Playing with Data

In this document, we will cover the following:
- How to conduct correlation analysis in PASW/SPSS
- How to explain (i.e., "interpret") the findings of a correlation analysis
- Tips, suggestions, and recommendations regarding data analysis and interpretion

Correlation in SPSS

This is not intended to be a technical report. The goal here is merely to show students how to make sense of numerically-based information in a commonly-used statistical package called "Statistical Package for the Social Sciences" (or "SPSS" for short).[1] SPSS is available on any of the computers in the campus library and computer labs, and it can be purchased for $45 from the Wisconsin Integrated Software Catalog (http://wiscsoftware.wisc.edu/wisc/).

We will show you how to open the program, how to put information (nerds call it "data") into it, and how to figure out (and, more importantly, articulate) what stories the data tell.

To do so, we use a simple example. Consider the following information:

X	Y
47	30
39	74
52	37
32	75
52	40
52	12
9	52
36	96
84	80
47	30

The table to the left contains two variables (let's call them "X," and "Y").[2] Both variables contain 10 values apiece; in this case, the values are randomly-picked numbers that range from 0 to 100. Suppose we ask the following question: is a relationship between the values in X and the values in Y? This is a common question to ask in social science research because much of what we study has to do with examining relationships between variables (like capitalism and economic development, income level and political ideology, etc.).

Anyway, our first step is to put the information here into SPSS. Find SPSS in the programs in the "Start" menu (if you are using a PC), or search for the SPSS logo on your Mac. Either way, double-click it and enter the following information into SPSS. We recommend that you copy the stuff in this table and paste into the "Data View" of SPSS. It would look like this:

---

[1] Since 2009, SPSS has had a new name: PASW (Predictive Analytics SoftWare), but it will always be SPSS to us!
[2] By "variables" we mean collections of information. Think of "X" and "Y" as "containers" that hold different values. For example, variable X contains the numbers you see in the left-hand column, and variable Y has the numbers in on the right-hand side.
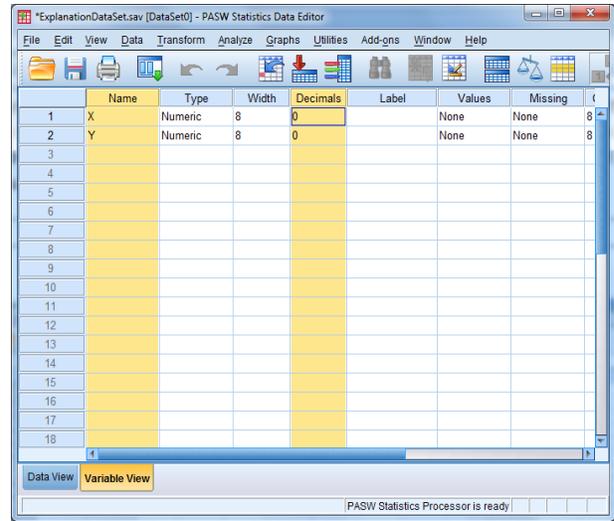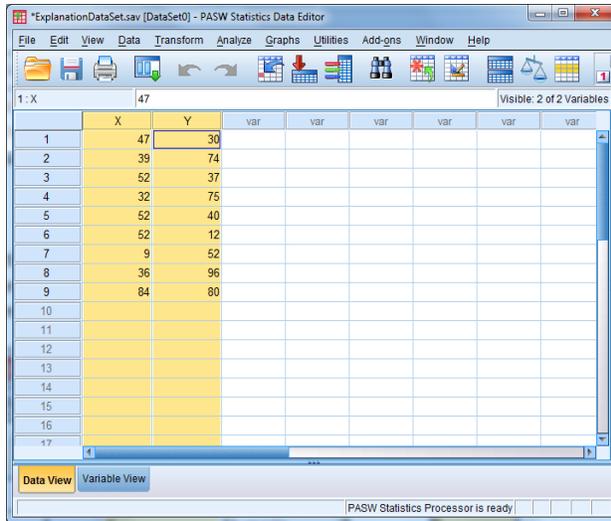
We highlighted the relevant information for you. Notice that we labeled the columns "X" and "Y" just like in the above table. You can do this easily by clicking the "Variable View tab in SPSS and re-naming the variables. Also, it would be a good idea to tell SPSS that your numbers have no (i.e., zero) decimal points:

Now, we can ask SPSS to look at the relationship between X and Y by using the pull-down menus. An easy thing to do is create a "scatterplot" of the data. Use the following commands:

How to create a scatterplot:

Graphs ➔ Legacy Dialogue ➔ Scatter/Dot
- Select "Simple Scatter" ➔ click "Define"
- Y goes into field labeled "Y Axis"
- X goes into field labeled "X Axis"
- Click "OK"

How to add a "line" to the scatterplot:
- Double click chart (in output) ➔ Add a "Fit Line"
- Click "Close" (the line will show up from now on)
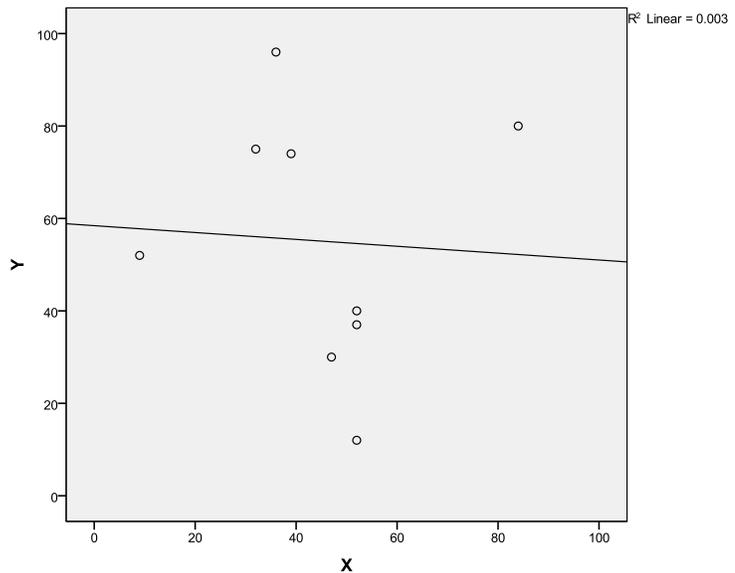
Scatterplots are great for determining how relationships between variables look. For example, the line you see above points slightly downward from left to right. This means that there is a weak but negative relationship between X and Y: higher values of Y tend to go with lower values of X and vice-versa. For example, if we look at the first "dot" to the left side of the scatterplot, we see that it represents the intersection of two values: a value of 52 for Y, and a value of 9 for X. In other words, as values of Y increase, we can generally expect values of X to decrease, but only a little bit.

As noted, the relationship between X and Y is a weak one, but is it "too weak" to care about? In social science research, whether a result is noteworthy or not depends on whether the result reaches what nerds call "statistical

significance." Given the weakness of this correlation, one might wonder whether it is statistically meaningful. In the following section, we will offer tips for interpreting a correlation result like this one.

<div align="center">

Interpreting/Explaining the Results

</div>

Notice that there is a number at the upper right-hand portion of the scatterplot. That number is called the R^2, which tells us the amount of variation in one variable that is accounted for by the variation in another variable. In this case, the R2 value here is .003, which means that only .3% variance in X is explained by Y and vice-versa. What does this mean? The more variance explained, the better, and since larger R^2 values signal stronger correlations, we can safely conclude that the relationship between X and Y here is rather weak.

Another approach to interpreting correlations is to do a hypothesis test to determine whether the correlation between X and Y is a meaningful one. Specifically, we want to determine whether the association between X and Y is merely due to chance. Correlation values range from +1 (perfect positive correlation) to -1 (perfect negative correlation) at their extremes, and a correlation value of zero (0) indicates that two variables are not correlated at all. When doing hypothesis tests, it is useful to state them in terms of whether the correlation in question is greater than zero (positive), smaller than zero (negative), or zero (no correlation). In our case, our "research hypothesis" is that a relationship exists between X and Y is negative, and we would have to run tests to rule out what statisticians call the "null hypothesis" (which, in our example, would be that the correlation between these variables is really zero). If we can rule out this null hypothesis, then we have evidence showing that there is a statistically meaningful correlation between X and Y. Here are the commands for testing correlation hypotheses:

Correlations		X	Y
X	Pearson Correlation	1	-.054
	Sig. (2-tailed)		.890
	N	9	9
Y	Pearson Correlation	-.054	1
	Sig. (2-tailed)	.890	
	N	9	9

How to do correlation analysis:
- Analyze ➔ Correlate ➔ Bivariate
- Variables: [add X and Y]
- Correlation Coefficients: [click "Pearson"]
- Test of Significance: [click "Two-tailed"]
- Flag significant correlations [check box]
- Click "OK"

There is a trick reading SPSS' correlation tables: Any variable correlated with itself will obviously have a value of "1". Therefore, the useful correlations appear in upper right and lower left "cells" of table. Here, the Pearson's correlation (our value of interest) is -.054, which confirms that our two variables are both negatively and weakly related to each other. The second set of information in the correlation table is labeled "Sig. (2-tailed)." It tells us the probability of us being wrong about ruling out the null hypothesis of no correlation between X and Y. As a general rule, we want the significance level to be as small as possible because that would mean that we did not make a mistake when we rejected the null hypothesis. The convention is to use significance values of .05 or lower as evidence of statistically meaningful correlations (put differently, statistically significant relationships are those for which we feel only 5% uncertain [or, if we invert the logic, 95% confident] that we have not committed an error). The significance value here is much larger than .05 (it suggests an 89% probability that ruling out the null hypothesis would be a mistake). Therefore, we cannot reject the null hypothesis of no correlation between X and Y, and we have to conclude that the relationship between these variables is meaningless.

Tips, Suggestions, and Recommendations

Data analysis is a form of storytelling. Your goal is to make the numbers come alive for your reader, and carefully-designed tables and figures are useful for summarizing data. When interpreting statistical results, always remember to speak plainly (avoid jargon when possible), write clearly, and weave technical statements about variables, hypotheses, etc. with common-sense discussions about what you expected to find, what you actually found, and whether you are surprised by your results.