



EvalFest How-Tos for Numeric Exploratory Data Analysis in SPSS

Numeric Exploratory Data Analysis in SPSS

Hi, this is Jane. In this video, we will talk about exploring the data numerically using SPSS version 25. This video builds on others 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.

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. Click on the Data menu at the top of the screen.
 - d. Choose Sort Cases.
 - e. Highlight the variable you want to explore, in this case, OverallCode.
 - f. Click the arrow to move the variable to the Sort by: box.
 - g. You can choose to sort ascending or descending.
 - h. Click OK.
 - i. The output window will appear, so just minimize that.
 - j. Take a look at the Data View of your data file.
 - k. Find the variable OverallCode and highlight it by clicking on the variable name.
 - l. You can quickly scroll through the OverallCode data to see that there are many more 4s and 5s, or Very Good and Excellent ratings, than ratings in the lower categories. Note that a dot indicates missing data, or no response.
2. Frequency Tables
 - a. Another way to view frequencies of responses is to create a frequency table.
 - b. Click on the Analyze menu at the top of the screen.

- c. Bring your mouse to hover over Descriptive Statistics.
- d. Choose Frequencies from the menu that appears on the right.
- e. Again, we will highlight OverallCode on the left and use the arrow to place it in the Variable(s) box on the right.
- f. Click OK.
- g. In the output window that appears, you will see a table of the counts, or frequencies, associated with each Overall Rating Code, numbered 1 through 5.

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, Median, and Standard Deviation

- a. When you create a frequency table, you can ask SPSS to compute other descriptive statistics, such as the mean, median, and standard deviation.
- b. Most of us understand the mean to be the average value. 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.
- c. 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.

4. Computing the mean, median and standard deviation along with a frequency table

- a. Let's go back to the Analyze menu and choose Frequencies again.
- b. You can see that SPSS remembers the last variable(s) we used, and automatically put them in the Variable(s) box.
- c. If you click Statistics, you will see options for several descriptive statistics that SPSS will compute.
- d. Check Mean, Median, and Std. deviation.
- e. Click Continue, and OK.
- f. Now, above the frequency table, you see the values of the descriptive

statistics you requested.

5. Questions to ask and interpretations

- a. A question to ask at this point is if these values make sense for the data. Let's consider the mean. 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 4.39, which is between 1 and 5, and this matches our expectations.
- b. The value of the median, 5, might seem odd. But recall that most of the festival-goers rated the festival as Excellent, or 5.
- c. Why might you want to report both the mean and the median? 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 you will see the data 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.
- d. For these data, we have a standard deviation of 0.759. We can interpret the standard deviation as the average number of units our entire set of ratings is from our mean of 4.39. 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.



UNC
MOREHEAD PLANETARIUM
AND SCIENCE CENTER



karen peterman
CONSULTING

