**STAT12S\_pspp: Exercise Using PSPP to Explore Spuriousness**

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 RECODE to combine categories of a variable, FREQUENCIES to see how respondents answered the questions, and CROSSTABS to test for spuriousness.  In CROSSTABS students are asked to use percentages, Chi Square, and an appropriate measure 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_STAT12S_pspp.docx) (MS Word; .docx format).
* [**PSPP syntax file**](http://ssric.org/files/PSPP_Syntax_for_STAT12S_pspp.sps) (.sps format)
* [**This page**](http://ssric.org/files/STAT12S_pspp.docx) (MS Word; .docx format).

**Goals of Exercise**

The goal of this exercise is to explore the concept of spuriousness.  We will consider the relationship of religiosity and control of the distribution of pornography and test for the possibility that this relationship is spurious due to sex.   The exercise also gives you practice in using RECODE to combine categories of a variable and CROSSTABS in PSPP to explore the relationships among variables and to test for spuriousness.

**Part I—Religiosity and Control of the Distribution of Pornography**

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.  For this exercise we’re going to use a subset of the 2014 GSS survey. Your instructor will tell you how to access this data set which is called gss14\_subset\_for\_classes\_STATISTICS\_pspp.sav.

Let’s look at the relationship between the strength of a person’s religious affiliation and how a person feels about controlling the distribution of pornography.  One of the variables in the data set is PORN1\_PORNLAW.  This question asks respondents what type of laws they think we ought to have regulating the distribution of pornography.  Should pornography be illegal for everyone or should it be illegal only for those under the age of 18 or should it be legal for everyone?  We can draw a parallel to laws governing the distribution of drugs such as cocaine (illegal for everyone) and laws governing the distribution of alcohol and tobacco (illegal only for those under a certain age).  So it’s really a social control issue.

What is going to be our measure or indicant of religiosity?  Religiosity refers to the strength of a person’s attachment to their religious preference.  One of the questions in the GSS asks respondents how strong they consider themselves to be in their chosen religion.  The response categories are strong, somewhat strong, not very strong, or they have no religious preference.  This variable is R8\_RELITEN in the data set.

We’re going to recode R8\_RELITEN for this exercise.  The value 1 stands for those who say they are strong in their religious preference.  We’re going to leave this category as it is.  Then we’re going to combine somewhat strong (2), not very strong (3) and no religion (4) into one category and assign it a value of 2.  When you use RECODE in PSPP, you can recode in two different ways—into the same variable or into different variables.  If you recode into the same variable, be careful.  It’s easier, but if you make a mistake, you will not be able to go back and recode it again.  You will have to close PSPP without saving the data set and then reopen the data set to get a fresh, clean copy of the data.  So for this exercise recode into different variables.  You’ll have to give your recoded variable a new name.  Let’s call it R8\_RELITEN1.  To make your output more readable, add value labels for this variable.  The label for value 1 will be strong and for value 2 not strong.

After you have recoded this variable, run FREQUENCIES for your unrecoded variable (R8\_RELITEN) and your recoded variable (R8\_RELITEN1).  Compare the two frequency distributions to make sure you didn’t make an error recoding.  If you did make a mistake, you’ll need to do the recoding again.

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.

We’ll start by developing a hypothesis.  The stronger a person’s religious affiliation, the more likely they are to feel that pornography ought to be illegal for everyone regardless of their age. However, the weaker the person’s religious affiliation, the more likely they are to feel that pornography ought to be illegal only for those under the age of 18.  Imagine that you have told your hypothesis to a friend and your friend asks “Why?”  You need to explain why you think your hypothesis is true.  In other words, you need to develop an argument.  What is the link between religiosity and respondent’s opinion about pornography laws? Why should more religious individuals be more likely to think that pornography should be illegal for everyone?

Once you have developed your argument, then you should construct a dummy table showing what the relationship between R8\_RELITEN1 and PORN1\_PORNLAW should look like if your hypothesis is true.  Use “Tables” in Word to construct the table below.  It’s customary to put the independent variable in the column and the dependent variable in the row.  Add arrows to the table to show what your hypothesis would predict.  For example, compare cells a and b.  Would your hypothesis predict that cell a would be greater than cell b or would it predict that a would be less than b?  Do the same thing for cells c and d.  Does your hypothesis make any prediction about cells e and f?  If it doesn’t, then don’t insert an arrow for these two cells.

|  |  |
| --- | --- |
| Distribution ofpornography | Recoded Religiosity |
| Strong | Notstrong |
| Illegal to all | a | b |
| Illegal under 18 | c | d |
| Legal | e | f |

Now that you have constructed your dummy table, it’s time to find out what the relationship actually looks like. To do this you will need to run CROSSTABS in PSPP.  Be sure to put the independent variable (R8\_RELITEN1) in the column and the dependent variable (PORN1\_PORNLAW) in the row.  You also need to be sure to ask for the percents, Chi Square, and an appropriate measure of association.  Since the independent variable is the column variable, you will want the column percents.

All that is left is to interpret the table.  Since the independent variable is the column variable, we had PSPP compute the column percents.  It’s important to compare the percents across.  What does the table tell you about the relationship between religiosity and control over the distribution of pornography?  Use the percents, Chi Square, and the measure of association to help you interpret the table.

Remember not to make too much out of small percent differences. The reason we don’t want to make too much out of small differences is because of sampling error.  No sample is a perfect representation of the population from which the sample was selected.  There is always some error present.  Small differences could just be sampling error.  So we don’t want to make too much out of small differences.

**Part II—Adding a Third Variable into the Analysis**

At this point we have only considered two variables.  We need to consider other variables that might be related to religiosity and pornography control.  For example, sex may be related to both these variables.  Women may be more likely to say that they are strong in their religion and women may also be more likely to feel that pornography ought to be illegal for all regardless of age.  This raises the possibility that the relationship between self-reported strength of religion and how one feels about pornography laws might be due to sex.  In other words, it may be spurious due to sex.

Let’s check to see if sex is related to both our independent and dependent variables.  This is important because the relationship can **only** be spurious if the third variable (sex) is related to both your independent and dependent variables.  Use CROSSTABS to get two tables – one table should cross tabulate D5\_SEX and PORN1\_PORNLAW and the other table should cross tabulate D5\_SEX and R8\_RELITEN1.  Be sure to get the percents, Chi Square, and an appropriate measure of association.  If sex is related to both variables, then we need to check further to see if the original relationship between religiosity and pornography control is spurious as a result of sex.

**Part III—Checking for Spuriousness**

How are we going to check on the possibility that the relationship between strength of religion and pornography laws is due to the effect of sex on the relationship?  What we can do is to separate males and females into two tables and look at the relationship between strength of religion and pornography laws separately for men and for women.  Sex is called the 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=porn1\_pornlaw BY r8\_reliten1 BY d5\_sex
  /STATISTICS=CHISQ GAMMA
  /CELLS=COUNT COLUMN.

Notice the formal of the “TABLES” subcommand.  It lists the table you want to run in the following order – dependent variable BY independent variable BY control variable.  To run the CROSSTABS command, click on “RUN” in the menu bar and then on “ALL”

Check to see what happens to the relationship between strength of religion and opinion on pornography laws when we hold sex constant.  If the original relationship is spurious then it either ought to go away or to decrease substantially for **both** males and females.  So look carefully at the two tables – one for males and the other for females.  But how can we tell if the relationship goes away or decreases for both males and females?  One clue will be the percent differences.  Compare the percent differences between those who are more religious (i.e., strong) and those who are less religious (i.e., not strong) for males and then for females with the percent differences in the original two-variable table.  Did the percent difference stay about the same or did they decrease substantially?  Another clue is your measure of association.   Did the measures of association stay about the same for males and for females or do they decrease substantially from that in the original two-variable table?

If the relationship had been due to sex, then the relationship between strength of religion and opinion on pornography laws would have disappeared or decreased substantially for **both** males and females when we took out the effect of sex by holding it constant.  In other words, the relationship would be spurious.  Spurious means that there is a statistical relationship, but not a causal relationship. It important to note that just because a relationship is not spurious due to sex doesn’t mean that it is not spurious at all.  It might be spurious due to some other variable such as age.

**Part IV—Conclusions**

Summarize what you learned in this exercise.  What was the original two-variable relationship between religiosity and control over the distribution of pornography?  What happened when you introduced sex into the analysis as a control variable?  Was the original relationship spurious or not?  What does it mean to say a relationship is spurious?