Statistics Definitions

ID1050– Quantitative & Qualitative Reasoning

Population vs. Sample

- We can use statistics when we wish to characterize some particular aspect of a group, merging each individual's data into a corporate value. The definitions on this and the following slides are useful in describing statistical analyses.
- Population
  - This is the entire group that we wish to characterize using statistics.
- Sample
  - It is often not practical or possible to collect data from every individual in the population. In this case, we collect data from a sample of the population.
  - The sample is usually a small subset of the population. In the case that the sample is the entire population, we call this a census.
  - The hope is that the statistics of the sample accurately represent the statistics of the population, had we collected data from every member.

Parameter vs. Variable

- The aspect we wish to characterize may have a numerical value, or it may be just the number of members that have some property.
- Parameter
  - The aspect we wish to characterize from the population is called the parameter.
- Variable
  - When we actually collect data from a sample of the population, we call that data the variable.
  - If we do our job well, the variable from the sample will be similar (though it is not expected to be identical) to that parameter of the population.

Discrete & Continuous Data

- The data we collect is either a mathematical measurement or the count of members with some property. In either case, we have a number that has one of two properties.
- Discrete variable
  - This type of variable can only take on discrete values. Often, these are integer values. There are no instances of members with a value between the discrete values.
  - Examples are number of children in a family, the state you were born in, the number of faulty computer chips from a factory, or a ranking of high/medium/low probability of some risk.
- Continuous variable
  - With this type of variable, a measurement from the sample can take on any real value. Typically, the number falls within some reasonable range.
  - Examples are heights of people, battery voltages from a manufacturer, or annual income.

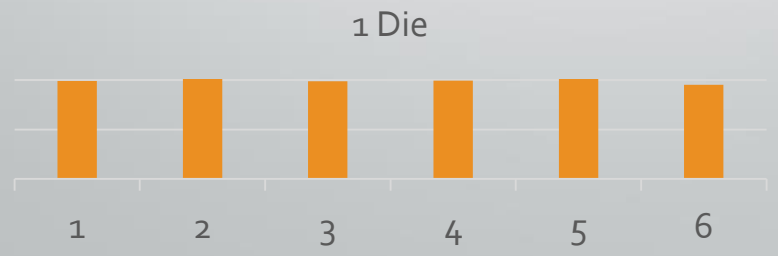
Scales (Data Type)

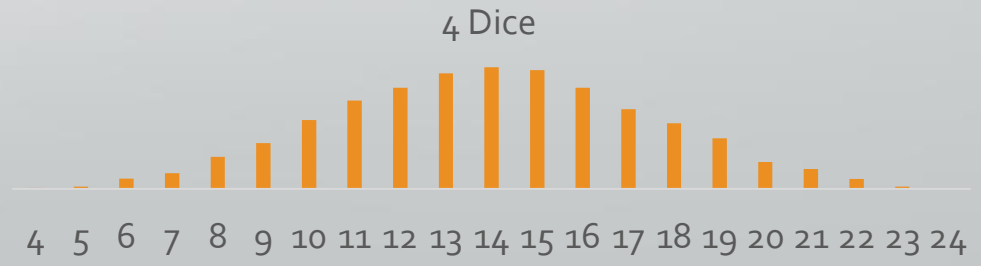
There are four types of classes or groups we can make from a population's parameter:

- Nominal scale – This classification is based on name only. There is no numerical value to the data.
  - Examples include *hair color*, *state of birth*, and *gender*.
- Ordinal scale – We use this scale when we rank or order a group based on how much of the characteristic they possess. Often the classes are *high/medium/low* or the like.
  - Examples: *dominance in monkeys* or *attractiveness of a painting*.
- Interval scale – If we have a numerical value, but what we define as 'zero' for that value is arbitrarily chosen, then the only meaningful comparison between values is the difference between them, or intervals.
  - The best example is *time*. We commonly count from a convenient starting point, but time existed before that point. So the interval between *start* and *stop* is what is relevant.
- Ratio scale – If the numerical value is a measurement of some continuous value that cannot be negative, then the 'zero' is not arbitrary, and meaningful comparisons between data can be made, such as 'how many times larger is a value' or 'what is half of this value'.
  - Some examples: *height or age of a person*, *corporate income*, *battery voltage*, etc.

Distributions

- For a large sample, we may see a particular distribution of members among the classes or categories. Some often-encountered distributions have names:
- Uniform distribution – The chances of a member being in any of the categories is equal. With a large sample, all classes should have approximately equal numbers.
  - The roll of a single six-sided die is expected to be uniformly distributed. The chances are equal that a roll will be any number between 1 and 6.
- Normal distribution – The values measured tend to be near a central value. There are more members near the middle, and fewer members at values farther from the middle.
  - The height of a population is expected to be normally distributed. There is an average height, and most people are near this value. There are few exceptionally tall or exceptionally short individuals. This is due to physiological constraints of the human body.
- Uniform and normal distributions are idealized distributions; real data may or may not approximate one of these distributions (or one of several other named distributions.)
- A single die should be uniform. Two dice tend to give more values near 7 and fewer near 2 and 12. The more dice that are rolled, the more the distribution begins to look like a normal distribution.





Measures of Central Tendency

- When data tends toward the middle, we have to ask "what does 'middle' mean?" There is no one definition of central tendency; instead, what the middle means may depend on context or desired use.
  - What is the middle of the title 'One Fish, Two Fish, Red Fish, Blue Fish'? Is it the middle letter (R), the middle word (there are two), or the word 'Fish', which appears the most?
- There are three common measures of central tendency:
  - Mean – Sometimes called 'average', this is the sum of the data divided by the number of data values.
  - Median – This is the middle element of an ordered list.
  - Mode – This is the class that is most represented.

Measures of Spread and Measure of Symmetry

- Once we have identified the middle of some data, we may then ask "how far is the data spread about the middle?".
  - There are two common measures of the spread: Variance and standard deviation.
  - If these values are large, the data is spread wide, and if they are small, the data occupies a narrow range.
  - The two differ from each other only slightly; one is the square of the other.
- After considering the spread, we may ask "how symmetric is the spread of the data about the middle?".
  - The value we will used for this measure is called *skewness*.
  - A value near 'zero' indicates high symmetry; the data tails off similarly to the left and to the right.
  - An absolute value near 'one' or higher indicates that one tail is distinctly longer than the other. The sign of the skewness indicates the direction of the tail (- = left, + = right).

Conclusion

- When performing statistical analysis, some vocabulary is helpful
  - Population vs. Sample – What is the group we are analyzing?
  - Parameter vs. Variable – The name for the data we collect from a population or sample.
  - Discrete & Continuous – The data can be the number of members, it could consist of integers, or it could be any real numerical value.
  - Nominal, Ordinal, Interval, and Ratio Scales – How do we categorize the collected data?
  - Uniform & Normal Distributions – How is the data spread among the groups?
  - Statistical Variables – The numerical results from analyzing the data
    - Mean, Median, & Mode – Measures of the middle
    - Standard Deviation & Variance – Measures of the spread
    - Skewness – Measure of symmetry