CHAPTER 5 Relationships between Quantitative Variables

Chapter Outline

5.1 Scatterplots and Linear Correlation

Learning Objectives

- Understand the concepts of bivariate data and correlation, and the use of scatterplots to display bivariate data.
- Understand when the terms 'positive', 'negative', 'strong', and 'perfect' apply to the correlation between two variables in a scatterplot graph.
- Calculate the linear correlation coefficient and coefficient of determination.
- Understand properties and common errors of correlation.

Introduction

So far we have learned how to describe distributions of a single variable. But what if we notice that two variables seem to be related? We may notice that the values of two variables, such as verbal SAT score and GPA, behave in the same way and that students who have a high verbal SAT score also tend to have a high GPA (see table below). In this case, we would want to study the nature of the connection between the two variables.

TABLE 5.1: A Table of Verbal SAT Values and GPAs for Seven Students

Student	SAT Score	GPA
1	595	3.4
2	520	3.2
3	715	3.9
4	405	2.3
5	680	3.9
6	490	2.5
7	565	3.5

These types of studies are quite common, and we can use the concept of correlation to describe the relationship between the two variables.

Bivariate Data, Correlation Between Values, and the Use of Scatterplots

Correlation measures the linear relationship between two quantitative variables. Correlation is possible when we have **bivariate data**. In other words, when the subjects in our dataset have scores on two separate quantitative variables, we have bivariate data. In our example above, we notice that there are two observations (verbal SAT score and GPA) for each subject (in this case, a student). Can you think of other scenarios when we would use bivariate data?

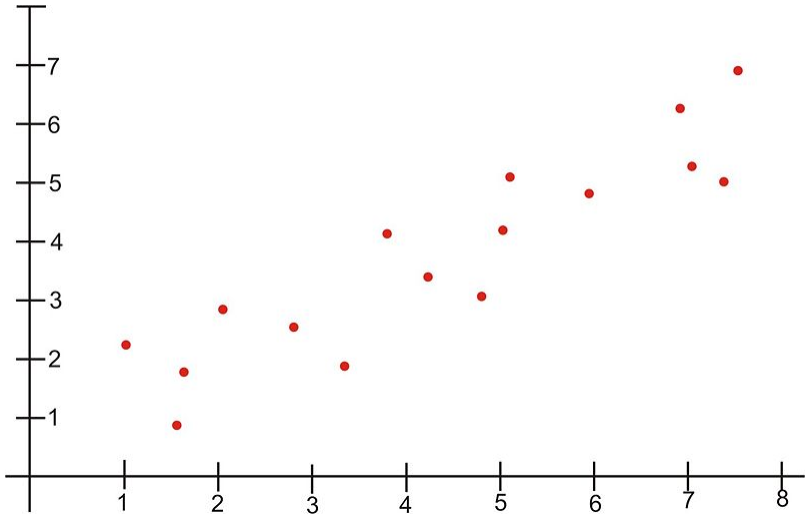
If we carefully examine the data in the example above, we notice that those students with high SAT scores tend to have high GPAs, and those with low SAT scores tend to have low GPAs. In this case, there is a tendency for students to score similarly on both variables, and the performance between variables appears to be related.

Correlation Patterns in Scatterplot Graphs

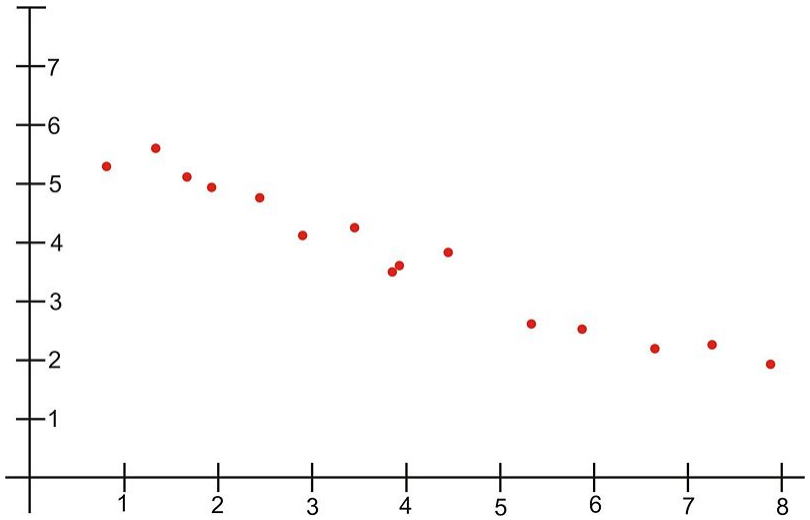
Scatterplots, like the one below, display bivariate data and provide a visual representation of the relationship between the two variables. In a scatterplot, each point represents a paired measurement of two variables for a specific subject, and each subject is represented by one point on the scatterplot.

Direction of Relationship

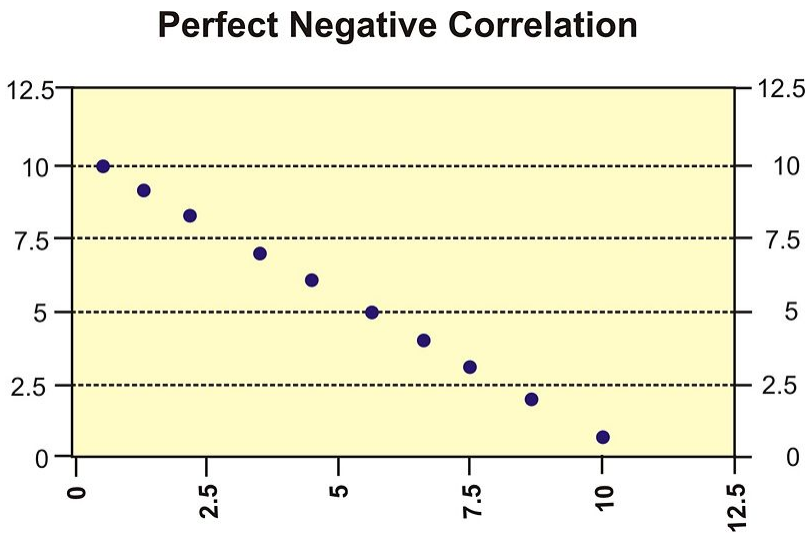
Examining a scatterplot graph allows us to obtain some idea about the relationship between two variables. When the points on a scatterplot graph produce a lower-left-to-upper-right pattern (see below), we say that there is a **positive correlation** between the two variables. This pattern means that when the score of one observation is high, we expect the score of the other observation to be high as well, and vice versa.



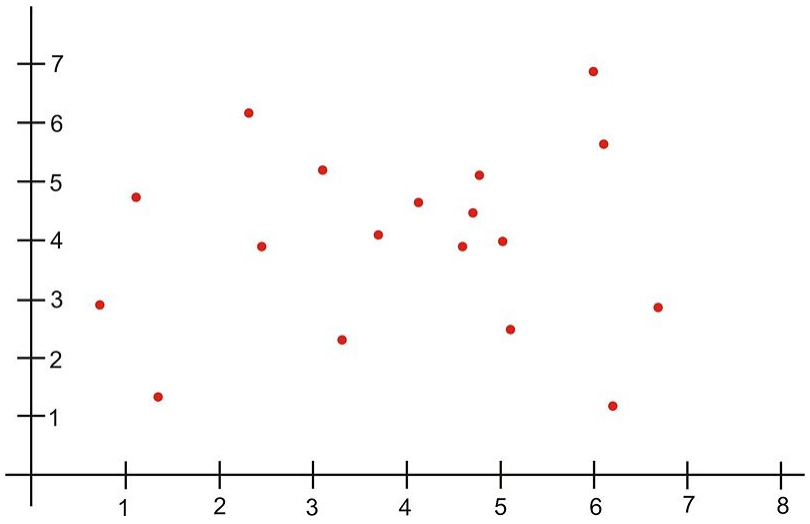
When the points on a scatterplot graph produce a upper-left-to-lower-right pattern (see below), we say that there is a **negative correlation** between the two variables. This pattern means that when the score of one observation is high, we expect the score of the other observation to be low, and vice versa.



When all the points on a scatterplot lie on a straight line, you have what is called a **perfect correlation** between the two variables (see below).

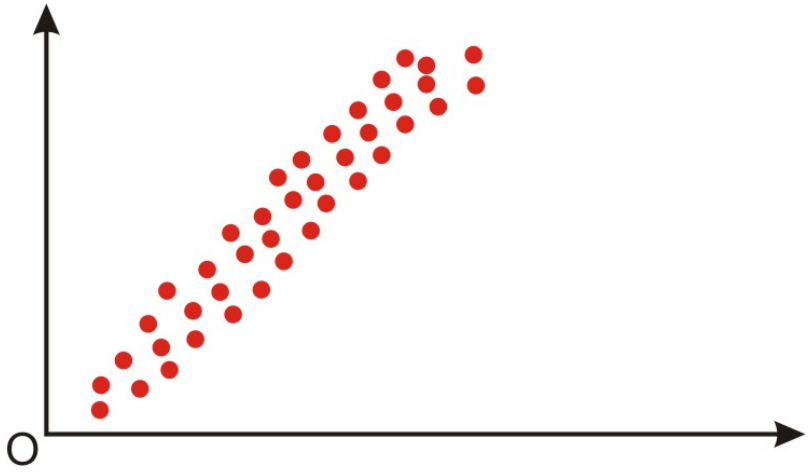


A scatterplot in which the points do not have a linear trend (either positive or negative) is called a **zero correlation** or a **near-zero correlation** (see below).

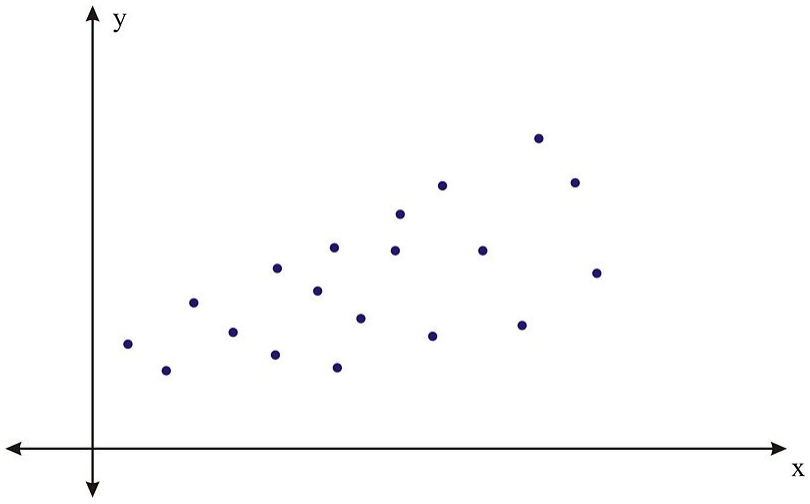


Magnitude of Relationship

When examining scatterplots, we also want to look not only at the direction of the relationship (positive, negative, or zero), but also at the **magnitude** of the relationship. If we drew an imaginary oval around all of the points on the scatterplot, we would be able to see the extent, or the magnitude, of the relationship. If the points are close to one another and the width of the imaginary oval is small, this means that there is a strong correlation between the variables (see below).



However, if the points are far away from one another, and the imaginary oval is very wide, this means that there is a weak correlation between the variables (see below).



Correlation Coefficient

While examining scatterplots gives us some idea about the relationship between two variables, we use a statistic called the **Pearson correlation coefficient** to give us a more precise measurement of the relationship between the two variables. We use r to denote the correlation coefficient, and r has the following properties:

- r is always a value between -1 and +1
- The further an r value is from zero, the stronger the relationship between the two variables. The absolute value of the coefficient indicates the magnitude, or strength, of the relationship.
- The sign of r indicates the nature of the relationship: A positive r indicates a positive relationship, and a negative r indicates a negative relationship.

If two variables have a perfect linear relationship (meaning they fall on a straight line), then r is equal to 1.0 or -1.0, depending on the direction of the relationship. When there is no linear relationship between two variables, r=0. It is important to remember that a correlation coefficient of 0 means that there is no linear relationship, but there may still be a relationship between the two variables. For example, there could be a quadratic relationship between them.

The name of this statistics is the **Pearson product-moment correlation coefficient**. It is symbolized by the letter r.

Generally speaking, you may think of the values of r in the following manner:

- If $|r|$ is between 0.85 and 1, there is a strong correlation.
- If $|r|$ is between 0.5 and 0.85, there is a moderate correlation.
- If $|r|$ is between 0.1 and 0.5, there is a weak correlation.
- If $|r|$ is less than 0.1, there is no apparent correlation.

Coefficient of Determination

At the risk of overloading you with new terms, there is one more that is worth learning in this lesson, the **coefficient of determination**. The coefficient of determination is very simple to calculate if you know the correlation coefficient, since it is just r^2. The coefficient of determination can be interpreted as the percentage of variation of the y variable that can be attributed to the relationship. In other words, a value of $r^2 = .63$ can be interpreted as "63% of the variation in Y can be attributed to the variation in X."

Thus, the correlation coefficient not only provides a measure of the relationship between the variables, but it also gives us an idea about how much of the total variance of one variable can be associated with the variance of the other. The higher the correlation we have between two variables, the larger the portion of the variance that can be explained by the independent variable.

Example A

Elaina is curious about the relationship between the weight of a dog and the amount of food it eats. Specifically, she wonders if heavier dogs eat more food, or if age and size factor in. She works at the Humane Society, and does some research. After some calculation, she determines that dog weight and food weight exhibit an r-value of 0.73. What can Elaina say about the relationship, based on her research? What percentage of the increases in food intake can she attribute to weight, according to her research?

Solution

The calculated r-value of 0.73 tells us that Elaina's data demonstrates a moderate to strong correlation between the variables.

Since the coefficient of determination tells us the percentage of changes in the output variable that can be attributed to the input variable, we need to calculate r^2:

$$r^2 = (0.73)^2 = .5329$$

Approximately 53% of increases in food intake can be attributed to the linear relationship between food intake and the weight of the dog. Since weight explains less than 100% of the difference in food intake, this suggests that other factors, perhaps age and size, are also involved.

Example B

Tuscany wonders if barrel racing times are related to the age of the horse. Specifically, she wonders if older horses take longer to complete a barrel racing run. As a member of the Pony Club, she does some research, and determines that horse age to barrel run time exhibits an r-value of 0.52.

What can Tuscany say about horse age vs barrel race time, according to her research?

Solution

Tuscany's research suggests that there is a moderate to weak correlation between horse age and barrel run time. In other words, the research suggests that $(0.52)^2 = .27 = 27\%$ of the differences between barrel run times could be attributable to the linear relationship between barrel run time and the age of the horse.

How to Calculate r

To understand how this coefficient is calculated, let's suppose that there is a positive relationship between two variables, X and Y. If a subject has a score on X that is above the mean, we expect the subject to have a score on Y that is also above the mean. Pearson developed his correlation coefficient by computing the sum of cross products. He multiplied the two scores, X and Y, for each subject and then added these cross products across the individuals. Next, he divided this sum by the number of subjects minus one. This coefficient is, therefore, the mean of the cross products of scores.

Pearson used standard scores (z-scores, t-scores, etc.) when determining the coefficient.

Therefore, the formula for this coefficient is as follows:

$$r_{XY} = \frac{\sum z_X z_Y}{n-1}$$

In other words, the coefficient is expressed as the sum of the cross products of the standard z-scores divided by the number of degrees of freedom. For correlation, df = n-1, where n is the number of bivariate data points in your analysis.

Calculating r from Descriptive Statistics

If you have the raw scores for your data, along with the mean and standard deviation of each variable, you can also calculate r using the following formula. This formula is less computationally intensive when you are trying to calculate r by hand:

$$r_{XY} = \frac{SP}{(n-1)(s_x)(s_y)}$$

where SP stands for sum of products. To calculate SP, use the following formula:

$$SP = \sum xy - \frac{(\sum x)(\sum y)}{n}$$

Calculating r from Raw Scores Only

An equivalent formula that uses the raw scores only is called the raw score formula and is written as follows:

$$r_{XY} = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]}\sqrt{\left[n\sum y^2 - (\sum y)^2\right]}}$$

Note that n is used instead of $n-1$, because we are using actual data and not z-scores. Let's use our example from the introduction to demonstrate how to calculate the correlation coefficient using the raw score formula.

Example C

What is the Pearson product-moment correlation coefficient for the two variables represented in the table below?

TABLE 5.2: Table of Values for this Example

Student	SAT Score	GPA
1	595	3.4
2	520	3.2
3	715	3.9
4	405	2.3
5	680	3.9
6	490	2.5
7	565	3.5

In order to calculate the correlation coefficient, we need to calculate several pieces of information, including xy, x^2, and y^2. Therefore, the values of xy, x^2, and y^2 have been added to the table.

TABLE 5.3:

Student	SAT Score (X)	GPA (Y)	xy	x^2	y^2
1	595	3.4	2023	354025	11.56
2	520	3.2	1664	270400	10.24
3	715	3.9	2789	511225	15.21
4	405	2.3	932	164025	5.29
5	680	3.9	2652	462400	15.21
6	490	2.5	1225	240100	6.25
7	565	3.5	1978	319225	12.25
Sum	3970	22.7	13262	2321400	76.01

Method 1

We can calculate SP from the raw data as follows:

$$SP = 13262 - \frac{(3970)(22.7)}{7} = 387.86$$

After calculating the values of s_x= 107.89 and s_y= 0.632, we then calculate r as:

$$r_{XY} = \frac{387.86}{(6)(107.89)(0.632)} = 0.948$$

Method 2

If we were to apply the raw formula to solve this problem, we find the following:

$$r_{XY} = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]}\sqrt{\left[n\sum y^2 - (\sum y)^2\right]}} = \frac{(7)(13262) - (3970)(22.7)}{\sqrt{[(7)(2321400) - 3970^2][(7)(76.01) - 22.7^2]}}$$
$$= \frac{2715}{2864.22} \approx 0.95$$

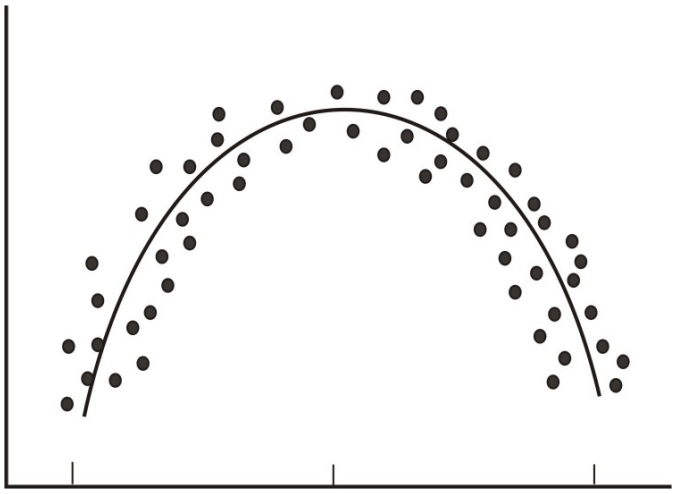
The Properties and Common Errors of Correlation

Correlation is a measure of the linear relationship between two variables-it does not necessarily state that one variable is caused by another. For example, a third variable or a combination of other things may be causing the two correlated variables to relate as they do. Therefore, it is important to remember that we are interpreting the variables and the variance not as causal, but instead as relational.

When examining correlation, there are four things that could affect our results: lack of linearity, outliers, homogeneity of the group, and sample size.

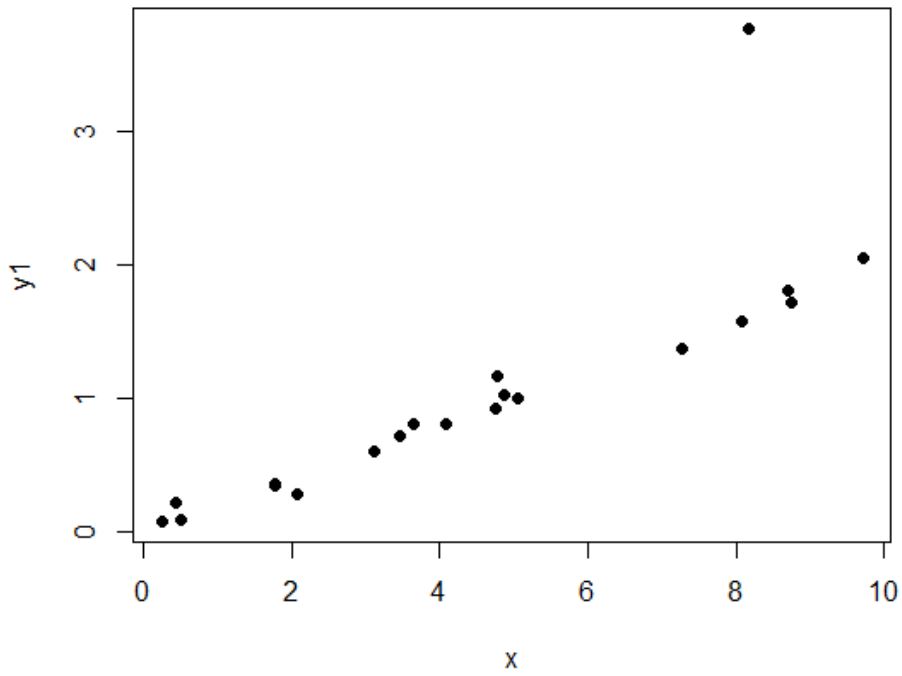
As mentioned, the correlation coefficient is the measure of the **linear** relationship between two variables. However, while many pairs of variables have a linear relationship, some do not. For example, let's consider performance anxiety. As a person's anxiety about performing increases, so does his or her performance up to a point. (We sometimes

call this good stress.) However, at some point, the increase in anxiety may cause a person's performance to go down. We call these non-linear relationships **curvilinear relationships**. We can identify curvilinear relationships by examining scatterplots (see below). One may ask why curvilinear relationships pose a problem when calculating the correlation coefficient. The answer is that if we use the traditional formula to calculate these relationships, it will not be an accurate index, and we will be underestimating the relationship between the variables. If we graphed performance against anxiety, we would see that anxiety has a strong affect on performance. However, if we calculated the correlation coefficient, we would arrive at a figure around zero. Therefore, the correlation coefficient is not always the best statistic to use to understand the relationship between variables.



Outliers

The correlation coefficient is also very sensitive to outliers, or points that fall far away from the general trend of the data. A single point that can greatly change the value of r, which means that the existence of outliers can either mask or inflate the apparent strength of the linear relationship between two variables. In the graph below, $r = 0.84$ if you include the outlier that falls well above the general trend of the rest of the data. However, if you ignore that single point, r increases to 0.99. Because of this sensitivity, it is always important to plot your data before running a correlation analysis and investigate any outlying points carefully.



Homogeneity of the Group (Restriction of Range)

Another error we could encounter when calculating the correlation coefficient is homogeneity of the group. When a group is homogeneous, or possesses similar characteristics, the range of scores on either or both of the variables is restricted. For example, suppose we are interested in finding out the correlation between IQ and salary. If only members of the Mensa Club (a club for people with IQs over 140) are sampled, we will most likely find a very low correlation between IQ and salary, since most members will have a consistently high IQ, but their salaries will still vary. This does not mean that there is not a relationship-it simply means that the restriction of the sample limited the magnitude of the correlation coefficient. This is sometimes referred to as a "restriction of range" problem.

Sample Size

Finally, we should consider sample size. One may assume that the number of observations used in the calculation of the correlation coefficient may influence the magnitude of the coefficient itself. However, this is not the case. Yet while the sample size does not affect the correlation coefficient, it may affect the accuracy of the relationship. The larger the sample, the more accurate of a predictor the correlation coefficient will be of the relationship between the two variables.

Lesson Summary

Bivariate data are data sets with two observations that are assigned to the same subject. Correlation measures the direction and magnitude of the linear relationship between bivariate data. When examining scatterplot graphs, we can determine if correlations are positive, negative, perfect, or zero. A correlation is strong when the points in the scatterplot lie generally along a straight line.

The correlation coefficient is a precise measurement of the relationship between the two variables. This index can take on values between and including -1.0 and $+1.0$.

To calculate the correlation coefficient, we most often use one of the following two formulas when calculating the coefficient by hand.

$$r_{XY} = \frac{SP}{(n-1)(s_x)(s_y)}$$

where

$$SP = \sum xy - \frac{(\sum x)(\sum y)}{n}$$

or, alternatively from raw data alone,

$$r_{XY} = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]}\sqrt{\left[n\sum y^2 - (\sum y)^2\right]}}.$$

When calculating the correlation coefficient, there are several things that could affect our computation, including curvilinear relationships, homogeneity of the group, and the size of the group.

Review Questions

1. Give 2 scenarios or research questions where you would use bivariate data sets.
2. In the space below, draw and label four scatterplot graphs. One should show:
    a. a positive correlation
    b. a negative correlation
    c. a perfect correlation
    d. a zero correlation
3. In the space below, draw and label two scatterplot graphs. One should show:
    a. a weak correlation
    b. a strong correlation.
4. What does the correlation coefficient measure?
5. The following observations were taken for five students measuring grade and reading level.

TABLE 5.4: A Table of Grade and Reading Level for Five Students

Student Number	Grade	Reading Level
1	2	6
2	6	14
3	5	12
4	4	10
5	1	4

f. a. Draw a scatterplot for these data. What type of relationship does this correlation have?
   b. Use the raw score formula to compute the Pearson correlation coefficient.

6. A teacher gives two quizzes to his class of 10 students. The following are the scores of the 10 students.

TABLE 5.5: Quiz Results for Ten Students

Student	Quiz 1	Quiz 2
1	15	20
2	12	15
3	10	12
4	14	18
5	10	10
6	8	13
7	6	12
8	15	10
9	16	18
10	13	15

a. a. Compute the Pearson correlation coefficient, r, between the scores on the two quizzes.
   b. Find the percentage of the variance, r^2, in the scores of Quiz 2 associated with the variance in the scores of Quiz 1.
   c. Interpret both $r$and r^2in words.

7. What are the three factors that we should be aware of that affect the magnitude and accuracy of the Pearson correlation coefficient?