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 central tendency and dispersion.  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_STAT2S_SDA.docx) (MS Word; .docx format).
* [This page](http://ssric.org/files/STAT2S_SDA.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, standard deviation, and variance). The exercise also gives you practice in using FREQUENCIES in SDA.

### ****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]](%22%20%5Cl%20%22_ftn1%22%20%5Co%20%22)

The median is the middle value of a distribution.[[2]](%22%20%5Cl%20%22_ftn2%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 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 1-1.  This tells SDA to select out the 2014 data from the cumulative file.


**Figure 2-1**

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.

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.

Run FREQUENCIES in SDA for the variable sibs.  To run the frequency distribution, enter the variable name, sibs, in the ROW box.  Your screen should like Figure 2-2.  Notice that the SELECTION FILTER(S) box and the WEIGHT box are both filled in.

|  |
| --- |
|  This image shows the SDA Frequencies dialog box with the row,  selection filter(s) and weight boxes filled in. |

**Figure 2-2**

Once you have selected this variable, click on the arrow next to OUTPUT OPTIONS and check the box for SUMMARY STATISTICS.  Then click on CHART OPTIONS and click the arrow next to TYPE OF CHART.  Select BAR CHART and now click on RUN THE TABLE at the bottom.

Your output will display the frequency distribution for sibs, the summary statistics, and the bar chart.  Three of the summary statistics are commonly used measures of central tendency – mode, median, and mean.

* 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 bar chart which is a chart or graph of the frequency distribution.  The bar chart clearly shows that one, two, and three are the most common values (i.e., the highest bars in the bar chart).  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.  If you added up the percents for all values less than 3 and the percents for all values less than or equal to  3, you would find 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\_SDA](http://www.ssric.org/node/598)).

* 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]](%22%20%5Cl%20%22_ftn3%22%20%5Co%20%22)  Go back and look at the bar chart for sibs. Notice that there is a long tail to the right of the distribution.  Most of the values are at the lower end – 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 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.

Run FREQUENCIES for the following variables.  Once you have selected the variables in the ROW box, ask for the SUMMARY STATISTICS and a BAR CHART.  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.  For some variables there will be more than one appropriate measure of central tendency.

* happy
* partyid
* reliten
* nummen
* numwomen
* age

### ****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\_SDA](http://www.ssric.org/node/598).)[[4]](%22%20%5Cl%20%22_ftn4%22%20%5Co%20%22) We’re going to discuss three such measures – the 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 age and compute the range by looking at the frequency distribution.  You can also ask SDA to compute it for you.  Once you have selected this variable click on the arrow next to OUTPUT OPTIONS and check the box for SUMMARY STATISTICS.   Now click on RUN THE TABLE at the bottom.

The range should equal 71 which is 89 – 18.  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.

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.  SDA will also calculate it for you.  The variance should equal 297.29 and the standard deviation will equal 17.24.

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 (297.29) 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 in the ROW box, ask for the SUMMARY STATISTICS.  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 (nummen) and the number of female sex partners (numwomen) by comparing the CRV’s for each variable.

* nummen
* numwomen?
* sibs

[[1]](%22%20%5Cl%20%22_ftnref1%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 (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]](%22%20%5Cl%20%22_ftnref2%22%20%5Co%20%22) In a grouped frequency distribution the median would be the category that contains the middle value.

[[3]](%22%20%5Cl%20%22_ftnref3%22%20%5Co%20%22) See Exercise [STAT3S\_SDA](http://www.ssric.org/node/600) for a more through discussion of skewness.

[[4]](%22%20%5Cl%20%22_ftnref4%22%20%5Co%20%22) The Interquartile Range can also be used to measure variation for an interval or ratio variable and the Index of Qualitative Variation can be used to measure variation for nominal variables.