Lab 3: Relationships between variables

We will explore in this lab how to use SPSS to inspect the relationships between variables.

As we saw in class, the kind of analysis we do depends on the types of variables we have:

1. If both are categorical, we look at 100% stacked bar graphs, and cross-tabulations.
2. If one is categorical and one scalar, we look at boxplots, as well as separate numerical summaries for each category of the categorical variable.
3. If they are both scalar, we use scatterplots, and we will also learn in class some numerical techniques that we can use.

So let us get started

Categorical - Categorical

We will discuss here how to deal with the situation where both variables are categorical. For this example we will use the cars dataset again, and we will focus on the two variables: Car Type and Cyl.

100% stacked bar graphs

We start by learning how to make stacked bar graphs.

- Go to “Graphs -> Chart Builder”
- Select the “Bar” section at the bottom left, then double-click the third graph, with the stacked bars.
- You should see, in addition to x and y-axis spots, a “Stack set color” spot on the top right.
- It is really important that you do NOT do anything to the y-axis. Instead, use the “Stack set color” place for the variable that corresponds to the different stacks.
- Add one variable to the x axis and the other to the stack set color spot. In this case, we want Car Type in the x axis and cylinder number in the stack set color spot.
- Click OK. This gives you a stacked bar graph. It is not yet 100%.
- Double-click the graph to edit it. Then choose “Options -> Scale to 100%”.
- In the interest of making the graph a bit more readable, we decide to remove some categories. You should consider whether this is a good idea or not in each case; we mostly do this here so you know how:
  - Click one of the bars. The “properties” window that pops on the right should have a “Categories” tab.

  - Make sure Cyl is the selected variable there. You will see a lot of values in the “Order” box.
  - You can use that area to either rearrange those values, or remove them altogether.
  - Select the number 3 and press the x. Repeat for 5, 10 and 12, leaving us with just 4, 6, 8. Hit Apply to see the effect.
  - Change the variable selected in the Categories tab to Car Type. Remove Minivan and Truck from the list in the same way, and hit Apply.
  - Rearrange “Wagon” so that it is next to Standard.
- These are some of the ways you can use to rework stacked bar graphs to better illustrate the point you are trying to make.
- Before we move on, try to change the colors of the 3 bars by selecting one of the colors in the legend, then looking at the “Fill&Border” tab in the Properties window.

Close the Chart Editor window, and take a moment to think about what conclusions we can draw from this graph. The way to read it is that each bar represents the totality of cars of a given type, and the sizes of the individual bars indicate *what percent of cars of that type have the given number of cylinders*. We can tell from this for example, that while 4 cylinders is fairly typical amongst Standard cars models, it is not typical amongst Sports cars and SUVs. We can translate that into saying that if we know a car is say SUV, then we know that it is less likely to have 4 cylinders. In other words the two variables are fairly related.

When you make 100% stacked bar graphs, you have to always choose which variable goes to which side. This matters because you can ask different questions:

1. The way it is set up right now, we can ask for example what percent of standard cars have 6 cylinders. You should be estimating that to be around 40%.
2. Try to do the graph the other way around, switching the locations of the two variables in the Chart Builder. Don’t forget to scale it to 100%.
   - In this graph you can read for example about the percent of 6-cylinder cars that are standard. You should find that to be close to 50%.
   - Make sure you understand how these two are very different questions. It will be important later on.

Crosstabulation

A cross-tabulation is essentially a frequency table that has two one variable in the rows and one in the columns, and each cell provides a count for the items that belong to the corresponding categories.

To do a crosstabulation:

- Go to “Analyze -> Descriptive Statistics -> Crosstabs”

- Select one variable for the rows (Car Type for us) and one for the columns (Cyl for us).
- Under "Cells" decide if you want to include "Row" percentages or "Column" percentages. Row percentages are set so each row adds up to 100%, and analogously for Column percentages. Do Row percentages for now.
- Click Continue, and then OK.

You should now be seeing a crosstabulation matrix. It has one row for each car type, and one column for each cylinder count, as well as totals at the ends. We can see from this for example that 39.6% of standard cars have 4 cylinders, and only 23.4% of sports cars have 4 cylinders.

If we had used column statistics, we would have seen the kinds of percentages that our second stacked bar graph tells us.

Categorical - Scalar

When comparing a categorical and a scalar variable, a good first step is a boxplot.

Boxplots

A boxplot will allow you to get a comparison of key numerical summaries between the various categories that the categorical variable introduces.

As an example, we will look at the number of cylinders compared to engine size. We would probably expect that the more cylinders a car has, the larger its engine size. Let us see if that is true.

You already know how to make these graphs. Go to the Chart Builder, choose boxplot from the left, and double-click the first of the three graphs there. Put Cyl in the x axis, and Eng in the y-axis. Then click OK.

When looking at boxplots:

- It is important to watch out for values that have too few values corresponding to them. You should have an idea before looking at the boxplot of how many values there are in each category. In this case we can get that information from the crosstabulation we made a moment ago - just look at the column totals.
- Boxes corresponding to very few elements are extremely unreliable. For instance 3, 5, 10 and 12 cylinders falls into that category.
- Take a look at the example for 10 cylinders in the boxplot graph. This is actually just two values, and they end up becoming the first and third quartiles, as well as the min and max, and the whole thing looks more important than it actually is.
- Let us remove the 3,5,10,12 cylinder cases. This is similar to what we did for the bargraphs a moment ago: Double-click the graph to edit it, select the bars, and you should see in the Properties window a "Categories" tab. In it you will find the

cases to remove. Then click Apply, and you should be seeing only the 4, 6 and 8 cylinder cases.

Let us compare those cases: You can see a marked difference: The engine size depends a lot on the number of cylinders. You can see this by comparing the boxes in each category. The middle 50% of the 8-cylinder cars are in the range of 4-5 engine size, while the 6-cylinder cars barely make it past 4, and the 4-cylinder cars don't even get close.

Summaries by category: Splitting the file

In such a situation we would like for SPSS to compute summaries for us broken by each category.

In order to achieve this, we will employ a process called "Splitting the file".

Splitting the file makes it so that every future work you do until you reset it will be done separately for each category of a categorical variable you provide.

To split the file:

- Go to "Data -> Split File". Do NOT choose "Split into Files".
- Select "Compare Groups".
- In the "Groups Based on" area, place the variable that you want to use as the splitting variable. In our case we want separate reports for each number of cylinders, so place Cyl there.
- Click OK.
- Not much appears to happen, but it did it. From now on any analysis or graph you ask of it will be done separately for each Cylinder count.
- When you are ready to go back to looking at all cases as one whole, go back to this menu and select "Analyze all cases, do not create groups". Do not do this now as we want to compute some statistics.

Now go ahead and compute summary statistics. If you recall, to do that we:

- go to "Analyze -> Descriptive Statistics -> Frequencies"
- Choose the variable we want (Engine size in this instance)
- select under "Statistics" which summaries to show (at least mean, median, quartiles, standard dev for now)
- deselect "Display Frequency Tables"

You should now be seeing separate statistics for each number of cylinders. You can now use those numbers to try and compare the different cylinder sizes in terms of their engine size behavior. As you can tell these are not easy to digest at first glance, so you typically want to combine them with graphics. And you definitely do not want to include those tables as is in a report. But you can take numbers out of it and create a brand new table in Powerpoint to show these numbers.

Here is an alternative:

- go to “Analyze -> Descriptive Statistics -> Descriptives”
- Choose the variable you want to see (Engine size for us)
- Choose the ones you want in the Options tab.
- Click OK

This produces a more compact array that is easier to visualize.

Before we move on, make sure to go back to the “Data -> Split File” menu and remove the splitting.

Scalar - Scalar

We will do a lot more later on with how to compare two scalar variables. For now we will just take a look at scatterplots.

Scatterplots

Scatterplots are 2-dimensional plots with a “dot” for each pair of x-y values appearing in your data.

To make a scatterplot:

- Go to “Graphs -> Chart Builder”
- Choose “Scatter/Dot” from the bottom left
- Double-click on the first option
- Put the two scalar variables you want to compare in the two axes. We will do Eng in the x-axis and CMpG in the y-axis.
- Click OK

You should be noticing a somewhat downward trend: As engine size becomes larger, city mileage becomes lower.

One nice feature you can add to a scatterplot is a “Fit line”.

- Double-click the graph to edit it
- Under Elements, choose “Fit Line at Total”
- By default “Linear” is chosen in the Properties window. This does what we call “linear regression” and is only suitable if the data trully behaves in a linear way.
- Switch the option to “Loess”. this tries to put a “smooth” line that best fits the data, but not necessarily a straight line. This is a good option to use to get a sense of the trend in your data.

We will have more to say about scatterplots later. For now, do a graph that shows the relation between length and width, and describe what you see.

To identify interesting values:

- When in the Chart Builder, go to “Groups/Point ID” and check “Point ID label”.
- Drag an appropriate variable to the “Point Label Variable” section. Car Name will do well for us.
- Click OK. This will add all names.
- Double-click the graph to edit it, right-click one of the dots and select “Hide Data Labels”.
- Click at the “target/crosshairs” button at the top left side of the graph, then click on those dots in the graph that you want to have labels (typically those outliers that deviate from the overall pattern).
- You can move the labels around if they overlap with other things.