Review of Standard Deviation, Skew and Kurtosis

Standard Deviation
large SD => wide distribution => heterogeneity
Small SD => narrow distribution => homogeneity

Skew
Positive => lots of bigger values
Negative => lots of smaller values

Kurtosis
Positive => More outliers than normal distribution
Negative => Less outliers than normal distribution

The height distribution taken from Computer Science class in Queen College will have a mean __similar__ (higher/lower/similar) than the whole college and a _________ (positive/zero/negative) skews

The height distribution taken from the basketball Team in Queen College will have a mean _higher___ (higher or lower) than the whole college and a ___positive or zero______ (positive/zero/negative) skews

mean larger (in terms of x-axis) than median means POSITIVE skew.

The height distribution taken from Computer Science class in Queen College will have a mean _higher___ (higher or lower) than the whole college and _____positive______ (positive/zero/negative) skews if we know many are also in the basketball Team

Questions

What are the factors that drive house prices?

Questions

What are the factors that drive house prices in a city?

Mortgage Rates
Unemployment Rates
Local School performance
…

Questions

How would you determine which factors are really important in 5 minutes (ie without developing any models)?

Covariance and Correlation

Covariance measures the linear relationship between two variables.

- **Positive covariance**: Indicates that two variables tend to move in the same direction.
- **Negative covariance**: Reveals that two variables tend to move in inverse directions

Covariance can range from negative infinity to positive infinity.

Correlation is the scaled measure of covariance. It is dimensionless. In other words, the correlation coefficient is always a pure value and not measured in any units.

Correlation is between -1 and +1

Correlation between -1 and +1

Check link below

ρ(X,Y) – the correlation between X and Y
Cov(X,Y) – the covariance between X and Y
σ_X – the standard deviation of X
σ_Y – the standard deviation of Y

$$\mathrm{Cov}\,(X, Y) = \frac{\sum(X_i - \overline{X})(Y_j - \overline{Y})}{n}$$

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

https://corporatefinanceinstitute.com/resources/knowledge/finance/covariance/

Covariance and Correlation

Measures the linear portion (slope) of the data

Pearson product moment correlation

The Pearson correlation evaluates the linear relationship between two continuous variables. A relationship is linear when a change in one variable is associated with a proportional change in the other variable.

For example, you might use a Pearson correlation to evaluate whether home price increase in a city is related to the unemployment rate in that area.

monotonic - whether a function preserves its order or reverses the given order.

Spearman rank-order correlation

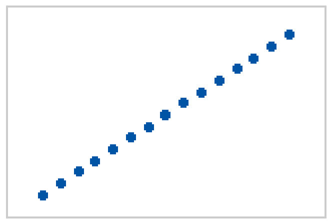
The Spearman correlation evaluates the monotonic relationship between two continuous or ordinal variables. In a monotonic relationship, the variables tend to change together, but not necessarily at a constant rate. The Spearman correlation coefficient is based on the ranked values for each variable rather than the raw data.

Spearman correlation is often used for ordinal variables. For example, you might use a Spearman correlation to study how the order in which employees complete a test exercise is related to the months they have been employed.

In a scatterplot, Pearson Correlation coefficients measure linear relationship while Spearman is more concerned on whether the relationships is monotonic or not.**:IMPORTANT —> pearson cares for linear relationship while Spearman cares for monotonic**

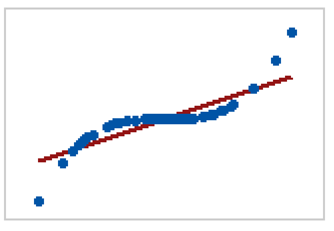
Pearson vs Spearman Correlation

Fig 1



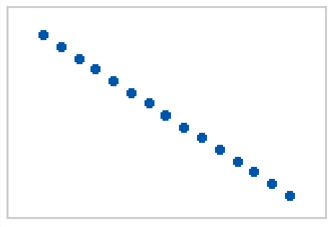
Pearson: +1
Spearman: +1

Fig 2



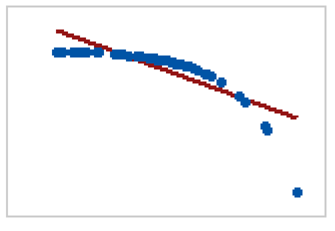
Pearson: ?
Spearman: ?

Fig 3



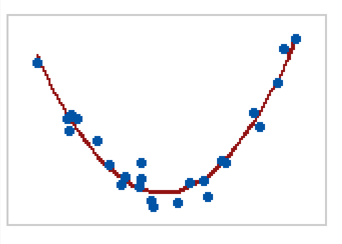
Pearson: -1
Spearman: -1

Fig 4



Pearson: ?
Spearman: ?

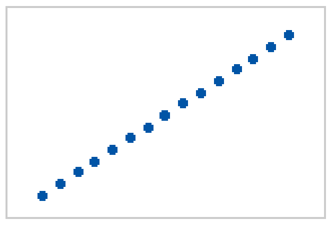
Fig 5



Pearson: ?
Spearman: ?

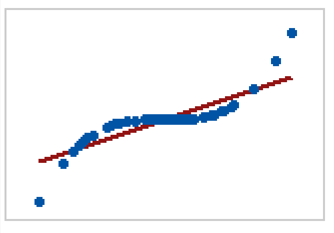
Pearson vs Spearman Correlation

Fig 1



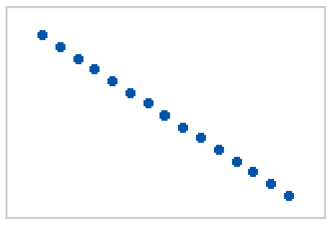
Pearson: +1
Spearman: +1

Fig 2



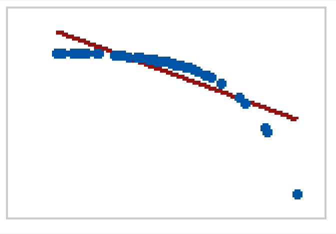
Pearson: +0.85
Spearman: +1

Fig 3



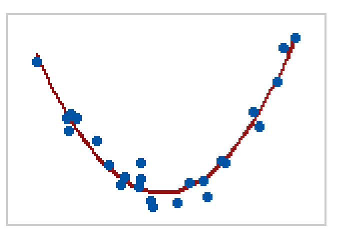
Pearson: -1
Spearman: -1

Fig 4



Pearson: -0.85
Spearman: -1

Fig 5



Pearson: 0
Spearman: 0

IMPORTANT: This means that 2 variables w/ zero/low correlation CAN be dependent on one another and it also means that 2 variables that have correlation can still be INDEPENDENT

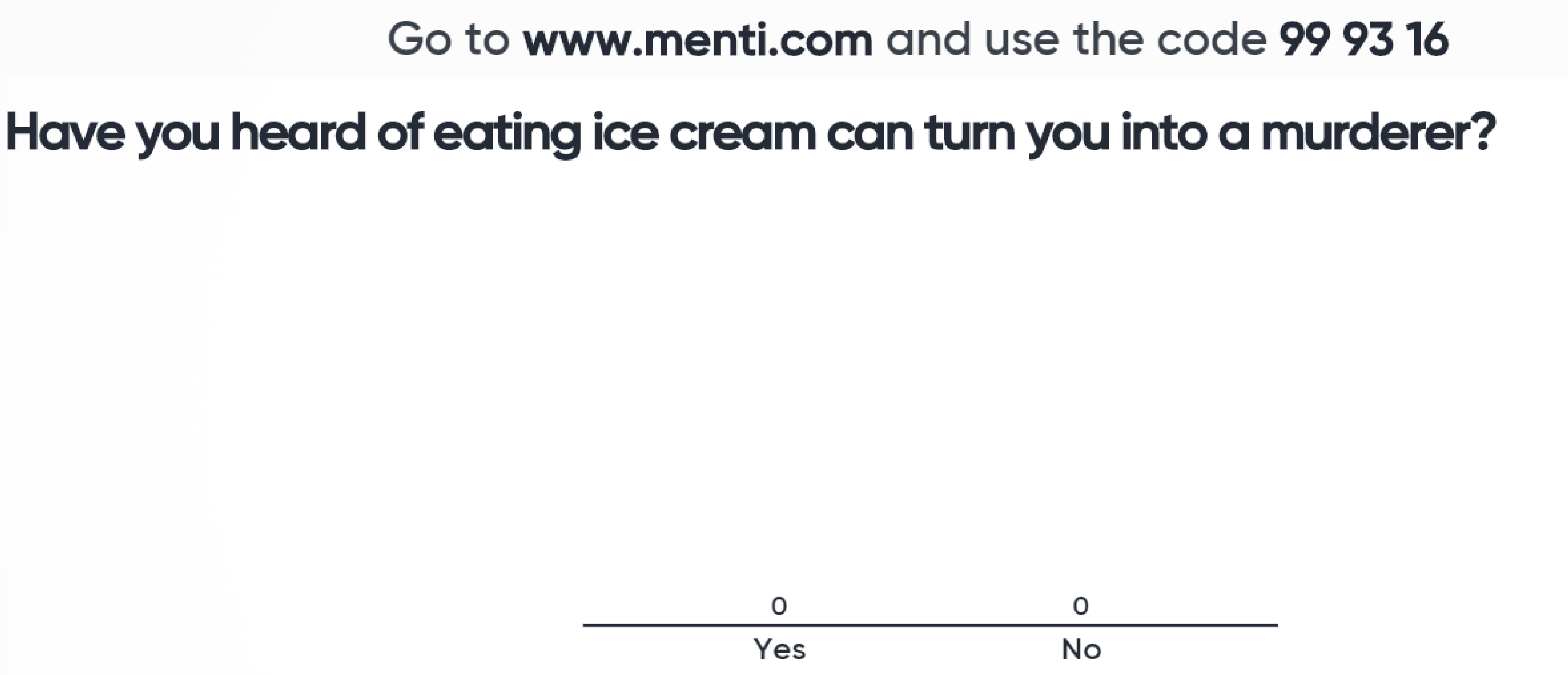
Zero correlation does not mean the variables are independent

Low correlation does not mean there is no dependence between two variables

CHECK LINK BELOW FOR EXAMPLES and explianation

https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/supporting-topics/basics/a-comparison-of-the-pearson-and-spearman-correlation-methods/

Questions



Go to **www.menti.com** and use the code **99 93 16**

Have you heard of eating ice cream can turn you into a murderer?

Correlation and Causation

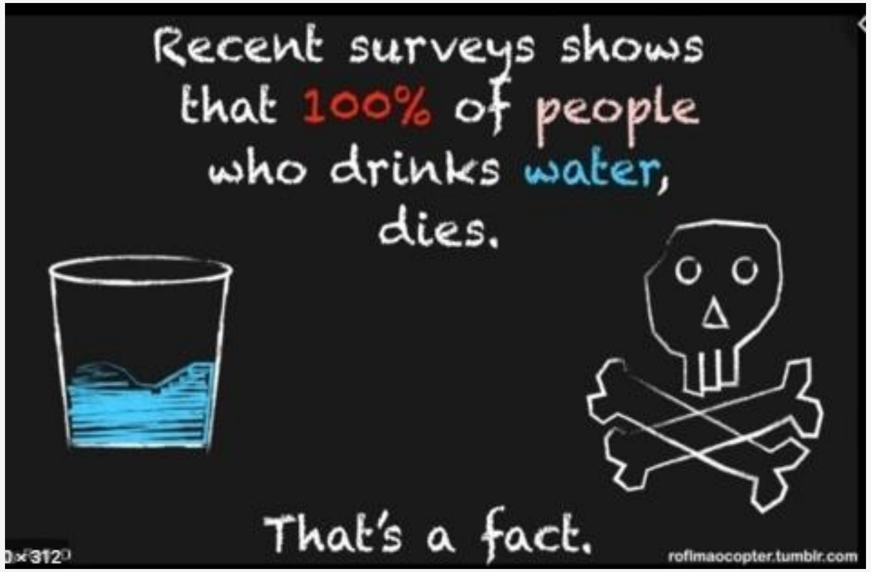
Causation will lead to high correlation, but high correlation may not necessarily imply causation relationship

this situation is just a concidence

Classic Example: Murder rates goes up when ice cream sales go up

The rates of violent crime and murder have been known to jump when ice cream sales do. But, presumably, buying ice cream doesn't turn you into a killer (unless they're out of your favorite kind?)

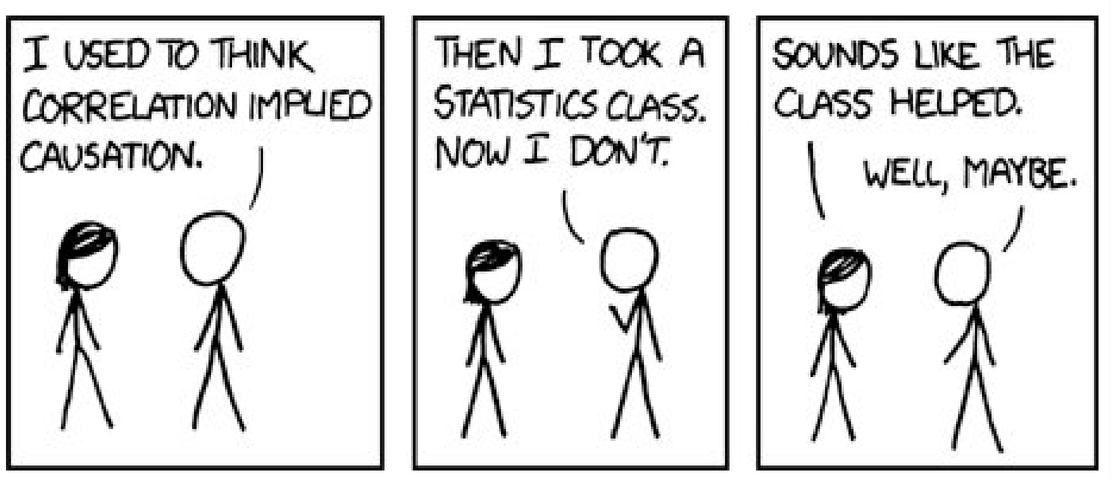
But, correlation is still one good tool to identify driving factors.


Recent surveys shows that 100% of people who drinks water, dies.
That's a fact.
roflmaocopter.tumblr.com

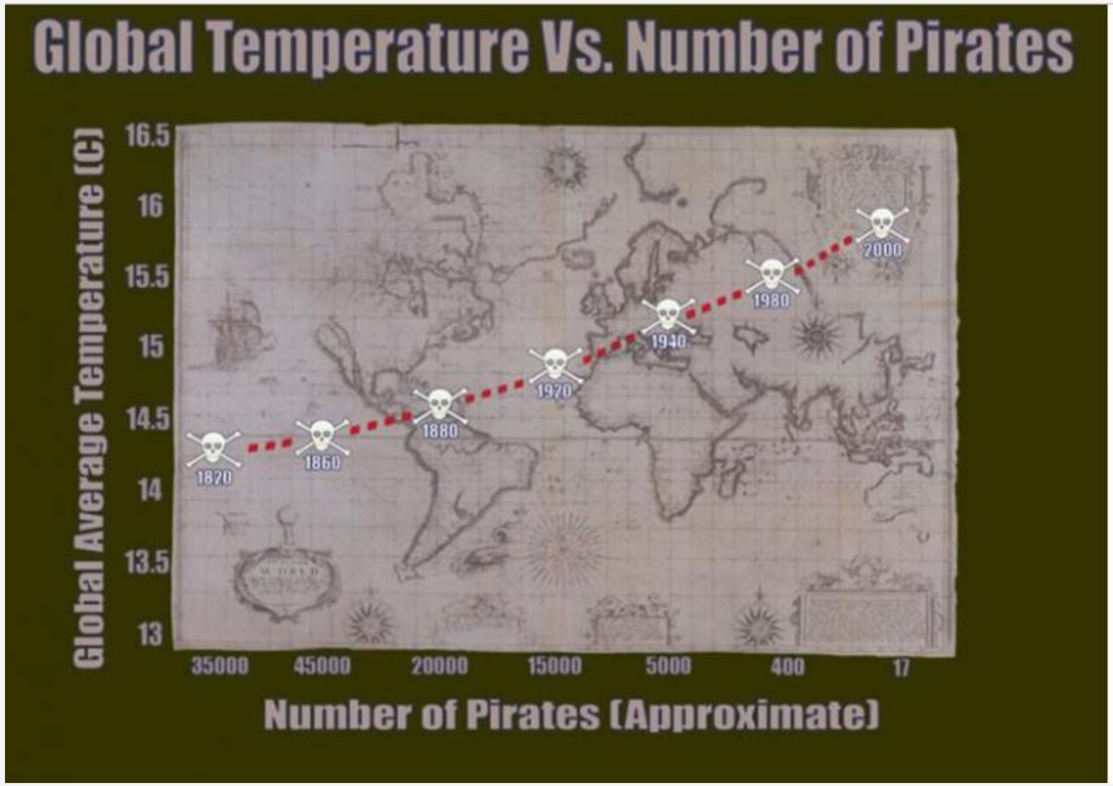
https://science.howstuffworks.com/innovation/science-questions/10-correlations-that-are-not-causations.htm

https://www.georanker.com/correlation-vs-causality-differences-and-examples

Correlation and Causation

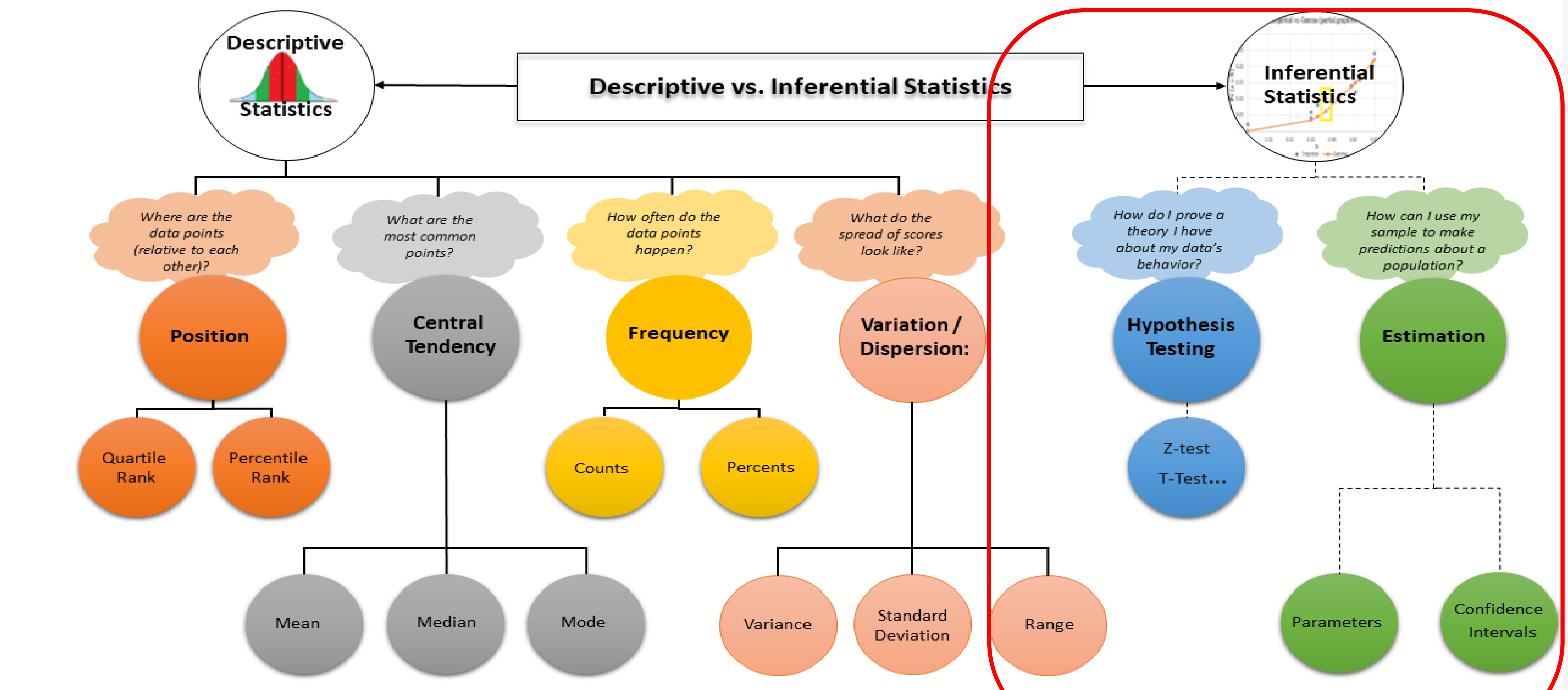

I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Global Warming caused by Lack of Pirate



https://www.sisense.com/blog/global-warming-caused-lack-pirates-bad-graph-lessons/

Inferential Statistics / Predictive Statistics


Descriptive Statistics
Descriptive vs. Inferential Statistics
Inferential Statistics
Where are the data points (relative to each other)?
What are the most common points?
How often do the data points happen?
What do the spread of scores look like?
How do I prove a theory I have about my data's behavior?
How can I use my sample to make predictions about a population?
Position
Central Tendency
Frequency
Variation / Dispersion:
Hypothesis Testing
Estimation
Quartile Rank
Percentile Rank
Counts
Percents
Z-test
T-Test…
Mean
Median
Mode
Variance
Standard Deviation
Range
Parameters
Confidence Intervals

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

Inferential Statistics – making estimations of the population from samples

Parameters: A characteristic that describes a population is called a parameter. Because it is often difficult (or impossible) to measure an entire population, parameters are most often estimated

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

Check out all the links for each topic

Statistic: A characteristic that describes a sample is called a statistic. Statistics are most often used to estimate the value of unknown parameters

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

- Distribution of Sample Mean:

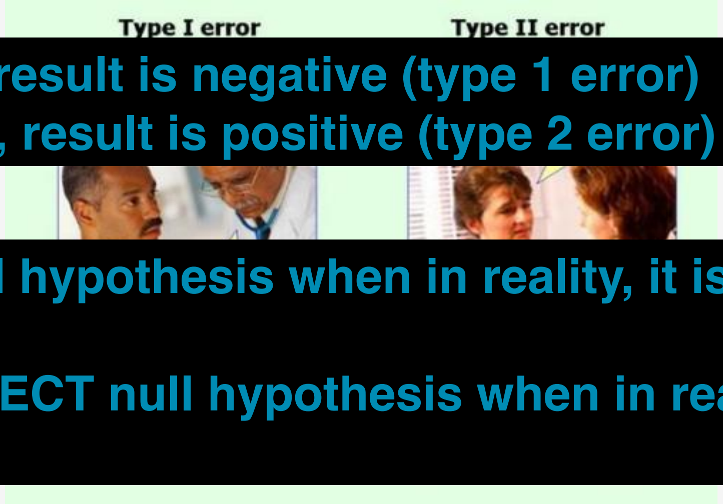
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

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

- The Central Limit Theorem: Independent of the actual distribution of the population, if we take a big enough sample size, when we repeat taking sample again and again, the distribution of the sample mean follows a normal distribution.
- That is why we can often use the normal distribution behind hypothesis testing

Hypothesis Testing

- Type I error (false positive, too excited to claim something non-existence)
- Type II error (false negative, failed to realize something real is going no)



False Positive = Prediction is positive, result is negative (type 1 error)
False Negative = Prediction is negative, result is positive (type 2 error)

- Null Hypothesis (nothing to see, life is as usual)
- Alternate Hypothesis (something is going on)

We REJECT null hypothesis when in reality, it is false
We do NOT REJECT null hypothesis when in reality, it is true

1. Define Null and Alternative Hypotheses
2. State Alpha
3. State Decision Rule
4. Calculate Test Statistic
5. State Results
6. State Conclusion

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

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

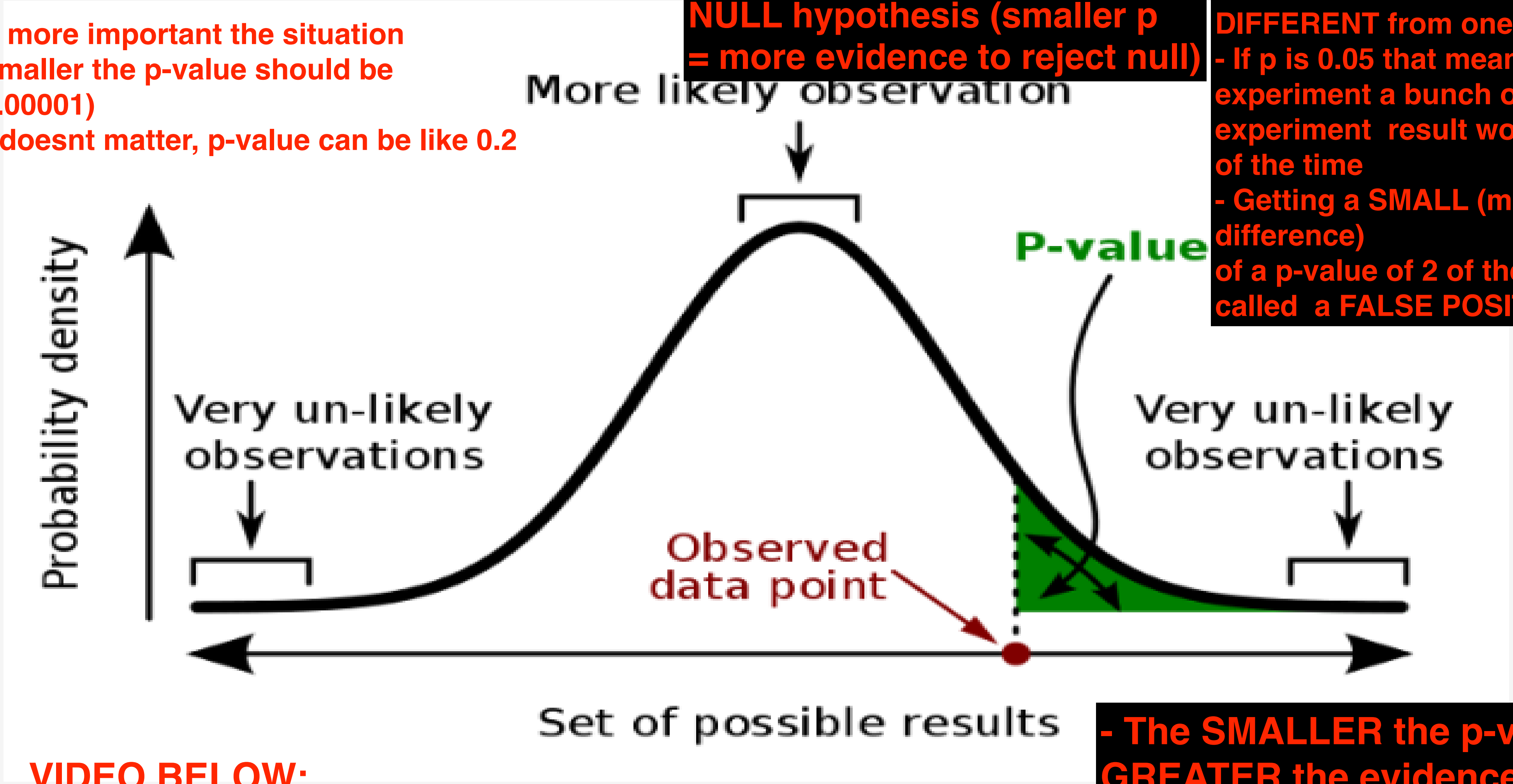
http://www.statisticslectures.com/topics/onesamplez/

P-value and Confidence interval

P-Value = probability we get the observed data point. How much evidence there is to REJECT the NULL hypothesis (smaller p = more evidence to reject null)

P-Value: Numbers between 0 and 1 that quantify how confidence you should be in the data. CLOSER to 0, the more confident (better probability) that we are sure that the 2 values are DIFFERENT from one another
- If p is 0.05 that means if we did the experiment a bunch of times, the experiment result would be wrong 5% of the time
- Getting a SMALL (meaning theres a difference) of a p-value of 2 of the SAME groups is called a FALSE POSITIVE

- The more important the situation the smaller the p-value should be (like .00001)
- if it doesnt matter, p-value can be like 0.2



- The SMALLER the p-value, the GREATER the evidence against the NULL hypothesis

VIDEO BELOW:

https://www.youtube.com/watch?v=vemZtEM63GY

Online Statistics Review

Watch this online Statistics Lectures as much as you can

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

TO-DO Task

Read Chapter 4 Data Mining of the Textbook

(first part of the chapter, especially on data cleansing and preparation)