Objectives:

- To explore the basic properties of the ***correlation coefficient*** as a measure of the degree of association between two variables.
- To discover some of the limitations of the ***correlation coefficient*** as a summary of the relationship between two variables.
- To recognize the important distinction between the ***association*** and ***causation***.
- To become familiar with judging ***correlation*** values from scatterplots
- To learn to use scatterplots and correlations to look for and to describe relationships between variables when analyzing genuine data.

Correlation coefficient denoted by ***r***, is a measure of the degree at which two variables are associated. (This is not a calculation you would want to do by hand - if you thought Standard Deviation was bad you should see this)

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

where x_i denotes the i^{th} observation of the dependent variable and y_i denotes the i^{th} observation of the independent variable, x and y the respective sample means and, s_x and s_y the respective sample deviations, and n the sample size.

A problem with the correlation coefficient is that it only illustrates the linear relationship between two variables. If there is a curvilinear relationship than you need to use a different calculation.

Association vs. causation A very common mistake made by people in statistics is to assume that when there is a relationship that means that one of the variables caused the other. *e.g.* When Ice cream sales increase, so does the murder rate - one might make the assumption that ice cream causes people to murder - does this make sense. There is an association but no causation (that I know of anyway)

If there is an association without causation than there is more than likely a third variable that both are related to and causes both in the above example it would be that they both increase in the summer, The hot weather caused both. Another is older women that have babies live longer. A reasonable explanation of the increased life expectancy: Since the woman was able to have a baby late in life shows that the reproductive system is in good health longer which is an indication that the rest of the woman's body is in good shape and therefore will live longer - both are caused by the healthfulness of the woman involved.

The third variables in the above examples would be called ***confounding*** or ***Lurking variables***