Lesson 17: Scatterplots

Daily Data Collection

Foot length and forearm length

A scatterplot shows the relationship between two quantitative variables measured on 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 in the plot fixed by the values of both variables for that individual.

A **response variable** measures the outcome of a study. An **explanatory variable** helps explain or influences change in a response variable.

You will often find explanatory variables called *independent variables*, and response variable called *dependent variables*. The idea behind this language is that the response variable *depends* on the explanatory variable. Because the words independent and dependent have other, unrelated meanings in statistics, we won't use them here.

Interpreting a Scatterplot

Direction

- Is the graph moving up or down as it moves from left to right?
- Rising is called a positive association.
- Falling is called a negative association.

Form

- Is the pattern linear or does it follow another type of function?

Strength

- How close does the data follow the given function/form?
- Are there outliers or striking deviations from the pattern?

Correlation

In order to strengthen the analysis when comparing two variables, we can attach a number, called the correlation coefficient (*r*), to describe the linear relationship between two variables. This number helps remove any subjectivity in reading a linear scatterplot and helps us avoid being fooled by axis manipulation.

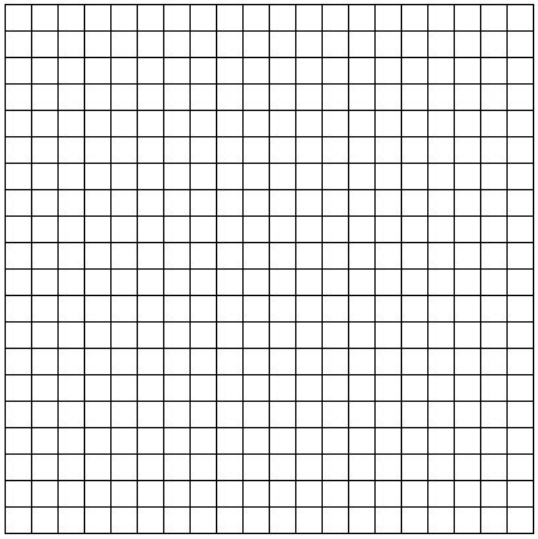
The correlation measures the strength and direction of the **linear** relationship between two quantitative variables:

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

Essentially, the correlation coefficient, r, finds the average of the product of the standardized scores.

Class Data:

Create a scatterplot.
Explanatory Variable: Foot length
Response variable: Forearm length



Describe the Direction

Describe the Form

Describe the Strength, including the correlation

Conclusion/Analysis

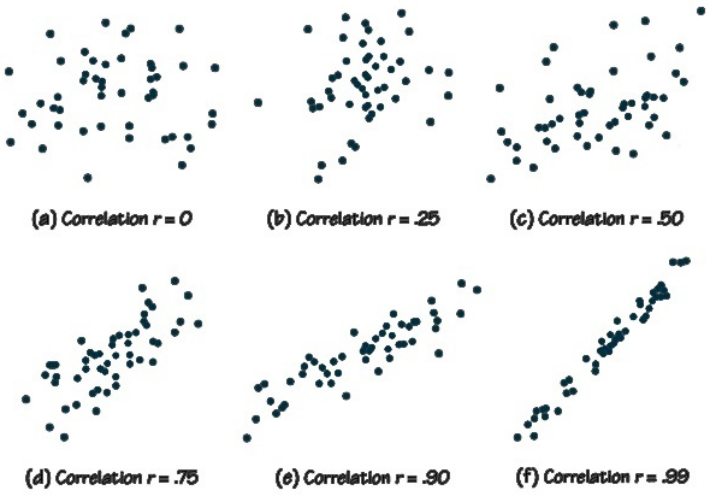
Facts about correlation

- Between -1 and 1
- 1 is a perfect positive association. As the explanatory variable goes up, the response variable goes up.
- -1 is a perfect negative association. As the explanatory variable goes down, the response variable goes down.
- 0 means there is NO association.
- Correlation does not imply causation. A strong correlation does not guarantee a cause and effect relationship.
- Correlation does not have units.
- Correlation requires both variables to be quantitative data.
- Correlation is affected by outliers.
- Correlation only describes linear relationships.

Value of r	Strength of relationship
-1.0 to -0.5 or 1.0 to 0.5	Strong
-0.5 to -0.3 or 0.3 to 0.5	Moderate
-0.3 to -0.1 or 0.1 to 0.3	Weak
-0.1 to 0.1	None or very weak

The following scatter plots show other types of relationships that might occur between two variables:
Describe the Direction, Form, and Strength:

Examples of Correlation:

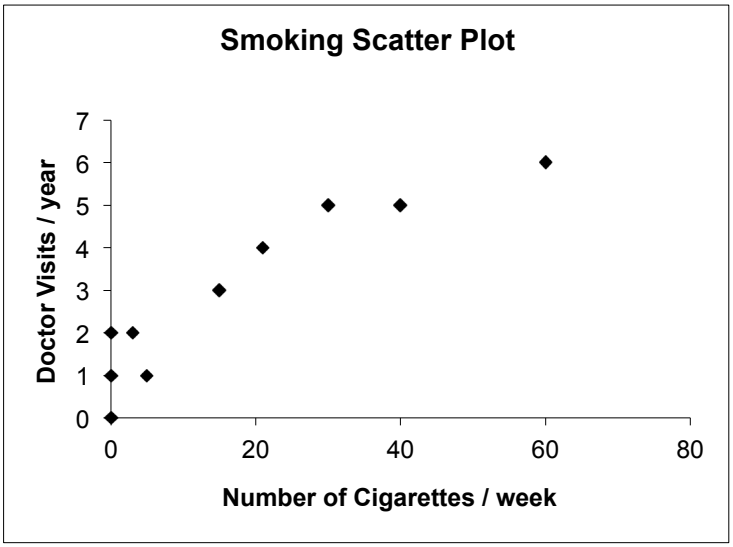


Example:

Suppose we hypothesize that the number of doctor visits a person has can be explained by the amount of cigarettes they smoke. So we want to see if there is a relationship between the number of cigarettes one smokes a week (the explanatory or independent variable) and the number of times per year one visits a doctor (response or dependent variable). We ask 10 random people and get the following information:

# of Cigarettes Per Week	0	3	21	15	30	5	40	60	0	0
Number of doctor visits per year	1	2	4	3	5	1	5	6	2	0

If we were to plot the following data point on an X-Y coordinate plane, we would get a scatter plot that looks like this:



Note that the graph shows us that as you smoke more cigarettes / week, you also tend to go to the doctor more often. This result demonstrates that there is a positive relationship between the two variables.

HW 17 - Section 3-1: 1,5,7,11,13