**STAT6S\_pspp: Exercise Using PSPP to Explore Hypothesis Testing – Independent-Samples**

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 (means and independent-samples t test) to explore hypothesis testing.  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 SPSS**](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_STAT6S_pspp.docx) (MS Word; docx format).
* [**PSPP syntax file**](http://ssric.org/files/PSPP_Syntax_for_STAT6S_pspp.sps) (.sps format)
* [**This page**](http://ssric.org/files/STAT6S_pspp.docx) (MS Word; docx format).

**Goals 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 2014 GSS. Your instructor will tell you how to access this data set which is called gss14\_subset\_for\_classes\_STATISTICS\_pspp.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 d4\_educ and d18\_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 d5\_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.

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.

Men and women differ very little in the number of years of school they completed.  Men have completed a little less than one-tenth of a year more than women.  But men worked quite a bit more than women in the last week – a difference of almost 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 (STAT5S\_PSPP).

**Part II – Now it’s Your Turn**

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

* Do men and women differ in the number of hours per day they have to relax?  This is variable d20\_hrsrelax in the GSS.
* Do men and women differ in the number of hours per day they watch television?  This is variable tv1\_tvhours in the GSS.

Use PSPP to get the sample means and then compare them to begin answering these questions.

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

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

* d4\_educ
* d18\_hrs1

Now we want to determine if that difference is 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, d5\_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 f4\_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 d5\_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 d4\_educ.

* t is the value of the t test which is 0.58 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 = 2,535.  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) = 13.72 – 13.64 = .08.  Instead of using the rounded values, PSPP carries the computation out to more decimal points which results in a mean difference of .07. In other words, males have .07 of a year more education than females which is a very small difference.
* The standard error of the difference which is .12 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 2014 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 0.559 for both t tests.  That tells you that the probability of being wrong if you rejected the null hypothesis is just about .56 or 56 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 .09 then the one-tailed significance value is half that or .09 divided by two or .045.

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 [**STAT2S\_pspp**](http://ssric.org/node/562)).  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 appropriate test would be the F test which we’re not going to discuss until a later exercise (STAT8S\_pspp).

**PART IV – Now It's Your Turn Again**

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

* Do men and women differ in the number of hours per day they have to relax (d20\_hrsrelax)?
* Do men and women differ in the number of hours per day they watch television (tv1\_tvhours)?

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 2014 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 next exercise (STAT7S\_pspp).