Relationships between two variables

Notes

When we have two variables **measured on the same individuals**, we can ask how they interact with each other:

> We say that the two variables x, y are **associated**, or **related**, if certain values of the one variable tend to appear with particular values of the other variable.

For instance, height and weight tend to be related, as taller people will also tend to have more weight. This does not mean that there are no shorter people with more weight than one might expect, just that there is a certain trend.

As another example, suppose 60% of female students are in greek life, while only 45% of male students are in greek life. Then we could say that the variables “Greek Status” and “Gender” are associated, since the “Female” value of the “Gender” variable tends to show a higher preference for being paired with the “Greek” value of the “Greek Status” variable than the “Male” value of the “Gender” variable does.

In general, depending on the types of the three variables, we use different means to determine whether they are related or not.

Categorical - Categorical If both variables are categorical, graphically we would use “100% stacked bar graphs”. Numerically we would use “row or column percentages” in a cross-tabulation table.

Quantitative - Categorical If one variable is categorical and another scalar, graphically we would use “box plots”. Numerically we would try to compare summary statistics computed separately for each separate value of the categorical variable.

Quantitative - Quantitative If both variables are scalar, graphically we would use “scatterplots”. Numerically we can discuss regression lines or other bivariate techniques.

Categorical - Categorical

Here is an example of looking for a relationship based on a stacked bar graph. We are comparing car types and number of cylinders. Notice that some car types have higher percent of specific numbers of cylinders than others. For instance, about 40% of standard cars have 4 cylinders, while only 25% or so of the sports cars have 4 cylinders. This is an indication that the two variables are associated.

Here is a numeric table that will tell us the same percentages that the graph does:

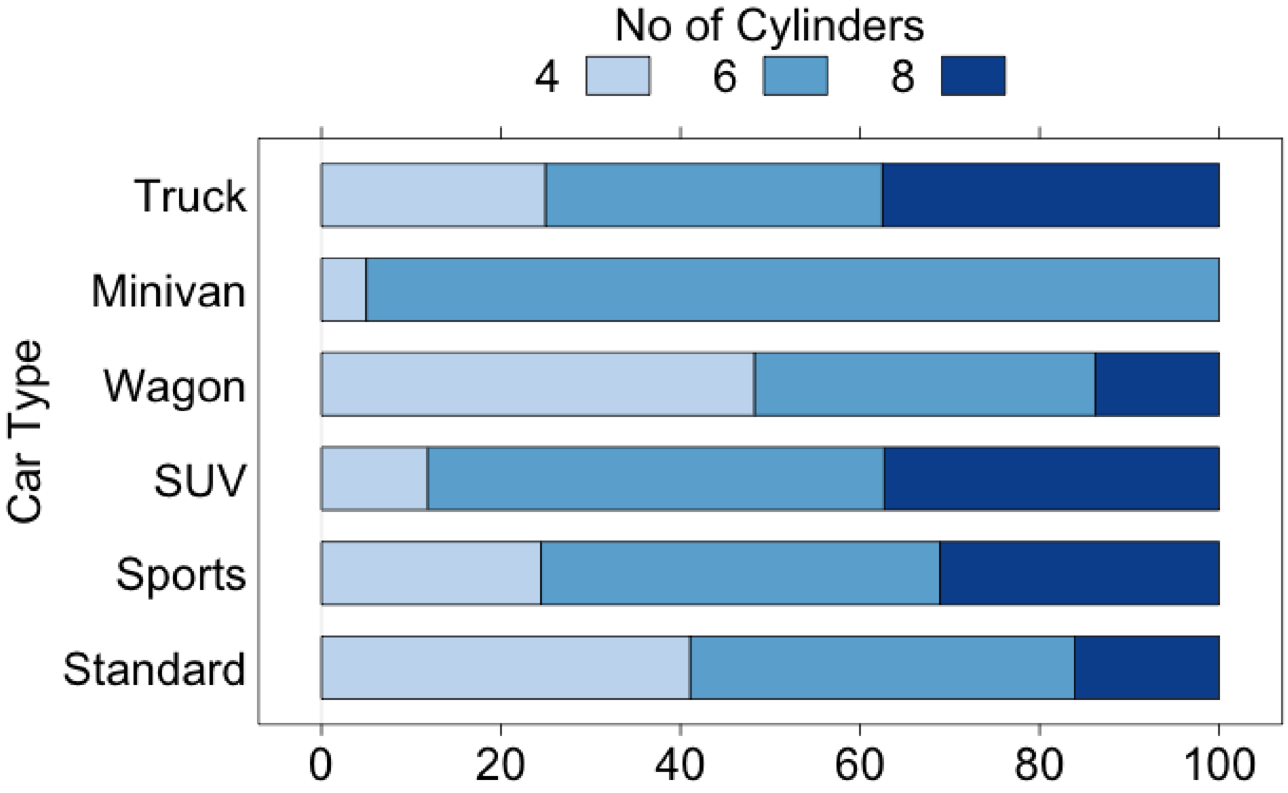

Figure 1: Example of a Stacked Bar Graph

Car Type	4	6	8
Standard	41.1%	42.8%	16.1%
Sports SUV	24.4%	44.4%	31.1%
Wagon	11.9%	50.8%	37.3%
Minivan	48.3%	37.9%	13.8%
Truck	5.0%	95.0%	0.0%
	25.0%	37.5%	37.5%

Each row's percentages add up to 100%. What we are looking for is sufficiently different distribution of the percentages across each column.

Categorical - Scalar

When comparing categorical and scalar variables, a good first step is a boxplot:

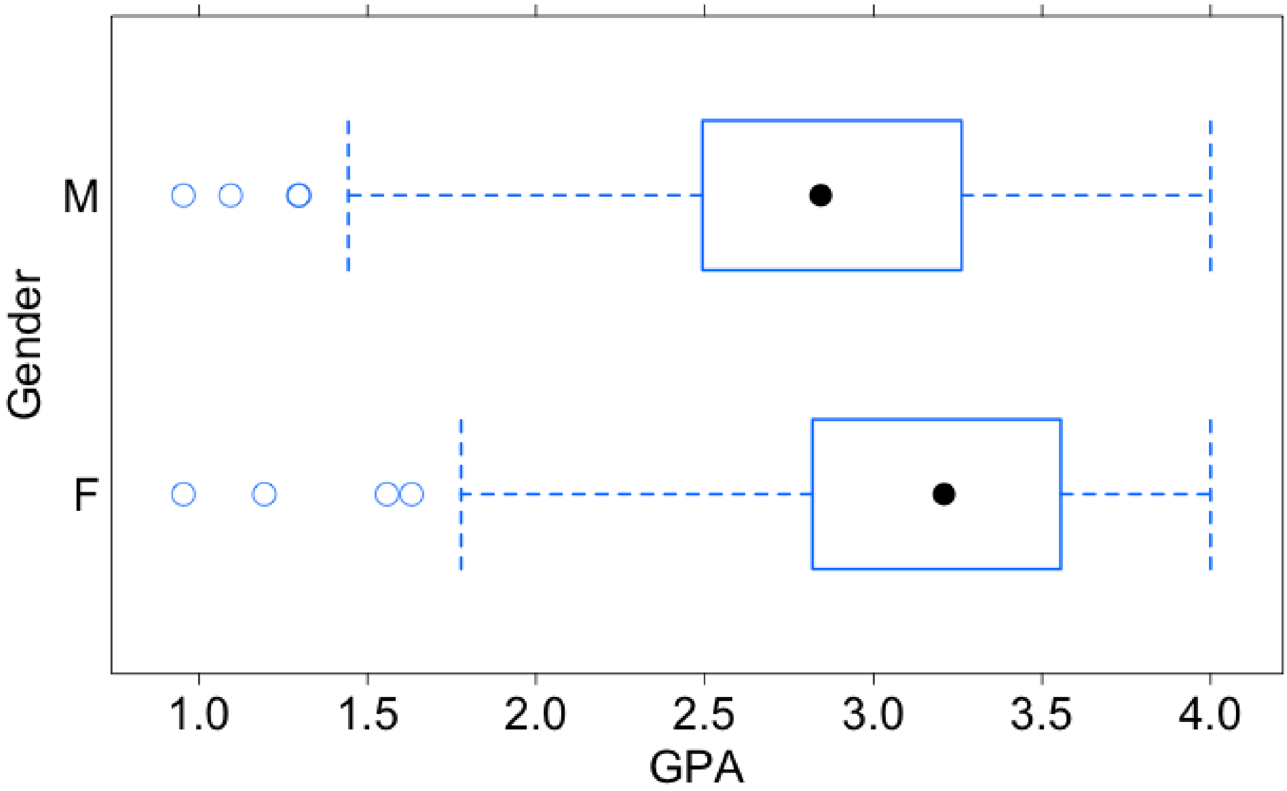

Figure 2: Example of a Boxplot

This compares the gender of Hanover students against their GPAs. What we can see here is that even though both sets of students have a wide range of values, the GPAs for female students tend to concentrate a bit higher than the GPAs for the male students. In fact the medians are about 0.3 units apart. So there appears to be some relation between a student's gender and their GPA.

Even though this difference might appear to be small, it comes from fairly large samples, which tends to make it more considerable. We will discuss these issues more extensively later.

Along with the boxplot, some numerical summaries are helpful:

Gender	Q1	Median	Q3	IQR	Mean	Std. dev.
Female	2.82	3.21	3.56	0.735	3.15	0.511
Male	2.49	2.84	3.26	0.768	2.86	0.554

Scalar - Scalar

We will discuss this extensively in the next section.