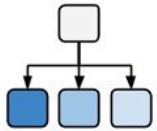
R Workshop III

Carlos Utrilla Guerrero
Institute of Data Science - Researcher



Course: VSK1004 Applied Researcher

Recap...

- Use descriptive statistics to explore your data
- Demonstrate types of graphs and interpret them
- Practice exploring data

What we are covering today

- Basics concepts of inferential statistics.
- Null hypothesis, significance testing for comparing means
- Correlation and interpretation.
- Demo using R Markdown

What we aren't doing today

- Covering in depth inferential statistics
- Deep dive into statistical formulas
- Expecting you to 100% get all what we'll cover today!

From describing to inferring

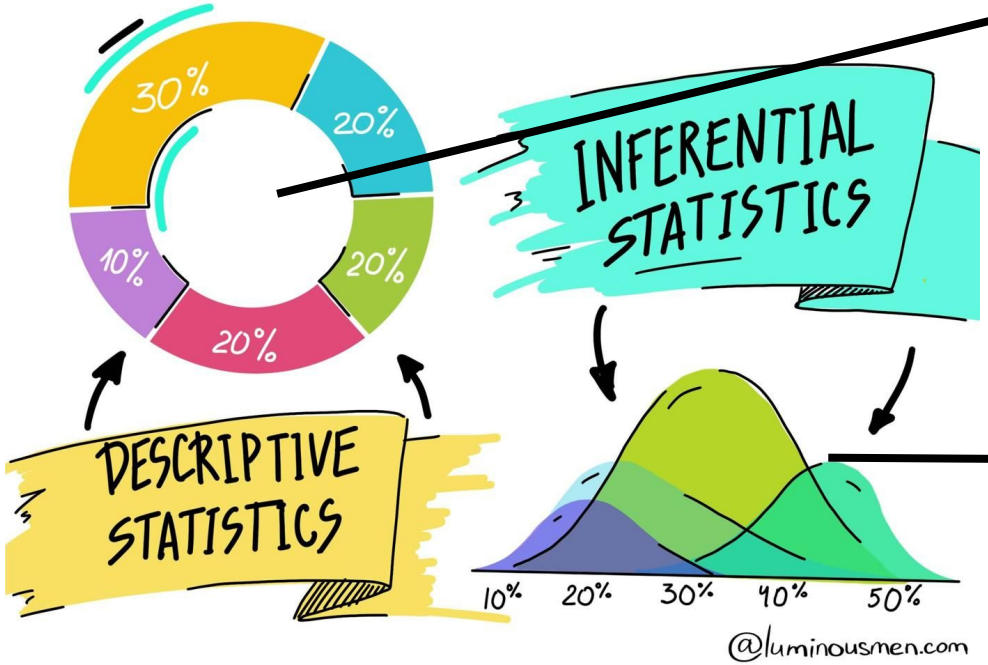


Table 1: Descriptive Statistics

	Body height
Mean value	1.67
Median	1.655
Mode	1.64
Sum	16.7
Standard deviation	0.066
Variance	0.004
Minimum	1.55
Maximum	1.78
Range	0.23

One sample t-test

Is there differences between a group of individuals and population?

Unpaired t-test

Is there a difference between two independent groups (samples)

We use these ideas to perform statistical statistics when we want to infer something about a population based on observations of a sample of that population.

For example, you would like to know the mean height of all the students enrolled on the Applied Research program.

However, the current coronavirus restrictions mean that I can only meet one student per day, so it will take ~2 months to measure the height of all the students.

Instead you would like to estimate the mean based on a subset of the students.

We use these ideas to perform statistical statistics when we want to infer something about a population based on observations of a sample of that population.

For example, you would like to know the mean height of all the students enrolled on the Applied Research program.

However, the current coronavirus restrictions mean that I can only meet one student per day, so it will take ~2 months to measure the height of all the students.

Instead you would like to estimate the mean based on a subset of the students.

We use these ideas to perform statistical statistics when we want to infer something about a population based on observations of a sample of that population.

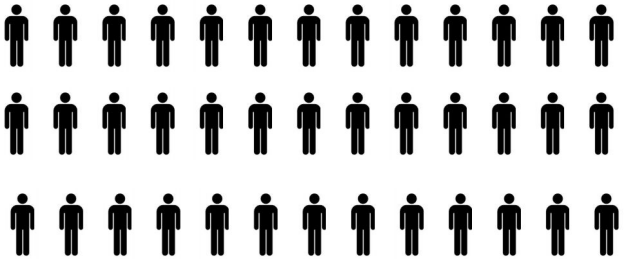
For example, you would like to know the mean height of all the students enrolled on the Applied Research program.

However, the current coronavirus restrictions mean that I can only meet one student per day, so it will take ~2 months to measure the height of all the students.

Instead you would like to estimate the mean based on a subset of the students.

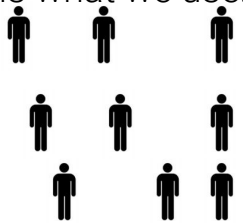
"...The science of drawing conclusion about population from a random sample..."

This is what we want to know...



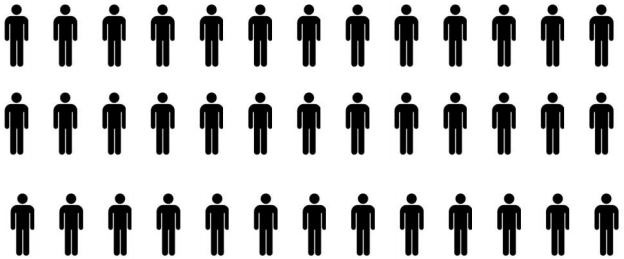
Random selection

This is what we use...



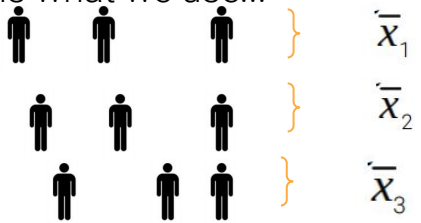
"...The science of drawing conclusion about population from a random sample..."

This is what we want to know...



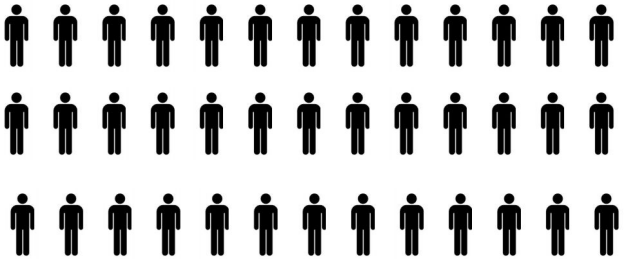
Random selection

This is what we use...



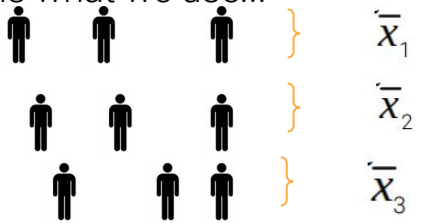
"...The science of drawing conclusion about population from a random sample..."

This is what we want to know...

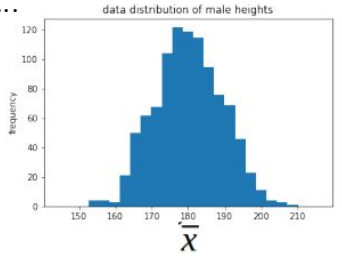


Random selection

This is what we use...

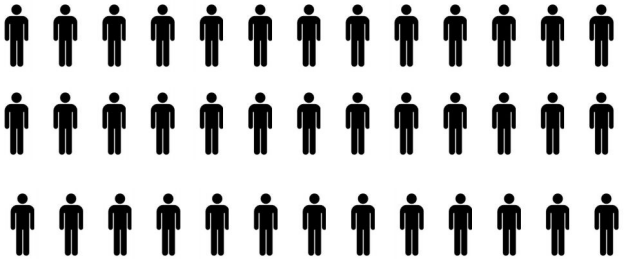


The sampling distribution of the mean....



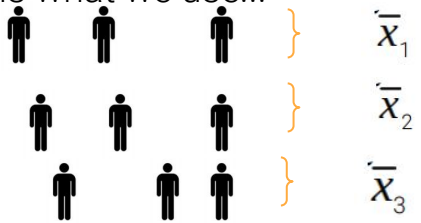
"...The science of drawing conclusion about population from a random sample..."

This is what we want to know...

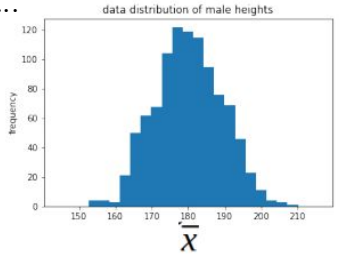


Random selection

This is what we use...



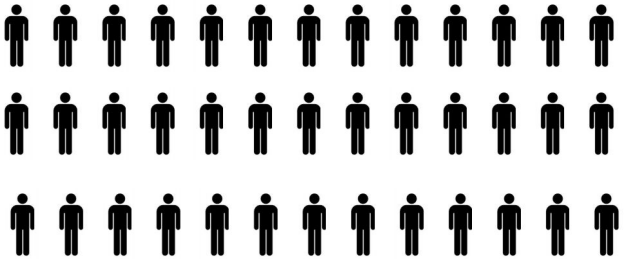
The sampling distribution of the mean....



Drawn inferences on

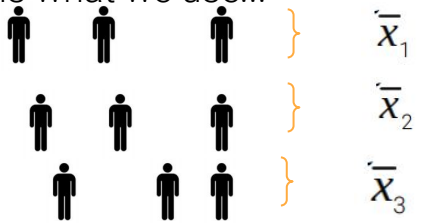
"...The science of drawing conclusion about population from a random sample..."

This is what we want to know...

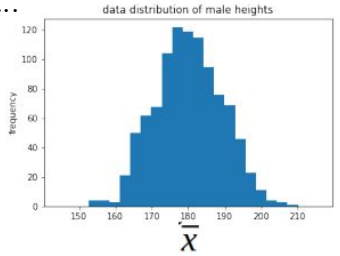


Random selection

This is what we use...

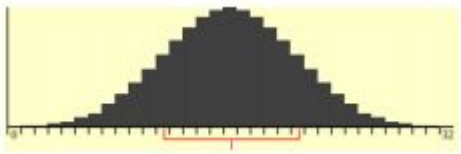


The sampling distribution of the mean....



Drawn inferences on

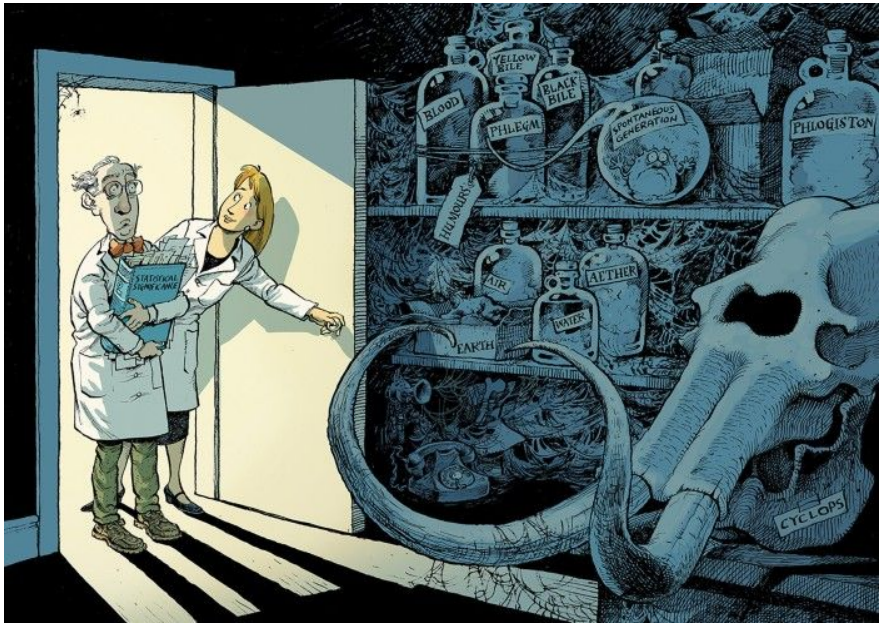
Population distribution of the mean



"...putting all together..."

https://onlinestatbook.com/stat_sim/sampling_dist/

Hypothesis testing and significance test for comparing means



Read: [Statistics, probability, significance, likelihood: words mean what we define them to mean]

Reminder: Using inferential statistics we want to infer something about a population based on observations of a sample of that population.

However, it might be that we know, or assume, something (e.g. the mean value) about the population and we would like to test if an observed sample matches this assumption.

The statistics of the population are typically not the same as the statistics of the sample.

With a null **hypothesis significance test**, we aim to test if the difference between population and sample occur due to chance (i.e., sampling error) or if there could be a systematic reason (e.g., the sample is from different population).

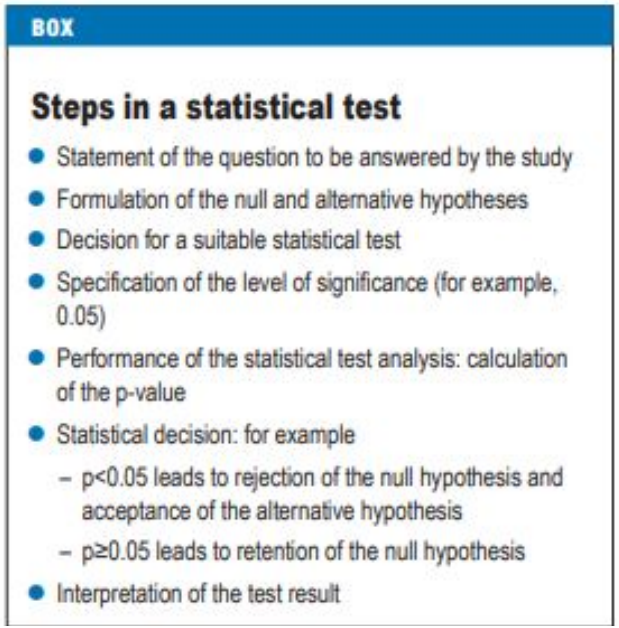
Example

..Maastricht has many nice bars like "*Coffee lovers*" that serve juices to take away, especially in warm months...

The one in the city center, one of my favorites, offers officially 33cl juices (or they say so..)

"I am convinced that Coffee Lovers orange juices are not truly 33cl."

How can I find out?
How can I formulate my previous belief into a formal statistical hypothesis?
How can I collect data to test the hypothesis?



BOX

Steps in a statistical test

- Statement of the question to be answered by the study
- Formulation of the null and alternative hypotheses
- Decision for a suitable statistical test
- Specification of the level of significance (for example, 0.05)
- Performance of the statistical test analysis: calculation of the p-value
- Statistical decision: for example
  - $p<0.05$ leads to rejection of the null hypothesis and acceptance of the alternative hypothesis
  - $p \geq 0.05$ leads to retention of the null hypothesis
- Interpretation of the test result

Read: Read: Choosing Statistical Tests (Jean-Baptist et al., 2010)

Step. Formulate statistical hypothesis

- Null hypothesis testing is a statistical framework where one hypothesis (H_0) is tested to defend the other, alternative hypothesis (H_a).

- There are two kinds of hypothesis:

H_0: $\mu = \mu_0$ (this represents the null hypothesis)

H_A: $\mu \neq \mu_0$ (this represents the alternative hypothesis)

The objective is to check/test if the observed mean value from a sample is the same as the assumed population.

Step. Formulate statistical hypothesis

The mean amount poured in 33cl orange juice by the shops is *not* equal to 33cl:

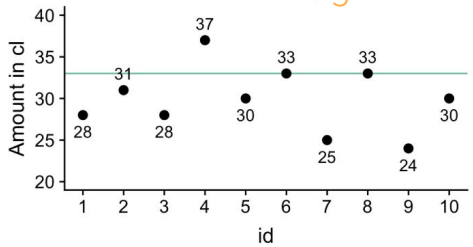
H_0: μ = 33,

H_a: μ ≠ 33

Orange Data

I ordered 10 oranges, and measured the exact amount in each cup, here the results:

Collected data 10 oranges from Shop



Step. Formulate statistical hypothesis

- Given a particular null hypothesis, we need to state an alternative hypothesis, which we assume to be true in the case that data do not support the null hypothesis (H_a).

- Three cases of alternative hypothesis are possible:

H_A: $\mu \neq \mu_0$ **(two-tailed test)**

H_A: $\mu \geq \mu_0$ **(one-tailed test)**

H_A: $\mu \leq \mu_0$ **(one-tailed test)**

Step. Formulate statistical hypothesis

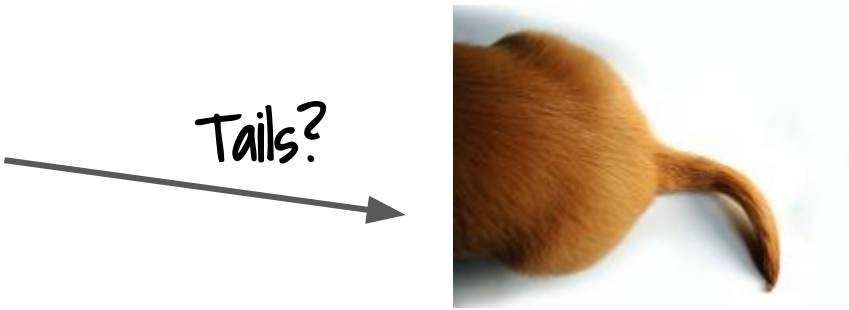
- Given a particular null hypothesis, we need to state an alternative hypothesis, which we assume to be true in the case that data do not support the null hypothesis (H_a).

- **Three cases of alternative hypothesis are possible:**

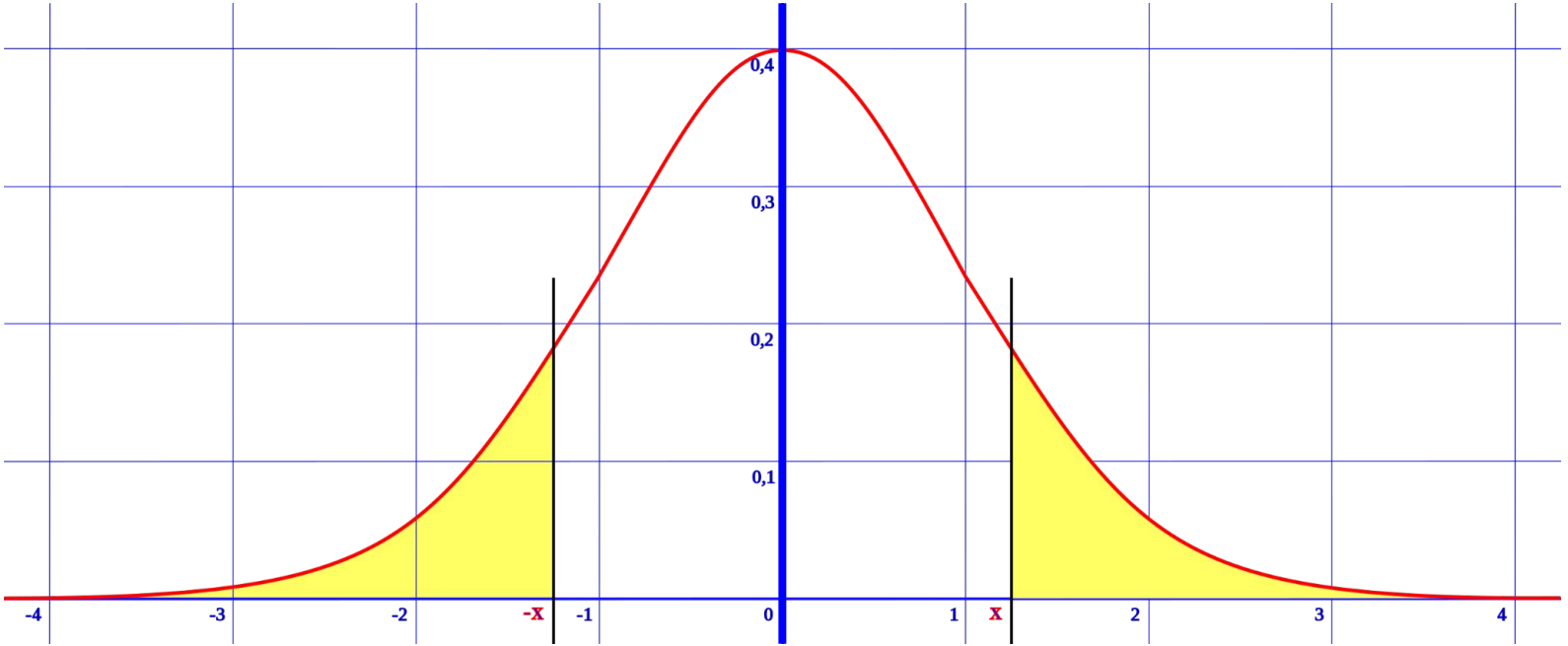
H_A: $\mu \neq \mu_0$ **(two-tailed test)**

H_A: $\mu \geq \mu_0$ **(one-tailed test)**

H_A: $\mu \leq \mu_0$ **(one-tailed test)**

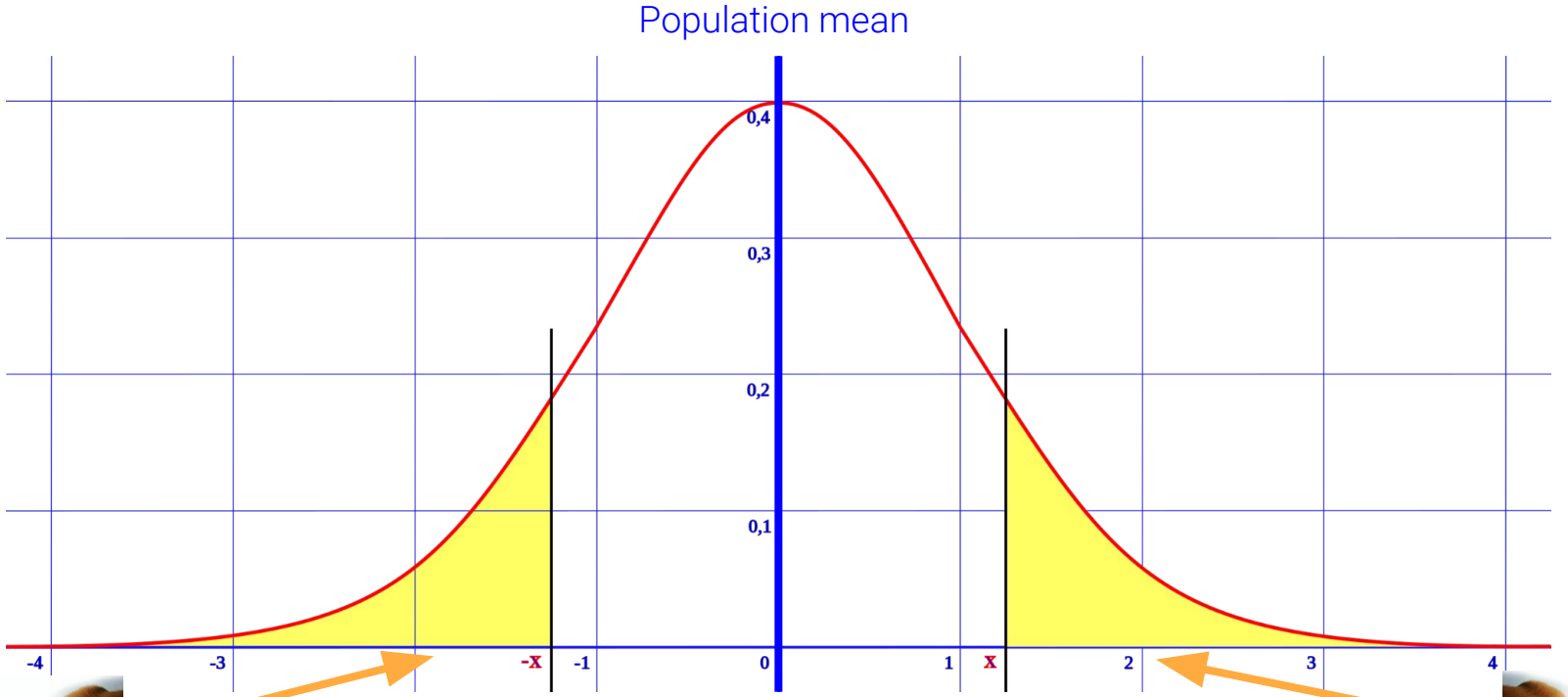


Population mean

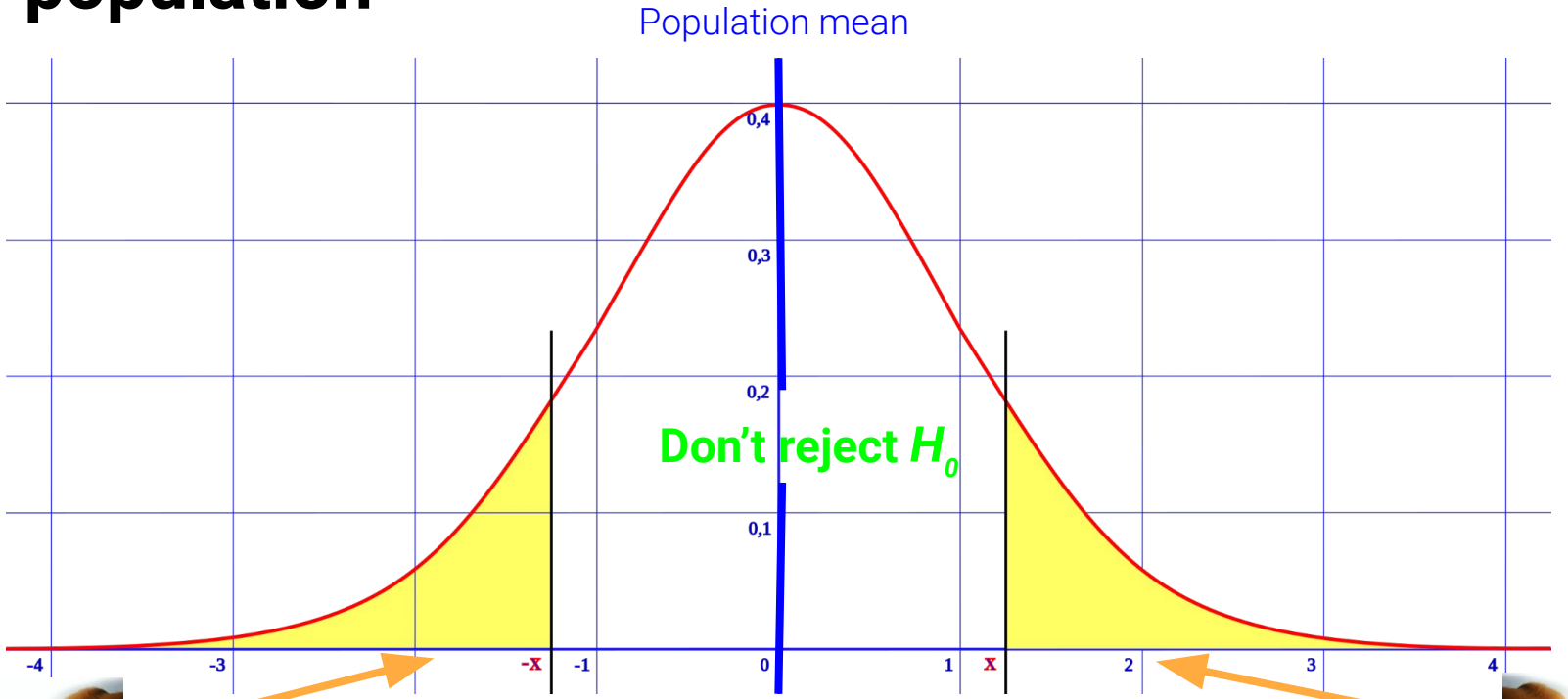


95% of sample means

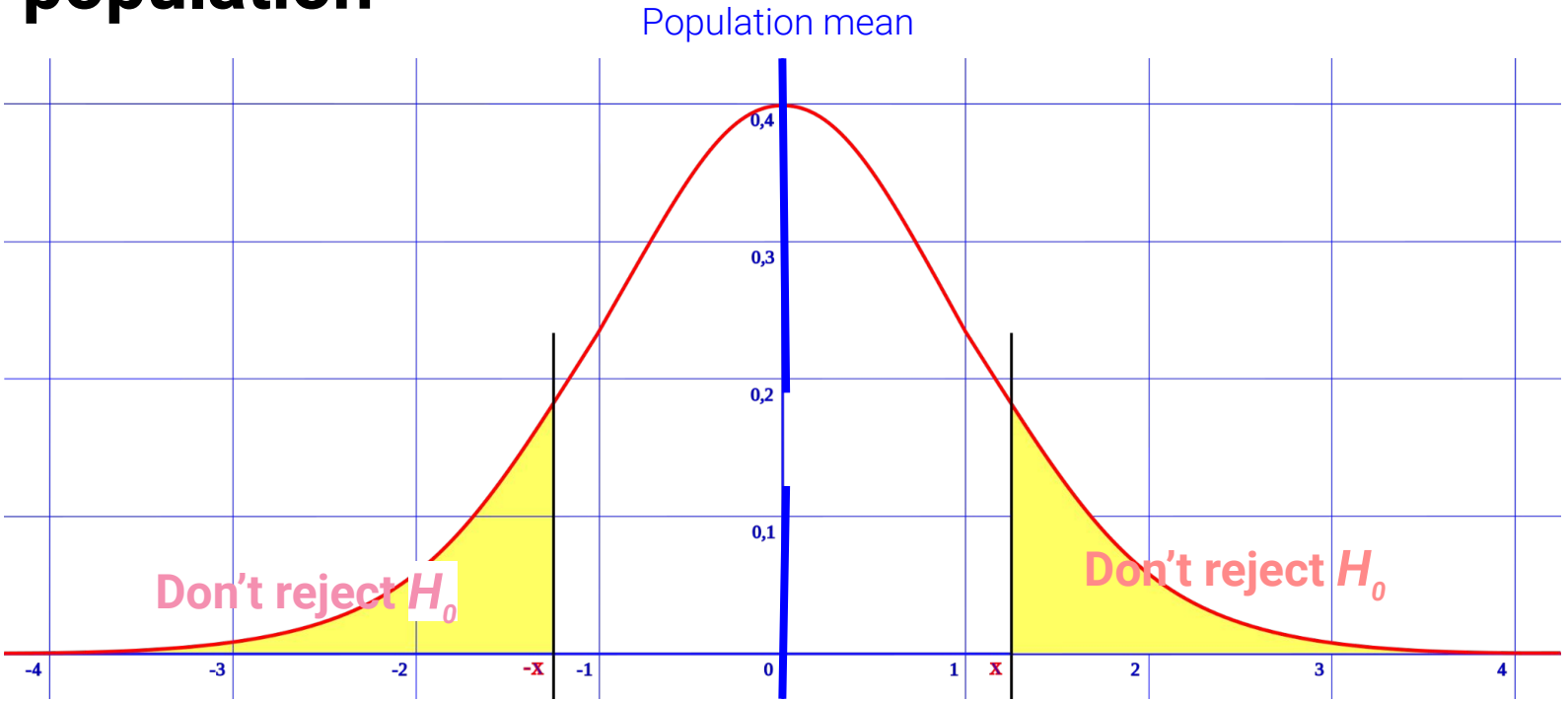
Tails as extreme values from a normal distribution



Acceptance region, sample mean has been drawn from population



Critical region, sample mean has not been drawn from population



Step. State *threshold* as the level of significance

The level of probability that determines which sample means are considered 'acceptable' is denoted α and its know as significance level

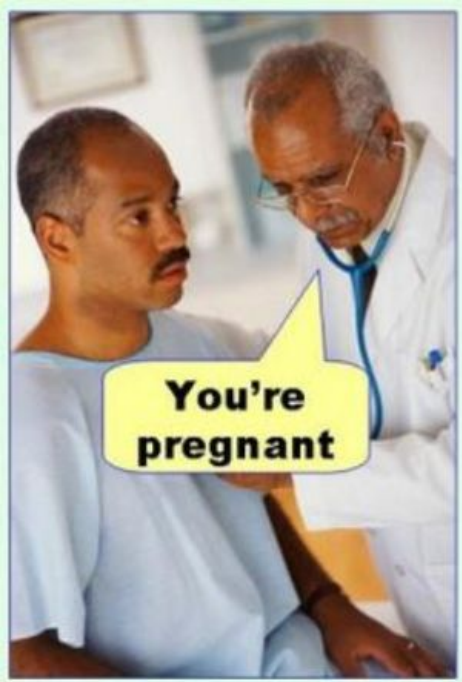
α = 0.05 means that we will reject the null hypothesis
5% of the time when the null hypothesis is actually true

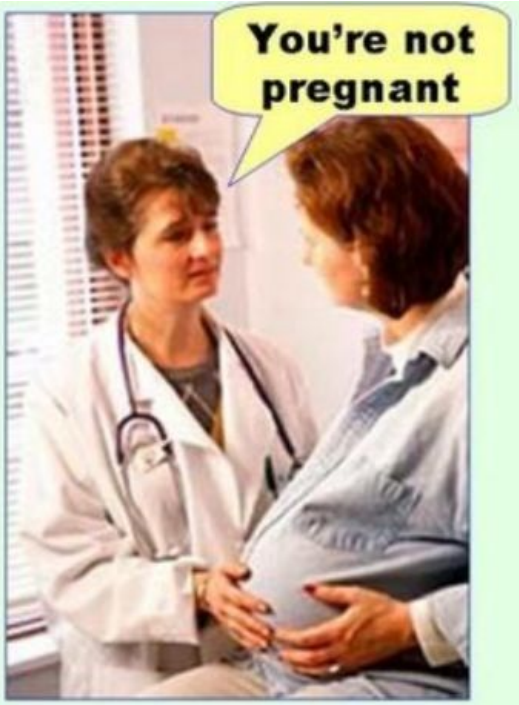
	don't reject (H_0)	reject (H_0)
H_0 is true	correct inference	*type I error (false positive)*
H_0 is false	*type II error (false negative)*	correct inference

$$\alpha = \frac{\text{False positives}}{\text{True negative} + \text{False positives}}$$

1. Formulate statistical hypothesis: Applied

- Null hypothesis is: *"You are not pregnant"* (commonly accepted as *'boring'* result)



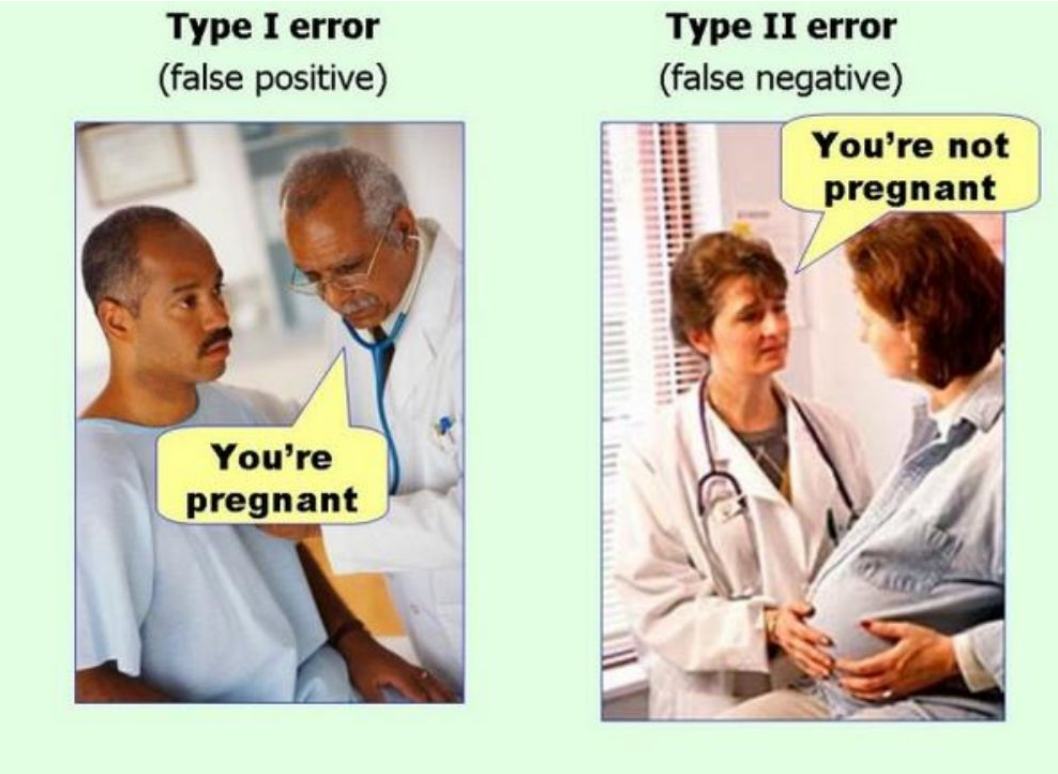


?

Type I error
(false positive)

Type II error
(false negative)

1. Formulate statistical hypothesis: Applied



Step. Compute your test statistic and p-value

- T-test is a inferential statistical procedure that determines whether there is statistically significant difference between the means.

T-test for comparing means	Applications
One-sample	Is females score higher than average 70 on a exam?
Independent samples (unpaired)	Different participants in the two groups
Dependent samples (paired)	Same participants in *before* group and *after* group.

```
> ?t.test
```

t.test {stats} R Documentation

Student's t-Test

Student's t-test

The t-statistic was developed and published in 1908 by William Gosset, a chemist/statistician working for the Guinness brewery in Dublin. Guinness employees were not allowed to publish under their own name, so Gosset published under the pseudonym "Student".

Description

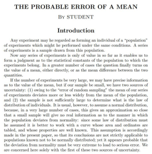
Performs one and two sample t-tests on vectors of data.

Usage

```
t.test(x, ...)

## Default S3 method:
t.test(x, y = NULL,
       alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = FALSE, var.equal = FALSE,
       conf.level = 0.95, ...)

## S3 method for class 'formula'
t.test(formula, data, subset, na.action, ...)
```



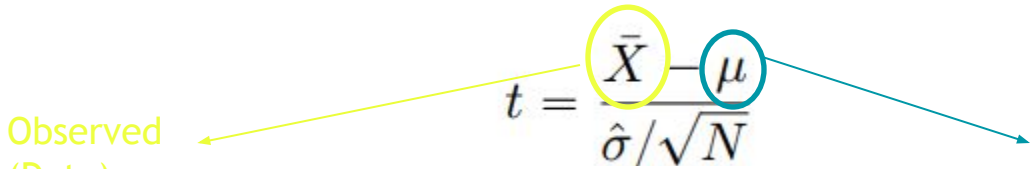
Arguments

`x`	a (non-empty) numeric vector of data values.
`y`	an optional (non-empty) numeric vector of data values.
`alternative`	a character string specifying the alternative hypothesis, must be one of `"two.sided"` (default), `"greater"` or `"less"`. You can specify just the initial letter.
`mu`	a number indicating the true value of the mean (or difference in means if you are performing a two sample test).

Step. Compute your test statistic

Equation for a one-sample* *t*-test

*one-sample = is the sample mean different from a known or predefined population mean (e.g. an exam score of 70)



where

t = the t statistic

$\bar{x}$ = the mean of the sample

μ = the comparison mean

$\hat{\sigma}$ = the sample standard deviation

n = the sample size

Step. Compute your test statistic

Equation for an independent samples* *t*-test

*Independent samples = different participants in the two groups (two samples)

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\mathrm{SE}}$$

$$\mathrm{SE} = \frac{\sigma}{\sqrt{n}}$$

SE = standard error of the sample

σ = sample standard deviation

n = number of samples

Step. Compute your test statistic

The mean amount poured in 33cl orange juice by the shops is *less* than 33cl:

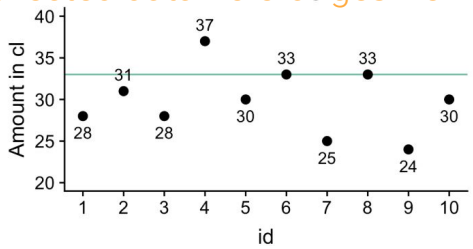
H_0: μ = 33,

H_a: μ < 33

Orange Data

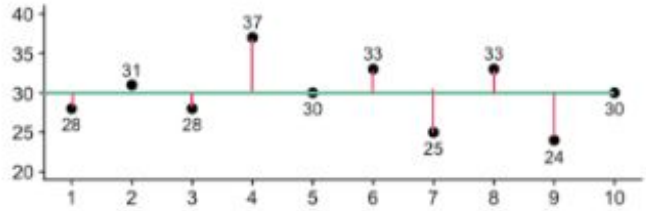
I ordered 10 oranges, and measured the exact amount in each cup, here the results:

Collected data 10 oranges from Shop



Models $H_0 : \mu = 33,\ H_1 : \mu < 33$

Raw Data



Sample Statistics

Mean $\bar{x} = 29.9$

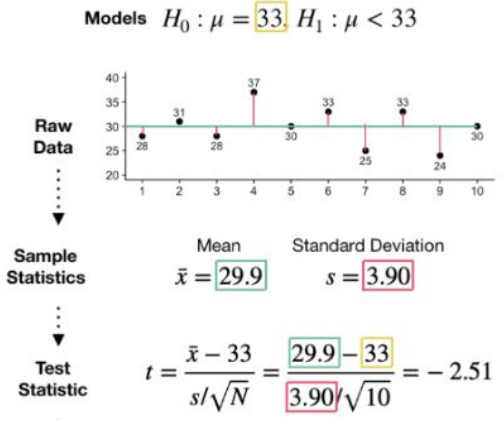
Standard Deviation $s = 3.90$

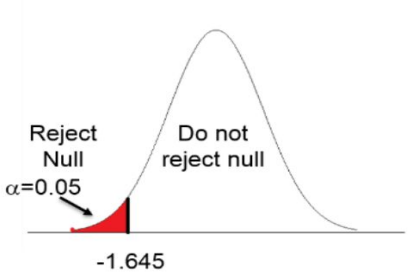
Test Statistic

$$t = \frac{\bar{x} - 33}{s/\sqrt{N}} = \frac{29.9 - 33}{3.90/\sqrt{10}} = -2.51$$

Make a statistical decision with the test statistics

Steps 1 through 3	Result
Null (H_0)	Mean is equal to 33
Alternative (H_a)	Mean is less 33
Level significance (α)	0.05 level
Test statistic (t)	-2.51
Critical values	[-1.645]
Conclusion	*t fall into critical region, we reject the Ho in favor of Ha. We conclude that the null hypothesis is likely to be wrong and they are pondering less than 33cl*





The infamous P-value

P-VALUE

The probability, computed assuming that H_0 is true, that the test statistic would take a value as extreme or more extreme than that actually observed is called the ***P*-value** of the test. The smaller the *P*-value, the stronger the evidence against H_0 provided by the data.

Definition, pg 405
Introduction to the Practice of Statistics, Fifth Edition

https://www.tldrpharmacy.com/content/everything-you-know-about-the-p-value-is-wrong

Step. Make a statistical decision with t statistics and p-values

Making decisions regarding the significant level (alpha) and p-value

Scenario	Description	Decision	Interpretation
p-value < α	We have an evidence to reject Ho in favor of Ha	Reject Ho	Our results are statistically significant
p - value > α	We do not have an evidence to reject Ho in favor of Ha	Do not reject Ho	Our results are not statistically significant from the Ho

Live demo

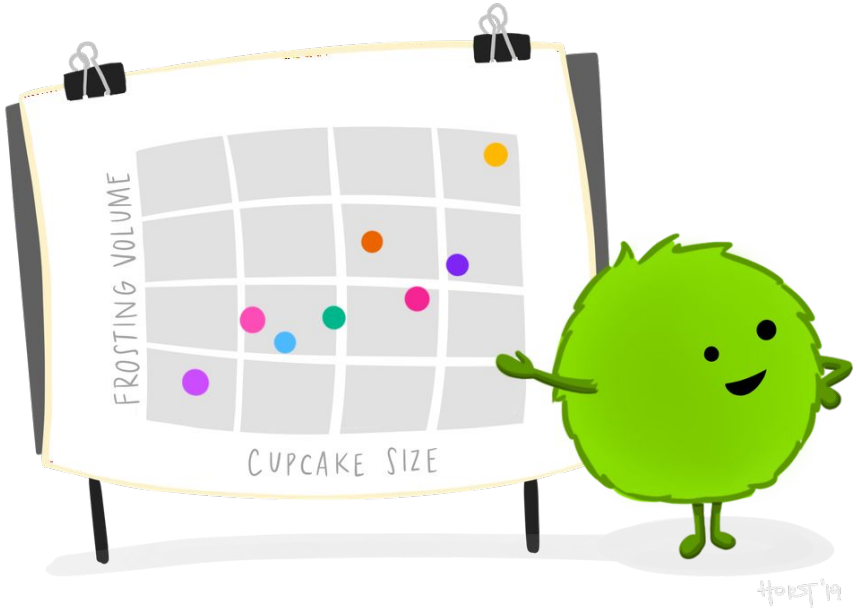
Is there any (statistical) difference in the mean orange juice poured in a glass of 33cl between **Maastricht** and **Amsterdam** coffee lovers?

N = 10 (sample size)
Mean.M = Mean volume (cl)
Maastricht Orange juice

N = 10 (sample size)
Mean.A = Mean volume (cl)
Amsterdam Orange juice

Is there any (statistical) difference in the mean Life Expectancy between **South Africa** and **Ireland**?

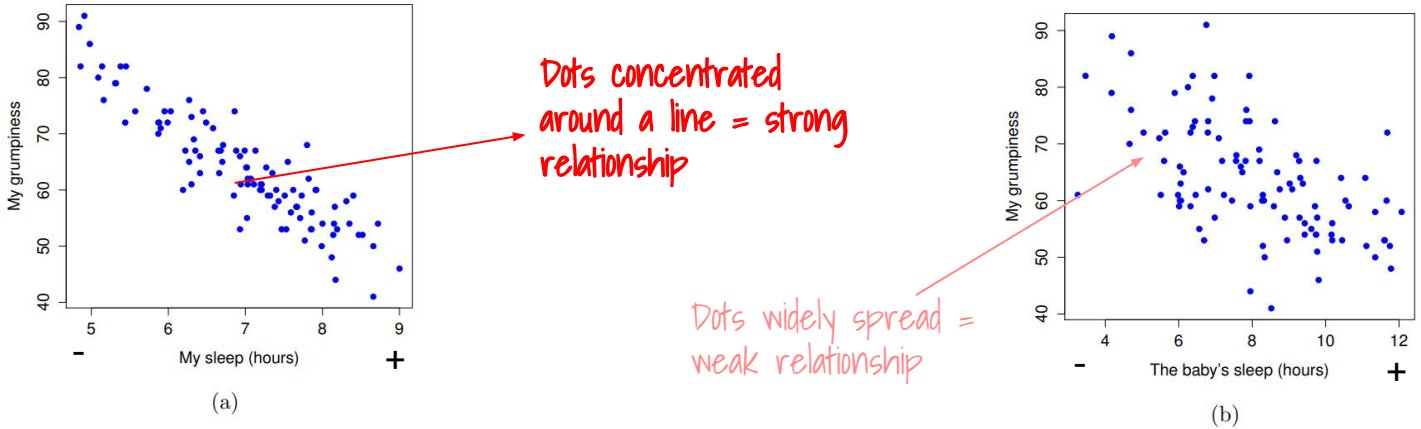
Correlations and interpretation



So far we have asked research questions such as *"is there any differences between"* What if I ask question like *"is there any relationship or association between"*?

The strength and direction of a relationship

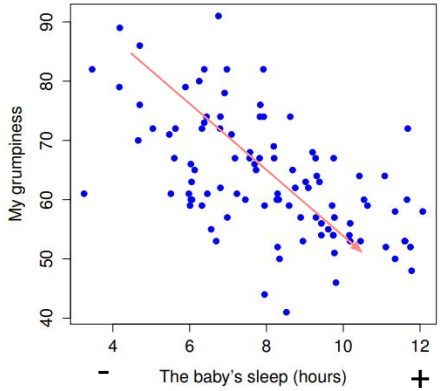
- The relationship is qualitatively the same in both cases: more sleep equals less grumpy mood!
- Relationship between my sleep hours and my grumpy mood (*figure a*) is stronger than my nieces sleep hours and my grumpy mood (*figure b*).



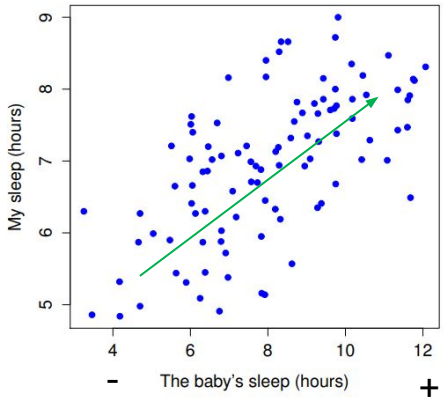
Scatterplots showing the relationship between my sleep hours and my grumpy mood (a) and the relationship between my niece sleep hours and my grumpy mood (b)

The strength and direction of a relationship

- The overall strength relationship is the same, but the direction is different.
- If she sleeps more then, I get less grumpy - negative relationship - figure (a)
- If my niece sleeps more, I get more sleep - positive relationship -figure (b)



(a)

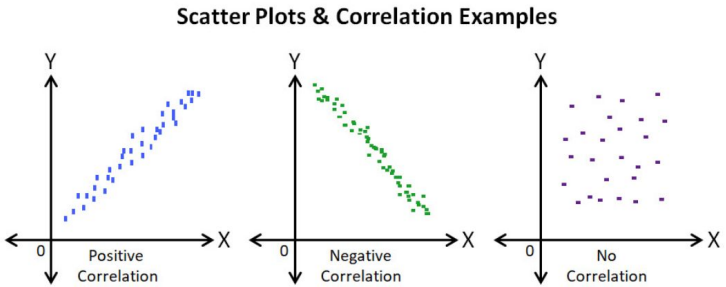


(b)

Scatterplots showing the relationship between my niece sleep hours and my grumpy mood (a) and the relationship between my niece sleep hours and my sleep hours (b)

The correlation coefficient

- The correlation coefficient (or Pearson's correlation coefficient *r*) measures the strength of the linear relationship between two continuous variables (sometimes denoted r_{xy})
- r is always a number between -1 and 1
- r > 0, indicates a positive association
- r < 0, indicates a negative association
- r = -1, indicates perfect negative relationship
- r = 1, indicates perfect positive relationship
- r = 0, indicates there is not relationship



Pearson's correlation coefficient (r_{xy})

- Pearson's correlation coefficient between two variables is defined as the covariance of the two variables divided by the product of their standard deviations:

$$r_{XY} = \frac{\operatorname{Cov}(X,Y)}{\hat{\sigma}_X \, \hat{\sigma}_Y}$$

- Covariance is a measure of the (average) co-variation between two variables, say x and y. In other words, it measures the degree to which two variables are linearly associated.

$$\operatorname{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Co-variance (x,y)

Pearson's correlation coefficient (r_{xy})

- The covariance captures the basic qualitative idea of correlation:
    - if the relationship is negative then, the covariance is also negative
    - if the relationship is positive then, the covariance is also positive
- The covariance is difficult to interpret: expressed in X and Y units
- Thus Pearson correlation r fixed this interpretation problem with meaningful scale:
    - r = 1 implies a perfect positive relationship
    - r = -1 implies a perfect negative relationship

Live demo

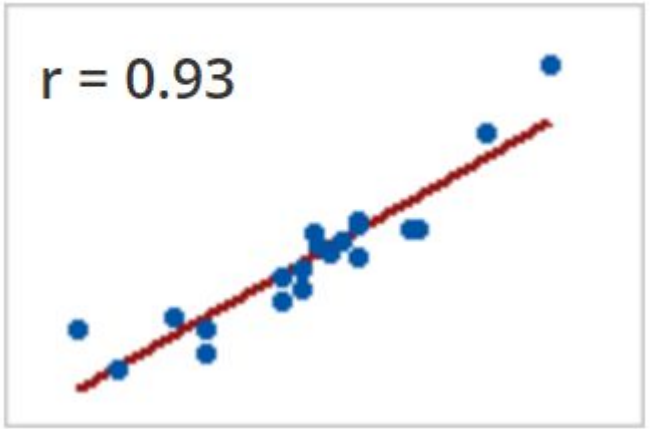
Interpreting the correlation

Correlation	Strength	Direction
-1.0 to -0.9	Very strong	Negative
-0.9 to -0.7	Strong	Negative
-0.7 to -0.4	Moderate	Negative
-0.4 to -0.2	Weak	Negative
-0.2 to 0	Negligible	Negative
0 to 0.2	Negligible	Positive
0.2 to 0.4	Weak	Positive
0.4 to 0.7	Moderate	Positive
0.7 to 0.9	Strong	Positive
0.9 to 1.0	Very strong	Positive

No relationship: Pearson r = 0

Interpreting the correlation

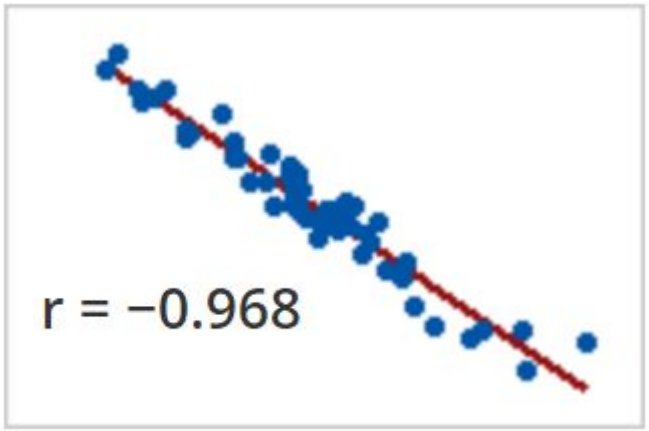
Correlation	Strength	Direction
-1.0 to -0.9	Very strong	Negative
-0.9 to -0.7	Strong	Negative
-0.7 to -0.4	Moderate	Negative
-0.4 to -0.2	Weak	Negative
-0.2 to 0	Negligible	Negative
0 to 0.2	Negligible	Positive
0.2 to 0.4	Weak	Positive
0.4 to 0.7	Moderate	Positive
0.7 to 0.9	Strong	Positive
0.9 to 1.0	Very strong	Positive



Large positive relationship

Interpreting the correlation

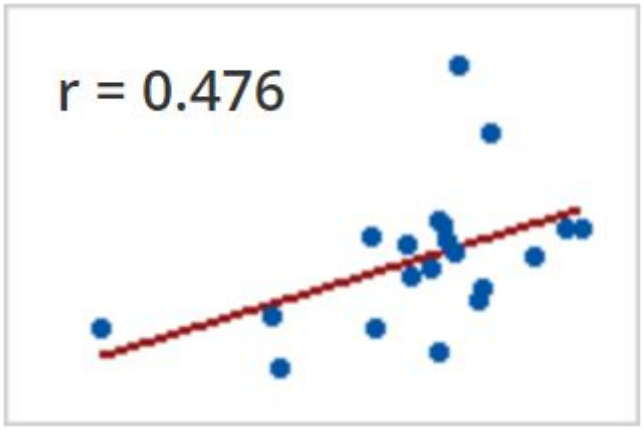
Correlation	Strength	Direction
-1.0 to -0.9	Very strong	Negative
-0.9 to -0.7	Strong	Negative
-0.7 to -0.4	Moderate	Negative
-0.4 to -0.2	Weak	Negative
-0.2 to 0	Negligible	Negative
0 to 0.2	Negligible	Positive
0.2 to 0.4	Weak	Positive
0.4 to 0.7	Moderate	Positive
0.7 to 0.9	Strong	Positive
0.9 to 1.0	Very strong	Positive



Large negative relationship

Interpreting the correlation

Correlation	Strength	Direction
-1.0 to -0.9	Very strong	Negative
-0.9 to -0.7	Strong	Negative
-0.7 to -0.4	Moderate	Negative
-0.4 to -0.2	Weak	Negative
-0.2 to 0	Negligible	Negative
0 to 0.2	Negligible	Positive
0.2 to 0.4	Weak	Positive
0.4 to 0.7	Moderate	Positive
0.7 to 0.9	Strong	Positive
0.9 to 1.0	Very strong	Positive



Moderate positive relationship

The takeaway

- A t-test is a statistical procedure to comparing one (or two) means.
- A one-sample t-test determines the differences between one sample and the population true mean.
- An independent sample t-test determines the differences between two groups with different participants in each group.
- The Pearson correlation coefficient (r) is a numerical expression of the relationship between two numerical variables.
- r can be vary from -1.0 to 1.0, and the closer it is to -1.0 or 1.0, the stronger correlation.