# Exercises for an Introductory Statistics Course Using PSPP

# Edward Nelson

# California State University, Fresno

# Table of Contents

 Page

Preface 3

Introduction to PSPP and the Exercises 4

Exercise 1 – Levels of Measurement 7

Exercise 2 – Frequency Distributions 11

Exercise 3 – Measures of Central Tendency and Dispersion 13

Exercise 4 – Measures of Skewness and Kurtosis 18

Exercise 5 – Hypothesis Testing – One-Sample T Test 21

Exercise 6 – Hypothesis Testing – Independent Samples T test 26

Exercise 7 – Hypothesis Testing – Paired Samples T Test 31

Exercise 8 – Hypothesis Testing – One Way Analysis of Variance 35

Exercise 9 – Crosstabulation 39

Exercise 10 – Chi-Square 44

Exercise 11 – Measures of Association 49

Exercise 12 – Spuriousness 54

Exercise 13 –Correlation 58

Exercise 14 – Bivariate Linear Regression 62

Exercise 15 – Multivariate Linear Regression 65

Exercise 16 – Dummy Variable Regression 69

Appendix – Codebook for GSS18A.SAV 73

# Exercises for an Introductory Statistics Course Using PSPPEdward 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 these statistics.

# The exercises assume a basic familiarity with PSPP although there is quite a bit of explanation of how to run the various statistical procedures in PSPP. There is a [tutorial](http://ssric.org/node/699) on using PSPP which is freely available on the Social Science Research and Instructional Center’s [website](http://ssric.org).

These exercises use a subset of the 2018 General Social Survey conducted by the National Opinion Research Center. The data set for these exercises can be downloaded by clicking on this [link](http://ssric.org/files/2020-02/GSS18A_1.sav).

The exercises were written so that each exercise was 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. Please contact him with any questions you might have.

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

# Introduction to PSPP and the Exercises

PSPP is a statistical analysis program made available at no charge to users by the *Free Software Foundation*. There are two versions: the syntax version and the less comprehensive but more user-friendly graphical interface. These exercises use PSPP and the graphical interface with a few exceptions. We’ll use the syntax version for selecting particular cases for analysis (see Exercises 14, 15, 16), for running crosstabulations with more than two variables (see Exercises 9, 11, 12), and for creating new variables using IF commands (see Exercise 16).

For more information on PSPP, click [here](http://www.gnu.org/software/pspp/). Their website says that “GNU PSPP is a program for statistical analysis of sampled data. It is a Free replacement for the proprietary program SPSS and appears very similar to it with a few exceptions.”

PSPP is similar to [SPSS](https://www.ibm.com/analytics/spss-statistics-software) which has gone through many iterations and served as a widely used standard for analyzing quantitative data. Over time SPSS has become much more user-friendly, so that even students with no background in statistical analysis can master it as part of a single introductory statistics or research methods course. It has also added more and more features including the ability to produce a wide range of graphs.

Despite its many advantages, one thing that SPSS is not is *free*. As of this writing, the cost of a base subscription to SPSS starts at $99 *per month*. This need not trouble you if you are a student at a college or university that has purchased a site license. Even if this is not the case, you can obtain a version available only to faculty and students at a deep discount. But still not free. (Check Amazon.com or other vendors for details.)

The easiest way to download PSPP is to click [here](http://pspp.awardspace.info/) and look for the “Downloads” box. Then download the latest version in either 32-bit or 64-bit format. If you’re not sure which version to download, go to your control panel and click on “System” and look for your system type. Then follow the instructions to download.

You can open SPSS data files (both .sav and .por) in PSPP. You can also open Excel files (both .csv and .txt). And, of course, you can create your own data files in PSPP.

PSPP will list the variables in your data file and you can select those variables you want to use. It’s easier to find the variables if they are listed by variable names. You can change the way PSPP lists the variables by right clicking anywhere on the list of variables and checking or unchecking the box for PREFER VARIABLE LABELS.

An even better way to do this is to click on EDIT in PSPP and then on OPTIONS. Click on DISPLAY NAMES and then on SORT BY NAME. If you are using your own computer, PSPP will remember your choices and you won’t have to do it each time you open PSPP. However, if you are working in a computer lab, you may have to do it each time you open PSPP. SORT BY NAME means that all your variables will be arranged alphabetically which will make it easier to find the variables you want to use.

All variable names in this series of exercises are in lower case with italics and all PSPP commands and dialog choices are in upper case.

## Data Set Used in These Exercises

In this tutorial, we’ll be using a subset of the General Social Survey (GSS). The GSS is a biannual national survey conducted by the National Opinion Research Center and used for teaching and research in a variety of disciplines since 1972. We’ve created a subset of the 2018 survey. Some new variables were created and a few were recoded. You can download the subset by clicking  [here](http://ssric.org/files/2020-02/GSS18A_1.sav) . The data file that you download has already been weighted so that the sample better represents the population from which it was selected. One consequence of this is that frequencies will appear as decimal values. That’s just a function of the weighting procedure and can be ignored. In these exercises the frequencies are rounded to the nearest whole number.

# **Other Resources**

There is a freely available [tutorial](http://ssric.org/node/699) for PSPP on the Social Science Research and Instructional Center’s website. There are also tutorials for SPSS [version 25](http://ssric.org/node/686) and [version 26](http://ssric.org/node/696). While you are at the website, take a look at the other [modules and exercises](http://ssric.org/trd/exercises) that are freely available.

The graphical interface version of PSPP is most limited in its minimal coverage of graphics, offering only pie charts, bar charts, histograms, and scatterplots, and these with very few options. Fortunately, another package, also freely available, is [*Statistics Open for All*](http://sofastatistics.com/home.php) *(SOFA),* which includes much more extensive graphics capabilities.George Self has developed a comprehensive lab manual for this package. Designed for use by his own students, he has not published it on the Internet. He has, however, generously granted us permission to post it [here](https://ssric.org/files/2019-10/G_SELF_LabManual.pdf) on the site of the California State University Social Science Research and Instructional Center. He has copyrighted his manual under the Creative Commons “Attribution-ShareAlike 4.0 International license,” which is even more open than the “Attribution-NonCommercial-ShareAlike 3.0 license” which, except where noted, governs our own site.

For another tutorial on the graphical interface version of PSPP, see Gary Fisk, [*PSPP for Beginners*](https://garyfisk.com/pspp/). GNU has produced a comprehensive [users’ guide](https://www.gnu.org/software/pspp/manual/) for the syntax version of PSPP.

## Exercises in This Module

* Exercise 1 – Levels of Measurement
* Exercise 2 – Frequency Distributions
* Exercise 3 – Measures of Central Tendency and Dispersion
* Exercise 4 – Measures of Skewness and Kurtosis
* Exercise 5 – Hypothesis Testing – One-Sample T Test
* Exercise 6 – Hypothesis Testing – Independent Samples T test
* Exercise 7 – Hypothesis Testing – Paired Samples T Test
* Exercise 8 – Hypothesis Testing – One Way Analysis of Variance
* Exercise 9 – Crosstabulation
* Exercise 10 – Chi-Square
* Exercise 11 – Measures of Association
* Exercise 12 – Spuriousness
* Exercise 13 – Correlation
* Exercise 14 – Bivariate Linear Regression
* Exercise 15 – Multivariate Linear Regression
* Exercise 16 – Dummy Variable Regression
* Appendix – Codebook for GSS18A.SAV

## First Exercise

Exercise 1 will focus on levels of measurement and get you started using PSPP.

# Exercises for an Introductory Statistics CourseEdward Nelson, California State University, Fresno

# Exercise 1Levels of Measurement

## Goal 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 PSPP.

## 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” 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 must select measures or indicants of these concepts.  Religiosity can be measured in a number of different ways including, for example, 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 a subset of the 2018 GSS. Your instructor will tell you how to access this data set which is called GSS18A.SAV.

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 a 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 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 PSPP for the variable *marital* so you can see the frequency distribution for a nominal level variable.[[2]](#footnote-2)  It wouldn’t matter how we ordered these categories.

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 PSPP 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 them 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.[[3]](#footnote-3)

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 3.

## Part III – Now It’s Your Turn

Run FREQUENCIES for the following variables in the GSS. 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 what PSPP calls the valid responses.  The missing responses represent missing data (e.g., don’t know or no answer responses).

* *satfin*
* *happy*
* *hrs1*
* *partyid*
* *relig*
* *denom*
* *reliten*
* *age*

## Next Exercise

Exercise 2 will focus on measures of frequency distributions.

# Exercises for an Introductory Statistics CourseEdward Nelson, California State University, Fresno

# Exercise 2Frequency Distributions

Frequency distributions show you the number of cases for each category of your variable. They also convert these frequencies to percents and tell you how many cases had missing information. Click on ANALYZE, then on DESCRIPTIVE STATISTICS, and finally on FREQUENCIES.

Select the variable(s) for which you want to get a frequency distribution by clicking on them and moving them to the VARIABLES box and then clicking on the arrow that should be pointing to the right. We’re going to use *educ* in this example so move *educ* over to the VARIABLES box.

To get the frequency distribution click on OK. You’ll notice that PSPP also gives you some additional information such as the name of the data file you are using and various statistics which we’ll discuss next. It shows you the PSPP commands that it just ran in syntax form. There are two ways to run PSPP. You can use the interface by clicking on the commands you want to run and then filling in the dialog boxes or by writing out the command in syntax form. We’re only going to use the interface most of the time in these exercises with a few exceptions.

### **Part 1 – Statistics**

In the lower right-hand of the dialog box, you’ll see a list of statistics that PSPP will compute. This list is longer that what fits in the box, so you’ll have to scroll down to see the full list. Notice that the first four statistics are already checked (i.e., mean, standard deviation, minimum value, maximum value). These are the default statistics which you will get automatically when you run a frequency distribution. You can choose not to get these statistics by unchecking them and you can also choose to get other statistics by checking them.

The statistics you choose to run are partially dictated by the level of measurement of the variables which are often classified as nominal, ordinal, interval, and ratio (see Exercise 1).

Since *educ* is a ratio variable, we could use the mean, median, and mode as our measures of central tendency and the standard deviation and variance as our measures of variability. If our variable was class (i.e., ordinal), then we couldn’t use the mean but could use the median and the mode as our measures of central tendency; the standard deviation and variance wouldn’t be appropriate measures of variability. If our variable was *marital* (i.e., nominal), then we could only use the mode as our measure of central tendency.

So, for *educ* we’re going to ask for the mean, median, mode, minimum value, maximum value, and standard deviation. Notice that in the output the median is listed as the 50th percentile.

There are some other options which we will briefly note.

* You can choose to include the missing values in the computation of the various statistics. However, you would rarely want to do that.
* You can select CHARTS. We’ll discuss charts and graphs next.
* You can select FREQUENCY TABLES and change the way the frequency distributions are displayed. We’ll be using the defaults here, so you won’t need to worry about this.

## **Part II – Charts and Graphs**

As part of FREQUENCIES, PSPP will construct pie charts, bar charts, and histograms. (Bar charts and histograms can also be produced by choosing GRAPHS in the menu bar.)

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 but can be used with interval or ratio variables which have a small number of categories.

A bar chart is a chart that shows the frequencies or percents of a variable and is presented as a series of vertical bars that do not touch each other.  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.

A histogram is a graph that shows the frequencies or percents of a variable with a larger number of categories. It is presented as a series of vertical bars that touch each other. The height of each bar is proportional to the number of cases or the percent of cases in each category. It is used with interval or ratio variables.

To get a chart from PSPP, click on the CHARTS button and check the box for the type of chart you want. If you don’t want to get the frequency distribution, click on FREQUENCIES TABLES and select NEVER under DISPLAY FREQUENCIES TABLES.

## Part III – Now It’s Your Turn

Run FREQUENCIES to get a frequency distribution for the following variables: *educ*, *marital*, degree, *class*, and *age*. For each variable answer the following questions. Which measure(s) of central tendency did you choose and why? Use the percentages and the measure of central tendency to write a paragraph describing each variable.

# Exercises for an Introductory Statistics CourseEdward Nelson, California State University, Fresno

# Exercise 3Measures of Central Tendency and Dispersion

## Goal of Exercise

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

## 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[[4]](#footnote-4).  The median is the middle value of a distribution.[[5]](#footnote-5) 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 a subset of the 2018 GSS. Your instructor will tell you how to access this data set which is called GSS18A.SAV.

Run FREQUENCIES in PSPP for the variable *sibs*. Once you have selected this variable, then check the boxes for mode, median, and mean in the STATISTICS box and click on the CHARTS button.  Select DRAW HISTOGRAM and check the box for SUPERIMPOSE NORMAL CURVE.  Click on CONTINUE and then click on OK. To see your output click on the PSPP icon at the bottom of your screen (i.e., looks like a red circle with a blue cutout in the upper right of the circle) and click on the output window.  PSPP will open the Output window and display the results that you requested.

Your output will display the frequency distribution for *sibs* and a box showing the mode, median, and mean with the following values displayed.

* Mode = 1 meaning that one brother or sister was the most common answer (21.45%) from the 2,343 respondents who answered this question.  However, not far behind are those with two siblings (20.57%).  So while technically one sibling is the mode, what you really found is that the most common values are one and two siblings.  Another part of your output is the histogram which is a chart or graph of the frequency distribution.  The histogram clearly shows two most common values (i.e., the highest bars in the histogram).  So we would want to report that these two 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.  The cumulative percent column of the frequency distribution tells you that 46.39% of the cases have two or fewer siblings and that 62.63% 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 8,142  Dividing that by the number of responses or 2,343 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?[[6]](#footnote-6) Go back and look at the histogram for *sibs*. Notice that there is a long tail to the right of the distribution.  Most of the values are at the lower level – one, two, and three siblings.  But there are quite a few respondents who report having four or more siblings and about 4% 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 and two siblings.  But what if there is more than one clearly defined peak?  For example, consider a hypothetical distribution of 100 cases in which there are 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.  Here we would report all three measures of central tendency.  Go back to your PSPP output and look at the histogram for *sibs*.  When you told PSPP to give you the histogram you checked the box that said, “Superimpose normal curve.”  The normal curve doesn’t fit the histogram perfectly, particularly at the lower end, but it does suggest that it approximates a normal curve, particularly at the upper end.

## Part III – Now It’s Your Turn

Run FREQUENCIES for the variables below.  Once you have selected the variables, then check the boxes for mode, median, and mean in the STATISTICS box and click on the CHARTS button.  Select either DRAW HISTOGRAM or DRAW BAR CHART depending on what is indicated to the right of the variable name.  Click on CONTINUE and then click on OK. PSPP will open the Output window and display the results of what you requested.  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.

* *happy* – draw bar chart
* *partyid* – draw bar chart
* *reliten* – draw bar chart
* *age* – draw histogram and superimpose normal curve on histogram

## Part IV – 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 appropriate for interval and ratio level variables (see Exercise 4).[[7]](#footnote-7)  We’re going to discuss four such measures – the range, the inter-quartile 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 PSPP to compute it for you.  Click on RANGE in the STATISTICS box.  You should get 71 which is 89 – 18.[[8]](#footnote-8)  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.

A more stable measure of dispersion is the interquartile range which is the difference between the third quartile (Q3) and the first quartile (Q1).  The third quartile is the same as the seventy-fifth percentile which is the value that has 25% of the cases above it and 75% of the cases below it.  The first quartile is the same as the twenty-fifth percentile which is the value that has 75% of the cases above it and 25% of the cases below it.  Look at the cumulative percent column in the frequency distribution for age.  The first quartile will be the category than contains the cumulative percent of 25.0 and the third quartile will be the category that contains the cumulative percent of 75.0.  Once you know Q3 and Q1 you can calculate the interquartile range by subtracting Q1 from Q3.  Since it’s not based on the most extreme values, it will be more stable from sample to sample.  From the cumulative percent column you should see that Q3 will equal 60 and Q1 will equal 32 and the interquartile range will equal 60 – 32 or 28.

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.  Select standard deviation and variance from the STATISTICS box to tell PSPP to calculate both the standard deviation and the variance.  Note that the variance is the standard deviation squared.

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 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.

## Part V – Now It’s Your Turn Again

Run FREQUENCIES for the variables *hrs1* and *hrs2*.  Once you have selected the variables then check the boxes for range, standard deviation, and mean in the STATISTICS box.  Click on OK and click on the Output window and PSPP will display the results of what you requested.  For each variable write a sentence or two indicating what the values of these statistics are and what they mean.

## Next Exercise

Exercise 4 will focus on measures of skewness and kurtosis.

# Exercises for an Introductory Statistics CourseEdward Nelson, California State University, Fresno

# Exercise 4Measures of Skewness and Kurtosis

## Goal 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 PSPP.

## Part I – 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.



The best way to determine the skewness of a distribution is to tell PSPP to give you a histogram along with the mean and median.  PSPP will also compute a measure of skewness.  We’re going to use the General Social Survey (GSS) for this exercise.  The GSS is a national probability sample of adults in the United States conducted by the National Opinion Research Center (NORC).  The GSS started in 1972 and has been an annual or biannual survey ever since. For this exercise we’re going to use a subset of the 2018 GSS. Your instructor will tell you how to access this data set which is called GSS18A.SAV.

Let’s use age as our variable. Click on ANALYZE, then DESCRIPTIVE STATISTICS, and finally on FREQUENCIES. Select *age* and check the boxes for mean, median, skewness, and kurtosis in the STATISTICS box and click on the CHARTS button and select DRAW HISTOGRAM and SUPERIMPOSE NORMAL CURVE.  Click on CONTINUE and then on OK.  We’ll talk about kurtosis in a little bit.

Notice that the mean is larger than the median for both variables.  This means that the distribution is positively skewed.  Look at the histograms and you’ll see that 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.

## Part II – Now It’s Your Turn

Run FREQUENCIES for the variables *hrs1* and *tvhours*.  Tell PSPP to give you the histogram and to superimpose the normal curve on the histogram. Also ask for the mean, median, and skewness.  Write a paragraph for each variable explaining what these statistics tell you about the skewness of the variables.

## Part III -- 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.



PSPP 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 5.14 indicating a more peaked distribution.  To see this visually look at your histograms.

## Part IV – Now It’s Your Turn Again

Run FREQUENCIES for the variables *maeduc* and *paeduc*.  Tell PSPP to give you the histogram and to superimpose the normal curve on the histogram. Also ask for kurtosis.  Write a paragraph for each variable explaining what this statistic tell you about the kurtosis of the variables.

## Next Exercise

Exercise 5 will focus on hypothesis and the one-sample t test.

# Exercises for an Introductory Statistics CourseEdward Nelson, California State University, Fresno

# Exercise 5Hypothesis Testing and the One-Sample T Test

## Goal of Exercise

The goal of this exercise is to explore hypothesis testing and the one-sample t test. The exercise also gives you practice in using COMPARE MEANS (one-sample t test) and SELECT CASES in PSPP.

## Part I – Simple Random Sampling

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.

There are many ways of selecting a probability sample but the most basic type of probability sample is a simple random sample in which everyone in the sample has the same chance of being selected in the sample.  PSPP will select a simple random sample for you.  We’re going to use the General Social Survey (GSS) for this exercise.  The GSS is a national probability sample of adults in the United States conducted by the National Opinion Research Center (NORC).  The GSS started in 1972 and has been an annual or biannual survey ever since. For this exercise we’re going to use a subset of the 2018 GSS. Your instructor will tell you how to access this data set which is called GSS18A.SAV.  It’s a large sample of a little more than 2,300 individuals.  To illustrate simple random sampling, we’re going to select a simple random sample of 30% of all the individuals in the GSS.[[9]](#footnote-9)

Start by getting a frequency distribution for the variable *educ* which is the last year of school completed by the respondent.  You’ll see that there are a total of 2,348 cases.  A few of the respondents said they didn’t know or refused to answer the question.  There were 2,345 valid cases that actually answered the question.

Now click on DATA in the menu bar at the top of the screen.  This will open a drop-down box. Click on SELECT CASES.  Then click on RANDOM SAMPLE OF CASES and then on SAMPLE in the box below.  One of the options will already be selected and will say APPROXIMATELY \_\_\_ % OF ALL CASES.  Fill in 30 in the box indicating that you want to select a simple random sample of 30% of all the cases in the GSS.  Click on CONTINUE and then on OK.  Now run FREQUENCIES again for the variable, *educ*.  Your sample will be smaller than before.  This is a random sample of all the cases in the GSS.

## Part II.  Hypothesis Testing – the One-Sample T test

Let’s think about our variable, *educ*.  What do we know about education in the United States?  One thing we know is that the average years of school completed has been increasing over the twentieth and twenty-first centuries.  It used to be that many people stopped after completing high school which would be 12 years of education.  Now more go on to college.  So we would hypothesize that the mean years of school completed is now greater than 12.  How could we test that hypothesis?  We need a statistical procedure to do that. The t test is one of a number of statistical tests that we can use to test hypotheses.

Notice how we are going about this.  We have a sample of adults in the United States (i.e., the 2018 GSS).  We can calculate the mean years of school completed by all the adults in the **sample** who answered the question.  But we want to test the hypothesis that the mean years of school completed in the **population** of all adults is greater than 12.  We’re going to use our sample data to test a hypothesis about the population.

What do we know about sampling?  We know that 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.  Another thing we know is that the larger the sample size, the less the sampling error.

So the hypothesis we want to test is that the mean years of school completed in the population is greater than 12.  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.[[10]](#footnote-10) In our case, the null hypothesis would be that the mean years of school completed in the population is equal to 12. If we can reject the null hypothesis then we have evidence to support the research hypothesis. If we can’t reject the null hypothesis then we don’t have any evidence in support of the research hypothesis.  You can see why this is called a method of indirect proof. We can’t prove the research hypothesis directly but if we can reject the null hypothesis then we have indirect evidence that supports the research hypothesis.

Here are our two hypotheses.

* research hypothesis – the population mean is greater than 12
* null hypothesis – the population mean is equal to 12

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

Before we carry out the t test, let’s make sure we are using the full GSS sample and not the 30% simple random sample.  Click on DATA and on SELECT CASES.  Select ALL CASES and then click on OK.  Now you are using all the cases.

Now click on ANALYZE in the menu bar which will open a drop-down menu.  Click on COMPARE MEANS which will open another drop-down menu and click on ONE-SAMPLE T TEST.  Move the variable, *educ*, over to the TEST VARIABLE(S) box on the right.  Below the box on the right you will see a box called TEST VALUE.  This is where we enter the value specified in the null hypothesis which in our case is 12.  All you have to do now is click on OK.

You should see two output boxes.  The first box will have four values in it.

* N is the number of cases for which we have valid information[[11]](#footnote-11) (i.e., the number of respondents who answered the question).  In this problem, N equals 2,345.
* Mean is the mean years of school completed by the respondents in the sample who answered the question (see Exercise 3).  In this problem, the sample mean equals 13.73.
* Standard Deviation is a measure of dispersion (see Exercise 3).  In this problem, the standard deviation equals 3.02.
* Standard Error of the Mean is an estimate of how much sampling error there is.  In this problem, the standard error equals .06.

The second box will have five values in it.

* t is the value of the t test
* df is the number of degrees of freedom
* Significance (2-tailed) value
* Mean Difference
* 95% Confidence Interval of the Difference which we’re not going to discuss in this exercise
* 95% Confidence Interval of the Difference which we’re not going to discuss in this exercise

There is a formula for calculating the value of t in the t test.  Your instructor may or may not want you to learn how to calculate the value of t.  I’m going to leave it to your instructor to do this.  In this problem t equals 27.77.

Degrees of freedom (df) is the number of values that are free to vary.  If the sample mean equals 13.73 then how many values are free to vary?  The answer is N – 1 which is 2,345 – 1 or 2,344.  See if you can figure out why it’s 2,344.  Your instructor will help you if you are having trouble figuring it out.

The significance value is a probability.  It’s the probability that you would be wrong if you rejected the null hypothesis.  It’s .000 which you would think is telling you that there is no chance of being wrong if you rejected the null hypothesis.  But it’s actually a rounded value and it means that the probability is less than .0005 or less than five in ten thousand.  So there is a chance of being wrong but it’s really, really small.

The mean difference is the difference between the sample mean (13.73) and the value specified in the null hypothesis (12).  So it’s 13.73 – 12 or 1.73. That’s the amount that your sample mean differs from the value in the null hypothesis.  If it’s positive, then your sample mean is larger than the value in the null and if it’s negative, then your sample mean is smaller than the value in the null.

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 significance value which is less than .0005.  That tells you that the probability of being wrong if you rejected the null hypothesis is less than five out of ten thousand.  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.

But wait a minute.  The PSPP output said this was a two-tailed significance value. What does that mean?  Look back at the research hypothesis which was that the population mean was greater than 12.  We’re actually predicting the direction of the difference.  We’re predicting that the population mean will be greater than 12.  That’s called a one-tailed test and we have to use a one-tailed significance value.  It’s easy to get the one-tailed significance value if we know the two-tailed significance value.  If the two-tailed significance value is less than .0005 then the one-tailed significance value is half that or .0005 divided by two or .00025.  We still reject the null hypothesis which means that we have evidence to support our research hypothesis. We haven’t proven the research hypothesis to be true but we have evidence to support it.

## Part III.  Now It’s Your Turn

There is another variable in the 2018 GSS called *hrs1* which is the number of hours that the respondent worked last week if he or she was employed.  Many people have suggested that Americans are working longer hours than they used to.  Since the traditional work week is 40 hours, if it’s true that we’re working more hours our research hypothesis would be that the mean number of hours worked last week would be greater than 40.  Do a one-sample t test to test this hypothesis.  For each value in the output that we discussed, explain what it means.  Then decide whether you should reject or not reject the null hypothesis and what this tells you about the research hypothesis.

I’ll tell you that you should reject the null hypothesis even though the mean difference was just a little more than one hour.  You might wonder why you reject the null hypothesis when the mean difference is so small.  Notice that we have a large sample (N = 1,417).  Let’s see what happens when we have a sample that’s only 10% of that size.  Take a simple random sample of 10% of the total sample.  (Look back at Part I to see how to do this.)  Now we have a much smaller sample size.  Rerun your t test and see what happens with a smaller sample.  For each value in the output that we discussed, explain what it means.  Then decide whether you should reject or not reject the null hypothesis and what this tells you about the research hypothesis.

Now you probably won’t be able to reject the null hypothesis.[[12]](#footnote-12)  Why?  Remember that we said the smaller the sample, the more the sampling error.  If there is more sampling error, it’s going to be harder to reject the null hypothesis.  You can see this by looking at the standard error of the mean.  It will probably be smaller in the larger sample and bigger in the smaller sample.  So when you have a really large sample don’t get too excited when you reject the null hypothesis even though you have only a small mean difference.

## Next Exercise

Exercise 6 will focus on hypothesis testing and the independent-samples t test.

# Exercises for an Introductory Statistics CourseEdward Nelson, California State University, Fresno

# Exercise 6Hypothesis Testing and the Independent-Samples T Test

## Goal of Exercise

The goal of this exercise is to explore hypothesis testing and the independent-samples t test. The exercise also gives you practice in using COMPARE MEANS in PSPP.

## 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 a subset of the 2018 GSS. Your instructor will tell you how to access this data set which is called GSS18A.SAV.

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?

Click on ANALYZE in the menu bar and then on COMPARE MEANS and finally on MEANS.  Select the variables *educ* and *hrs1* and move them to the DEPENDENT LIST box.  These are the variables for which you are going to compute means.  Then select the variable *sex* and move it to the INDEPENDENT LIST 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 PSPP will show you the mean, number of cases, and standard deviation for men and women for these two variables.

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 as we did in the previous exercise see (see Exercise 5).

## Part II – Now it’s Your Turn

In this part of the exercise you want to compare men and women to determine whether men or women watch more television. This is variable *tvhours* in the GSS. Use PSPP to get the sample means and then compare them to begin answering this question.

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

In Part I we compared the mean scores for men and women for the variable educ and hrs1. Now we want to determine if those differences are statistically significant by carrying out the independent-samples t test.

A t test is used when you want to compare **two** groups.  The grouping variable defines these two groups.  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 establishing a cut point or 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). The cut point is the value that makes this into a dichotomy.  All values less than the cut point are in one category and all values equal to or larger than the cut point are in the other category.  If your cut point is 3, then values 1 and 2 are in one category and value 3 is in the other category.

Click on ANALYZE and then on COMPARE MEANS and finally on INDEPENDENT-SAMPLES T TEST.” Move the two variables listed above into the TEST VARIABLE(S) box.  These are the variables for which you want to compute the mean scores.  Right below the TEST VARIABLE(S) box is the GROUPING VARIABLE box.  This is where you indicate which variable defines the groups you want to compare.  In this problem the grouping variable is *sex*.  Once you have entered the grouping variable, then enter either the values of the two groups or the cut point.

In our case, you would enter 1 for male into Group 1 and 2 for females into Group 2.  It wouldn’t matter which was Group 1 and which was Group 2.  Finally click on OK.

You should see two boxes in the output screen. The first box gives you four pieces of information.

* N which is the number of males and females on which the t test is based.  This includes only those cases with valid information.  In other words, cases with missing information (e.g., don’t know, no answer) are excluded.
* Means for males and females.
* Standard deviations for males and females.
* Standard error of the mean for males and females which is an estimate of the amount of sampling error for the two samples.

The second box has more information in it.  The first thing you notice is that there are two t tests for each variable.  One assumes that the two populations (i.e., all males and all females) have equal population variances and the other doesn’t make this assumption.  In our two examples, both t tests give about the same results.  We’ll come back to this in a little bit.  The rest of the second box has the following information.  Let’s look at the t test for *educ*.

* t is the value of the t test which is -0.47 for both t tests.  There is a formula for computing t which your instructor may or may not want to cover in your course.
* Degrees of freedom in the first t test is (Nmales – 1) + (Nfemales – 1) = Nmales + Nfemales - 2 = 2,343.  In the second t test the degrees of freedom is estimated.
* The significance (two-tailed) value which we’ll cover in a little bit.
* The mean difference is the mean for the first group (males) – the mean for the second group (females) = 13.70 – 13.76 = -.06.  In other words, females have .06 of a year more education than males which is a very small difference.
* The standard error of the difference which is .13 is an estimate of the amount of sampling error for the difference score.
* 95% confidence interval of the difference which we’re not going to talk about in this exercise.

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 significance value which is .635 for both t tests.  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.

But wait a minute.  The SPSS output said this was a two-tailed significance value. What does that mean?  Look back at the research hypothesis which was that the population mean for men minus the population mean for women does not equal 0.   We’re not predicting that one population mean will be larger or smaller than the other.  That’s called a two-tailed test and we have to use a two-tailed significance value.  If we had predicted that one population mean would be larger than the other that would be a one-tailed test.  It’s easy to get the one-tailed significance value if we know the two-tailed significance value.  If the two-tailed significance value is .635 then the one-tailed significance value is half that or .635 divided by two or .3175.

We still haven’t explained why there at two t tests.  As we said earlier, one assumes that the two populations (i.e., all males and all females) have equal population variances and the other doesn’t make this assumption.  To compute the t value we need to estimate the population variances (see Exercise 3).  If the population variances are about the same, we can pool our two samples to estimate the population variance.  If they are not about the same, we wouldn’t want to do this.  So how do we decide which t test to use?  Here’s where we’ll talk about the Levene’s test for the equality of variances which is in the second box in your SPSS output.  For this test, the null hypothesis is that the two population variances are equal.  The F value is .50 and the significance value is .481. Since the significance value is not less than .05, we don’t reject the null hypothesis that the population values are equal. That means we want to use t test that assumes equal population variances.

## PART IV – Now It's Your Turn Again

In this part of the exercise you want to compare men and women to whether men or women watch more 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.  We’re going to discuss the paired-samples t test in the Exercise 7.

## Next Exercise

Exercise 7 will focus on hypothesis testing and the paired-samples (dependent) t test.

# Exercises for an Introductory Statistics CourseEdward Nelson, California State University, Fresno

# Exercise 7Hypothesis Testing and the Paired-Samples (Dependent) T Test

## Goal of Exercise

The goal of this exercise is to explore hypothesis testing and the paired-samples t test. The exercise also gives you practice in using COMPARE MEANS in PSPP.

## 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 a subset of the 2018 GSS. Your instructor will tell you how to access this data set which is called GSS18A.SAV.

In Exercise 6 we compared means from two independent samples.  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.

In this exercise we’re going to compare means from two dependent samples.  Dependent samples are samples in which the composition of one sample influences the composition of the other sample.  The 2018 GSS includes questions about the years of school completed by the respondent’s parents – *maeduc* and *paeduc*.  If the respondent’s mother is in one sample, then the respondent’s father must be in the other sample.  The composition of the samples is therefore dependent on each other.  PSPP calls these paired samples so we’ll use that term from now on.

Let’s start by asking whether respondents’ fathers or mothers have more years of school?  Click on ANALYZE in the menu bar and then on COMPARE MEANS and finally on MEANS.  Select the variables *maeduc* and *paeduc* and move them to the DEPENDENT LIST box.  These are the variables for which you are going to compute the mean years of school of the fathers and mothers for each respondent.  The output from PSPP will show you the mean, number of cases, and standard deviation for fathers and mothers.

Mothers and fathers have about the same amount of education with mothers having less one-tenth of a year more than fathers. But how are we going to interpret this. We could say that it’s such a small difference that mothers and fathers have about the same number of years of school completed. Or we could say that mothers have slightly more education than fathers. If we were just describing the **sample**, both these interpretations would be correct.  But what we want to do is to make inferences about fathers and mothers in the **population**.  We have a sample of fathers and a sample of mothers 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 as we did in the two previous exercises (see Exercises 5 and 6).

## Part II – Hypothesis Testing – Paired-Samples t Test

In Part I we compared the mean years of school completed by fathers and mothers.  Now we want to determine if this difference is statistically significant by carrying out the paired-samples t test. We don’t think it is significant because of the small difference but since our samples are so large, we want to be sure.

Click on ANALYZE and then on COMPARE MEANS and finally on PAIRED-SAMPLES T TEST.  Move the two variables listed above into the TEST PAIR(S) box.  Do this by selecting *maeduc* and click on the arrow to move it into the VAR 1 box.  Then you will need to click on the slider at the bottom of the TEST PAIR(S) box and move it to the right until you see VAR 2.  Now select the other variable, *paeduc*, and click on the arrow to move it into the VAR 2 box.  Finally click on OK and PSPP will carry out the paired-samples t test.  It doesn’t matter which variable you put in the VARIABLE 1 and VARIABLE 2 boxes.

You should see three boxes in the output screen. The first box gives you four pieces of information.

* Means for mothers and fathers.
* N which is the number of mothers and fathers on which the t test is based.  This includes only those cases with valid information.  In other words, cases with missing information (e.g., don’t know, no answer) are excluded.
* Standard deviations for mothers and fathers.
* Standard error of the mean for mothers and fathers which is an estimate of the amount of sampling error for the two samples.

The second box gives you the paired sample correlation which is the correlation between mother’s and father’s years of school completed for the paired samples.  If you haven’t discussed correlation yet, don’t worry about what this means.

The third box has more information in it.  With paired samples what we do is subtract the years of school completed for one parent in each pair from the years of school completed for the other parent in the same pair.  Since we put mother’s years of school completed in variable 1 and father’s education in variable 2, PSPP will subtract father’s education from mother’s education.  So if the father completed 12 years and the mother completed 10 years, we would subtract 12 from 10 which would give you -2.  For this pair the father completed two more years than the mother.

The third box gives you the following information.

* The mean difference score for all the pairs in the sample which is 0.04.  This means that mothers had an average of less than one-tenth of a year more education than the fathers.
* The standard deviation of the difference scores for all these pairs which is 3.06.
* The standard error of the mean which is an estimate of the amount of sampling error.
* The 95% confidence interval for the mean difference score.  If you haven’t talked about confidence intervals yet, just ignore this.  We’re not going to discuss this in the exercise.
* The value of t for the paired-sample t test which is 0.55.  There is a formula for computing t which your instructor may or may not want to cover in your course.
* The degrees of freedom for the t test which is 1,597. This is the number of pairs minus one or 1,598 – 1 or 1,597.  In other words, 1,597 of the difference scores are free to vary.  Once these difference scores are fixed, then the final difference score is fixed or determined.
* The two-tailed significance value which is .585 which we’ll cover next.

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 respondent’s fathers and mothers in the **sample** who answered the question.  But we want to test the hypothesis that the mean years of school completed by mothers is greater than the mean for fathers in the **population**.  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 mothers is greater than the mean years of school completed by fathers 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 mean difference score in the population is positive.  In other words, the mean years of school completed by mothers is greater than the mean years for fathers for all pairs in the population.
* null hypothesis – the mean difference score for all pairs in the population is equal to 0.

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 significance value which is 0.585.  That tells you that the probability of being wrong if you rejected the null hypothesis is almost 6 times out of ten.  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.

But wait a minute.  The PSPP output said this was a two-tailed significance value. What does that mean?  Look back at the research hypothesis which was that the mean difference score for all pairs in the population was less than 0.   We’re predicting that the mean difference score for all pairs in the population will be positive.  That’s called a one-tailed test and we have to use a one-tailed significance value.  It’s easy to get the one-tailed significance value if we know the two-tailed significance value.  If the two-tailed significance value is .585 then the one-tailed significance value is half that or .585 divided by two or .2925.  We still do not reject the null hypothesis which means that we do not have evidence to support our research hypothesis.

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

In this part of the exercise you want to compare occupational prestige scores for fathers and mothers of each respondent. These variables are called *masei10* and *pasei10*. The 10 simply refers to the year that the index was developed. High values indicate high occupational prestige and low scores indicate low occupational prestige.

Use the paired-sample t test to compare the respondents’ mothers and fathers. State the research and the null hypotheses for the t test. Do you reject or not reject the null hypotheses?  Explain why. What can you conclude from this test?

## Next Exercise

Exercise 7 will focus on One-Way Analysis of Variance.

# Exercises for an Introductory Statistics CourseEdward Nelson, California State University, Fresno

# Exercise 8One-Way Analysis of Variance

## Goal 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 COMPARE MEANS in PSPP.

## 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 a subset of the 2018 GSS. Your instructor will tell you how to access this data set which is called GSS18A.SAV.

In Exercise 6 we compared means from two independent samples while in Exercise 7 they were from paired samples.  In both cases the variable that defined the groups to be compared was a dichotomy so we were comparing two groups (e.g., males and females).  But what if we wanted to compare more than two groups?  For that we need to use a statistical test called 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 looking at the mean number of hours that people watch television broken down by highest educational degree.  Click on ANALYZE in the menu bar and then on COMPARE MEANS and finally on MEANS.  Select the variable *tvhours* and move it to the DEPENDENT LIST box.  This is the variable for which you are going to compute means.  Then select the variable *degree* and move it to the INDEPENDENT LIST box.  The output from PSPP will show you the mean, number of cases, and standard deviation for the different levels of education.

Respondents with more education watch less television than those with fewer years of 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 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 the three previous exercises (Exercises 5, 6, and 7).

## 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 PSPP 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 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 ANALYZE in the menu bar and then on COMPARE MEANS and finally on ONE-WAY ANOVA. Select the variable *tvhours* and move it to the DEPENDENT VARIABLE(S) box.  Then select the variable *degree* and move it to the FACTOR box.  Click on DESCRIPTIVES in the STATISTICS box and finally click on OK.

The output gives you the following results for the one-way analysis of variance.  We’re not going to explain these statistics in this exercise.  Your instructor will decide how much to cover on the calculation and meaning of these statistics.

* Between groups and within groups sum of squares.
* Degrees of freedom for the between groups and within groups sum of squares.
* Mean square for the between groups and within groups sum of squares.
* F statistic.
* Significance value.

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 0.000.  Since this is a rounded value it tells you that the probability of being wrong if you rejected the null hypothesis is less than .0005 or less than 5 times out of ten thousand.  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?

## Next Exercise

Exercise 9 will focus on crosstabulation as a way to describing the relationship between two variables.

# Exercises for an Introductory Statistics CourseEdward Nelson, California State University, Fresno

# Exercise 9Crosstabulation

## Goal 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 PSPP.

## Part I—Relationships between Variables

In Exercises 5 through 8 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.  The mean assumes interval or ratio level measurement (see Exercise 1).

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*),
* voting (*pres16*), and
* religiosity (*reliten*).

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 would be how people feel about abortion, fear of crime, gun control, voting and 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 PSPP 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.  By default PSPP will give you the counts as well as the row, column, and total percents.  In this case you want only the counts and column percents.  That means you will want to uncheck the boxes for the row and total percents so you won’t have unnecessary and perhaps confusing numbers in your output.

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.  Approximately 50.3% of men think abortion should be legal for any reason compared to 49.6% of women.  There’s a difference of less than 1% 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.  About 21% of men are afraid to walk alone at night in their neighborhood compared to 43% of women.  This is a difference of 22.38%.  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*),
* voting (*pres16*), and
* 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

So far we have only looked at variables two at a time.  Often we want to add other variables into the analysis.  Let’s focus on the difference between men and women (*sex*) in terms of how they feel about gun control (*gunlaw*).  First, let’s get the two-variable table which should look like this.



Men were more likely to oppose guns permits by 10.91 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 variable is named *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 *gunlaw* separately for each of these four groups.  Social class will be our control variable since we are going to hold it constant.

In order to run a table with a control variable, we need to create a blank syntax file.  To do this click on FILE in the menu bar and then on NEW and finally on SYNTAX.  A blank syntax file should open.  Enter the following commands into the syntax file.  It’s easiest to do this by copying and pasting the commands into the syntax file.

CROSSTABS
  /TABLES=gunlaw BY sex BY class
  /STATISTICS=CHISQ GAMMA
  /CELLS=COUNT COLUMN.

To run this command click on RUN in the menu bar and then on ALL.  You should see the following table in your output window.

**

This table is more complicated.  Notice that the table is actually divided into four tables with one on top of the other.  At the top we have 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: 19.37, 7.55, 12.78, and 17.37.  The percent differences vary but are all in the same direction (i.e., women are more likely to favor gun permits) and don’t vary too much from the percent difference in the two-variable table. Remember that we don’t want to make too much out of small differences because of sampling error.

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 113 respondents in the lower-class table and even fewer, 50 respondents, in the upper-class table.  We’ll have more to say about this in Exercise 10.

## 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? Was the relationship the same for all categories of the control variable?

Recall from Part IV that to run a table with a control variable, we need to create a blank syntax file.  To do this click on FILE in the menu bar and then on NEW and finally on SYNTAX.  A blank syntax file should open.  Enter the following commands into the syntax file.  It’s easiest to do this by copying and pasting the commands into the syntax file. The only difference between this and what you did in Part IV is that you have substituted *fear* for *gunlaw*.

CROSSTABS
  /TABLES=fear BY sex BY class
  /STATISTICS=CHISQ GAMMA
  /CELLS=COUNT COLUMN.

## Next Exercise

Exercise 10 will focus on Chi-Square.

# Exercises for an Introductory Statistics CourseEdward Nelson, California State University, Fresno

# Exercise 10Chi-Square

## Goal 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 PSPP.

## Part I—Relationships between Variables

We’re going to use the General Social Survey (GSS) for this exercise.  The GSS is a national probability sample of adults in the United States conducted by the National Opinion Research Center (NORC).  The GSS started in 1972 and has been an annual or biannual survey ever since. For this exercise we’re going to use a subset of the 2018 GSS. Your instructor will tell you how to access this data set which is called GSS18A.SAV.

The 2018 GSS is a sample from the population of all adults in the United States at the time the survey was done.  In Exercise 9 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 and is the appropriate test when your variables are nominal or ordinal (see Exercise 1).

In Exercise 9 we started by using crosstabulation to look at the relationship between sex and opinion about abortion.  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.  Run CROSSTABS to produce the table.  You want to get the crosstabulation of *sex* and *abany*.  Put the independent variable in the column and the dependent variable in the row.   Since your independent variable is in the column, you want to use the column percents.  Uncheck the boxes for row and total percents so PSPP will not give you unnecessary output.

## Part II – Interpreting the Percents

Your table should look like this.

**

Since your percents sum down to 100% (i.e., column percents), you want to compare the percents across.  Look at the first row.  Approximately 49.6% of men think abortion should be legal for any reason compared to 50.3% of women.  There’s a difference of 0.73 percentage points which is quite small.  We never want to make too much of small differences.  Why not?  No sample is ever a perfect representation of the population from which the sample is drawn.  This is because every sample contains some amount of sampling error.  Sampling error in inevitable.  There is always some amount of sampling error present in every sample.  The larger the sample size, the less the sampling error and the smaller the sample size, the more the sampling error.

But what is a small percent difference?  Probably you would agree that any difference less than four percentage points 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.[[13]](#footnote-13) 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.

PSPP 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.  You don’t want the row and total percents so uncheck those boxes.  Then click on the STATISTICS button and check the box for CHI-SQUARE and click on CONTINUE and then on OK.

Now you will see another output box below the crosstabulation called CHI-SQUARE TESTS.  We want the test that is called PEARSON CHI-SQUARE in the first row of the box.  Ignore all the other rows in this box.[[14]](#footnote-14)



* You should see three values to the right of Pearson Chi-Square. The value of Chi Square is 0.08.  Your instructor may or may not want to go into the computation of the Chi Square value but we’re not going to cover the computation in this exercise.
* The degrees of freedom (df) is 1.  Degrees of freedom is number of values that are free to vary.  In a table with two columns and two rows only one of the cell frequencies is free to vary assuming the marginal frequencies are fixed.  The marginal frequencies are the values in the margins of the table.  There are 689 males and 837 females in this table and there are 763 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 so there is one degree of freedom.
* The two-tailed significance value is 0.775.[[15]](#footnote-15) This tells us that there is a probability of .775 that we would be wrong if we rejected the null hypothesis.  In other words, we would be wrong 775 out of 1,000 times.  With odds like that, of course, we’re not going to reject the null hypothesis.  A common rule is to reject the null hypothesis if the significance value is less than .05 or less than five out of one hundred.  Since .775 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*),
* voting (*pres16*), and
* 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 PSPP gives you) and the cell frequencies that you would expect by chance assuming the null hypothesis was true.  PSPP will also compute these expected frequencies for you.  Rerun the crosstabulation for *sex* and *abany* remembering to ask for the column percents and Chi Square.  But this time when you click on the CELLS button to ask for the column percents look in the upper left of the dialog box where it says COUNTS which is selected as the default.  These are the observed cell frequencies.  The boxes for column, row, and total are also checked as the default.  Uncheck the boxes for row and total percents since you don’t want them.  Now click on the EXPECTED box so you will get the expected cell frequencies.

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

Chi Square assumes that all the expected cell frequencies are greater than five.  We can see from the table that this is the case for this table.  If they are just a little bit below five, that’s no problem.  But if they get down around three, then you have a problem.  What you’ll have to do is to combine rows or columns that have small marginal frequencies.

For example, run the crosstabulation of *sex* and *sibs* which is the number of brothers and sisters that the respondent has and ask for the counts, expected frequencies, and column percents.[[16]](#footnote-16) Some of the minimum expected frequencies are close to 0.  That’s because there are only a few respondents with more than 13 siblings.  You will need to recode the number of siblings into fewer categories making sure that you don’t have any categories with a really small number of cases.

## Part VI – Now it’s Your Turn Again

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

## Next Exercise

Exercise 11 will focus on measures of association.

# Exercises for an Introductory Statistics CourseEdward Nelson, California State University, Fresno

# Exercise 11Measures of Association

## Goal of Exercise

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

## Part I—Relationships between Variables

We’re going to use the General Social Survey (GSS) for this exercise.  The GSS is a national probability sample of adults in the United States conducted by the National Opinion Research Center (NORC).  The GSS started in 1972 and has been an annual or biannual survey ever since. For this exercise we’re going to use a subset of the 2018 GSS. Your instructor will tell you how to access this data set which is called GSS18A.SAV.

The 2018 GSS is a sample from the population of all adults in the United States at the time the survey was done.  In Exercise 9 we used crosstabulation and percents to describe the relationship between pairs of variables in the sample.  In Exercise 10 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 a later exercise (see Exercise 13) we’ll talk about measures for interval and ratio variables.

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

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.[[17]](#footnote-17)
* Some measures specify which variable is dependent and which is independent.  The independent variable is some variable that you think might help you explain 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 PSPP will compute. 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.[[18]](#footnote-18)
	+ Contingency Coefficient
	+ Phi and Cramer’s V
	+ Lambda
* If both of the variables are ordinal, then choose from this list.
	+ Gamma
	+ Somer’s d
	+ 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*
* region in which respondent lives – *region*
* religious preference of respondent – *relig*
* marital status – *marital*

When one or both of your variables are nominal, you have a choice among the following measures – Contingency Coefficient, Phi and Cramer’s V, and Lambda.  Let’s start with the Contingency Coefficient (C).  One of the problems of 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 (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.  If your table has two columns and two rows V is the same as the Phi Coefficient which is another measure of association so PSPP refers to it as Phi and Cramer’s V.[[19]](#footnote-19)

Lambda is a very useful measure because it has a clear and intuitive interpretation.  The value of Lambda tells you the degree to which knowing one of the variables helps you predict the other variable. A Lambda of .25 means that you can reduce the error in predicting one of the variables by 25% if you take into account the other variable.  Moreover, there are actually three versions of Lambda – one that you would use when one variable is the dependent variable, another that you would use if the other variable was dependent, and a third you would use if you don’t want to designate either of the variables as dependent.  The problem with Lambda is that it often underestimates the strength of the relationship.

Let’s look at an example to help us better understand measures of association for nominal variables.  Use CROSSTABS in PSPP to get the table for *marital* and *region*.  The first variable is the respondent’s current marital status and the second is the region of the country in which the respondent lives.  It would make sense to think of *marital* as the independent variable since respondents’ marital status might influence where they currently live.  Remember to put the dependent variable in the row and the independent variable in the column.  Ask PSPP to compute the column percents, Chi Square and the three measures we just discussed.

Notice that C and V are fairly low.  C is 0.16 and V is 0.08.  Ignore Phi since Phi is only used for a table with two columns and two rows.  You can see that both measures tells us that the relationship is fairly weak.

Since we said that it was reasonable to think of where the respondent currently lives as the dependent variable, the appropriate value for lambda is .01 meaning that knowing the respondent’s marital status doesn’t help you at all when predicting where they currently life.

## Part V – Now it’s Your Turn

Use CROSSTABS 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 helps us predict 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 PSPP to compute Chi Square and the three measures of association we discussed.  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*
* trust of other people – *trust*

You have a choice from four measures that PSPP 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 to get the crosstabulation of *degree* and *satfin*.  If we think that education influences how satisfied respondents are with their financial situation, then education would be our independent variable.  Put *degree* in the column and *satfin* in the row and run the table. 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 also -14.  Gamma (-0.21) is larger.  Gamma will usually 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 we would want the symmetric value of Somer’s d which equals 0.50.  Tab-b is 0.50 and tau-c is .45.  Gamma as usually larger (0.66).  The relationship between these two variables is clearly much 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 use them to determine if it is positive or negative.

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

Use CROSSTABS to give you a table for *degree* and *trust*.  We want to find out if the respondent’s education helps us understand why some say they trust people and other respondents feel they can’t trust others.   Decide which variable is independent and dependent.  Get the correct percents and tell PSPP 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.  Compare Gamma values to Gamma values and Lambda values to Lambda values.  Rerun one of the tables that you created in Parts 5 and 7 but this time hold sex constant.  In other words, sex would be your control variable.

In order to run a table with a control variable, we need to create a blank syntax file.  To do this click on FILE in the menu bar and then on NEW and finally on SYNTAX.  A blank syntax file should open.  Enter the following commands into the syntax file.  It’s easiest to do this by copying and pasting the commands into the syntax file.

CROSSTABS
  /TABLES=dependent variable BY independent variable BY sex
  /STATISTICS=CHISQ BTAU CTAU GAMMA D
  /CELLS=COUNT COLUMN.

Be sure to replace “dependent variable” and “independent variable” with the names of your variables.  To run this command click on RUN in the menu bar and then click on ALL.  Your table should appear in the output window.

Now compare the appropriate measure of association to determine if the relationship is stronger for male or females or whether it doesn’t vary by sex.  Remember not to make too much out of small differences in the measures.

## Next Exercise

Exercise 12 will focus on spuriousness.

# Exercises for an Introductory Statistics CourseEdward Nelson, California State University, Fresno

# Exercise 12Spuriousness

## Goal of Exercise

The goal of this exercise is to explore the concept of spuriousness.  We will consider the relationship of religiosity and control of the distribution of pornography and test for the possibility that this relationship is spurious due to sex.   The exercise also gives you practice in using CROSSTABS to explore the relationships among variables and 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.  For this exercise we’re going to use a subset of the 2018 GSS survey. Your instructor will tell you how to access this data set which is called GSS18A.SAV.

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 called *reliten* in the data set.

We’re going to recode *reliten* for this exercise.  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.  Let’s call the recoded variable *reliten1*.  To make your output more readable, we’ll add value labels for this variable.  The label for value 1 will be strong and for value 2 not strong.

After recoding *reliten1*, we’ll run FREQUENCIES for our unrecoded variable (*reliten*) and our recoded variable (*reliten1*) and compare the two frequency distributions to make sure we didn’t make an error recoding.  If we made a mistake, then we’ll need to do the recoding again. Since you may not know how to recode in PSPP, we have done it for you and the recoded variable *reliten1* is part of the data set.

We’ll start by developing a hypothesis.  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 respondent’s opinion about pornography laws? Why should more religious individuals be more likely to think that pornography should be illegal for everyone? You might argue that the moral codes of more religious individuals tend to be absolute while the moral codes of those who are less religious tend to be more situational.

Once you have developed your argument, then you would construct a dummy table showing what the relationship between *reliten1* and *pornlaw* should look like if your hypothesis is true.  It’s customary to put the independent variable in the column and the dependent variable in the row.  Let’s 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?  We’ll 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 those two cells. The figure below shows what your dummy table should look like.



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.  Be sure to put the independent variable (*reliten1*) in the column and the dependent variable (*pornlaw*) in the row.  You also need to ask for the percents, Chi-Square, and an appropriate measure of association.  Since the independent variable is the column variable, you will want the column percents.

All that is left is to interpret the table.  Since the independent variable is the column variable, we computed the column percents.  It’s important to compare the percents straight 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 an appropriate measure of association 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 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 control of pornography.  For example, sex may be related to both these variables.  Research has shown that women are more religious than men. They are more likely to attend worship services, more likely to pray, and more likely to say that they are strongly religious than men. Women may also be more likely to want stricter controls over pornography. Why do you think that might be? 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 to get two tables – one table should cross tabulate *sex* and *pornlaw* and the other table should cross tabulate *sex* and *reliten1*.  Be sure to get the percents, Chi-Square, and an appropriate measure of association.  If sex is related to both variables, then we need to check further to see if the original relationship between religiosity and control of pornography 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.  Sex is referred to as the control variable.

In order to run a table with a control variable, we need to create a blank syntax file.  To do this click on FILE in the menu bar and then on NEW and finally on SYNTAX. A blank syntax file should open.  Enter the following commands into the syntax file.  It’s easiest to do this by copying and pasting the commands into the syntax file.

CROSSTABS
  /TABLES=pornlaw BY reliten1 BY sex
  /STATISTICS=CHISQ BTAU CTAU GAMMA D
  /CELLS=COUNT COLUMN.

Notice the formal of the “TABLES” subcommand.  It lists the table you want to run in the following order – dependent variable BY independent variable BY control variable.  To run the CROSSTABS command, click on RUN in the menu bar and then on ALL.

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 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 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 they decrease substantially?  Another clue is your measure of association.   Did the measures of association stay about the same for males and for females or did they decrease substantially from that in the original two-variable table? Remember to not make too much of small differences.

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—Now It’s Your Turn

Now that we have seen that the relationship between *reliten1* and *pornlaw* is not spurious due to sex, let’s see if it might be spurious due to some other variable such as age. The variable *age1* is a recode of *age*. Age has been recoded into three categories – 35 or younger, 36 to 54, and 55 and over. Repeat the analysis carried out in Part III to determine if the relationship is spurious due to age. Here’s a brief summary of what you should do.

* Crosstabulate *reliten1* and *pornlaw* to determine what the relationship looked like before you introduced *age1* as a control variable.
* Crosstabulate *age1* with both *reliten1* and *pornlaw* to see if the control variable is related to both your independent variable and your dependent variable. A relationship can only be spurious if the control variable is related to both the independent and dependent variables.
* Crosstabulate *reliten1* and *pornlaw* controlling for *age1* to see what effect controlling for *age1* was on the original two-variable relationship. Use the syntax from Part III substituting *age1* for *sex*.
* Summarize what you learned.
	+ What was the original two-variable relationship between *reliten1* and *pornlaw*?
	+ What happened when you introduced *age1* into the analysis as a control variable.
	+ Was the relationship spurious or not? How did you decide?
	+ What does it mean to say a relationship is spurious?

## Next Exercise

Exercise 13 will focus on correlation.

# Exercises for an Introductory Statistics CourseEdward Nelson, California State University, Fresno

# Exercise 13Correlation

## Goal of Exercise

The goal of this exercise is to introduce measures of correlation. The exercise also gives you practice using GRAPHS, BIVARIATE CORRELATION, and COMPARE MEANS in PSPP.

## Part I – Correlation

We’re going to use the General Social Survey (GSS) for this exercise. The GSS is a national probability sample of adults in the United States conducted by the National Opinion Research Center (NORC). The GSS started in 1972 and has been an annual or biannual survey ever since. For this exercise we’re going to use a subset of the 2018 GSS. Your instructor will tell you how to access this data set which is called GSS18A.SAV.

In Exercise 11 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.

Before we look at these measures let’s talk about a type of graph that is used to display the relationship between two variables called a scatterplot.  Click on GRAPHS in the menu bar and then on SCATTERPLOT.  So all our scatterplots will look the same let’s put *maeduc* on the X-axis and *paeduc* on the Y-Axis.  Click on OK and PSPP will display your graph.

Now look for the general pattern to the scatterplot. You see more cases in the upper right and lower left of the plot and fewer cases in the upper left and lower right. In general, as one of the variables increases, the other variable tends to increase as well. Moreover, you can imagine drawing a straight line that represents this relationship. The line would start in the lower left and continue towards the upper right of the plot. That’s what we call a positive linear relationship.[[20]](#footnote-20)  But how strong is the relationship and where exactly would you draw the straight line? The Pearson Correlation Coefficient will tell us the strength of the linear relationship and linear regression will show us the straight line that best fits the data points. We’ll talk about the Pearson Correlation Coefficient in part 3 of this exercise and linear regression in Exercise 14.

## Part II – Now it’s Your Turn

Use GRAPHS to create the scatterplot for years of school completed by the respondent (*educ*) and spouse’s years of school completed (*speduc*). So all our plots will look the same, put *speduc* on the X-Axis and *educ* on the Y-Axis. Look at your scatterplot and decide if the scatterplot has a pattern to it. What is that pattern? Do you think it is a linear relationship? Is it a positive linear or a negative linear relationship?

## Part III - 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 Coefficient for *maeduc* and *paeduc*. Click on ANALYZE in the menu bar and then click on BIVARIATE CORRELATION. Bivariate just means that you want to compute a correlation for two variables. Move these two variables into the VARIABLE(S) box. Notice that the circle for TWO-TAILED is filled in for TEST OF SIGNIFICANCE. A two-tailed significance test is used when you don’t make any prediction as to whether the relationship is positive or linear. In our case, we would expect that the relationship to be positive so we would want to use a one-tailed test. Click on the circle for one-tailed to change the selection. Check the box labelled FLAG SIGNIFICANT CORRELATIONS so PSPP will tell you when a relationship is statistically significant. Now click OK and PSPP will display your correlation coefficient.

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 11) meaning that we don’t designate one of the variables as the dependent variable and the other as the independent variable. Notice that the Pearson r is statistically significant using a one-tailed test at the .01 level of significance. A Pearson r of 0.71 is really pretty large. 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 14 fits the data points reasonably well.

## Part IV – Now it’s Your Turn Again

Use BIVARIATE CORRELATION 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 V – 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 1 through 4 as a set. That means that we want to see the values for r for each pair of variables. This time move all four of the variables into the VARIABLE(S) box (i.e., *educ*, *maeduc*, *paeduc*, and *speduc*) and click on OK. 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 VI – The Correlation Ratio or 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 8. Click on ANALYZE in the menu bar and then on COMPARE MEAN and finally on ONE-WAY ANOVA. Select the variable *tvhours* and move it to the DEPENDENT LIST box. This is the variable for which you are going to compute means. Then select the variable *degree* and move it to the FACTORS box. In this example, the independent variable is ordinal and the dependent variable is ratio. Notice that we’re using our independent variable to predict our dependent variable. Now click on OK.

The F test in one-way analysis of variance tells us to reject the null hypothesis that all the population means are 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.  Eta-Squared is similar to the squared Pearson Correlation Coefficient (i.e., the Coefficient of Determination).  PSPP doesn’t compute Eta-Squared for us but it gives us the information we need to compute it ourselves.  Eta-Squared is equal to the between groups sum of squares divided by the total sum of squares.  For this problem that would be 488.63 divided by 10530.04 which equals .046.  This tells us that 5.1% 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 VII – Your Turn Again

In Exercise 8 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 that 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?

## Next Exercise

Exercise 14 will focus on bivariate (two-variable) regression.

# Exercises for an Introductory Statistics CourseEdward Nelson, California State University, Fresno

# Exercise 14Bivariate Linear Regression

## Goal of Exercise

The goal of this exercise is to introduce bivariate linear regression. The exercise also gives you practice using LINEAR REGRESSION, FREQUENCIES, and SELECT CASES in PSPP.

## Part 1 – 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 a subset of the 2018 GSS. Your instructor will tell you how to access this data set which is called GSS18A.SAV.

In Exercise 13 we considered 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 to get a frequency distribution for the variable *tvhours* which is the number of hours that a respondent watches television per day. Look in the STATISTICS box and check the boxes for SKEWNESS and KURTOSIS. Notice that there are only a few people who watch 13 or more hours of television per day. There are even some who say they watch 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 out only those cases for which tv1\_tvhours is less than or equal to 12. That way the outliners will be excluded from the analysis. To do this you will have to create a PSPP syntax file and then execute the command.  Click on FILE in the menu bar and then on NEW and then on SYNTAX.  This will open a blank syntax file.  In the syntax file enter the following command.  You can do this by cutting and pasting this command into the PSPP syntax file.

SELECT IF tvhours <= 12.

Once you have done this click on RUN in the menu bar and then click on ALL.  Note that you have removed these cases from your data file for this exercise.  So when you complete this exercise do **NOT** save the data file because you will want to use the entire data set for future exercises.  To see your output click on the PSPP icon at the bottom of your screen (i.e., looks like a red circle with a blue cutout at the top).  This will open the output window where you will see your results.

Now use FREQUENCIES to get a frequency distribution for *tvhours*. Remember to ask for SKEWNESS and KURTOSIS (see Exercise 4) by checking these boxes in the STATISTICS box.

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.

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.[[21]](#footnote-21)

Let’s think 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.[[22]](#footnote-22)

## 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 PSPP to carry out the calculations by running the REGRESSION command.

Click on ANALYZE in the menu bar of PSPP and then click on REGRESSION which will open another dropdown menu. Click on LINEAR in the menu. Your dependent variable will be *tvhours*. Enter *age* as your independent variable and click on OK.

You should see three output boxes.

* The first box tells you the value of the Pearson Correlation Coefficient (R and the correlation coefficient squared (R2). Age explains or accounts for 5.0% 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 square. But in Exercise 15 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 second box is the analysis of variance table that tests the null hypothesis that the Pearson Correlation Coefficient Squared in the population is 0. In this example we reject the null hypothesis since the significance value is less than .05 (or whatever level of significance you’re using which is usually .05 or .01 or .001). This means that age explains more than 0 percent of the variation.
* The third box gives you the regression coefficients.
	+ The slope of the line (b) is equal to 0.03.
	+ The point at which the regression line crosses the Y-Axis is 1.25. 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 tests which test 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 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 .03 units (i.e., number of hours watching television). So our regression equation is Y = 1.25 + .03X. Thus for a person that is 20 years old, the predicted number of hours that he or she watches television 1.25 + (.03) (20) or 1.25 + 0.6 or 1.85 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 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.

* 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?

## Next Exercise

Exercise 15 will focus on multivariate linear regression.

# Exercises for an Introductory Statistics CourseEdward Nelson, California State University, Fresno

# Exercise 15Multivariate Linear Regression

## Goal of Exercise

The goal of this exercise is to introduce multivariate (i.e., more than two variables) linear regression.  The exercise also gives you practice using LINEAR REGRESSION, FREQUENCIES, SELECT CASES, and BIVARIATE CORRELATION in PSPP.

## 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 a subset of the 2018 GSS. Your instructor will tell you how to access this data set which is called GSS18A.SAV.

In Exercise 14 we considered 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 the last exercise 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 want 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 who had more education tended to watch less television.

Let’s start by using FREQUENCIES to get the frequency distribution for *tvhours*.  In the previous exercise 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.

To remove these outliers you will have to create a PSPP syntax file and then run the file.  Click on FILE in the menu bar and then on NEW and then on SYNTAX.  This will open a blank syntax file.  In the syntax file enter the following command.  You can do this by cutting and pasting this command into the PSPP syntax file.

SELECT IF tvhours <= 12.

Once you have done this click on RUN in the menu bar and then click on ALL.  Note that you have removed these cases from your data file for this exercise.  So when you complete this exercise do **NOT** save the data file because you will want to use the entire data set for other exercises.

To see your output click on the PSPP icon at the bottom of your screen (i.e., looks like a red circle with a blue cutout at the top).  This will open the output window where you will see your results.

Now use FREQUENCIES again to get the frequency distribution for *tvhours* and make sure that you correctly removed the outliers.  You should not see any cases with more than 12 hours.

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.[[23]](#footnote-23)

## 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 PSPP to carry out the calculations.

Click on ANALYZE in the menu bar and then click on REGRESSION which will open another dropdown menu.  Click on LINEAR in the menu.  Your dependent variable will be *tvhours*.  In the previous exercise 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* as your independent variables and click on OK.

You should see four output boxes.

* The first box lists the variables you entered and reminds you which is your dependent variable.
* The second box tells you the value of the Pearson Multiple Correlation Coefficient (R) and the Squared Multiple Correlation (R2) which is usually referred to as the Coefficient of Determination.  How does this differ from the Pearson Correlation Coefficient (r) and r squared?  The Multiple R squared tells us that *age* and *educ* together explain or account for 8.52% of the total variation in the number of hours per day that respondents watch television.[[24]](#footnote-24)  In the previous exercise we saw that *age* by itself explained 5.29% (.23 squared) of the variation in tvhours and that *educ* by itself explained 3.61% (i.e., .19 squared).  Why can’t we just add 5.29% and 3.61% and say that 8.9% of the total variation is explained by these two variables together.  It’s because the variation explained by these two independent variables overlap and because of this overlap they might account for less of the variation in the dependent variable.
* The second box also gives us the Adjusted R squared which “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 Publishers, p. 293)  The standard error is an estimate of the amount of sampling error for this statistic.
* The third box is the analysis of variance table that tests the null hypothesis that the Squared Multiple R **in the population** is 0.  In this example we reject the null hypothesis since the significance value is less than .05 (or whatever level of significance you’re using which is usually .05 or .01 or .001).  This means that age and respondent’s education together explain more than 0 percent of the variation.
* Recall that the equation for predicting Y is a + b1X1 + b2X2 where X1 is age and X2 is respondent’s education.  The fourth box gives you more information.
	+ The constant (a) is 3.04.
	+ The unstandardized multiple regression coefficient (b1) for *age* is .03.  This means that an increase of one unit in *age* results in an average increase of .03 units in *tvhours* after statistically adjusting for *educ*.  Or, to put this in more easily understood terms, an increase of one year in the respondent’s age results in an average increase of .03 hours of television viewing after statistically adjusting for father’s education.
	+ The unstandardized multiple regression coefficient (b2) for *educ* is -.13.  This means that an increase of one unit in *educ* results in an average decrease of -.13 units in *tvhours* after statistically adjusting for *age*.  Or, to put it another way, an increase of one year in respondents’ education results in an average decrease of .13 hours of television viewing after statistically adjusting for the respondent’s age.
	+ So what does it mean to statistically adjust for something?  Suffice it to say that it means that b1 tells us the effect 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 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 -.18 and for *age* is .22 which means that age is slightly more important in predicting hours of television viewing than is respondent’s education.
	+ The t test which tests the null hypotheses that the **population** constant and **population** multiple regression coefficients are equal to 0.
	+ The significance value for each test.  As you can see, in this example we reject all three null hypotheses.  However, we’re usually only interested in the t test for the population multiple regression coefficients.

So our multiple regression equation for predicting Y is 3.04 + .03X1 - .13X2 where X1 is age and X2 is education. 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 3.04 + (.03) (20) - .13 (12) or 3.04 + 0.6 – 1.56 or 2.08 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 *sibs* to your list of independent variables.  Now you will have three independent variables – *age*, *educ*, and *sibs*.  The variable *sibs* is the number of siblings the respondent has.

* 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? Pay particular attention to the Beta value for *sibs*.
* 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? Pay particular attention t the t test for *sibs*.

## 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.  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. Use BIVARIATE CORRELATION to get the bivariate Pearson Correlation matrix for the three independent variables you used in Part 3 – *age*, *educ*, and *sibs*.  If any of these correlations is really high, then you would have a problem but in this example, that clearly isn’t the case.  There are other ways to detect multicollinearity but for this exercise we’ll stop here.

## Next Exercise

Exercise 16 will focus on dummy variable multiple regression.

# Exercises for an Introductory Statistics CourseEdward Nelson, California State University, Fresno

# Exercise 16Dummy Variable Multiple Regression

## Goal of Exercise

The goal of this exercise is to introduce dummy variable regression.  The exercise also gives you practice using LINEAR REGRESSION, FREQUENCIES, SELECT CASES, and COMPUTE in PSPP.

## 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 a subset of the 2018 GSS. Your instructor will tell you how to access this data set which is called GSS18A.SAV.

In Exercise 14 we considered linear regression for one independent and one dependent variable which is often referred to as bivariate linear regression.  Multiple linear regression (see Exercise 15) 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 respondent’s education.  The variable *degree1* has two categories – 1 for those without a college degree and 2 for those with a college degree.  What we do is to create two dummy variables – one for those with a college degree and the other for those without a college degree.  Here’s how we do it:

* *without\_degree* = 1 if *degree1* = 1 and 0 if *degree1* = 2, and
* *with\_degree* = 1 if *degree1* = 2 and 0 if *degree1* = 1.

If there are k categories, then you use k – 1 of the dummy variables in your regression analysis.  The category that you omit becomes your comparison group.  We’re going to enter *with\_degree* into the analysis and omit *without\_degree*.  That means that those without a college degree 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 would create nine dummy variables and omit one of them.  Actually, you wouldn’t need to create all nine dummy variables since you’re going to omit one.  If we decide to omit the category for the Pacific region (value 9), then you would create eight dummy variables, one for each of the other categories, and the Pacific region would be our comparison group.

Neither of these two variables – *tvhours* and *degree1* – have missing data but if the variable for which you are creating dummy variables has missing data you need to be careful to exclude those cases with missing data from the analysis.  You want to be careful not to include them in one of your dummy variables.

Let’s use *tvhours* as our dependent variable as we did in the previous two exercises (see Exercises 14 and 15).  Run FREQUENCIES to get the frequency distribution for *tvhours*.  In the previous exercises we discussed outliers and noted that there are a few individuals 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.

To remove these outliers you will have to create a PSPP syntax file and then execute the file.  Click on FILE in the menu bar and then on NEW and then on SYNTAX.  This will open a blank syntax file.  In the syntax file enter the following command.  You can do this by cutting and pasting this command into the PSPP syntax file.

SELECT IF tvhours <= 13.

Once you have done this click on RUN in the menu bar and then click on ALL.  Note that you have removed these cases from your data file for this exercise.  So when you complete this exercise do **NOT** save the data file because you will want to use the entire data set for future exercises.

To see your output click on the PSPP icon at the bottom of your screen (i.e., looks like a red circle with a blue cutout at the top).  This will open the output window where you will see your results.

Now use FREQUENCIES again to get the frequency distribution for *tvhours* and make sure that you correctly removed the outliers.  You should not see any cases with more than 12 hours.

To create the dummy variable for those with a degree (*with\_degree*), click on TRANSFORM in the menu bar for PSPP and then click on COMPUTE VARIABLE.  Enter the variable name, *with\_degree*, in the target variable box and enter 0 in the NUMERIC EXPRESSION box.  Then click on OK.  This will assign the value 0 to all cases.

Now you want to change the value for *with\_degree* to 1 for all with a degree in the sample.  To this you will have to create a PSPP syntax file and then execute the file.  Click on FILE in the menu bar and then on NEW and then on SYNTAX.  This will open a blank syntax file.  In the syntax file enter the following command.  Remember that you can do this by cutting and pasting this command into the PSPP syntax file.

IF (degree1 = 2) with\_degree=1.

Once you have done this click on RUN in the menu bar and then click on ALL.

## Part II – Regression with Dummy Variables

Click on ANALYZE in the menu bar of PSPP and then click on REGRESSION which will open another dropdown menu.  Click on LINEAR in the menu.  Your dependent variable will be *tvhours*.  Enter the dummy variable for those with a college degree (*with\_degree*) as your independent variable.  Remember that you are omitting the dummy variable for those without a degree (*without\_degree*) so this becomes your comparison group.

Let’s look at the output box that contains your unstandardized regression coefficient.  From this you can see that your regression equation for predicting *tvhours* is 2,94 – 1.04X where X is your dummy variable.  Remember that your dummy variable, *with\_degree*, equals 1 if the person has a degree and 0 if the person does not have a degree.  So for those with a degree the predicted number of hours watching television is 2.94 – 1.04 (1) or 1.90.  For those without a degree the predicted number of hours is 2.94 – 1.04 (0) or 2.94.  Since we left the dummy variable for without a degree (*without\_degree*) out of the regression equation, those without a degree become our comparison group.  The unstandardized regression coefficient (1.04) is the mean number of hours that those with a degree watch television (1.90) minus the mean for those without a degree (2.94) which is -1.04.

PSPP 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 test value is -8.66 and the significance value is .000.  This means that we reject the null hypotheses.  In other words, we can conclude that the population slope is significantly different from zero.  The Pearson Correlation Coefficient Squared (Coefficient of Determination) tells us that the dummy variable for sex explains approximately 5% of the variation in the dependent variable.

## Part III – Now it’s Your Turn

Use PSPP 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:

* *race\_whites* – equals 1 if the person is white and 0 if the person is non-white, and
* *race\_nonwhites* – equals 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 *race\_nonwhites* out of the regression equation.  Use COMPUTE to create the *race\_whites* dummy variable following the procedures we used in Part 2.

Now run the regression analysis with *tvhours* as your dependent variable and *race\_whites* as your independent variable.

* Write out the regression equation.
* What does the unstandardized multiple regression coefficient 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

Let’s add more variables into our analysis. This time we’re going to include two independent variables – *age* and *with\_degree*..  Use PSPP to carry out the regression analysis for this model.

The regression equation for predicting *tvhours* is 1.50 + 0.03 (*age*) - 1.11 (*with\_degree*). 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 and those with a college degree watch less television that without a degree. The t tests show that all the unstandardized coefficients are statistically significant meaning that that we can reject the null hypotheses that they are zero in the population.  The Pearson Multiple Correlation Coefficient Squared (Coefficient of Determination) tells us that together the independent variables explain or account for 11% of the variation in television viewing.  The F test in the analysis of variance table is also statistically significant meaning that we can reject the null hypothesis that our independent variables explain none of the variation in the dependent variable.  The Beta values show the relative importance of the independent variables in predicting television viewing and tell us that both *age* and *without\_degree* are equally in importance.

##  Part V – Now it’s Your Turn Again

Repeat the regression analysis you did in Part 4 but this time add *race\_whites* into the analysis, This means you will have three independent variables – *age*, *with\_degree*, and *race\_whites*.

* 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?

# **Appendix A**

# **Codebook for the Subset of the 2018 General Social Survey**

The General Social Survey (GSS) is a large, national probability sample of adults in the United States. It began in 1972 and continued on an almost yearly basis until 1996. In 1996, the GSS became a biannual survey and the sample size increased. Many questions are asked on each survey, while other questions are rotated from survey to survey. This subset from the 2018 GSS includes all the cases (2,348) and 99 variables. This data set has already been weighted using the weight variable supplied by the GSS (*wtss*). Some of the original GSS variables were recoded and a few new variables created. Some of the new variables have names similar to those in the original GSS data set. The data set was created to accompany these exercises and is named GSS18A.SAV.

**Variable** **Description of Variable**

ABANY Abortion if woman wants for any reason

ABDEFECT Abortion if strong chance of serious defect

ABHLTH Abortion if woman's health seriously endangered

ABNOMORE Abortion if married and wants no more children

ABPOOR Abortion if low income and can't afford more children

ABRAPE Abortion if pregnant as result of rape

ABSINGLE Abortion if not married

ADULTS Household members 18 years and older

AGE Age of respondent

AGE1 Age recoded into three categories

AGED Should aged live with their children?

AGEKDBRN Respondent's age when first child born

ATTEND How often respondent attends religious services

BIBLE Feelings about the bible

BLACK Is respondent African American?

CAPPUN Favor or oppose death penalty for murder

CHILDS Number of children

CHLDIDEL Ideal number of children

CLASS Subjective class identification

CLASS1 Recoded from subjective class identification [CLASS]

COHORT Cohort

COLATH Allow anti‑religionist to teach

COLCOM Allow communist to teach

COLHOMO Allow homosexual to teach

COLMIL Allow militarist to teach

COLMSLM Allow anti-American Muslim Clergyman to teach in college

COLRAC Allow racist to teach

DEGREE Respondent's highest degree

DEGREE1 Recoded from R’s highest degree [DEGREE] – does respondent have a four-year college or postgraduate degree

DENOM Specific Protestant denomination

EDUC Highest year of school completed

ETHNICITY Respondent’s race/ethnicity[[25]](#footnote-25)

FAIR People fair or try to take advantage

FEAR Afraid to walk at night in neighborhood

FINRELA Opinion of family income

FUND Fundamentalism of respondent’s religion

GRASS Should marijuana be made legal?

GUNLAW Favor or oppose gun permits

HAPMAR Happiness of marriage

HAPPY General happiness

HEALTH Condition of health

HOMPOP Number of persons in household

HRS1 Number of hours respondent worked last week

HRS2 Number of hours respondent usually works a week

ID Respondent’s identification (id) number

INCOME16 Total family income (2017)

INCOME16.1 Recoded from total family income (2017) [INCOME16]

LATINO Is respondent Latino?

LIBATH Allow anti‑religious book in library

LIBCOM Allow communist's book in library

LIBHOMO Allow homosexual's book in library

LIBMIL Allow militarist's book in library

LIBMSLM Allow anti-American Muslim clergyman's book in library

LIBRAC Allow racist's book in library

MADEG Mother's highest degree

MAEDUC Highest year school completed, mother

MARITAL Marital status

MASEI10 Mother’s socioeconomic status using scale developed in 2010

PADEG Father's highest degree

PAEDUC Highest year school completed, father

PARTYID Political Party Affiliation

PARTYID1 Recoded from Political Party Identification [PARTYID]

PASEI10 Father’s socioeconomic status using scale developed in 2010

POLVIEWS Think of self as liberal or conservative

POLVIEWS1 Recoded from Think of self as liberal or conservative [POLVIEWS]

PORNLAW Feelings about pornography laws

POSTLIFE Belief in life after death

PRAY How often does respondent pray?

PRAYER Support Supreme Court Decision on prayer in public schools

PRES12 Vote for Romney or Obama in 2012

PRES16 Vote for Clinton or Trump in 2016

RACE Race of respondent

REGION Region of interview

RELIG Respondent's religious preference

RELIG1 More detailed breakdown of religious preference

RELITEN Strength of religious affiliation

RELITEN1 Strength of religious affiliation recoded into two categories

RELPERSN Respondent considers self a religious person

SATFIN Satisfaction with financial situation

SEI10 Respondent’s socioeconomic status using scale developed in 2010

SEX Respondent's sex

SIBS Number of brothers and sisters

SIZE Size of place respondent lives in thousands

SPDEG Spouse's highest degree

SPEDUC Highest year school completed, spouse

SPKATH Allow anti‑religionist to speak

SPKCOM Allow communist to speak

SPKHOMO Allow homosexual to speak

SPKMIL Allow militarist to speak

SPKMSLM Allow anti-American Muslim clergyman to speak

SPKRAC Allow racist to speak

SPSEI10 Spouse’s socioeconomic status using scale developed in 2010

TRUST Can people be trusted?

TVHOURS Hours per day watching television

VOTE12 Did respondent vote in 2012?

VOTE16 Did respondent vote in 2016?

WTSS Weight variable for GSS18 (data subset already weighted by the variable WTSS)

YEAR Year of survey (2018 for all respondents)

ZODIAC Respondent's astrological sign

1. Stanley Smith Stevens, 1946, “On the Theory of Scales of Measurement,” Science 103 (2684), pp. 677-680. See also Dan Osherson and David M. Lane, Levels of Measurement, <http://onlinestatbook.com/2/introduction/levels_of_measurement.html> [↑](#footnote-ref-1)
2. Frequency distributions are described in more detail in Exercise 2. Ignore the statistics that PSPP prints out and focus on the percentages displayed.  We’re not ready to discuss the use of the other statistics.  We’ll do that in later exercises. [↑](#footnote-ref-2)
3. You might wonder why we didn’t use an example from the GSS.  There aren’t any. All the variables in the GSS are nominal, ordinal, or ratio.   [↑](#footnote-ref-3)
4. 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 in the 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-4)
5. In a grouped frequency distribution the median would be the category that contains the middle value. [↑](#footnote-ref-5)
6. See Exercise 4 for a more through discussion of skewness.  [↑](#footnote-ref-6)
7. The Index of Qualitative Variation can be used to measure variation for nominal variables. [↑](#footnote-ref-7)
8. Actually the highest age is larger than 89.  The GSS combines all respondents that are 89 or older into one category which is given the value of 89.  But for our purposes that is close enough.  [↑](#footnote-ref-8)
9. The GSS is itself not a simple random sample but rather is an example of a multistate cluster sample. [↑](#footnote-ref-9)
10. The null hypothesis is often called the hypothesis of no difference.  We’re saying that the population mean is still equal to 12.  In other words, nothing has changed.  There is no difference. [↑](#footnote-ref-10)
11. Missing cases would include those who said they didn’t know or refused to answer the question. [↑](#footnote-ref-11)
12. Why probably? Because by chance you could get a much higher or lower mean which will produce a larger t value and could mean that your significance would be low enough to reject the null hypothesis. [↑](#footnote-ref-12)
13. 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-13)
14. Unfortunately there is no way to tell PSPP to just give us the “Pearson Chi-Square.” [↑](#footnote-ref-14)
15. What do we mean by two-tailed? We’re not predicting the direction of the relationship. We’re not predicting that men are more likely to think abortion should be legal or that women are more likely.  So it’s a two-tailed test. [↑](#footnote-ref-15)
16. 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-16)
17. 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-17)
18. PSPP will also compute the Uncertainty Coefficient but we’re not going to consider this measure. [↑](#footnote-ref-18)
19. In the dialog box where you tell PSPP which measures of association to compute, it is referred to as Phi.  But in the output, both the values of Phi and Cramer’s V are displayed.  Phi is only used in a table with two rows and two columns.  In such a table, Phi and Cramer’s V are the same.  In a table with more than two rows or two columns, Phi should never be used. [↑](#footnote-ref-19)
20. This assumes that the variables are coded low to high (or high to low) on both the X-Axis and the Y-Axis. [↑](#footnote-ref-20)
21. 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-21)
22. When you square a value the result is always a positive number. [↑](#footnote-ref-22)
23. When you square a value the result is always a positive number.  [↑](#footnote-ref-23)
24. PSPP rounds this to the nearest second decimal point. It’s actually .0852 which rounds to .09. To get .0852 divide the between sum of squares (624.30) by the total sum of squares (7323.35). [↑](#footnote-ref-24)
25. This variable was created by combining responses to a question asking the respondent’s race (coded as White, Black, and Other), and another question asking whether the respondent is Hispanic. Any respondent identifying as Hispanic was so classified, regardless of race. [↑](#footnote-ref-25)