Back to Basics

Statistics and Data Science

Hosted by the QSCU

The QSCU is a course union focused on helping students in MATH, COSC, PHYS, DATA, and STAT on campus

- QSCU website: http://qscu.org
- QSCU discord: https://discord.gg/xH48R8K
- QSCU facebook: https://www.facebook.com/ubcoqscu/
- QSCU instagram: @ubcoqscu

What will this be covering?

I'm going to try to aim for a broad overview of concepts, aiming on understanding of what these are, rather than any particular equations or practice problems.

Think STAT 230 but very fast and no homework.

If you are looking for solutions to practice problems, you can find review material for courses on our website :)

Topics Covered:

- Data
- Set Theory
- Introduction to Probability
- Hypothesis Testing

Data

Types of Data:

Qualitative Data:

- Data representing various attributes of a subject
  - Type
  - rating/ranking
- Oftentimes more difficult to run analysis on
  - need special algorithms or techniques to manipulate
  - can encode groups

Quantitative Data:

- measures of values or counts and are expressed as numbers.
- Easy to run analysis on

Types of Qualitative Data

Categorical/Nominal

- Can take on one of a limited number of possible values
- ex) Nationality: Irish, English, Canadian, American, French, etc.

Ordinal

- Exist on an ordinal scale
- Often rating on a scale

Binary

- Two possible values
- Often yes/no
- Represented with 1/0

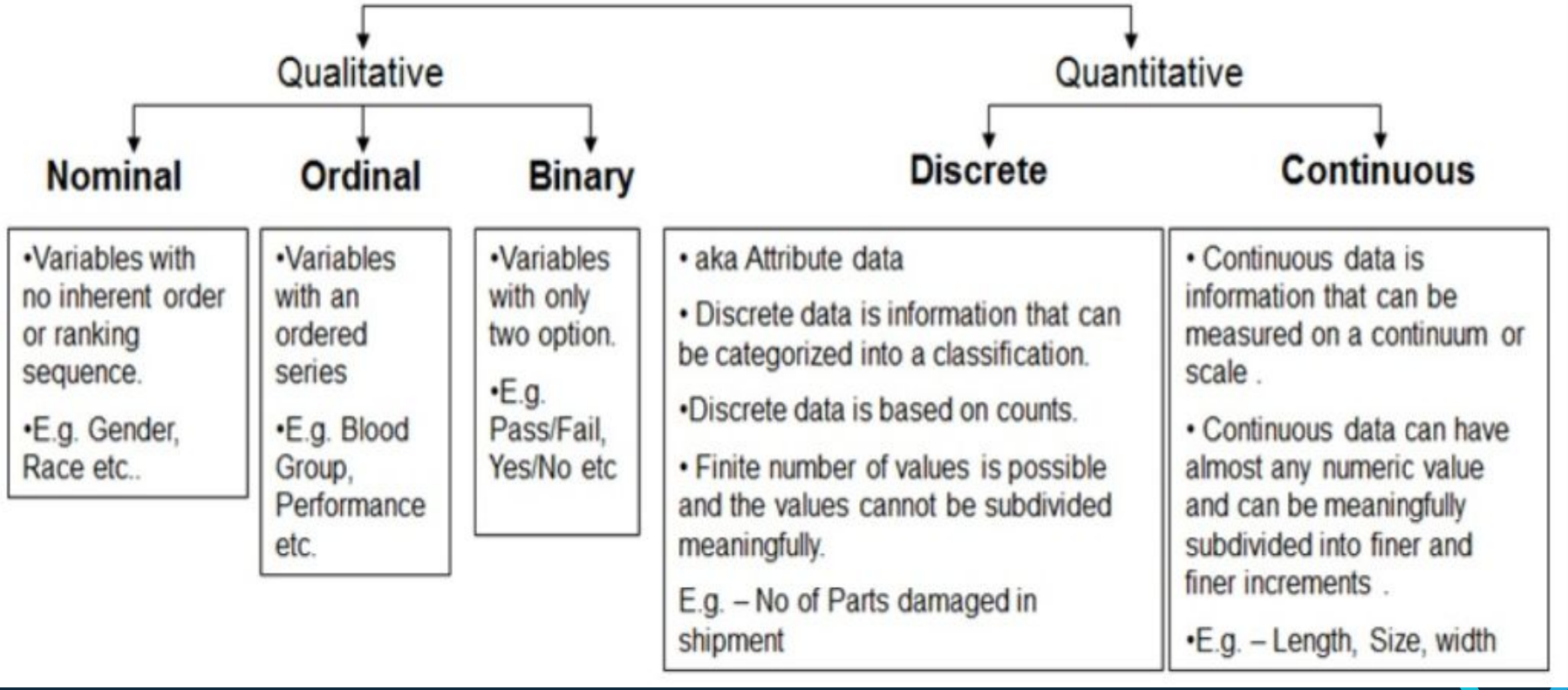
Types of Quantitative Data

Continuous

- Data arises when all values are possible inside some interval on the real line
- ex) distance between cities

Discrete

- Data arises when the possible items are countable
- ex) number of flips needed to get 10 heads with a flip of a coin


Qualitative
Quantitative
Nominal
Ordinal
Binary
Discrete
Continuous
•Variables with no inherent order or ranking sequence.
•E.g. Gender, Race etc..
•Variables with an ordered series
•E.g. Blood Group, Performance etc.
•Variables with only two option.
•E.g. Pass/Fail, Yes/No etc
• aka Attribute data
• Discrete data is information that can be categorized into a classification.
•Discrete data is based on counts.
• Finite number of values is possible and the values cannot be subdivided meaningfully.
E.g. – No of Parts damaged in shipment
• Continuous data is information that can be measured on a continuum or scale .
• Continuous data can have almost any numeric value and can be meaningfully subdivided into finer and finer increments .
•E.g. – Length, Size, width

Population vs Sample

Population

- It is an all-encompassing group of interest
- Usually unobservable in its entirety for one reason or another(most often the cost of measurement)
- ex) UBC Okanagan Students

Sample

- individual observations
- ex) people in this zoom call is a sample of UBC students

Percentile & Quartile & Interquartile Range

Percentile

- Value that has 100p% of ordered data falling below it
- Value that has 100(1-p)% of the ordered data falling above it

Quartile (Q)

- 25th percentile = first quartile (Q!)
- 75th percentile = third quartile (Q3)

Interquartile Range (IQR)

- Size of the gap between the first and third quartile
- ‘Distance’ over which the ‘middle half’ of the data is spread
- IQR = Q3 - Q1

Data Representation

Attributes

sample

	State	Account length	Area code	International plan	Voice mail plan	Number vmail messages	Total day minutes	Total day calls	Total day charge
0	KS	128	415	No	Yes	25	265.1	110	45.07
3	OH	84	408	Yes	No	0	299.4	71	50.90
4	OK	75	415	Yes	No	0	166.7	113	28.34
5	AL	118	510	Yes	No	0	223.4	98	37.98
6	MA	121	510	No	Yes	24	218.2	88	37.09

Individual observation

Set Theory

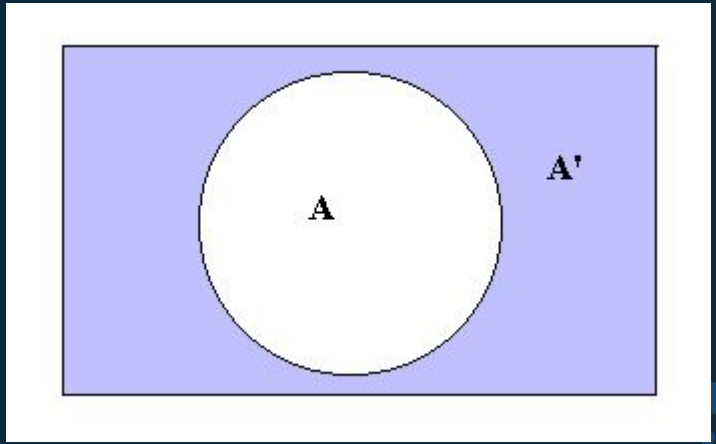
What is a set

A set is a well defined collection of unique objects.

- think of a list of unique numbers (no duplicates)
- In Statistics, a set is often used to describe possible outcomes to an experiment

Compliment

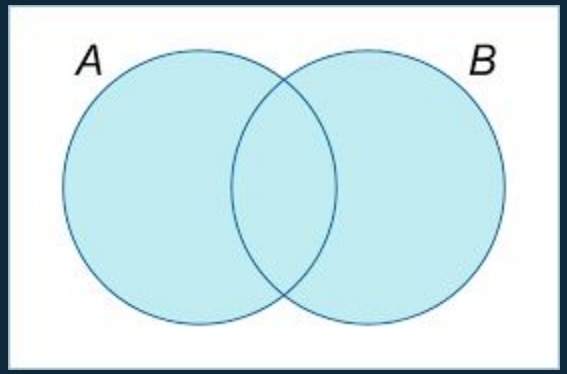
A’ is the set of all outcomes that are not contained in A



Union

A∪B is the event consisting of all outcomes in either A or in B or in both events.

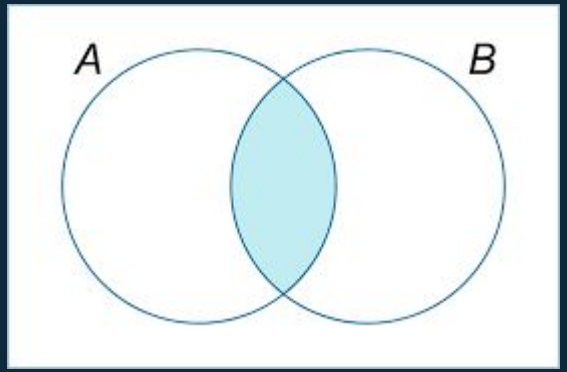
- you can think of a union as an OR operation



Intersection

A∩B is the event consisting of all outcomes that are in both events A and B.

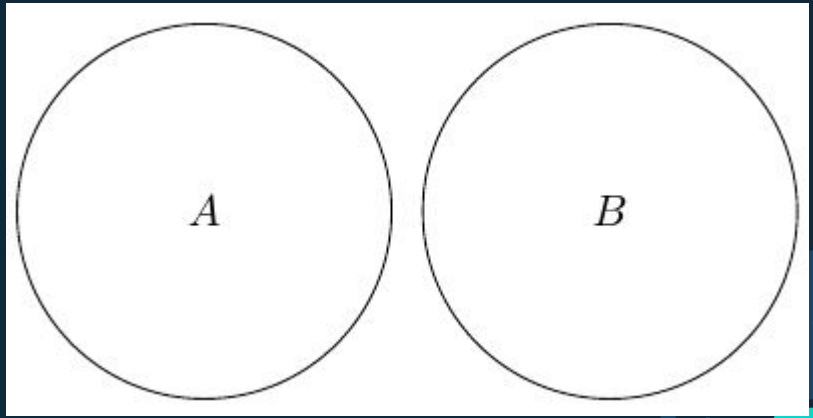
- you can think of an intersection as an AND operation

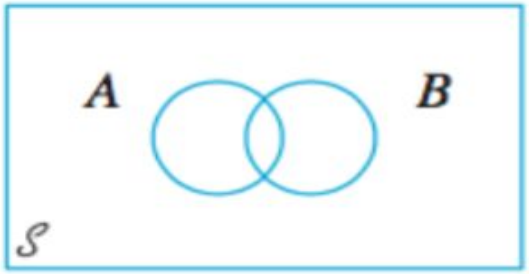


Disjoint Sets

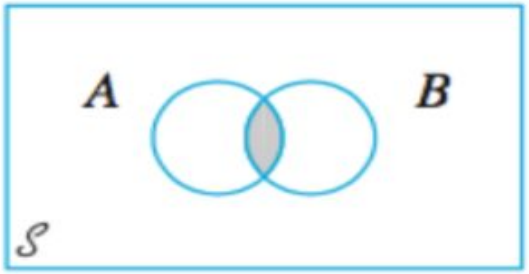
When A and B hae no outcomes in common, they are **disjoint** or **mutually exclusive**.

- $A \cap B = \emptyset$
- $\emptyset$ is the **null set**, or an empty set

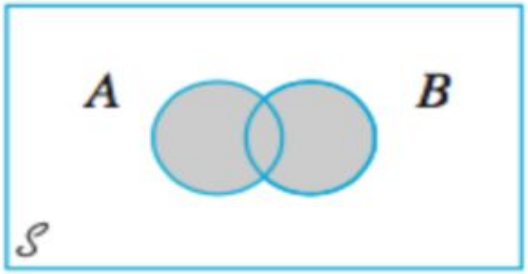




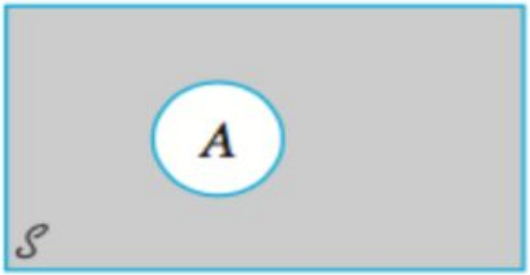
(a) Venn diagram of events A and B



(b) Shaded region is $A \cap B$



(c) Shaded region is $A \cup B$



(d) Shaded region is A'

(e) Mutually exclusive events

Probability

What is Probability?

Probability is the study of randomness and uncertainty in any situation in which one of a number of possible outcomes may occur.

The **Theory of Probability** provides methods of quantifying chances/likelihoods associated with various outcomes.

Sample Space, Experiments, and Events

An **experiment** is any action or process whose outcome is subject to uncertainty

The **Sample Space** (S) of an experiment is the set of all possible outcomes of that experiment

In the study of probability, we are interested in both individual outcomes and any collection of outcomes in the sample space.

An **event** is any subset of outcomes contained in the sample space.

- a simple event consists of one outcome
- a compound event consists of more than one outcome

Independent vs Mutually Exclusive

Mutually Exclusive

- Events cannot happen at the same time
- ex) a dice cannot be 2 and 3 at the same time
- Check if disjoint
  - A and B = 0
  - $P(A \cap B) = 0$
  - Check the formula
    $P(A \cup B) = P(A)+P(B)$
    Check LHS = RHS

Independent Events

- The outcome of A does not affect B
- ex) rolling a dice, then rolling again
- Check formula

  $P(A \text{ and } B) = P(A)P(B)$

  Check LHS = RHS

Axioms of Probability

Axiom 1: Non Negative Probability

- For any event A, $P(A) \geq 0$

Axiom 2:

- $P(S) = 1$

Axiom 3:

- If $A_1, A_2, \ldots$ is an infinite collection of disjoint events, then $P(A_1 \cup A_2 \cup \ldots) = \sum_{i=1}^{\infty} P(A_i)$
- If $A_i \cup A_j$ are mutually exclusive for all $i \neq j$ then $P(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$

Interpreting Probability

If an experiment is performed n times, on some of the replications event A will occur and on others it will not. The **relative frequency** of the occurrence of A is then*(number of times A occurs) / n*

As n increases, the relative frequency stabilizes.

For sufficiently large n, the relative frequency is an objective interpretation of probability.

Any event that is not repeatable therefore cannot stabilize, so it is a subjective interpretation.

Some other properties...

- For any event A

  $P(A) = 1 - P(A')$

- For any event A

  $P(A) \leq 1$

- For any events A and B

  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

  Think of this as not counting the intersection twice

Counting

Permutation

- The number of ways of arranging n DISTINCT objects in a line is

  $n! = n*(n-1*)(n-2)*...*3*2*1$

- Each of n! Arrangement of n distinct objects is called a permutation of the n objects.
- ORDER DOES MATTER
- eg)
  - anagrams of a word

Combination

- Order does not matter
- keyword is: CHOSE
- eg)
  - choosing cards from a deck
    - probability of drawing a heart

$$ {}^{n}C_{r} = \frac{n!}{r!(n-r)!} = \binom{n}{r} $$

Conditional Probability

The probability that an event E occurs given that an event F has already occured.

Second event (E) is not independent from the first event (F)

$$P(F \mid E) = \frac{P(E \cap F)}{P(E)}$$

$$P(F \mid E) = \frac{P(E \mid F)P(F)}{P(E)}$$

Bayes Theorem

Theorem 3.6 (Bayes' Theorem)

If $F_1, F_2, \ldots, F_n$ are mutually exclusive events of which one must occur, and $P(F_i) \neq 0$ for $i = 1, 2, ..., n$ then for any event E for which $P(E) > 0$

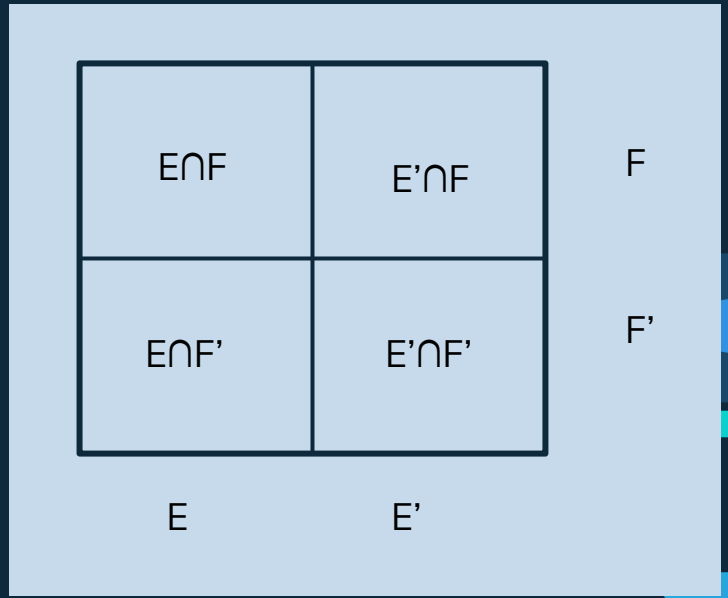
$$P(F_j \mid E) = \frac{P(E \mid F_j)P(F_j)}{\sum_{i=1}^{n} P(E \mid F_i)P(F_i)}, \quad j = 1, 2, \ldots, n.$$

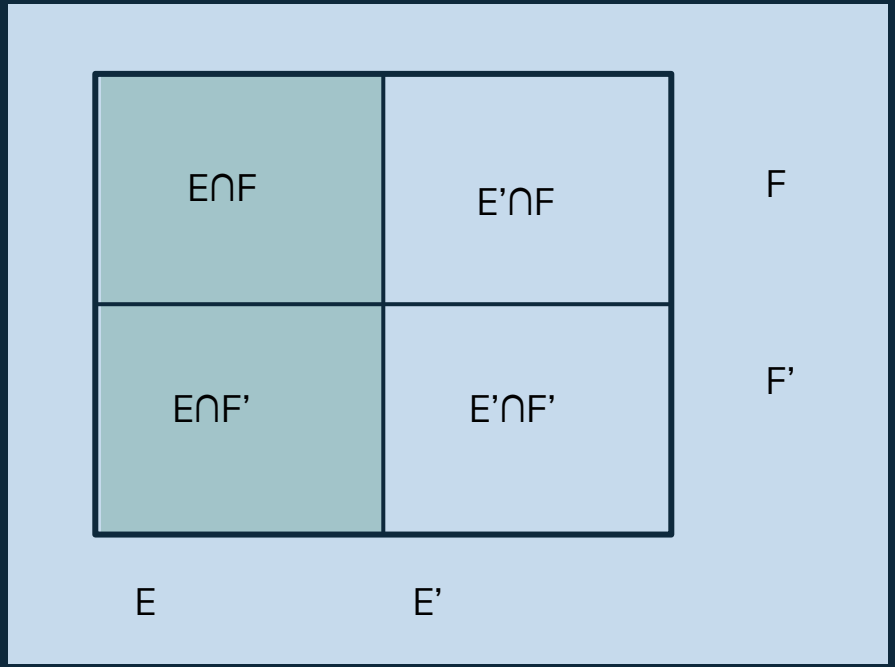
Understanding Bayes

When thinking about Bayes, one must consider that for any event, you must account for the situation that it occurs with the other event and the situation that it occurs without it.

For any probability E:

$$P(E) = \sum_{i=1}^{n} P(E \mid F_i)P(F_i)$$





$$P(E) = P(E|F)*P(F) + P(E|F')*P(F')$$

Hypothesis Testing

Why do we need Hypothesis Testing

When we conduct an experiment there is random variation that comes from the differences within the population. This is called the **variance** of the population.

When we try to compare two populations, we must consider how their variance comes into account. If the difference in mean between the two populations is small in comparison to large variance, there is insufficient evidence to suggest a difference.

Sampling Variance

When we conduct an experiment, we cannot feasibly evaluate the entire population. To maintain a controllable experiment, we use **sampling** to make a representation of the population, infer truths about the population from the sample, and then adjust any values to account for the **sampling variance**.

eg. a t test will have an adjusted estimate to account for the sampling variance.

Central Limit Theorem

If we take a random sample of size *n* from any population, with mean μ and standard deviation σ, then when *n* is large, the distribution of the sample mean is approximately normal with mean μ and variance σ^2/n.

For sufficiently large n, the sample mean is approximately a normal distribution with the sample variance

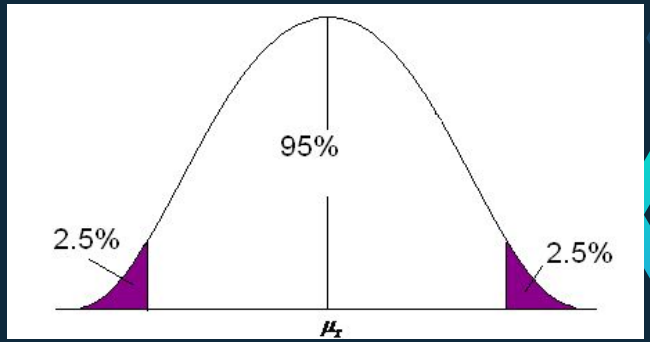
$$\overline{X} \text{ is approximately } N(\mu, \sigma_{\overline{x}}^2),$$

The 2 Sample T Test

When comparing two independent samples, we compare their means to determine if the differences between both are only due to variations form random sampling (H_0) or reflect a difference in the actual population (H_A).

We reject our null hypothesis if we consider the probability of observing the null hypothesis to be below a threshold (alpha) that is considered too unlikely to occur naturally assuming the null hypothesis is true

$$t = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

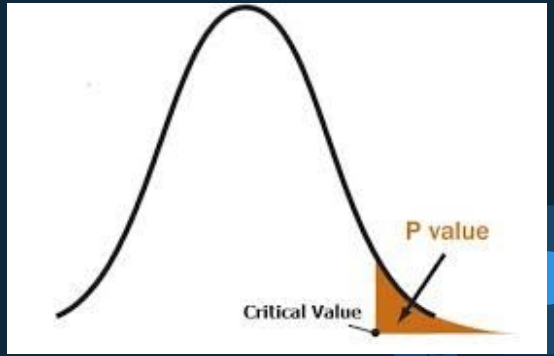


What is a P Value

The p value is the probability of observing a test statistic as or more extreme than the one observed, assuming the null hypothesis is true.

We use p values to determine how likely it is to see the results observed.

If the p value is below a threshold (alpha) then we claim that the results seen are too unlikely to occur if the null hypothesis is true.



The 6 Steps of a T Test

1. State the null and alternative hypothesis
2. State the assumptions of your test, as well as the alpha
3. Calculate the test statistic
4. Calculate the p value
    a. from a t table
    b. from R
5. Make a decision about the null hypothesis
    a. this is NOT about the alternative hypothesis
6. State what this test implies about your alternative hypothesis

Can a hypothesis test be wrong?

A hypothesis test can absolutely be wrong. In statistics, nothing has a zero percent probability.

We can replicate tests to prove that the results are correct

We can determine the **power** of a test, in which the pvalue and alpha can be used to say that the results are valid. This is a pretty complex topic, though

Statistical Computing

R

R is a free software environment for statistical computing and graphics.

It basically exists to keep any statisticians and Data Scientists out of a hell of Excel spreadsheets. It functions like a programming language, just focused on statistics.

Thank you! :)