Chem 197
Fall 2019

measurement exercise ii

Part A: Ways to Describe Measurements

At the heart of any analysis are measurements. Sometimes our measurements are categorical and sometimes they are numerical; sometimes our measurements convey order and sometimes they do not; sometimes our measurements have an absolute reference and sometimes they have an arbitrary reference; and sometimes our measurements take on discrete values and sometimes they take on continuous values. To give meaning to these descriptive terms, let's consider the set of measurements in Table 1, where each column is a different measurement, or variable, and each row is a record for a single sample. The result is an aggregation of a set of measurements on individual samples into a single dataset. As a group, consider each of the following prompts and be prepared to share your conclusions.

Table 1. Distribution of Yellow and Red M&Ms

bag id	year	weight (oz)	type	# yellow M&Ms	% red M&Ms	total M&Ms	rank (total M&Ms)
a	2006	1.74	peanut	2	27.8	18	sixth
b	2006	1.74	peanut	3	4.3	23	fourth
c	2000	0.80	plain	1	22.7	22	fifth
d	2000	0.80	plain	5	20.8	24	third
e	1994	10.0	plain	56	23.0	331	second
f	1994	10.0	plain	63	21.9	333	first

Prompt 1. Of the variables included in Table 1, some are categorical and some are numerical. Define these terms and assign each of the variables to one of these terms.

Prompt 2. Suppose we decide to code the type of M&Ms using 1 for plain and 2 for peanut. Does this change your answer to the previous prompt? Why or why not?

Prompt 3. Categorical variables are described as nominal or ordinal. Define the terms nominal and ordinal and assign each of the categorical variables in Table 1 to one of these terms.

Prompt 4. A numerical variable is useful because we can use it to make quantitative comparisons between samples; for example, there are approximately 14 times more plain M&Ms in a 10-oz. bag as there are in a 0.8-oz. bag. Although we can complete a meaningful calculation with any numerical variable, the types of calculations we can perform depend on whether the variable has an absolute reference or an arbitrary reference. A numerical variable is described as either ratio or interval depending on whether it has (ratio) or does not have (interval) an absolute reference. Explain what it means for a variable to have an absolute reference and assign each of the numerical variables in Table 1 as either a ratio variable or an interval variable. Why might this difference be important?

Finally, the granularity of a variable's possible values provides one more way to describe our data.

Prompt 5. Numerical variables also are described as discrete or continuous. Define the terms discrete and continuous and assign each of the numerical variables in Table 1 to one of these terms.

Part B: Ways to Summarize Aggregate Data

For a numerical variable, we can reduce the individual measurements into single representations of their central tendency and their dispersion, the most common of which are the mean, the median, the standard deviation, and the interquartile range; these summary statistics are defined below in Table 2.

Table 2. Summary Statistics for n Samples

statistic	description	formula
mean	average of the *n* measurements	$\overline{X} = \frac{\sum_{i=1}^{i=n} X_i}{n}$
median	middle value when the *n* measurements are ranked from smallest-to-largest; for an even number of measurements, the median is the average of the middle two values	—
standard deviation	average deviation of the *n* individual measurements relative to their mean	$s = \sqrt{\frac{\sum_{i=1}^{i=n}(X_i - \overline{X})^2}{n-1}}$
interquartile range	the range for the middle 50% of values when the *n* measurements are ranked from smallest-to-largest	—

Prompt 6. Open the .csv file with our aggregate data for M&Ms and use the formulas in Table 3 to report the mean, median, standard deviation, and interquartile range for any two colors of M&Ms. What conclusions can you draw from these values?

Table 3. Formulas for Summary Statistics

statistic	formula in Excel or Google Sheets
mean	=average(*start:end*)
median	=median(*start:end*)
standard deviation	=stdev(*start:end)*
interquartile range	=quartile(*start:end, 3*) – quartile(*start:end, 1*)

Prompt 7. We describe a statistic as either robust or as non-robust based on how sensitive it is to an individual measurement that is an outlier. Suppose the number of blue M&Ms in the first bag of our aggregate data was recorded in error as 54. Using the definitions and formulas in Table 2, decide which of the four statistics are insensitive to this error (robust) and which are sensitive to this error (non-robust). You can check your conclusions by changing the value in your spreadsheet and observing how it affects each of the summary statistics.

Part C. One Way to Visualize Aggregate Data

The statistics introduced in Part B provide a powerful summary of our aggregated data, but at the cost of losing information about the individual measurements. A boxplot is a simple way to report the mean, median, interquartile range, and to show the individual measurements and possible outliers. You can access the on-line tool BoxPlotR to create boxplots at http://shiny.chemgrid.org/boxplotr/. To use BoxPlotR, copy and paste our data for the counts of M&Ms by color, including the column headers, into the space for entering data, selecting comma as the separator if using a .csv file.

Prompt 8. Compare the boxplots to our summary statistics and deduce the general structure of the boxplot. You might begin by tabulating the number of samples that fall above, below, and within the box, including its boundaries, and by playing with the controls. Are there any apparent discrepencies between the boxplots and our summary statistics?