Chapter 2 Looking at Data--Relationships

In the previous chapter, we described the distribution of the individual variables of a data set. Suppose we want to determine if there is a relationship between two variables measured on the same group of individuals. The first step is to create a table or graph that will display the relationship (provided one exists) between the variables.

There are three cases to consider:

1. One categorical and one quantitative variable
2. Two quantitative variables
3. Two categorical variables

To display, a relationship between one categorical variable and on quantitative variable consider using side-by-side box plots or back-to-back stem-n-leaf plots from chapter 1. We will look at two categorical variables in section 2.5.

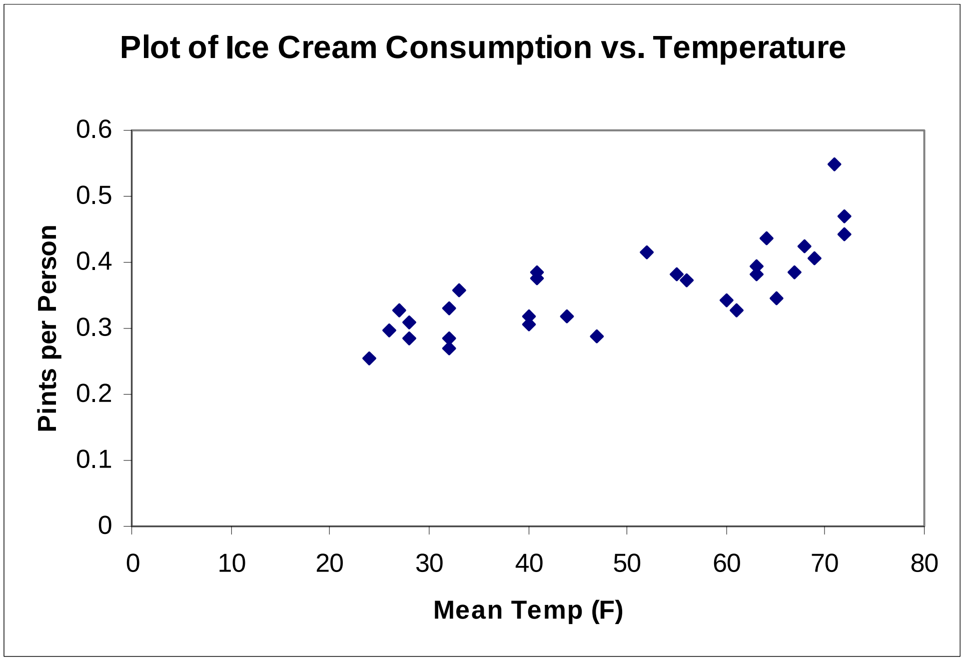
To investigate the relationship between two quantitative random variables, we will start by labeling the variables as the explanatory variable and the response variable. The response variable measures the outcome of a study and the explanatory variable explains the changes in the response variable. The explanatory variable is typically denoted by X while the response variable is denoted by Y. Then we will create a scatterplot of the data with X on the horizontal axis and Y on the vertical axis.

We will use the scatterplot to provide insight into whether there is an association between the two variables. If there is an association, then knowing something about the value of one variable will tell us something about the value of the other variable.

Example: Ice Cream Consumption Suppose we are thinking of opening an ice cream business and we want to know whether there is a relationship between the amount of ice cream consumed and the temperature. We decided to collect data over a 30-week time period from March to July. For each week, we recorded the average amount of ice cream consumed (per person) as well as the mean temperature. The data are presented on the last page of the handout.

Notice the data is paired. For each of the 30 weeks, we have two pieces of information, ice cream consumption and temperature. We would like to determine if there is a pattern or predictable relationship that occurs between the two variables? In other words, as temperature increases, what happens to ice cream consumption (does it increase, decrease, or stay the same)?

If we do see a pattern that is strong enough, we can build a model from our sample data that allows us to make predictions about the population, and help our ice cream business be more successful.



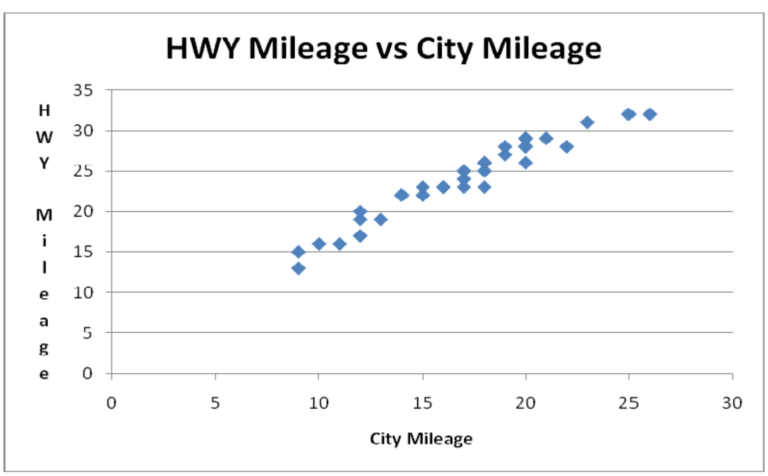
Looking at the data, do you see a relationship between temperature and ice cream consumption? If so, describe this relationship.

Types Relationships between Quantitative Variables

Two quantitative variables can be related to each other in a number of ways. Some variables have a linear relationship; others have relationships that are best described as non-linear.

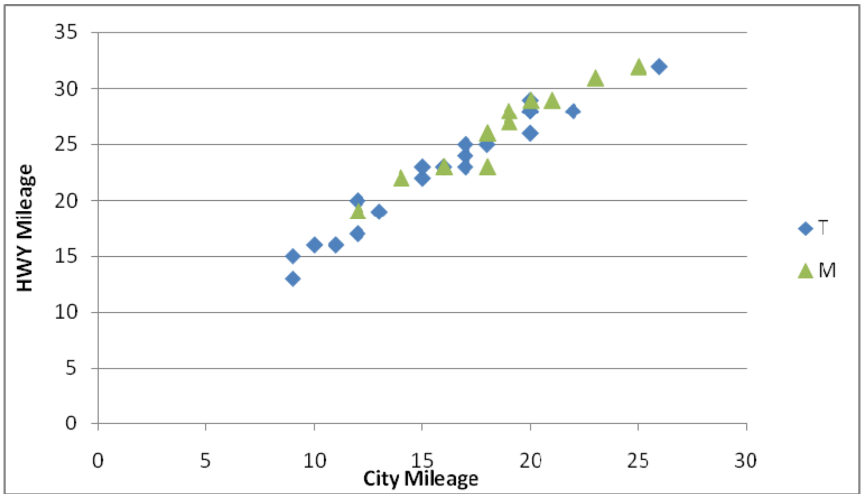
Variables that have a non-linear relationship might be related through a simple curve (such as the number of fruit flies that multiply over time, or the relationship might be much more involved, requiring more complicated functions to describe (such as stock market prices over time.)

Example City and Highway Gas Mileage (problem 2.16): Table 10 page 31 gives the city and highway mileage for minicompact and two-seater cars. Is there a relationship between city and highway mileage?



Is there a relationship between highway and city mileage?

Look at the scatterplot between highway and city mileage using different symbols to indicate type (two-seater (T) or minicompact(M))



Measuring the Strength of the Linear Relationship

A scatterplot can help give us a general idea as to whether or not there is a linear relationship between two variables, how strong the relationship appears to be in the sample, and what the direction of the relationship is.

Direction of the Linear Relationship. If the pattern goes uphill, we say the linear relationship is positive (both variables increase together or decrease together.) For example, height and weight of male adults exhibit a positive linear relationship. If the pattern goes downhill, we say the linear relationship is negative (as one variable increases, the other decreases. For example, we hope that as the number of police officers increases, the number of crimes decreases.

Strength of the Linear Relationship. The strength of the linear relationship is a measure of how close the pattern of observed values resembles a straight line. If the data points line up perfectly, we say there is a perfect linear relationship. If the points lie quite close to the line overall, we say the relationship is strong. If the points don't have too much of a pattern, yet seem to resemble a cloud going uphill, the relationship is weak. If the points are scattered everywhere (or in cases where a different type of pattern exists) we say there is no linear relationship.

Section 2.2 Correlation

To measure the strength and direction of the linear relationship between two quantitative variables we will calculate **correlation coefficient,** denoted by r.

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

where $\bar{x}$ and s_x are the mean and standard deviation of the *x*-values and $\bar{y}$ and s_y are mean and standard deviation of the *y*-values.

The Correlation Coefficient has the following properties:

- The correlation coefficient is "unitless." It does not depend on the variables examined.
- The correlation coefficient always lies between -1 and 1.
- Correlations of -1 and +1 indicate perfect negative and perfect positive linear relationships, respectively.
- The closer the correlation coefficient lies to either -1 or +1, the stronger the linear relationship is perceived to be. (A correlation coefficient at or beyond +/- 0.60 is considered meaningful to most researchers.)
- A correlation coefficient of 0 indicates no linear relationship. (Note: there could still be another type of relationship present.)
- Not resistant to outliers