**RESEARCH METHODS 12RM - 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:** This is the twelfth 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_12RM.docx) in MS Word (.docx) format is attached.

**Goals of Exercise**

The goal of this exercise is to explore the concept of spuriousness.  We will consider the relationship between students’ high school grades and their expectation of graduating from a four-college in the future.  Then we will test for the possibility that this relationship is spurious.   The exercise also gives you practice in using CROSSTABS in SDA to explore relationships among variables and to test for spuriousness.

**Part I—Relationship of Grades in High School and Expectation of Graduating from Four-Year College**

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 12-1.  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 12-1

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 previous exercises we looked at variables one at a time (i.e., univariate analysis) and at the relationship between two variables (i.e., bivariate analysis).  In this exercise we’re going to add a third variable into the analysis (i.e., multivariate analysis) and consider the possibility that our two-variable relationship might be spurious due to this third variable.  Spuriousness means that there is a statistical relationship between two variables, but it is not a causal relationship.  The statistical relationship is due to the third variable which we typically call the control variable.

To illustrate the idea of spuriousness, think about children in elementary, middle, and high school.  Every year children take standardized tests at the end of the school year to measure their achievement in areas such as mathematics, reading, and science.  Did you know that children with small feet score lower on these tests than children with big feet?  There is a relationship between children’s foot size and their test scores.  There is a clear statistical relationship between these two variables.  But is it a causal relationship?  Of course not!  No parent ever says I hope my kids have big feet so they will do better in school.  What we’re saying is that we think this relationship is spurious.  There is a statistical relationship but it’s not causal.

But why is this relationship spurious?  There must be some third variable that is creating this relationship.  One possibility is children’s grade level.  Children in lower grades have smaller feet and lower test scores.  Children in higher grades have bigger feet and higher test scores.  So, the relationship between foot size and test scores might be due to grade level.

How are we going to test this hypothesis?  What we do is to hold the third variable constant.  Let’s say that we have test scores for children in grades 6 through 12.  We’ll start with the sixth grade and look at the relationship of foot size and test scores for only the sixth graders.  Then we’ll repeat this for the seventh graders and for each successive grade level.  If the relationship is spurious, then we ought to find that the relationship between foot size and test scores goes away or is considerably reduced for **each** grade level. If the relationship is not spurious, then we ought to find that the relationship does not change much for the different grade levels.

Now let’s turn to an example from the MTF survey.  One of the questions in the survey asks students “How likely is it that you will graduate from a four-year college after high school?”  This is variable *v2183* in the data set.  The response categories are definitely won’t, probably won’t, probably will, and definitely will.  This will be our dependent variable.  In other words, we’re trying to explain why some students think they will graduate from college and others don’t.

It might be that students’ expectations about college are, in part, based on how they have done academically in high school.  Another question in the survey asks, “Which of the following best describes your average grade so far in high school?”  This is variable v2179 and will be our independent variable.  In other words, this is the variable that we think might explain why some students think they will graduate from college and others don’t.

Run CROSSTABS in SDA to see the relationship between students’ grades (v*2179*) and their expectation of graduating from a four-college in the future (*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 the summary statistics.  Write a paragraph interpreting this relationship using the percents, Chi Square, and an appropriate measure of association.

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 (i.e., bivariate analysis).  We need to consider other variables that might be related to grades and expectations about college.  For example, sex may be related to both these variables.  Women may report higher grades and women may also be more likely to think they will graduate from college.  This raises the possibility that the relationship between grades and expectations for graduating from college might be due to sex.  In other words, it may be spurious due to sex.

Let’s check to make sure that sex is related to both our independent and dependent variables.  This is important because the relationship can **only** be spurious if the third variable (*v2150*) is related to both your independent and dependent variables.  Use CROSSTABS in SDA to get two tables – one table should cross-tabulate *v2150* and *v2179* and the other table should cross-tabulate *v2150* and *v2183*.  Be sure to get the SUMMARY STATISTICS so you will be able to use Chi Square and whichever measure of association you think is appropriate.  If sex is related to both variables, then we need to check further to see if the original relationship between grades and expectations for graduating from college 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 grades in high school and expectations about college 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 grades and expectations about college separately for men and for women. Sex is variable 2150 in our data set.

We can do that in SDA by getting a crosstab putting *v2183* in the ROW box (our dependent variable), *v2179* in the COLUMN box (our independent variable), and *v2150* in the CONTROL box.  In this case, *v2150* is the variable we are holding constant and is often called the control variable. You will get two tables – one for males and the other for females. Sometimes we call these partial tables since each partial table contains part of the sample.

We’re going to check to see what happens to the relationship between grades and expectations about college 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 markedly for both males and females?  One clue will be the percent differences between those who get high grades and those who get lower grades. Did the percent differences stay about the same or did they decrease substantially?  But there are so many percent differences that it’s hard to make sense of them. That’s where the summary statistics come in handy. Did the measures of association for males and females stay about the same or did they decrease substantially from that in the original two-variable table?

If the relationship had been due to sex, then the relationship between grades and expectations about college 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.

The first thing we notice is that the Chi Square is significant for both males and females.  That tells us that there is probably a relationship between grades and expectations for graduating from college for **both** males and females.

Look at the pattern to the percents and the measures of association in the tables for males and females.  For both males and females, the higher the grades in high school, the more likely students are to feel they will graduate from college.  Additionally, the measures of association aren’t identical, but they aren’t very different.  Remember we don’t want to make too much of small differences because of sampling error.  So, in this analysis we would conclude that the relationship is not spurious due to sex.  But keep in mind that it might be spurious when we control for a different variable.

**Part IV – Now it’s Your Turn**

Let’s stick with the same two-variable relationship – *v2179* and *v2183* – but this time let’s use a different control variable.  This time let’s use father’s education as our control variable.  Father’s education is coded into the following categories: 1=completed grade school or less, 2=some high school, 3=completed high school, 4=some college, 5=completed college, 6=graduate or professional school after college, and 7=don't know or does not apply. This is variable v2163 in our data set.

We’re going to make things simpler by recoding father’s education (*v2163*) into two categories – not college grad and college grad.  Recoding just means to combine categories of the variable.  Follow these steps to recode in SDA.

* Enter the variable name in the row box.  The variable name in this example is *v2163*.  (Don’t enter the period.)
* After the variable name, enter (r: where r stands for recode.
* Enter the new value you want to assign to the first recode followed by the values you want to combine.  In our case, we want to assign the new value 1 to the old values of 1 through 4.  So, this would be 1=1-4.  (Don’t enter the period.)
* We also want to assign a label to each category.  Enter the label in double quotation marks. For example, our recode for the first category would look like this – *v2163* (r:1=1-4"not college grad";.  (Don’t enter the period.)
* Separate the recodes by a semicolon.
* Repeat this process for each recode.  For the second category it would look this – 2=5-6"college grad".  (Don’t enter the period.)
* After the last recode, end the statement with a right parenthesis.
* This is what our recode statement would look like – v2163(r:1=1-4"not college grad";2=5-6"college grad").  (Don’t enter the period.)
* One more thing.  You’ll notice that we didn’t enter the value 7 (don’t know) into the recode.  That’s because we want to treat 7’s as missing data.  These respondents didn’t really answer the question.  They said they didn’t know.  Missing data are automatically excluded from the table.

Follow the same procedure that we used in parts 1, 2, and 3.  Interpret the tables and decide if the relationship is spurious or not.

**Part V—Conclusions**

Summarize what you learned in this exercise.  What happened when you introduced *sex* into the analysis as a control variable?  What happened when you used father’s education as the control variable?  Were the original relationships spurious or not?  What does it mean to say a relationship is spurious?