**STAT11S\_pspp: Exercise Using PSPP to Explore Measures of Association**

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.  This exercise uses CROSSTABS in PSPP to explore measures of association.  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_STAT11S_pspp.docx) (MS Word; .docx format).
* [**PSPP syntax file**](http://ssric.org/files/PSPP_Syntax_for_STAT11S_pspp.sps) (.sps format)
* [**This page**](http://ssric.org/files/STAT11S_pspp.docx) (MS Word; .docx format)

 **Goals of Exercise**

The goal of this exercise is to introduce measures of association.  The exercise also gives you practice in using CROSSTABS in PSPP.

**Part I—Relationships between Variables**

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.

The 2014 GSS is a sample from the population of all adults in the United States at the time the survey was done.  In a previous exercise (STAT9S\_pspp) we used crosstabulation and percents to describe the relationship between pairs of variables in the sample.  In exercise STAT10S\_pspp we went beyond simple description.  We used the sample data to make inferences about the population from which the sample was selected.  Chi Square was used to test hypotheses about the population.  Chi Square is the appropriate test when your variables are nominal or ordinal (see [**STAT1S\_pspp).**](http://ssric.org/node/560)

Chi Square is a test of the null hypothesis that two variables are unrelated to each other.  Another way to put this is that the two variables are independent of each other.  If we can reject the null hypothesis then we have support for our research hypothesis that the two variables are related to each other.  But showing that two variables are related is not the same thing as determining the strength of the relationship.  The strength of a relationship is actually a continuum from very weak to very strong.  To measure the strength of a relationship we need to select and compute a measure of association.  In this exercise we’re going to focus on nominal and ordinal variables.  In a later exercise (STAT13S\_pspp) we’ll talk about measures for interval and ratio variables.

**Part II – What is a Measure of Association?**

A measure of association is a numerical value that tells us how strongly related two variables are.  There are several characteristics of a good measure of association.

* They range from a value of 0 (i.e., no relationship) to 1 (i.e., the strongest possible relationship).
* For variables that have an underlying order from low to high they can be positive or negative.  A positive value indicates that as one variable increases, the other variable also increases.  A negative value indicates that as one variable increases, the other variable decreases.**[[1]](http://ssric.org/node/574/edit%22%20%5Cl%20%22_edn1%22%20%5Co%20%22)**
* Some measures specify which variable is dependent and which is independent.  The independent variable is some variable that you think might help you explain variation in the dependent variable.  For example, if your two variables were education and voting you might choose education as the independent variable and voting as your dependent variable because you think that education will help you explain why some people vote Democrat and others vote Republican. Measures of association that specify which variable is dependent and which is independent are called asymmetric measures and measures that don’t specify which is dependent and which is independent are called symmetric measures.

**Part III – Choosing a Measure of Association**

There are many measures of association to choose from. We’re going to limit our discussion to those measures that PSPP will compute. When choosing a measure of association we’ll start by considering the level of measurement of the two variables (see [**STAT1S\_pspp**](http://ssric.org/node/560)).

* If one or both of the variables is nominal, then choose one of these measures.**[[2]](http://ssric.org/node/574/edit%22%20%5Cl%20%22_edn2%22%20%5Co%20%22)**
	+ Contingency Coefficient
	+ Phi and Cramer’s V
	+ Lambda
* If both of the variables are ordinal, then choose from this list.
	+ Gamma
	+ Somer’s d
	+ Kendall’s tau-b
	+ Kendall’s tau-c
* Dichotomies should be treated as ordinal. Most variables can be recoded into dichotomies. For example, marital status can be recoded into married or not married. Race can be recoded as white or non-white. All dichotomies should be considered ordinal.

**Part IV – Measures of Association for Nominal Variables**

There are a number of nominal level variables in the 2014 GSS.  Here are a few examples.

* race of respondent – d6\_race
* race of household – d7\_hhrace
* region in which respondent lives – d25\_region
* region in which respondent lived at age 16 – d26\_reg16
* religious preference of respondent – r1\_relig
* religious preference of respondent at age 16 – r2\_relig16
* if the respondent is Protestant, denomination of respondent – r3\_denom1
* if the respondent is Protestant, denomination of respondent at age 16 – r5\_denom16

When one or both of your variables are nominal, you have a choice among the following measures – Contingency Coefficient, Phi and Cramer’s V, and Lambda.  Let’s start with the Contingency Coefficient (C).  One of the problems of this measure is that it varies from 0 to some value less than 1.  The larger the number of categories, the closer the maximum value is to 1.  For a table with two rows and two columns, the maximum value is .707 but for a table with three rows and three columns the maximum value is .816.  So you can’t use C to compare the strength of the relationship unless the tables have the same number of rows and columns.

Cramer’s V (V) is an extremely useful measure because it can vary between 0 and 1 regardless of the number of rows and columns.  Values of V can therefore be compared for tables with different number or rows and columns.  If your table has two columns and two rows V is the same as the Phi Coefficient which is another measure of association so PSPP refers to it as Phi and Cramer’s V.**[[3]](http://ssric.org/node/574/edit%22%20%5Cl%20%22_edn3%22%20%5Co%20%22)**

Lambda is a very useful measure because it has a clear and intuitive interpretation.  The value of Lambda tells you the degree to which knowing one of the variables helps you predict the other variable. A Lambda of .25 means that you can reduce the error in predicting one of the variables by 25% if you take into account the other variable.  Moreover there are actually three versions of Lambda – one that you would use when one variable is the dependent variable, another that you would use if the other variable was dependent, and a third you would use if you don’t want to designate either of the variables as dependent.  The problem with Lambda is that it often underestimates the strength of the relationship.

Let’s look at an example to help us better understand measures of association for nominal variables.  Use CROSSTABS in PSPP to get the table for d25\_region and d26\_reg16.  The first variable is the region of the country in which the respondent currently lives and the second is where the respondent lived at the age of 16.  It would make sense to think of d25\_region as the dependent variable since where respondents lived at age 16 might influence where they currently live.  Remember to put the dependent variable in the row and the independent variable in the column.  Ask PSPP to compute the column percents, Chi Square and the three measures we just discussed.

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.

Notice that C and V are quite high.  C is 0.90 and V is 0.71.  Ignore Phi since Phi is only used for a table with two columns and two rows.  You can see that C tells us that there is a very strong relationship between these two variables as does V.

Since we said that it was reasonable to think of where the respondent currently lives as the dependent variable, the appropriate value for lambda is .64 meaning that we can eliminate 64% of the errors in predicting the region in which one currently lives by taking into account where they lived at age 16.

**Part V – Now it’s Your Turn**

Use CROSSTABS to give you the table for d6\_race and d25\_region.  The variable d6\_race classifies the respondents as white, black, or other.  We want to find out whether the respondent’s race helps us predict where the respondent currently lives.  Decide which variable is independent and dependent.  Remember to put the dependent variable in the row and the independent variable in the column.  Get the correct percents and tell PSPP to compute Chi Square and the three measures of association we discussed.  Use all this information to describe the relationship between these two variables.

**Part VI – Measures of Association for Ordinal Variables**

There are a number of ordinal level variables in the 2014 GSS.  Here are a few examples.

* respondent’s highest educational degree – d3\_degree
* spouse’s highest educational degree – d28\_spdeg
* satisfaction with current financial situation – f4\_satfin
* happiness with life – hap2\_happy
* political views – p4\_polviews
* trust of other people – tf1\_trust

You have a choice from four measures that PSPP will compute for ordinal variables – Gamma, Somer’s d, Kendall’s tau-b, and Kendall’s tau-c.  Let’s start with Somer’s d.   This measure is the only one of the four that is an asymmetric measure.  That means that Somer’s d allows you to specify one of the variables as independent and the other as dependent.  Use CROSSTABS to get the crosstabulation of d3\_degree and f4\_satfin.  If we think that education influences how satisfied respondents are with their financial situation, then education would be our independent variable.  Put d3\_degree in the column and f4\_satfin in the row and run the table. Be sure to get the column percents, Chi Square, and the four measures of association we listed above.

Chi Square tells us that we should reject the null hypothesis that the two variables are unrelated which provides support for our research hypothesis that the variables are related to each other.  Since f4\_satfin is our dependent variable the appropriate value of Somer’s d is
-.15.  Tau-b and tau-c are very close to each other (-0.16 and -0.15).  Gamma (-0.24) is larger.  Gamma will always be larger because of the way it is computed.

Now let’s run a table using d3\_degree and d28\_spdeg.  It doesn’t seem reasonable to treat one spouse’s education as independent and the other spouse’s education as dependent so we would want the symmetric value of Somer’s d which equals 0.52.  Tab-b is 0.52 and tau-c is .46.  Gamma as always is larger (0.69).  The relationship between these two variables is clearly stronger than in the previous example.

You probably noticed that these measures for ordinal variables can be both positive and negative.  The problem is that it’s hard to interpret the sign.  We would like to be able to say that a positive value indicates that as one variable increases the other variable increases and a negative value indicates that as one variable increases the other variable decreases.  But that depends on how the values are coded.  So to determine whether a relationship is positive or negative it’s better to look at the percentages and let them tell you if it is positive or negative.

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

Use CROSSTABS to give you a table for d3\_degree and tf1\_trust.  We want to find out if the respondent’s education helps us understand why some say they trust people and other respondents feel they can’t trust others.   Decide which variable is independent and dependent.  Get the correct percents and tell PSPP to compute Chi Square and the four measures of association we discussed.  Use all this information to describe the relationship between these two variables.

**Part VIII – Using Measures of Association to Compare Tables**

The primary use of measures of association is to compare the strength of a relationship in several tables.  You want to make sure that you compare the same measure of association across tables.  Compare Gamma values to Gamma values and Lambda values to Lambda values.  Rerun one of the tables that you created in Parts 5 and 7 but this time hold sex constant.  In other words, sex would be your control variable.

In order to run a table with a control variable, we need to create a blank syntax file.  To do this click on “File” in the menu bar and then on “New” and finally on “Syntax.”  A blank syntax file should open.  Enter the following commands into the syntax file.  It’s easiest to do this by copying and pasting the commands into the syntax file.
CROSSTABS
  /TABLES=dependent variable BY independent variable BY d5\_sex
  /STATISTICS=CHISQ GAMMA
  /CELLS=COUNT COLUMN.
Be sure to replace “dependent variable” and “independent variable” with the names of these variables.  To run this command click on “Run” in the menu bar and then click on “All.”  Your table should appear in the output window.

Now compare the appropriate measure of association to determine if the relationship is stronger for male or females or whether it doesn’t vary by sex.  Remember not to make too much out of small differences in the measures.

**[[1]](http://ssric.org/node/574/edit%22%20%5Cl%20%22_ednref1%22%20%5Co%20%22)** See exercise STAT1S\_pspp for a discussion of levels of measurement.  Nominal variables have no underlying order and ordinal variables have an underlying order.  Measures of association for nominal variables range from 0 to 1 while measures for ordinal variables range from -1 to +1.

**[[2]](http://ssric.org/node/574/edit%22%20%5Cl%20%22_ednref2%22%20%5Co%20%22)** PSPP will also compute the Uncertainty Coefficient but we’re not going to consider this measure.

**[[3]](http://ssric.org/node/574/edit%22%20%5Cl%20%22_ednref3%22%20%5Co%20%22)** In the dialog box where you tell PSPP which measures of association to compute, it is referred to as Phi.  But in the output, both the values of Phi and Cramer’s V are displayed.  Phi is only used in a table with two rows and two columns.  In such a table, Phi and Cramer’s V are the same.  In a table with more than two rows or two columns, Phi should never be used.