

Scatterplots and Correlation

Diana Mindrila, Ph.D.

Phoebe Balentyne, M.Ed.

Based on Chapter 4 of The Basic Practice of Statistics (6th ed.)

Concepts:

- Displaying Relationships: Scatterplots
- Interpreting Scatterplots
- Adding Categorical Variables to Scatterplots
- Measuring Linear Association: Correlation
- Facts About Correlation

Objectives:

- Construct and interpret scatterplots.
- Add categorical variables to scatterplots.
- Calculate and interpret correlation.
- Describe facts about correlation.

References:

Moore, D. S., Notz, W. I, & Flinger, M. A. (2013). *The basic practice of statistics* (6th ed.). New York, NY: W. H. Freeman and Company.

Scatterplot

- The most useful graph for displaying the relationship between two quantitative variables is a **scatterplot**.

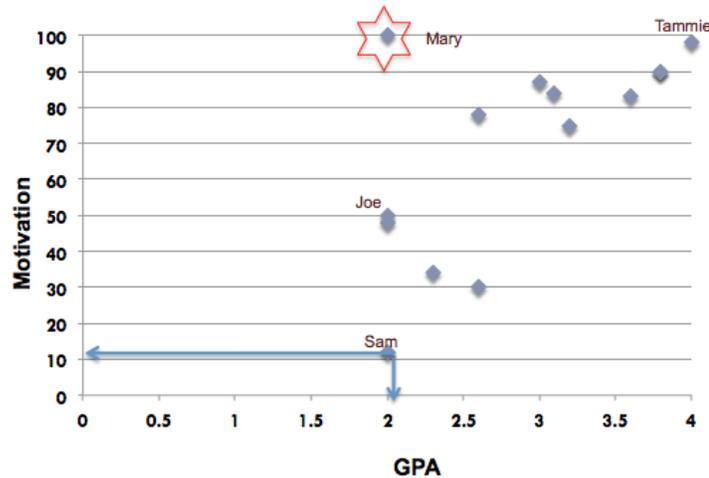
A **scatterplot** shows the relationship between two quantitative variables measured for the same individuals. The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis. Each individual in the data appears as a point on the graph.

- Many research projects are **correlational studies** because they investigate the relationships that may exist between variables. Prior to investigating the relationship between two quantitative variables, it is always helpful to create a graphical representation that includes both of these variables. Such a graphical representation is called a **scatterplot**.

Scatterplot Example

What is the relationship between students' achievement motivation and GPA?

Student	Student GPA	Motivation
Joe	2.0	50
Lisa	2.0	48
Mary	2.0	100
Sam	2.0	12
Deana	2.3	34
Sarah	2.6	30
Jennifer	2.6	78
Gregory	3.0	87
Thomas	3.1	84
Cindy	3.2	75
Martha	3.6	83
Steve	3.8	90
Jamell	3.8	90
Tammie	4.0	98



- In this example, the relationship between students' achievement motivation and their GPA is being investigated.
- The table on the left includes a small group of individuals for whom GPA and scores on a motivation scale have been recorded. GPAs can range from 0 to 4 and motivation scores in this example range from 0 to 100. Individuals in this table were ordered based on their GPA.
- Simply looking at the table shows that, in general, as GPA increases, motivation scores also increase.
- However, with a real set of data, which may have hundreds or even thousands of individuals, a pattern cannot be detected by simply looking at the numbers. Therefore, a very useful strategy is to represent the two variables graphically to illustrate the relationship between them.
- A graphical representation of individual scores on two variables is called a **scatterplot**.
- The image on the right is an example of a scatterplot and displays the data from the table on the left. GPA scores are displayed on the horizontal axis and motivation scores are displayed on the vertical axis.
- Each dot on the scatterplot represents one individual from the data set. The location of each point on the graph depends on both the GPA and motivation scores. Individuals with higher GPAs are located further to the right and individuals with higher motivation scores are located higher up on the graph.
- Sam, for example, has a GPA of 2 so his point is located at 2 on the right. He also has a motivation score of 12, so his point is located at 12 going up.
- Scatterplots are not meant to be used in great detail because there are usually hundreds of individuals in a data set.

- The purpose of a scatterplot is to provide a general illustration of the relationship between the two variables.
- In this example, in general, as GPA increases so does an individual's motivation score.
- One of the students in this example does not seem to follow the general pattern: Mary. She is one of the students with the lowest GPA, but she has the maximum score on the motivation scale. This makes her an exception or an outlier.

Interpreting Scatterplots

How to Examine a Scatterplot

As in any graph of data, look for the *overall pattern* and for striking *departures* from that pattern.

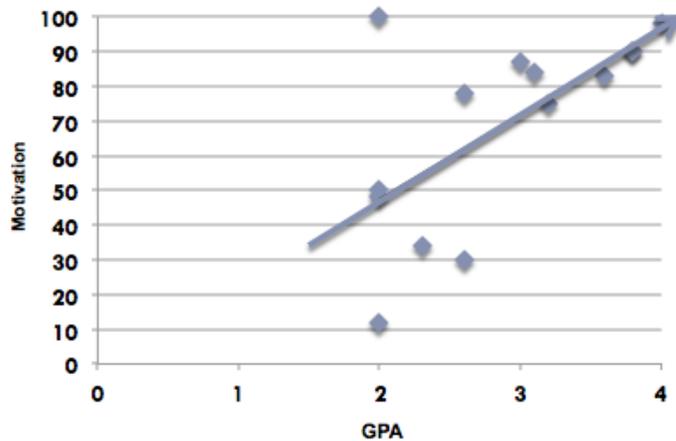
- The overall pattern of a scatterplot can be described by the **direction**, **form**, and **strength** of the relationship.
- An important kind of departure is an **outlier**, an individual value that falls outside the overall pattern of the relationship.

Interpreting Scatterplots: Direction

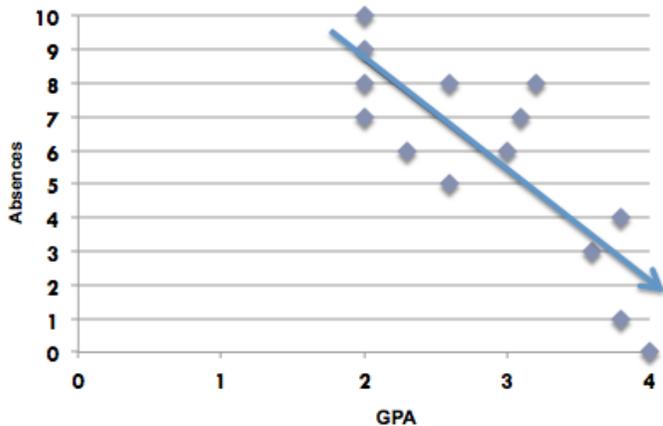
- One important component to a scatterplot is the **direction** of the relationship between the two variables.

Two variables have a **positive association** when above-average values of one tend to accompany above-average values of the other, and when below-average values also tend to occur together.

Two variables have a **negative association** when above-average values of one tend to accompany below-average values of the other.



This example compares students' achievement motivation and their GPA. These two variables have a **positive association** because as GPA increases, so does motivation.

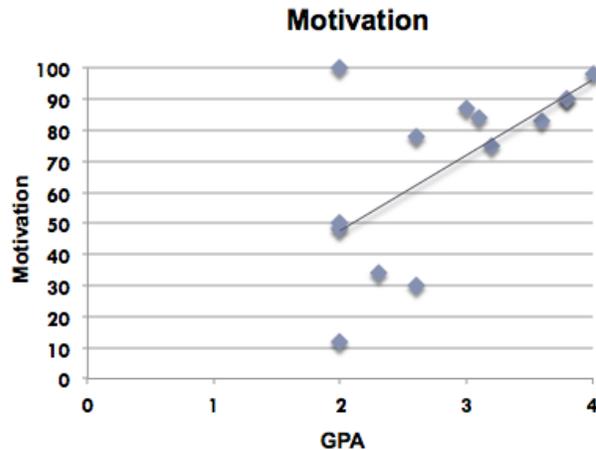


This example compares students' GPA and their number of absences. These two variables have a **negative association** because, in general, as a student's number of absences decreases, their GPA increases.

Interpreting Scatterplots: Form

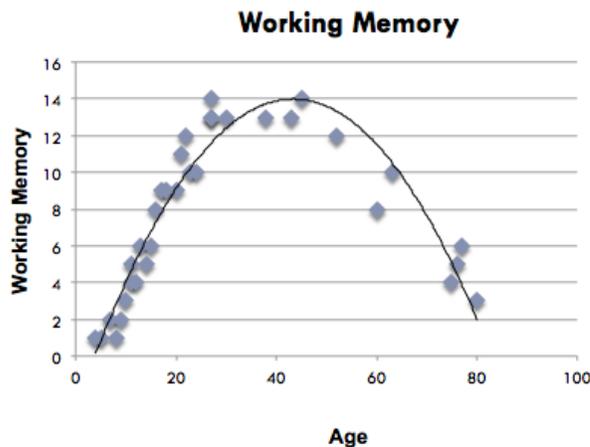
- Another important component to a scatterplot is the **form** of the relationship between the two variables.

Linear relationship:



This example illustrates a linear relationship. This means that the points on the scatterplot closely resemble a straight line. A relationship is linear if one variable increases by approximately the same rate as the other variables changes by one unit.

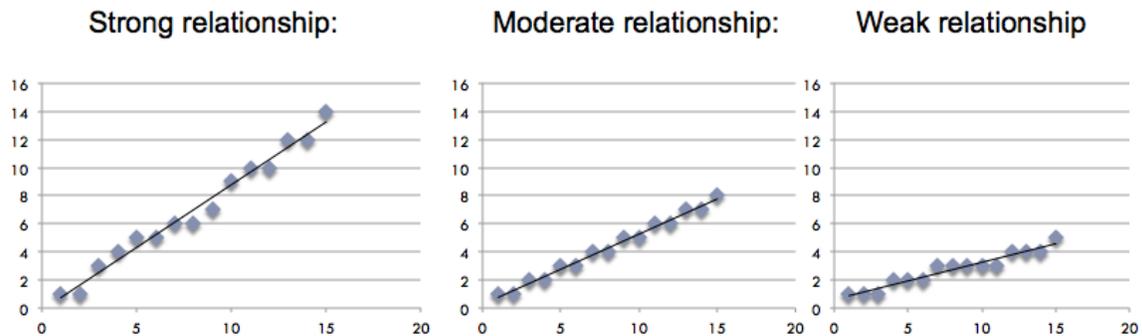
Curvilinear relationship:



This example illustrates a relationship that has the form of a curve, rather than a straight line. This is due to the fact that one variable does not increase at a constant rate and may even start decreasing after a certain point. This example describes a curvilinear relationship between the variable "age" and the variable "working memory." In this example, working memory increases throughout childhood, remains steady in adulthood, and begins decreasing around age 50.

Interpreting Scatterplots: Strength

- Another important component to a scatterplot is the **strength** of the relationship between the two variables.
- The **slope** provides information on the strength of the relationship.



- The strongest linear relationship occurs when the slope is 1. This means that when one variable increases by one, the other variable also increases by the same amount. This line is at a 45 degree angle.
- The strength of the relationship between two variables is a crucial piece of information. Relying on the interpretation of a scatterplot is too subjective. More precise evidence is needed, and this evidence is obtained by computing a coefficient that measures the strength of the relationship under investigation.

Measuring Linear Association

- A scatterplot *displays* the strength, direction, and form of the relationship between two quantitative variables.
- A correlation coefficient *measures* the strength of that relationship.

The **correlation r** measures the strength of the linear relationship between two quantitative variables.

Pearson r :

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

- r is always a number between -1 and 1.
 - $r > 0$ indicates a positive association.
 - $r < 0$ indicates a negative association.
 - Values of r near 0 indicate a very weak linear relationship.
 - The strength of the linear relationship increases as r moves away from 0 toward -1 or 1.
 - The extreme values $r = -1$ and $r = 1$ occur only in the case of a perfect linear relationship.
- Calculating a Pearson correlation coefficient requires the assumption that the relationship between the two variables is linear.
 - There is a rule of thumb for interpreting the strength of a relationship based on its r value (use the absolute value of the r value to make all values positive):

Absolute Value of r

$r < 0.3$

$0.3 < r < 0.5$

$0.5 < r < 0.7$

$r > 0.7$

Strength of Relationship

None or very weak

Weak

Moderate

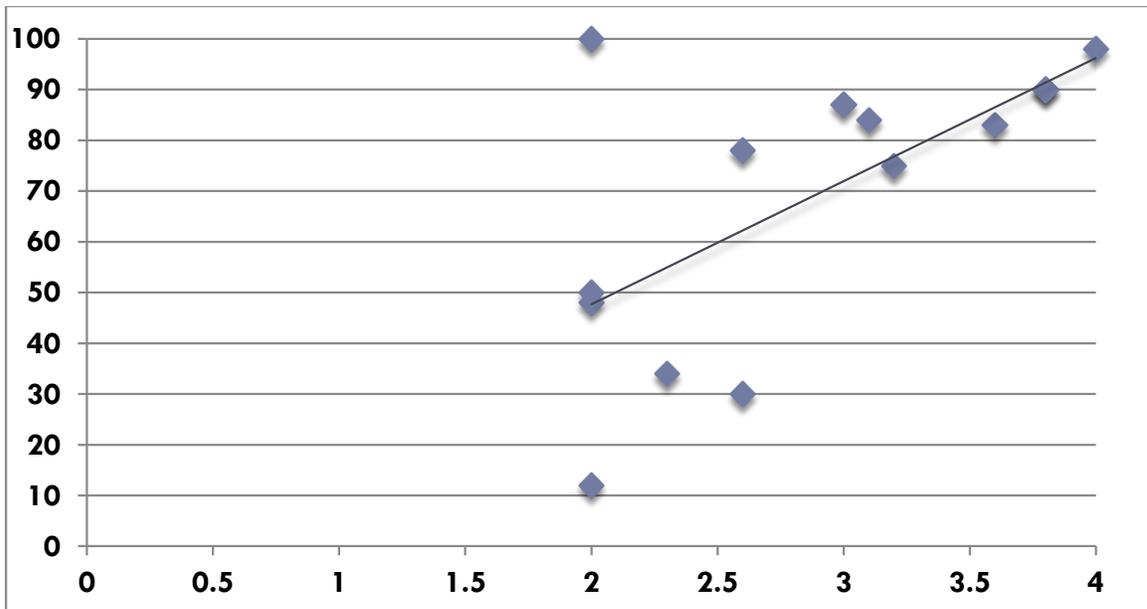
Strong

- The relationship between two variables is generally considered strong when their r value is larger than 0.7.

Correlations

Example: There is a moderate, positive, linear relationship between GPA and achievement motivation.

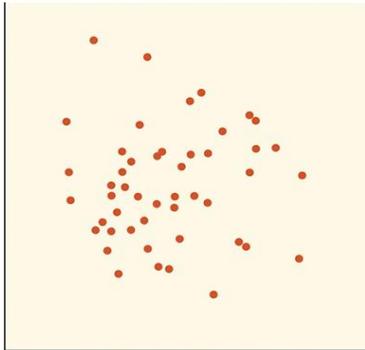
$$r = 0.62$$



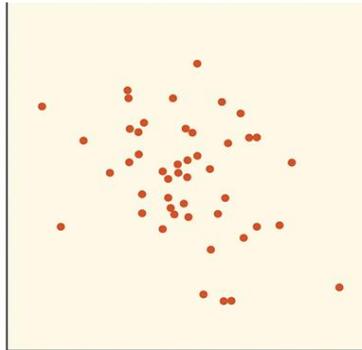
- Based on the criteria listed on the previous page, the value of r in this case ($r = 0.62$) indicates that there is a positive, linear relationship of **moderate** strength between achievement motivation and GPA.

Correlation

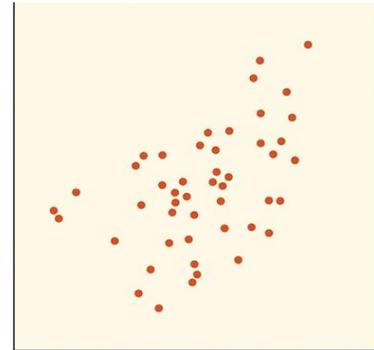
- The images below illustrate what the relationships might look like at different degrees of strength (for different values of r).



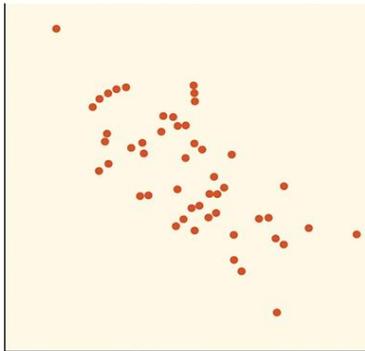
Correlation $r = 0$



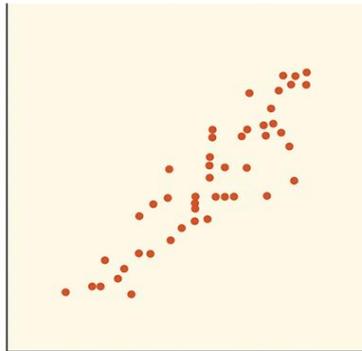
Correlation $r = -0.3$



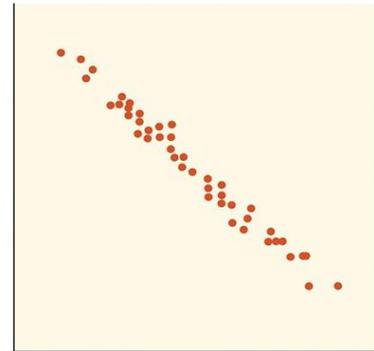
Correlation $r = 0.5$



Correlation $r = -0.7$



Correlation $r = 0.9$



Correlation $r = -0.99$

- For a correlation coefficient of zero, the points have no direction, the shape is almost round, and a line does not fit to the points on the graph.
- As the correlation coefficient increases, the observations group closer together in a linear shape.
- The line is difficult to detect when the relationship is weak (e.g., $r = -0.3$), but becomes more clear as relationships become stronger (e.g., $r = -0.99$)

Correlation Coefficients

The Statistical Significance of Correlation Coefficients:

- Correlation coefficients have a probability (p-value), which shows **the probability that the relationship between the two variables is equal to zero** (null hypotheses; no relationship).
- **Strong** correlations have **low** p-values because the probability that they have no relationship is very low.
- Correlations are typically considered statistically significant if the p-value is lower than 0.05 in the social sciences, but the researcher has the liberty to decide the p-value for which he or she will consider the relationship to be significant.
- The value of p for which a correlation will be considered statistically significant is called the **alpha level** and must be reported.
- SPSS notation for p values: Sig. (2 tailed)

In the previous example, $r = 0.62$ and $p\text{-value} = 0.03$. The p-value of 0.03 is less than the acceptable alpha level of 0.05, meaning the correlation is statistically significant.

Four things must be reported to describe a relationship:

- 1) The **strength** of the relationship given by the correlation coefficient.
- 2) The **direction** of the relationship, which can be positive or negative based on the sign of the correlation coefficient.
- 3) The **shape** of the relationship, which must always be linear to compute a Pearson correlation coefficient.
- 4) Whether or not the relationship is **statistically significant**, which is based on the p-value.

Facts About Correlation

- 1) The order of variables in a correlation is not important.
- 2) Correlations provide evidence of association, not causation.
- 3) r has no units and does not change when the units of measure of x , y , or both are changed.
- 4) Positive r values indicate positive association between the variables, and negative r values indicate negative associations.
- 5) The correlation r is always a number between -1 and 1.

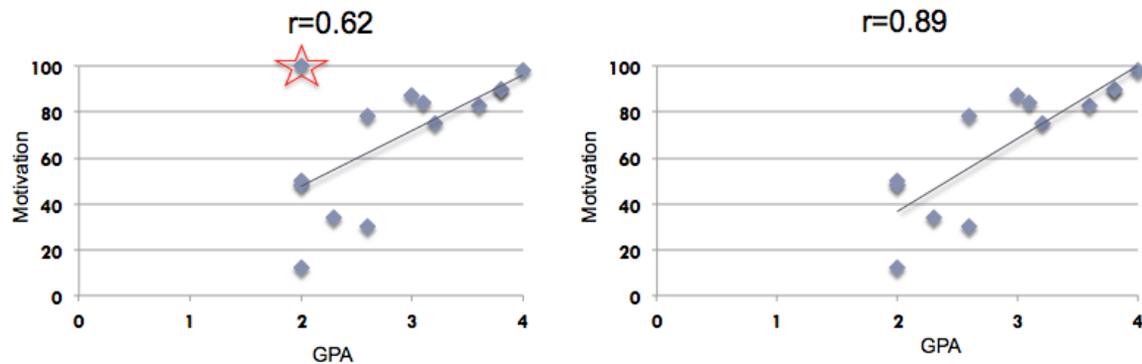
Pearson r : Assumptions

Assumptions:

- Correlation requires that both variables be quantitative.
- Correlation describes *linear* relationships. Correlation does not describe curve relationships between variables, no matter how strong the relationship is.

Cautions:

- Correlation is not resistant. r is strongly affected by outliers.
- Correlation is not a complete summary of two-variable data.
- For example:



- The correlation coefficient is based on means and standard deviations, so it is not robust to outliers; it is strongly affected by extreme observations. These individuals are sometimes referred to as *influential observations* because they have a strong impact on the correlation coefficient.
- For instance, in the above example the correlation coefficient is 0.62 on the left when the outlier is included in the analysis. However, when this outlier is removed, the correlation coefficient increases significantly to 0.89.
- This one case, when included in the analysis, reduces a strong relationship to a moderate relationship.
- This case makes such a big difference in this example because the data set contains a very small number of individuals. As a general rule, as the size of the sample increases, the influence of extreme observations decreases.
- When describing the relationship between two variables, correlations are just one piece of the puzzle. This information is necessary, but not sufficient. Other analyses should also be conducted to provide more information.