Questions

What are the factors that drive house prices?

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

Mortgage Rates
Unemployment Rates
Local School performance

…

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
σ_X – the standard deviation of X
σ_Y – the standard deviation of Y

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

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

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 <u>proportional change</u> 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 <u>ranked values</u> 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.
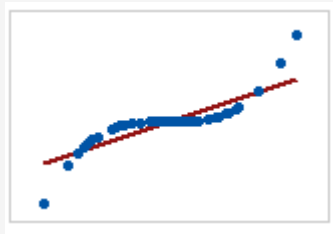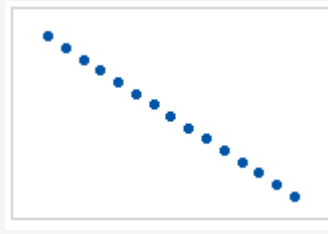
Pearson vs Spearman Correlation
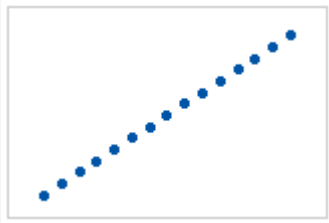
Fig 1

Pearson: +1
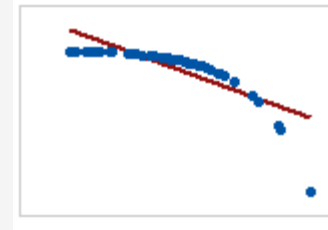Spearman: +1

Fig 2

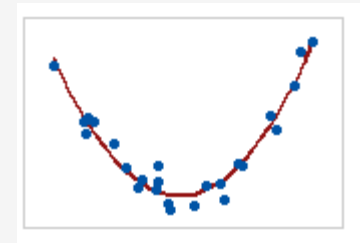Pearson: ?
Spearman: ?

Fig 3

Pearson: -1
Spearman: -1

Fig 4

Pearson: ?
Spearman: ?

Fig 5

Pearson: ?
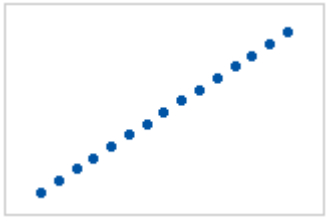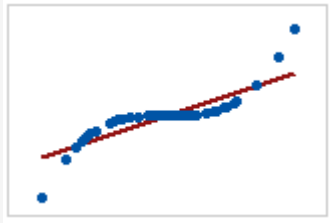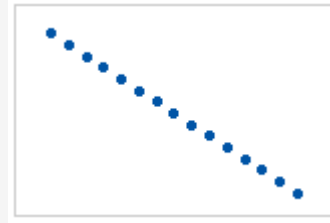Spearman: ?

Pearson vs Spearman Correlation

Fig 1

Fig 2

Fig 3

Fig 4

Fig 5

Pearson: +1
Spearman: +1

Pearson: +0.85
Spearman: +1
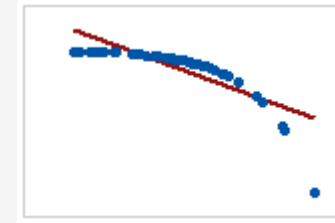
Pearson: -1
Spearman: -1

Pearson: -0.85
Spearman: -1
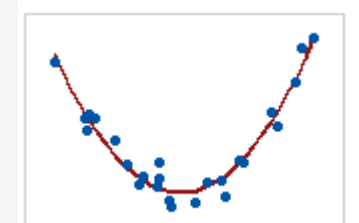
Pearson: 0
Spearman: 0

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	0
Yes	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 goes 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.

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

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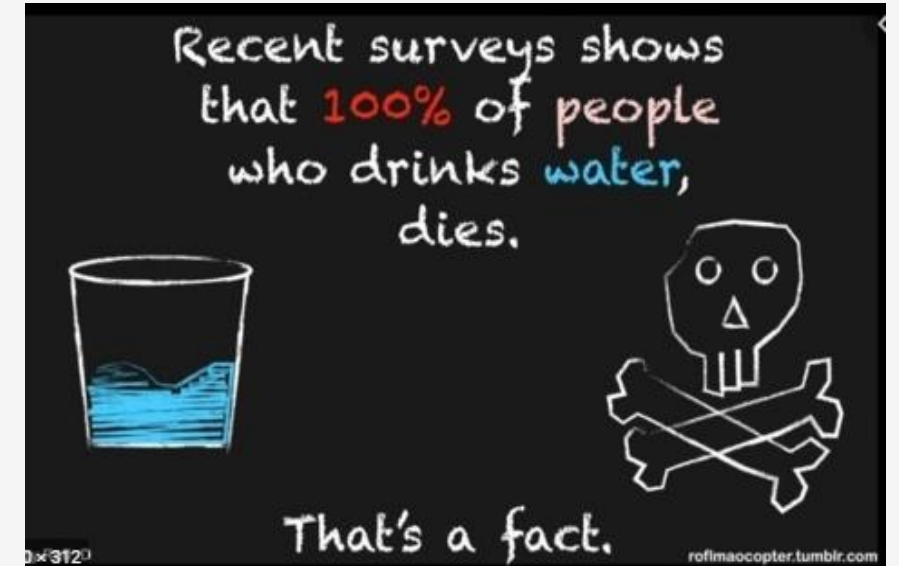
Inferential Statistics / Predictive Statistics

Descriptive vs. Inferential Statistics

Descriptive Statistics

Inferential Statistics

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

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/

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

- Distribution of Sample Mean: $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$

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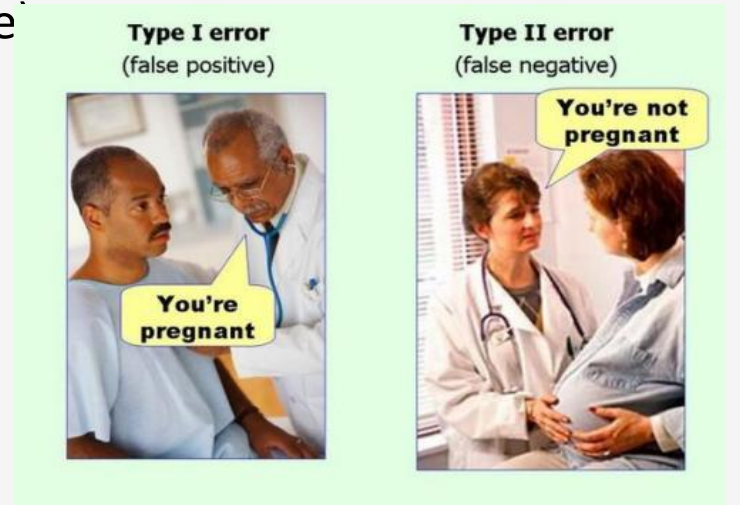
- Null Hypothesis (nothing to see, life is as usual)
- Alternate Hypothesis (something is going on)

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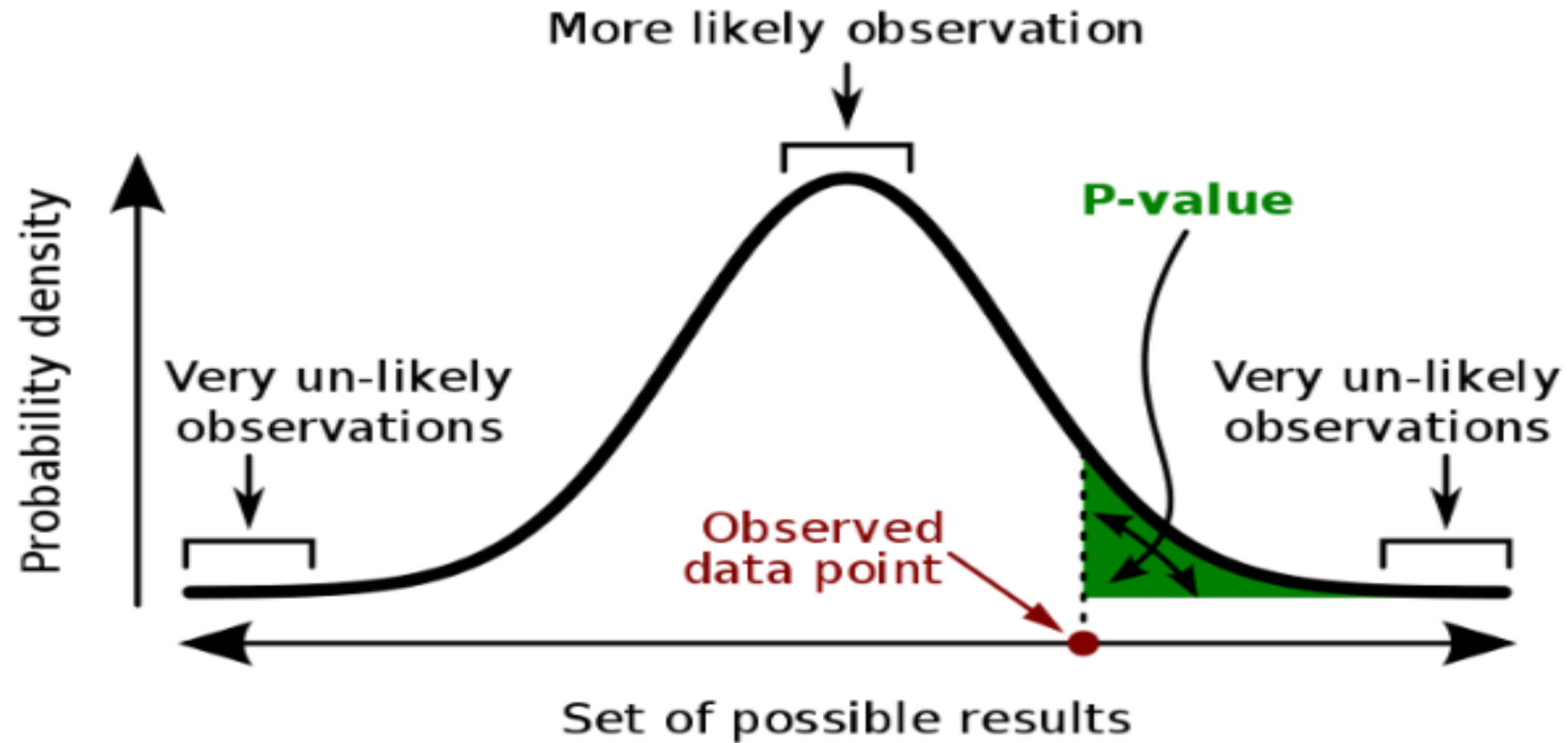5. State Results
6. State Conclusion

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

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

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

P-value and Confidence interval

Online Statistics Review

Watch this online Statistics Lectures as much as you can

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

TO-DO Task

Email your HW 1 if you haven't done so.
Please rename your notebook with your name

Read Chapter 4 Data Mining of the Textbook

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