**STAT3S\_pspp: Exercise Using PSPP to Explore Measures of Skewness and Kurtosis**

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 FREQUENCIES in PSPP to explore measures of skewness and kurtosis.  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_STAT3S_pspp.docx) (MS Word; docx format).
* [**PSPP syntax file**](http://ssric.org/files/PSPP_Syntax_for_STAT3S_pspp.sps) (.sps format)
* [**This page**](http://ssric.org/files/STAT3S_pspp.docx) (MS Word; docx format).

 **Goals of Exercise**

The goal of this exercise is to explore measures of skewness and kurtosis. The exercise also gives you practice in using FREQUENCIES in PSPP.

**Part I – Measures of Skewness**

A normal distribution is a unimodal (i.e., single peak) distribution that is perfectly symmetrical.  In a normal distribution the mean, median, and mode are all equal.  Here’s a graph showing what a normal distribution looks like.



The horizontal axis is marked off in terms of standard scores where a standard score tells us how many standard deviations a value is from the mean of the normal distribution.  So a standard score of +1 is one standard deviation above the mean and a standard score of -1 is one standard deviation below the mean.  The percents tell us the percent of cases that you would expect between the mean and a particular standard score if the distribution was perfectly normal.  You would expect to find approximately 34% of the cases between the mean and a standard score of +1 or -1.  In a normal distribution, the mean, median, and mode are all equal and are at the center of the distribution.  So the mean always has a standard score of zero.

Skewness measures the deviation of a particular distribution from this symmetrical pattern.  In a skewed distribution one side has longer or fatter tails than the other side.  If the longer tail is to the left, then it is called a negatively skewed distribution.  If the longer tail is to the right, then it is called a positively skewed distribution.  One way to remember this is to recall that any value to the left of zero is negative and any value to the right of zero is positive.  Here are graphs of positively and negatively skewed distributions compared to a normal distribution.



The best way to determine the skewness of a distribution is to tell PSPP to give you a histogram along with the mean and median.  PSPP will also compute a measure of skewness.  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.

Run FREQUENCIES in PSPP for the variables d1\_age and d9\_sibs.  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.

Once you have selected the variables, then check the boxes for mean, median, skewness, and kurtosis in the “Statistics” box and click on the “Charts” button and select “Draw histogram” and “Superimpose normal curve.”  Click on “Continue” and then on “OK.”  We’ll talk about kurtosis in a little bit.

Notice that the mean is larger than the median for both variables.  This means that the distribution is positively skewed.  But also notice that the mean for d9\_sibs is quite a bit larger than the median in a relative sense than is the case for d1\_age.  This suggests that the distribution for d9\_sibs is the more skewed of the two variables.  Look at the histograms and you’ll see the same thing.  Both variables are positively skewed but d9\_sibs is the more skewed variable.  Now look at the skewness values — 1.73 for d9\_sibs and .24 for d1\_age.  The larger the skewness value, the more skewed the distribution.  Positive skewness values indicate a positive skew and negative values indicate a negative skew.  There are various rules of thumb suggested for what constitutes a lot of skew but for our purposes we’ll just say that the larger the value, the more the skewness and the sign of the value indicates the direction of the skew.

Run FREQUENCIES for the following variables.  Tell PSPP to give you the histogram and to superimpose the normal curve on the histogram. Also ask for the mean, median, and skewness.  Write a paragraph for each variable explaining what these statistics tell you about the skewness of the variables.

* d20\_hrsrelax
* tv1\_tvhours

 **Part II – Measures of Kurtosis**

Kurtosis refers to the flatness or peakness of a distribution relative to that of a normal distribution.  Distributions that are flatter than a normal distribution are called platykurtic and distributions that are more peaked are called leptokurtic.



PSPP will compute a kurtosis measure.  Negative values indicate a platykurtic distribution and positive values indicate a leptokurtic distribution.  The larger the kurtosis value, the more peaked or flat the distribution is.

Look back at the output for d1\_age and d9\_sibs.  For d1\_age the kurtosis value was -.80 indicating a flatter distribution and for d9\_sibs kurtosis was 4.41 indicating a more peaked distribution.  To see this visually look at your histograms.

Run FREQUENCIES for the following variables.  Tell PSPP to give you the histogram and to superimpose the normal curve on the histogram. Also ask for kurtosis.  Write a paragraph for each variable explaining what these statistics tell you about the kurtosis of the variables.

* d22\_maeduc
* d24\_paeduc