Lecture 18

Association between two variables

The main purpose of a data analysis with two variables is to investigate whether there is an **association** and to describe the nature of that association. An association exists between two variables if a particular value for one variable is more likely to occur with certain values of the other variable.

Example:
When there's an association, the likelihood of a particular value for one variable depends on the value of the other variable. The chance of a college GPA above 3.5 is greater for those with high school GPA=4.0 than for those with high school GPA=3.0. So high school GPA and college GPA have an association.

Looking for a trend: The Scatterplot

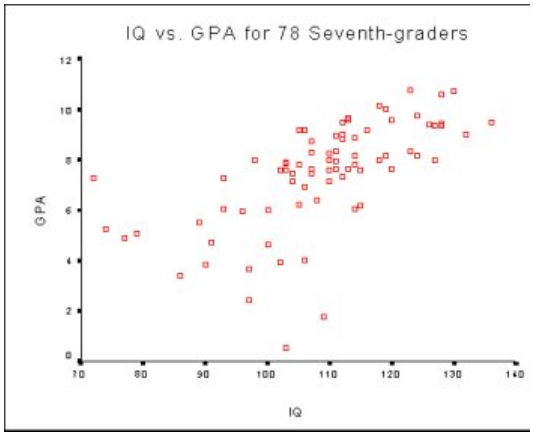
With two quantitative variables it is common to denote the response variable by y and the explanatory variable by x. A **scatterplot** is a graphical display for two quantitative variables. It uses the horizontal axis for the explanatory variable x and the vertical axis for the response variable y. The values of x and y for a subject are represented by a point relative to the two axes. The observations for the n subject are n points on the scatterplot.

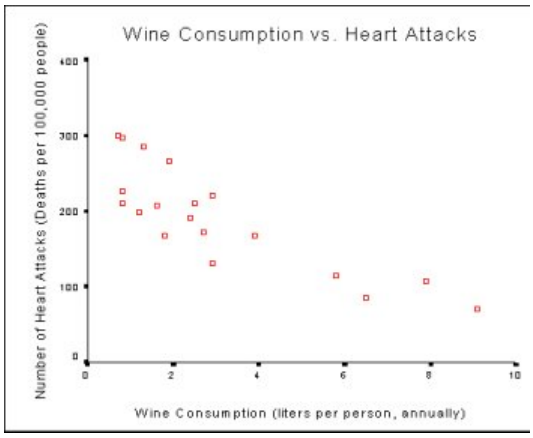
Questions to answer regarding scatterplots:

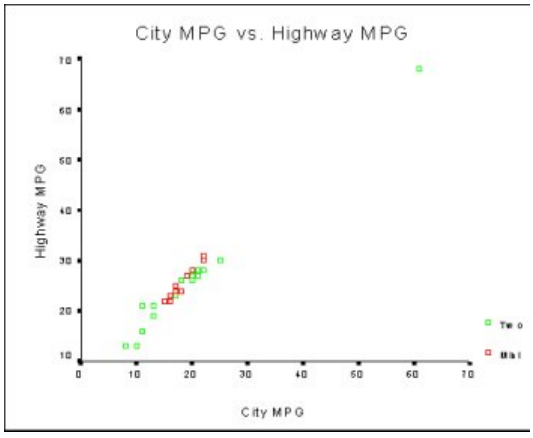
Form: Are all the points in one cluster or multiple clusters?
Direction: Is there evidence of a linear (straight-line) association? If so, in which direction (positive or negative)?
Strength: How close are the points to the line or curve? The closer the points are to the line or curve, the stronger the association is.
Outliers: Are there any values that are far away from the rest?

Two quantitative variables have a **positive association** when high values of x tend to occur with high values of y, and when low values of x tend to occur with low values of y.

They are said to have a **negative association** when high values of one variable tend to occur with low values of the other variables, and when low values of one variable tend to occur with high values of the other variable.







Summarizing the association through the correlation

When the data points follow roughly a straight line trend, the variables are said to have an approximately **linear** relationship. In some cases the points fall close to a straight line, but more often there is quite a bit of variability of the points around the straight-line trend. A summary measure called the **correlation** describes the strength of the linear association.

We denote the correlation by r and it takes values between -1 and +1.

- A positive value for r indicates a positive linear association between the two variables while a negative value of r indicates a negative linear relationship between the two variables.

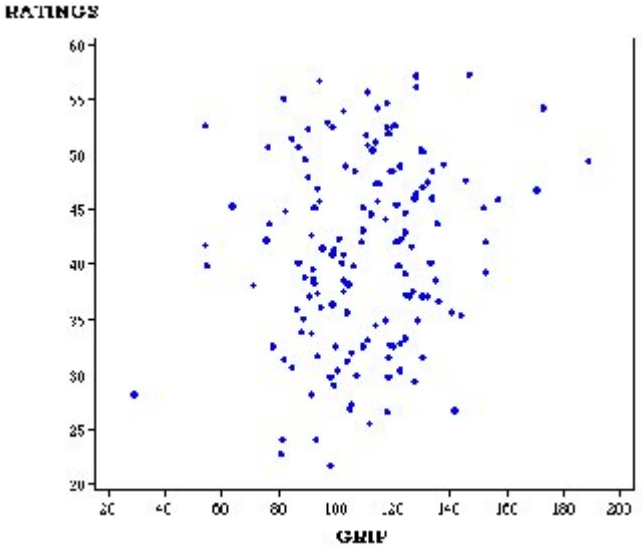
- The closer r is to ± 1, the closer the data points fall to a straight line, and the stronger the linear association. The close r is to 0, the weaker the linear association.

- There is no need to make a distinction between the explanatory variable (x) and the response variable (y)

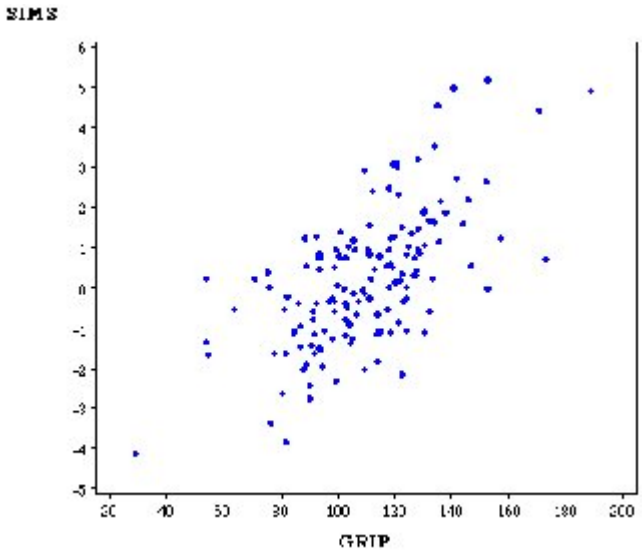
- Correlation is sensitive to outliers just like the mean and standard deviation.

Example: A study by Blakely, Quinones, and Jago (1995) published in the journal *Personnel Psychology* examines the association between physical strength and job ability in people with labor intensive jobs. With each person four different variables were measured. The two variables used to measure strength were grip strength and arm strength. The two variables used to measure job ability were supervisor's rating and score on a simulation test. Below are two scatterplots representing the association between the variables.

$r = 0.18$



$r = 0.64$



To calculate the correlation

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x - \bar{x}}{\sigma_x} \right) \left(\frac{y - \bar{y}}{\sigma_y} \right) \tag{1}$$

Before you calculate the correlation you should always graph the data to see if the correlation is an appropriate measure. Correlation only measures the linear association between variables.

Example:
Let x be age of a person and y be annual medical expenses. What do you think the scatterplot of these two variables would look like?

This relationship is not linear (it's quadratic) so the correlation would be near zero but that doesn't mean that the variables are not associated it just means they are not linearly associated.