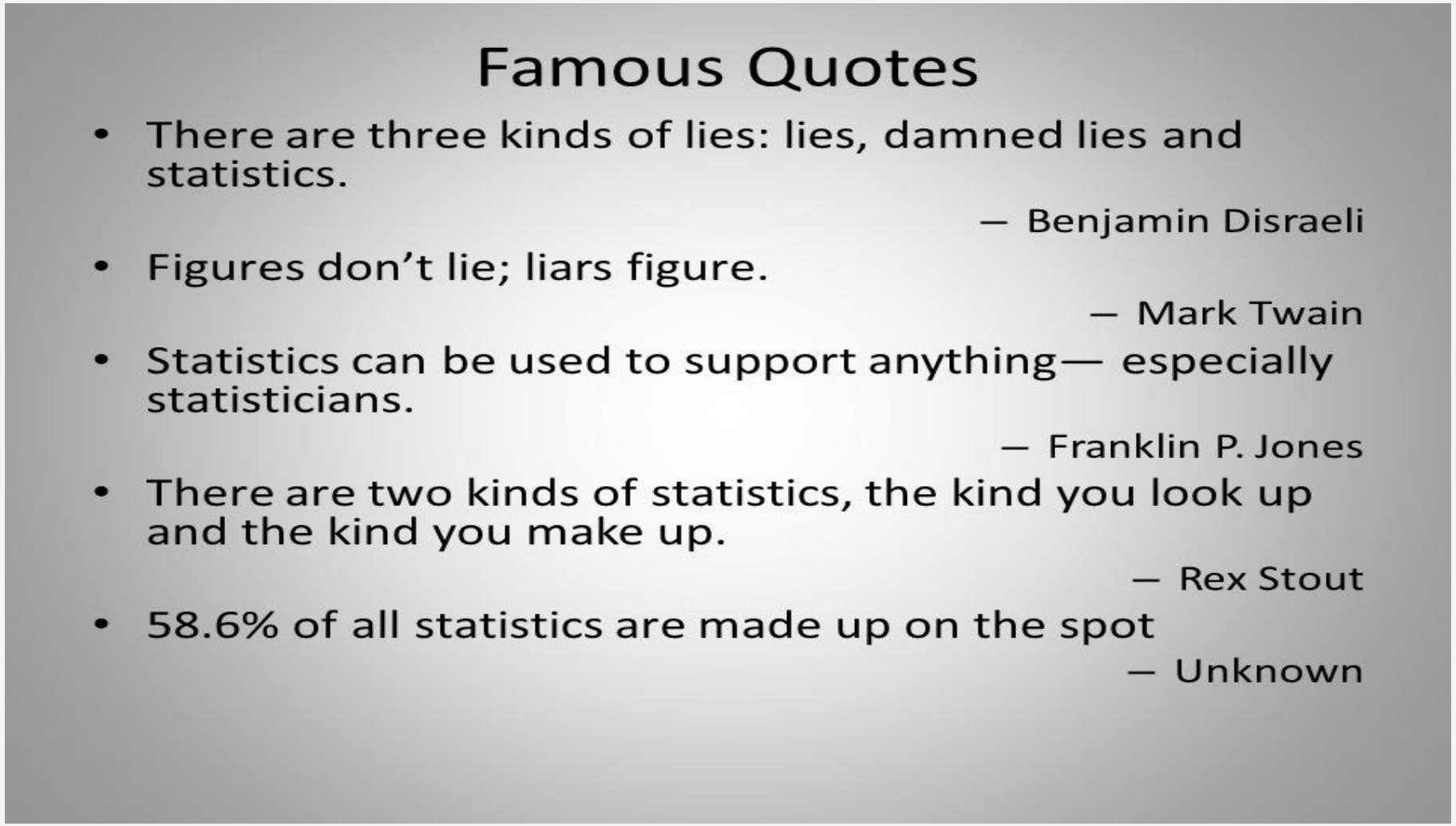
Probability and Statistic Review



Then why do we still care?

A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician". – Josh Wills on Quora

http://www.mastersindatascience.org/careers/data-scientist/

Why Statistics Again?

- The world is not deterministic, in other words we need to deal with randomness
- We only live once. What we observe is only ONE realization of many possibility. In another universe, you may be the professor while I may be the student
- However, if there are some underlying truth (such as the sun will always rise from the east or a human being will not grow more than 8 feet), you will see pretty much the same thing again and again if you can make multiple observations from the same underlying mechanism.
- Observations will have a lot of noise. The Signal to Noise ratio will be extremely important.
- Question to ask is whether the conclusion you draw from your observation is statistically significant.

Importance of Statistics

- Cannot just say I am a genius. I know what I am doing. Believe me!
- Statistics is the language of Randomness
- Can only qualify your statements with certain probability
- Interested only in conclusions that are Statistically Significance
- This is what Data Analytics is all about.
- Science behind drawing meaningful conclusions from observations

In fact … Data science first appeared in a statistics paper

In 2001, William S. Cleveland published a research paper that coined the term "Data Science" the first time

Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics

Article *in* International Statistical Review 69(1) · March 2001 *with* 410 Reads

An action plan to enlarge the technical areas of statistics focuses on the data analyst. The plan sets out six technical areas of work for a university department and advocates a specific allocation of resources devoted to research in each area and to courses in each area. The value of technical work is judged by the extent to which it benefits the data analyst, either directly or indirectly. The plan is also applicable to government research labs and corporate research organizations. 1 Summary of the Plan This document describes a plan to enlarge the major areas of technical work of the field of statistics. Because the plan is ambitious and implies substantial change, the altered field will be called "data science." The focus of the plan is the practicing data analyst. A basic premise is that technical areas of data science should be judged by

Computer Science + data mining = Make statistics a lot more technical
= Data Science

Importance of Statistics

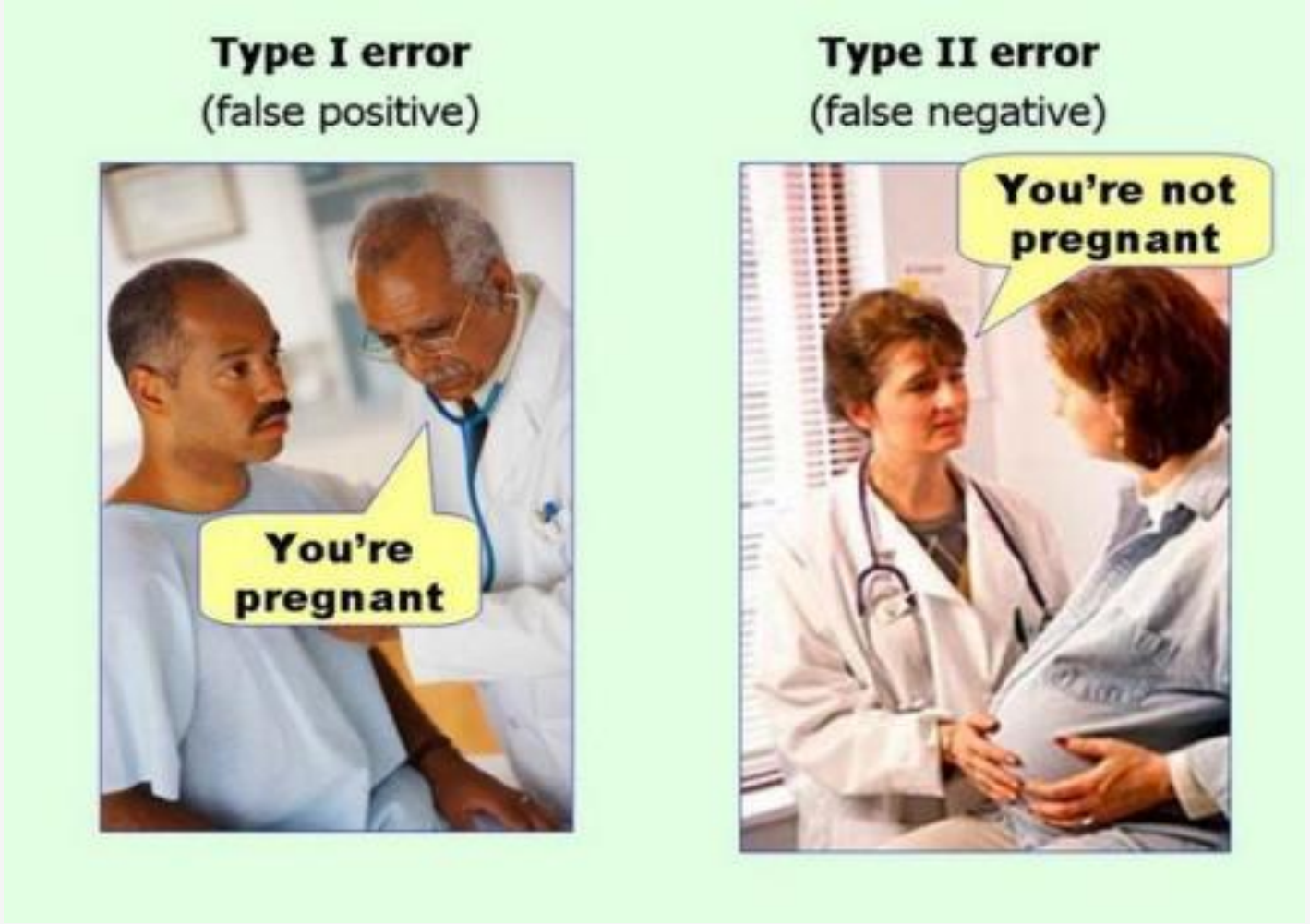
Some even prefers referring the study as Statistical Learning over Machine Learning

From Introduction To Statistical Learning:

> Statistical learning refers to a set of tools for modeling and understanding complex datasets. It is a recently developed area in statistics and blends with parallel developments in computer science and, in particular, machine learning.
>
> Since that time, inspired by the advent of *machine learning* and other disciplines, statistical learning has emerged as a new subfield in statistics, focused on supervised and unsupervised modeling and prediction.

Know your Statistics (terms you that need to understand well)

- Mean, Standard Deviation
- Distribution
- Law of large numbers
- Statistical Significance
- Survival Bias
- Bias vs Variance



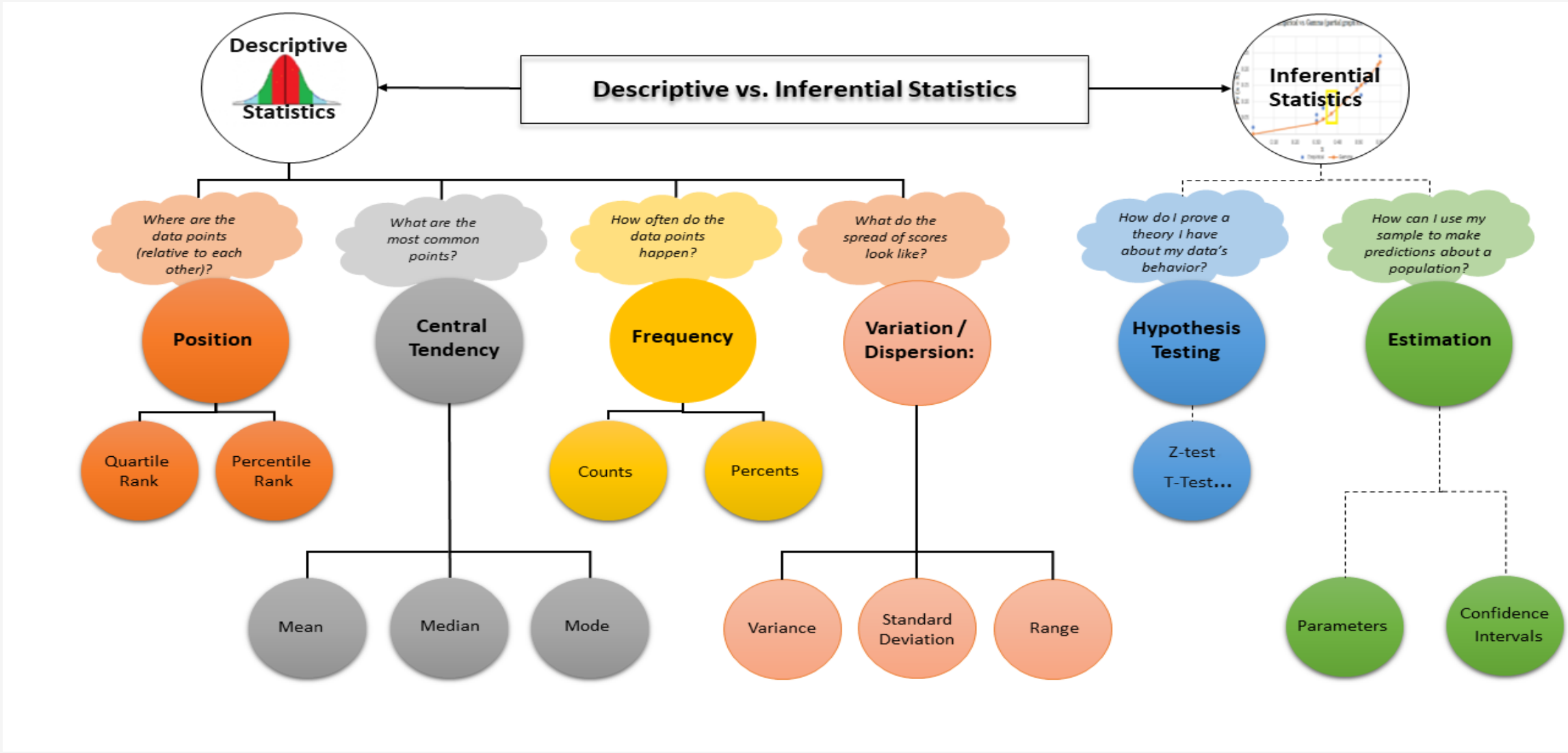
Two kind of Statistics

Descriptive Statistics vs Inferential Statistics

- Descriptive Statistics consists of organizing and summarizing data.
  - Mean, Standard Deviation, Skew, Quantile, Ranks

- Inferential Statistics (Predictive Statistics) consists of using data you have collected to form conclusions
  - Hypothesis testing
  - Estimation (Use sample mean to predict population mean)

Two kind of Statistics



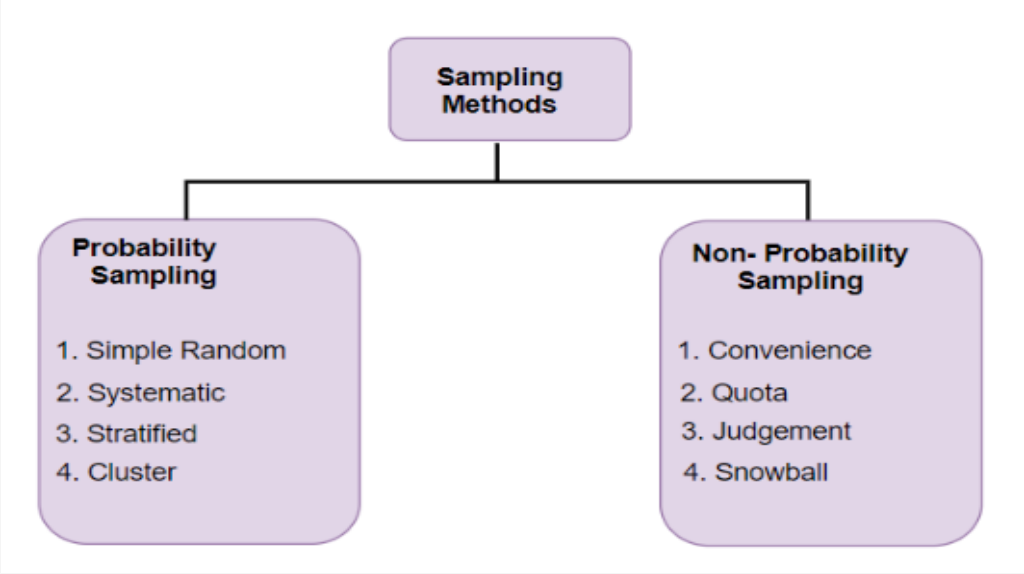
Sampling

Population vs Sample

- The population is the entire group you are interested in studying.
- A sample is a subset of the population. That is to say, it is a select group of information taken from a population.

Sampling Methods

- Random Sampling
- Stratified Sampling
- Systematic Sampling
- Convenience Sampling



https://www.analyticsvidhya.com/blog/2019/09/data-scientists-guide-8-types-of-sampling-techniques/?utm_source=feedburner&utm_medium=email&utm_campaign=Feed%3A+AnalyticsVidhya+%28Analytics+Vidhya%29

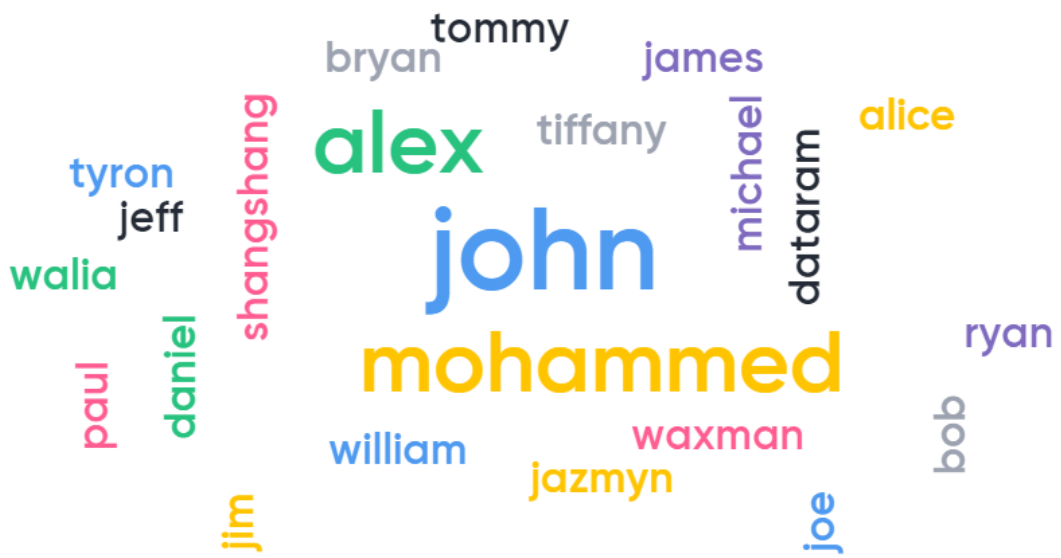
Sampling

Let’s answer the question:
What is the most common name in US?

Sampling

Go to **www.menti.com** and use the code **12 12 64**

To help answer the question of what is the most common name in US, please submit your first name.



Importance of Correct Sampling

What is the problem of what we just did?

Important consideration in Sampling

- Systematic Bias
- Survival Bias
- Size of the samples (cost)

Excellent Online Resource for Statistics Review

Assume you have the Math 241 (Probability and Statistics) pre-requisites

http://www.statisticslectures.com/topics/statistics/ is an excellent online resource for quick review

- Basics of Probability
- Discrete and Continuous Random Variables
- Probability Distribution
- Mean and Expected Value
- Law of Large Numbers
- Central Limit Theorem
- Normal Distributions
- Sampling
- Hypothesis Testing
- Type I and Type II Errors
- P-value
- One-tail tests
- Conditional Probability
- Bayes Rules

Descriptive Statistics

Central tendency refers to the measure used to determine the center of a distribution of data. It is used to find a single score that is most representative of an entire data set

Mean, Median and Mode

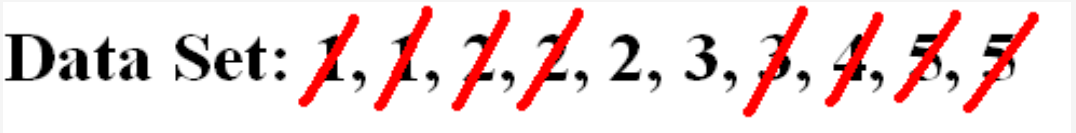
- Mean is the most common single statistics (number) to describe an entire data set.
- However, it is very sensitive to outliers.
  - Example: mean of the age of students in a college class: { 18, 19, 19, 17, 60}
- Median is the number that lies in the middle after the data set is sorted.
- Mode is simply the most frequently occurring value

Example Dataset:

1, 1, 2, 2, 2, 3, 3, 4, 5, 5

$$\bar{X} = \frac{\sum x}{n} \qquad \frac{1+1+2+2+2+3+3+4+5+5}{10} = \frac{28}{10} = 2.8$$

Mean is 2.8
Median is (2+3)/2 = 2.5
Mode is 2 because it occurs the most frequently

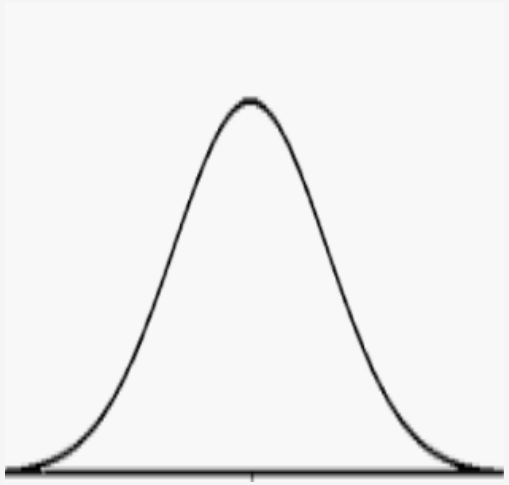


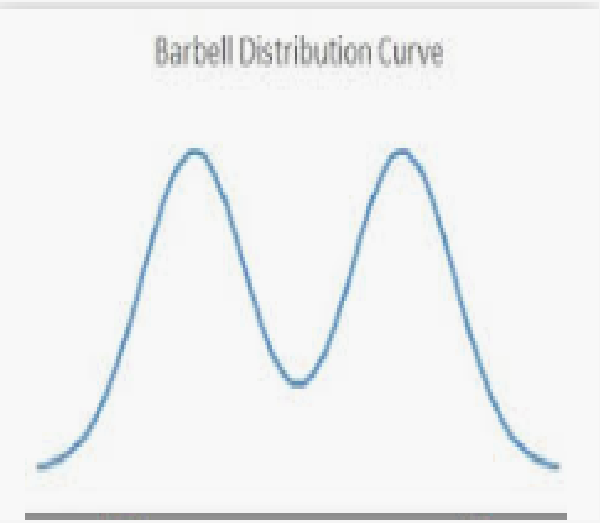
Descriptive Statistics

Two very different distribution could have the same mean and median

Example: dataset1 { 16, 18, 18, 18, 18, 18, 18, 20 }
dataset2 { 8, 8, 8, 18, 18, 28, 28, 28 }

So, we need more descriptive statistics to describe a distribution





Standard Deviation, Skew, Kurtosis

Standard Deviation or Variance

Dispersion refers to how spread out a data set is about the mean.

Variance and Standard Deviation are two measures of dispersion within a data set.

Example Dataset: {1, 2, 2, 3, 4, 5}

μ denotes the mean

$$3.35 + 0.69 + 0.69 + 0.03 + 1.37 + 4.71 = 10.84$$

$$\sigma^2 = \frac{10.84}{6} = 1.81$$

Population Variance
$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$

Population Standard Deviation
$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$

x	μ	x - μ	$(x - \mu)^2$
1	2.83	1 – 2.83 = (-1.83)	(-1.83)^2 = 3.35
2	2.83	2 – 2.83 = (-0.83)	(-0.83)^2 = 0.69
2	2.83	2 – 2.83 = (-0.83)	(-0.83)^2 = 0.69
3	2.83	3 – 2.83 = (0.17)	(0.17)^2 = 0.03
4	2.83	4 – 2.83 = (1.17)	(1.17)^2 = 1.37
5	2.83	5 – 2.83 = (2.17)	(2.17)^2 = 4.71

Second Equivalent Formula For Variance

Population Variance

$$\sigma^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{N}}{N}$$

Population Standard Deviation

$$\sigma = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{N}}{N}}$$

Figure 3.

ı this problem, N is the size of our data set(6). The other values are calculated like this:

$$\sum x^2 = 1^2 + 2^2 + 2^2 + 3^2 + 4^2 + 5^2 = 59$$

$$(\sum x)^2 = (1 + 2 + 2 + 3 + 4 + 5)^2 = (17)^2 = 289$$

After plugging in all the values, we again find a variance of 1.81, and a standard deviation of 1.35.

Sample Variance vs Population Variance

N denotes the Population size, n is the sample size

Sample Variance

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$

Sample Standard Deviation

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

Population Variance	Population Standard Deviation
$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$	$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$

Sample Variance

$$s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n - 1}$$

Sample Standard Deviation

$$s = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n - 1}}$$

Population Variance

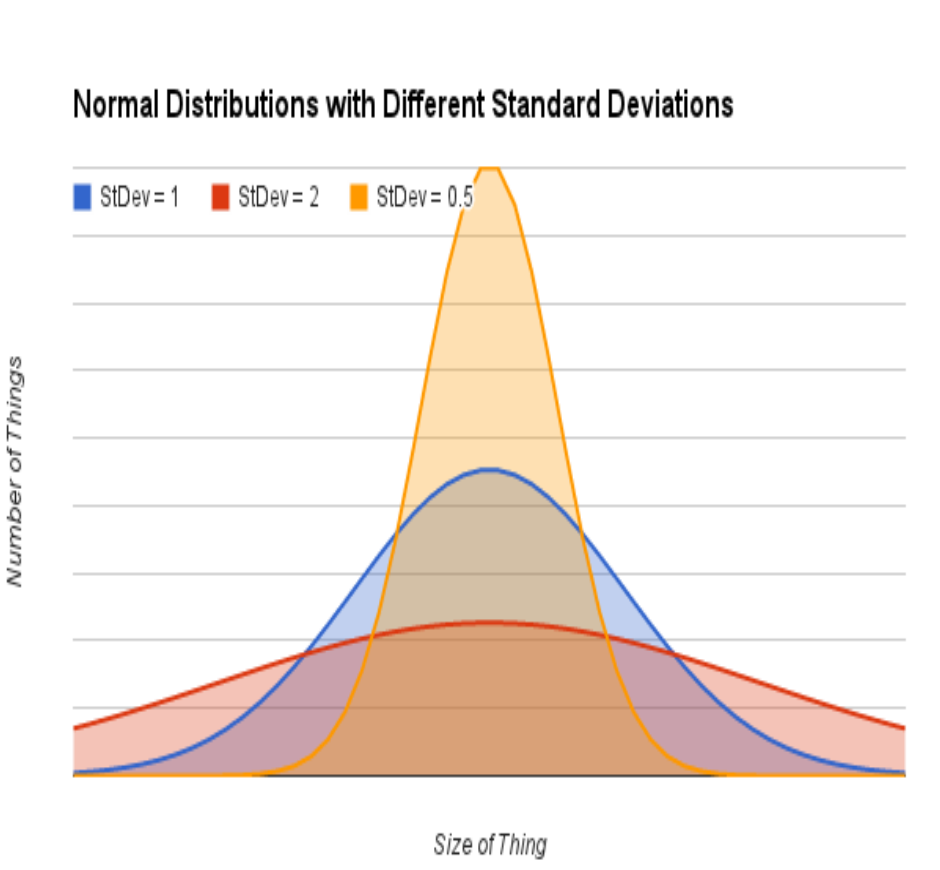
$$\sigma^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{N}}{N}$$

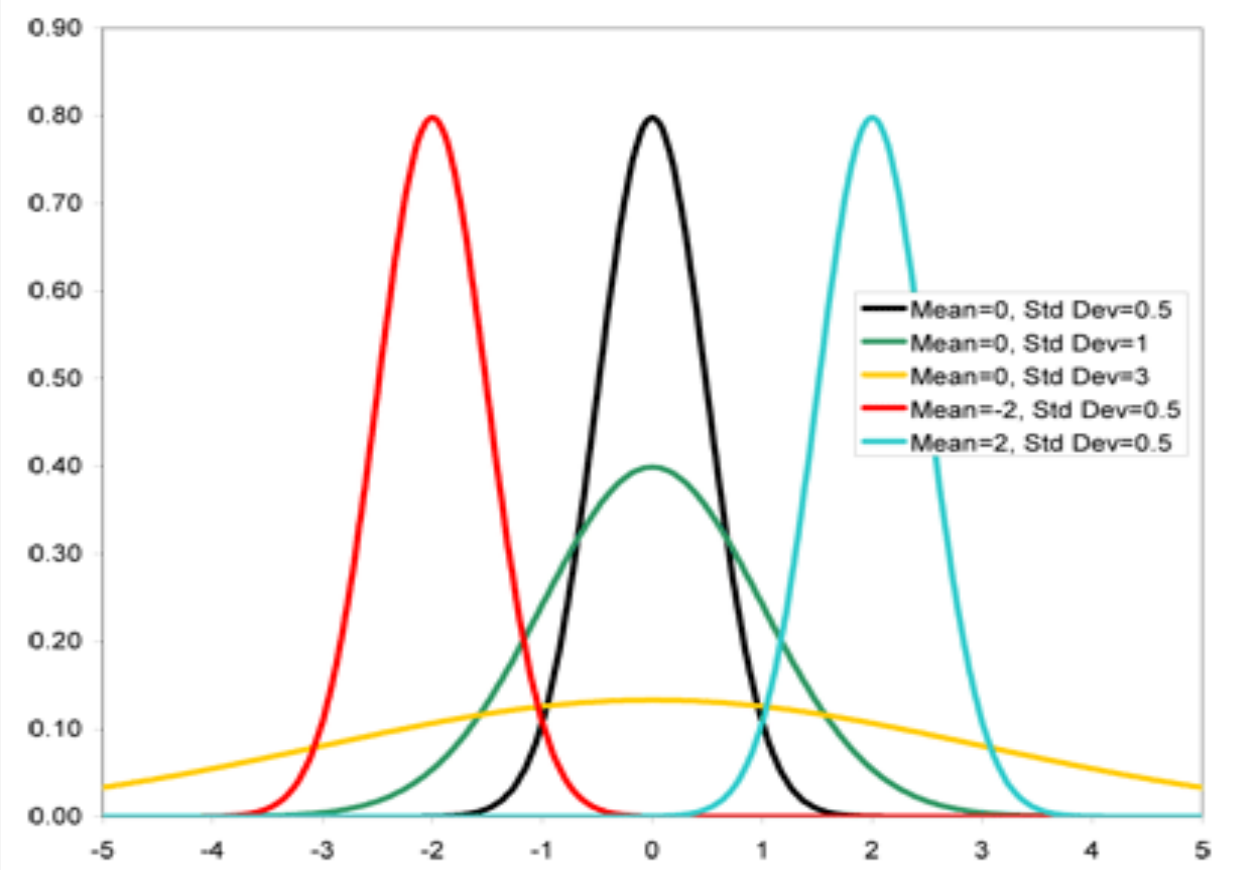
Population Standard Deviation

$$\sigma = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{N}}{N}}$$

Standard Deviation

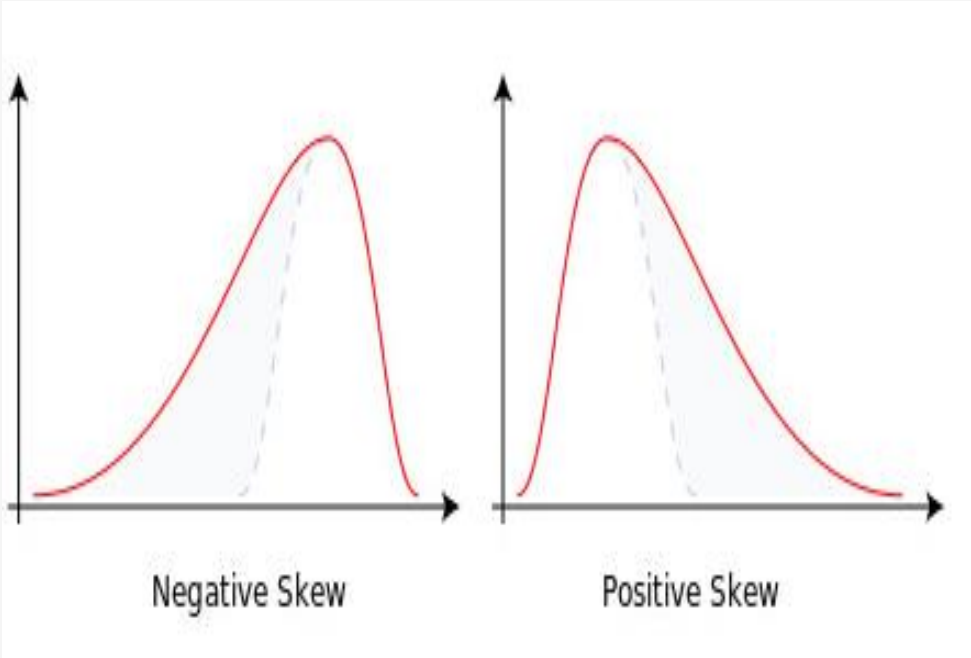
Standard Deviation as a metric to describe how “wide” or “spread-out” the distribution is.

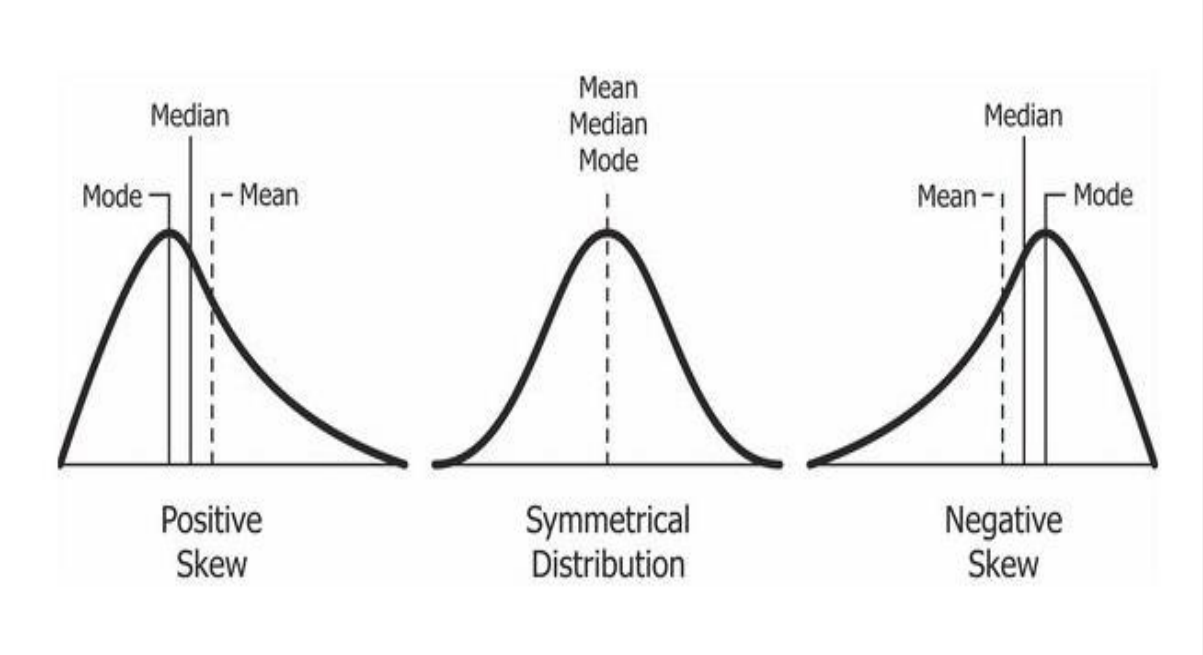




Skew

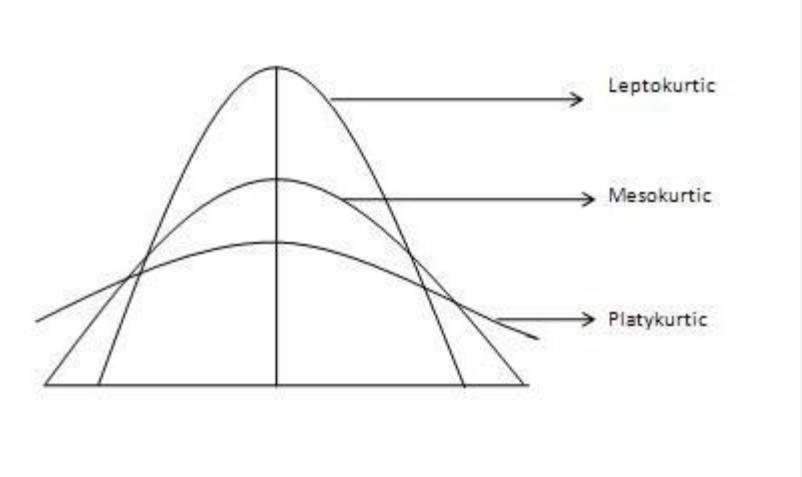
Skewness is the degree of distortion from the symmetrical bell curve or the normal curve. It measures the lack of symmetry in data distribution. It differentiates extreme values in one versus the other tail. A symmetrical distribution will have a skewness of 0.





Kurtosis

Kurtosis, on the other hand, refers to the pointedness of a peak or the tails in the distribution curve. The main difference between skewness and kurtosis is that the former talks of the degree of symmetry, whereas the latter talks of the degree of peakedness (or tailedness) in the frequency distribution.



Mesokurtic: This distribution has kurtosis statistic similar to that of the normal distribution. The standard normal distribution has a *kurtosis of three.*
Leptokurtic (*Kurtosis > 3*): tails are fatter, has more outliers. Peak is higher and sharper than Mesokurtic
Platykurtic: (*Kurtosis < 3*): tails are thinner, has less outliers than the normal distribution. The peak is lower

In Pandas the kurtosis definition is slightly different. Normal distribution has a zero Kurtosis. Leptokurtic kurtosis is > 0 and Platykurtic kurtosis is < 0

Appendix: Statistics Review

Descriptive and Inferential Statistics http://www.statisticslectures.com/topics/descriptiveinferential/

Population vs Sample http://www.statisticslectures.com/topics/samplingmethods/

Parameters, Statistics and Sampling Errors http://www.statisticslectures.com/topics/parametersstatistics/

Distribution of Sample Mean http://www.statisticslectures.com/topics/distributionsamplemean/

Mean and Expected value of a probability http://www.statisticslectures.com/topics/meanexpectedvaluediscrete/

Variance and Standard deviations http://www.statisticslectures.com/topics/variancestandarddeviationdiscrete/

Law of Large Numbers, Central Limit Theorem http://www.statisticslectures.com/topics/centrallimittheorem/

Skew and Kurtosis explained:
https://keydifferences.com/differences-between-skewness-and-kurtosis.html
https://codeburst.io/2-important-statistics-terms-you-need-to-know-in-data-science-skewness-and-kurtosis-388fef94eeaa
https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-whats-with-the-different-formulas-for-kurtosis/

Statistics Review

Watch the Recommended Statistics Lectures as much as you could

http://www.statisticslectures.com/topics/statistics