Lesson 4-3 Relations in Categorical Data

***A two-way table*** of counts describes the relationship between two categorical variables. Values of the **row variable** label the rows of the table, and the values of the **column variable** label the columns. Two-way tables are often used to summarize large amounts of data by grouping outcomes into categories.

The row totals and the column totals give the **marginal distributions** of the two variables separately. They do not give any information about the relationship between the variables.

To find the **conditional distribution** of the row variable for one specific variable of the column variable, look only at that one column in the table. Find each entry in the column as a percent of the column total. There is one conditional distribution of the row variable for each column in the table. Comparing these conditional distributions is one way of showing the association between the row and the column variables. We generally use bar graphs to display the relationships between the variables. Pie charts can also be used.

A comparison between two variables that holds for each individual value of a third variable can be changed or reversed when the data for all the values of the third variable are combined. This is **Simpson’s paradox.**

# What can go wrong?

* Make sure that the areas represented in your graphs actually represent the relationship between variables.
* Make sure the total percents are 100% (with perhaps some rounding error).
* Don’t use unfair or silly averages. Sometimes averages are not appropriate and sometimes categories don’t make sense.
* Read the conditions of a distribution carefully.