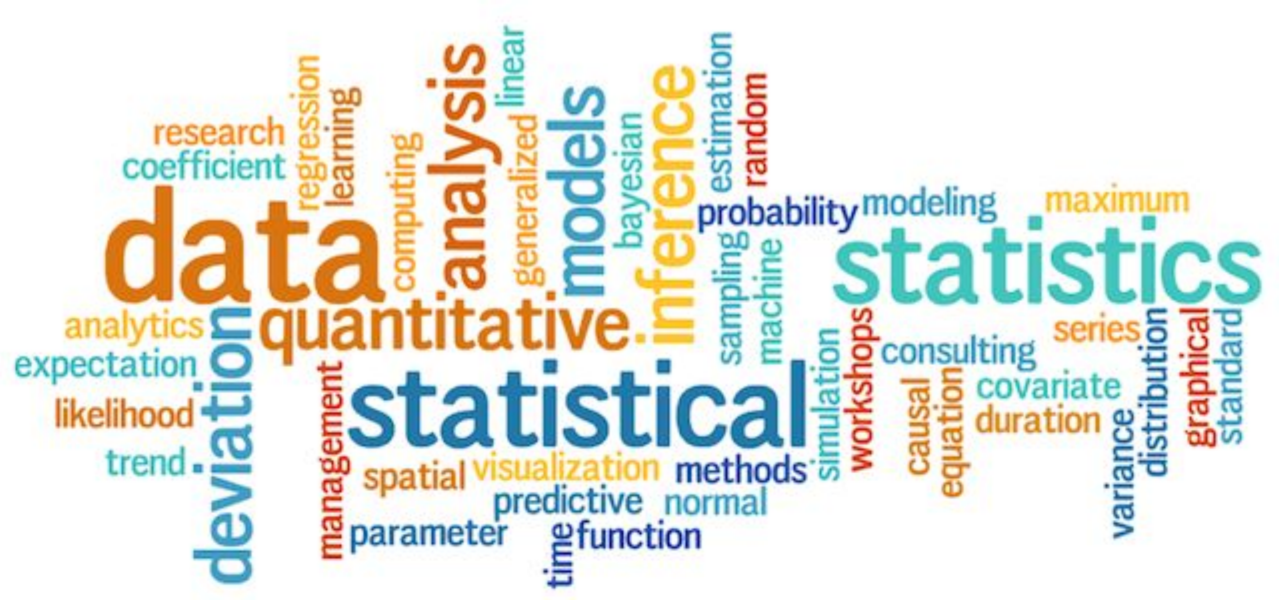
Rudimentary Statistics

The Basics

SoftStack Factory
2018

Types of Statistics



Word cloud credit: Cal. State University

Statistics

Definition:

- Statistics is the field of study concerned with
  Collecting,
  Analyzing,
  and Communicating data.

Statistics

A well-trained statistician is able to:

- draw conclusions from data through the process of statistical analysis
- is also able to communicate their findings to non-statisticians
- allowing others to make better decisions.

Two “Activities” in Statistics

Descriptive Statistics

- Describing data.

Inferential Statistics

- Learn from data.

Descriptive Statistics

- Encompasses the many tools used to “Describe” data.
- Finding averages, variance, standard deviations, etc.
- Visualizations.

Inferential Statistics

- Encompasses the many tools used to “learn” from data.
- “Infer” Definition: To draw conclusions after a period of reasoning.
- Finding relevant correlations / causations.

Data Types Key Terms

Continuous - Data that take on any value in an interval.

Discrete - Data that can only take on integer values, such as counts.

Categorical - Data that can take on only a specific set of values representing a set of possible categories.

Binary - Categorical data with just two categories of values (0/1, true/false)

Ordinal - Categorical data that has an explicit order (1,2,3,4,5)

SoftStack Factory 2018

Central Tendency

Mean vs. Median vs. Mode

- MEAN
  - The *average* value of a set.
- MEDIAN
  - The *middle* value of an ordered set.
- Mode
  - The most common value, the value that appears the most in a set.

Central Tendency

Mean

- Average: (sum of all values of set) DIVIDED by (# of values in set.)
- Given: X = {x1, x2, x3, … , xn}
- Mean = sum(x1, x2, x3, … , xn) / (n)

Central Tendency

Median

- Middle Value: Set must be an *ordered* set. I.e: ascending or descending order.
- Given: An ordered set X = {x1, x2, x3, … , xn} with n entries.
- If n = odd:
    - If n is odd, there IS a clear middle index.
    - Median = Value at middle index
- If n = even:
    - If n is even, there IS NOT a clear middle index, but two...
    - Median = Average between the two middle values.

Sample Vs. Population MEAN

Sample MEAN

- The MEAN of the SAMPLE
- After a sufficient number of trials,
  - the sample MEAN can be thought of as the EXPECTED VALUE

Population MEAN

- The *real-world* MEAN of the POPULATION you've sampled.

★ Sample mean can only APPROXIMATE the population MEAN.

Sample Mean - Expected Value

- The sample mean is is equivalent to the expected value of an experiment.
- Let’s say we are performing a series of coin flips:
    - We want to find the sample mean for the coin coming up HEADS
    - Well, first… what is the expected value?
        - We know there are 2 possible outcomes: H or T, therefore expected value for H is 50%
    - As we begin the experiment we’ll notice that our sample mean may be above or below our expected value
    - However, As we increase the number of coin flips, the sample mean will begin to more closely approximate the expected value.

Data Spread

Variance vs. Standard Deviation

VARIANCE (σ2)

- Average distance between points?

STD DEVIATION (σ)

- Square root of variance.

Variance

- The variance is a measure of spread of the set, it uses the mean to calculate the average distance between the points of a set.
- Calculation: Take each difference (Xi - MEAN). Square it. Then average the result:

Standard Deviation

- The standard deviation is also a measure of spread of the set. It is the square root of the variance.
- Expressed in same units as the mean
- Can generally be used interchangeably with variance to describe the spread of a set, but only if ALL other relevant calculations ALSO refer to the standard deviation over variance.