MATH 130: 4/14 WORKSHEET
STATISTICS: SUMMARIZING DATA

One use of statistics is to understand complicated data sets. Perhaps you have the complete data, perhaps all you can do is randomly sample to get a representative picture. Either way, you want to summarize what's going on to get something a human understand.

Types of data.

Quantitative data are those that can be represented numerically. *Qualitative* data are those which merely categorize.

- Age, weight, and income are examples of quantitative data.
- Gender, religion, and occuptation are examples of qualitative data.

It's common to use n for the number of entries in a data set and x_i for an individual value.

Averages.

The notions of averages we looked at within probability also make sense when talking about data samples.

- The *mode* is the value x_i which occurs the most often. If there are ties, there are multiple modes. This notion makes sense for both quantitative and qualitative data.
- The *median* is the value x_i which occurs in the middle of the data set. If n is odd, x_i is the value precisely in the middle. If n is even, x_i is the average of two closest to the middle. This only makes sense for quantitative data.
- The *mean* $\bar{x}$ is the sum of the values, divided by the number of values:

$$\bar{x} = \frac{\sum_i x_i}{n}.$$

($\sum x_i$ is shorthand for "the sum of the x_is". The capital Greek letter Sigma is chosen as 's' for sum.) This only makes sense for quantitative data.

Spreadsheets.

A useful tool for looking at data sets is a spreadsheet such as in Google Docs or Microsoft Office. You can both have your data set and calculate statistics of it.

- Data is usually organized so that each row represents the data associated with one entry, where each column represents a different measurement. For example, a data set consisting of info about different community colleges might have a row for each college, with the columns being values like total enrollment or average credit load.
- A formula can be typed in a cell by starting off the input with `=`. The functions `average`, `median`, and `mode` give the mean, median, and mode. The input to the functions should be a range of cells.
- A *range* of cells can be specified as the starting cell, then a colon, then the ending cell. For instance, `B2:B11` will select the cells from B2 to B11.
- Putting this together, typing `=average(B2:B11)` in a cell will show you the average of the values in the cells from B2 to B11.

Measures of spread-outness.

Averages measure the center of a data set, but you might also want to know how spread out it is. Is it closely clustered around the center? Or are the values far apart? For technical reasons, a useful measure of this is to look at the squares of distances from the center.

- The *variance* s^2 of a data sample (`var` in your spreadsheet) is an average of the squares of distances from the center:
$$s^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2.$$
- The *standard deviation* s (`stdev` in your spreadsheet) is the square root of the variance:
$$s = \sqrt{\frac{1}{n-1}\sum_i (x_i - \bar{x})^2}.$$

You divide by $n-1$ instead of n for reasons related to *Bessel's correction*, which is too difficult to get into in this class.

Sample questions

(1) Suppose you are interning for a market research company that conducts polls of the population. List three different pieces of qualitative data and three pieces of quantitative data you might ask for from respondents.
(2) Explain why it doesn't make sense to take the median of a sample of qualitative data.
(3) Explain why it doesn't make sense to take the mean of a sample of qualitative data.
(4) Calculate the mean, median, and mode of the following data set
$$2, 4, 4, 6, 6, 6, 8, 10, 12$$
(5) Calculate the mean of this data set, either by hand or using a calculator:
$$1.3, 4.2, 3.3, 5.6, 8.1, 10.3, 11.4, 3.8, 9.7, 7.7$$
(6) Calculate the variance and standard deviation of the above data set.
(7) To illustrate how standard deviation is a measure of spread-outness, calculate the standard deviations of the following data sets, each with three values:
$$9, 10, 11; \qquad 7, 10, 13; \qquad 4, 10, 16; \qquad 0, 10, 20.$$
(8) Go to `https://data.mass.gov` and download a dataset that's provided as a spreadsheet file. (Look for either a `.xls` or `.csv` file.) Use the spreadsheet functionality to calculate the mean and standard deviation of different columns. Report back on what the data set contained, what columns you looked at, and what the values of your statistics were. Write a few paragraphs explaining what the values you calculated show about the data.