**STAT10S: Exercise Using SPSS to Explore Chi Square**

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.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 SPSS to explore the Chi Square test.  A good reference on using SPSS is *SPSS for Windows Version 23.0 A Basic Tutorial* by Linda Fiddler, John Korey, Edward Nelson (Editor), and Elizabeth Nelson.  The online version of the book is on the [**Social Science Research and Instructional Council's Website**](http://ssric.org/node/582).  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, the SPSS syntax necessary to carry out the exercise (SPSS syntax file), and the SPSS output for the exercise (SPSS output file). Please contact the author for additional information.

I’m attaching the following files.

* [**Data subset**](http://ssric.org/files/gss14_subset_for_classes_STATISTICS.sav) (.sav format)
* [**Extended notes for instructors**](http://ssric.org/files/Extended_Notes_for_Instructors_for_STAT10S.docx) (MS Word; .docx format)
* [**Syntax file**](http://ssric.org/files/SPSS_Syntax_for_STAT10S.sps) (.sps format)
* [**Output file**](http://ssric.org/files/SPSS_Output_for_STAT10S.spv) (.spv format)
* [**This page**](http://ssric.org/files/STAT10S.docx) (MS Word; .docx format)

**Goals of Exercise**

The goal of this exercise is to introduce Chi Square as a test of significance.  The exercise also gives you practice in using CROSSTABS in SPSS.

**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.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 the previous exercise (STAT9S) we used crosstabulation and percents to describe the relationship between pairs of variables in the sample.  But we want to go beyond just describing the sample.  We want to use the sample data to make inferences about the population from which the sample was selected.  Chi Square is a statistical test of significance that we can use to test hypotheses about the population.  Chi Square is the appropriate test when your variables are nominal or ordinal (see STAT1S).

In STAT9S we started by using crosstabulation to look at the relationship between sex and opinion about abortion.  We’re going to use a1\_abany as our measure of opinion about abortion.  Respondents were asked if they thought abortion ought to be legal for any reason.  Run CROSSTABS to produce the table.  (See Chapter 5, Crosstabs in the online SPSS book mentioned on page 1.)  You want to get the crosstabulation of d5\_sex and a1\_abany.  Put the independent variable in the column and the dependent variable in the row.   Since your independent variable is in the column, you want to use the column percents.

**Part II – Interpreting the Percents**

Your table should look like this.****

Since your percents sum down to 100% (i.e., column percents), you want to compare the percents across.  Look at the first row.  Approximately 47% of men think abortion should be legal for any reason compared to 44% of women.  There’s a difference of 3.6% which seems small.  We never want to make too much of small differences.  Why not?  No sample is ever a perfect representation of the population from which the sample is drawn.  This is because every sample contains some amount of sampling error.  Sampling error in inevitable.  There is always some amount of sampling error present in every sample.  The larger the sample size, the less the sampling error and the smaller the sample size, the more the sampling error.

But what is a small percent difference?  Probably you would agree that a one to four percent difference is small.  But what about a five or six or seven percent difference?  Is that small?  Or is it large enough for us to conclude that there is a difference between men and women **in the population.**  Here’s where we can use Chi Square.

**Part III – Chi Square**

Let’s assume that you think that sex and opinion about abortion are related to each other.  We’ll call this our research hypothesis.  It’s what we expect to be true.  But there is no way to prove the research hypothesis directly.  So we’re going to use a method of indirect proof.  We’re going to set up another hypothesis that says that the research hypothesis is not true and call this the null hypothesis.  In our case, the null hypothesis would be that the two variables are unrelated to each other.**[[1]](http://ssric.org/node/507/edit%22%20%5Cl%20%22_ftn1%22%20%5Co%20%22)**  In statistical terms, we often say that the two variables are independent of each other.  If we can reject the null hypothesis then we have evidence to support the research hypothesis. If we can’t reject the null hypothesis then we don’t have any evidence in support of the research hypothesis.  You can see why this is called a method of indirect proof. We can’t prove the research hypothesis directly but if we can reject the null hypothesis then we have indirect evidence that supports the research hypothesis.

Here are our two hypotheses.

* research hypothesis – sex and opinion about abortion are related to each other
* null hypothesis – sex and opinion about abortion are unrelated to each other; in other words, they are independent of each other

It's the null hypothesis that we are going to test.

SPSS will compute Chi Square for you.  Follow the same procedure you used to get the crosstabulation between d5\_sex and a1\_abany.  Remember to get the column percents.  Then click on the “Statistics” button in the upper right of the dialog box.  Check the box for “Chi-Square” and then click on “Continue” and then on “OK.”

Now you will see another output box below the crosstabulation called “Chi-Square Tests.”  We want the test that is called “Pearson Chi-Square” in the first row of the box.  Ignore all the other rows in this box.**[[2]](http://ssric.org/node/507/edit%22%20%5Cl%20%22_ftn2%22%20%5Co%20%22)**  You should see three values to the right of “Pearson Chi-Square.”

* The value of Chi Square is 2.147.  Your instructor may or may not want to go into the computation of the Chi Square value but we’re not going to cover the computation in this exercise.
* The degrees of freedom (df) is 1.  Degrees of freedom is number of values that are free to vary.  In a table with two columns and two rows only one of the cell frequencies is free to vary assuming the marginal frequencies are fixed.  The marginal frequencies are the values in the margins of the table.  There are 766 males and 898 females in this table and there are 752 that think abortion should be legal for any reason and 912 who think abortion should not be legal for any reason.  Try filling in any one of the cell frequencies in the table.  The other three cell frequencies are then fixed assuming we keep the marginal frequencies the same so there is one degree of freedom.
* The two-tailed significance value is 0.143.**[[3]](http://ssric.org/node/507/edit%22%20%5Cl%20%22_ftn3%22%20%5Co%20%22)** This tells us that there is a probability of .143 that we would be wrong if we rejected the null hypothesis.  In other words, we would be wrong 143 out of 1,000 times.  With odds like that, of course, we’re not going to reject the null hypothesis.  A common rule is to reject the null hypothesis if the significance value is less than .05 or less than five out of one hundred.  Since .143 is not smaller than .05, we don’t reject the null hypothesis.  Since we can’t reject the null hypothesis, we don’t have any support for our research hypothesis.

**Part IV – Now it’s Your Turn**

Choose any two of the tables from the following list and compare men and women using crosstabulation and Chi Square:

* satisfaction with current financial situation (f4\_satfin),
* opinion about gun control (g1\_gunlaw),
* gun ownership (g2\_owngun),
* voting (p6\_pres12), and
* religiosity (r8\_reliten).

Make sure that you put the independent variable in the column and the dependent variable in the row.  Be sure to ask for the correct percents and Chi Square.  What are the research hypothesis and the null hypothesis?  Do you reject the null hypothesis?  How do you know?  What does that tell you about the research hypothesis?

**Part V – Expected Values**

We said we weren’t going to talk about how you compute Chi Square but we do have to introduce the idea of expected values.  The computation of Chi Square is based on comparing the observed cell frequencies (i.e., the cell frequencies that you see in the table that SPSS gives you) and the cell frequencies that you would expect by chance assuming the null hypothesis was true.  SPSS will also compute these expected frequencies for you.  Rerun the crosstabulation for d5\_sex and a1\_abany remembering to ask for the column percents and Chi Square.  But this time when you click on the “Cells” button to ask for the column percents look in the upper left of the dialog box where it says “Counts.”  “Observed” is selected as the default.  These are the observed cell frequencies.  Click on the “Expected” box to get the expected cell frequencies.

Now you will see both the observed and the expected cell frequencies in your output table.  Notice that they aren’t very different.  The closer they are to each other, the smaller Chi Square will be.  The more different they are, the larger Chi Square will be.  The larger Chi Square is, the more likely you are to be able to reject the null hypothesis.

Chi Square assumes that all the expected cell frequencies are greater than five.  We can see from the table that this is the case for this table.  But we don’t have to get the expected frequencies to see this.  Look back at the “Chi-Square Tests” table in your output.  Look at footnote a.  It tells you that the smallest expected cell frequency is 346.17.  So clearly all four expected cell frequencies are at least five.  If it’s just a little bit below five, that’s no problem.  But if it gets down around three you have a problem.  What you’ll have to do is to combine rows or column that have small marginal frequencies.

For example, run the crosstabulation of d5\_sex and d9\_sibs which is the number of brothers and sisters that the respondent has.**[[4]](http://ssric.org/node/507/edit%22%20%5Cl%20%22_ftn4%22%20%5Co%20%22)** The minimum expected frequency is so small that it rounds to 0.00.  That’s because there are only a few respondents with more than 15 siblings.  You will need to recode the number of siblings into fewer categories making sure that you don’t have any categories with a really small number of cases.

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

Look back at the two tables you ran in Part III and see if any of your expected frequencies were less than five.  What does that tell you?

**[[1]](http://ssric.org/node/507/edit%22%20%5Cl%20%22_ftnref1%22%20%5Co%20%22)** The null hypothesis is often called the hypothesis of no difference.  We’re saying that there is no relationship between these two variables.  In other words, there’s nothing there.

**[[2]](http://ssric.org/node/507/edit%22%20%5Cl%20%22_ftnref2%22%20%5Co%20%22)** Unfortunately there is no way to tell SPSS to just give us the “Pearson Chi-Square.”

**[[3]](http://ssric.org/node/507/edit%22%20%5Cl%20%22_ftnref3%22%20%5Co%20%22)** What do we mean by two-tailed? We’re not predicting the direction of the relationship. We’re not predicting that men are more likely to think abortion should be legal or that women are more likely.  So it’s a two-tailed test.

**[[4]](http://ssric.org/node/507/edit%22%20%5Cl%20%22_ftnref4%22%20%5Co%20%22)** Number of siblings is a ratio level variable.  You can use Chi Square with ratio level variables but usually there are better tests.  We’re just using this as an example.