# Scatterplots and Correlation (Page 97-107, Chapter 4)

**TODAY YOU WILL BE ABLE TO…**

* Define explanatory and response variables
* Construct and interpret scatterplots
* Add categorical variables to scatterplots

A **response variable** measures an outcome of study.

An **explanatory variable** may explain or influence changes in a response variable.

The most useful graph for displaying the relationship between two quantitative variables, such as a response and an explanatory, is a **scatterplot.**

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

***Note:*** If there is no explanatory-response distinction between the variables, either one can be plotted on either axis.

**CONSTRUCT A SCATTERPLOT**

1. Decide which variable should go on each axis. If a distinction exists, plot the explanatory variable on the x-axis and the response variable on the y-axis.
2. Label and scale your axes.
3. Plot individual data values.

***Example 1:***

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  | Make a scatterplot of the relationship between body weight and pack weight for a group of hikers.  X: Body Weight (lb.)  Y: Backpack Weight (lb.)  Body weight “explains” backpack weight. | | | | | | | | |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  | X: | 120 | 187 | 109 | 103 | 131 | 165 | 158 | 116 |
|  |  |  |  |  |  |  |  |  |  |  | Y: | 26 | 30 | 26 | 24 | 29 | 35 | 31 | 28 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

**INTERPRET A SCATTERPLOT**

To interpret a scatterplot, look for patterns and important departures from those patterns.

* You can describe the overall pattern of a scatterplot by the **form**, **direction**, and **strength** of the relationship.
* You may also find one or more **outliers**, individual values that fall outside the overall pattern of the relationship.

**Form**

The form of the relationship between body weight and backpack weight is linear, i.e, you can draw a straight line through the data with most points lying fairly close to the line. Outliers fall outside this trend. There are two points that appear to be possible outliers.

**Direction**

There are two possible directions that data with a linear trend can follow:

*Positive direction:*

Two variables are positively associated when above-average values of one tend to accompany above-average values of the other. Likewise, below-average values of one tend to accompany below-average values of the other. As values of one variable **increase**, values of the other also **increase**, and vice versa.

In the backpack weight data, as body weight **increases**, backpack weight also **increases**. People with above-average weight tend to carry backpacks that are also above average in weight. Backpack weight has a positive relationship with body weight.

*Negative direction:*

Two variables are negatively associated when above-average values of one tend to accompany below-average values of the other, and vice versa. As values of one variable **increase**, values of the other **decrease**.

**Strength**

The relationship between two variables can be weak, moderate, or strong. The strength is determined by how closely the points follow the trend line.

Backpack weight and body weight have a **strong**, **positive**, **linear** relationship. Only two points are potential outliers, but they still follow the trend fairly closely.

**FORMS OF SCATTERPLOTS**

**Linear**



**Clusters**



**Curved**



**ADD CATEGORICAL VARIABLES TO SCATTERPLOTS**

You can distinguish between different categorical groups by using different symbols or colors for each group on the plot. For example, you could visualize differences in the relationship between body weight and backpack for men and women by plotting men with a circle and women with a triangle or by plotting men with green and women with yellow symbols.

# Scatterplots and Correlation (Page 108-114, Chapter 4)

**TODAY YOU WILL BE ABLE TO…**

* Calculate and interpret correlation
* Describe facts about correlation

**NOTATION**

x the *variable* usually used to represent the individual data values of the explanatory variable

y the *variable* usually used to represent the individual data values of the response or outcome variable

r represents the correlation between two quantitative variables, x and y

**WHAT IS CORRELATION?**

* The **correlation**, *r,* measures the strength and direction of a *linear* relationship between two *quantitative* variables.
* Correlation makes no distinction between explanatory and response variables. The statements, “x is related to y” and “y is related to x,” are the same because there is only one correlation between the two variables.
* The correlation has no units so the value of r does not change if the units of measurement of x and/or y change.
* Correlation is affected by outliers. Any calculation that involves the actual data points is influenced by extreme values.

**CALCULATE THE CORRELATION**

**INTERPRET THE CORRELATION**

* *r* is always a number between -1 and 1.
* Values of *r* near 0 indicate a very weak linear relationship. (1)
* *r* > 0 indicates a positive association. (2), (3)
* *r* < 0 indicates a negative association. (4), (5), (6)
* The strength of the linear relationship increases as *r* moves away from 0 toward -1 or 1. Figure (3) shows a stronger linear relationship than (2). Figure (6) shows a stronger linear relationship than (5) or (4).
* The extreme values *r* = -1 and r = 1 occur only in the case of a perfect linear relationship. Figure (6) is shows a relationship that is very close to perfect.

|  |  |  |
| --- | --- | --- |
| F3.06.jpg | F3.06.jpg | F3.06.jpg |
| (1) r = 0 (weak) | (2) r = 0.5 (moderate) | (3) r = 0.9 (strong) |

|  |  |  |
| --- | --- | --- |
| F3.06.jpg | F3.06.jpg | F3.06.jpg |
| (4) r = -0.3  (weak) | (5) r = -0.7  (moderate toward strong) | (6) r = -0.99 (very strong) |