EvalFest How-Tos for Numeric Exploratory Data Analysis in R

Numeric Exploratory Data Analysis in R

Here, we will explore the data numerically using R. I use RStudio as my editor. This builds 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.

> **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. In the .csv data file, click in the top left corner of the spreadsheet to select all rows and columns.
   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. We will sort from largest to smallest.
   h. Click OK.
   i. 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. You should recognize lines 2, 3, and 5 as the code to set up the data for R.

    c. Line 8 tells RStudio to create a frequency table of Overall Rating codes, and line 9 displays the table.
    d. Again, we see the higher frequencies in the 4 and 5 categories, as opposed to the lower ratings.

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. Most of us understand the mean to be the average value. Line 12 is the code to compute the mean Overall Rating. The code at the end of the line, na.rm=T tells R to remove any missing cases. If you don't include this piece of code, you won't be able to compute these sample statistics.
    b. In case your memory is a little fuzzy on the median, this value 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. Line 13 is the code to compute the median Overall Rating.
    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.
    d. Line 14 computes the standard deviation for Overall Festival Rating.
4. 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 we can see how 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.