**STAT5S\_pspp: Exercise Using PSPP to Explore Hypothesis Testing – One-Sample t Test**

Author:   Ed Nelson
Department of Sociology M/S SS97
California State University, Fresno
Fresno, CA 93740
Email:  **ednelson@csufresno.edu**

**Note to the Instructor:** The data set used in this exercise is gss14\_subset\_for\_classes\_STATISTICS\_pspp.sav which is a subset of the 2014 General Social Survey. Some of the variables in the GSS have been recoded to make them easier to use and some new variables have been created.  The data have been weighted according to the instructions from the National Opinion Research Center.  This exercise uses COMPARE MEANS (one-sample t test) and SELECT CASES in PSPP to explore hypothesis testing and the one-sample t test.  I prepared two documents to help you with PSPP – “[**Notes on Using PSPP**](http://www.ssric.org/files/notes_on_using_PSPP.docx)” and “[**Differences between PSPP and SPSPS**](http://www.ssric.org/files/differences_between_PSPP_and_SPSS.docx)” which should answer many of your questions about PSPP. You have permission to use this exercise and to revise it to fit your needs.  Please send a copy of any revision to the author. Included with this exercise (as separate files) are more detailed notes to the instructors and the PSPP syntax necessary to carry out the exercise. These, of course, will need to be removed as you prepare the exercise for your students.  Please contact the author for additional information.

I’m attaching the following files.

* [**Data subset**](http://ssric.org/files/gss14_subset_for_classes_STATISTICS_pspp.sav) (.sav format)
* [**Extended notes for instructors**](http://ssric.org/files/Extended_Notes_for_Instructors_for_STAT5S_pspp.docx) (MS Word; docx format).
* [**PSPP syntax file**](http://ssric.org/files/PSPP_Syntax_for_STAT5S_pspp.sps) (.sps format)
* [**This page**](http://ssric.org/files/STAT5S_pspp.docx) (MS Word; docx format).

**Goals 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 2014 GSS. Your instructor will tell you how to access this data set which is called gss14\_subset\_for\_classes\_STATISTICS\_pspp.sav.  It’s a large sample of about 2,500 individuals.  To illustrate simple random sampling, we’re going to select a simple random sample of 30% of all the individuals in the GSS.**[[1]](http://ssric.org/node/564/edit%22%20%5Cl%20%22_edn1%22%20%5Co%20%22)**

Start by getting a frequency distribution for the variable d4\_educ which is the last year of school completed by the respondent.  PSPP will list the variables and you will select those variables you want to use.  PSPP lists the variables using the variable labels.  However, 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 selecting “Prefer variable labels” and that will list the variables by name.  However, you will have to do this each time you encounter a list of variables.  There is no way to do this permanently.

You’ll see that there are a total of 2,538 cases.  One of those cases said he or she didn’t know.  That means there are 2,537 valid cases that 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 [box] % 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, d4\_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, d4\_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 such hypotheses.

Notice how we are going about this.  We have a sample of adults in the United States (i.e., the 2014 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.**[[2]](http://ssric.org/node/564/edit%22%20%5Cl%20%22_edn2%22%20%5Co%20%22)**

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.**[[3]](http://ssric.org/node/564/edit%22%20%5Cl%20%22_edn3%22%20%5Co%20%22)** 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, d4\_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**[[4]](http://ssric.org/node/564/edit%22%20%5Cl%20%22_edn4%22%20%5Co%20%22)** (i.e., the number of respondents who answered the question).  In this problem, N equals 2,537.
* Mean is the mean years of school completed by the respondents in the sample who answered the question (see STAT2S).  In this problem, the sample mean equals 13.68.
* Standard Deviation is a measure of dispersion (see STAT2S).  In this problem, the standard deviation equals 3.07.
* 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

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 27.51.

Degrees of freedom (df) is the number of values that are free to vary.  If the sample mean equals 13.68 then how many values are free to vary?  The answer is N – 1 which is 2,537 – 1 or 2,536.  See if you can figure out why it’s 2,536.  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 (13.68) and the value specified in the null hypothesis (12).  So it’s 13.68 – 12 or 1.68.**[[5]](http://ssric.org/node/564/edit%22%20%5Cl%20%22_edn5%22%20%5Co%20%22)** 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 2014 GSS called d18\_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, 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 less 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 = 1,579).  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, 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.**[[6]](http://ssric.org/node/564/edit%22%20%5Cl%20%22_edn6%22%20%5Co%20%22)**  Why?  Remember that we said the larger the sample, the less the sampling error.  If there is less sampling error, it’s going to be easier 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.

**[[1]](http://ssric.org/node/564/edit%22%20%5Cl%20%22_ednref1%22%20%5Co%20%22)** The GSS is itself not a simple random sample but rather is an example of a multistate cluster sample.

**[[2]](http://ssric.org/node/564/edit%22%20%5Cl%20%22_ednref2%22%20%5Co%20%22)** Characteristics of a sample are called statistics while characteristics of a population are called parameters.

**[[3]](http://ssric.org/node/564/edit%22%20%5Cl%20%22_ednref3%22%20%5Co%20%22)** 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.

**[[4]](http://ssric.org/node/564/edit%22%20%5Cl%20%22_ednref4%22%20%5Co%20%22)** Missing cases would include those who said they didn’t know or refused to answer the question.

**[[5]](http://ssric.org/node/564/edit%22%20%5Cl%20%22_ednref5%22%20%5Co%20%22)** By the way, the value of the mean (13.68) is a rounded value so that’s why the mean difference isn’t exactly 1.68.

**[[6]](http://ssric.org/node/564/edit%22%20%5Cl%20%22_ednref6%22%20%5Co%20%22)** 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 value would be low enough to reject the null hypothesis.