EvalFest How-Tos for Numeric Exploratory Data Analysis in Excel

Numeric Exploratory Data Analysis in Excel

These instructions build on videos in our Data Management series. In the Database Organization and Password Protection and Data Cleaning and Recoding videos, we discuss how to prepare your data for analysis. And the Pivot Tables and Basic Analysis video demonstrates how to compute basic frequencies for each answer category associated with a variable (or item).

> **Now you might be thinking, why bother?** As evaluators, we do our best to ensure that the data gathered will allow us to tell the story of the data. Before you perform any statistical tests, you want to make sure you know your data, and that you have good data with which to tell the story. This is also the time to confirm your understanding of how the data were collected and coded so that you know how to interpret your results.

Exploring the data numerically: Descriptive statistics

1. Frequencies (visual exploration)
    a. There are a number of ways to check that the data are, for the most part, as you expected. The first is by simply sorting the data and taking a quick look through the responses for a particular variable of interest.
    b. Let's look at Overall rating, for example. We know that, typically, festivals are rated on the high, or positive, end of the scale, so we would expect to see many more 5s than 1s or 2s in the OverallCode variable.
    c. First, click in the top left corner of the spreadsheet.
    d. Choose the Data menu.
    e. Click Sort.
    f. Under Column, choose the column on which you want to sort the data. Here, we'll choose OverallCode.
    g. Under Order, you can choose to sort from Smallest to Largest, Largest to Smallest, or create custom sort criteria. Go ahead and sort from smallest to largest.
    h. You can quickly look down the list of OverallCode values to see that there are many more 4s and 5s, or Very Good and Excellent ratings, than ratings in the lower categories.
2. Frequency Tables
    a. Another way to view frequencies of responses is to create a frequency table.
    b. First, you'll want to open a new sheet in your workbook. You can do this by

clicking the plus sign at the bottom of the workbook.

c. Next, you want to create a list of the answer categories in this new workbook. We'll work with Overall rating, so I'll go ahead and type in the categories of Poor through Excellent in my new workbook.
d. In the cell next to the first category name, begin to type =COUNTIF. Choose COUNTIF from the list.
e. First, highlight the range of cells you want Excel to search and, subsequently, count. These are located in the worksheet that contains the data.
f. Type a comma and then click on the cell that contains the first category name on our new worksheet.
g. Press Return or Enter.
h. The cell should populate with the count of the category, in this case, Poor.
i. You can drag the small green square to the cells below to produce counts for the other categories, in this case Fair through Excellent.
j. You can see that the large majority of the responses are either Very Good or Excellent, as expected.

Besides using frequencies to ensure that your data are ready for analysis, there are certain statistics, or measures of your sample data, that you may want to explore. Typically, these include measures of center, such as the mean and/or the median, as well as measures of variation, like the standard deviation.

3. Mean
    a. In order to compute the mean for a numeric variable, you will want to first sort the data set by the variable of interest. This step isn't completely necessary, but makes things a bit easier. In this case, we are going to compute the mean of the Overall rating, so we will sort by OverallCode.
    b. Next, you want to go to the end of the list of answer items for OverallCode.
    c. In the cell below the last entry, begin typing '=AVERAGE'. Once you begin typing, this will appear as an option to choose.
    d. Choose AVERAGE.
    e. Highlight all of the cells that contain data for the variable for which you want to compute a mean.
        i. A quick way to do this is to click on the last entry, hold down Shift and Command (or Control on a PC), and press the up arrow.
        ii. This will also select the column header, which you don't want in the range. Release the Command (or Control) key, and click on the first entry in the list.
    f. Click Return (or Enter) on your keyboard, and the mean appears at the end of the list of entries.
    g. A question to ask at this point is if the value of the mean makes sense for the data. In this example, we know that these ratings were entered as 1 – 5 and so our mean rating should be within that range. If it isn't, something is amiss in the data. Our mean is 4.388, which is between 1 and 5, and this matches our expectations.
4. Median
    a. Next, you might want to look at the median. In case your memory is a little fuzzy on this term, the median is determined by lining up all the ratings in order from lowest to highest and figuring out the rating that is in the very center of your

data.

b. In the case that you want to report the median as the measure of center (either on its own or in conjunction with the mean), you can compute the median using the same method as we did to compute the mean.

c. Instead of beginning to type =AVERAGE, you would begin to type =MEDIAN and follow the same steps as above. You can see that the median is 5, which may seem odd. But recall that most of our responses were Excellent, or 5.

d. Why report both? The relationship between the mean and the median tells us about the distribution of data, as long as the data are on a scale from low to high. This can be numeric data or categorical data. If the mean and median are very similar to one another, we can assume our data are symmetrically distributed. If the mean is greater than the median, our data are typically right skewed. If the mean is less than the median, that typically indicates that our data are left skewed. For example, let's consider the overall festival rating, which is on a scale from 1 to 5, or Poor to Excellent. If the data are left skewed, you will have more high ratings than low ratings, and the data will trail off to the left (hence, "left skewed"). Right skewed data are just the opposite – there will be more low values than high values. Ratings data for festival events tend to be left skewed, indicating that more respondents rated the event on the high end of the scale than on the low end.

5. Standard deviation
    a. We often want to report a standard deviation, or a measure of the variation in the data, along with the mean. The variation in the data means how different the ratings are from one another, and from the mean. Did most respondents choose the same rating, for example, Excellent? If so, there is little variation in the data. However, if you have several ratings of Poor, some ratings of Fair, some ratings of Good, a few ratings of Very Good and a several ratings of Excellent, you have a lot of variation in the data. The standard deviation tells us if our data contain just a few ratings (low standard deviation) or a variety of ratings (high standard deviation).
        i. Why do we care about standard deviation? Think about an event that was rated as Good, on average, so it had a mean of 3 out of 5 on our scale. This could happen if all 100 respondents rated it as Good (3), and you would compose your report or presentation based on this information. The event could also have an average rating of Good if 50 respondents rated it as Poor (1) and 50 rated it as Excellent (5) - and you would probably tell a very different story given this scenario. The mean doesn't tell you about the variability in ratings - the standard deviation does.
    b. The formula to compute the standard deviation is specific to the mean (the average), and you would not report the standard deviation as the measure of variation along with the median - only with the mean.
    c. You have two options here. You can compute a standard deviation for the population (or STDEV.P), or standard deviation for the sample (or STDEV.S). If your sample is very large, there will be very little difference in these two values. First, let's compute STDEV.P. for Overall rating.
    d. As with the mean, begin to type =STDEV.P in the cell below the mean (or

another cell of your choice). You will see both STDEV.P and STDEV.S appear. Choose STDEV.P.

e. Again, choose the full list of entries for the variable (or item), in this case, OverallCode.
f. Click Return (or Enter), and the population standard deviation appears.
g. You can repeat this process to compute the standard deviation for the sample, using the STDEV.S command if you wish.
h. For these data, we have a sample standard deviation of 0.7586 and a population standard deviation of 0.7585. You can see that, because we have a very large sample, these two are very close in value. We can interpret the sample standard deviation as the average number of units our entire set of ratings is from our mean of 4.388. Thinking about our scale from 1 to 5, and the magnitude of our standard deviation, we can think of this as saying that there is very little variation in our data, and most of the ratings are in the same 1 or 2 categories. We can see this from this graph of the data as well. Most of the ratings are in the Very Good or Excellent categories, with little representation for Good, Fair, or Poor.