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 Field\_2013\_subset\_for\_classes\_GUN\_CONTROL.sav which is a subset of a Field Poll conducted in February, 2013.  Some of the variables in this Field Poll 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 Field Research Corporation.  This exercise uses FREQUENCIES to get frequency distributions and CROSSTABS to explore relationships between variables.  In CROSSTABS students are asked to use percentage, Chi Square, and a measure of association.  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 Center'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). 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 which is a SPSS data file**](http://ssric.org/files/field_2013_subset_for_classes_GUN_CONTROL.sav) (.sav format).
* [**Extended notes for instructors**](http://ssric.org/files/Extended_Notes_for_GUN_CONTROL4G.docx). MS Word (.docx) format.
* [**SPSS syntax file**](http://ssric.org/files/SPSS_Syntax_for_GUN_CONTROL4G.sps) (.sps format).
* [**SPSS output file**](http://ssric.org/files/SPSS_Output_for_GUN_CONTROL4G.spv) (.spv format).
* [**This page**](http://ssric.org/files/GUN_CONTROL.docx) in MS Word (.docx) format.

## Goals of Exercise

The goal of this exercise is to explore the relationship between region of the state and opinion on gun control.  The exercise also gives you practice in using several SPSS commands –FREQUENCIES and CROSSTABS.  Percentages, Chi Square, and measures of association are used in the analysis.

### Part I—Region of the State

We’re going to use a Field Poll conducted in 2013 for this exercise.  The Field Poll is a statewide poll of registered voters in California conducted by the Field Research Corporation.  For this exercise we’re going to use a subset of this Field Poll. Your instructor will tell you how to access this data set which is called Field\_2013\_subset\_for\_classes\_GUN\_CONTROL.sav.

The Field Poll contains several variables that measure the region of the state in which the respondent lives.

* North vs. South (D\_REGION1\_region\_1)
* Inland vs. Coastal (D\_REGION2\_region\_2)
* Counties (D\_REGION3\_calcount)

Run FREQUENCIES in SPSS to get the frequency distribution for these variables. (See Chapter 4, FREQUENCIES, in the online SPSS book mentioned on page 1 of this exercise.)

There are five columns in the output that SPSS gives you.

* The first column is the value label for the response category.
* The second column is the number of cases or frequency for each response.
* The third column is the percent.  The denominator for the percent is the total number of cases in the sample (834).
* The fourth column is the valid percent.  Here the denominator is the number of valid cases.  This is the number of respondents who actually answered the question.  The number of valid cases is the total number of cases in the sample (834) minus the number of cases with missing data.
* The fifth column is the cumulative percent.  Notice that these percents cumulate and eventually equal 100.0 for the last of the valid response categories.

The percents and valid percents are identical for these variables because there aren’t any cases with missing information.  The respondent’s county of residence is part of the information from the list of registered voters from which the sample was selected.  When there are cases with missing information, these percents can be quite different.

Write a paragraph describing what these frequency distributions tell you about the regional distribution of respondents across the state.  You can view a map of counties in California by clicking on this [link](http://www.mapsofworld.com/usa/states/california/california-county-map.html).  Scroll down to see the population of each county.

### Part II – Region – North vs. South

When you ask respondents where they live in California, they often say “North” or South.”  The Field Poll used the county in which respondents live to divide respondents into these two groups.  In Part I you discovered that about 60% live in the south and 40% in the North.

Now let’s see if living in the North or South is related to opinion on gun control.  Run CROSSTABS in SPSS to get the crosstabulation of D\_REGION1\_region\_1 and G1\_q13.  Think carefully about which is your independent variable and which is your dependent variable.  Be sure to put the independent variable in the columns of the table and the dependent variable in the rows.  Make sure that you get the column percents, Chi Square, and Cramer’s V.

Chi Square is a test of significance that tests the null hypothesis that the two variables in your crosstabulation are unrelated to each other.  Another way of saying this is that the two variables are statistically independent of each other.  If you can reject the null hypothesis, then you have reason to believe that the two variables are related to each other.  If you can’t reject the null hypothesis, then you have no reason to believe that they are related.

But how do we decide if we should reject the null hypothesis?  Look at the output that SPSS gave you for the Chi Square Test.  The test that we want to use is the Pearson Chi-Square test which is on the first line of the output box.  It lists the Chi Square value, the degrees of freedom (df), and the significance value (labelled “Asymptotic Significance 2-sided”).  What we want to look at is the significance value.  This is the probability that we would be wrong if we rejected the null hypothesis.  For this pair of variables, it’s .044.  That means that there is a 4.4% chance that we would be wrong if we rejected the null hypothesis. We’re going to reject the null hypothesis if this probability is low which we will define as less than 5%.  We often refer to this 5% value as the level of significance.  Since .044 is less than .05, we reject the null hypothesis which means that there is probably a relationship between these two variables.  When we reject the null hypothesis, we often say that Chi Square is statistically significant.

Chi Square is a test of the null hypothesis that the two variables are unrelated to each other.  If you reject the null hypothesis, then you have evidence that the two variables are related.  But it doesn’t tell you anything about the strength of the relationship.  We can get an idea of the strength by looking at the percent differences in the crosstabulation.  Look again at the table you just ran.  The independent variable (i.e., D\_REGION1\_region\_1) ought to be in the column and the dependent variable (i.e., G1\_q13) ought to be the row and you should have asked SPSS to give you the column percents.  Remember that if the percents sum down to 100 (as they should if you asked for the column percents), then you want to compare straight across. Look at the first row (i.e., “right to own guns”).  The percents (36.7% and 32.5%) are quite similar.  If you subtract one from another (i.e., 36.7 – 32.5 = 4.2), the difference is not much larger than zero.  And that’s true for the other rows of the table too.  That’s telling you that there isn’t much of a difference between those who live in the North and those who live in the South in terms of how they feel about gun control.

A measure of association is a statistic that measures the strength of the relationship between two variables.  G1\_q13 is an example of a nominal measure.  A **nominal measure** is one in which objects (i.e. in our survey, these would be the respondents) are sorted into a set of categories which are qualitatively different from each other.  The categories in a nominal measure have no inherent order to them.  This means that it wouldn’t matter how we ordered the categories.  They could be arranged in several different ways.  For our variable, G1\_q13 we could have arranged the three response categories (i.e., right to own guns, control gun ownership, no opinion) in any order.  When one of your variables is nominal, then one of the measures of association you might use is Cramer’s V.  V varies between 0 and 1.  The closer it is to 0, the weaker the relationship and the closer it is to one, the stronger the relationship.  In this example, V = .086, which is not much larger than zero indicating a quite weak relationship.

Write a paragraph describing the relationship between these two variables.  Use the percents, Chi Square, and V to help you interpret the table.

### Part III – Region – Inland vs. Coastal

Another way to measure region is to classify counties as either coastal or inland counties.  You ran a frequency distribution for D\_REGION2\_region\_2 in Part I.  Now run CROSSTABS to get the crosstabulation of this variable and G1\_q13.  Think carefully about which is your independent variable and which is your dependent variable.  Be sure to put the independent variable in the columns of the table and the dependent variable in the rows.  Make sure that you get the column percents, Chi Square, and Cramer’s V.

Look at the Chi Square box in the output.  What is the significance value?  Do you reject or not reject the null hypothesis that the two variables are unrelated to each other?  What does that tell you?

Now look at the percent differences for each row of the table as you did in Part II.  Are they larger, smaller, or about the same as the percent differences you found in Part II?  What does this tell you about the strength of the relationship between the variables in Parts II and III?

Also, consider the values of V in Part II and Part III.  What does that tell you?

### Part IV – Region – Counties

California is a very large state both in terms of geographical size and population.  There are 58 counties in California that vary widely in terms of location, geographical size, and population size.  D\_REGION3\_calcount classifies respondents in terms of the county in which they reside.  In Part I you ran FREQUENCIES to see how many respondents live in each county.  Which counties have 50 or more respondents?  Which county has the most respondents?  Which counties have the fewest respondents?

Run CROSSTABS to get the crosstabulation of this variable and G1\_q13.  Think carefully about which is your independent variable and which is your dependent variable.  Be sure to put the independent variable in the columns of the table and the dependent variable in the rows.  Make sure that you get the column percents, Chi Square, and Cramer’s V.

There are some clear problems with this table.  One problems is that it’s too large – 58 columns and 3 rows.  It’s much too large to be useful.  But there’s a more important problem.  Look at the total row in the table that shows how many respondents live in each county.  There are a number of counties that have as few as one or two respondents.  Why is that a problem?  (Hint: think about how the percents are computed.)

There’s another problem.  Look at the Chi Square box in the output.  In the footnote to the table, it says that “The minimum expected count is .05.”  The expected counts are the number of cases that you would expect in each cell of the crosstabulation assuming that the two variables were unrelated to each other.  The “minimum expected count” is the smallest of the expected counts in all cells of the table.  An important assumption of the Chi Square test is that these expected counts are all at least five.  As long as they are not too much smaller than five, you don’t have a problem.  But if they drop too much below three, then you need to combine columns and/or rows to increase the expected counts.  And .05 is much, much too low.  So clearly you have to combine counties to in some meaningful way. And that is exactly what the Field Poll did in D\_REGION1\_region\_1 and D\_REGION2\_region\_2.

### Part V -- Conclusions

Write a paragraph describing what you learned about the relationship between region and how people feel about gun control. Be sure to explain why you can’t use the table in Part IV.