# RESEARCH METHODS 10RM - 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:** This is the tenth in a series of 13 exercises that were written for an introductory research methods class.  The first exercise focuses on the research design which is your plan of action that explains how you will try to answer your research questions.  Exercises two through four focus on sampling, measurement, and data collection.  The fifth exercise discusses hypotheses and hypothesis testing.  The last eight exercises focus on data analysis.  In these exercises we’re going to analyze data from one of the [Monitoring the Future Surveys](http://monitoringthefuture.org/) (i.e., the 2017 survey of high school seniors in the United States).  This data set is part of the collection at the Inter-university Consortium for Political and Social Research at the University of Michigan.  This data set is freely available to the public and you do not have to be a member of the Consortium to use it.  We’re going to use SDA (Survey Documentation and Analysis) to analyze the data which is an online statistical package written by the Survey Methods Program at UC Berkeley and is available without cost wherever one has an internet connection.  A weight variable is automatically applied to the data set so it better represents the population from which the sample was selected.  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 so I can see how people are using the exercises. Included with this exercise (as separate files) are more detailed notes to the instructors and the exercise itself.  Please contact the author for additional information.

[This page](http://ssric.org/files/Research_Methods_10RM.docx) in MS Word (.docx) format is attached.

**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 SDA.

**Part I—Relationships between Variables**

We’re going to use the Monitoring the Future (MTF) Survey of high school seniors for this exercise.  The MTF survey is a multistage cluster sample of all high school seniors in the United States.  The survey of seniors started in 1975 and has been done annually ever since. To access the MTF 2017 survey follow the instructions in the Appendix.   Your screen should look like Figure 10-1.  Notice that a weight variable has already been entered in the WEIGHT box.  This will weight the data so the sample better represents the population from which the sample was selected.

​
Figure 10-1

MTF is an example of a social survey.  The investigators selected a sample from the population of all high school seniors in the United States.  This particular survey was conducted in 2017 and is a relatively large sample of a little more than 12,000 seniors.  In a survey we ask respondents questions and use their answers as data for our analysis.  The answers to these questions are used as measures of various concepts.  In the language of survey research these measures are typically referred to as variables.

In exercise [9RM](http://ssric.org/node/637) 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 exercise [6RM](http://ssric.org/node/628)).

Before we look at the relationship between variables, we need to talk about independent and dependent variables.  The dependent variable is whatever you are trying to explain.  For example, we could be trying to explain why some students have consumed alcoholic beverages during the last year and others have not.  The independent variable is some variable that you think might help you explain why some students have consumed alcoholic beverages.  We’re going to use sex as our independent variable.  Normally we put the dependent variable in the row and the independent variable in the column of our table.  We’ll follow that convention in this exercise.

Run CROSSTABS in SDA to produce the crosstabulation of *v2105 (how often consumed alcoholic beverages in last year)* and *v2150 (sex)*.  Look at PERCENTAGING under TABLE OPTIONS.  Since your independent variable is in the column, you want to use the column percents.  By default, the box for column percents is already checked. Your screen should look like Figure 10-2. Notice that the WEIGHT box is filled in.  Click on RUN THE TABLE to produce the crosstabulation.

​
Figure 10-2

**Part II – Interpreting the Percents**

Your table should look like Figure 10-3.

​
Figure 10-3

One of the first things you notice is that a large percent of students (43.8%) have not consumed alcoholic beverages in the last 12 months and a much smaller percent have consumed them a large number of times (8.3% engaged in drinking 20 or more times).  It might be easier to divide *v2105* into two categories – never and one or more times.  Fortunately, the data set has such a variable.  The dichotomous variable is *v2105d*.  (The d stands for dichotomy.)  Rerun the table you just ran replacing *v2015* with *v2015d*.  Your screen should look like Figure 10-4.

​
Figure 10-4

Since your percents sum down to 100% (i.e., column percents), you want to compare the percents across.  Look at the first row. Approximately 45.2% of men have not consumed alcoholic beverages in the last year compared to 42.5% of women which is a difference of only 2.7 percentage points. 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 is 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?  If you think that three percent is a small difference, what about a four or 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 males and females have different drinking patterns.  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]](%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 – the two variables, sex and drinking, are related to each other
* null hypothesis – the two variables, sex and drinking, 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.

SDA will compute Chi Square for you.  Follow the same procedure you used to get the crosstabulation between *v2150 (sex)* and *v2105d (drinking)*.  Remember to get the column percents.  Check the box for SUMMARY STATISTICS under TABLE OPTIONS.  Finally, click on RUN THE TABLE.

In the SUMMARY STATISTICS part of the output, you’ll see two Chi Squares – Chisq-P and Chisq-LR.  We want to use the first one listed – Chisq-P.  This is usually referred to as the Pearson Chi Square.  The number in parentheses which in this case is 1 is the degrees of freedom.

The value of the Pearson Chi Square is 7.99.  Your instructor may or may not want to go into the computation of the Chi Square value but we’re not going to cover it in this exercise.

The degrees of freedom (df) is 1.  That’s the number inside the parentheses following the Chi Square value.  Degrees of freedom is the 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 5,517.3 males and 5,984.4 females in this table and there are 5,037.1 that have not consumed alcoholic beverages and 6,464.6 who have.  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.

Now we have to decide if we should reject the null hypothesis that the two variables are unrelated (or statistically independent) based on the Chi Square value and the degrees of freedom.  Look at your output again and you’ll see that after the Chi Square value it says (p=0.00).  That is the probability that you would be wrong if you rejected the null hypothesis.  It looks like the probability is zero but it’s not.  This is actually a rounded value.  The probability is less than 0.005.  There is some very small chance that you would be wrong if you rejected the null hypothesis.  With odds like that, of course, we’re 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 <0.005 is smaller than .05, we reject the null hypothesis.  That means that there is support for our research hypothesis that sex is related to drinking.

You might be wondering how such a small percent difference (i.e., 3%) could be statistically significant.  It’s because the sample is so large. The larger the sample, the less the sampling error.  The smaller the sample, the more the sampling error.  With a sample of almost 12,000, there won’t be much sampling error and even a small percent difference could be significant.

You might be wondering why our sample is now a little less than 12,000 while before it was more than 12,000.  It's because of missing data. Some cases have missing data on one or both of the variables and these cases are omitted from the table.

**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.

* ever smoked cigarettes (*v2101d*)
* ever used marijuana or hashish (*v2115d*)
* how often attend religious services (*v2169*)
* how important religion is in life (*v2170*)
* rate self on school ability (*v2173*)
* how often skipped classes in last four weeks (*v2178*)
* how likely to graduate from four-year college (*v2183*)

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 to 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 SDA gives you) and the cell frequencies that you would expect by chance assuming the null hypothesis was true.

Computing expected frequencies is not difficult. Look back at Figure 10.4 and focus on the upper left cell of the table. This cell is for males who have not consumed alcoholic beverages in the last year. To get the expected frequency for this cell multiply the row marginal (5037.1) and the column marginal (5517.3) and divide the product by the total number of cases in the table (11501.6). Your answer should be 2416.3. This is the expected frequency for this cell. You would do this for all four cells in the table to get the expected frequencies.

Chi Square assumes that all the expected cell frequencies are greater than five.  Small expected frequencies might occur when you have a column or a row with a small number of cases in it.  What you’ll have to do is to combine rows or columns that have small marginal frequencies in order to increase the expected frequencies values.  You can do that in SDA but we're not going to go into it in this exercise.

[[1]](%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.