7 Chapter 3 - Describing Bivariate Data

1. Motivation - Does the knowledge of the values of one variable help you understand the distribution of values of another variable? How do you measure and assess this association?
   (a) Economic theory tells us that price demanded falls as the available quantity of a good increase. Supply of a good increases as prices rise. Increasing the income of consumers, for some goods, can lead to an increase in price demanded for a good.
2. Discussion on causality - does the association between two variables always imply causality?
3. Independent and dependent variables.
   (a) Dependent variable - variable of interest - variations in the variable described by exploratory (independent) variable(s). When theory describes a casuality between two variables, values found in the dependent variable are heavily influenced ("caused") by the exploratory or independent variables.
   (b) Independent variable(s) - values found in an independent variable are not perceived to be influenced by other variables.
4. Does every association between variables imply that there exists an independent and dependent variable - **No, No, No**.
5. Graphical presentation of bivariate data
   (a) Comparative bar chart (All variables)
   (b) Comparative pie chart (All variables)
   (c) Stacked bar chart (All variables)
   (d) Scatterplot (Interval variables)
      i. The type of graphs used in economic textbooks are scatterplots.
   (e) Tools to analyze the values in a scatterplot
      i. Is there a pattern? (downward sloping pattern, upward sloping pattern, other patterns)
      ii. Clustering of the points along the pattern
      iii. The presence of outliers
6. Numerical Measures of Bivariate data (Interval variables x,y)

(a) Correlation coefficient

i. Formula

$$r = \frac{s_{xy}}{s_x s_y}$$

where

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

ii. Range of r $\rightarrow$ -1 to 1

iii. Interpretation of r $\rightarrow$ 1 implies a very strong, positive correlation (scatterplot would show an upward sloping pattern with a tight clustering of points along the pattern). -1 implies a very strong, negative pattern. 0 implies no pattern. **Note: the value of r is not a measure of the slope.**

7. Using the correlation value to estimate the slope and intercept of the line that best represents the linear relationship between two variables.

Best fitting line is called the regression line. The best fitting line is the line that gives the minimum distance between the points of data and the line. The econometric course will examine how to produce the best fitting line under different circumstances.

Let's take b to be the slope of the line and a as the intercept value (the value where x=0). The first variable is x and its mean is $\bar{x}$. The second variable is y and its mean is $\bar{y}$. The formulas for b & a are:

$$b = r\left(\frac{s_y}{s_x}\right)$$

$$a = \bar{y} - b\bar{x}$$