**STAT7S\_pspp: Exercise Using PSPP to Explore Hypothesis Testing – Paired-Samples 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 (paired-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_STAT7S_pspp.docx) (MS Word; docx format).
* [**PSPP syntax file**](http://ssric.org/files/PSPP_Syntax_for_STAT7S_pspp.sps) (.sps format)
* [**This page**](http://ssric.org/files/STAT7S_pspp.docx) (MS Word; docx format).

**Goals 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 2014 GSS. Your instructor will tell you how to access this data set which is called gss14\_subset\_for\_classes\_STATISTICS\_pspp.sav.

In STAT6S\_pspp 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 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.

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 2014 GSS includes questions about the years of school completed by the respondent’s parents – d22\_maeduc and d24\_paeduc.  Let’s assume that we think that respondent’s fathers have more education than respondent’s mothers.  We would compare the mean years of school completed by mothers with the mean years of school completed by fathers.  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 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 d22\_maeduc and d24\_paeduc and move them to the “Dependent List” box.  These are the variables for which you are going to compute means.  The output from PSPP will show you the mean, number of cases, and standard deviation for fathers and mothers.

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.

Fathers have about two-tenths of a year more education than mothers.  Why can’t we just conclude that fathers have more education than mothers?  If we were just describing the **sample**, we could.  But what we want to do is to make inferences about differences between 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 (STAT5S\_pspp and STAT6S\_pspp).

**Part II – Now it’s Your Turn**

In this part of the exercise you want to compare the years of school completed by respondents and their spouses to determine whether men have more education than their spouses or whether women have more education than their spouses.

Use PSPP to get the sample means as we did in Part I and then compare them to begin answering this question.  But we need to be careful here.  Respondents could be either male or female.  We need to separate respondents into two groups – men and women – and then separately compare male respondents with their spouses and female respondents with their spouses.  We can do this by putting the variables d4\_educ and d29\_speduc into the “Dependent List” box and d5\_sex into the “Independent List” box.

Write a paragraph describing the difference between the education of the respondent and the spouse’s education for both males and females.  Use the means to illustrate your answer.

**Part III – 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.

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 d22\_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, d24\_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 -0.18.  This means that fathers had an average of almost two-tenths of a year more education than the mothers.  By the way, in Part I when we compared the means for d22\_maeduc and d24\_paeduc the difference was 0.22.  Here the mean difference score is .18.  Why aren’t they the same?  See if you can figure this out.  (Hint: it has something to do with comparing differences for pairs.)
* The standard deviation of the difference scores for all these pairs which is 3.21.
* 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 -2.32.  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 1,795 which is the number of pairs minus one or 1,796 – 1 or 1,795.  In other words, 1,795 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 .020 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 2014 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 fathers is greater than the mean for mothers 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 fathers is greater than the mean years of school completed by mothers 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 negative.  In other words, the mean years of school completed by fathers is greater than the mean years for mothers 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 0.020.  That tells you that the probability of being wrong if you rejected the null hypothesis is. 02 or 2 times out of one hundred.  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 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 negative.  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 .020 then the one-tailed significance value is half that or .020 divided by two or .010.  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 IV – Now it’s Your Turn Again**

In this part of the exercise you want to compare the years of school completed by respondents and their spouses to determine if women have more education than their spouses but this time you want to test the appropriate null hypotheses.

Remember from Part II that we have to test this hypothesis first for men and then for women.  We’re going to do this by selecting all the men and then computing the paired-samples t test.   Then we’re going to repeat this but this time by selecting all the women and then computing the paired-samples t test.  To do 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 commands.  You can do this by cutting and pasting these commands into the PSPP syntax file.  Once you have done this click on “Run” in the menu bar and then click on “All.”  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.

TEMPORARY.
SELECT IF d5\_sex = 1.
T-TEST PAIRS=d4\_educ WITH d29\_speduc (PAIRED)
  /CRITERIA=CI(.9500)
  /MISSING=ANALYSIS.
TEMPORARY.
SELECT IF d5\_sex = 2.
T-TEST PAIRS=d4\_educ WITH d29\_speduc (PAIRED)
  /CRITERIA=CI(.9500)
  /MISSING=ANALYSIS.

For each paired-sample t test, state the research and the null hypotheses.  Do you reject or not reject the null hypotheses?  Explain why.