

1. Statistics vocabulary - please define / describe and/or give example(s):
 - a. descriptive vs. inferential stats
 - i. Descriptive statistics only provide the values and properties of a dataset, while inferential statistics use sample groupings to make educated predictions about the population.
 - b. normal distribution
 - i. A normal distribution is symmetrical data that resembles a bell. Datasets modeled this way are required for inferential analysis to be meaningful and correct.
 - c. measures of center (central tendency): mean; median; mode
 - i. There are many different ways of finding the middle of data. Mean is the average of a dataset. The Median represents the middle-most value when the numbers are lined up in order. Mode is the value that occurs most often, which is the least correlative with center. Normal distributions all have the exact same values for mean, median, and mode.
 - d. fractiles: percentiles; quartiles; etc.
 - i. Fractiles are slices or samples of a dataset. Percentiles represent how a specific value compares to the total set, represented as a ranking out of 100. Quartiles are dynamic fractiles, grouping data into 4 subsets.
 - e. measures of spread (variability): range & IQR; variance & stdev
 - i. The interquartile range is the middle-most 50% of data, which is where the majority of data points lie. The range is the difference between the biggest and smallest values. Variance is a computed value that uses the mean to gauge how different each point is from it. Variance is used to find the standard deviation, which describes how close or far apart the dataset is.
 - f. empirical rule (a.k.a. 68.3-95.4-99.7)
 - i. When working with normally distributed data, 68.3% of data should fall between 1 standard deviation length away from the mean. The same goes for 2(95.4%) and 3(99.7%) standard deviations.
 - g. other measures: modality; skewness
 - i. Modality refers to the shape of the graph of data, and if there are any peaks or outliers. Skewness refers to the slope of the graph of data, and which could vary based of the distance of the mean from the first and last values.
 - h. data types in stats
 - i. Data types are descriptive representations of the underlying value, and display what is being quantified.
 - i. categorical: nominal; ordinal
 - i. Categorical data is representative of a characteristic of the subject, and are usually qualities. Nominal data is categorical data that has no underlying structure, and is often represented in any order. Ordinal data is qualitative rankings of data, like stars, or "good", "great", "excellent"
 - j. numerical: interval; ratio

- i. Numerical data is a quantitative representation of data. Interval data is equally spread out data points, that represent a quantity or scale. Ratios are also equally spread, though they compare two values together.
 - k. numerical data: discrete vs. continuous
 - i. Discrete data is data derived from taking it, like number of test results, or flips heads, while continuous data doesn't have a defined cap to the value.
2. Statistics methods - please show / describe a few examples of these activities:
- a. calculating measures of center and spread

1	1	Mean
2	1	6.08
3	2	Median
4	3	6
5	3	Mode
6	4	5
7	4	Range
8	5	9
9	5	
10	5	
11	5	
12	6	
13	6	
14	7	
15	7	
16	7	
17	8	
18	8	
19	8	
20	9	
21	9	
22	9	
23	10	
24	10	

25	10
----	----

- b. using the empirical rule
 - i. For using the empirical rule, you'd first need to find the standard deviation of the data set, and your dataset needs to have a normal distribution. The empirical rule is important in statistics as it allows you to group the set and make inferences. If a new piece of data were to be introduced into the set, it would be least likely to be outside of 3 standard deviations (as it comprises 0.3% of all data in the set) and most likely to fall within 2.
- c. classifying data
 - i. Normalizing data can help for better conclusions through classification. Classifying data can help more accurately make predictions based off of statistical regularities in the data.
- d. choosing the right methods of analysis / visualization
 - i. Displaying data is the main driving force behind statistics. Having good and easy ways for outsiders to understand data allows for everyone to draw conclusions from data. If displayed incorrectly, data may lead to confusion, or improper readings, which can further propagate bias or lead to improper health diagnoses.