# Exercises for an Introductory Statistics Course Using SDA

# Edward Nelson

# California State University, Fresno

# Table of Contents

Page

Preface 3

Exercise 1 – Levels of Measurement 5

Exercise 2 – Measures of Central Tendency and Dispersion 9

Exercise 3 - Measures of Skewness and Kurtosis 15

Exercise 4 – Graphs and Charts 19

Exercise 5 – Comparing Means and Testing Hypotheses 22

Exercise 6 – Hypothesis Testing – One Way Analysis of Variance 27

Exercise 7 – Crosstabulation 32

Exercise 8 – Hypothesis Testing – Chi Square 38

Exercise 9 – Measures of Association 44

Exercise 10 – Spuriousness 50

Exercise 11 – Correlation 56

Exercise 12 – Comparison of Correlations 61

Exercise 13 – Bivariate Linear Regression 66

Exercise 14 – Multivariate Linear Regression 71

Exercise 15 – Dummy Variable Regression 77

Appendix – Notes on Using SDA 83

# Exercises for an Introductory Statistics Course Using SDA Edward Nelson, California State University, Fresno

# Preface

# These exercises were written for introductory statistics and research methods courses although they could be used in any class that has a quantitative component. They could also be used by individuals who want practice with the various statistical procedures. The exercises do not discuss all aspects of the statistics covered nor do they describe how to compute most of these statistics.

This is the third in a series of exercises for statistics and methods courses. The first set uses SPSS while the second set uses PSPP. These two sets of exercises can be found on the Social Science Research and Instructional Council’s [website](http://ssric.org/tr/onlinetextbooks).

This series uses SDA (Survey Documentation and Analysis) which is an online statistical package written by the Survey Methods Program at UC Berkeley. SDA can be used without cost wherever one has an internet connection. Students can be shown how to use SDA in approximately ten minutes making it unnecessary to spend valuable class time learning how to use a statistical package. There is also an extensive help menu available to users of SDA.

The data set used in this series of exercises is the General Social Survey’s 2018 Cumulative Data File (1972 to 2018) which is available without cost by clicking [here](https://sda.berkeley.edu/sdaweb/analysis/?dataset=gss18). For these exercises we will only be using the 2018 General Social Survey. The exercises show students how to select the 2018 survey from the cumulative data set. A weight variable is automatically applied to the data so it better represents the population from which is sample was selected.

The General Social Survey is a large, national probability sample of adults (18 years and older) living in the United States conducted by the National Opinion Research Center (NORC) at the University of Chicago. The GSS started in 1972 and was conducted annually through 1994 and biannually since then. Many of the questions in the GSS have been repeated from previous years providing important trend data. The most recent GSS was in 2018. The sample size for the 2018 survey was a little more than 2,300 individuals.

More information about the GSS can be found on the [NORC - General Social Survey website](http://gss.norc.org/). At the website you will find documentation for the survey, survey questionnaires, a bibliography, useful Frequently Asked Questions, and more. You can also download the complete GSS in either SPSS or Stata format.

In these exercises variable names appear in italics and SDA commands are in all capitals to make them easily recognizable. You can modify this if you wish.

The exercises were written so that each exercise is independent of the other exercises. That means that there is some redundancy across the exercises. If you choose to use several exercises you may want to remove some of the redundant material.

You have permission to edit the exercises in whatever way you desire. You can freely delete and add materials of your own. Please cite the original source of the exercises. I would like to hear from you about your experiences using them. If you find any errors, please let me know and I’ll correct them.

## About the Author

Ed Nelson is Professor Emeritus of Sociology at California State University, Fresno. Before retiring he taught courses in research methods, statistics, and critical thinking. After retiring he continues to teach a course in critical thinking. He can be reached by email at [ednelson@csufresno.edu](mailto:ednelson@csufresno.edu). Please contact him with any questions you might have.

# Exercises for an Introductory Statistics Course Using SDA Edward Nelson, California State University, Fresno

# Exercise 1 Levels of Measurement

## Goals of Exercise

The goal of this exercise is to explore the concept of levels of measurement (nominal, ordinal, interval, and ratio measures) which is an important consideration for the use of statistics. The exercise also gives you practice in using FREQUENCIES in SDA.

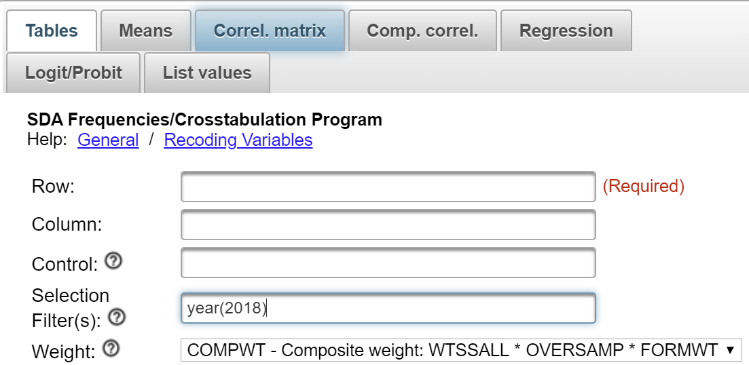
## Part I—Introduction to Levels of Measurement

We use concepts all the time. We all know what a book is. But when we use the word “book” often we’re not talking about a particular book that we’re reading. We’re talking about books in general. In other words, we’re talking about the concept to which we have given the name “book.” There are many different types of books – paperback, hardback, small, large, short, long, and so on. But they all have one thing in common – they all belong to the category “book.”

Let’s look at another example. Religiosity is a concept which refers to the degree of attachment that individuals have to their religious preference. It’s different than religious preference which refers to the religion with which they identify. Some people say they are Lutheran; others say they are Roman Catholic; still others say they are Muslim; and others say they have no religious preference. Religiosity and religious preference are both concepts.

A concept is an abstract idea. So there are the abstract ideas of book, religiosity, religious preference, and many others. Since concepts are abstract ideas and not directly observable, we select measures or indicants of these concepts. Religiosity can be measured in a number of different ways – how often people attend church, how often they pray, and how important they say their religion is to them.

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 the 2018 GSS. To access the GSS cumulative data file in SDA format click [here](https://sda.berkeley.edu/sdaweb/analysis/?dataset=gss18). The cumulative data file contains all the data from each GSS survey conducted from 1972 through 2018. We want to use only the data that was collected in 2018. To select out the 2018 data, enter *year (2018)* in the SELECTION FILTER(S) box. Your screen should look like this.

  
  
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.

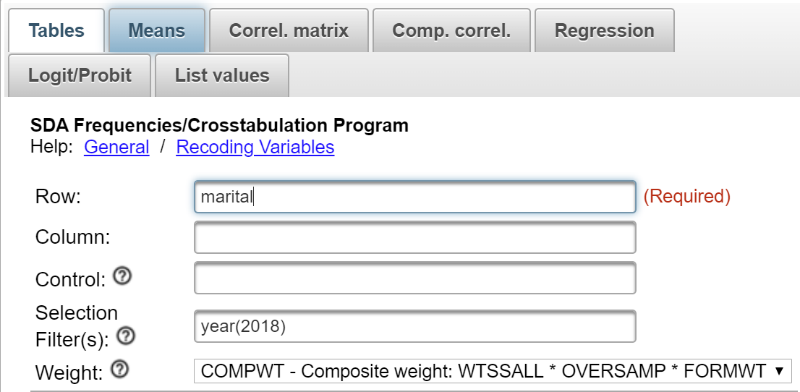
The GSS is an example of a social survey. The investigators selected a sample from the population of all adults in the United States. This particular survey was conducted in 2018 and is a relatively large sample of a little more than 2,300 adults. In a survey we ask respondents questions and use their 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. Often we want to describe respondents in terms of social characteristics such as marital status, education, and age. These are all variables in the GSS.

These measures are often classified in terms of their levels of measurement. S. S. Stevens described measures as falling into one of four categories – nominal, ordinal, interval, or ratio.[[1]](#footnote-1)

Here’s a brief description of each level.

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. For example, we could classify individuals by their marital status. Individuals could be married or widowed or divorced or separated or never married. Our categories should be mutually exclusive and exhaustive. Mutually exclusive means that every individual can be sorted into one and only one category. Exhaustive means that every individual can be sorted into a category. We wouldn’t want to use single as one of our categories because some people who are single can also be divorced and therefore could be sorted into more than one category. We wouldn’t want to leave widowed off our list of categories because then we wouldn’t have any place to sort these individuals.

The categories in a nominal level measure have no inherent order to them. This means that it wouldn’t matter how we ordered the categories. They could be arranged in any number of different ways. Run FREQUENCIES in SDA for the variable marital so you can see the frequency distribution for a nominal level variable. It wouldn’t matter how we ordered these categories. To run the frequency distribution, enter the variable name, marital, in the ROW box. Notice that the SELECTION FILTER(S) box and the WEIGHT box are both filled in. Your screen should look like this. To run the table, click RUN at the bottom.

  
  
An ordinal measure is a nominal measure in which the categories are ordered from low to high or from high to low. We could classify individuals in terms of the highest educational degree they achieved. Some individuals did not complete high school; others graduated from high school but didn’t go on to college. Other individuals completed a two-year junior college degree but then stopped college. Still others completed their bachelor’s degree and others went on to graduate work and completed a master’s degree or their doctorate. These categories are ordered from low to high.

But notice that while the categories are ordered they lack an equal unit of measurement. That means, for example, that the differences between categories are not necessarily equal. Run FREQUENCIES in SDA for degree. Look at the categories. The GSS assigned values (i.e., numbers) to these categories in the following way:

0 = less than high school,  
1 = high school degree,   
2 = junior college,  
3 = bachelors, and  
4 = graduate.

The difference in education between the first two categories is not the same as the difference between the last two categories. We might think they are because 0 minus 1 is equal to 3 minus 4 but this is misleading. These aren’t really numbers. They’re just symbols that we have used to represent these categories. We could just as well have labeled them a, b, c, d, and e. They don’t have the properties of real numbers. They can’t be added, subtracted, multiplied, and divided. All we can say is that b is greater than a and that c is greater than b and so on.

An interval measure is an ordinal measure with equal units of measurement. For example, consider temperature measured in degrees Fahrenheit. Now we have equal units of measurement – degrees Fahrenheit. The difference between 20 degrees and 40 degrees is the same as the difference between 70 degrees and 90 degrees. Now the numbers have the properties of real numbers and we can add and subtract them. But notice one thing about the Fahrenheit scale. There is no absolute zero point. There can be both positive and negative temperatures. That means that we can’t compare values by taking their ratios. For example, we can’t divide 80 degrees Fahrenheit by 40 degrees and conclude that 80 is twice as hot at 40. To do that we would need a measure with an absolute zero point.[[2]](#footnote-2)

A ratio measure is an interval measure with an absolute zero point. Run FREQUENCIES for sibs which is the number of siblings. This variable has an absolute zero point and all the properties of nominal, ordinal, and interval measures and therefore is a ratio variable.

Notice that level of measurement is itself ordinal since it is ordered from low (nominal) to high (ratio). It’s what we call a cumulative scale. Each level of measurement adds something to the previous level.

Why is level of measurement important? One of the things that helps us decide which statistic to use is the level of measurement of the variable(s) involved. For example, we might want to describe the central tendency of a distribution. If the variable was nominal, we would use the mode. If it was ordinal, we could use the mode or the median. If it was interval or ratio, we could use the mode or median or mean. Central tendency will be the focus of Exercise 2.

Run FREQUENCIES for the following variables:

* satfin,
* wealth.
* happy,
* partyid,
* relig,
* denom,
* reliten,
* nummen,
* numwomen,
* premarsx, and
* age.

For each variable, decide which level of measurement it represents and write a sentence or two indicating why you think it is that level. Keep in mind that we’re only considering those responses that actually answered the question. Not all respondents answered the question. Some said they didn’t know or refused to answer the question. These respondents are assigned missing values to indicate they didn’t answer the question and are not included in the frequency distribution.

# Exercises for an Introductory Statistics Course Using SDA Edward Nelson, California State University, Fresno

# Exercise 2 Measures of Central Tendency and Dispersion

## Goals of Exercise

The goal of this exercise is to explore measures of central tendency (mode, median, and mean) and dispersion (range, standard deviation, and variance). The exercise also gives you practice in using FREQUENCIES in SDA.

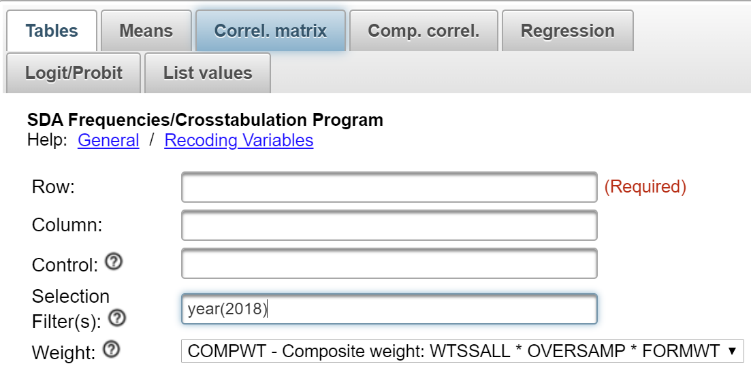
## Part I – Measures of Central Tendency

Data analysis always starts with describing variables one-at-a-time.  Sometimes this is referred to as univariate (one-variable) analysis.  Central tendency refers to the center of the distribution.

There are three commonly used measures of central tendency – the mode, median, and mean of a distribution.  The mode is the most common value or values in a distribution.[[3]](#footnote-3)

The median is the middle value of a distribution.[[4]](#footnote-4)  The mean is the sum of all the values divided by the number of values.

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 the 2018 GSS.  To access the GSS cumulative data file in SDA format click [here](https://sda.berkeley.edu/sdaweb/analysis/?dataset=gss18).  The cumulative data file contains all the data from each GSS survey conducted from 1972 through 2018.  We want to use only the data that was collected in 2018.  To select out the 2018 data, enter *year(2018)* in the Selection Filter(s) box.  Your screen should look like the following.

  
  
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.

The GSS is an example of a social survey.  The investigators selected a sample from the population of all adults in the United States.  This particular survey was conducted in 2018 and is a relatively large sample of a little more than 2,300 adults.  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.  Often we want to describe respondents in terms of social characteristics such as marital status, education, and age.  These are all variables in the GSS.

|  |
| --- |
| Run FREQUENCIES in SDA for the variable *sibs*.  To run the frequency distribution, enter the variable name, *sibs*, in the ROW box.  Notice that the SELECTION FILTER(S) box and the WEIGHT box are both filled in. Your screen should look like the following.  This is the Tables dialog box with sibs entered in the Row box and year(2018) entered in the Selection Filter box. |

Once you have selected this variable, click on the arrow next to OUTPUT OPTIONS and check the box for SUMMARY STATISTICS.  Then click on CHART OPTIONS and click the arrow next to TYPE OF CHART.  Select BAR CHART and now click on RUN THE TABLE at the bottom.

Your output will display the frequency distribution for *sibs*, the summary statistics, and the bar chart.  Three of the summary statistics are commonly used measures of central tendency – mode, median, and mean.

* Mode = 1 meaning that two brothers and sisters was the most common answer (21.5%) from the 2,342 respondents who answered this question.  However, not far behind are those with two siblings (20.6%) and those with three siblings (16,2%).  So while technically one sibling is the mode, what you really found is that the most common values are one, two, and three siblings.  Another part of your output is the bar chart which is a chart or graph of the frequency distribution.  The bar chart clearly shows that one, two, and three are the most common values (i.e., the highest bars in the bar chart).  So we would want to report that these three categories are the most common responses.
* Median = 3 which means that three siblings is the middle category in this distribution.  The middle category is the category that contains the 50th percentile which is the value that divides the distribution into two equal parts.   In other words, it’s the value that has 50% of the cases above it and 50% of the cases below it.  If you added up the percents for all values less than 3 and the percents for all values less than or equal to  3, you would find that 46.5% of the cases have two or fewer siblings and that 62.7% of the cases have three or fewer siblings.  So the middle case (i.e., the 50th percentile) falls somewhere in the category of three siblings.  That is the median category.
* Mean = 3.47 which is the sum of all the values in the distribution divided by the number of responses.  If you were to sum all these values that sum would be 8,137.  Dividing that by the number of responses or 2,342 will give you the mean of 3.47.

Part II – Deciding Which Measure of Central Tendency to Use

The first thing to consider is the level of measurement (nominal, ordinal, interval, ratio) of your variable (see Exercise 1).

* If the variable is nominal, you have only one choice.  You must use the mode.
* If the variable is ordinal, you could use the mode or the median.  You should report both measures of central tendency since they tell you different things about the distribution.  The mode tells you the most common value or values while the median tells you where the middle of the distribution lies.
* If the variable is interval or ratio, you could use the mode or the median or the mean.  Now it gets a little more complicated.  There are several things to consider.
  + How skewed is your distribution?[[5]](#footnote-5)  Go back and look at the bar chart for *sibs*. Notice that there is a long tail to the right of the distribution.  Most of the values are at the lower end – one, two, and three siblings.  But there are quite a few respondents who report having four or more siblings and about 6% said they have ten or more siblings.  That’s what we call a positively skewed distribution where there is a long tail towards the right or the positive direction. Now look at the median and mean.  The mean (3.47) is larger than the median (3.0).  The respondents with lots of siblings pull the mean up.  That’s what happens in a skewed distribution.  The mean is pulled in the direction of the skew.  The opposite would happen in a negatively skewed distribution.  The long tail would be towards the left and the mean would be lower than the median.  In a heavily skewed distribution the mean is distorted and pulled considerably in the direction of the skew.  So consider reporting only the median in a heavily skewed distribution.  That’s why you almost always see median income reported and not mean income.  Imagine what would happen if your sample happened to include Bill Gates.  The income distribution would have this very, very large value which would pull the mean up but not affect the median.
  + Is there more than one clearly defined peak in your distribution?   The number of siblings has one clearly defined peak – one, two and three siblings.  But what if there is more than one clearly defined peak?  For example, consider a hypothetical distribution of 100 cases in which there 50 cases with a value of two and fifty cases with a value of 8.  The median and mean would be five but there are really two centers of this distribution – two and eight.  The median and the mean aren’t telling the correct story about the center. You’re better off reporting the two clearly defined peaks of this distribution and not reporting the median and mean.
  + If your distribution is normal in appearance then the mode, median, and mean will all be about the same.  A normal distribution is a perfectly symmetrical distribution with a single peak in the center.  No empirical distribution is perfectly normal but distributions often are approximately normal.

Run FREQUENCIES for the following variables.  Once you have selected the variables in the ROW box, ask for the SUMMARY STATISTICS and a BAR CHART.  For each variable write a sentence or two indicating which measure(s) of central tendency would be appropriate to use to describe the center of the distribution and what the values of those statistics mean.  For some variables there will be more than one appropriate measure of central tendency.

* *happy*
* *partyid*
* *reliten*
* *nummen*
* *numwomen*
* *age*

## Part III – Measures of Dispersion or Variation

Dispersion or variation refers to the degree that values in a distribution are spread out or dispersed.  The measures of dispersion that we’re going to discuss are only appropriate for interval and ratio level variables (see Exercise 1.)[[6]](#footnote-6) We’re going to discuss three such measures – the range, the variance, and the standard deviation.

The range is the difference between the highest and the lowest values in the distribution.  Run FREQUENCIES for age and compute the range by looking at the frequency distribution.  You can also ask SDA to compute it for you.  Once you have selected this variable click on the arrow next to OUTPUT OPTIONS and check the box for SUMMARY STATISTICS.   Now click on RUN THE TABLE at the bottom.

The range should equal 71 which is 89 – 18.  The range is not a very stable measure since it depends on the two most extreme values – the highest and lowest values.  These are the values most likely to change from sample to sample.

The variance is the sum of the squared deviations from the mean divided by the number of cases minus 1 and the standard deviation is just the square root of the variance.  Your instructor may want to go into more detail on how to calculate the variance by hand.  SDA will also calculate it for you.  The variance should equal 313.65 and the standard deviation will equal 17.71.

The variance and the standard deviation can never be negative.  A value of 0 means that there is no variation or dispersion at all in the distribution.  All the values are the same.  The more variation there is, the larger the variance and standard deviation.

So what does the variance (313.65) and the standard deviation (17.71) of the age distribution mean?  That’s hard to answer because you don’t have anything to compare it to.  But if you knew the standard deviation for both men and women you would be able to determine whether men or women have more variation.  Instead of comparing the standard deviations for men and women you would compute a statistic called the Coefficient of Relative Variation (CRV).  CRV is equal to the standard deviation divided by the mean of the distribution.   A CRV of 2 means that the standard deviation is twice the mean and a CRV of 0.5 means that the standard deviation is one-half of the mean.  You would compare the CRV’s for men and women to see whether men or women have more variation relative to their respective means.

You might also have wondered why you need both the variance and the standard deviation when the standard deviation is just the square root of the variance.  You’ll just have to take my word for it that you will need both as you go further in statistics.

Run FREQUENCIES for the following variables.  Once you have selected the variables in the ROW box, ask for the SUMMARY STATISTICS.  For each variable write a sentence or two indicating what the values of these statistics are for each of the variables and what the values of those statistics mean.  Compare the relative variation for the number of male sex partners and the number of female sex partners by comparing the CRV’s for each variable.

* *nummen*
* *numwomen*
* *sibs*

# Exercises for an Introductory Statistics Course Using SDA Edward Nelson, California State University, Fresno

# Exercise 3 Measures of Skewness and Kurtosis

## **Goals of Exercise**

The goal of this exercise is to explore measures of skewness and kurtosis. The exercise also gives you practice in using FREQUENCIES in SDA.

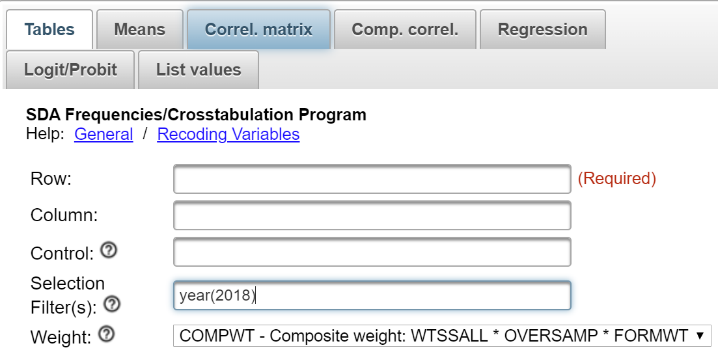
## **Part I – Measures of Skewness**

A normal distribution is a unimodal (i.e., single peak) distribution that is perfectly symmetrical.  In a normal distribution the mean, median, and mode are all equal.  Here’s a graph showing what a normal distribution looks like.

The horizontal axis is marked off in terms of standard scores where a standard score tells us how many standard deviations a value is from the mean of the normal distribution.  So a standard score of +1 is one standard deviation above the mean and a standard score of -1 is one standard deviation below the mean.  The percents tell us the percent of cases that you would expect between the mean and a particular standard score if the distribution was perfectly normal.  You would expect to find approximately 34% of the cases between the mean and a standard score of +1 or -1.  In a normal distribution, the mean, median, and mode are all equal and are at the center of the distribution.  So the mean always has a standard score of zero.

Skewness measures the deviation of a particular distribution from this symmetrical pattern.  In a skewed distribution one side has longer or fatter tails than the other side.  If the longer tail is to the left, then it is called a negatively skewed distribution.  If the longer tail is to the right, then it is called a positively skewed distribution.  One way to remember this is to recall that any value to the left of zero is negative and any value to the right of zero is positive.  Here are graphs of positively and negatively skewed distributions compared to a normal distribution.

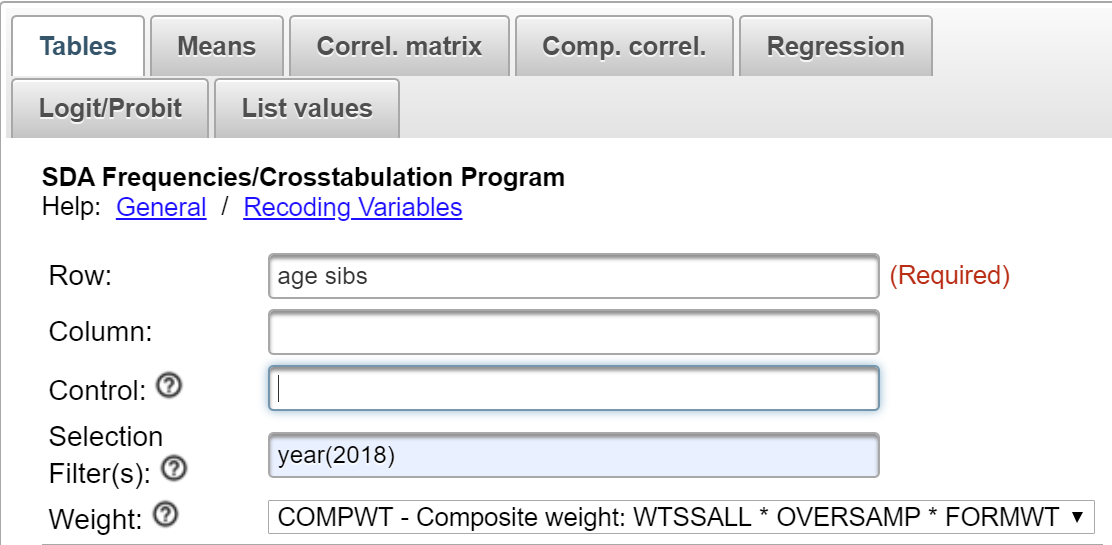
We’re going to use the General Social Survey (GSS) for this exercise.  The GSS is a national probability sample of adults in the United States conducted by the National Opinion Research Center (NORC).  The GSS started in 1972 and has been an annual or biannual survey ever since. For this exercise we’re going to use the 2018 GSS.  To access the GSS cumulative data file in SDA format click [here](https://sda.berkeley.edu/sdaweb/analysis/?dataset=gss18).  The cumulative data file contains all the data from each GSS survey conducted from 1972 through 2018.  We want to use only the data that was collected in 2018.  To select out the 2018 data, enter year(2018) in the Selection Filter(s) box.  Your screen should look like the following.



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.

The GSS is an example of a social survey.  The investigators selected a sample from the population of all adults in the United States.  This particular survey was conducted in 2018 and is a relatively large sample of a little more than 2,300 adults.  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.  Often we want to describe respondents in terms of social characteristics such as marital status, education, and age.  These are all variables in the GSS.

Run FREQUENCIES in SDA for the variables age and sibs. To run the frequency distributions, enter the variable names, age and sibs, in the ROW box.  Separate the variable names by either a space or a comma.  Notice that the SELECTION FILTER(S) box and the WEIGHT box are both filled in. Your screen should like the following.

  
  
Once you have selected these variables click on the arrow next to OUTPUT OPTIONS and check the box for SUMMARY STATISTICS.  Then click on CHART OPTIONS and click the arrow next to TYPE OF CHART.  Select BAR CHART and now click on RUN THE TABLE at the bottom.  SDA will compute the mean and median (plus other statistics) for each variable along with the bar chart.

Notice that the mean is larger than the median for both variables.  This means that the distribution is positively skewed.  But also notice that the mean for sibs is quite a bit larger than the median in a relative sense than is the case for age.  This suggests that the distribution for sibs is the more skewed of the two variables.  Look at the bar charts and you’ll see the same thing.  Both variables are positively skewed but sibs is the more skewed variable.  Now look at the skewness values — 1.72 for sibs and .30 for age.  The larger the skewness value, the more skewed the distribution.  Positive skewness values indicate a positive skew and negative values indicate a negative skew.  There are various rules of thumb suggested for what constitutes a lot of skew but for our purposes we’ll just say that the larger the value, the more the skewness and the sign of the value indicates the direction of the skew.

Run FREQUENCIES for the following variables.  Tell SDA to give you the bar chart along with the summary statistics.  Write a paragraph for each variable explaining what these statistics tell you about the skewness of the variables.

* hrsrelax
* tvhours

### **Part II – Measures of Kurtosis**

Kurtosis refers to the flatness or peakness of a distribution relative to that of a normal distribution.  Distributions that are flatter than a normal distribution are called platykurtic and distributions that are more peaked are called leptokurtic.

SDA will compute a kurtosis measure.  Negative values indicate a platykurtic distribution and positive values indicate a leptokurtic distribution.  The larger the kurtosis value, the more peaked or flat the distribution is.

Look back at the output for age and sibs.  For age the kurtosis value was -.82 indicating a flatter distribution and for sibs kurtosis was 5.12 indicating a more peaked distribution.  To see this visually look at your bar charts.

Run FREQUENCIES for the following variables.  Tell SDA to give you the summary statistics and a bar chart.  Write a paragraph for each variable explaining what these statistics tell you about the kurtosis of the variables.

* maeduc
* paeduc
* sexfreq

# Exercises for an Introductory Statistics Course Using SDA Edward Nelson, California State University, Fresno

# Exercise 4 Graphs and Charts

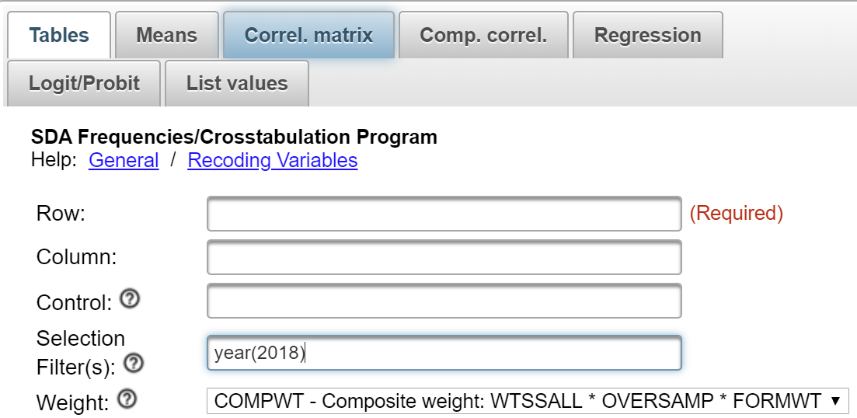
## **Goals of Exercise**

The goal of this exercise is to explore graphs and charts for frequency distributions. The exercise also gives you practice in using FREQUENCIES in SDA.

## **Part I – Pie Charts**

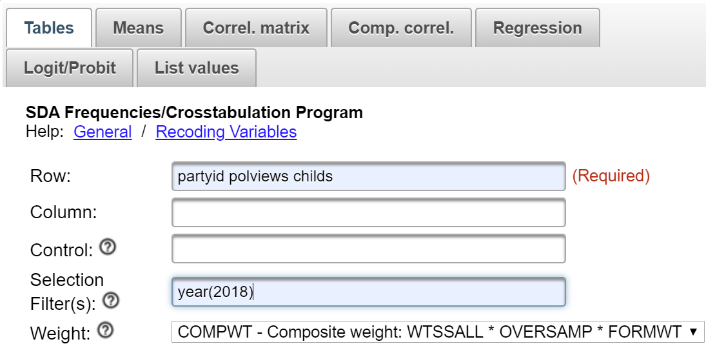
A pie chart is a chart that shows the frequencies or percents of a variable with a small number of categories.  It is presented as a circle divided into a series of slices.  The area of each slice is proportional to the number of cases or the percent of cases in each category.  It is normally used with nominal or ordinal variables (see Exercise 1).

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 the 2018 GSS.  To access the GSS cumulative data file in SDA format click [here](https://sda.berkeley.edu/sdaweb/analysis/?dataset=gss18).  The cumulative data file contains all the data from each GSS survey conducted from 1972 through 2018.  We want to use only the data that was collected in 2019.  To select out the 2018 data, enter year(2018) in the Selection Filter(s) box.  Your screen should look like the following.



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.

The GSS is an example of a social survey.  The investigators selected a sample from the population of all adults in the United States.  This particular survey was conducted in 2018 and is a relatively large sample of a little more than 2,300 adults.  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.  Often we want to describe respondents in terms of social characteristics such as marital status, education, and age.  These are all variables in the GSS.

Run FREQUENCIES in SDA for the variables partyid, polviews, and childs. To run the frequency distributions, enter the variable names in the ROW box.  Separate the variable names by either a space or a comma.  Notice that the SELECTION FILTER(S) box and the WEIGHT box are both filled in. Your screen should like the following.  
  
  
  
Once you have selected the variables, click on the arrow next to CHART OPTIONS and click the arrow next to TYPE OF CHART and select PIE CHART.  Click also on the box to SHOW PERCENTS so SDA will print the percents on the pie chart.  Now click on RUN THE TABLE at the bottom.  SDA will draw the pie chart for each of these variables.  Write a sentence or two for each variable describing the distributions based on these pie charts.

If you are wondering why you shouldn’t use pie charts for variables with a large number of categories, create a pie chart for age and you’ll see why.  There are so many categories for the variable age that the pie chart is unreadable.  The solution is to combine the various ages into a smaller number of categories.  This is called recoding.  Let’s divide age into four categories:  18 (the youngest age in the sample) to 34, 35 to 49, 50 to 64, and 65 to 89 (the oldest age in the sample).  We can do this be entering the following in the ROW box – age(r:1=18-34 "under 35"; 2=35-49 "35 to 49"; 3=50-64 "50 to 64"; 4=65-89 "65 and older").  (Note: there is no period at the end of the command.)

Note the syntax for the recode statement since you will want to use it later.

* First you enter r: followed by the variable name which is age.  The r stands for recode.
* Then you indicate the new value you want to assign to the first category which is 1.
* Then you put the values that you are combining which is 18-34 for the first category.  These values must be separated by a dash (i.e., hyphen).
* This is followed by the label you want to assign to this category enclosed in double quotation marks which is “under 35” for the first category.  This is free form meaning you can put what you want for the label.
* This is separated from another recoded category by a semi-colon.
* Finally, the entire recode specification is in parentheses.

Enter this into the ROW box and then click on RUN THE TABLE to produce your recoded pie chart for age.  Write a sentence or two for this variable describing the distribution based on the bar chart.

## **Part II – Bar Charts**

A bar chart is a chart that shows the frequencies or percents of a variable and is presented as a series of vertical bars.  The height of each bar is proportional to the number of cases or the percent of cases in each category.  It is normally used with nominal or ordinal variables (see Exercise 1).

Run FREQUENCIES for the variables partyid, polviews, childs, age.  Be sure to enter the recode instruction for age.  This time click on the arrow next to CHART OPTIONS and click the arrow next to TYPE OF CHART and select BAR CHART.  Click also on the box to SHOW PERCENTS so SDA will print the percents on the bar chart.  Now click on RUN THE TABLE to produce the bar charts.  There is a small problem with childs.  One of the categories is “eight or more” children.  That means we don’t know what these values actually are.  They could be 8 or 10 or 12 or 14 or something else.  Since there are so few cases in this category, we’re going to ignore this problem.

Write a sentence or two for each variable describing the distributions based on the bar charts.

## **Part III – Stacked Bar Charts**

SDA will also produce what it calls a stacked bar chart.  To get a stacked bar chart, click on the arrow next to CHART OPTIONS and click the arrow next to TYPE OF CHART and select STACKED BAR CHART.  Now click on RUN THE TABLE to get the stacked bar charts for the variables in Part II.  Write a short paragraph describing the stacked bar chart and how it is different from a bar chart.  Which do you prefer – bar charts or stacked bar charts?  Why?

## **Part IV – Line Charts**

The last kind of chart that SDA will produce is called a LINE CHART.  To get a line chart, click on the arrow next to CHART OPTIONS and click the arrow next to TYPE OF CHART and select LINE CHART.  Now click on RUN THE TABLE to get the line charts for the variables in Parts II and III.  Write a short paragraph describing the line chart and how it is different from the other types of charts.  Do you think a line chart is clearer than or not as clear as the other types of charts?  Why?

## **Part V – Conclusions**

We have talked about four different types of graphs – pie charts, bar charts, stacked bar charts, and line charts.  Are there limitations on when you should use a particular type of chart?  Why?

# Exercises for an Introductory Statistics Course Using SDA Edward Nelson, California State University, Fresno

# Exercise 5 Comparing Means and Testing Hypotheses

## **Goals of Exercise**

The goal of this exercise is to compare means and test hypotheses. The exercise also gives you practice in using MEANS in SDA.

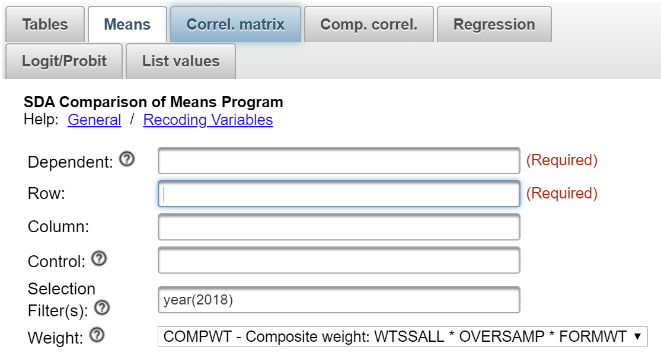
## **Part I – Computing Means**

Populations are the complete set of objects that we want to study.  For example, a population might be all the individuals that live in the United States at a particular point in time.  The U.S. does a complete enumeration of all individuals living in the United States every ten years (i.e., each year ending in a zero).  We call this a census.  Another example of a population is all the students in a particular school or all college students in your state.  Populations are often large and it’s too costly and time consuming to carry out a complete enumeration.  So what we do is to select a sample from the population where a sample is a subset of the population and then use the sample data to make an inference about the population.

A statistic describes a characteristic of a sample while a parameter describes a characteristic of a population.  The mean age of a sample is a statistic while the mean age of the population is a parameter.   We use statistics to make inferences about parameters.  In other words, we use the mean age of the sample to make an inference about the mean age of the population.  Notice that the mean age of the sample (our statistic) is known while the mean age of the population (our parameter) is usually unknown.

There are many different ways to select samples.  Probability samples are samples in which every object in the population has a known, non-zero, chance of being in the sample (i.e., the probability of selection).  This isn’t the case for non-probability samples.  An example of a non-probability sample is an instant poll which you hear about on radio and television shows.  A show might invite you to go to a website and answer a question such as whether you favor or oppose same-sex marriage.  This is a purely volunteer sample and we have no idea of the probability of selection.

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 the 2018 GSS.  To access the GSS cumulative data file in SDA format click [here](https://sda.berkeley.edu/sdaweb/analysis/?dataset=gss18).  The cumulative data file contains all the data from each GSS survey conducted from 1972 through 2018.  We want to use only the data that was collected in 2018.  To select out the 2018 data, enter year(2018) in the Selection Filter(s) box.  Your screen should look like the following.



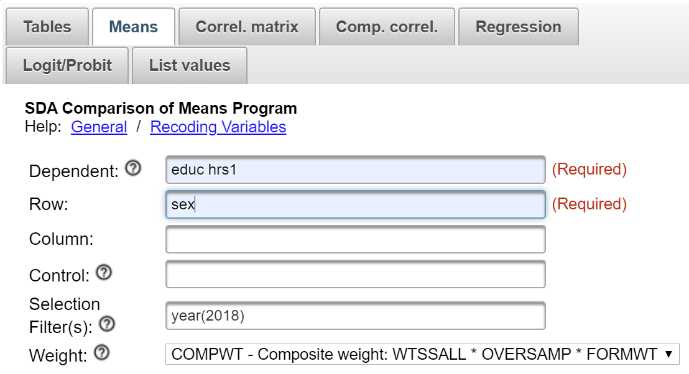
Notice that a weight variable has already been entered in the WEIGHT box.  This will weight the data so the sample better represents the population from which the sample was selected.  Note also that in the SAMPLE DESIGN line SRS should be selected.

The GSS is an example of a social survey.  The investigators selected a sample from the population of all adults in the United States.  This particular survey was conducted in 2018 and is a relatively large sample of a little more than 2,300 adults.  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.  Often we want to describe respondents in terms of social characteristics such as marital status, education, and age.  These are all variables in the GSS.

Let’s start by asking two questions.

* Do men and women differ in the number of years of school they have completed?
* Do men and women differ in the number of hours they worked in the last week?

Let’s start by getting the means for these variables.  Click on MEANS in the menu bar at the top of SDA and enter the variables educ and hrs1 in the DEPENDENT box.  The dependent variable will always be the variable for which you are going to compute means.  Then enter the variable *sex* in the ROW box.  This is the variable which defines the groups you want to compare.  In our case we want to compare men and women.  The output from SDA will show you the means for the two dependent variables for both men and women.  Notice that you must enter one or more variables in both the DEPENDENT and ROW boxes.  That what it means when it says REQUIRED next to these boxes.  Separate the variable names by either a space or a comma.  Notice that the SELECTION FILTER(S) box and the WEIGHT box are both filled in.  Click RUN THE TABLE to produce the means. Your screen should like the following.



Men and women differ very little in the number of years of school they completed.  Women have completed a little less than one-tenth of a year more than men.  But men worked quite a bit more than women in the last week – a difference of a little more than six hours.  By the way, only respondents who are employed are included in this calculation but both part-time and full-time employees are included.

Why can’t we just conclude that men and women have about the same education and that men work more than women?  If we were just describing the **sample**, we could.  But what we want to do is to make inferences about differences between men and women in the **population**.  We have a sample of men and a sample of women and some amount of sampling error will always be present in both samples.  The larger the sample, the less the sampling error and the smaller the sample, the more the sampling error.  Because of this sampling error we need to make use of hypothesis testing.

## **Part II – Now it’s Your Turn**

In this part of the exercise you want to compare men and women to answer these two questions.

* Do men and women differ in the number of hours per day they have to relax?  This is variable hrsrelax in SDA.
* Do men and women differ in the number of hours per day they watch television?  This is variable tvhours in SDA.

Use SDA to get the sample means and then compare them to begin answering these questions.  Write a couple sentences describing the differences between men and women for both variables.

## **Part III – Hypothesis Testing – Independent-Samples t Test**

In Part I we compared the mean scores for men and women for the following variables.

* educ
* hrs1

A t test is used when you want to compare **two** groups.  That means that your row variable must be a dichotomy consisting of only two categories.  The variable sex is a dichotomy.  It has only two categories – male (value 1) and female (value 2).  But any variable can be made into a dichotomy by recoding.  For example, the variable *satfin* (satisfaction with financial situation) has three categories – satisfied (value 1), more or less satisfied (value 2), and not at all satisfied (value 3). You could recode satfin to combine values 1 and 2 which would then give you a dichotomy (i.e., satisfied and not satisfied).

Click on MEANS in the menu bar at the top and enter the variables educ and hrs1 in the DEPENDENT box as you did in Part I.  The SELECTION FILTER(S) box and the WEIGHT box should both be filled in.  Be sure to click on OUTPUT OPTIONS and both check SRS STD ERRS and uncheck COMPLEX STD ERRS.  Now we want to determine if the differences between men and women are statistically significant by carrying out the independent-samples t test.  Click on OUTPUT OPTIONS and check the box for ANOVA STATS under OTHER OPTIONS.  Finally, click RUN THE TABLE to carry out the procedure.

You’re probably wondering why we requested the ANOVA table.  (By the way, ANOVA means Analysis of Variance.[[7]](#footnote-7))  The t test is a special case of One-Way Analysis of Variance[[8]](#footnote-8) when your independent variable is a dichotomy.  In this special case, the F statistic in the Analysis of Variance table for education is the square of the t statistic.  So to get t all you have to do is to take the square root of 0.225 which is 0.474.  The degrees of freedom (df) is 2,343 which you can get from the ANOVA table.  The P value is the probability that you would be wrong if you rejected the null hypothesis that there was no difference between the means for the population of all men and the population of all women.[[9]](#footnote-9)

Notice how we are going about this.  We have a sample of adults in the United States (i.e., the 2018 GSS).  We calculate the mean years of school completed by men and women in the **sample** who answered the question.  But we want to test the hypothesis that the mean years of school completed by men and women in the **population** are different.  We’re going to use our sample data to test a hypothesis about the population.

The hypothesis we want to test is that the mean years of school completed by men in the population is different than the mean years of school completed by women in the population.  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.  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. We haven’t proven the research hypothesis, but we have support for this hypothesis.

Here are our two hypotheses.

* research hypothesis – the population mean for men minus the population mean for women does not equal 0.  In other words, they are different from each other.
* null hypothesis – the population mean for men minus the population mean for women equals 0.  In other words, they are not different from each other.

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

Now all we have to do is figure out how to use the t test to decide whether to reject or not reject the null hypothesis.  Look again at the P value which is 0.6350 for the t test.  That tells you that the probability of being wrong if you rejected the null hypothesis is just about 63 or 64 times out of one hundred.  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.

## **Part IV – Now it’s Your Turn Again**

In this part of the exercise you want to compare men and women to answer these two questions but this time you want to test the appropriate null hypotheses.

* Do men and women differ in the number of hours per day they have to relax?
* Do men and women differ in the number of hours per day they watch television?

Use the independent-sample t test to carry out this part of the exercise.  What are the research and the null hypotheses?  Do you reject or not reject the null hypotheses?  Explain why.

## **Part V – What Does Independent Samples Mean?**

Why do we call this t test the independent-samples t test?  Independent samples are samples in which the composition of one sample does not influence the composition of the other sample.  In this exercise we’re using the 2018 GSS which is a sample of adults in the United States.  If we divide this sample into men and women we would have a sample of men and a sample of women and they would be independent samples.  The individuals in one of the samples would not influence who is in the other sample.

Dependent samples are samples in which the composition of one sample does influence the composition of the other sample.  For example, if we have a sample of married couples and divide that sample into two samples of men and women, then the men in one of the samples determines who the women are in the other sample.  The composition of the samples is dependent on each other.  SDA does not include a test for dependent samples.

# Exercises for an Introductory Statistics Course Using SDA Edward Nelson, California State University, Fresno

# Exercise 6 Hypothesis Testing – One Way Analysis of Variance

## **Goals of Exercise**

The goal of this exercise is to explore hypothesis testing and one-way analysis of variance (sometimes abbreviated one-way anova). The exercise also gives you practice in using MEANS in SDA.

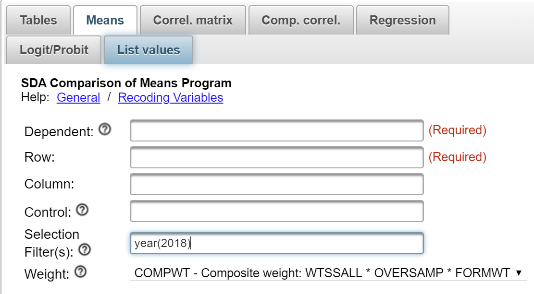
## **Part I – Populations and Samples**

Populations are the complete set of objects that we want to study.  For example, a population might be all the individuals that live in the United States at a particular point in time.  The U.S. does a complete enumeration of all individuals living in the United States every ten years (i.e., each year ending in a zero).  We call this a census.  Another example of a population is all the students in a particular school or all college students in your state.  Populations are often large and it’s too costly and time consuming to carry out a complete enumeration.  So what we do is to select a sample from the population where a sample is a subset of the population and then use the sample data to make an inference about the population.

A statistic describes a characteristic of a sample while a parameter describes a characteristic of a population.  The mean age of a sample is a statistic while the mean age of the population is a parameter.   We use statistics to make inferences about parameters.  In other words, we use the mean age of the sample to make an inference about the mean age of the population.  Notice that the mean age of the sample (our statistic) is known while the mean age of the population (our parameter) is usually unknown.

There are many different ways to select samples.  Probability samples are samples in which every object in the population has a known, non-zero, chance of being in the sample (i.e., the probability of selection).  This isn’t the case for non-probability samples.  An example of a non-probability sample is an instant poll which you hear about on radio and television shows.  A show might invite you to go to a website and answer a question such as whether you favor or oppose same-sex marriage.  This is a purely volunteer sample and we have no idea of the probability of selection.

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 the 2018 GSS.  To access the GSS cumulative data file in SDA format click [here](https://sda.berkeley.edu/sdaweb/analysis/?dataset=gss18).  The cumulative data file contains all the data from each GSS survey conducted from 1972 through 2018.  We want to use only the data that was collected in 2018.  To select out the 2018 data, enter year(2018) in the Selection Filter(s) box.  Your screen should look like the following.



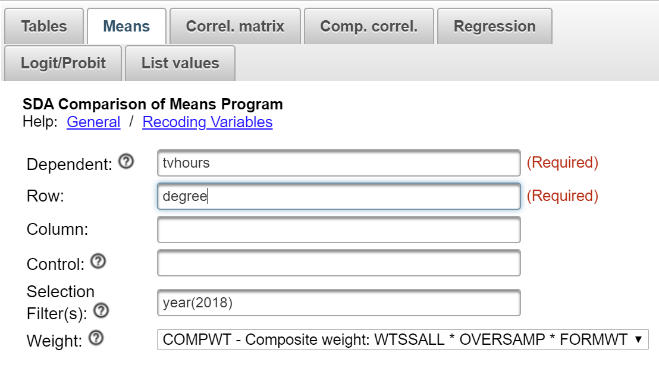
Notice that a weight variable has already been entered in the WEIGHT box.  This will weight the data so the sample better represents the population from which the sample was selected.  Note also that in the SAMPLE DESIGN line SRS should be selected.

The GSS is an example of a social survey.  The investigators selected a sample from the population of all adults in the United States.  This particular survey was conducted in 2018 and is a relatively large sample of a little more than 2,300 adults.  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.  Often we want to describe respondents in terms of social characteristics such as marital status, education, and age.  These are all variables in the GSS.

In Exercise 5 we used the t test to compare means from two independent samples.  But what if we wanted to compare means from more than two samples?  For that we need to use a statistical test called analysis of variance.  In fact, the t test is a special case of analysis of variance.

The 2018 GSS includes a variable (degree) that describes the highest degree in school that the person achieved.  The categories are less than high school, high school, junior college, bachelor’s degree, graduate degree.  Another variable is the number of hours per day that respondents say they watch television (tvhours).  We want to find out if there is any relationship between these two variables.  One way to answer this question would be to see if respondents with different levels of education watch different amounts of television.  For example, you might suspect that the more education respondents have, the less television they watch.

Let’s start by getting the means for tvhours broken down by degree.  Click on MEANS in the menu bar at the top of SDA and enter the variable tvhours in the DEPENDENT box.  The dependent variable will always be the variable for which you are going to compute means.  Then enter the variable degree in the ROW box.  This is the variable which defines the groups you want to compare.  In our case we want to compare respondents with different levels of education.  The output from SDA will show you the mean number of hours respondents watched television for each level of education.  Notice that you must enter one or more variables in both the DEPENDENT and ROW boxes.  That what it means when it says REQUIRED next to these boxes.  Your screen should like the following.



Notice that the SELECTION FILTER(S) box and the WEIGHT box are both filled in. Be sure to click on OUTPUT OPTIONS and both check SRS STD ERRS and uncheck COMPLEX STD ERRS.  Click RUN THE TABLE to produce the means.

Respondents with more education watch less television than those with less education.  For example, respondents with a graduate degree watch an average of 1.73 hours of television per day while those who haven’t completed high school watch an average of 3.42 hours – a difference of a little less than two hours.  Why can’t we just conclude that those with more education watch less television that those with less education?  If we were just describing the **sample**, we could. But what we want to do is to make inferences about differences in the **population**.  We have five samples from five different levels of education and some amount of sampling error will always be present in all these samples.  The larger the samples, the less the sampling error and the smaller the samples, the more the sampling error.  Because of this sampling error we need to make use of hypothesis testing as we did in Exercise 5.

## **Part II – Now it’s Your Turn**

In this part of the exercise you want to determine whether people who live in some regions of the country (region) watch more television (tvhours) than people in other regions.   Use SDA to get the sample means as we did in Part I and then compare them to begin answering this question.  Write one or two paragraphs describing the regions in which people watch more and less television.

## **Part III – Hypothesis Testing – One-Way Analysis of Variance**

In Part I we compared the mean number of hours of television watched per day for different levels of education.  Now we want to determine if these differences are statistically significant by carrying out a one-way analysis of variance.

Click on MEANS in the menu bar at the top of SDA and enter the variable tvhours in the DEPENDENT box as you did in Part I.  This time enter the variable degree in the ROW box.  The SELECTION FILTER(S) box and the WEIGHT box should both be filled in. Be sure to click on OUTPUT OPTIONS and both check SRS STD ERRS and uncheck COMPLEX STD ERRS.  Now we want to determine if the differences between levels of education are statistically significant by carrying out a one-way analysis of variance.  Click on OUTPUT OPTIONS and check the box for ANOVA STATS under OTHER OPTIONS.  Finally, click RUN THE TABLE to carry out the procedure.

Notice how we are going about this.  We have a sample of adults in the United States (i.e., the 2018 GSS).  We calculate the mean number of hours per day that respondents watch television for each level of education in the **sample**.  But we want to test the hypothesis that the amount respondents watch television varies by level of education in the **population**.  We’re going to use our sample data to test a hypothesis about the population.

Our hypothesis is that the mean number of hours watching television is higher for some levels of education than for other levels in the population. 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 mean number of hours watching television is the same for all levels of education in the population and call this the null 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. We haven’t proven the research hypothesis, but we have support for this hypothesis.

Here are our two hypotheses.

* research hypothesis – the mean number of hours watching television for at least one level of education is different from at least one other population mean.
* null hypothesis – the mean number of hours watching television is the same for all five levels of education in the population.

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

Now all we have to do is figure out how to use the F test to decide whether to reject or not reject the null hypothesis.  Look again at the significance value which is 0.0000.  By the way, it isn’t exactly 0.  This is a rounded value.  It means it is less than 0.00005.  That tells you that the probability of being wrong if you rejected the null hypothesis is really, really small.  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.

So what have we learned?  We learned that the mean number of hours watching television for at least one of the populations is different from at least one other population.  But which ones?  There are statistical tests for answering this question.  But we’re not going to cover that although your instructor might want to discuss these tests.

## **Part IV – Now it’s Your Turn Again**

In Part II you computed the mean number of hours that respondents watched television for each of the nine regions of the country.  Now we want to determine if these differences are statistically significant by carrying out a one-way analysis of variance as described in Part III.  Indicate what the research and null hypotheses are and whether you can reject the null hypothesis.  What does that tell you about the research hypothesis?

# Exercises for an Introductory Statistics Course Using SDA Edward Nelson, California State University, Fresno

# Exercise 7 Hypothesis Testing – Crosstabulation

## **Goals of Exercise**

The goal of this exercise is to introduce crosstabulation as a statistical tool to explore relationships between variables.  The exercise also gives you practice in using CROSSTABS in SDA.

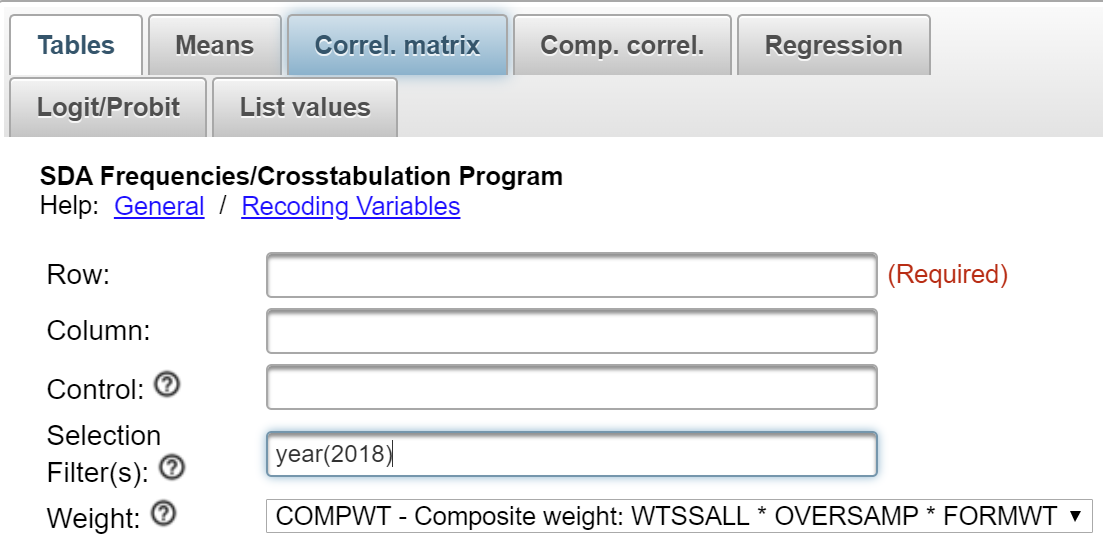
## **Part I—Relationships between Variables**

In Exercises 5 and 6 we used sample means to analyze relationships between variables.  For example, we compared men and women to see if they differed in the number of years of school completed and the number of hours they worked in the previous week and discovered that men and women had about the same amount of education but that men worked more hours than women.  We were able to compute means because years of school completed and hours worked are both ratio level variables (see Exercise 1). The mean assumes interval or ratio level measurement (see Exercise 2).

But what if we wanted to explore relationships between variables that weren’t interval or ratio?  Crosstabulation can be used to look at the relationship between nominal and ordinal variables.  Let’s compare men and women (sex) in terms of the following.

* opinion about abortion (abany)
* fear of crime (fear)
* satisfaction with current financial situation (satfin)
* opinion about gun control (gunlaw)
* gun ownership (owngun)
* voting (pres16),
* religiosity (reliten)

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 the 2018 GSS.  To access the GSS cumulative data file in SDA format click [here](https://sda.berkeley.edu/sdaweb/analysis/?dataset=gss18).  The cumulative data file contains all the data from each GSS survey conducted from 1972 through 2018.  We want to use only the data that was collected in 2018.  To select out the 2018 data, enter year(2018) in the SELECTION FILTER(S) box.  Your screen should look like the following.

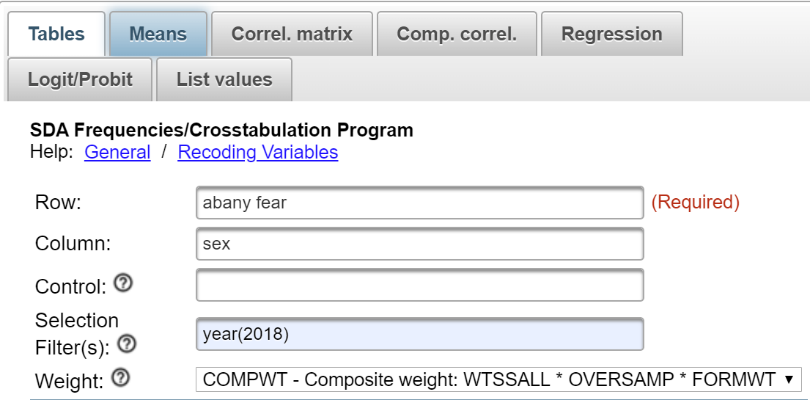


Notice that a weight variable has already been entered in the WEIGHT box.  This will weight the data so the sample better represents the population from which the sample was selected.

The GSS is an example of a social survey.  The investigators selected a sample from the population of all adults in the United States.  This particular survey was conducted in 2018 and is a relatively large sample of a little more than 2,300 adults.  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.  Often we want to describe respondents in terms of social characteristics such as marital status, education, and age.  These are all variables in the GSS.

Before we look at the relationship between sex and these other variables, we need to talk about independent and dependent variables.  The dependent variable is whatever you are trying to explain.  In our case, that could be how people feel about abortion, fear of crime, gun control and ownership, voting or religiosity.  The independent variable is some variable that you think might help you explain why some people think abortion should be legal and others think it shouldn’t be legal or any of the other variables in our list above.  In our case, that would be sex.  Normally we put the dependent variable in the row and the independent variable in the column.  We’ll follow that convention in this exercise.

Let’s start with the first two variables in our list.  We’re going to use abany as our measure of opinion about abortion.  Respondents were asked if they thought abortion ought to be legal for any reason.  And we’re going to use fear as our measure of fear of crime.  Respondents were asked if they were afraid to walk alone at night in their neighborhood.  Run CROSSTABS in SDA to produce two tables.  One will be for the relationship between sex and abany.  The other will be for sex and fear.  Put the independent variable in the column and the dependent variable in the row.  SDA can compute the row percents, column percents, and total percents.  Click on OUTPUT OPTIONS and look at PERCENTAGING.  By default, the box for column percents is already checked. Your screen should look like the following.



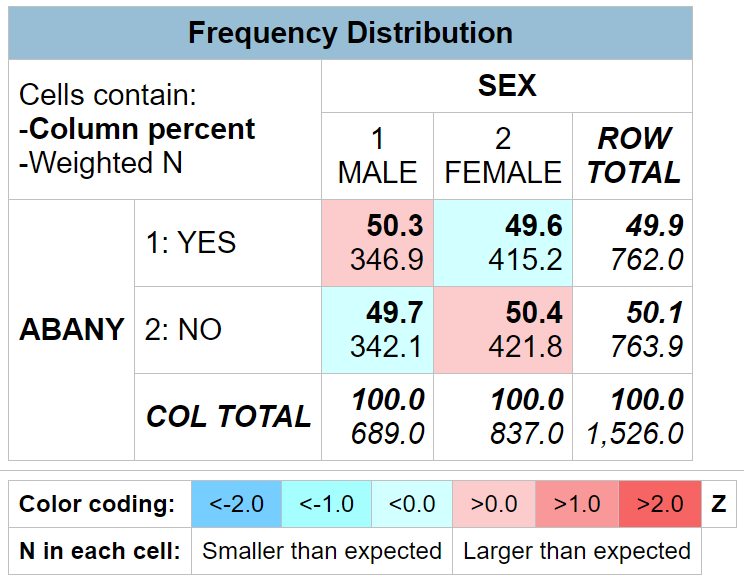
Your instructor will probably talk about how to compute these different percents.  But how do you know which percents to ask for?  Here’s a simple rule for computing percents.

* If your independent variable is in the column, then you want to use the column percents.
* If your independent variable is in the row, then you want to use the row percents.

Since you put the independent variable in the column, you want the column percents.

## **Part II – Interpreting the Percents**

Your first table should look like this.

  
  
It’s easy to make sure that you have the correct percents.  Your independent variable (sex) should be in the column and it is.  Column percents should sum down to 100% and they do.

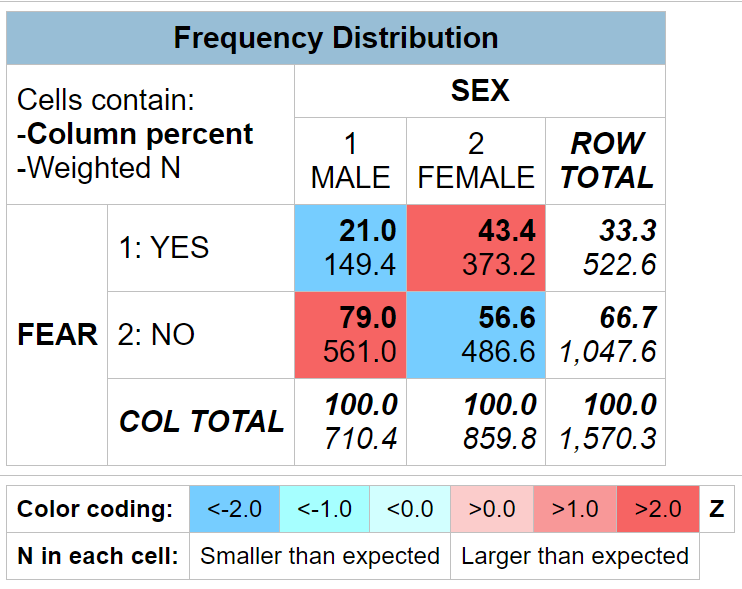
How are you going to interpret these percents?  Here’s a simple rule for interpreting percents.

* If your percents sum down to 100%, then compare the percents across.
* If your percents sum across to 100%, then compare the percents down.

Since the percents sum down to 100%, you want to compare across.

Look at the first row.  The table shows that 50.3% of men think abortion should be legal for any reason compared to 49.6% of women.  There’s a difference of 0.7% which is really 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.  So in this case we would conclude that there probably isn’t any difference in the population between men and women in their approval of abortion for any reason.

Now let’s look at your second table.



This time the percent difference is quite a bit larger.  The table shows that 21.0% of men are afraid to walk alone at night in their neighborhood compared to 43.4% of women.  This is a difference of 22.4%.  This is a much larger difference and we have reason to think that women are more fearful of being a victim of crime than men.

## **Part III – Now it’s Your Turn**

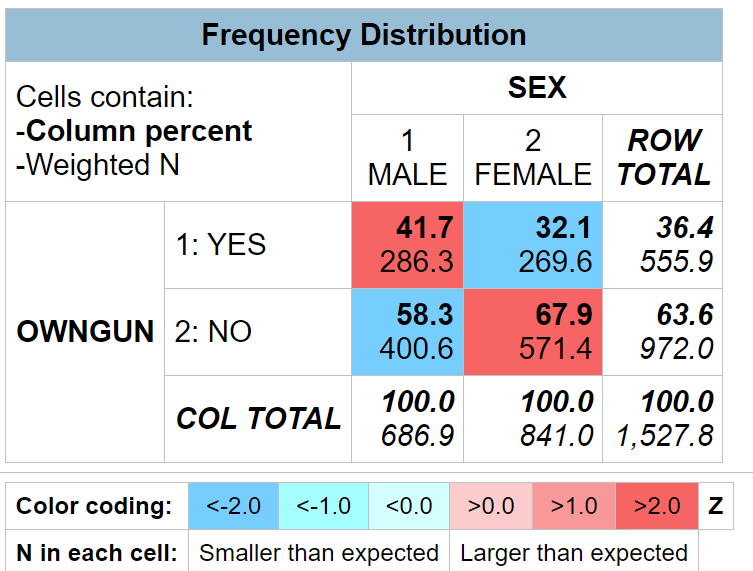
Choose two of the tables from the following list and compare men and women.

* satisfaction with current financial situation (satfin)
* opinion about gun control (gunlaw)
* gun ownership (owngun)
* religiosity (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.  What are values of the percents that you want to compare?  What is the percent difference?  Does it look to you that there is much of a difference between men and women in the variables you chose?

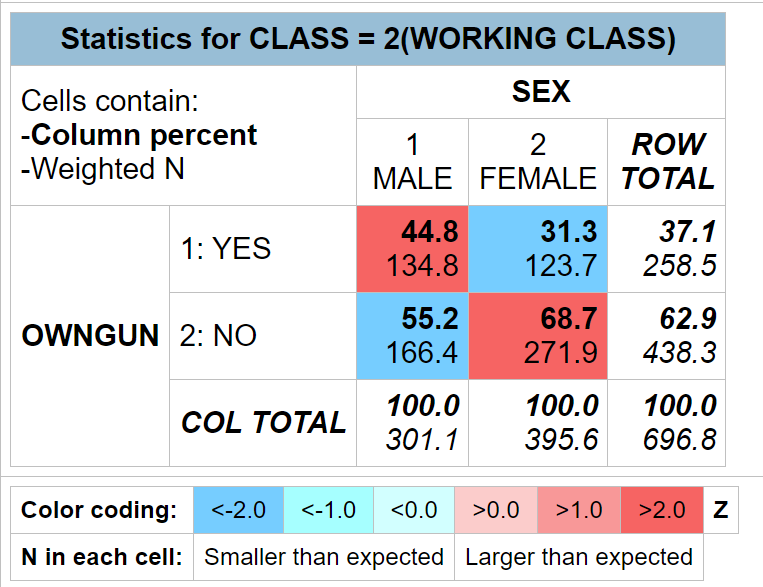
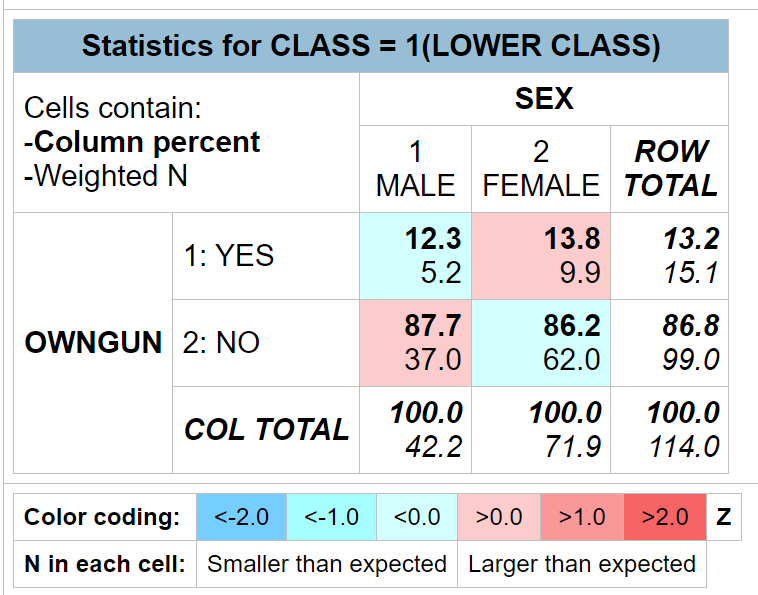
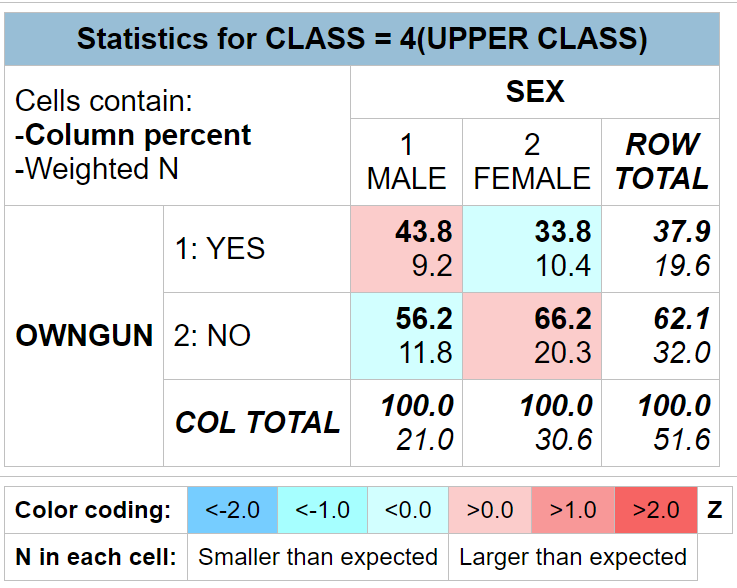
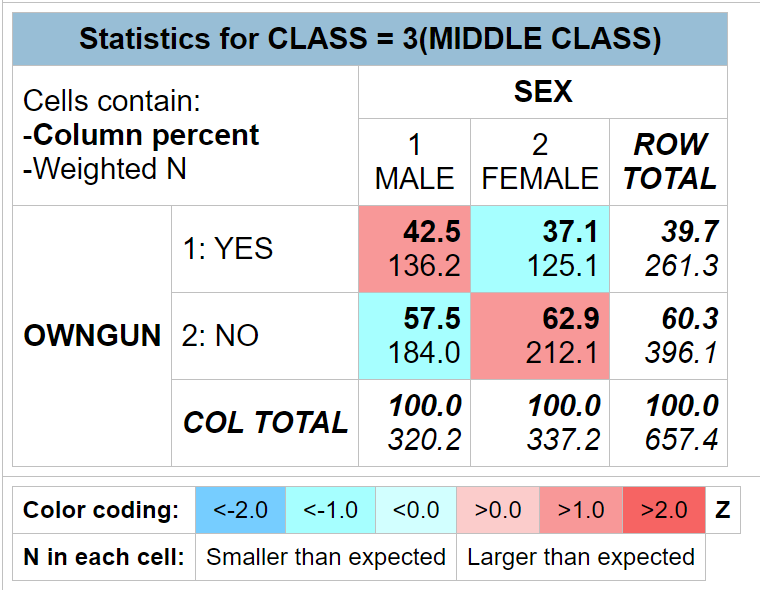
## **Part IV – Adding another Variable into the Analysis**

This time let’s focus on the difference between men and women (sex) in terms of gun ownership (owngun).  You might have run this table in Part III.  If not, run the crosstab now.  Here’s what your table should look like.



Men were more likely to own guns by 9.6 percentage points.  But what if we wanted to include social class in this analysis?  The 2018 GSS asked respondents whether they thought of themselves as lower, working, middle, or upper class.  This is the variable class.  What we want to do is to hold constant perceived social class.  In other words, we want to divide our sample into four groups with each group consisting of one of these four classes and then look at the relationship between sex and owngun separately for each of these four groups.

We can do this by going back to the SDA dialog box where we requested the crosstabulation and putting the variable class in the CONTROL box right below the COLUMN box.  Click on RUN THE TABLE to produce the results.  Your tables should look like the following.

This table is more complicated.  Notice that there are actually four tables.  We often call them partial tables since each table contains part of the total sample.  One of the tables is for those who said they were lower class, then working, middle and upper class.  Let’s look at the percent differences for each of these tables: -1.5 percentage points (lower), 13.5 percentage points (working), 5.4 percentage points (middle), and 10.0 percentage points (upper).  The biggest difference is for working class individuals, a moderate difference for middle and upper class, and virtually no difference for lower class. We need to remember not to make too much out of small differences because of sampling error.  In other words, when we look at only those who see themselves as lower class, there really isn’t much of a difference between men and women in terms of gun ownership buy there are larger differences for the other classes with men more likely than women to own guns.

But notice something else.  There are fewer people who say they are lower and upper class than say they are working or middle class.  There are only 114 respondents in the lower-class table and even fewer, 52 respondents, in the upper-class table.  We’ll have more to say about this in the next exercise (Exercise 8).

## **Part V – Now it’s Your Turn Again**

In Part II we compared men and women (sex) in terms of fear of crime (fear).  Run this table again but this time add social class (class) into the analysis as a control variable as we did in Part IV.  What happens to the percent difference when you hold constant class?  What does this tell you?

# Exercises for an Introductory Statistics Course Using SDA Edward Nelson, California State University, Fresno

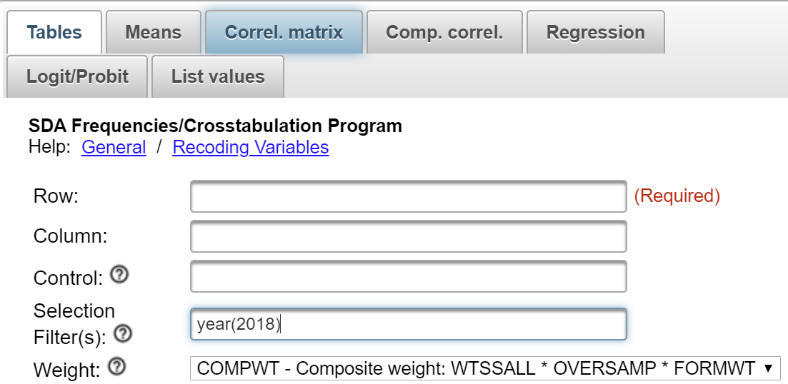
# Exercise 8 Hypothesis Testing – Chi Square

## 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 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 the 2018 GSS.  To access the GSS cumulative data file in SDA format click [here](https://sda.berkeley.edu/sdaweb/analysis/?dataset=gss18).  The cumulative data file contains all the data from each GSS survey conducted from 1972 through 2018.  We want to use only the data that was collected in 2018.  To select out the 2018 data, enter year(2018) in the Selection Filter(s) box.  Your screen should look like the following.



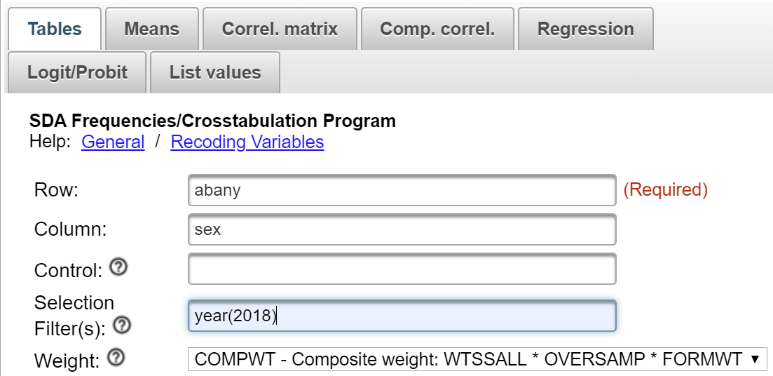
Notice that a weight variable has already been entered in the WEIGHT box.  This will weight the data so the sample better represents the population from which the sample was selected.

The GSS is an example of a social survey.  The investigators selected a sample from the population of all adults in the United States.  This particular survey was conducted in 2018 and is a relatively large sample of a little more than 2,300 adults.  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.  Often we want to describe respondents in terms of social characteristics such as marital status, education, and age.  These are all variables in the GSS.

In the previous exercise (Exercise 7) 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 1).

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.  We could be trying to explain how people feel about abortion.  The independent variable is some variable that you think might help you explain why some people think abortion should be legal and others think it shouldn’t be legal.  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.  We’ll follow that convention in this exercise.

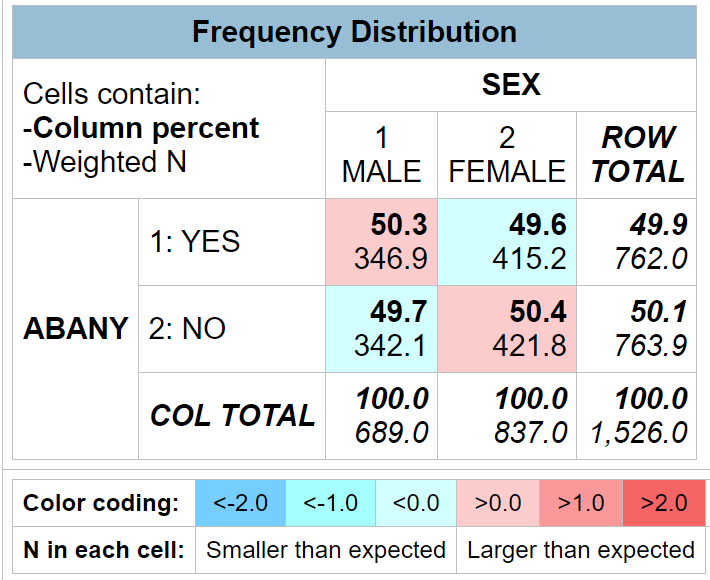
Run CROSSTABS in SDA to produce the crosstabulation of abany and sex.  Click on OUTPUT OPTIONS and look at PERCENTAGING.  Since your independent variable is in the column, you want to use the column percents.  By default, the box for column percents is already checked. Your screen should look like the following.



Notice that the SELECTION FILTER(S) box and the WEIGHT box are both filled in.  Click on RUN THE TABLE to produce the crosstabulation.

## 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.  The table shows that 50.3% of men think abortion should be legal for any reason compared to 49.6% of women.  This is a difference of 0.7 percentage points which seems rather 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 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?  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.[[10]](#footnote-10)   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.

SDA will compute Chi Square for you.  Follow the same procedure you used to get the crosstabulation between sex and abany.  Remember to get the column percents.  Click on OUTPUT OPTIONS and then check the box for SUMMARY STATISTICS.  Be sure to select SRS in the SAMPLE DESIGN line.  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. SDA will give you the value of Chi Square (.08 in this example) and the significance value in parentheses. 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.

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 689 males and 837 females in this table and there are 762 that think abortion should be legal for any reason and 764 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.

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=0.78.  That is the probability that you would be wrong if you rejected the null hypothesis.  In other words, we would be wrong almost 8 out of 10 times.  With odds like that 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 .08 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 (satfin)
* opinion about gun control (gunlaw)
* religiosity (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 SDA gives you) and the cell frequencies that you would expect by chance assuming the null hypothesis was true.  Your instructor may want to show you how to calculate the expected values by hand.  We’re not going to go into it in this exercise.

Chi Square assumes that all the expected cell frequencies are greater than five.  For the crosstabulation of abany and sex this is not a problem.  All the expected frequencies are greater than five.  If they are just a little bit below five, that’s no problem.  But if they get down to around three you have a problem.  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.

Let’s look at an example where the expected frequencies are considerably less than five.  Run the crosstabulation of sex and sibs*[[11]](#footnote-11)* which is the number of brothers and sisters that respondents have. Some of the expected frequencies are so small that they are just about 0.  That’s because there are aren't many respondents with more than 7 siblings.  You will need to recode the number of siblings into fewer categories to increase the size of the expected frequencies.

We’ll recode sibs by combining all values from 8 through 25 into a single category and leaving the other categories the same.  We can do this be entering the following in the ROW box – sibs(r:0=0;1=1;2=2;3=3;4=4;5=5;6=6;7=7;8=8-25 "8 or more") (Note: there is no period at the end of the command.)

Note the syntax for the recode statement since you will want to use it later.

* First you enter r: followed by the variable name which is sibs.  The r stands for recode.
* Then you indicate the new value you want to assign to the recoded category.
* Then you put the old values that you are combining (or leaving as is).  If there is a range of values that you are combining, then these values must be separated by a dash (i.e., hyphen).
* This is followed by the label you want to assign to the new category enclosed in double quotation marks which is “8 or more” for the category 8.  This is optional and is free form meaning you can put what you want for the label.
* Recoded categories are separated by a semi-colon.
* Finally, the entire recode specification is in parentheses.
* So your recode statement looks like this: sibs(r:0=0;1=1;2=2;3=3;4=4;5=5;6=6;7=7;8=8-25 "8 or more")

Enter this into the ROW box and then click on RUN THE TABLE to rerun the crosstab using the recoded sibs.  Now your expected frequencies will be large enough to meet the assumption that all expected values are greater than 5.  Compare the value of Chi Square for the unrecoded sibs with the recoded sibs.  Are they the same?  Why or why not?

# Exercises for an Introductory Statistics Course Using SDA Edward Nelson, California State University, Fresno

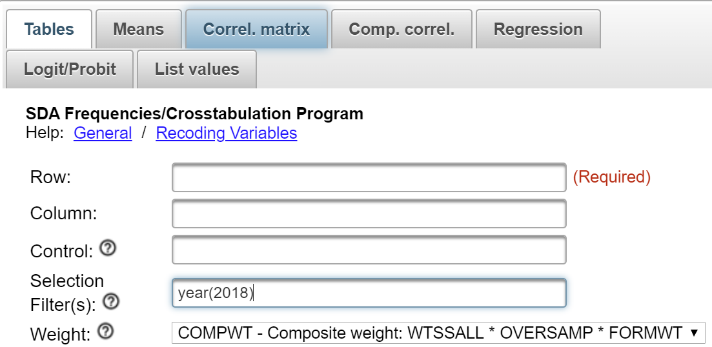
# Exercise 9 Measures of Association

## Goals of Exercise

The goal of this exercise is to introduce measures of association.  The exercise also gives you practice in using CROSSTABS in SDA.

## 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 the 2018 GSS.  To access the GSS cumulative data file in SDA format click [here](https://sda.berkeley.edu/sdaweb/analysis/?dataset=gss18).  The cumulative data file contains all the data from each GSS survey conducted from 1972 through 2018.  We want to use only the data that was collected in 2018.  To select out the 2018 data, enter year(2018) in the Selection Filter(s) box.  Your screen should look like the following.



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.

There’s one other thing that it’s important to do.  Click on the arrow next to OUTPUT OPTIONS and look at the line that says SAMPLE DESIGN.  On your screen COMPLEX will be selected.  Click on the circle next to SRS to select it.

The GSS is an example of a social survey.  The investigators selected a sample from the population of all adults in the United States.  This particular survey was conducted in 2018 and is a relatively large sample of a little more than 2,300 adults.  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.  Often we want to describe respondents in terms of social characteristics such as marital status, education, and age.  These are all variables in the GSS.

In Exercise 7 we used crosstabulation and percents to describe the relationship between pairs of variables in the sample.  In Exercise 8 we went beyond simple description.  We used the sample data to make inferences about the population from which the sample was selected.  Chi Square was used to test hypotheses about the population.  Chi Square is the appropriate test when your variables are nominal or ordinal (see Exercise 1).

Chi Square is a test of the null hypothesis that two variables are unrelated to each other.  Another way to put this is that the two variables are independent of each other.  If we can reject the null hypothesis then we have support for our research hypothesis that the two variables are related to each other.  But showing that two variables are related is not the same thing as determining the strength of the relationship.  The strength of a relationship is actually a continuum from very weak to very strong.  To measure the strength of a relationship we need to select and compute a measure of association.  In this exercise we’re going to focus on nominal and ordinal variables.  In Exercises 11 and 12 we’ll talk about measures for interval and ratio variables.

## Part II – What is a Measure of Association?

Before we discuss measures of association, we need to talk about independent and dependent variables.  The dependent variable is whatever you are trying to explain.  For example, let’s say we want to find out why some people think that abortion should be legal and others think it should be illegal.  The independent variable is some variable that you think might help you answer this question.  Perhaps we decide to use sex as our independent variable.

A measure of association is a numerical value that tells us how strongly related two variables are.  There are several characteristics of a good measure of association.

* They range from a value of 0 (i.e., no relationship) to 1 (i.e., the strongest possible relationship).
* For variables that have an underlying order from low to high they can be positive or negative.  A positive value indicates that as one variable increases, the other variable also increases.  A negative value indicates that as one variable increases, the other variable decreases.[[12]](#footnote-12)
* Some measures specify which variable is dependent and which is independent.  The independent variable is some variable that you think might help explain the variation in the dependent variable.  For example, if your two variables were education and voting you might choose education as the independent variable and voting as your dependent variable because you think that education will help you explain why some people vote Democrat and others vote Republican. Measures of association that specify which variable is dependent and which is independent are called asymmetric measures and measures that don’t specify which is dependent and which is independent are called symmetric measures.

## Part III – Choosing a Measure of Association

There are many measures of association to choose from. We’re going to limit our discussion to those measures that SDA will compute plus a couple others. When choosing a measure of association we’ll start by considering the level of measurement of the two variables (see Exercise 1).

* If one or both of the variables is nominal, then choose one of these measures.
  + Contingency Coefficient – SDA doesn’t compute this but it’s easy to compute by hand.
  + Cramer’s V – SDA doesn’t compute this either but it’s also easy to compute by hand and we’ll show you how.
* If both of the variables are ordinal, then choose from this list.
  + Gamma
  + Somer’s d with the row variable as the dependent variable
  + Kendall’s tau-b
  + Kendall’s tau-c
* Dichotomies should be treated as ordinal. Most variables can be recoded into dichotomies. For example, marital status can be recoded into married or not married. Race can be recoded as white or non-white. All dichotomies should be considered ordinal.

## Part IV – Measures of Association for Nominal Variables

There are a number of nominal level variables in the 2018 GSS.  Here are a few examples.

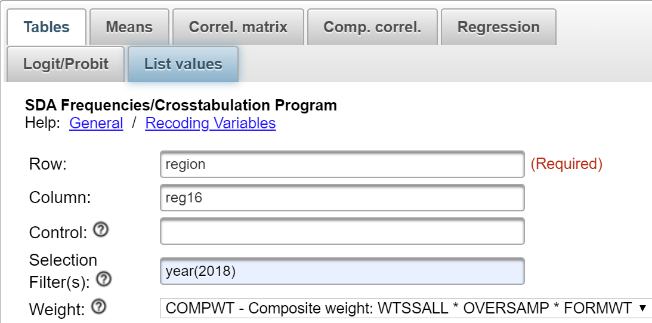
* race of respondent – *race*
* race of household – *hhrace*
* region in which respondent lives – *region*
* region in which respondent lived at age 16 – *reg16*

When one or both of your variables are nominal, you have a choice of the following measures[[13]](#footnote-13) – Contingency Coefficient and Cramer’s V.   Let’s start with the Contingency Coefficient (C).  One of the problems with this measure is that it varies from 0 to some value less than 1.  The larger the number of categories, the closer the maximum value is to 1.  For a table with two rows and two columns, the maximum value is .707 but for a table with three rows and three columns the maximum value is .816.  So you can’t use C to compare the strength of the relationship unless the tables have the same number of rows and columns.

Cramer’s V is an extremely useful measure because it can vary between 0 and 1 regardless of the number of rows and columns.  Values of V can therefore be compared for tables with different number or rows and columns.[[14]](#footnote-14)

Let’s look at an example to help us better understand measures of association for nominal variables.  We’re going to use two variables – region and reg16.  The first variable – region – is the region of the country in which the respondent currently lives and the second – reg16 – is where the respondent lived at the age of 16.  It would make sense to think of region as the dependent variable since where respondents lived at age 16 might influence where they currently live.  Always put the dependent variable in the row and the independent variable in the column.[[15]](#footnote-15)

Run CROSSTABS in SDA to produce the crosstabulation of region and reg16.  Click on OUTPUT OPTIONS and look at PERCENTAGING.  Since your independent variable is always in the column, you want to use the column percents.  By default, the box for column percents is already checked. Also, click on OUTPUT OPTIONS and check the box for SUMMARY STATISTICS.  Your screen should look like Figure 11-2. Notice that the SELECTION FILTER(S) box and the WEIGHT box are both filled in.  Make sure that you selected SRS in the SAMPLE DESIGN line.  Click on RUN THE TABLE to produce the crosstabulation.



Calculating C and V is easy. All you have to do is follow these simple steps.

* C equals the square root of the following: Chi Square divided by the sum of the number of cases in the table and Chi Square.
  + Chi Square is the Pearson Chi Square.  SDA expresses this as Chisq-P.  (See Exercise 8)
  + Look at the SUMMARY STATISTICS that SDA gives you.  The Pearson Chi Square is 9,384.91 and the number of cases in the table is 2,348.
  + So divide 9,384.91 by the sum of 2,348 and 9,348.91.  This equals 9,384.91 divided by 11,696.91 or 0.8023.
  + Now take the square root of .8023 which equals 0.896.
* V equals the square root of the following: Chi Square divided by the product of the number of cases in the table and the smaller of two values – the number of rows minus 1 and the number of columns minus 1.
  + The Pearson Chi Square is 9,384.91, the number of cases in the table is 2,348, the number of rows minus 1 is 9-1 or 8, the number of columns minus 1 is 10 – 1 or 9.
  + The smaller of the number of rows minus 1 and the number of columns minus 1 is 8 since 9 -1 is smaller than 10 – 1.
  + So divide 9,384.91 by the product of 2,348 and 8. This equals 9,384.91 divided by 18,784 or .4996.
  + Now take the square root of .5057 which equals 0.707.

Notice that C and V are quite high.  C is 0.896 and V is 0.707.  You can see that C tells us that there is a very strong relationship between these two variables as does V.

## Part V – Now it’s Your Turn

Use CROSSTABS in SDA to give you the table for *race* and *region*.  The variable *race* classifies the respondents as white, black, or other.  We want to find out whether the respondent’s race influences where the respondent currently lives.  Decide which variable is independent and dependent. Remember to put the dependent variable in the row and the independent variable in the column.  Get the correct percents and tell SDA to compute Chi Square. Then compute C and V by hand.  Use all this information to describe the relationship between these two variables.

## Part VI – Measures of Association for Ordinal Variables

There are a number of ordinal level variables in the 2018 GSS.  Here are a few examples.

* respondent’s highest educational degree – degree
* spouse’s highest educational degree – spdeg
* satisfaction with current financial situation – satfin
* happiness with life – happy
* political views – polviews

You have a choice from four measures that SDA will compute for ordinal variables – Gamma, Somer’s d, Kendall’s tau-b, and Kendall’s tau-c.  Let’s start with Somer’s d.   This measure is the only one of the four that is an asymmetric measure.  That means that Somer’s d allows you to specify one of the variables as independent and the other as dependent.  Use CROSSTABS in SDA to get the crosstabulation of *degree* and satfin.  If we think that education influences how satisfied respondents are with their financial situation, then satisfaction with financial situation would be our dependent variable and would go in the row and education would go in the column.  Be sure to get the column percents, Chi Square, and the four measures of association we listed above.

Chi Square tells us that we should reject the null hypothesis that the two variables are unrelated which provides support for our research hypothesis that the variables are related to each other.  Since satfin is our dependent variable the appropriate value of Somer’s d is -.14. Tau-b and tau-c are the same (-.14).  Gamma (-0.21) is larger.  Gamma will always be larger because of the way it is computed.

Now let’s run a table using degree and spdeg.  It doesn’t seem reasonable to treat one spouse’s education as independent and the other spouse’s education as dependent so let’s ignore Somer’s d.[[16]](#footnote-16)   In this example, it doesn’t matter which variable you put in the column and which you put in the row. Tab-b is 0.50 and tau-c is .45.  Gamma as always is larger (0.66).  The relationship between these two variables is clearly stronger than in the previous example.

You probably noticed that these measures for ordinal variables can be both positive and negative.  The problem is that it’s hard to interpret the sign.  We would like to be able to say that a positive value indicates that as one variable increases the other variable increases and a negative value indicates that as one variable increases the other variable decreases.  But that depends on how the values are coded.  So to determine whether a relationship is positive or negative it’s better to look at the percentages and let them tell you if it is positive or negative.

## Part VII – Now it’s Your Turn Again

Use CROSSTABS to give you a table for degree and happy.  We want to find out if the respondent’s education helps us understand why some say they are happier than others.  Decide which variable is independent and dependent.  Get the correct percents and tell SDA to compute Chi Square and the four measures of association we discussed.  Use all this information to describe the relationship between these two variables.

## Part VIII – Using Measures of Association to Compare Tables

The primary use of measures of association is to compare the strength of a relationship in several tables.  You want to make sure that you compare the same measure of association across tables.  For example, compare Gamma values to Gamma values and V values to V values.  Rerun one of the tables that you created in Parts 5 and 7 but this time hold sex constant.  Do this by moving sex to the control box which is right below the COLUMN box in the crosstabs dialog box.  Now compare the appropriate measure of association to determine if the relationship is stronger for males or females or whether it doesn’t vary much by sex.  Remember not to make too much out of small differences in the measures.

# Exercises for an Introductory Statistics Course Using SDA Edward Nelson, California State University, Fresno

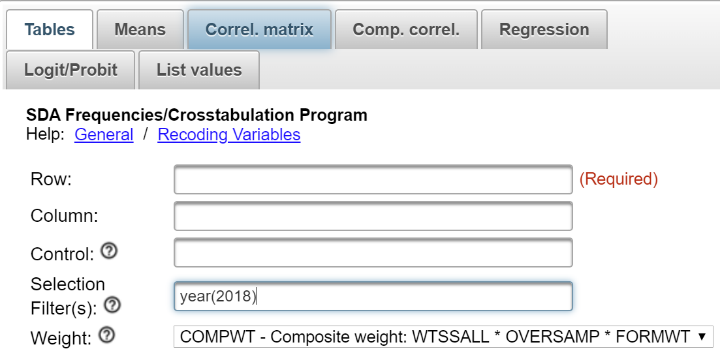
# Exercise 10 Spuriousness

## **Goals of Exercise**

The goal of this exercise is to explore the concept of spuriousness.  We will consider the relationship of religiosity and how respondents feel about controlling the distribution of pornography and test for the possibility that this relationship is spurious due to sex.   The exercise also gives you practice in combining categories of a variable (i.e., recoding) and using CROSSTABS in SDA to explore 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 (NORC).  The GSS started in 1972 and has been an annual or biannual survey ever since. For this exercise we’re going to use the 2018 GSS.  To access the GSS cumulative data file in SDA format click [here](https://sda.berkeley.edu/sdaweb/analysis/?dataset=gss18).  The cumulative data file contains all the data from each GSS survey conducted from 1972 through 2018.  We want to use only the data that was collected in 2018.  To select out the 2018 data, enter year(2018) in the Selection Filter(s) box.  Your screen should look like the following.



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.

There’s one other thing that it’s important to do.  Click on the arrow next to OUTPUT OPTIONS and look at the line that says SAMPLE DESIGN.  On your screen COMPLEX will be selected.  Click on the circle next to SRS to select it.

The GSS is an example of a social survey.  The investigators selected a sample from the population of all adults in the United States.  This particular survey was conducted in 2018 and is a relatively large sample of a little more than 2,300 adults.  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.  Often we want to describe respondents in terms of social characteristics such as marital status, education, and age.  These are all variables in the GSS.

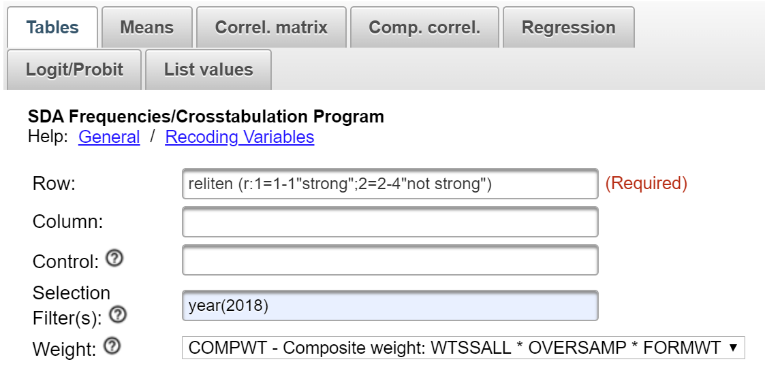
Let’s look at the relationship between the strength of a person’s religious affiliation and how a person feels about controlling the distribution of pornography.  One of the variables in the data set is pornlaw.  This question asks respondents what type of laws they think we ought to have regulating 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 reliten in the data set.

We’re going to recode reliten for this exercise.  Recoding means to combine categories of the variable.  Before we start recoding, run FREQUENCIES in SDA for the variable reliten so you will know what the frequency distribution looks like before you recode.  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.  Follow these steps to recode in SDA

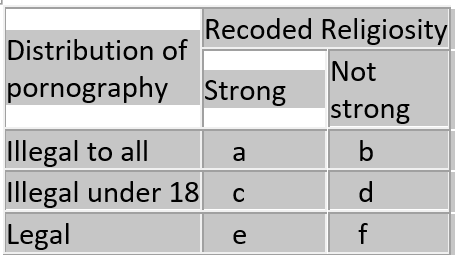
* Enter the variable name in the appropriate row or column or control box.  The variable name in this example is reliten.  (Don’t enter the period.)
* After the variable name, enter (r: where r stands for recode.
* Enter the new value you want to assign to the first recode followed by the recode.  In our case we want to assign the new value 1 to the old value 1 so this would be 1=1-1.  (Don’t enter the period.)
* Enter the label you want to assign to this recode in double quotation marks so that would be "strong" followed by a semi-colon.  So far our recode would look like this – reliten (r:1=1-1"strong";.  (Don’t enter the period.)
* Repeat this process for each recode.  If you want to recode a range of values into a new value, it would look this – 2=2-4.  (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 – reliten (r:1=1-1"strong";2=2-4"not strong").  (Don’t enter the period.)

After you have recoded this variable, run FREQUENCIES in SDA for the recoded variable reliten.  Your screen should look like the following.

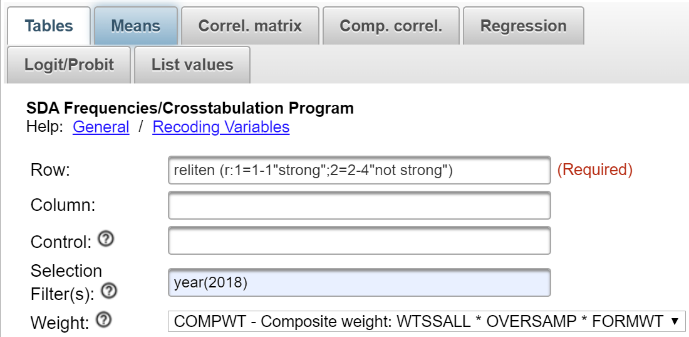


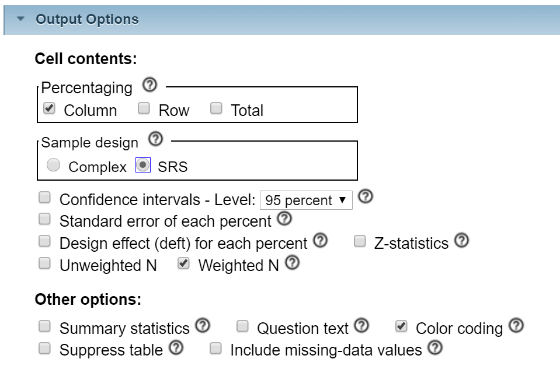
Click RUN THE TABLE to produce the frequency distribution.  Compare the two frequency distributions to make sure you didn’t make an error recoding.  If you did make a mistake, you’ll need to do the recoding again.  
  
Now that we’ve taken care of recoding reliten, let’s start by developing a hypothesis.  A hypothesis states the relationship that you expect to find between your two variables.  In this case, our hypothesis could be that 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 the respondent’s opinion about pornography laws? Why should more religious individuals be more likely to think that pornography should be illegal for everyone?  Write a clear argument explaining why you think your hypothesis is true.

Once you have developed your argument, then you should construct a dummy table showing what the relationship between the recoded variable reliten and pornlaw should look like if your hypothesis is true.  Use “Tables” in Word to construct the table below.  We’re going to always 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.  Use Tables in Word to create the table below and add the arrows indicating the relationship that you expect to find.



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 SDA.  Be sure to put the recoded variable reliten in the column and the dependent variable pornlaw in the row.  Click on OUTPUT OPTIONS and make sure that you have checked the box for column presents and that you have selected SRS for SAMPLE DESIGN. You also need to check the box for SUMMARY STATISTICS under OTHER OPTIONS. Your screen ought to look like the following.



  
  
Click RUN THE TABLE to produce your table.

All that is left is to interpret the table.  Since the independent variable is the column variable, we had SDA 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 whichever measure of association you think is appropriate 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 ever 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 in SDA to get two tables – one table should cross tabulate sex and pornlaw and the other table should cross tabulate sex and the recoded variable reliten.  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 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.

We can do that in SDA by getting a crosstab putting pornlaw in the ROW box (our dependent variable), the recoded version of reliten in the COLUMN box (our independent variable),  and sex in the CONTROL box.  In this case, sex 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.

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 markedly 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 it decrease substantially?  Another clue is your measure of association.   Did the measure or association for males and females stay about the same or did it 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—Summarizing Your Results**

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?

# Exercises for an Introductory Statistics Course Using SDA Edward Nelson, California State University, Fresno

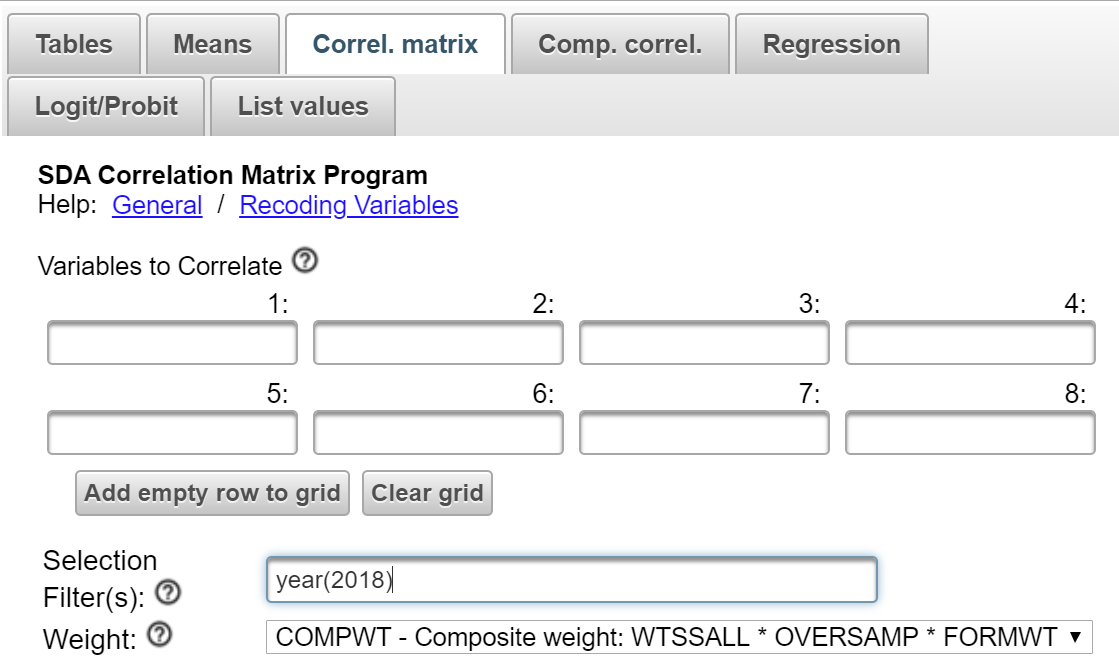
# Exercise 11 Correlation

## Goals of Exercise

The goal of this exercise is to introduce measures of correlation. The exercise also gives you practice using CORRELATION MATRIX in SDA.

## Part I – Getting Started

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 the 2018 GSS.  To access the GSS cumulative data file in SDA format click [here](https://sda.berkeley.edu/sdaweb/analysis/?dataset=gss18).  The cumulative data file contains all the data from each GSS survey conducted from 1972 through 2018.  Click on the CORRELATION MATRIX tab at the top of your screen. Since we want to use only the data that was collected in 2018, enter year(2018) in the Selection Filter(s) box.  Your screen should look like the following.



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.

The GSS is an example of a social survey.  The investigators selected a sample from the population of all adults in the United States.  This particular survey was conducted in 2018 and is a relatively large sample of a little more than 2,300 adults.  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.  Often we want to describe respondents in terms of social characteristics such as marital status, education, and age.  These are all variables in the GSS.

In Exercise 9 we considered different measures of association that can be used to determine the strength of the relationship between two variables that have nominal or ordinal level measurement (see Exercise 1). In this exercise we’re going to look at two different measures that are appropriate for interval and ratio level variables. The terminology also changes in the sense that we’ll refer to these measures as correlations rather than measures of association.

## Part II - Pearson Correlation Coefficient

The Pearson Correlation Coefficient (r) is a numerical value that tells us how strongly related two variables are. It varies between -1 and +1. The sign indicates the direction of the relationship. A positive value means that as one variable increases, the other variable also increases while a negative value means that as one variable increases, the other variable decreases. The closer the value is to 1, the stronger the linear relationship and the closer it is to 0, the weaker the linear relationship.

The usual way to interpret the Pearson Coefficient is to square its value. In other words, if r equals .5, then we square .5 which gives us .25. This is often called the Coefficient of Determination. This means that one of the variables explains 25% of the variation of the other variable. Since the Pearson Correlation is a symmetric measure in the sense that neither variable is designated as independent or dependent we could say that 25% of the variation in the first variable is explained by the second variable or reverse this and say that 25% of the variation in the second variable is explained by the first variable. It’s important not to read causality into this statement. We’re not saying that one variable causes the other variable. We’re just saying that 25% of the variation in one of the variables can be accounted for by the other variable.

The Pearson Correlation Coefficient assumes that the relationship between the two variables is linear. This means that the relationship can be represented by a straight line. In geometric terms, this means that the slope of the line is the same for every point on that line. Here are some examples of a positive and a negative linear relationship and an example of the lack of any relationship.

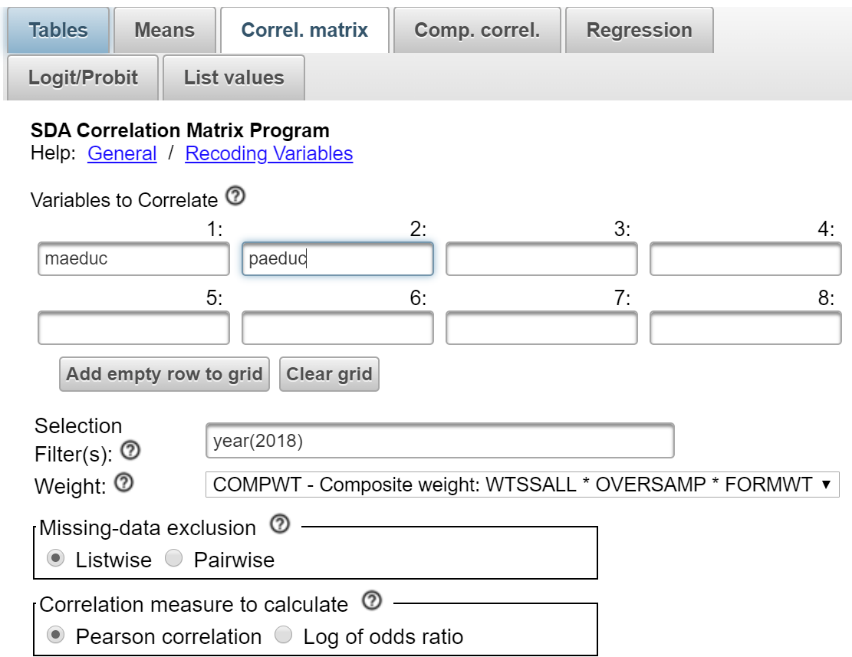
Pearson r would be positive and close to 1 in the left-hand example, negative and close to -1 in the middle example, and closer to 0 in the right-hand example. You can search for “free images of a positive linear relationship” to see more examples of linear relationships.

But what if the relationship is not linear? Search for “free images of a curvilinear relationship” and you’ll see examples that look like this.

Here the relationship can’t be represented by a straight line. We would need a line with a bend in it to capture this relationship. While there clearly is a relationship between these two variables, Pearson r would be closer to 0. Pearson r does not measure the strength of a curvilinear relationship; it only measures the strength of linear relationships.

Another way to think of correlation is to say that the Pearson Correlation Coefficient measures the fit of the line to the data points. If r was equal to +1, then all the data points would fit on the line that has a positive slope (i.e., starts in the lower left and ends in the upper right). If r was equal to -1, then all the data points would fit on the line that has a negative slope (i.e., starts in the upper left and ends in the lower right).

Click on CORRELATION MATRIX at the top of your screen. We’re going to get the Pearson Coefficient for two variables that measure the amount of education of the respondent’s father and mother.  These variables are named maeduc and paeduc in the GSS.  Enter the two variables in the CORRELATION MATRIX dialog box.  It doesn’t matter which you enter first.  Notice that the SELECTION FILTER(S) and the WEIGHT boxes are filled in.  Also notice that SDA has checked the box for the Pearson Correlation which is what we want.  Listwise has been selected in the MISSING-DATA EXCLUSION box.  That means that any case with missing data for either of these two variables will be excluded from analysis.  Your screen should look like the following.



Now click on RUN CORRELATIONS to produce the Pearson Correlations. You should see four correlations. The correlations in the upper left and lower right will be 1 since the correlation of any variable with itself will always be 1. The correlation in the upper right and lower left will both be 0.71. That’s because the correlation of variable X with variable Y is the same as the correlation of variable Y with variable X. Pearson r is a symmetric measure (see Exercise 9) meaning that we don’t designate one of the variables as the dependent variable and the other as the independent variable. You don’t see r’s that big very often. That’s telling us that the linear regression line that we’re going to talk about in Exercise 13 fits the data points reasonably well.

## Part III – Now it’s Your Turn Again

Use CORRELATION MATRIX in SDA to get the Pearson Correlation Coefficient for the years of school completed by the respondent (educ) and the spouse’s years of school completed (speduc). What does this Pearson Correlation Coefficient tell you about the relationship between these two variables?

## Part IV – Correlation Matrices

What if you wanted to see the values of r for a set of variables? Let’s think of the four variables in Parts II and III as a set. That means that we want to see the values of r for each pair of variables. This time move all four of the variables into the VARIABLES TO CORRELATE box (i.e., educ, maeduc, paeduc, and speduc) and click on RUN CORRELATIONS.   That would mean we would calculate six coefficients. (Make sure you can list all six.)

What did we learn from these correlations? First, the correlation of any variable with itself is 1. Second, the correlations above the 1’s are the same as the correlations below the 1’s. They’re just the mirror image of each other. That’s because r is a symmetric measure. Third, all the correlations are fairly large. Fourth, the largest correlations are between father’s and mother’s education and between the respondent’s education and the spouse’s education.

## Part V – Eta-Squared

The Pearson Correlation Coefficient assumes that both variables are interval or ratio variables (see Exercise 1). But what if one of the variables was nominal or ordinal and the other variable was interval or ratio? This leads us back to one-way analysis of variance which we discussed in Exercise 6.

Click on MEANS in the menu bar at the top of SDA and enter the variable tvhours in the DEPENDENT box and degree in the ROW box.  You’ll need to enter *year(2018)* in the SELECTION FILTER(S) box. The WEIGHT box should still be filled in.  In Exercise 6 we wanted to determine if the differences between levels of education are statistically significant so we carried out a one-way analysis of variance.  Make sure that the MEAN STATISTIC TO DISPLAY box says MEANS.  This is the default so it should.  Then click on OUTPUT OPTIONS and check the box for ANOVA STATS under OTHER OPTIONS.  Be sure to uncheck the box for COMPLEX STANDARD ERRORS and check the box for SRS STANDARD ERRORS.  Finally, click RUN THE TABLE to carry out the procedure.

The F test in the one-way analysis of variance tells us to reject the null hypothesis that all the population means are all equal. So we know that at least one pair of population means are not equal. But that doesn’t tell us how strongly related these two variables are. The SDA output tells us that eta-squared is equal to .046. You should be able to locate this in the Analysis of Variance output table.  This tells us that 4.6% of the variation in the dependent variable, number of hours the respondent watches television, can be explained or accounted for by the independent variable, highest education degree. This doesn’t seem like much but it’s not an atypical outcome for many research findings.

## Part VI – Your Turn

In Exercise 6 you computed the mean number of hours that respondents watched television (tvhours) for each of the nine regions of the country (region). Then you determined if these differences were statistically significant by carrying out a one-way analysis of variance. Repeat the one-way analysis of variance but this time focus on eta-squared. What percent of the variation in television viewing can be explained by the region of the country in which the respondent lived?

# Exercises for an Introductory Statistics Course Using SDA Edward Nelson, California State University, Fresno

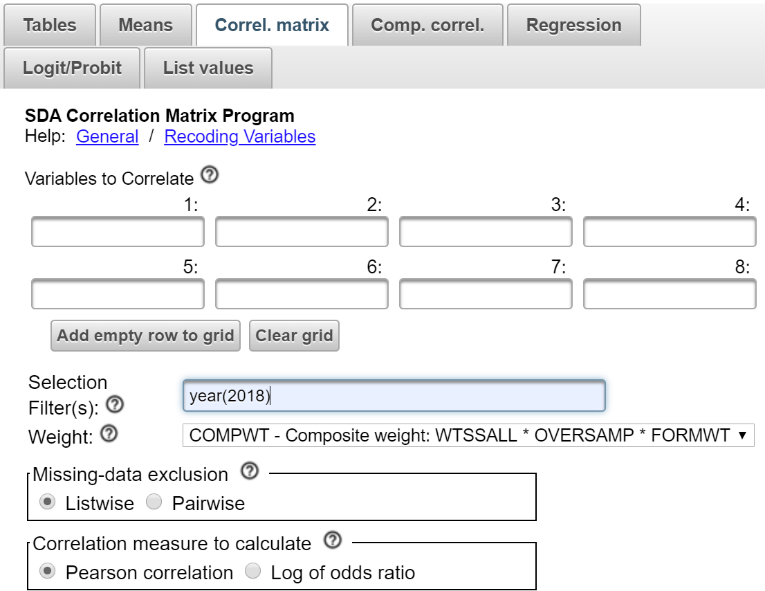
# Exercise 12 Comparison of Correlations

## Goals of Exercise

The goal of this exercise is to explore how to compare correlations. The exercise also gives you practice using COMPARISON OF CORRELATIONS in SDA.

## Part I – Getting Started

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 the 2018 GSS.  To access the GSS cumulative data file in SDA format click [here](https://sda.berkeley.edu/sdaweb/analysis/?dataset=gss18).  The cumulative data file contains all the data from each GSS survey conducted from 1972 through 2018.  Click on the COMPARISON OF CORRELATIONS tab at the top of your screen. Since we want to use only the data that was collected in 2018, enter year(2018) in the Selection Filter(s) box.  Your screen should look like the following.



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.

The GSS is an example of a social survey.  The investigators selected a sample from the population of all adults in the United States.  This particular survey was conducted in 2018 and is a relatively large sample of a little more than 2,300 adults.  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.  Often we want to describe respondents in terms of social characteristics such as marital status, education, and age.  These are all variables in the GSS.

In a previous Exercise 9 we considered different measures of association that can be used to determine the strength of the relationship between two variables that have nominal or ordinal level measurement (see Exercise 1). In this exercise we’re going to look at two different measures that are appropriate for interval and ratio level variables. The terminology also changes in the sense that we’ll refer to these measures as correlations rather than measures of association.

## Part II - Pearson Correlation Coefficient

The Pearson Correlation Coefficient (r) is a numerical value that tells us how strongly related two variables are. It varies between -1 and +1. The sign indicates the direction of the relationship. A positive value means that as one variable increases, the other variable also increases while a negative value means that as one variable increases, the other variable decreases. The closer the value is to 1, the stronger the linear relationship and the closer it is to 0, the weaker the linear relationship.

The usual way to interpret the Pearson Coefficient is to square its value. In other words, if r equals .5, then we square .5 which gives us .25. This is often called the Coefficient of Determination. This means that one of the variables explains 25% of the variation of the other variable. Since the Pearson Correlation is a symmetric measure in the sense that neither variable is designated as independent or dependent we could say that 25% of the variation in the first variable is explained by the second variable or reverse this and say that 25% of the variation in the second variable is explained by the first variable. It’s important not to read causality into this statement. We’re not saying that one variable causes the other variable. We’re just saying that 25% of the variation in one of the variables can be accounted for by the other variable.

The Pearson Correlation Coefficient assumes that the relationship between the two variables is linear. This means that the relationship can be represented by a straight line. In geometric terms, this means that the slope of the line is the same for every point on that line. Here are some examples of a positive and a negative linear relationship and an example of the lack of any relationship.

Pearson r would be positive and close to 1 in the left-hand example, negative and close to -1 in the middle example, and closer to 0 in the right-hand example. You can search for “free images of a positive linear relationship” to see more examples of linear relationships.

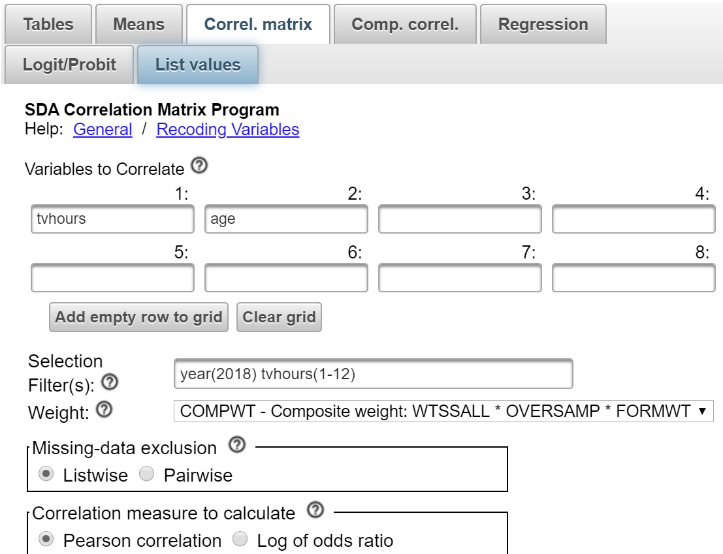
But what if the relationship is not linear? Search for “free images of a curvilinear relationship” and you’ll see examples that look like this.

Here the relationship can’t be represented by a straight line. We would need a line with a bend in it to capture this relationship. While there clearly is a relationship between these two variables, Pearson r would be closer to 0. Pearson r does not measure the strength of a curvilinear relationship; it only measures the strength of linear relationships.

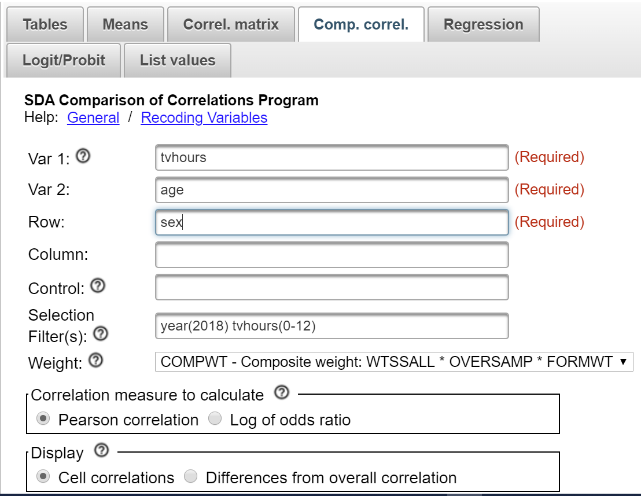
Another way to think of correlation is to say that the Pearson Correlation Coefficient measures the fit of the line to the data points. If r was equal to +1, then all the data points would fit on the line that has a positive slope (i.e., starts in the lower left and ends in the upper right). If r was equal to -1, then all the data points would fit on the line that has a negative slope (i.e., starts in the upper left and ends in the lower right).

Let’s get the Pearson Correlation Coefficient for the variables tvhours (i.e., number of hours that respondents watch television per day) and age.  The variable tvhours has some extreme values or outliers that we need to eliminate from the data file before computing the correlation.  We’re going to define extreme values as any value of 13 or larger.  Let’s exclude these individuals by selecting only those cases for which tvhours is less than 13. That way the extreme values will be excluded from the analysis. To do this add tvhours(0-12) to the SELECTION FILTER(S) box.  Be sure to separate year(2018) and tvhours(0-12) with a space or a comma.  This will tell SDA to select out only those cases for which year is equal to 2018 **and** tvhours is less than 13.  Rerun FREQUENCIES in SDA to get a frequency distribution for tvhours after eliminating the outliers and check to make sure that you did it correctly.

Now we’re ready to get the correlation.  Click on CORRELATION MATRIX at the top of the SDA screen and enter the two variables in the dialog box.  It doesn’t matter which you enter first.  You should have already filled in the SELECTION FILTER(S) box so it reads *year(2018) tvhours(0-12).* The WEIGHT box is already filled in.  Also notice that SDA has checked the box for the Pearson Correlation which is what we want.  Listwise has been selected in the MISSING-DATA EXCLUSION box.  That means that any case with missing data for either of these two variables will be excluded from analysis.  Your screen should look like the following.

  
  
Now click on RUN CORRELATIONS to produce the Pearson Correlations. You should see four correlations. The correlations in the upper left and lower right will be 1 since the correlation of any variable with itself will always be 1. The correlation in the upper right and lower left will both be 0.19. That’s because the correlation of variable X with variable Y is the same as the correlation of variable Y with variable X. Pearson r is a symmetric measure (see Exercise 9) meaning that we don’t designate one of the variables as the dependent variable and the other as the independent variable. The correlation of .19 indicates that you have a weak to moderate correlation in the positive direction.  In other words, the older the respondent is, the more television they watch.

Now click on COMP CORREL at the top of the SDA screen.  This stands for comparison of correlations.  Let’s say we wanted to compare the correlation of tvhours and age for men and for women. In other words, we want to separate the males and the females and get two correlations – one for males and one for females.  Enter tvhours and age in the VAR 1: and VAR2: boxes.  It doesn’t make any difference which you put in the first and second boxes.  Enter sex in the ROW box. You will have to fill in the SELECTION FILTER(S) box so it reads *year(2018) tvhours(0-12). The* WEIGHT boxes should still be filled in as before. This will tell SDA to compute the Pearson Correlation for males and females separately.  Your screen should look like the following



Now click on RUN THE TABLE to produce the correlations. Notice that the correlations are about the same for males (.24) as for females (.22).  Remember not to make too much of small differences.

Now let’s break the data down more finely taking into account sex and race.  Sex has two categories – male and female – while race has three categories – white, black, other.  If we break our data down by both sex and race we’ll have six categories – white males, black males, white females, black females, other males, and other females.  To do this put sex in the ROW box and race in the COLUMN box and rerun the correlations.  Now click on RUN THE TABLE to produce the correlations.  Now we see more variation in the correlations from .11 for other females to .28 for other males.

## Part III – Now it’s Your Turn

Use CORRELATION MATRIX to get the Pearson correlation between educ (i.e., respondent’s years of school completed) and speduc (i.e., spouse’s years of school completed).  Then use COMPARISON OF CORRELATIONS to get the correlation for males and for females.  There is no need to select those cases for which tvhours is less than 13 since that variable is not part of this analysis.  So delete that from the SELECTION FILTER(S) box.  Write a paragraph describing your findings.

Now do the same thing but this time break the data down by both sex and race.  Write a paragraph describing your findings.

# Exercises for an Introductory Statistics Course Using SDA Edward Nelson, California State University, Fresno

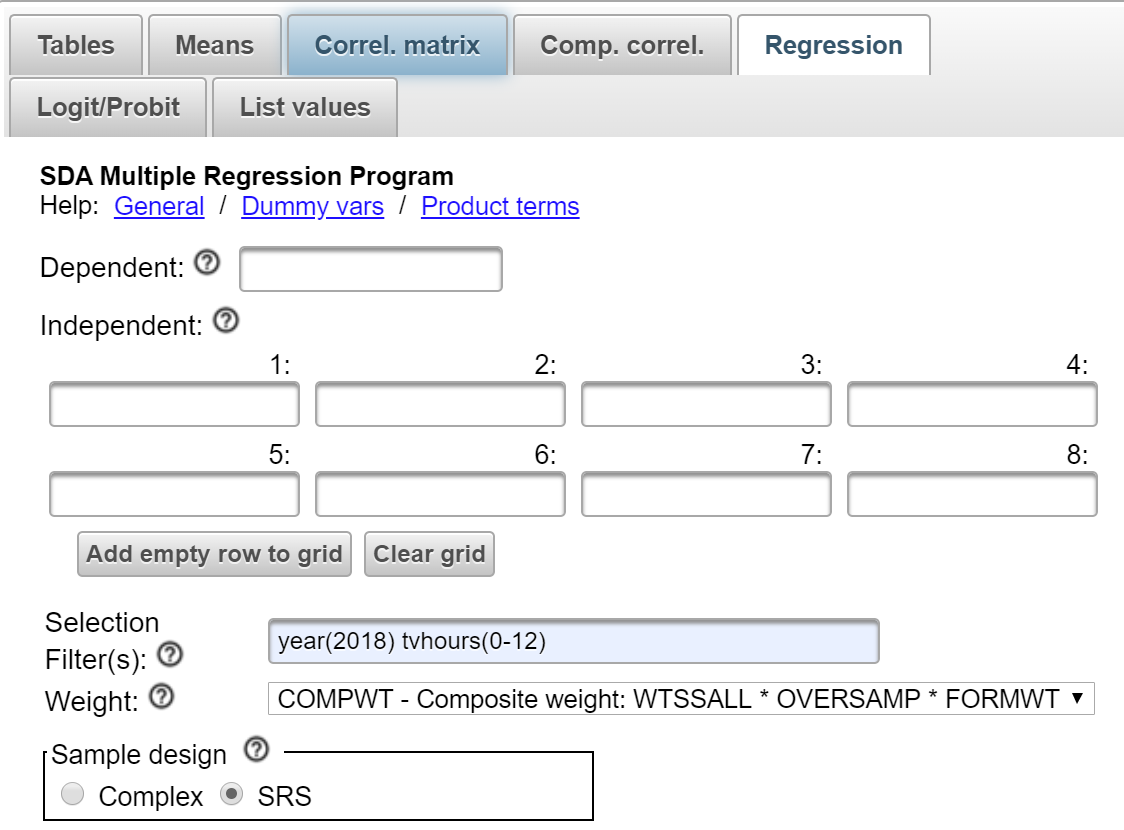
# Exercise 13 Bivariate Linear Regression

## Goals of Exercise

The goal of this exercise is to introduce bivariate linear regression. The exercise also gives you practice using REGRESSION in SDA.

## Part I – Finding the Best Fitting Line to a Scatterplot

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 the 2018 GSS.  To access the GSS cumulative data file in SDA format click [here](https://sda.berkeley.edu/sdaweb/analysis/?dataset=gss18).  The cumulative data file contains all the data from each GSS survey conducted from 1972 through 2018.  We want to use only the data that was collected in 2018.  Click on REGRESSION at the top of your screen. To select out the 2018 data, enter year(2018) in the Selection Filter(s) box.  Your screen should look like the following.



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.

The GSS is an example of a social survey.  The investigators selected a sample from the population of all adults in the United States.  This particular survey was conducted in 2018 and is a relatively large sample of a little more than 2,300 adults.  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.  Often we want to describe respondents in terms of social characteristics such as marital status, education, and age.  These are all variables in the GSS.

In Exercises 11 and 12 we discussed the Pearson Correlation Coefficient which is a measure of the strength of the linear relationship between two interval or ratio variables. In this exercise we’re going to look at linear regression for two interval or ratio variables. An important assumption is that there is a linear relationship between the two variables.

Before we look at these measures let’s talk about outliers. Use FREQUENCIES in SDA to get a frequency distribution for the variable tvhours which is the number of hours that a respondent watches television per day. Click on OUTPUT OPTIONS and check the box for SUMMARY STATISTICS so you will get the skewness and kurtosis measures.  Click also on CHART OPTIONS and then click on TYPE OF CHART and select BAR CHART.  Now click on RUN THE TABLE to produce the frequency distribution and the summary statistics.  Notice that there are only a few people (i.e., 13) who watch 13 or more hours of television per day. There’s even one person who says he or she watches television 24 hours per day. These are what we call outliers and they can affect the results of our statistical analysis.

Let’s exclude these individuals by selecting only those cases for which tvhours is less than 13. That way the outliners will be excluded from the analysis. To do this add tvhours(0-12) to the SELECTION FILTER(S) box.  Be sure to separate year(2018) and tvhours(0-12) with a space or a comma.  This will tell SDA to select out only those cases for which year is equal to 2018 **and** tvhours is less than 13.  Rerun FREQUENCIES in SDA to get a frequency distribution for tvhours after eliminating the outliers and check to make sure that you did it correctly.

Now let’s compare the frequency distribution before we eliminated the outliers with the distribution after eliminating them. Notice that the skewness and kurtosis values are considerably lower for the distribution after eliminating the outliers than they were before the outliers were dropped. This is because outliers affect our statistical analysis. (See Exercise 3 a discussion of skewness and kurtosis.)

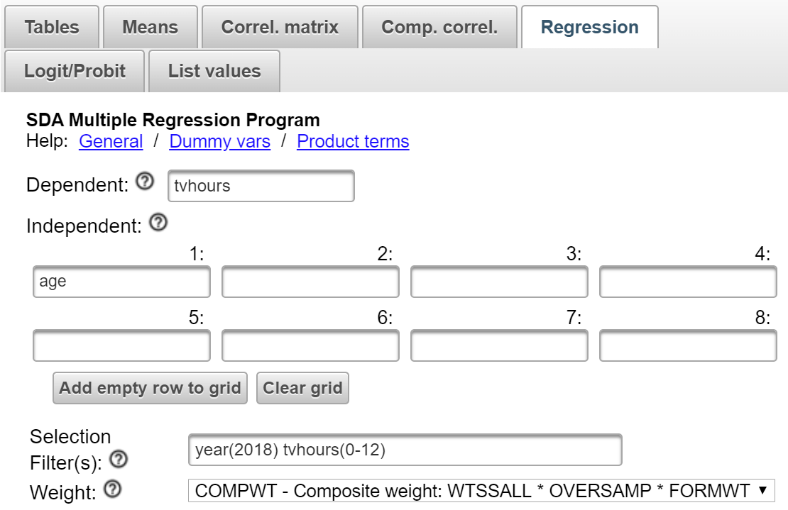
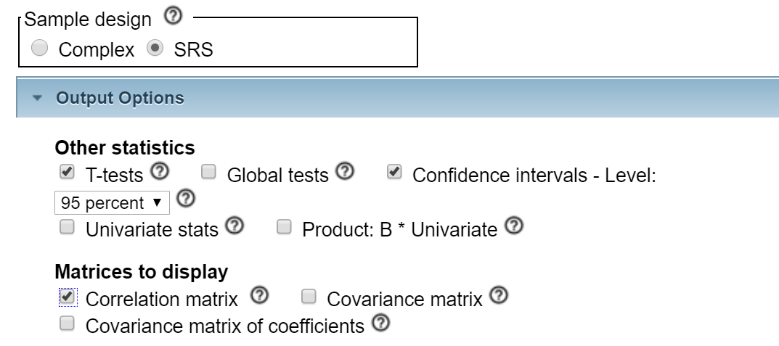
Now we’re ready to find the straight line that best fits the data points. The equation for a straight line is Y = a + bX where a is the point where the line crosses the Y-Axis, b is the slope of the line, and Y is the predicted value of Y. Think of the slope as the average change in Y that occurs when X increases by one unit.[[17]](#footnote-17)

Let’s think about how we’re going to do that. The best fitting line will be the line that minimizes error where error is the difference between the observed values and the predicted values based on our regression equation. So if our regression equation is Y = 10 + 2X, we can compute the predicted value of Y by substituting any value of X into the equation. If X = 5, then the predicted value of Y is 10 + (2)(5) or 20. It turns out that minimizing the sum of the error terms doesn’t work since positive error will cancel out negative error so we minimize the sum of the squared error terms.[[18]](#footnote-18)

## Part II – Getting the Regression Coefficients

The regression equation will be the values of a and b that minimize the sum of the squared errors. There are formulas for computing a and b but usually we leave it to SDA to carry out the calculations.

Click on REGRESSION at the top of the SDA page and enter your dependent variable (tvhours) in the DEPENDENT box.  We’re going to use age as our independent variable so enter age into the INDEPENDENT box.  You’ll need to enter *year(2018) tvhours(0-12)* in the SELECTION FILTER(S) box. The WEIGHT box should automatically be filled in. In the SAMPLE DESIGN line select SRS.  Under OTHER STATISTICS uncheck the box for GLOBAL TESTS.  We’ll use this in the next exercise.  Under MATRICES TO DISPLAY, check the box for CORRELATION MATRIX.  Your screen should look like the following.

Now click RUN REGRESSION to produce the regression analysis. The first three boxes in your output are what you want to look at.

* The first box lists the variables you entered as your dependent, independent, weight, and filter variables.
* The second box gives you the regression coefficients.
  + The slope of the line (B) is equal to 0.029.
  + The point at which the regression line crosses the Y-Axis is 1.249. This is often referred to as the constant since it always stays the same regardless of which value of X you are using to predict Y.
  + The standard error of these coefficients which is an estimate of the amount of sampling error.
  + The standardized regression coefficient (often referred to as Beta). We’ll have more to stay about this in the next exercise but with one independent variable Beta always equals the Pearson Correlation Coefficient (r).
  + The t test which tests the null hypotheses that the population constant and population slope are equal to 0.
  + The significance value for each test. As you can see in this example, we reject both null hypotheses. However, we’re usually only interested in the t test for the slope.
  + The value of the Multiple R and Multiple R-Squared.  With only one independent variable this is the same as the Pearson Correlation and the Pearson Correlation squared. Age explains or accounts for 5.2% of the variation in tvhours. The Adjusted R Square “takes into account the number of independent variables relative to the number of observations.” (George W. Bohrnstedt and David Knoke, *Statistics for Social Data Analysis*, 1994, F.E. Peacock, p. 293)  The standard error is an estimate of the amount of sampling error in this statistic. By the way, notice the output refers to R square. In our example with only one independent variable this is the same as r. But in Exercise 14 we’ll talk about multivariate linear regression where we have two or more independent variables and we’ll explain why this is called R square and not r square.
* The third box shows the Pearson Correlation for our two variables. Note that it is the same as the Beta value when you have only one independent variable.

The slope is what really interests us. The slope or B tells us that for each increase of one unit in the independent variable (i.e., one year of age) the value of Y increases by an average of .029 units (i.e., number of hours watching television). So our regression equation is Y = 1.249 + .029X. Thus for a person that is 20 years old, the predicted number of hours that he or she watches television is 1.249 + (.029) (20) or 1.249 + 0.58 or 1.829 hours.

It’s really important to keep in mind that everything we have done assumes that there is a linear relationship between the two variables. If the relationship isn’t approximately linear, then this is all meaningless.

## Part III – It’s Your Turn Now

Use the same dependent variable, tvhours, but this time use educ as your independent variable. This refers to the years of school completed by the respondent. Tell SDA to give you the regression coefficients as we did in Part II.

* Write out the regression equation.
* What do the constant and the slope tell you?
* What are the values of r and r2 and what do they tell you?
* What are the different tests of significance that you can carry out and what do they tell you?

# Exercises for an Introductory Statistics Course Using SDA Edward Nelson, California State University, Fresno

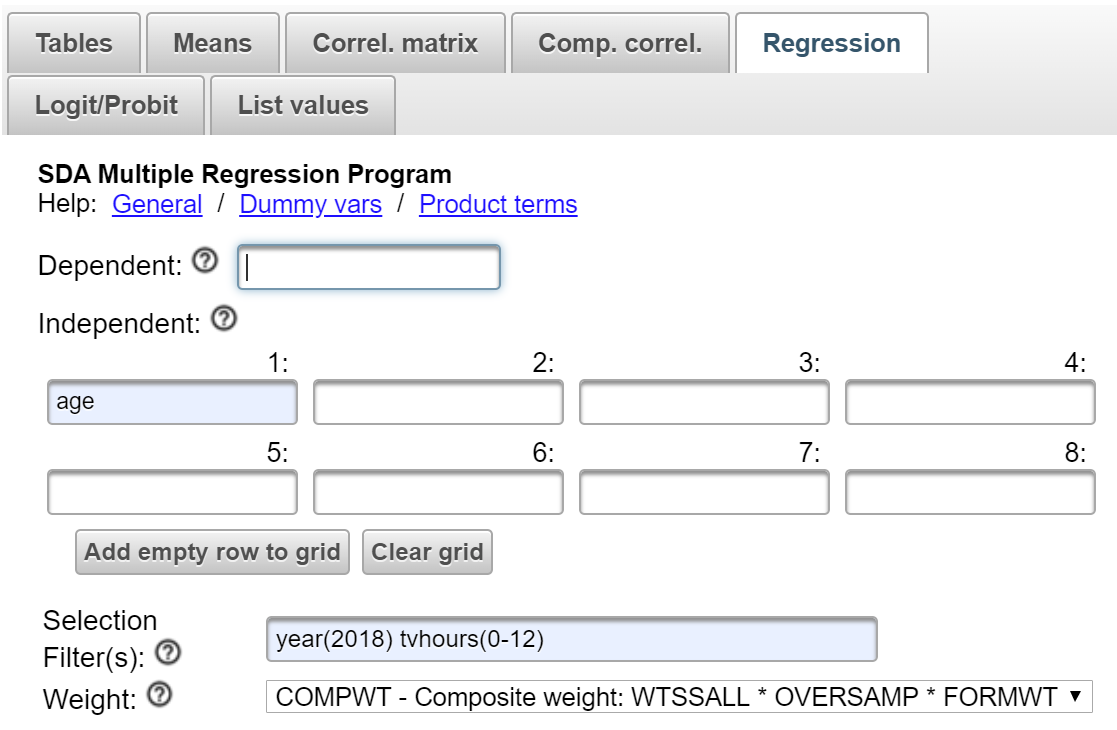
# Exercise 14 Multivariate Linear Regression

## **Goals of Exercise**

The goal of this exercise is to introduce multivariate linear regression.  The exercise also gives you practice using REGRESSION in SDA.

## **Part I – Linear Regression with Multiple Independent 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 the 2018 GSS.  To access the GSS cumulative data file in SDA format click [here](https://sda.berkeley.edu/sdaweb/analysis/?dataset=gss18).  The cumulative data file contains all the data from each GSS survey conducted from 1972 through 2018.  We want to use only the data that was collected in 2018.  Click on REGRESSION at the top of your screen. To select out the 2018 data, enter year(2018) in the SELECTION FILTER(S) box.  Your screen should look like the following.



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.  You will also need to click on OUTPUT OPTIONS and select SRS in the SAMPLE DESIGN box.

The GSS is an example of a social survey.  The investigators selected a sample from the population of all adults in the United States.  This particular survey was conducted in 2018 and is a relatively large sample of a little more than 2,300 adults.  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.  Often we want to describe respondents in terms of social characteristics such as marital status, education, and age.  These are all variables in the GSS.

In Exercise 13 we discussed linear regression for one independent and one dependent variable which is often referred to as bivariate linear regression.  Multivariate linear regression expands the analysis to include multiple independent variables.  In the first part of this exercise we’re going to focus on two independent variables.  Then we’re going to add a third independent variable into the analysis.  An important assumption is that “the dependent variable is seen as a linear function of more than one independent variable.”  (Colin Lewis-Beck and Michael Lewis-Beck, Applied Regression – An Introduction, Sage Publications, 2015, p. 55)

In Exercise 13 we used tvhours as our dependent variable which refers to the number of hours that the respondent watches television per day.  In other words, we wanted to understand why some people watch more television than others.  We found that age was positively related to television viewing and respondent’s education was negatively related.  Older respondents tended to watch more television and respondents with more education tended to watch less television.

Let’s start by using FREQUENCIES in SDA to get the frequency distribution for tvhours.  In Exercise 13 we discussed outliers and noted that there are a few individuals (i.e., 13) who watch a lot of television which we defined as 13 or more hours per day.  We also noted that outliers can affect our statistical analysis so we decided to remove these outliers from our analysis.

Let’s exclude these individuals by selecting out only those cases for which tvhours is less than 13. That way the outliners will be excluded from the analysis. To do this add tvhours(0-12) to the SELECTION FILTER(S) box.  Be sure to separate year(2018) and tvhours(0-12) with a space or a comma.  This will tell SDA to select out only those cases for which year is equal to 2018 **and** tvhours is less than 13.  Run FREQUENCIES in SDA to get a frequency distribution for tvhours after eliminating the outliers and check to make sure that there are no values greater than 12.

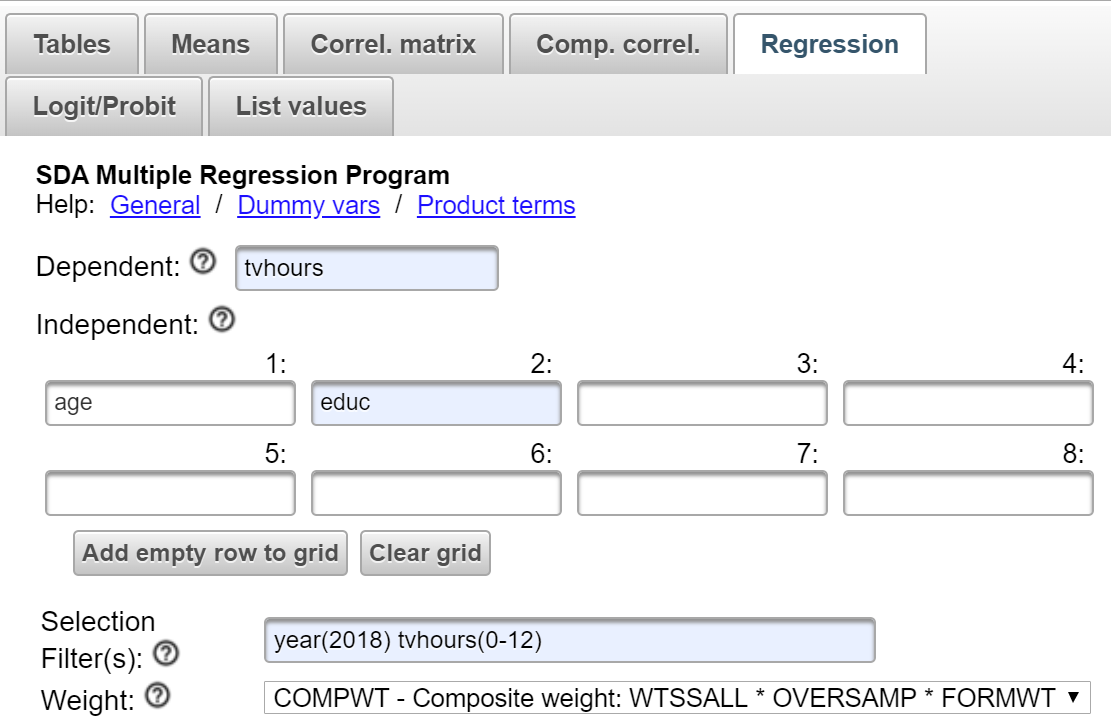
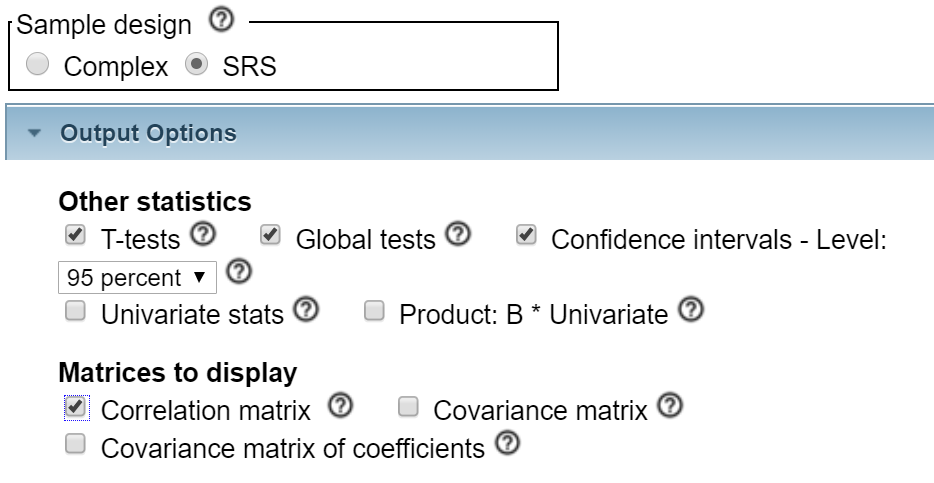
In bivariate linear regression we have one independent and one dependent variable.  So we are trying to find the straight line that best fits the data points in a two-dimensional space.  With two independent and one dependent variable we have a three-dimensional space.  So now we’re trying to find the plane that best fits the data points in this three-dimensional space.

With two independent variables our regression equation for predicting Y is a + B1X1 + B2X2 where a is the constant, B1 and B2 are the unstandardized multiple regression coefficients, and X1 and X2 are the independent variables.  As with bivariate linear regression we want to minimize error where error is the difference between the observed values and the predicted values based on our regression equation.  It turns out that minimizing the sum of the error terms doesn’t work since positive error will cancel out negative error so we minimize the sum of the squared error terms.[[19]](#footnote-19)

## **Part II – Getting the Regression Coefficients**

The regression equation will contain the values of a, B1, and B2 that minimize the sum of the squared errors.  There are formulas for computing these coefficients but usually we leave it to SDA to carry out the calculations.

Click on REGRESSION at the top of your screen and enter your dependent variable (tvhours) in the DEPENDENT box.  In Exercise 13 we ran two bivariate linear regressions – one with tvhours and age and a second with tvhours and educ.  In this exercise we’re going to use both independent variables simultaneously.  Enter both age and educ in the INDEPENDENT BOX so they become your independent variables.  Make sure that the SELECTION FILTER(S) box contains *year(2018) tvhours(0-12)*. The WEIGHT box should already be filled in. In the SAMPLE DESIGN line select SRS.  Under MATRICES TO DISPLAY, check the box for CORRELATION MATRIX.  Your screen should look like the following.[[20]](#footnote-20)

Now click RUN REGRESSION to produce the regression analysis. The first four boxes in your output are what you want to look at.

* The first box lists the variables you entered as your dependent, independent, weight, and filter variables.
* The second box gives you the regression coefficients.
  + The unstandardized regression coefficient for age (B1) is equal to 0.028 and the unstandardized regression coefficient for educ (B2) is -0.129.  This means that an increase of one unit in educ results in an average decrease of .129 units in tvhours after statistically adjusting for age.  Or, to put it another way, an increase of one year in the education results in an average decrease of .129 hours of television viewing after statistically adjusting for the respondent’s age.
  + An increase of one unit in age results in an average increase of .028 units in tvhours after statistically adjusting for education.  Or, to put it another way, an increase of one year in the age results in an average increase of .028 hours of television viewing after statistically adjusting for the respondent’s education.
  + So what does it mean to statistically adjust for something?  Suffice it to say that it means that B1 tells us the effect or influence of X1 on the dependent variable after taking into account the other independent variables (i.e., in this case X2).   The other regression coefficient, B2, would be similarly interpreted.
  + The regression constant is 3.039. This is referred to as the constant since it always stays the same regardless of which values of X1 and X2 you are using to predict Y.
  + The standard error of these coefficients which is an estimate of the amount of sampling error.
  + The standardized multiple regression coefficients (often referred to as Beta).  You can’t compare the unstandardized multiple regression coefficients (B1 and B2) because they have different units of measurement.  One year of age is not the same thing as one year of education.  The standardized multiple regression coefficients (Beta) are directly comparable.  You can see that the Beta for educ is -.181 and for age is .223 which means that age is relatively more important in predicting hours of television viewing than is education.
  + The t test which tests the null hypotheses that the population constant and population regression coefficients are equal to 0.
  + The significance value for each test. As you can see in this example, all three coefficients are statistically significant. However, we are usually not very interested in the constant coefficient.
  + The value of the Multiple R and Multiple R-Squared.  The Multiple R squared tells us that age and educ together explain or account for 8.5% of the total variation in the number of hours per day that respondents watch television.   In Exercise 13 we saw that age by itself explained 5.2% of the variation in tvhours and that educ by itself explained 3.6%.  Why can’t we just add 5.2% and 3.6% and say that 8.8% of the total variation is explained by these two variables together.  We see from the SDA output that this is not true.  In fact, 8.5% of the total variation is explained by these two variables together.  Why is that?  It’s because the variation explained by these two independent variables overlap and because of this overlap they only account for 8.5% of the total variation in the dependent variable.
  + The value of the Adjusted R-Squared which is 0.084.  The Adjusted R Square “takes into account the number of independent variables relative to the number of observations.” (George W. Bohrnstedt and David Knoke, *Statistics for Social Data Analysis*, 1994, F.E. Peacock, p. 293)
* The third box shows the Wald F-Statistic and its associated probability (P) which tests the null hypothesis that all the unstandardized and standardized regression coefficient are equal to 0.  Since the P value (.000) is less than .05 we can reject the null hypothesis and conclude that at least one coefficient is not 0.
* The fourth box shows the Pearson Correlation matrix for our three variables. As you can see, the correlations for age and tvhours, and for education and tvhours, are about the same. However, the correlation for age and education is virtually 0.

So our multiple regression equation for predicting Y is 3.039 + .028X1 - .129X2.  Thus, for a person that is 20 years old and whose father completed 12 years of school,  the predicted number of hours that he or she watches television 3.039 + (.028) (20) - .129 (12) or 3.039 + 0.56 – 1.548 or 2.051 hours.

It’s **important** to keep in mind that everything we have done assumes that our dependent variable is a “linear function of more than one independent variable.”  (Colin Lewis-Beck and Michael Lewis-Beck, Applied Regression – An Introduction, Sage Publications, 2015, p. 55)

## **Part III – It’s Your Turn Now**

Use the same dependent variable, tvhours, but this time add paeduc to your list of independent variables.  Now you will have three independent variables – age, educ, and maeduc.  The variable *pa*educ is the years of school completed by the respondent’s mother.  Use SDA to get the regression equation.

* Write out the regression equation.
* What do the unstandardized multiple regression coefficients (B1, B2, and B3) tell you?
* What do the standardized regression coefficients (Beta) tell you?
* What are the values of R and R2 and what do they tell you?
* What are the different tests of significance that you can carry out and what do they tell you?

## **Part IV – Do we have a problem?**

Multicollinearity occurs when the independent variables are highly correlated with each other.  If one of your independent variables is a perfect linear function of the other independent variables, then you would not be able to determine the regression coefficients.  But this is not typical.  What is more likely is that some of the independent variables might explain a large portion of the variation in another independent variable.  For example, in Part III what if both father’s education and age explained a very large portion of the variation in respondent’s education?  Then you would have high multicollinearity.  The problem that multicollinearity creates is that it tends to make your regression coefficients less reliable.  The standard errors of the regression coefficients increase which makes it harder to reject the null hypothesis in your t tests.

There are several ways to determine if multicollinearity is a problem in your analysis.  You can start by looking at the Pearson Correlation matrix for your independent variables.  Look at the correlation matrix in the SDA output to see if any of the independent variables are highly intercorrelated.   If they are, then you would have a problem but, in this example, it doesn’t appear this is the case.  There are other ways to detect multicollinearity but for this exercise we’ll stop here.

# Exercises for an Introductory Statistics Course Using SDA Edward Nelson, California State University, Fresno

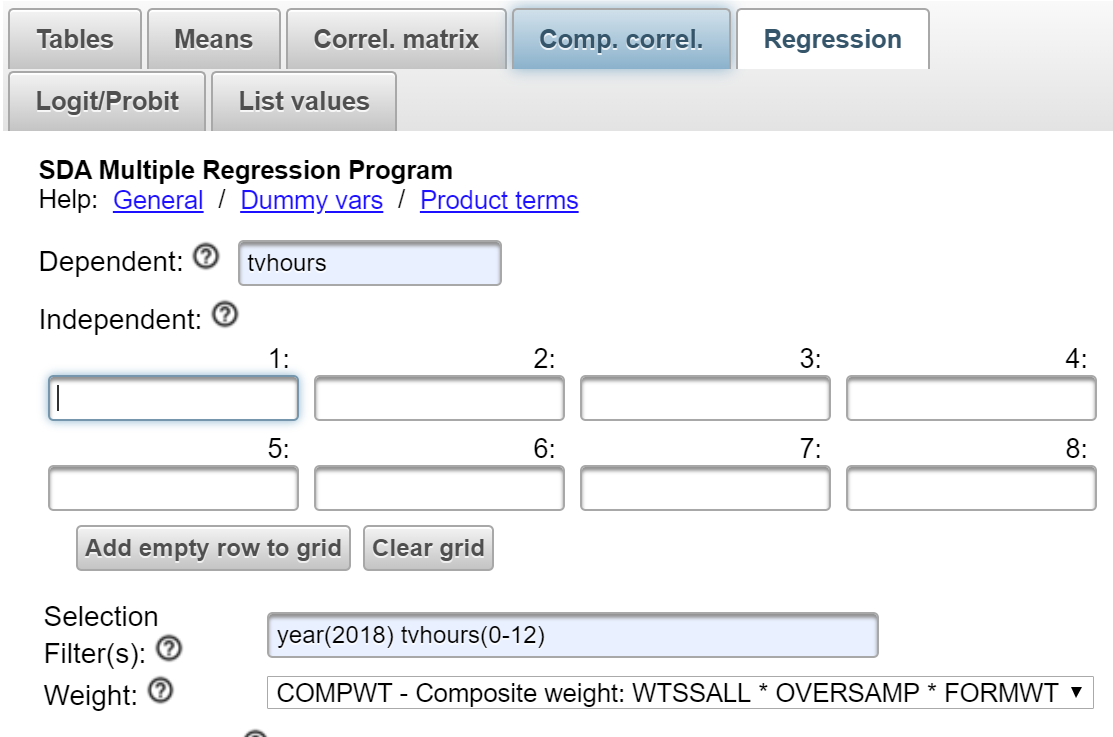
# Exercise 15 Dummy Variable Regression

## **Goals of Exercise**

The goal of this exercise is to introduce dummy variable regression.  The exercise also gives you practice using REGRESSION in SDA.

## **Part I –Dummy 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 the 2018 GSS.  To access the GSS cumulative data file in SDA format click [here](https://sda.berkeley.edu/sdaweb/analysis/?dataset=gss18).  The cumulative data file contains all the data from each GSS survey conducted from 1972 through 2018.  We want to use only the data that was collected in 2018.  Click on REGRESSION at the top of your screen. To select out the 2018 data, enter year(2018) in the Selection Filter(s) box.  Your screen should look like the following.



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.  You will need to click on SAMPLE OPTIONS and select SRS in the SAMPLE DESIGN box.   
  
The GSS is an example of a social survey.  The investigators selected a sample from the population of all adults in the United States.  This particular survey was conducted in 2018 and is a relatively large sample of a little more than 2,300 adults.  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.  Often we want to describe respondents in terms of social characteristics such as marital status, education, and age.  These are all variables in the GSS.

In Exercise 13 we considered linear regression for one independent and one dependent variable which is often referred to as bivariate linear regression.  Multivariate linear regression (see Exercise 14) expands the analysis to include multiple independent variables.  In both these exercises the variables in the regression analysis were interval or ratio (see Exercise 1). What do you do if you want to include a nominal or ordinal variable as one of your independent variables in the regression?

The answer is to create dummy variables.  Consider the respondent’s sex.  Since the variable sex has two categories (i.e., 1 for males and 2 for females), you could create two dummy variables – one for each value.  Your dummy variables would look like this.

* The dummy variable for males will equal 1 if male and 0 if female.
* The dummy variable for females will equal 1 if female and 0 if male.

In other words, you could create as many dummy variables as there are categories of your variable.

If there are k categories, then you would use k – 1 of the dummy variables in your regression analysis.  The category that you omit becomes your comparison group.  We could enter the dummy variable for males into the regression analysis and omit the dummy variable for females.  That means that females will be the comparison group.

What if you had more than two categories?  For example, the region where the respondent lives (region) has nine categories.  So you could create nine dummy variables and omit one of them.  If we decide to omit the category for the Pacific region (value 9), then you would use eight dummy variables in your regression analysis, one for each of the other categories, and the Pacific region would be our comparison group.

Neither of these two variables – sex and region – have missing data but if there had been missing data SDA would automatically assign a missing value to the dummy variable.

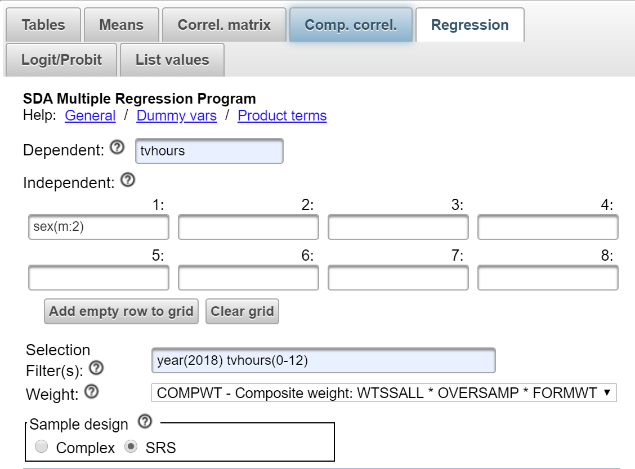
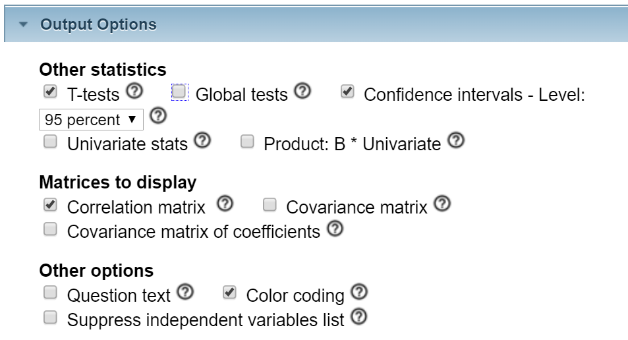
Let’s use tvhours as our dependent variable as we did in Exercises 13 and 14.  Run FREQUENCIES in SDA to get the frequency distribution for tvhours.  In the previous exercises we discussed outliers and noted that there are a few individuals (i.e., 13) who watched a lot of television which we defined as 13 or more hours per day.  We also noted that outliers can affect our statistical analysis so we decided to remove these outliers from our analysis.

Let’s exclude these individuals by selecting only those cases for which tvhours is less than 13. That way the outliners will be excluded from the analysis. To do this add tvhours(0-12) to the SELECTION FILTER(S) box.  Be sure to separate year(2018) and tvhours(0-12) with a space or a comma.  This will tell SDA to select out only those cases for which year is equal to 2018 **and** tvhours is less than 13.  Run FREQUENCIES in SDA to get a frequency distribution for tvhours after eliminating the outliers and check to make sure that there are no values greater than 12.

Click on REGRESSION at the top of the SDA page. To create the dummy variables you enter the variable name in the INDEPENDENT box.[[21]](#footnote-21)  For example, enter sex in the INDEPENDENT box.  Following the variable name, enter (m:X) where X stands for the code or value of the category that you want to omit.  For example, if you entered sex(m:2) that would mean that you are omitting the dummy variable for code or value 2 (i.e., females).  So females becomes your comparison group.

## **Part II – Regression with Dummy Variables**

Enter your dependent variable (tvhours) in the DEPENDENT box.  Enter the command for creating the dummy variable in the INDEPENDENT box.  In our case, we would enter *sex(m:2).*  (Don’t enter the period.)  Make sure that you enter *year(2018) tvhours(0-12)* in the SELECTION FILTER(S) box and that you have selected SRS in the SAMPLE DESIGN line.  The WEIGHT box should automatically be filled in. Under MATRICES TO DISPLAY, check the box for CORRELATION MATRIX.  Under OTHER STATISTICS, uncheck the GLOBAL TESTS box which we won’t need in this analysis.  Your screen should look like the following.

Now click RUN REGRESSION to produce the regression analysis. Let’s start by looking at the output that contains your unstandardized regression coefficients.  From this you can see that your regression equation for predicting tvhours is 2.584 + .106 X where X is your dummy variable.  Remember that your dummy variable equals 1 if the person is male and 0 if the person is female.  So for males the predicted number of hours watching television is 2.584 + .106 (1) or 2.690.  For females the predicted number of hours is 2.584 + .106 (0) or 2.584.  Since we left the dummy variable for females out of the regression equation, females become our comparison group.  The unstandardized regression coefficient (0.106) is the mean number of hours that males watch television (2.69) minus the mean for females (2.58) which is 0.11.[[22]](#footnote-22)

SDA will also calculate t tests to test the null hypotheses that the regression coefficients **in the population** are equal to 0.  Normally we’re only interested in the slope.  The t value for B is 0.937 and the significance value is .349.  This means that we can’t reject the null hypothesis.  In other words, we have no basis for asserting that the population slope is significantly different from zero.  The Pearson Correlation Coefficient Squared tells us that the dummy variable for sex explains virtually none of the variation in the dependent variable.

## **Part III – Now it’s Your Turn**

Use SDA to get the frequency distribution for race.  There are three categories for this variable – white (value 1), black (value 2), and other (value 3).  We want to compare whites with non-whites.  This means that there will be two dummy variables:

* The dummy variable for whites will equal 1 if the person is white and 0 if the person is non-white.
* The dummy variable for non-whites will equal 1 if the person is nonwhite and 0 if the person is white.

Let’s make non-whites our comparison group so that means that we’ll leave the dummy variable for non-whites out of the regression equation.  To do this in SDA, enter the following in the INDEPENDENT box – race(m:2-3).  (Don’t enter the period.)  This means that the values 2 and 3 will form the comparison group. Keep in mind that SDA will automatically assign a missing value to the dummy variable if *race* had any missing data.

Now run the regression analysis with tvhours as your dependent variable and race(m:2-3) as your independent variable.

* Write out the regression equation.
* What do the unstandardized multiple regression coefficients tell you?
* What are the values of R and R2 and what do they tell you?
* What are the different tests of significance that you can carry out and what do they tell you?

## **Part IV – Multiple Regression with Dummy Variables**

In Exercise 14 you did a regression analysis with tvhours as your dependent variable and age, maeduc, and educ as your independent variables.  This time we’re going to add the dummy variable for males into the analysis.  Use SDA to carry out the regression analysis for this model.

The regression equation for predicting tvhours is 2.701 + .031 (age) - .157 (educ) + .049 (maeduc) + .024 (dummy variable for males).  The unstandardized regression coefficients show the average change in the dependent variable when the independent variable increases by one unit after statistically adjusting for the other independent variables in the equation.  As age increases, television viewing increases. As mother’s education increases, television viewing also increases but as respondent’s education increases, television viewing decreases.  Males watch more television that females.  The t tests show that all the unstandardized coefficients are statistically significant with the exception of the dummy variable for males. The Pearson Multiple Correlation Coefficient Squared tells us that together the independent variables explain or account for 8.7% of the variation in television viewing.  The Adjusted Squared Correlation Coefficient adjusts for the number of independent variables and is slightly lower (8.5%).  The Beta values show the relative importance of the independent variables in predicting television viewing and tell us that age and respondent’s education are the most important variables.

## **Part V – Now it’s Your Turn Again**

Repeat the regression analysis you did in Part IV but instead of adding the dummy variable for males into the analysis, this time add the dummy variable you created in Part III for race.  This means you will have four independent variables -- age, maeduc, educ, and the dummy variable for race.

* Write out the regression equation.
* What do the unstandardized multiple regression coefficients tell you?
* What do the standardized regression coefficients (Beta) tell you?
* What are the values of R and R2 and what do they tell you?
* What are the different tests of significance that you can carry out and what do they tell you?

# Exercises for an Introductory Statistics Course Using SDA Edward Nelson, California State University, Fresno

# Appendix Notes on Using SDA

Survey Documentation and Analysis (SDA) was developed by the Survey Methods Program at UC Berkeley and is currently maintained by the Institute for Scientific Analysis. It’s an online statistical package that is freely available wherever you have internet access. SDA is easy to use and it doesn’t take very long to get up and running on SDA. It has a very good context-sensitive help menu that is easily accessible.

One of the limitations of SDA is that it’s not easy to develop data sets in SDA format. While SDA is freely available to use, it requires a site license to create your own SDA data sets. Developing SDA data sets is labor intensive and has a fairly steep learning curve. That means that one normally relies on data sets that others have developed. Fortunately, there are many high-quality SDA data sets available on the internet. Here are some places to go to find SDA data sets.

* [SDA Archive](http://sda.berkeley.edu/archive.htm) located at UC Berkeley including:
  + General Social Survey cumulative data file from 1972 through 2018
  + American National Election Studies from 1992 through 2016 and the ANES cumulative data file from 1952 through 2016
  + Census microdata
    - U.S. from 1990 through 2003
    - California for 1990 and 2000
  + Field Poll data from 1956 through 2006 .
* Inter-University Consortium for Political and Social Research (ICPSR) including:
  + many data set located in the ICPSR [general archive](http://www.icpsr.umich.edu)
  + [data-driven learning guides](http://www.icpsr.umich.edu/icpsrweb/instructors/biblio/resources) which are instructional exercises that can be used in the classroom
  + Investigating Community and Social Capital, an [instructional module](http://www.icpsr.umich.edu/icpsrweb/instructors/icsc/index.jsp) using SDA
  + Voting Behavior: the 2012 Election, another [instructional module](http://www.icpsr.umich.edu/icpsrweb/instructors/setups2012/index.jsp) using SDA
  + Voting Behavior: the 2016 Election, still another [instructional module](https://www.icpsr.umich.edu/icpsrweb/instructors/setups2016/) using SDA

In this series of exercises we’re going to use the 2018 General Social Survey. The GSS cumulative data file is freely available by clicking [here](https://sda.berkeley.edu/sdaweb/analysis/?dataset=gss18). The 2018 file can be extracted by using the SELECTION FILTER(S) box. All you need to do is to enter *year(2018)* in the box and SDA will extract the 2018 file for your analysis.

It’s important to weight the data so the sample better represents the population from which the sample is selected. SDA automatically enters the weight variable in the WEIGHTS box.

There are a couple of things that need to be mentioned about SDA. SDA doesn’t have a command to run a t test. The t test is a special case of one-way analysis of variance when the independent variable is a dichotomy.

The General Social Survey is not a simple random sample. Rather it is an area probability (cluster) sample. This means that standard errors for the GSS are larger than what you would get assuming simple random sampling. This, in turn, affects tests for statistical significance. SDA will compute standard errors using either complex standard errors for samples like the GSS or simple random sample (SRS) errors when the sample is a simple random sample. However, complex standard errors are beyond the scope of any introductory statistics course. After consultation with a statistician at UC Berkeley, his advice for introductory statistics classes was to assume simple random sampling even though we know it will produce an underestimate of the standard errors.

The default when you open the GSS data file is to compute complex standard errors. Therefore, it’s important that students click on the arrow next to OUTPUT OPTIONS and then click on the circle next to SRS to select it. This will be on the line that says SAMPLE DESIGN. When you use MEANS in SDA, students should check the box for SRS STD ERRS and uncheck the box for COMPLEX STD ERRS. However, they only need to make this change when they plan on using tests of significance. I incorporated this into the exercises but only when they were going to use tests of significance.

These exercises use the SDA 4.0 interface which is an update from the earlier 3.5 interface. The differences between these two interfaces is explained on the SDA [website](https://sda.berkeley.edu/archive.htm).

You have permission to use these exercises and to revise them to fit your needs. Feel free to revise the exercises in any way you want. Just recognize the source of the original exercises. I would be interested in hearing about your experiences when you use these exercises.

If you would like to contact me, please email me at [ednelson@csufresno.edu](mailto:ednelson@csufresno.edu). I’m Professor Emeritus at California State University, Fresno in the Sociology department. I taught research methods, statistics, and critical thinking before retiring and now teach a critical thinking course part time.

1. Stanley Smith Stevens, 1946, “On the Theory of Scales of Measurement,” Science 103 (2684), pp. 677-680. [↑](#footnote-ref-1)
2. You might wonder why we didn’t use an example from the GSS.  There isn’t one.  The variables in the GSS are nominal, ordinal, and ratio. [↑](#footnote-ref-2)
3. Frequency distributions can be grouped or ungrouped.  Think of age.  We could have a distribution that lists all the ages in years of the respondents to our survey.  One of the variables (*age*) in our data set does this.  But we could also divide *age* into a series of categories such as under 30, 30 to 39, 40 to 49, 50 to 59, 60 to 69, and 70 and older.  In a grouped frequency distribution the mode would be the most common category or categories.  [↑](#footnote-ref-3)
4. In a grouped frequency distribution the median would be the category that contains the middle value. [↑](#footnote-ref-4)
5. See Exercise 3 for a more through discussion of skewness.  [↑](#footnote-ref-5)
6. The Interquartile Range can also be used to measure variation for an interval or ratio variable and the Index of Qualitative Variation can be used to measure variation for nominal variables. [↑](#footnote-ref-6)
7. We’ll talk more about Analysis of Variance in Exercise 6. [↑](#footnote-ref-7)
8. One-way just means that there is only one independent variable. [↑](#footnote-ref-8)
9. Null means that it is a hypothesis of no difference.  In other words, there is no difference between the means for men and for women.  Notice that this refers to the mean of the population of all men and the mean of the population of all women.  In other words, we are using our sample data to test a hypothesis about population values.  [↑](#footnote-ref-9)
10. 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. [↑](#footnote-ref-10)
11. 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. [↑](#footnote-ref-11)
12. See Exercise 1 for a discussion of levels of measurement.  Nominal variables have no underlying order and ordinal variables have an underlying order.  Measures of association for nominal variables range from 0 to 1 while measures for ordinal variables range from -1 to +1. [↑](#footnote-ref-12)
13. There’s another popular measure called Lambda but SDA doesn’t compute it and it’s harder to compute by hand so we’re going to skip it. [↑](#footnote-ref-13)
14. If your table has two columns and two rows, V is equal to Phi which you might be familiar with.  Since Phi is a special case of V, we’re not going to discuss it. [↑](#footnote-ref-14)
15. That’s important because as we noted earlier some measures are asymmetric which means that it depends on which of the two variables is dependent and which is independent.  In these cases, SDA assumes that the row variable is the dependent variable.  [↑](#footnote-ref-15)
16. There is a symmetric value for Somer’s d which SDA does not compute. [↑](#footnote-ref-16)
17. For example, a unit could be a year in age or a year in education depending on the variable we are talking about. [↑](#footnote-ref-17)
18. When you square a value the result is always a positive number.  [↑](#footnote-ref-18)
19. When you square a value the result is always a positive number.  [↑](#footnote-ref-19)
20. In Exercise 13 we unchecked the box for GLOBAL TESTS in the OTHER STATISTICS line.  We’re going to need these global tests for this exercise so leave it checked. [↑](#footnote-ref-20)
21. We only use dummy variables for the independent variables and not for the dependent variable. [↑](#footnote-ref-21)
22. The difference between .106 and 0.11 is due to the fact that SDA calculated the means to two decimal points and the regression coefficient to three decimal points.  [↑](#footnote-ref-22)