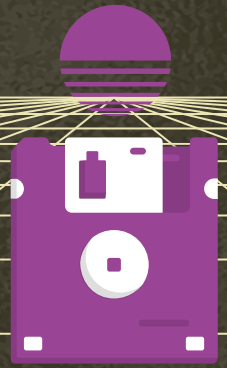
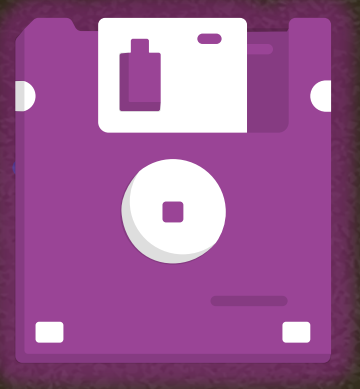
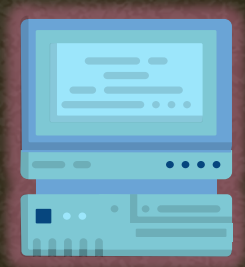

StaRtistics
introduction to statistics

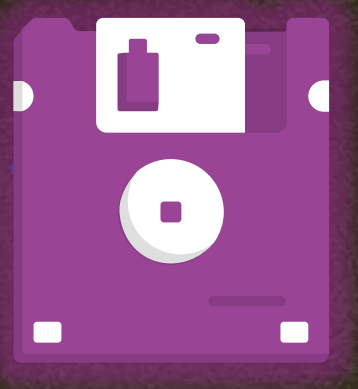
Class 1



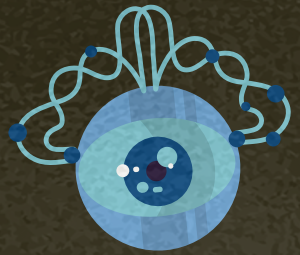


Statistics aim to understand your data by describing it and making predictions

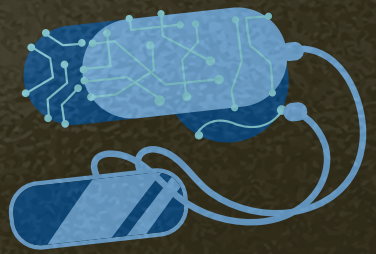




Descriptive statistics



Descriptive stats aims to describe data's distribution by central tendency (location in a plane) and dispersion (spread) around the location



Central tendency

- Mean : the arithmetic average
- Median: the middle value of array
- Mode: Most common value of array

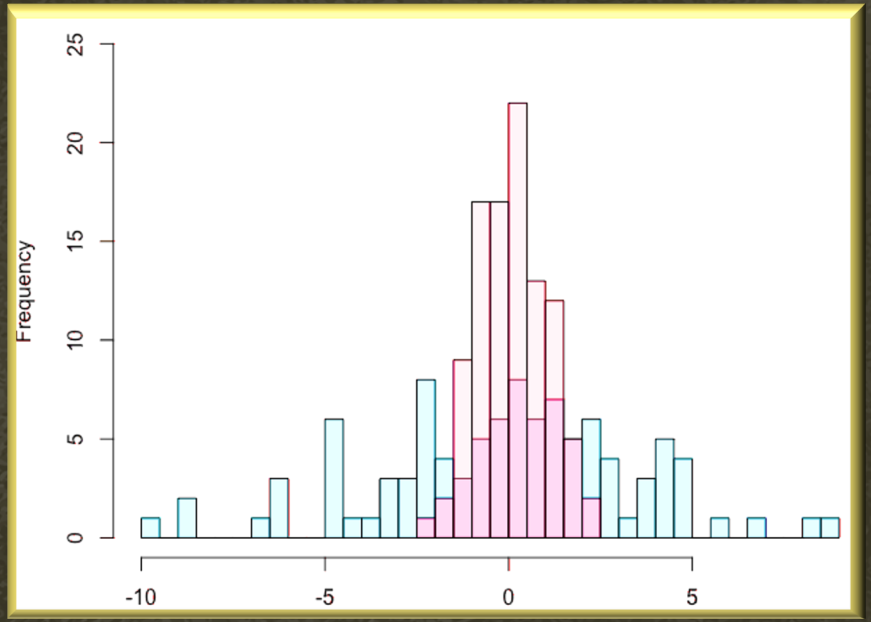
Note: some descriptive stats have associated functions in basic R e.g., rowMeans() see code

Data spread

- Max : largest value
- Min: smallest value
- Find k largest/ smallest elements (see code)
- Range: difference between largest and smallest value
- Percentile: value where X% of your data smaller than this value

Data spread

- Standard deviation: the spread of your data
- Variance: the square of standard deviation



Confidence Intervals

Not only do you want to DESCRIBE the general tendency of your data but also its SPREAD or uncertainty ...

Confidence Intervals

Not only do you want to DESCRIBE the general tendency of your data but also its SPREAD or uncertainty ...

$$CI = X \pm z * \frac{sd}{\sqrt{n}}$$

Where z is the confidence level (i.e., if 95%, it would reflect 1.96 SD)

Confidence Intervals

$$CI = X \pm z * \frac{sd}{\sqrt{n}}$$

Z is computed assuming a distribution (e.g., normal distribution)

We will discuss how we can compute CI WITHOUT explicitly assuming a distribution of the data

Confidence Intervals

$$CI = X \pm z * \frac{sd}{\sqrt{n}}$$

95% CI DO NOT signify that you have a 95% probability of having the parameter within the interval!!! This is a very common misconception

It means that the interval will contain the parameter estimate in 95 studies out of 100 when repeated

Correlation

Correlations is the relationship between any two variables

They can be described in different ways:

- **Pearson**—linear relationship between continuous variables
- **Spearman Rho**—nonparametric rank correlation, describing two variables as a monotonic function

Dealing with missing data

You can simply remove that datapoint

You can replace that datapoint for the mean, median, or mode

You can use regression to infer the value of this datapoint (see regression section)

Finding outliers

Some people use 3*Sd of your data as a cutoff for outliers…

because these values are very unlikely (refer to normal dist plot to see likelihood)

You can also other definitions like the Inter quantile range or Cook’s distance

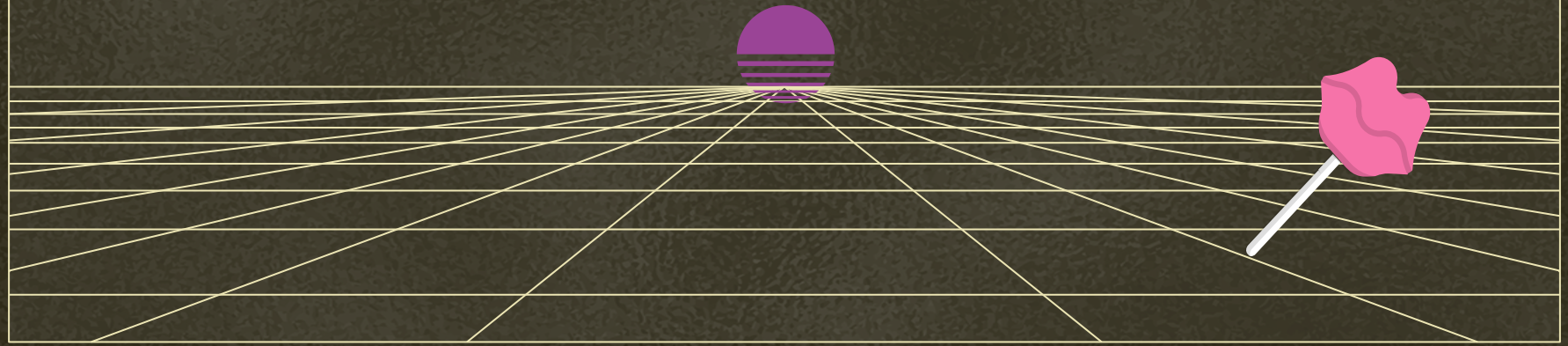
Dealing with outliers

You can remove outliers when they affect your interpretation of your results

But keep in mind you are removing data which is assuming that it does not reflect reality even though it is a sample

Always ask yourself how much you believe that this observation is abnormal

Inferential statistics



Inferential stats aims to predict future outcomes or observations based on your data

T-test

Student t-test allows you to test mean differences between normal distributions

There are many different **flavours** of t's

- one-sample vs two samples
- paired vs unpaired
- one tail vs two

T-test

T-tests assume that your data come from a **normal** distribution and the observations are sampled **independently** from one another

These assumptions apply for both paired and unpaired test

T-test

One-sample t-tests test the hypothesis that mean is different than a prespecified u0

Independent-sample t-tests test the hypothesis that the mean difference between both samples is not 0

T-test

Paired t-tests are used when the observations are repeated, i.e., each person is sampled twice thus each observation comes in a pair

Unpaired t-tests are used when the observations are independent

T-test

One tail tests are used when you have **directional** hypotheses (i.e., group A is bigger than group B)

Two tail tests are used when you have **non-directional** hypotheses (i.e., group A is different than group B, but you don’t care if it’s bigger or smaller)

Statistical power in Inferential stats

Statistical power

Statistical power is the probability that a given statistical test correctly rejects a null hypothesis H_0

For example: getting a negative COVID test when you do NOT have COVID

Note: it is commonly denoted 1-β

Effect size

Effect size is a (usually) standardized measure of a relationship between two variables. These can be for example:

Correlation coefficient (r)

Beta from a regression (β)

Cohen’s D (standardized mean difference)

Power analysis

Effect size, **statistical power**, and **sample size** are all RELATED

It is enough to know two of these, to compute the value of the third

i.e., compute sample size given an effect size and statistical power or compute power given a sample size and effect size etc..

Experiment

Given the same number of data which type of t-test is more stringent? Which one requires a bigger mean diff for the same value of t?

T-test

Let us assume we collect data from 20 people (paired) and observe a mean diff of 3.75, and a sd of 1

Then the t value for a paired test would be:

$$t = \frac{mean\ diff}{\frac{sd}{\sqrt{n}}} \text{ with df n} - 1$$

T-test

Let us assume we collect data from 20 people (paired) and observe a mean diff of 3.75, and a sd of 1

Then the t value for a paired test would be:

$$t = \frac{3.75}{\frac{1}{\sqrt{20}}} = \text{ with df } 19$$

T-test

Let us assume we collect data from 20 people (independent groups) and observe a mean diff of 3.75, and a sd of 1 (assuming equal variance)

Then the t value for an unpaired test would be:

$$t = \frac{mean\ diff}{sp * \sqrt{\frac{1}{n1} + \frac{1}{n2}}} \text{ with } \mathrm{df}\ \mathrm{n1} + \mathrm{n2} - 2$$

T-test

Let us assume we collect data from 20 people (independent groups) and observe a mean diff of 3.75, and a sd of 1 (assuming equal variance)

Then the t value for an unpaired test would be:

$$\mathrm{sp} = \sqrt{\frac{(n1-1) * std1^2 + (n2-1) * std2^2}{n1+n2-2}} \quad \mathrm{df\ n1 + n2 - 2}$$

T-test

Let us assume we collect data from 20 people (independent groups) and observe a mean diff of 3.75, and a sd of 1 (assuming equal variance)

Then the t value for an unpaired test would be:

$$\mathrm{sp} = \sqrt{\frac{(20-1)*1+(20-1)*1}{20+20-2}} \qquad t = \frac{3.75}{sp*\sqrt{\frac{1}{20}+\frac{1}{20}}}. \qquad \mathrm{df}\ 38$$

T-test

We can see that the paired t-test has a much higher t-value than the unpaired (note that t-values ARE NOT effect sizes, compute Cohen's D).

The Cohen's D in this situation would be the same given that we have the same mean diff and sd

$$Cohen's\ D = \frac{mean\ difference}{Std\ deviation\ samples}$$

T-test

Yet, since we use t-values to calculate the p value you can see that you are more likely to get a significant effect when using the paired test (since the t-value is bigger)

T-test

T value for paired is larger, in comparison to t values of unpaired or independent tests

This is because **within-person** designs have more **power** as they control for more noise by observing the same person twice