**STAT2S\_pspp:  Exercise Using PSPP to Explore Measures of Central Tendency and Dispersion**

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 central tendency and dispersion.  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_STAT2S_pspp.docx) (MS Word; docx format).
* [**PSPP syntax file**](http://ssric.org/files/PSPP_Syntax_for_STAT2S_pspp.sps) (.sps format)
* [**This page**](http://ssric.org/files/STAT2S_pspp.docx) (MS Word; docx format).

 **Goals of Exercise**

The goal of this exercise is to explore measures of central tendency (mode, median, and mean) and dispersion (range, interquartile range, standard deviation, and variance). The exercise also gives you practice in using FREQUENCIES in PSPP.

**Part I – Measures of Central Tendency**

Data analysis always starts with describing variables one-at-a-time.  Sometimes this is referred to as univariate (one-variable) analysis.  Central tendency refers to the center of the distribution.

There are three commonly used measures of central tendency – the mode, median, and mean of a distribution.  The mode is the most common value or values in a distribution**[[1]](http://ssric.org/node/562/edit%22%20%5Cl%20%22_edn1%22%20%5Co%20%22)**.  The median is the middle value of a distribution.**[[2]](http://ssric.org/node/562/edit%22%20%5Cl%20%22_edn2%22%20%5Co%20%22)** The mean is the sum of all the values divided by the number of values.

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 variable 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 this variable then check the boxes for mode, median, and mean in the “Statistics” box and click on the “Charts” button.  Select “Draw histogram” and check the box for “Superimpose normal curve.”  Click on “Continue” and then click on “OK.” 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 in the upper right of the circle) and click on the output window.  PSPP will open the Output window and display the results that you requested.

Your output will display the frequency distribution for d9\_sibs and a box showing the mode, median, and mean with the following values displayed.

* Mode = 2 meaning that two brothers and sisters was the most common answer (19.4%) from the 2,531 respondents who answered this question.  However, not far behind are those with one sibling (18.6%) and those with three siblings (17.9%).  So while technically two siblings is the mode, what you really found is that the most common values are one, two, and three siblings.  Another part of your output is the histogram which is a chart or graph of the frequency distribution.  The histogram clearly shows that one, two, and three are the most common values (i.e., the highest bars in the histogram).  So we would want to report that these three categories are the most common responses.
* Median = 3 which means that three siblings is the middle category in this distribution.  The middle category is the category that contains the 50th percentile which is the value that divides the distribution into two equal parts.   In other words, it’s the value that has 50% of the cases above it and 50% of the cases below it.  The cumulative percent column of the frequency distribution tells you that 41.4% of the cases have two or fewer siblings and that 59.3% of the cases have three or fewer siblings.  So the middle case (i.e., the 50th percentile) falls somewhere in the category of three siblings.  That is the median category.
* Mean = 3.74 which is the sum of all the values in the distribution divided by the number of responses.  If you were to sum all these values that sum would be 9,476.  Dividing that by the number of responses or 2,531 will give you the mean of 3.74.

**Part II – Deciding Which Measure of Central Tendency to Use**

The first thing to consider is the level of measurement (nominal, ordinal, interval, ratio) of your variable (see Exercise [**STAT1S\_pspp**](http://ssric.org/node/560)).

* If the variable is nominal, you have only one choice.  You must use the mode.
* If the variable is ordinal, you could use the mode or the median.  You should report both measures of central tendency since they tell you different things about the distribution.  The mode tells you the most common value or values while the median tells you where the middle of the distribution lies.
* If the variable is interval or ratio, you could use the mode or the median or the mean.  Now it gets a little more complicated.  There are several things to consider.
	+ How skewed is your distribution? **[[3]](http://ssric.org/node/562/edit%22%20%5Cl%20%22_edn3%22%20%5Co%20%22)** Go back and look at the histogram for d9\_sibs. Notice that there is a long tail to the right of the distribution.  Most of the values are at the lower level – one, two, and three siblings.  But there are quite a few respondents who report having four or more siblings and about 5% said they have ten or more siblings.  That’s what we call a positively skewed distribution where there is a long tail towards the right or the positive direction. Now look at the median and mean.  The mean (3.74) is larger than the median (3.0).  The respondents with lots of siblings pull the mean up.  That’s what happens in a skewed distribution.  The mean is pulled in the direction of the skew.  The opposite would happen in a negatively skewed distribution.  The long tail would be towards the left and the mean would be lower than the median.  In a heavily skewed distribution the mean is distorted and pulled considerably in the direction of the skew.  So consider reporting only the median in a heavily skewed distribution.  That’s why you almost always see median income reported and not mean income.  Imagine what would happen if your sample happened to include Bill Gates.  The income distribution would have this very, very large value which would pull the mean up but not affect the median.
	+ Is there more than one clearly defined peak in your distribution?   The number of siblings has one clearly defined peak – one, two and three siblings.  But what if there is more than one clearly defined peak?  For example, consider a hypothetical distribution of 100 cases in which there are 50 cases with a value of two and fifty cases with a value of 8.  The median and mean would be five but there are really two centers of this distribution – two and eight.  The median and the mean aren’t telling the correct story about the center. You’re better off reporting the two clearly defined peaks of this distribution and not reporting the median and mean.
	+ If your distribution is normal in appearance then the mode, median, and mean will all be about the same.  A normal distribution is a perfectly symmetrical distribution with a single peak in the center.  No empirical distribution is perfectly normal but distributions often are approximately normal.  Here we would report all three measures of central tendency.  Go back to your PSPP output and look at the histogram for d9\_sibs.  When you told PSPP to give you the histogram you checked the box that said “Superimpose normal curve.”  The normal curve doesn’t fit the histogram perfectly particularly at the lower end but it does suggest that it approximates a normal curve particularly at the upper end.

Run FREQUENCIES for the variables below.  Once you have selected the variables then check the boxes for mode, median, and mean in the “Statistics” box and click on the “Charts” button.  Select either “Draw histogram” or “Draw bar chart” depending on what is indicated to the right of the variable name.  Click on “Continue” and then click on “OK.” PSPP will open the Output window and display the results of what you requested.  For each variable write a sentence or two indicating which measure(s) of central tendency would be appropriate to use to describe the center of the distribution and what the values of those statistics mean.

* hap2\_happy – draw bar chart
* p1\_partyid – draw bar chart
* r8\_reliten – draw bar chart
* d1\_age – draw histogram and superimpose normal curve on histogram

**Part III – Measures of Dispersion or Variation**

Dispersion or variation refers to the degree that values in a distribution are spread out or dispersed.  The measures of dispersion that we’re going to discuss are appropriate for interval and ratio level variables (see Exercise STAT1S).**[[4]](http://ssric.org/node/562/edit%22%20%5Cl%20%22_edn4%22%20%5Co%20%22)**  We’re going to discuss four such measures – the range, the inter-quartile range, the variance, and the standard deviation.

The range is the difference between the highest and the lowest values in the distribution.  Run FREQUENCIES for d1\_age and compute the range by looking at the frequency distribution.  You can also ask PSPP to compute it for you.  Click on “Range” in the “Statistics” box.  You should get 71 which is 89 – 18.**[[5]](http://ssric.org/node/562/edit%22%20%5Cl%20%22_edn5%22%20%5Co%20%22)**  The range is not a very stable measure since it depends on the two most extreme values – the highest and lowest values.  These are the values most likely to change from sample to sample.

A more stable measure of dispersion is the interquartile range which is the difference between the third quartile (Q3) and the first quartile (Q1).  The third quartile is the same thing as the seventy-fifth percentile which is the value that has 25% of the cases above it and 75% of the cases below it.  The first quartile is the same as the twenty-fifth percentile which is the value that has 75% of the cases above it and 25% of the cases below it.  Look at the cumulative percent column in the frequency distribution for age.  The first quartile will be the category than contains the cumulative percent of 25.0 and the third quartile will be the category that contains the cumulative percent of 75.0.  Once you know Q3 and Q1 you can calculate the interquartile range by subtracting Q1 from Q3.  Since it’s not based on the most extreme values it will be more stable from sample to sample.  From the cumulative percent column you should see that Q3 will equal 60 and Q1 will equal 33 and the interquartile range will equal 60 – 33 or 27.

The variance is the sum of the squared deviations from the mean divided by the number of cases minus 1 and the standard deviation is just the square root of the variance.  Your instructor may want to go into more detail on how to calculate the variance by hand.  PSPP will calculate the standard deviation for you.  Select standard deviation from the “Statistics” box to tell PSPP to calculate the standard deviation.  Then square the standard deviation (i.e., 17.24) to get the variance (297.22).

The variance and the standard deviation can never be negative.  A value of 0 means that there is no variation or dispersion at all in the distribution.  All the values are the same.  The more variation there is, the larger the variance and standard deviation.

So what does the variance (297.22) and the standard deviation (17.24) of the age distribution mean?  That’s hard to answer because you don’t have anything to compare it to.  But if you knew the standard deviation for both men and women you would be able to determine whether men or women have more variation.  Instead of comparing the standard deviations for men and women you would compute a statistic called the Coefficient of Relative Variation (CRV).  CRV is equal to the standard deviation divided by the mean of the distribution.   A CRV of 2 means that the standard deviation is twice the mean and a CRV of 0.5 means that the standard deviation is one-half of the mean.  You would compare the CRV’s for men and women to see whether men or women have more variation relative to their respective means.

You might also have wondered why you need both the variance and the standard deviation when the standard deviation is just the square root of the variance.  You’ll just have to take my word for it that you will need both as you go further in statistics.

Run FREQUENCIES for the following variables.  Once you have selected the variables then check the boxes for range, standard deviation, and mean in the “Statistics” box.  Click on “OK” and then click on the Output window and PSPP will display the results of what you requested.  For each variable write a sentence or two indicating what the values of these statistics are for each of the variables and what the values of those statistics mean.  Compare the relative variation for the number of male sex partners since the age of 18 (s1\_nummen) and the number of female sex partners (s2\_numwomen) by comparing the CRV’s for each variable.

* s1\_nummen
* s2\_numwomen

**[[1]](http://ssric.org/node/562/edit%22%20%5Cl%20%22_ednref1%22%20%5Co%20%22)** Frequency distributions can be grouped or ungrouped.  Think of age.  We could have a distribution that lists all the ages in years of the respondents to our survey.  One of the variables (d1\_age) in our data set does this.  But we could also divide age into a series of categories such as under 30, 30 to 39, 40 to 49, 50 to 59, 60 to 69, and 70 and older.  In a grouped frequency distribution the mode would be the most common category or categories.

**[[2]](http://ssric.org/node/562/edit%22%20%5Cl%20%22_ednref2%22%20%5Co%20%22)** In a grouped frequency distribution the median would be the category that contains the middle value.

**[[3]](http://ssric.org/node/562/edit%22%20%5Cl%20%22_ednref3%22%20%5Co%20%22)** See Exercise STAT3S for a more through discussion of skewness.

**[[4]](http://ssric.org/node/562/edit%22%20%5Cl%20%22_ednref4%22%20%5Co%20%22)** The Index of Qualitative Variation can be used to measure variation for nominal variables.

**[[5]](http://ssric.org/node/562/edit%22%20%5Cl%20%22_ednref5%22%20%5Co%20%22)** Actually the highest age is larger than 89.  The GSS combines all respondents that are 89 or older into one category which is given the value of 89.  But for our purposes that is close enough.