Objectives:

- To learn to produce a ***two-way table*** as a summary of the information contained in a pair of categorical variables
- To develop skills of interpreting information presented in a two-way tables of counts
- To become familiar with concepts of ***marginals*** and ***conditional distributions*** of categorical variables
- To discover the ***segmented bar graph*** as a visual representation of onformation contained in such tables
- To explore the and understand the concepts of ***independence*** and ***relative risk***
- To gain experience in applying techniques of analyzing two-way tables to genuine data

Response variable and explanatory variable- Response variable is affected and/or predicted by the explanatory variable (response can be thought of as y, explanatory can be thought of as x.)

Two-way Tables: called two way tables because it categorizes each person (or object) according to two variables *e.g.* political affiliation and age. We use this to investigate the relationship of the two variables. This is similar to regression but since these are categorical variables we group the like occurrences (regression variables are continuous and there for we can use a linear equation to describe the relationship)

Marginals: When you find the totals of each individual variable it is called the marginal for that variable (total number of conservatives, total number of 30-50 year-olds etc.) To find the ***marginal distribution*** we find the proportion of the total that are included in a category (25% of subjects are conservative, 30% are between 30-and 50)

Conditional distributions: To find relationships between variables we find the proportions of subjects that are explained by each combination of the two variables. We must do this first before we can investigate any relationship between two variables. (20% of older people are liberal, 80% are conservative)

In a conditional distribution, it is important to note the direction of the conditional. For example, what is the difference between these proportions:

The proportion of females that are democrat

The proportion of democrats that are females

BIG DIFFERENCE!!!

Segmented bar graph: used to visually represent conditional distributions. A bar graph that uses equal sized boxes to represent the total (100%) of each category, these boxes are then broken up into proportional sized rectangles to represent the proportion of each category that fall into a certain value. The conditional distribution yields the percentages used in the construction of the segmented bar graph.

Simpson's Paradox: The phenomenon when each category may be higher percentage for one variable value but the aggregate percentage is higher for the other variable. For example, can you come up with an example of a batter in baseball having a higher batting average in both the first and second halves of the season but another batter having a better total batting average?

Independent variables: two variables are said to be independent if the conditional distributions of each are identical for every category. *e.g.* No matter what age group the proportion of democrats for that age group is 30%.

Association: If variables are not independent then we say there is an association between the two variables.

Relative risk: The relative risk is the ratio of the proportions having the disease between the two groups of the explanatory variable. For example in the book they give the example on page 145, AZT vs. Placebo pregnancies. The relative risk in this case is the number of times more at risk a placebo baby is to an AZT baby.