**STAT13S\_pspp: Exercise Using PSPP to Explore Correlation**

Author: Ed Nelson
Department of Sociology M/S SS97
California State University, Fresno
Fresno, CA 93740
Email: ednelson@csufresno.edu

Note to the Instructor: The data set used in this exercise is gss14\_subset\_for\_classes\_STATISTICS\_pspp.sav which is a subset of the 2014 General Social Survey. Some of the variables in the GSS have been recoded to make them easier to use and some new variables have been created. The data have been weighted according to the instructions from the National Opinion Research Center. I prepared two documents to help you with PSPP – “[**Notes on Using PSPP**](http://www.ssric.org/files/notes_on_using_PSPP.docx)” and “[**Differences between PSPP and SPSS**](http://www.ssric.org/files/differences_between_PSPP_and_SPSS.docx)” which should answer many of your questions about PSPP. You have permission to use this exercise and to revise it to fit your needs. Please send a copy of any revision to the author. Included with this exercise (as separate files) are more detailed notes to the instructors and the PSPP syntax necessary to carry out the exercise. These, of course, will need to be removed as you prepare the exercise for your students. Please contact the author for additional information.

I’m attaching the following files.

* [**Data subset**](http://ssric.org/files/gss14_subset_for_classes_STATISTICS_pspp.sav) (.sav format)
* [**Extended notes for instructors**](http://ssric.org/files/Extended_Notes_for_Instructors_for_STAT13S_pspp.docx) (MS Word; docx format).
* [**PSPP syntax file**](http://ssric.org/files/PSPP_Syntax_for_STAT13S_pspp.sps) (.sps format)
* [**This page**](http://ssric.org/files/STAT13S_pspp.docx) (MS Word; docx format).

**Goals of Exercise**

The goal of this exercise is to introduce measures of correlation. The exercise also gives you practice using GRAPHS, BIVARIATE CORRELATION, and COMPARE MEANS in PSPP.

**Part I – Correlation**

We’re going to use the General Social Survey (GSS) for this exercise. The GSS is a national probability sample of adults in the United States conducted by the National Opinion Research Center (NORC). The GSS started in 1972 and has been an annual or biannual survey ever since. For this exercise we’re going to use a subset of the 2014 GSS. Your instructor will tell you how to access this data set which is called gss14\_subset\_for\_classes\_STATISTICS\_pspp.sav.

In a previous exercise (STAT11S) we considered different measures of association that can be used to determine the strength of the relationship between two variables that have nominal or ordinal level measurement (see STAT1S). In this exercise we’re going to look at two different measures that are appropriate for interval and ratio level variables. The terminology also changes in the sense that we’ll refer to these measures as correlations rather than measures of association.

Before we look at these measures let’s talk about a type of graph that is used to display the relationship between two variables called a scatterplot.  Click on “GRAPHS” in the menu bar and then on “Scatterplot.”  So all our scatterplots will look the same let’s put d22\_maeduc on the X-axis and d24\_paeduc on the Y-Axis.  Click on “OK” and PSPP will display your graph.

PSPP will list the variables and you will select those variables you want to use.  PSPP lists the variables using the variable labels.  However, it’s easier to find the variables if they are listed by variable names.  You can change the way PSPP lists the variables by right clicking anywhere on the list of variables and selecting “Prefer variable labels” and that will list the variables by name.  However, you will have to do this each time you encounter a list of variables.  There is no way to do this permanently.

Now look for the general pattern to the scatterplot. You see more cases in the upper right and lower left of the plot and fewer cases in the upper left and lower right. In general, as one of the variables increases, the other variable tends to increase as well. Moreover, you can imagine drawing a straight line that represents this relationship. The line would start in the lower left and continue towards the upper right of the plot. That’s what we call a positive linear relationship.**[[i]](http://ssric.org/node/576/edit%22%20%5Cl%20%22_edn1%22%20%5Co%20%22)**  But how strong is the relationship and where exactly would you draw the straight line? The Pearson Correlation Coefficient will tell us the strength of the linear relationship and linear regression will show us the straight line that best fits the data points. We’ll talk about the Pearson Correlation Coefficient in part 3 of this exercise and linear regression in exercise STAT14S.

**Part II – Now it’s Your Turn**

Use GRAPHS in PSPP to create the scatterplot for the years of school completed by the respondent (d4\_educ) and the spouse’s years of school completed (d29\_speduc). So all our plots will look the same, put d29\_speduc on the X-Axis and d4\_educ on the Y-Axis. Look at your scatterplot and decide if the scatterplot has a pattern to it. What is that pattern? Do you think it is a linear relationship? Is it a positive linear or a negative linear relationship?

**Part III - Pearson Correlation Coefficient**

The Pearson Correlation Coefficient (r) is a numerical value that tells us how strongly related two variables are. It varies between -1 and +1. The sign indicates the direction of the relationship. A positive value means that as one variable increases, the other variable also increases while a negative value means that as one variable increases, the other variable decreases. The closer the value is to 1, the stronger the linear relationship and the closer it is to 0, the weaker the linear relationship.

The usual way to interpret the Pearson Coefficient is to square its value. In other words, if r equals .5, then we square .5 which gives us .25. This is often called the Coefficient of Determination. This means that one of the variables explains 25% of the variation of the other variable. Since the Pearson Correlation is a symmetric measure in the sense that neither variable is designated as independent or dependent, we could say that 25% of the variation in the first variable is explained by the second variable or reverse this and say that 25% of the variation in the second variable is explained by the first variable. It’s important not to read causality into this statement. We’re not saying that one variable causes the other variable. We’re just saying that 25% of the variation in one of the variables can be accounted for by the other variable.

The Pearson Correlation Coefficient assumes that the relationship between the two variables is linear. This means that the relationship can be represented by a straight line. In geometric terms, this means that the slope of the line is the same for every point on that line. Here are some examples of a positive and a negative linear relationship and an example of the lack of any relationship.



Pearson r would be positive and close to 1 in the left-hand example, negative and close to -1 in the middle example, and closer to 0 in the right-hand example. You can search for “free images of a positive linear relationship” to see more examples of linear relationships.



But what if the relationship is not linear? Search for “free images of a curvilinear relationship” and you’ll see examples that look like this.

Here the relationship can’t be represented by a straight line. We would need a line with a bend in it to capture this relationship. While there clearly is a relationship between these two variables, Pearson r would be closer to 0. Pearson r does not measure the strength of a curvilinear relationship; it only measures the strength of linear relationships.

Another way to think of correlation is to say that the Pearson Correlation Coefficient measures the fit of the line to the data points. If r was equal to +1, then all the data points would fit on the line that has a positive slope (i.e., starts in the lower left and ends in the upper right). If r was equal to -1, then all the data points would fit on the line that has a negative slope (i.e., starts in the upper left and ends in the lower right). Let’s get the Pearson Coefficient for d22\_maeduc and d24\_paeduc. Click on Analyze in the menu bar and then click on BIVARIATE CORRELATION. Bivariate just means that you want to compute a correlation for two variables. Move these two variables into the “Variable(s)” box. Notice that the circle for “Two-tailed” is filled in for “Test of Significance.” A two-tailed significance test is used when you don’t make any prediction as to whether the relationship is positive or linear. In our case, we would expect that the relationship would be greater than zero (i.e., positive) so we would want to use a one-tailed test. Click on the circle for one-tailed to change the selection. Check the box labelled “Flag significant correlations” so PSPP will tell you when a relationship is statistically significant. Now click “OK” and PSPP will display your correlation coefficient.

You should see four correlations. The correlations in the upper left and lower right will be 1 since the correlation of any variable with itself will always be 1. The correlation in the upper right and lower left will both be 0.71. That’s because the correlation of variable X with variable Y is the same as the correlation of variable Y with variable X. Pearson r is a symmetric measure (see [**STAT11S\_pspp**](http://ssric.org/node/574)) meaning that we don’t designate one of the variables as the dependent variable and the other as the independent variable. Notice that the Pearson r is statistically significant using a one-tailed test at the .01 level of significance. A Pearson r of 0.71 is really pretty large. You don’t see r’s that big very often. That’s telling us that the linear regression line that we’re going to talk about in STAT14S\_pspp fits the data points reasonably well.

**Part IV – Now it’s Your Turn Again**

Use BIVARIATE CORRELATION in PSPP to get the Pearson Correlation Coefficient for the years of school completed by the respondent (d4\_educ) and the spouse’s years of school completed (d29\_speduc). What does this Pearson Correlation Coefficient tell you about the relationship between these two variables?

**Part V – Correlation Matrices**

What if you wanted to see the values of r for a set of variables? Let’s think of the four variables in Parts 1 through 4 as a set. That means that we want to see the values for r for each pair of variables. This time move all four of the variables into the “Variable(s)” box (i.e., d4\_educ, d22\_maeduc, d24\_paeduc, and d29\_speduc) and click on “OK.” That would mean we would calculate six coefficients. (Make sure you can list all six.)

What did we learn from these correlations? First, the correlation of any variable with itself is 1. Second, the correlations above the 1’s are the same as the correlations below the 1’s. They’re just the mirror image of each other. That’s because r is a symmetric measure. Third, all the correlations are fairly large. Fourth, the largest correlations are between father’s and mother’s education and between the respondent’s education and the spouse’s education.

**Part VI – The Correlation Ratio or Eta-Squared**

The Pearson Correlation Coefficient assumes that both variables are interval or ratio variables (see [**STAT1S\_pspp**](http://ssric.org/node/560)). But what if one of the variables was nominal or ordinal and the other variable was interval or ratio? This leads us back to one-way analysis of variance which we discussed in exercise STAT8S\_pspp. Click on “Analyze” in the menu bar and then on “COMPARE MEANS” and finally on “One-Way ANOVA.” Select the variable tv1\_tvhours and move it to the “Dependent List” box. This is the variable for which you are going to compute means. Then select the variable d3\_degree and move it to the “Factors” box. In this example, the independent variable is ordinal and the dependent variable is ratio. Notice that we’re using our independent variable to predict our dependent variable. Now click on “OK.”

The F test in the one-way analysis of variance tells us to reject the null hypothesis that all the population means are equal. So we know that at least one pair of population means are not equal. But that doesn’t tell us how strongly related these two variables are.  Eta-Squared is similar to the squared Pearson Correlation Coefficient (i.e., the Coefficient of Determination).  PSPP doesn‘t compute Eta-Squared for us but it gives us the information we need to compute it ourselves.  Eta-Squared is equal to the between groups sum of squares divided by the total sum of squares.  For this problem that would be 507.59 divided by 10,049.89 which equals .051.  This tells us that 5.1% of the variation in the dependent variable, number of hours the respondent watches television, can be explained or accounted for by the independent variable, highest education degree. This doesn’t seem like much but it’s not an atypical outcome for many research findings.

**Part VII – Your Turn**

In Exercise STAT8S you computed the mean number of hours that respondents watched television (tv1\_tvhours) for each of the nine regions of the country (d25\_region). Then you determined if these differences were statistically significant by carrying out a one-way analysis of variance. Repeat the one-way analysis of variance but this time focus on eta-squared. What percent of the variation in television viewing can be explained by the region of the country in which the respondent lived?

**[[i]](http://ssric.org/node/576/edit%22%20%5Cl%20%22_ednref1%22%20%5Co%20%22)** This assumes that the variables are coded low to high (or high to low) on both the X-Axis and the Y-Axis.