# 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

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 <u>if we know many are also in the basketball Team</u>

# 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

**ρ(X,Y)** – the correlation between X and Y
**Cov(X,Y)** – the covariance between X and Y
$\sigma_X$ – the standard deviation of X
$\sigma_Y$ – the standard deviation of Y

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

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

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

# Covariance and Correlation

**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.

**Spearman rank-order correlation**

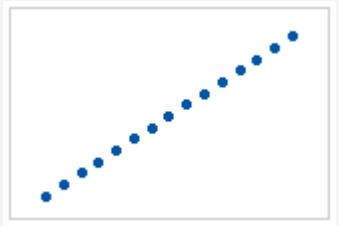
The Spearman correlation evaluates the monotonic relationship between two continuous or ordinal variables.
In a monotonic relationship, the variables tend to change together, but not necessarily at a constant rate.
The Spearman correlation coefficient is based on the ranked values for each variable rather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

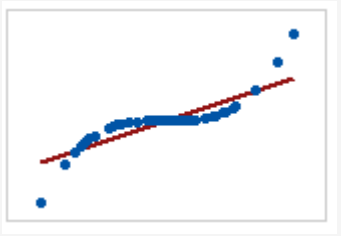
# Pearson vs Spearman Correlation

Fig 1



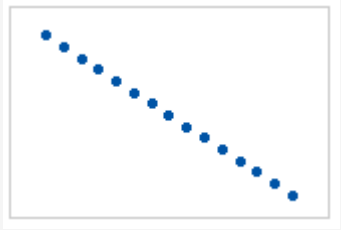
Pearson: +1
Spearman: +1

Fig 2



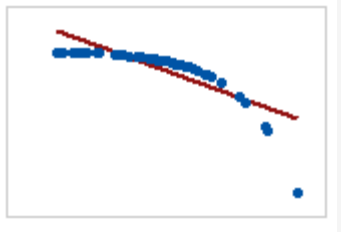
Pearson: ?
Spearman: ?

Fig 3



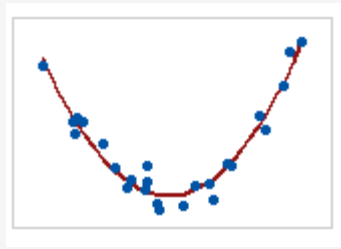
Pearson: -1
Spearman: -1

Fig 4



Pearson: ?
Spearman: ?

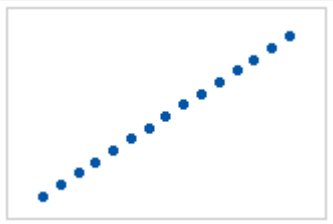
Fig 5



Pearson: ?
Spearman: ?

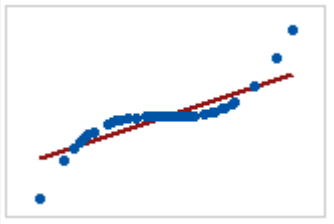
# Pearson vs Spearman Correlation

Fig 1



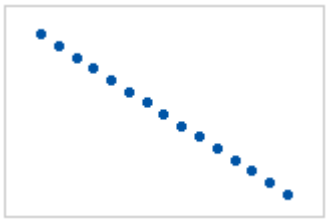
Pearson: +1
Spearman: +1

Fig 2



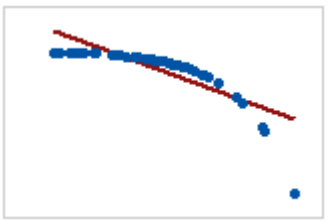
Pearson: +0.85
Spearman: +1

Fig 3



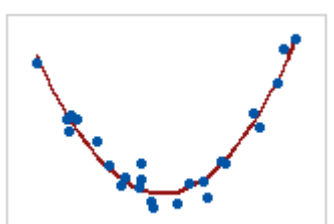
Pearson: -1
Spearman: -1

Fig 4



Pearson: -0.85
Spearman: -1

Fig 5



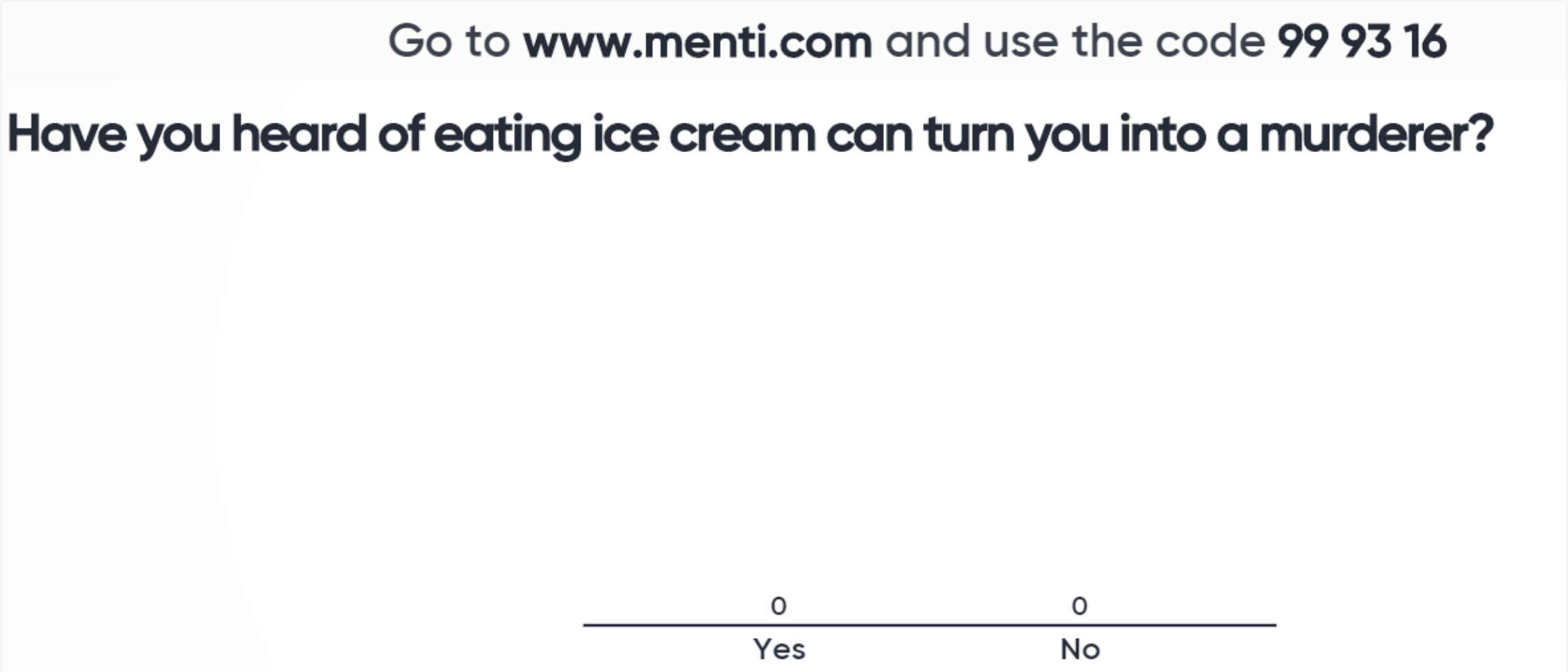
Pearson: 0
Spearman: 0

Zero correlation does not mean the variables are independent

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

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?
0
Yes
0
No

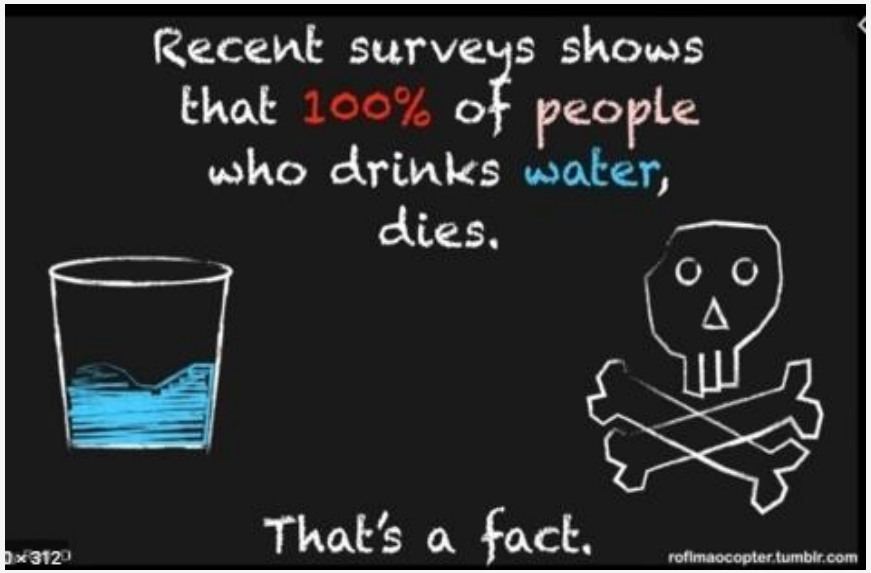
# Correlation and Causation

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

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?)

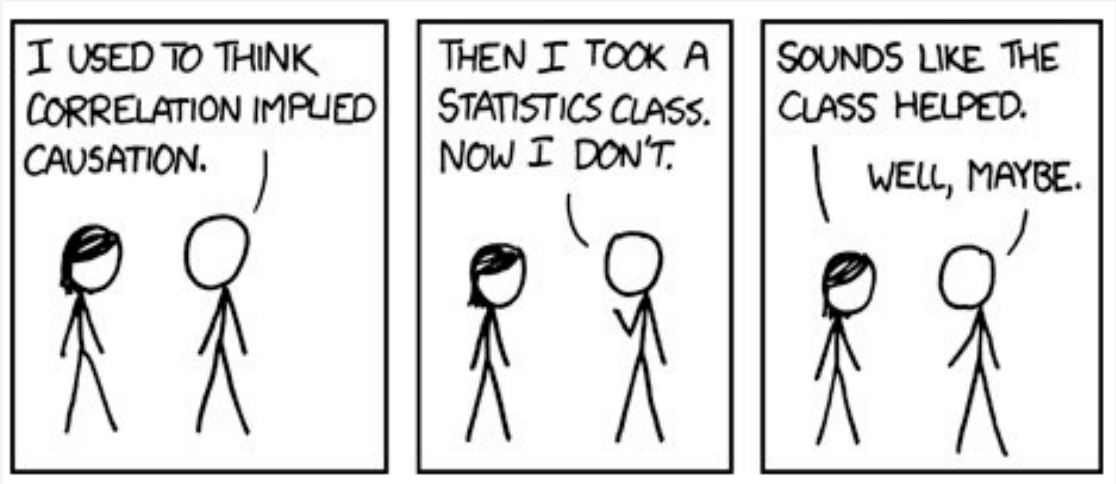
But, correlation is still one good tool to identify driving factors.



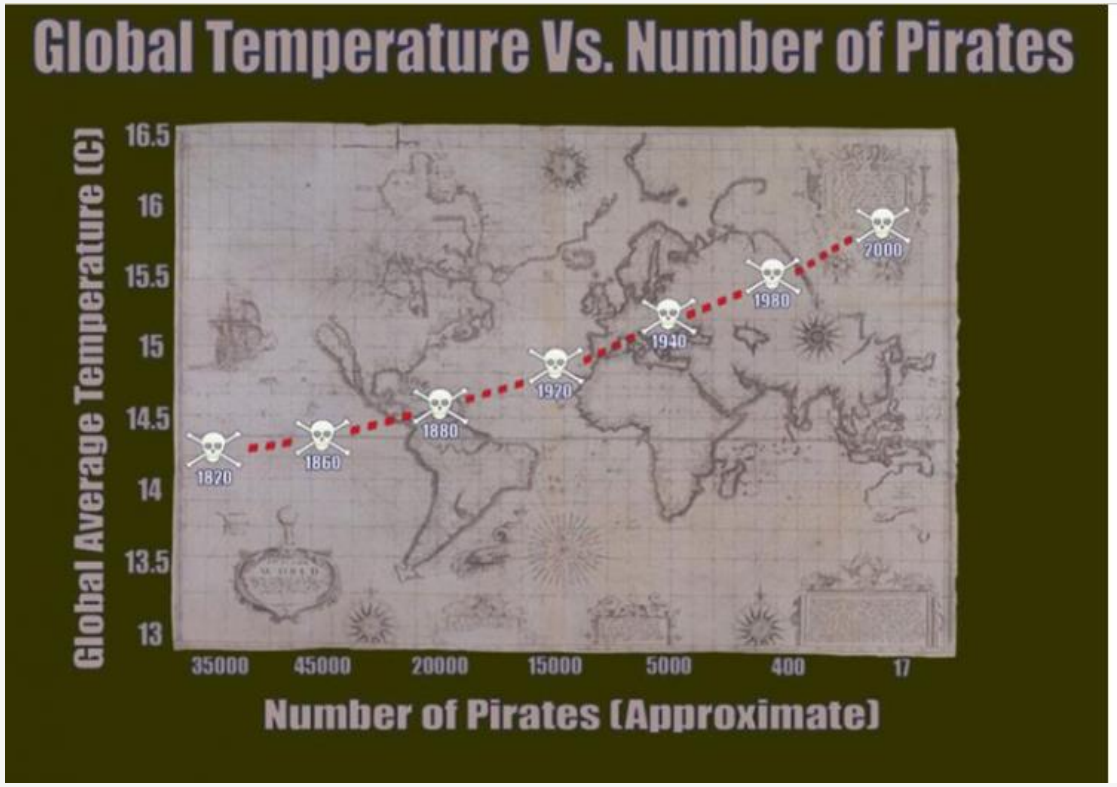

https://science.howstuffworks.com/innovation/science-questions/10-correlations-that-are-not-causations.htm

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

# Correlation and Causation

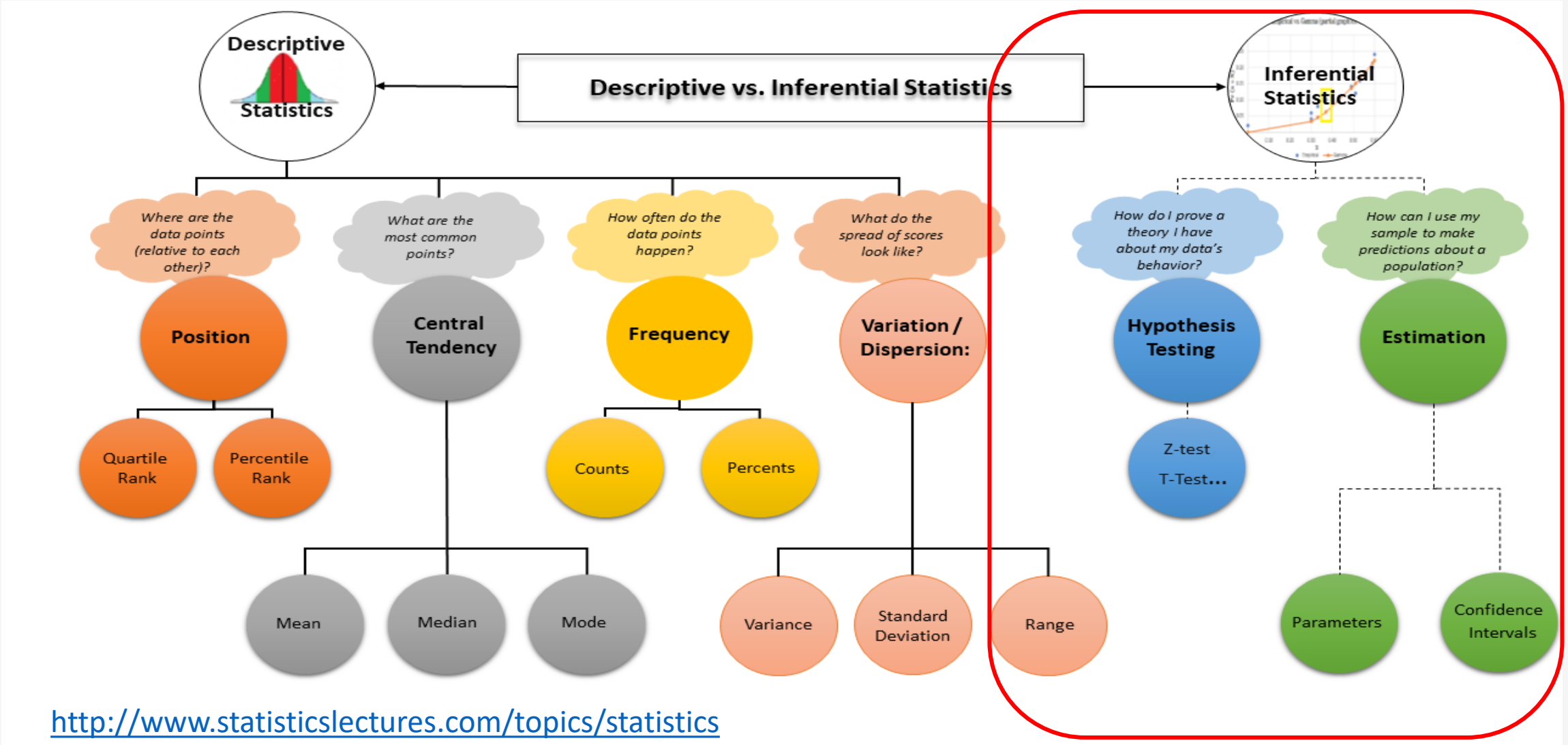



Global Warming caused by Lack of Pirate




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

# Inferential Statistics / Predictive Statistics




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/

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

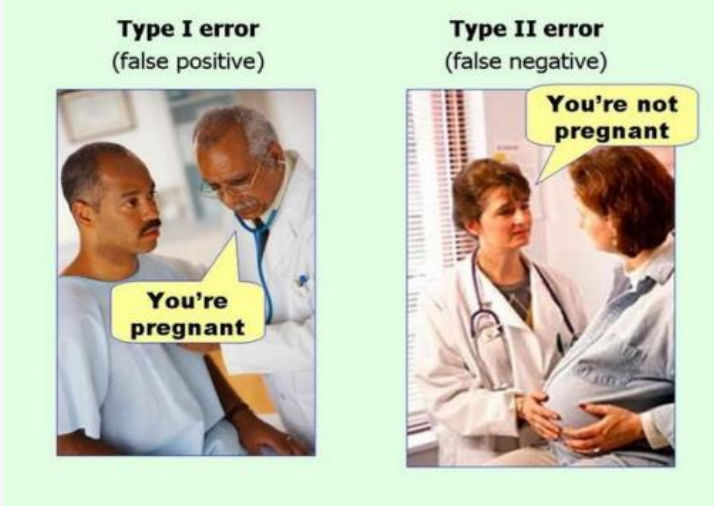
http://www.statisticslectures.com/topics/centrallimittheorem/

- 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)




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

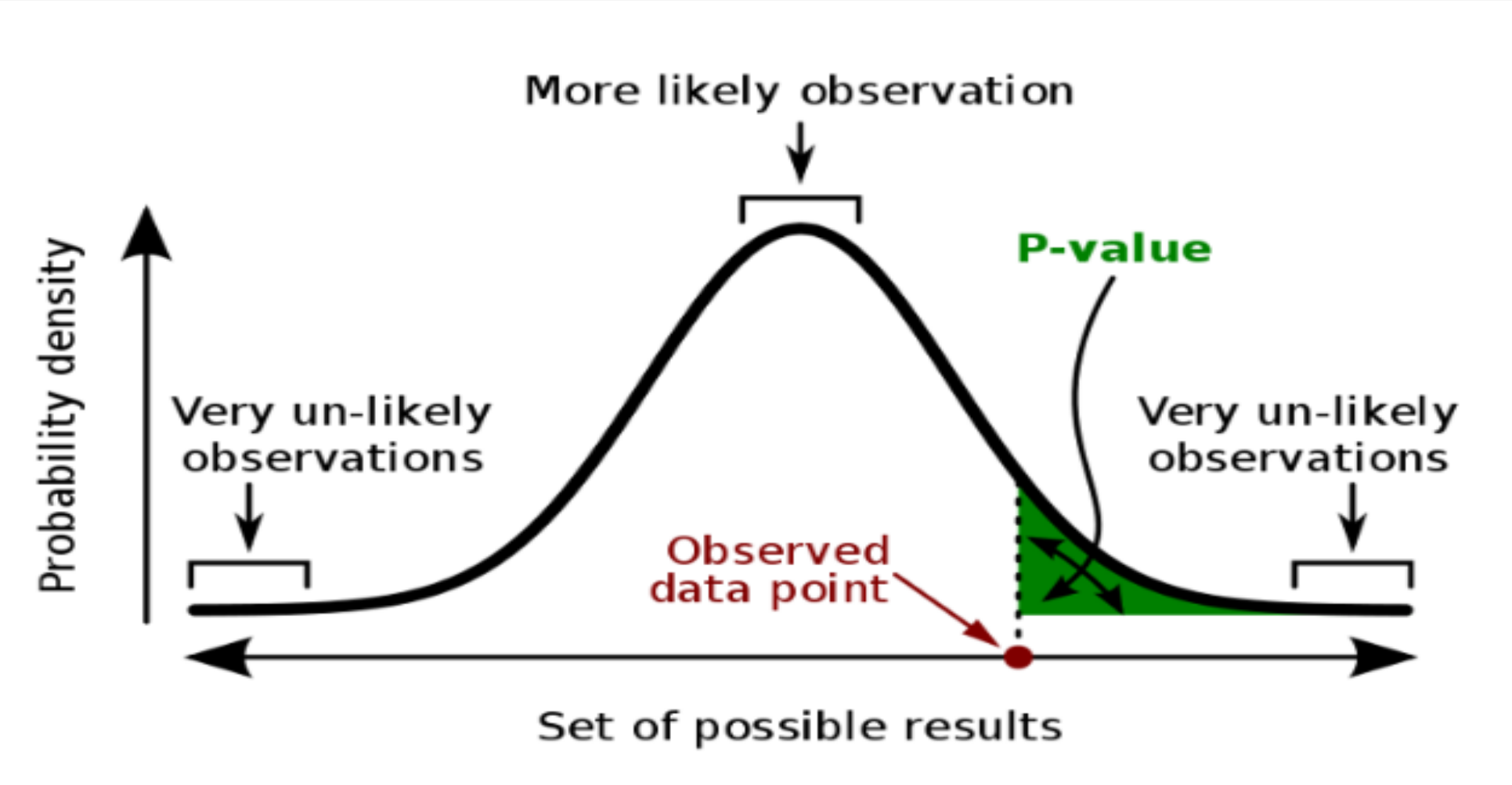
1. Define Null and Alternative Hypotheses
2. State Alpha
3. State Decision Rule
4. Calculate Test Statistic
5. State Results
6. State Conclusion

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

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

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

# P-value and Confidence interval




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)