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

**Note to the Instructor:** This exercise uses the 2014 General Social Survey (GSS) and SDA to explore hypothesis testing and one-way analysis of variance.  SDA (Survey Documentation and Analysis) is an online statistical package written by the Survey Methods Program at UC Berkeley and is available without cost wherever one has an internet connection.  The 2014 Cumulative Data File (1972 to 2014) is also available without cost by clicking [here](http://sda.berkeley.edu/sdaweb/analysis/?dataset=gss14).  For this exercise we will only be using the 2014 General Social Survey.  A weight variable is automatically applied to the data set so it better represents the population from which the sample was selected.  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 exercise itself.  Please contact the author for additional information.

I’m attaching the following files.

* [Extended notes for instructors](http://ssric.org/files/Extended_Notes_for_Instructors_for_STAT8S_SDA.docx) (MS Word; .docx format).
* [This page](http://ssric.org/files/STAT8S_SDA.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 MEANS in SDA.

### ****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 the 2014 GSS.  To access the GSS cumulative data file in SDA format click [here](http://sda.berkeley.edu/sdaweb/analysis/?dataset=gss14).  The cumulative data file contains all the data from each GSS survey conducted from 1972 through 2014.  We want to use only the data that was collected in 2014.  To select out the 2014 data, enter year(2014) in the Selection Filter(s) box.  Your screen should look like Figure 8-1.  This tells SDA to select out the 2014 data from the cumulative file.

Notice that a weight variable has already been entered in the WEIGHT box.  This will weight the data so the sample better represents the population from which the sample was selected.  Notice also that in the SAMPLE DESIGN line SRS has been selected.


**Figure 8-1**

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 2014 and is a relatively large sample of approximately 2,500 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.

In [STAT6S\_SDA](http://ssric.org/node/602/STAT6S_SDA) we used the t test to compare means from two independent samples.  But what if we wanted to compare means from more than two samples?  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 (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 getting the means for tvhours broken down by degree.  Click on MEANS in the menu bar at the top of SDA and enter the variable tvhours in the DEPENDENT box.  The dependent variable will always be the variable for which you are going to compute means.  Then enter the variable degree in the ROW box.  This is the variable which defines the groups you want to compare.  In our case we want to compare respondents with different levels of education.  The output from SDA will show you the mean number of hours respondents watched television for each level of education.  Notice that you must enter one or more variables in both the DEPENDENT and ROW boxes.  That what it means when it says REQUIRED next to these boxes.  Your screen should like Figure 8-2. Notice that the SELECTION FILTER(S) box and the WEIGHT box are both filled in. Be sure to click on OUTPUT OPTIONS and both check SRS STD ERRS and uncheck COMPLEX STD ERRS.  Click RUN THE TABLE to produce the means.



**Figure 8-2**

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 exercise [STAT6S\_SDA](http://ssric.org/node/602/STAT6S_SDA).

### ****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 SDA 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 number of 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 MEANS in the menu bar at the top of SDA and enter the variable tvhours in the DEPENDENT box as you did in Part I.  This time enter the variable degree in the ROW box.  The SELECTION FILTER(S) box and the WEIGHT box should both be filled in. Be sure to click on OUTPUT OPTIONS and both check SRS STD ERRS and uncheck COMPLEX STD ERRS.  Now we want to determine if the differences between levels of education are statistically significant by carrying out a one-way analysis of variance.  Click on OUTPUT OPTIONS and check the box for ANOVA STATS under OTHER OPTIONS.  Finally, click RUN THE TABLE to carry out the procedure.

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 0.0000.  By the way, it isn’t exactly 0.  This is a rounded value.  It means it is less than 0.00005.  That tells you that the probability of being wrong if you rejected the null hypothesis is really, really small.  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?