**STAT8S\_pspp: Exercise Using PSPP to Explore Hypothesis Testing – One-Way Analysis of Variance**

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 and one-way analysis of variance 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_STAT8S_pspp.docx) (MS Word; docx format).
* [**PSPP syntax file**](http://ssric.org/files/PSPP_Syntax_for_STAT8S_pspp.sps) (.sps format)
* [**This page**](http://ssric.org/files/STAT8S_pspp.docx) (MS Word; docx format).

**Goals of Exercise**

The goal of this exercise is to explore hypothesis testing and one-way analysis of variance (sometimes abbreviated one-way anova). The exercise also gives you practice in using 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 while in STAT7S\_pspp 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.  In fact, the t test is a special case of analysis of variance.

The 2014 GSS includes a variable (d3\_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 (tv1\_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 tv1\_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 d3\_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.

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.

Respondents with more education watch less television than those with less education.  For example, respondents with a graduate degree watch an average of 1.86 hours of television per day while those who haven’t completed high school watch an average of 3.91 hours – a difference of about 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 (STAT5S\_pspp, STAT6S\_pspp, and STAT7S\_pspp).

**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 (d25\_region) watch more television (tv1\_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 tv1\_tvhours and move it to the “Dependent Variable(s)” box.  Then select the variable d3\_degree and move it to the “Factor” box.  Click on “Descriptives” in the “Statistics” box and finally click on “OK.”

The output gives you results of 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 2014 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?