ha: $\rho \neq 0$

The value of $r$ and the number of pairs of scores are converted through a formula into a distribution (similar to the $z$ distribution) called the $t$ distribution (in Appendix B, page 276 of the required text, *Statistics: A Gentle Introduction*). The $t$ formula can only be used to test whether $r$ is equal to zero. It cannot be used to test to see whether $r$ might be equal to some number other than zero. It is also
important to note that the \( t \) distribution may be used to test other types of inferential statistics. Therefore, if someone says that a \( t \) test is being used, it would be a legitimate question to ask “why?” The \( t \) distribution is most commonly used to test whether two means are significantly different, however, it may also be used to test the significance of the correlation coefficient. The \( t \) distribution also has other uses. Interestingly, the \( t \) distribution becomes \( z \) distribution when the data is infinite but they are also strikingly visually similar when there are only several hundred numbers in the set of data.

The \( t \) test formula in order to test the null hypothesis for a correlation coefficient is:

\[
t = \frac{r}{\sqrt{1 - r^2}} \sqrt{N - 2}
\]

In the previous example on smoking, the research question was whether heavy smoking was related to longevity. A \( r = -.61 \) was obtained which meant that there was a strong negative relationship between smoking and longevity. In order to test whether this obtained \( r \) significantly differs from zero, the \( t \) formula is used.

\[
t = \frac{r}{\sqrt{1 - r^2}} \sqrt{N - 2}
\]

where \( N = \) the number of pairs of scores.

\[
t = \frac{-.6111}{\sqrt{1 - (.6111)^2}} \sqrt{9 - 2}
\]

\[
t = \frac{-.6111}{\sqrt{1 - .3734}}
\]

\[
t = -.6111
\]

\[
1 - .3734
\]
Obtaining the Critical Values of the $t$ Distribution

We will now obtain the critical values of $t$, which is obtained from the $t$ distribution in Appendix B.

**Step 1. Choose a one-tailed or two-tailed test of significance.** The alternative hypothesis establishes whether we will use a one-tailed or two-tailed significance test. If the alternative hypothesis is non-directional, as is the case in most studies, then a two-tailed test of significance is required.

**Step 2. Choose the level of significance.** The conventional level of significance is $p = .05$. Only in rare circumstances would one ever depart from $p = .05$ as a starting point.

**Step 3. Determine the degrees of freedom ($df$).** The $df$ is an advanced statistical concept related to sampling. We will keep things simple: The formula for this $t$ test statistic is $df = N - 2$ where $N$ is the number of pairs of scores. In this example, there were 9 pairs of scores so $df = 9 - 2$ or $df = 7$.

**Step 4. Determine whether the $t$ from the formula (called the derived $t$) exceeds the tabled critical values from the $t$ distribution.** For a two-tailed test of significance at $p = .05$ with $df = 7$, the critical values of $t$ are $t = +2.365$ and $t = -2.365$. If the derived $t$ is greater than $t = +2.365$ or less than $t$
= -2.365, then the null hypothesis will be rejected. In this example, the derived \( t = -2.042 \) is not less than \( t = -2.365 \), therefore, the null hypothesis is not rejected, and it will be concluded that \( r = -0.61 \) indicates a nonsignificant relationship. Curiously, although the strength of the relationship was strong (\( r = -0.61 \)), the test of significance indicated that the obtained relationship was likely due to chance or there was greater than 5 chances out of 100 than the relationship was due to chance.

In a research paper, the results might be reported as follows:

"There was a strong negative correlation found between smoking and longevity although the correlation was not statistically significant, \( r(7) = -0.61, p > .05 \). Note that the degrees of freedom appear in the parentheses to the right of \( r \).

**If the Null Hypothesis is Rejected**

Remember also that if the null hypothesis is rejected, the experimenter would report the lowest \( p \) level possible. In that case, the derived \( t \) would be compared to the critical values of \( t \) at .01 and .001. If the derived \( t \) exceeded the critical values at both .01 and .001, then the experimenter would report the \( r \) as significant at \( p < .001 \).

**Representing the Pearson Correlation Graphically: The Scatterplot**

A scatterplot is a graphic presentation of the pairs of scores involved in a correlation coefficient. It is also sometimes called a bivariate distribution. Let us produce a scatterplot of the example on page 161 of the number of cigarettes smoked and longevity. In order to construct a scatterplot, prepare to graph each of the variables along one of the graph's axes. Ultimately, it does not matter which variable is chosen for which axis, just as long as you prepare each axis for one of the variables (see Figure 6.1).

The order in which the pairs are graphed is not important. Thus, it does not matter if you begin with the last pair of scores or the first pair of scores. However, each pair of scores is very important, so be sure to plot each participant’s score on one variable with their corresponding score on the other variable.
For example, if we began with the participant who smoked an average of 85 cigarettes per day and lived 65 years, we would first locate the 85 cigarettes value on the cigarette axis and draw a horizontal line across the graph along this value. Next, locate this participant’s longevity score on the longevity axis and draw a line vertically up the graph. The intersection of the two lines is the graphic representation of the pair of values in mathematical two-dimensional space. Figure 6.2 shows a completed scatterplot.
Fitting the Points with a Straight Line: The Assumption of a Linear Relationship
When using the Pearson correlation coefficient, it is assumed that the cluster of points is best fit by a straight line. Look at the cluster in Figure 6.2, and imagine how a straight line would pass through the points with as little overall distance between all the points and the straight line as possible. Obviously, there is no single straight line that would pass through all the points. There is only a best-fitting line that minimizes all of the distances between the points and the line. If a straight line, as opposed to a curved line, best fits the points then the relationship between the two variables is said to be linear. If a curved line best fits the points, then the relationship between the two variables is said to be curvilinear. Remember, it is an assumption of the correlation coefficient that the best fitting line is linear, or in other words, it is assumed that the relationship between the two variables is linear. The violation of this assumption is typically not harmful, at least in terms of committing a Type I error. If we assume a relationship is linear when it is really curvilinear, it will result in a lower $r$ value, and statistical significance is less likely to be attained. Of course, it could conceivably be harmful in subtle ways to the experimenter if no relationship was found where one actually exists (Type II error). Figure 6.3 presents a curvilinear relationship. Curvilinear relationships are not uncommon. For example, there is a curvilinear relationship between strength and age: When we are younger, we are weaker; when we are older, we are stronger; but when we become very old, we are weaker again.
Interpretation of the Slope of the Best-Fitting Line

The best-fitting line may also be used to interpret the nature and strength of the relationship between two variables. In the previous example, the best-fitting line slopes from the upper left of the graph to the lower right. This indicates that there is a negative correlation between the two variables. The relatively small amount of overall distance between the points and the line indicates that this negative relationship is also strong; that is, it will be a large negative number.

A positive relationship will produce a best-fitting line that slopes from the lower left of the graph to the upper right. A weak or no relationship will produce a seemingly random cluster of points. There will be no best-fitting line, or it could be said that a straight horizontal line through the cluster is as good as good as any.

A perfect correlation ($r = 1.00$) would produce a scatterplot where the best-fitting straight line passes through all of the points, and thus, it is an extremely unlikely event in the real world of data. See Figure 6.4 for a graphic representation of positive, negative, weak, and perfect correlations.
Figure 6.4

Perfect Positive Correlation: \( r=1.00 \)

Perfect Negative Correlation: \( r=-1.00 \)

Figure 6.4 (continued overleaf)
Positive Correlation: \( r = 0.63 \)

Negative Correlation: \( r = -0.61 \)

Weak Correlation: \( r = -0.20 \)

Figure 6.4 (continued)
The Assumption of Homoscedasticity

A second assumption of the correlation coefficient is that of homoscedasticity. This assumption is met if the distance from the points to the line is relatively equal all along the line. The violation of the assumption is called heteroscedasticity, and a graphic representation is presented in Figure 6.5.

Figure 6.5

![Weak Correlation: r = -.20](image)

The effects of violating the assumption of homoscedasticity are the same as violating the assumption of linearity, and that is, the value of \( r \) is more likely to underestimate the population value of \( r \).

The Coefficient of Determination: How Much One Variable Accounts for Variation in Another Variable: The Interpretation of \( r^2 \)

Another use of the correlation coefficient is its squared value. \( r^2 \) is called the coefficient of determination, and it has two important interpretations. First, it explains the proportion of variance in one variable accounted for by the other variable. For example, if the correlation between two variables A and B is \( r = .25 \), then \( r^2 = (.25)(.25) = .0625 \). Therefore, the variable A explains approximately 6% of
the variation in variable B. Another way of looking at $r$ in this example, would be to say that 6% of the variation in variable B can be explained by its relationship to variable A, and 94% of the total variance between variables A and B remains unexplained. The former variance is called shared variance and the latter variance is called uncommon, unshared, or unexplained variance.

A second interpretation of $r^2$ is its use as a measure of strength between two or more $r$ values. For example, if given an $r = .25$ and $r = .50$, it is clear that the second $r$-value is twice as great as the first $r$ value. However, the $r^2$ values are 6.25% and 25% respectively. Thus, $r = .50$ actually explains four times as much variance as does an $r = .25$. The moral here is to be careful when interpreting and comparing $r$ values, particularly smaller values of $r$ where any $r < .31$ will explain less than 10% of the variance in two variables.

The same causality caution that was applied to the interpretation of the simple correlation coefficient $r$ is applied to the interpretation of $r^2$. The coefficient of determination does not provide an explanation for the observed relationship or nor does it account for the reason for the relationship. The coefficient of determination simply provides another way of viewing the relationship between two variables.

Quirks in the Interpretation of Significant and Nonsignificant Correlation Coefficients

There are some serious quirks in the interpretations of the correlation coefficients. We have witnessed one already in our example on smoking and longevity. The interpretation of the strength of the relationship between these two variables was actually independent of the significance testing. The derived $r = -.61$ indicated that there was a moderately strong negative correlation between smoking and longevity. However, we found that this relationship was not statistically significant at $p < .05$. The statistical quirk is that significance tests are artifactually dependent upon sample size: Larger sample sizes will more likely produce significance than smaller samples. In our example, $N = 9$ was an exceptionally small sample, thus, despite a moderately strong $r$-value, we were not able to reject the null
hypothesis. If we had had 30 pairs of scores, we would have obtained significance.

At this point, the coefficient of determination helps us to interpret this sample size quirk. When we square \( r = -0.61 \), we obtain \( 0.37 \) or 37% of the variance is shared by the two variables. This is not a small amount of variance, all things being equal. In a 1988 prejudice study, it was found that \( r = 0.06 \) was significant for the relationship between anti-Semitism and prejudice against smoking. How could the latter \( r \) be statistically significant when our \( r = -0.61 \) was not? In the prejudice study there were 5,977 pairs of scores. When we square \( r = 0.06 \), we obtain \( r^2 = 0.0036 \) or about one-third of one percent of the total variance is explained variance! This also indicates that well over 99% of the variance remains unexplained between the two variables. In these two examples, the coefficient of determination has helped clear away some of the confusion in the interpretation of significance or a lack of significance.

**Linear Regression**

Now that you have learned many of practical aspects of the Pearson product-moment correlation coefficient, let us delve into some other theoretical aspects. It can be said that in correlation to this point, we have used the \( x \) and \( y \) variables symmetrically, that is, the correlation between \( x \) and \( y \) was the same as the correlation between \( y \) and \( x \). Now, let us consider them asymmetrically, where the \( Y \) variable will be called the dependent variable and \( X \) will be labeled the independent variable. Our interest will be in seeing how various values of the independent variable predict corresponding values in the dependent variable. The dependent variable could also be called the response variable, and the independent variable could also be called the explanatory variable. This statistical technique is called **regression analysis**, and it is probably the most common statistical technique in business and economics. Because virtually no one calculates them by hand any longer, we will focus on its uses, meaning, and interpretation.

Regression analysis deals with the way one variable changes (\( Y \)) based on how one or more other
variables change (X₁ and X₂ etc.). If we are interested in how gold prices (Y) might vary as a function of the price of a barrel of oil (X₁), then this kind of relationship is called **simple regression** or **simple linear regression**. If we are interested in how gold prices (Y) vary as a function of the price of a barrel of oil (X₁), the Dow Jones Industrial Average (X₂), and, perhaps, additional variables then the investigation of these relationships is called **multiple regression analysis**.

In simple linear regression, a regression equation is used to plot a straight line through the middle of the scatter plot. The formula is:

\[ Y = a + bX \]

Where Y is the dependent variable or the value we are trying to predict:

- a is the Y intercept or the point at which the straight line crosses the ordinate or Y axis when X is 0
- b defines the slope or the angle of the straight line

The regression equation attempts to choose among an infinite number of straight lines to produce the single best-fitting line that predicts a Y score given an X score. The Least Squares Method is used to produce the best-fitting line. This method involves measuring the square of the distance from each point to a potential best-fitting line and then choosing the line, which produces the smallest value or “least squares.” The values are squared to control for the positive and negative distances that result from points above and below the best-fitting line. If these values were added without squaring them then the positive and negative distances would cancel each other out. By squaring the distances, negative distances are turned into positive values. The interest is not in the real distance of points to the best-fitting line but in a measure that estimates the best-fitting line.

**Reading the Regression Line**

In regression, we use X scores to predict Y scores. If there is a correlation between X and Y (if there
was no correlation between X and Y then X scores could not predict Y scores), then each value of X predicts a different value of Y. In order to find a specific value of Y given a specific value of X, find the X value on the horizontal axis (abscissa) and draw a line parallel to the vertical Axis (ordinate). When that straight line meets up with the regression line, another line is drawn at a right angle (parallel to the X axis) until it meets the vertical or Y axis. The value at which this line intersects the Y axis is the predicted value of Y given that value of X. See Figure 6.6.

Figure 6.6

The slope of the regression line indicates how many units the line rises (positive slope) or falls
(negative slope) on the Y axis for every unit moved to the right on the X axis. For example, if \( b = .75 \), then the regression line would rise .75 units of Y with every successive unit of X. If the regression line had a slope of +3 between gold prices and oil prices then a $1 increase in oil prices would result in a $3 increase in gold prices. Also, a positive slope means that there is a positive correlation, and a negative slope means there is a negative correlation. If there is no correlation, then the regression line runs parallel to the X axis (a flat line) and the slope is 0. Interestingly, when \( r = 0 \), the slope of the line \( b = 0 \), and the regression line intersects the Y axis at the mean value of Y for all values of X.

The world, however, is complex. It is much more common that single variable or score is dependent on multiple independent variables, thus, multiple regression analysis will often be more useful. Let’s see how multiple regression analysis can be applied to currency trading. A currency trader wishes to know what variables affect the value of the Indian rupee of the period of 10 months. The trader suspects that the rupee’s value against the US dollar) may vary as a function of the value of the Japanese yen (against the US dollar), the price of gold (per ounce) and the price of silver (per ounce). I am going to use statistical software Statistical Packages for the Social Sciences. I have entered each month’s data for the four variables on each row (rupee followed by yen, gold, and silver). There are ten rows of data, one for each month. In the Analyze menu, I chose Regression. In the submenu for Regression, I chose Linear. An options menu appeared, and I entered the rupee as my dependent variable, and then I entered yen, gold, and silver as the independent variables. I accepted the default method of regression, which is enter. As a point of information, there are many types of linear regression including enter, backward, forward, hierarchical, etc., and interestingly, most types yield similar results. Consult a multivariate statistics book or a book specifically on regression analyses to learn further about the differences between these various procedures.

The following is the initial part of the SPSS output.
This table explains three things. First, under variables entered, it is telling us that the three independent variables are silver, gold, and yen. Second, it is telling us that the dependent variable is the Indian rupee. Third, it tells which method of multiple regression was used, and it was the default, the enter method. This next table in the printout presents three of the most important aspects of the multiple regression analysis, R, R Square, and Adjusted R Square.

**R:** The coefficient R in this model summary can be interpreted just as we interpreted a Pearson product-moment correlation coefficient. In this case, \( R = .873 \) indicates that there is a very strong positive correlation between the dependent variable (Indian rupee) and the set of three independent variables (silver, gold, and yen). In general, statisticians view any values of R over .30 as reflecting the beginnings of a meaningful relationship between the dependent and independent variables.

**R Square:** The R Square value represents the percentage of variance accounted for in the dependent variable (Indian rupee) by the set of three independent variables (silver, gold, and yen). In other words, approximately 76% of the variance or changes in the value of the rupee can be accounted for by the confluence of the three independent variables.

**Adjusted R Square:** The Adjusted R Square value also represents the percentage of variance accounted for in the dependent variable by the set of three independent variables. However, the Adjusted R Square value is always slightly lower than the R Square value. This lowering of the estimate is similar to a penalty paid for having more than one independent variable. Typically, the larger the
number of independent variables, the lower the Adjusted R Square value will be than the R Square value. As in any distribution of samples, chance fluctuations may be larger with smaller sample sizes. There is a tendency then for R Square to be overestimated, so the Adjusted R Square is the adjustment made for this expected inflation in a sample.

<table>
<thead>
<tr>
<th>Model Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>1</td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), SILVER, GOLD, YEN

The next table in the output is labeled ANOVA, which stands for analysis of variance. This ANOVA tells us whether the three main regression coefficients were significantly different than zero.

Of primary interest is the F value and its significance level. If the three main regression coefficients were close to zero, then the F value would be approximately 1.00. In order to be statistically significant, the significance level must be less than the conventional level of statistical significance (i.e., .05). In the present case, the obtained F = 6.434 is statistically significant. Statistical convention dictates that one reports the lowest significance level possible. In a write-up of the present results so far, it would appear like this: the regression equation indicated that silver, gold, and yen were significantly related to the value of the Indian rupee, R = .87, R Square = .76, and Adjusted R Square = .64, F(3, 6) = 6.43, p < .03.

<table>
<thead>
<tr>
<th>ANOVA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), SILVER, GOLD, YEN
b. Dependent Variable: RUPEE
Now that we have established the significance of the regression equation, we need to examine the strength of the individual independent variables in the prediction of the dependent variable. The following is the final section of the multiple regression output.

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(Constant)</td>
<td></td>
<td>.337</td>
<td>.747</td>
</tr>
<tr>
<td></td>
<td>YEN</td>
<td>2.247E-03</td>
<td>.405</td>
<td>1.776</td>
</tr>
<tr>
<td></td>
<td>GOLD</td>
<td>2.470E-02</td>
<td>.625</td>
<td>2.766</td>
</tr>
<tr>
<td></td>
<td>SILVER</td>
<td>1.411</td>
<td>.254</td>
<td>1.244</td>
</tr>
</tbody>
</table>

a. Dependent Variable: RUPEE
The B column under Unstandardized Coefficients gives the actual value that would be multiplied against the specific values of the independent variables in order to make a prediction of the dependent variable. However, the B weights are not very useful in understanding the relative importance of the independent variables. These weights will be more interpretable if the dependent and independent variables have been standardized into Z scores with a mean of 0.00 and a standard deviation of 1.00. These standardized values appear in the Beta column under Standardized Coefficients. The Beta weights can be read like typical correlation coefficients. It can be seen that gold’s Beta weight of .63 appears to be the strongest of the three in the prediction of the Indian rupee. Next, examine the significance of the t values to see whether gold’s Beta weight is significantly different than zero. For gold, the t = 2.766, and the significance level for gold’s Beta weight is .033, which is less than the conventional level of significance (.05). Thus, it can be concluded that gold is a strong and significant predictor of the value of the Indian rupee. Interestingly, the yen Beta value (.405) represents a moderate relationship with the Indian rupee; however, it was not statistically significant (p = .126). The Beta value for silver (.254) would be interpreted as weak, and it was also not statistically significant (p = .260).

Final Thoughts About Regression Analyses

As noted earlier, regression analyses are common in business, finance, and economics. They can also become highly complex, and their interpretation can sometimes be controversial. If you wish to learn more about these procedures, consult an advanced statistical text. I would recommend Using Multivariate Statistics, 4th Edition, by Tabachnick and Fidell, published by Harper Collins.

Spearman's Correlation

Spearman's correlation coefficient (sometimes referred to as Spearman's rho or r_s where the sub s is in honor of Spearman) determines the degree of relationship for ranked data. Spearman's correlation
is also called the rank-order correlation coefficient. Although Spearman's correlation is far less common than Pearson's $r$, occasionally variables are ordered according to rank (like 1st through 10th), or variables may be subsequently ranked on the basis of a continuous variable. The formula for Spearman's $r$:

$$r_s = 1 - \frac{6 \sum D^2}{N(N^2 - 1)}$$

where $N$ = the number of pairs of scores.

Spearman's might be most typically used in situations where there are a number of variables and they are all ranked by two independent judges. For example, a grocery store owner wishes to know whether married couples have similar tastes in vegetables. The members of the couple were independently asked to rate their preference for seven vegetables from most preferred (#1 rank) to least preferred (#7). Their data is as follows:

<table>
<thead>
<tr>
<th></th>
<th>Husband</th>
<th>Wife</th>
<th>D score</th>
<th>$D^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Broccoli</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Cauliflower</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Brussel Sprouts</td>
<td>6</td>
<td>7</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>Okra</td>
<td>1</td>
<td>2</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>Cabbage</td>
<td>5</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Spinach</td>
<td>2</td>
<td>4</td>
<td>-2</td>
<td>4</td>
</tr>
<tr>
<td>Turnips</td>
<td>7</td>
<td>6</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

$\sum D^2 = 12$

Note that the $D$ score is the difference between the pairs of ranks on the first variable ranked, etc.

The number "6" in the formula is a **constant** and remains "6" regardless of the numbers of ranked
variables. \( N \) is the number of pairs of ranks (or the number of variables that are ranked).

In this example, \( N = 7 \).

\[
r_s = 1 - \frac{6 \sum D^2}{N(N^2 - 1)}
\]

\[
r_s = 1 - \frac{6(12)}{7(49 - 1)}
\]

\[
r_s = 1 - \frac{72}{7(48)}
\]

\[
r_s = 1 - \frac{72}{336}
\]

\[
r_s = 1 - .214
\]

\[
r_s = .786
\]

\[
r_s = .79
\]

Note that Pearson's \( r \) and Spearman's \( r \) are most typically reported to two decimal places.

Spearman's \( r \) may be interpreted as a measure of the linear correlation between ranks. Pearson's \( r \) will produce the same value as Spearman's \( r \) on the same set of ranked data. In the case where the variables are expressed in their original form as continuous measures, the Pearson's \( r \) will not equal the Spearman's \( r \) after they have been converted to ranks, but they will have similar values.

**Significance Test for Spearman's \( r \)**

Spearman's \( r \) cannot be tested for significance in the same manner as Pearson's \( r \). Refer to the significance table for Spearman's \( r \) in Appendix C page 277 of *Statistics: A Gentle Approach*. This table presents the minimum size of Spearman's \( r \) in order to reject the null hypothesis at \( p < .05 \) and \( p < .01 \). This table is unique because the degrees of freedom do not have to be calculated, because \( N \) is used
directly in the table. In the previous example, the null and alternative hypotheses are:

Ho: $r = 0$

Ha: $r \neq 0$

The Spearman's significance table reveals that an $r$ value of at least .786 is necessary to reject Ho at the $p < .05$ level. The obtained $r_s$ value equals this level exactly, therefore, Ho can be rejected at $p = .05$. According to APA format the derived $r_s$ value may be reported as:

$r_s(7) = .79, p = .05$

In conclusion, with respect to the previous example, the results indicate that there is a strong positive correlation between the husband's and the wife's preferences for the seven vegetables. The probability that $r_s$ would equal .79 by chance alone is equal to five chances out of a hundred or .05. Thus, the $r_s = .79$ may be reported as statistically significant.

**Ties in Ranks**

The following example comes from a study (Coolidge, 1983) of Wechsler Intelligence Scale for Children-Revised (WISC-R) profiles of emotionally disturbed children (EDC) and learning disabled children (LDC). The WISC-R contains 10 separate subtests. The focus of the study was whether the relative strengths and weaknesses within each group of children were similar between the two groups. The subtests were ranked from highest (#1) to lowest (#10) in terms of overall group performance. The data is tabled as follows:

<table>
<thead>
<tr>
<th>Subtest</th>
<th>EDC Rank</th>
<th>LDC Rank</th>
<th>D</th>
<th>$D^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>5.5*</td>
<td>1.5</td>
<td>2.25</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>9</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>7</td>
<td>-1</td>
<td>17</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>5.5*</td>
<td>-0.5</td>
<td>.25</td>
</tr>
</tbody>
</table>
\[ \sum D^2 = 13.5 \]

Note: Since there is a tie between two subtests at the 5th rank place, positions 5 and 6 are added together and divided by 2 for a 5.5 average rank for both subtests. An asterisk was added to indicate the tie. **Note that since the 5th and 6th ranked places are now taken, the next lowest subtest is given the 7th place rank.** Had a three-way tie occurred at the 5th rank place, then places 5, 6, and 7 would have been added together and divided by three. Thus, all three tied subtests would be given a 6. The next subtest would be 8th ranked.

\[
r = 1 - \frac{6 \sum D^2}{N(N^2 - 1)}
\]

\[
r = 1 - \frac{6(135)}{10(100 - 1)}
\]

\[
r = 1 - \frac{81}{990}
\]

\[
r = 1 - .0818
\]

\[
r = .918
\]

\[
r = .92
\]

The null and alternative hypotheses are:

**Ho:** \( r = 0 \)

**Ha:** \( r \neq 0 \)

According to the Spearman significance table, Ho is rejected at \( p < .01 \). It may be concluded that
there is a significant, strong positive correlation between the two sets of ranks, and the relative strengths and weaknesses within the groups are similar between the two groups. This means that knowing the rank of a subtest in one group will predict the rank of the same subtest in the other group $r_s = .92, p < .01$.

**Point-Biserial Correlation**

The point-biserial correlation ($r_{pb}$) gives an estimate of the degree of relationship between a dichotomous variable and a continuous variable. Before the advent of modern calculators, students having to use slow, noisy, mechanical ones, would dichotomize one of two continuous variables (like using the median score on the variable to be dichotomized) and then run the point-biserial correlation instead of Pearson’s because the formula was simpler. Interestingly, Pearson’s correlation coefficient yields the same value at the point-biserial correlation formula.

One typical use of $r_{pb}$ is correlating a single test item, which is dichotomous (like yes-no) with the overall test score, which is continuous. This might tell the researcher whether an individual item is a good predictor of the overall test score. The formula is as follows:

$$r_{pb} = \frac{\bar{X}_1 - \bar{X}_2}{S} \cdot \sqrt{pq}$$

where

$\bar{X}_1$ = the mean score on the continuous variable of just the participants on level one of the dichotomous variable

$\bar{X}_2$ = the mean score on the continuous variable of just the participants on level two of the dichotomous variable

$p$ and $q$ are the proportion of participants in each group, with $p + q = 1$.
The standard deviation of all the participants on the continuous variable

- the proportion of persons in level one of the dichotomous variable

- 1 - p

For example, a company owner wanted to hire exceptionally dexterous workers to assemble small electronic items. The owner did an experiment to see if fine motor movement, as measured by speed of finger tapping, was related to gender. In this case, gender is inherently a dichotomous variable while finger tapping (number of taps in 5 seconds) is measured as a continuous variable. The data is tabled as follows:

<table>
<thead>
<tr>
<th>Subjects</th>
<th>Finger Taps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>10</td>
</tr>
<tr>
<td>Female</td>
<td>12</td>
</tr>
<tr>
<td>Female</td>
<td>14</td>
</tr>
<tr>
<td>Male</td>
<td>9</td>
</tr>
<tr>
<td>Male</td>
<td>11</td>
</tr>
<tr>
<td>Female</td>
<td>13</td>
</tr>
<tr>
<td>Female</td>
<td>14</td>
</tr>
<tr>
<td>Female</td>
<td>10</td>
</tr>
<tr>
<td>Male</td>
<td>8</td>
</tr>
<tr>
<td>Female</td>
<td>11</td>
</tr>
</tbody>
</table>

Arbitrarily, consider the males as Level 1 and females as Level 2 of the dichotomous variable.

However, caution is in order! As previously noted it is entirely arbitrary which dichotomous group ends up in Level 1 or 2. However, the interpretation of the final test statistic very much depends on those two levels. It is an artifactual quirk, but Level 1 is considered the “higher” level and Level 2 is considered as “lower” on the dichotomous scale. Thus, if there is a positive phi correlation, then higher
on the dichotomous variable (meaning Level 1) is associated with higher scores on the continuous variable. If there is a negative phi correlation then higher scores on the dichotomous variable (Level 1) are associated with lower scores on the continuous variable.

\[
\bar{X}_1 = \text{mean score on continuous variable of the level one group}
\]

\[
\bar{X}_1 = \frac{(10+9+11+8)}{4} = 9.50
\]

\[
\bar{X}_2 = \text{mean score on continuous variable of the level two group}
\]

\[
\bar{X}_2 = \frac{(12 + 14 + 13 + 14 + 10 + 11)}{6} = 12.33
\]

\[
S = \text{standard deviation of continuous variable}
\]

\[
S = \sqrt{\frac{\sum x^2 - \left(\frac{\sum x}{N}\right)^2}{N-1}}
\]

\[
S = \sqrt{\frac{1292 - \left(\frac{112}{10}\right)^2}{9}}
\]

\[
p = \text{proportion of level one to total}
\]

\[
p = \frac{4}{10} = 0.4
\]

\[
q = 1 - p = 1 - 0.4 = 0.6
\]

\[
r_{pb} = \frac{9.50 - 12.33}{2.044} \cdot \sqrt{(0.4)(0.6)}
\]
\[ r_{pb} = -0.679 \]
\[ \Gamma_{pb} = -0.68 \]

**Testing for the Significance of the Point-Biserial Correlation Coefficient**

The significance of the point-biserial correlation coefficient is tested the same way as the Pearson's \( r \). The null and alternative hypotheses are as follows:

\[ H_0: r = 0 \]
\[ H_a: r \neq 0 \]

The value of \( r \) and the number of pairs of scores are converted to a \( t \) distribution with \( N - 2 \) degrees of freedom where \( N \) is the number of pairs of scores. The \( t \) statistic can only be used to test whether \( r \) is equal to zero. The formula is as follows:

\[
t = \frac{r}{\sqrt{\frac{1 - r^2}{N - 2}}}\]

where \( N \) = the number of pairs of scores and \( df = N - 2 \).

Thus:

\[
t = \frac{-0.6790}{\sqrt{\frac{1 - 0.4611}{8}}}\]

\[
t = \frac{-0.6790}{\sqrt{0.0674}}\]

\[
t = \frac{-0.6790}{0.2595}\]
The critical value of $t$ is obtained from the $t$ distribution in Appendix B of the text (page 276). In this case, the critical value of $t$ with $df = 8$ and $p = .05$ is $\pm 2.306$. Our formula-derived $t$ of $-2.62$ exceeds the critical value of $t$, therefore, we reject $H_0$ and conclude that our $r_{pb} = -0.68$ is statistically significant. This means that there is a significant relationship between gender and finger tapping. Since males were Level 1, the interpretation of the negative correlation would be that males are associated with lower levels of the continuous variable, and the lower level of the dichotomous variable (females) is associated with higher levels of the continuous variable.

In a research paper, the $r_{pb}$ might be reported as follows:

"There was a strong and significant negative correlation found between gender and fine motor movement $r_{pb}(8) = -0.68, p < .05$, in other words, males are associated with slower fine motor movements and females appear to have faster fine motor movements."

### Phi $\phi$ Correlation

The phi (rhymes with fee) correlation gives an estimate of the degree of relationship between two dichotomous variables. The value of the phi $\phi$ correlation coefficient is interpreted just like the Pearson $r$, that is, it can vary from $-1.00$ to $+1.00$.

For example, let us compare the industrial ratings (above average versus at or below average) of two rival accreditation companies (NHAC and UGAC) in Chapter 4 on a sample of 50 businesses. We are interested in determining whether their ratings are related to each other, i.e., do they both rate the same company highly? Does a high rating from NHAC on a company predict a high rating UGAC on the same company? Or is a high rating on a company by NHAC associated with an opposite rating from
UGAC on the same company (resulting in a negative phi correlation)? The frequencies of the companies who met the above and below criterion for both accreditation groups are tabled as follows:

<table>
<thead>
<tr>
<th></th>
<th>UGAC</th>
<th>NHAC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Above Average</td>
<td>Average or Below</td>
</tr>
<tr>
<td>Above Average</td>
<td>6 (a)</td>
<td>2 (b)</td>
</tr>
<tr>
<td>Average or Below</td>
<td>6 (c)</td>
<td>36 (d)</td>
</tr>
</tbody>
</table>

The individual cells in this matrix of numbers have been labeled "a" through "d" in order to identify the cells in the formula. The formula for phi is:

\[
\phi = \frac{ad - bc}{\sqrt{(a + b)(c + d)(a + c)(b + d)}}
\]

\[
\phi = \frac{(6)(36) - (2)(6)}{\sqrt{(6 + 2)(6 + 36)(6 + 6)(2 + 36)}}
\]

\[
\phi = \frac{216 - 12}{\sqrt{8x42x12x38}}
\]

\[
\phi = \frac{204}{\sqrt{153,216}}
\]

\[
\phi = \frac{204}{391.4281}
\]

\[
\phi = 0.52
\]
Testing for the Significance of Phi

The phi correlation can be tested for significance by converting the value of phi into a chi-square statistic ($\chi^2$) and comparing it to the chi-square distribution. The null and alternative hypotheses are as follows:

$$\begin{align*}
\text{Ho: } & \phi = 0 \\
\text{Ha: } & \phi \neq 0
\end{align*}$$

The formula for the conversion of phi to chi-square is:

$$\chi^2 = N (\phi)^2$$

where $N$ = the number of participants in the correlation and the $df$ is always equal to 1.

The critical values of chi square are in Appendix D page 278 of the text. The critical value of chi-square with $df = 1$ at $p = .05$ is 3.84. The obtained value of chi-square is $(50)(.52)^2 = 13.52$. The obtained value exceeds the critical value, and Ho is rejected. It is concluded that there is a significant positive relationship between the two accreditation companies and their ratings on the 50 companies.
History Trivia

Francis Galton (1822-1911) is credited with the first formal presentation of a statistical relationship. He was profoundly influenced by Charles Darwin, and Galton's work was devoted to prediction as a tool for the study of inheritance. For part of his research, Galton used sweet pea seeds, and he ranked the size of the parent and offspring seeds. He also studied height in fathers and sons. From this research he noted that the offspring at their mature height showed less variability and fewer extremes than the parents. He called this phenomenon "reversion" and thus was derived the symbol "r.r."

Karl Pearson (1857-1936) was a friend of Galton, and he is likewise famous in the early history of statistics. In 1893 Pearson presented the term "standard deviation." In 1895 he published an article deriving the current correlation coefficient formula and its test statistic. In the 1920s, Pearson's son, Egon, developed the idea of hypothesis testing, and his work led to the present definitions of the null and alternative hypotheses.

In 1904, Charles Spearman (1863-1945) published an article in which he stated a formula for rank-order correlation. However, the present formula was actually derived by Karl Pearson in 1907. In addition, Galton was probably the first to develop the concept of correlation with ranks when he rank-ordered his sweet pea seeds. Furthermore, it was Pearson who proposed to use rho (ρ) as the symbol for the rank order correlation coefficient, although most statistics books use rho (ρ) as the symbol for the population correlation coefficient. $r_s$ (where the s occurs in honor of Spearman) is also currently used as the rank-order correlation coefficient.
Key Symbols and Terms

**Correlation** - measures the degree of relationship between two variables.

**Strong Positive Relationship** - indicates that a high score on variable x will be associated with a high score on variable y, and a low score on variable x will be associated with a low score on variable y.

**Strong Negative Relationship** - indicates that a high score on variable x will be associated with a low score on variable y, and a low score on variable x will be associated with a low score on variable y.

**Archival Data** - is a type of study, which has already been conducted, and the data has already been gathered. The experimenter only has statistical power and no longer has any means of changing the original experimental design.

**Causal Relationship** - should never be inferred from a correlational design or a correlational test statistic. Causal relationships can be inferred from true experimental designs (like the archetypal experiment).

**Chance is Lumpy** - is a rule of statistics where a long and unusual occurrence in data should first be attributed to randomness or chance.

**Pearson Product Moment Correlation** - is a measure of the strength of a relationship between two continuous variables. It is measured by the coefficient $r$.

**Scatterplot** - is a graphic representation of the relationship of two continuous variables in correlational designs. It is also called a bivariate distribution.

**Linear Relationship** - is an assumption made between two continuous variables when graphically represented. The assumption is that a straight line best fits the bivariate distribution.

**Curvilinear Relationship** - is an assumption made between two continuous variables when graphically represented. The assumption is that a curved line best fits the bivariate distribution. The Pearson’s $r$ cannot be used when the relationship is curvilinear.

**Homo-scedasticity** - is an assumption of the Pearson Product Moment Correlation Coefficient that the bivariate distribution is evenly distributed throughout the length of the scatterplot best-fitting line.

**Heteroscedasticity** - is considered to be a violation of the assumption of homoscedasticity and is the condition where the bivariate distribution has greater variance on some lengths of the line than on others.

**Coefficient of Determination ($r^2$)** - is an interpretation made by squaring the Pearson $r$ and determines how much variance is accounted for in one variable by variation in another variable.

**Spearman Correlation** - is a measure of the relationship between two ordinal rankings of the same set of data.
**Point-Biserial Correlation** - is a measure of the strength of a relationship between one continuous variable and one dichotomous variable.

**Phi Correlation** - is a measure of the strength of a relationship between two dichotomous variables.

**Regression** – measures the strength of a relationship between an independent variable and a dependent variable.

**Multiple Regression** – measures the strength of a relationship between multiple independent variables and a single dependent variable.

**Degrees of Freedom** - is a complicated statistical concept, but is approximately and usually less than N of a set of numbers.

**Chi Square (χ²) Distribution** - is a distribution which can be used to test the significance of a Phi correlation and also to test nonparametric frequency data sets.

**Constant** - is a specific number that always stays the same in a statistical formula.
Practice Problems for Chapter 6

Chapter 6: AN INTRODUCTION TO CORRELATION

Problem 1: The owner of a shoe store wants to know if shoe size and weight are correlated in female adults. She measures the shoe size and asks the weight of 12 consecutive customers. They are as follows: size 6, 116 lbs; size 5.5, 121 lbs; size 10, 165 lbs; size 3, 101 lbs; size 9.5; 148 lbs; size 6, 150 lbs; size 10, 167 lbs; size 5.5, 134 lbs; size 7.7, 145 lbs; size 8, 201 lbs; size 8.5, 138 lbs; size 5.5, 123 lbs.

1. Conduct the Pearson correlation on this data.

2. Test for statistical significance.

3. Graph a scatterplot of the data.

Problem 2: Two sportswriters ranked ten pro football teams for best team in the 1990’s. The team and rankings by the two writers respectively are as follows:

<table>
<thead>
<tr>
<th>Team</th>
<th>#1</th>
<th>#2</th>
<th>#3</th>
<th>#4</th>
<th>#5</th>
<th>#6</th>
<th>#7</th>
<th>#8</th>
<th>#9</th>
<th>#10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fortyiners</td>
<td>#1</td>
<td>#2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Broncos</td>
<td>#2</td>
<td>#3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cowboys</td>
<td>#3</td>
<td>#5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Packers</td>
<td>#4</td>
<td>#1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raiders</td>
<td>#5</td>
<td>#4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Colts</td>
<td>#6</td>
<td>#8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dolphins</td>
<td>#7</td>
<td>#6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Titans</td>
<td>#8</td>
<td>#10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jaguars</td>
<td>#9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bills</td>
<td>#10</td>
<td>#7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1. Conduct a Spearman correlation on this data, and test for statistical significance.
Test Questions for Chapter 6 – Fundamentals of Business Statistics

1. A Pearson correlation above r = .50 may be labeled
   a. weak
   b. significant
   c. strong
   d. nonsignificant

2. Given a strong negative correlation, a score that is relatively high on variable x will be relatively _____ on variable y.
   a. low
   b. medium
   c. high
   d. significant

3. Variables may be analyzed and studied long after the data has been gathered. Such data is known as _______ data.
   a. correlation
   b. experimental
   c. archival
   d. archetypal

4. According to the course notes, a correlational analysis _______
   a. sometimes provides clues about cause-effect relationships
   b. often provides clues about cause-effect relationships
   c. never provides clues about cause-effect relationships
   d. is relatively useless compared to controlled experiments
5. According to the course notes, what are the consequences of the Type I error in the Mozart Effect?
   a. benign
   b. horrible
   c. could lead to mental disturbances
   d. more classical musicians

6. Milk has been found to be positively correlated to cancer because
   a. milk causes cancer
   b. drinking milk provides an environment for cancer
   c. milk drinkers clearly live longer
   d. there are suspicions but a correlation design doesn’t allow one to say definitely why two variables are correlated

7. Abelson’s first law of statistics is
   a. correlation never implies causation
   b. correlation sometimes implies causation
   c. chance is lumpy
   d. if something is too good to be true, it probably is too good to be true

8. ________________ correlation measures the strength of a relationship between two continuous variables.
   a. Pearson
   b. Spearman
   c. Point-biserial
   d. Phi
9. The _____________ hypothesis establishes whether a one-tailed or two-tailed significance test will be used.
   a. null
   b. alternative
   c. general
   d. specific

10. In the study of 30 pairs of scores in correlation, what are the degrees of freedom?
    a. 30
    b. 29
    c. 28
    d. 15

11. A _______________ is a graphic representation of the pairs of scores plotted along the x and y axes in a correlation coefficient.
    a. bivariate distribution
    b. scatterplot
    c. homoscedast
    d. heteroscedast

12. The coefficient of determination explains
    a. the relationship between two variables.
    b. the proportion of variance in one variable accounted for by another variable.
    c. the cause-effect relationship between two variables.
    d. all of the above.
13. The ____________ assumption is met if, in a scatterplot, the distance from the points to the line is relatively equal all along the line.
   a. homoscedasticity
   b. linearity
   c. curvilinearity
   d. heterolinearity

14. The strength of a relationship between two variables in correlation is ________________ of the significance testing.
   a. highly dependent
   b. actually independent (not completely, but mostly)
   c. highly independent
   d. the obverse

15. The interest in _______________ analysis is seeing how various values of the independent variable predict corresponding values in the dependent variable.
   a. correlational
   b. curvilinear
   c. topographical
   d. regression

16. In the formula $Y = a + bX$, the letter “b” defines
   a. the intercept of the x axis
   b. the intercept of the y axis
   c. the slope or angle of the straight line
   d. all of the above
17. In multiple regression, the coefficient R can be interpreted as
   a. the percentage of variance accounted for in the dependent variable by the set of independent variables.
   b. the percentage of variance accounted for in the dependent variable by a single independent variable.
   c. the strength of a relationship between the dependent variable and a set of independent variables
   d. all of the above.

18. In multiple regression, the R Square can be interpreted as
   a. the percentage of variance accounted for in the dependent variable by the set of independent variables.
   b. the percentage of variance accounted for in the dependent variable by a single independent variable.
   c. the strength of a relationship between the dependent variable and a set of independent variables
   d. the percentage of variance accounted for in the dependent variable by the set of independent variables minus an estimate penalty.

19. In multiple regression, the Adjusted R Square can be interpreted as
   a. the percentage of variance accounted for in the dependent variable by the set of independent variables.
   b. the percentage of variance accounted for in the dependent variable by a single independent variable.
c. the strength of the relationship between the dependent variable and the set of independent variables.

d. the percentage of variance accounted for in the dependent variable by the set of independent variables minus an estimate penalty.

20. In multiple regression, the ANOVA table tells us whether the three main regression coefficients (R, R Square, Adjusted R Square) are significantly different than

   a. zero.
   b. 1.00.
   c. .50.
   d. 1.50.

21. In a multiple regression analysis, the final section of the output contains the coefficients. Which of these coefficients is of primary concern?

   a. unstandardized B
   b. standard error of B
   c. standardized coefficient beta
   d. standard error of beta

22. The ______________ correlation gives an estimate of the degree of relationship between two dichotomous variables

   a. Pearson
   b. Spearman
   c. Point-biserial
   d. Phi
23. In order to test for the significance of a Pearson correlation, one must use the ____________ distribution.
   a.  t
   b.  F
   c.  $\chi^2$
   d.  z

24. In order to test for the significance of the phi correlation, one must use the ____________ distribution.
   a.  t
   b.  F
   c.  $\chi^2$
   d.  z

25. The formula for Spearman’s rank order correlation was actually derived by
   a.  Francis Galton
   b.  Charles Darwin
   c.  Karl Pearson
   d.  Egon Pearson

Problem 1: The owner of a shoe store wants to know if shoe size and weight are correlated in adult males. She measures the shoe size and asks the weight of 14 consecutive customers. They are as follows: size 9, 176 lbs; size 7.5, 141 lbs; size 10, 185 lbs; size 12, 202 lbs; size 9.5; 174 lbs; size 10, 150 lbs; size 10, 193 lbs; size 10.5, 237 lbs; size 13, 248 lbs; size 8, 159 lbs; size 8.5, 136 lbs; size 9.5, 174 lbs; size 9, 172 lbs; size 11, 183 lbs.
26. The Pearson correlation on this data is $r =$
   a. .61
   b. .71
   c. .81
   d. .91

27. The $r$ value can be reported significant (lowest possible) at $p <$
   a. > .05
   b. < .05
   c. < .01
   d. < .001

28. The $df$ for this problem are
   a. 14
   b. 12
   c. 10
   d. 8

29. The coefficient of determination for this problem is
   a. .65
   b. .81
   c. .809
   d. .99

30. The percentage of variance in shoe size that can be accounted by weight is
   a. 65
   b. 81
c. 90

d. 99