Intro to Inferential Statistics with R

Workshop 3

Course: VSK1004 Applied Researcher

Maastricht University
Campus Venlo

 c.utrillaguerrero@maastrichtuniversity.nl

Workshop structure

Intro to Statistic Inference (TODAY)	More about inferential statisitcs	Linear and Logistic regression
1. Descriptive vs Inferential statistics 2. Population, sample and sampling distribution 3. Hypothesis testing a. One-sample b. Independent t-test 4. Correlation and interpretation	1. Choosing statitistical model 2. Dependent t-test 3. ANOVA **4. Chi-squared distribution**	1. Model assumptions 2. Interpretations

Our goal in the next 40 min

In this session, we will cover some of the **basic principles of statistical inference**.

1. Descriptive vs Inferential statistics
2. Population, sample and sampling distribution
3. Hypothesis testing:
   a. One-Sample t-test
   b. Independent t-test
4. Correlation and interpretation

1. Descriptive vs. Inferential

From describing to infering

Descriptive A random sample of 10 basketball players will be drawn, whose height will be measured in mts. *Table 1* displays the relevant dispersion measures *(Covered in workshop 2)*

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

Inferential Investigate whether or not tennis players are smarter than volleyball players

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)

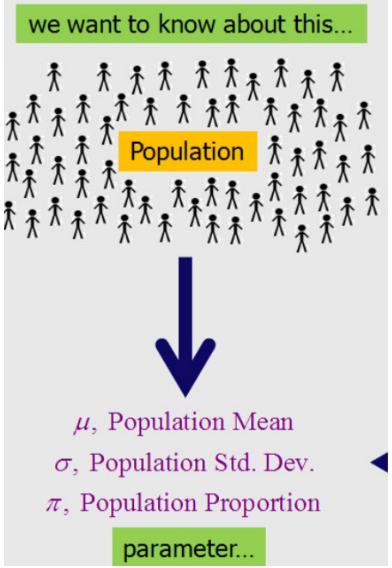
2. Population, sample and sampling distribution

Average height of the ALL residents of India

- Population (N)
  - all possible values, or individuals, you are interested in
  - exists, but the **parameters** are unknown
  - **population distribution**, the variation in the values in a population (e.g. population mean, Std Dev)

Sample
  - subset of values drawn from the population
  - exists, and its parameters are known
  - **sample distribution**, the variation in the values in the sample

Sampling distribution *(of the mean)*
  - the means we might get if we took infinite samples of the same size
  - exists, but purely theoretical
  - **sampling distribution** *(of the mean)*, the variation in the sample means



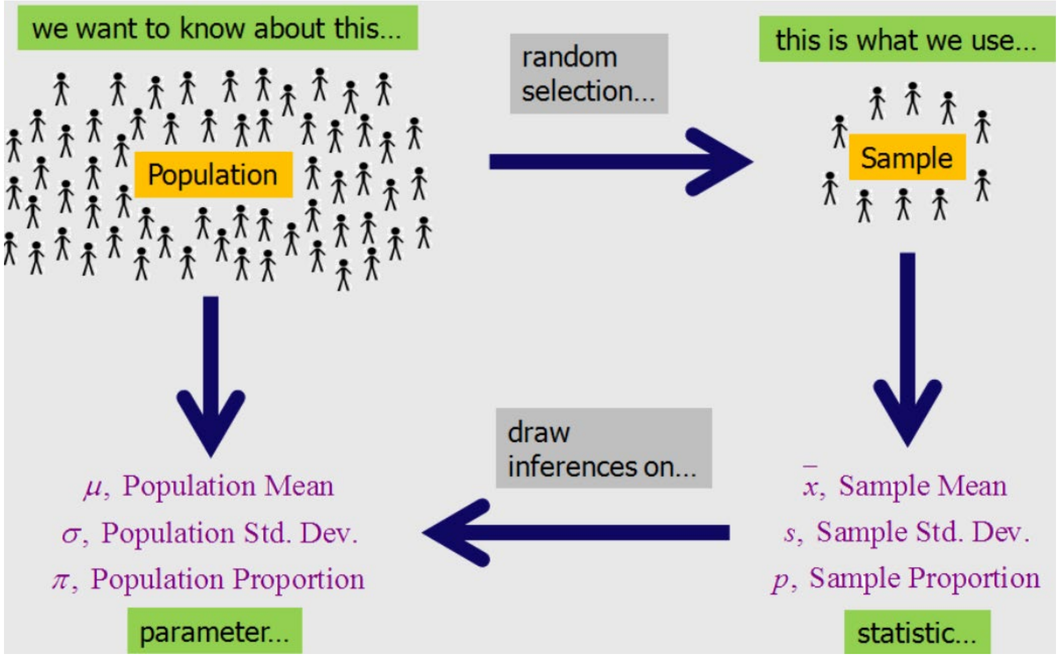
1000 individuals randomly selected from India

Population (n)

- all possible values, or individuals, you are interested in
- exists, but the parameters (e.g. mean of the entire numbers of newborns in North America) are unknown
- **population distribution**, the variation in the values in a population

- Sample (n)
  - subset of values drawn from the population
  - exists, and its parameters are known
  - **sample distribution**, the variation in the values in the sample (e.g. mean, std dv)

The science of drawing conclusion about population from a sample



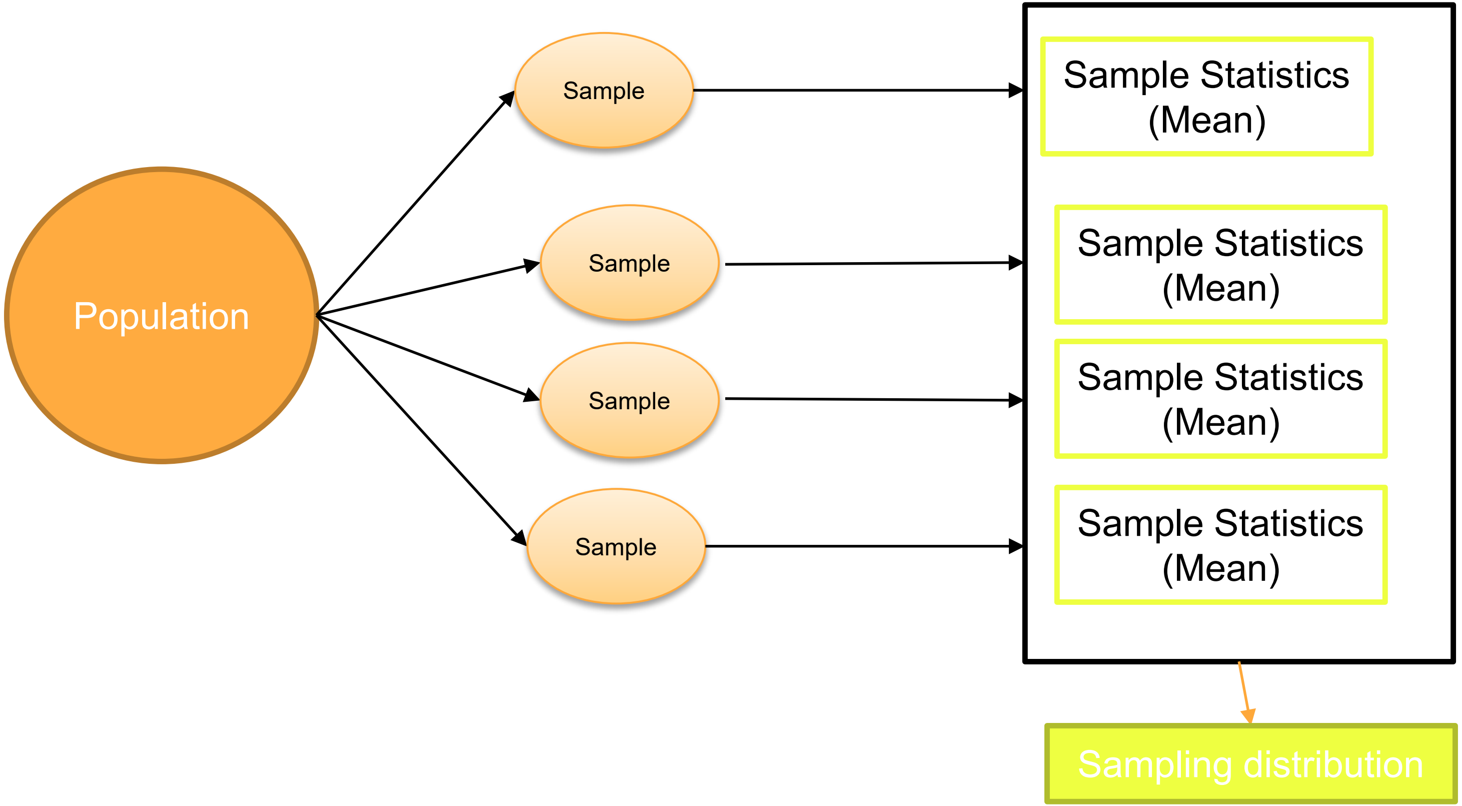
Indefinite number of samples of 1000 respondents

- Population
  - all possible values, or individuals, you are interested in
  - exists, but the parameters (e.g. mean of the entire numbers of newborns in North America) are unknown
  - **population distribution**, the variation in the values in a population

- Sample
  - subset of values drawn from the population
  - exists, and its parameters are known
  - **sample distribution**, the variation in the values in the sample

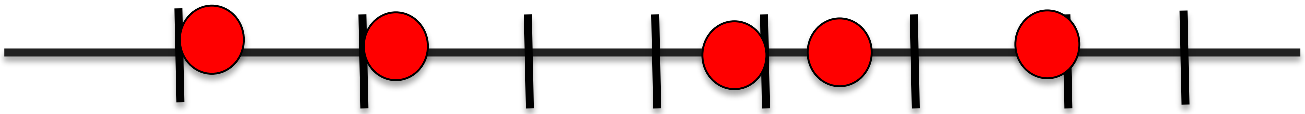
- Sampling distribution *(of the mean)*
  - the means we might get if we took infinite samples of the same size
  - exists, but purely theoretical
  - **sampling distribution** *(of the mean)*, the variation in the sample means (e.g. standard error)

Sampling distribution of the mean

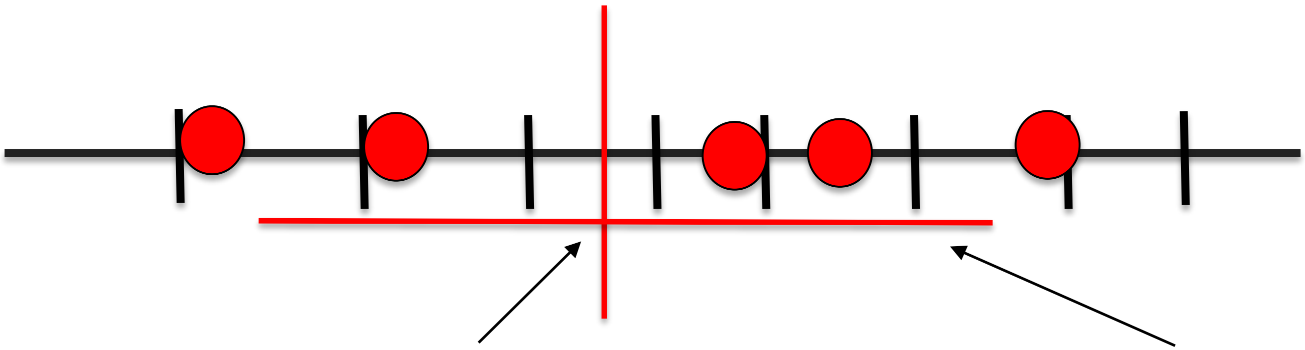


Standard Deviation and Standard Error:

Imagine we weighted 5 mice



We get the mean and standard deviation of the weights

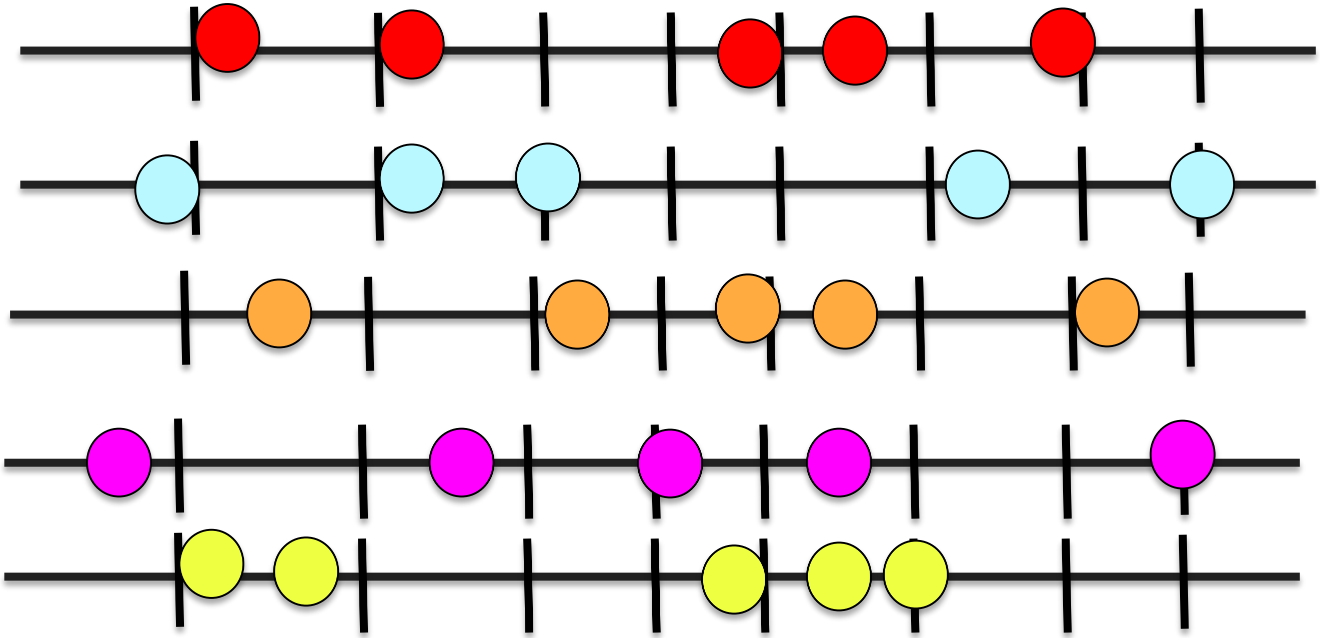


This is the average (mean)

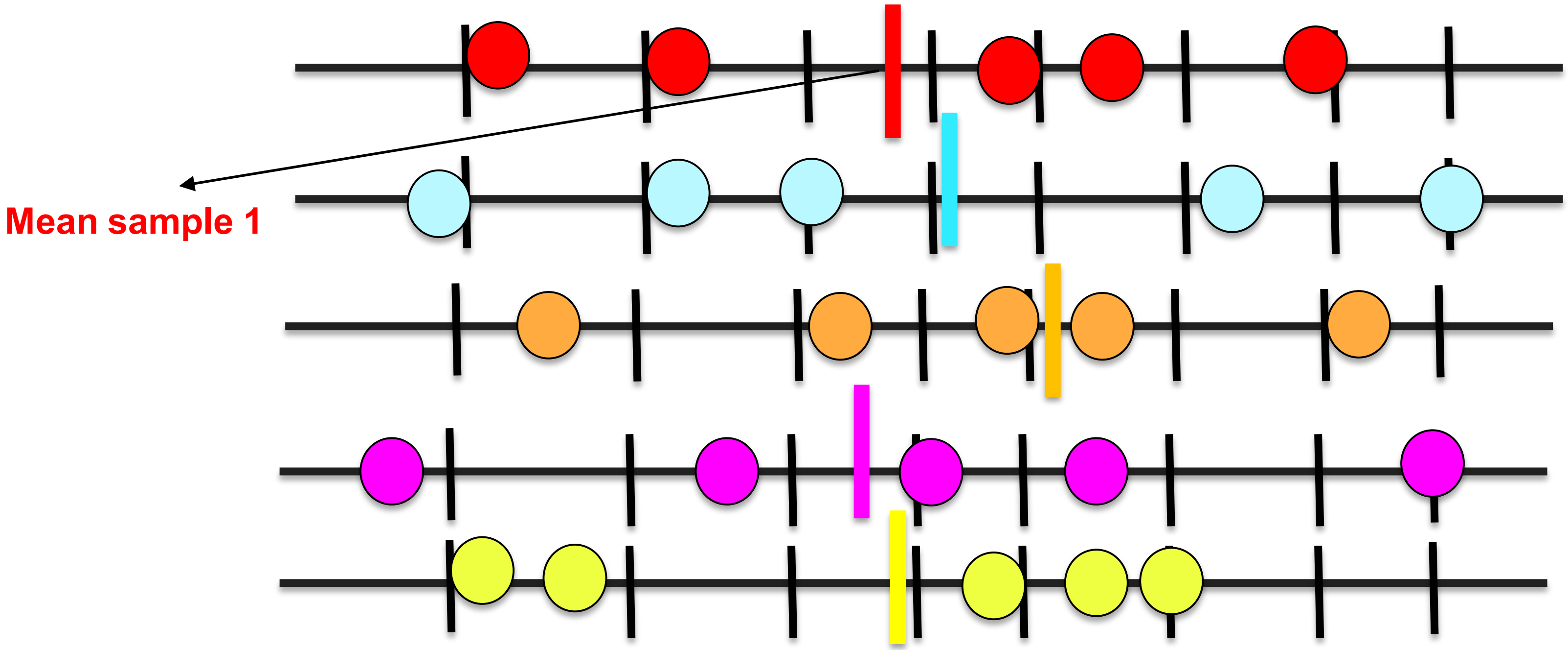
This is the **standard deviation** of both side of the mean

It quantifies how much the data are spread out

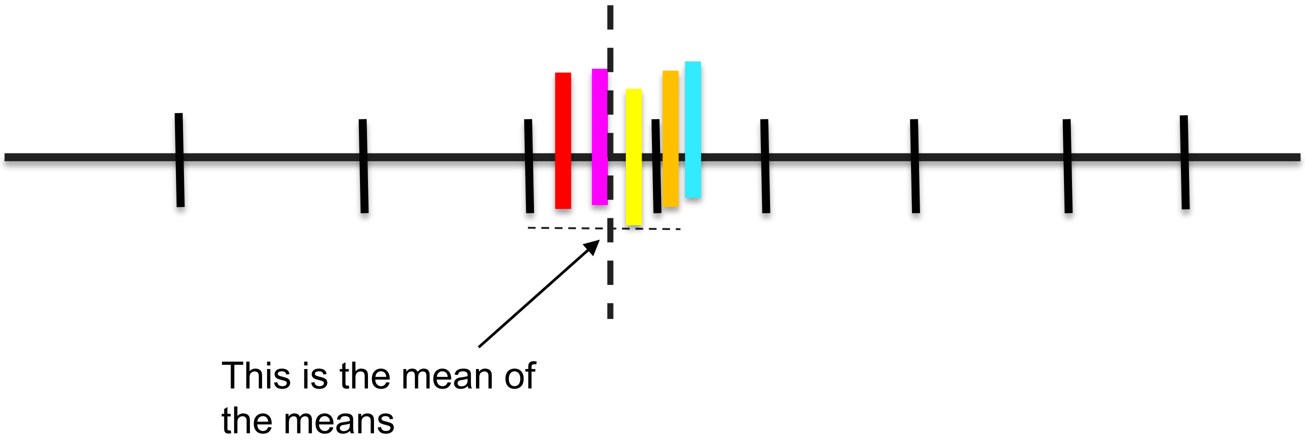
Imagine we did weight 5 mice, 5 separate times



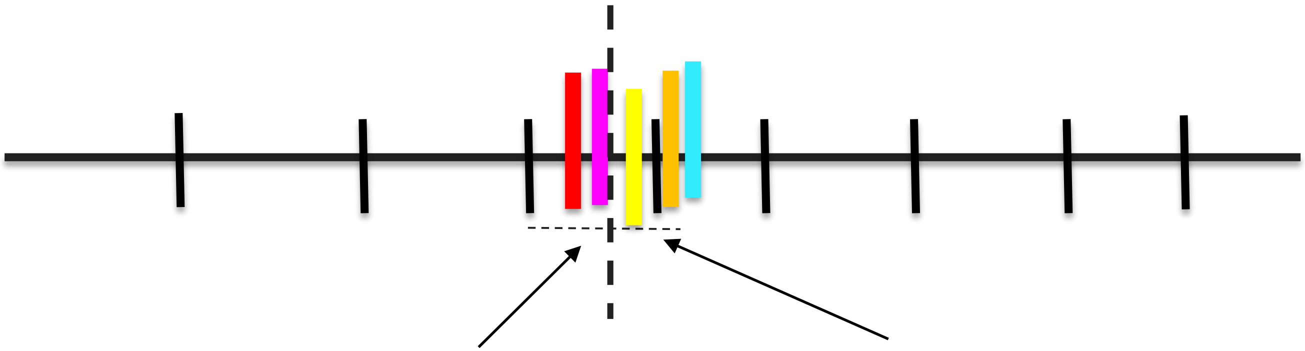
This would result 5 different means and standard deviations, one per sample



Plot 5 means on the same line



The standard deviation of the means is called The Standard Error



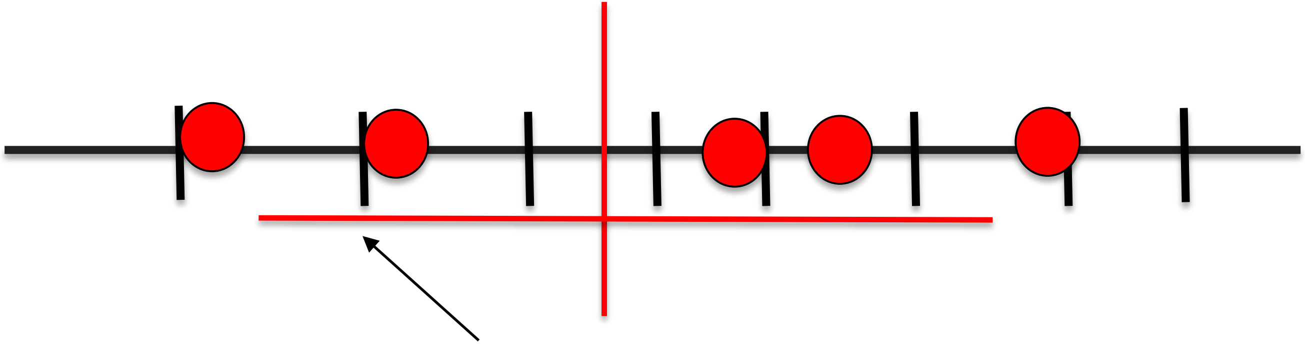
This is the mean of the means

This is the **standard deviation** of both side of the mean of the means

It quantifies how much the means are spread out

The Standard Deviation vs The Standard Error

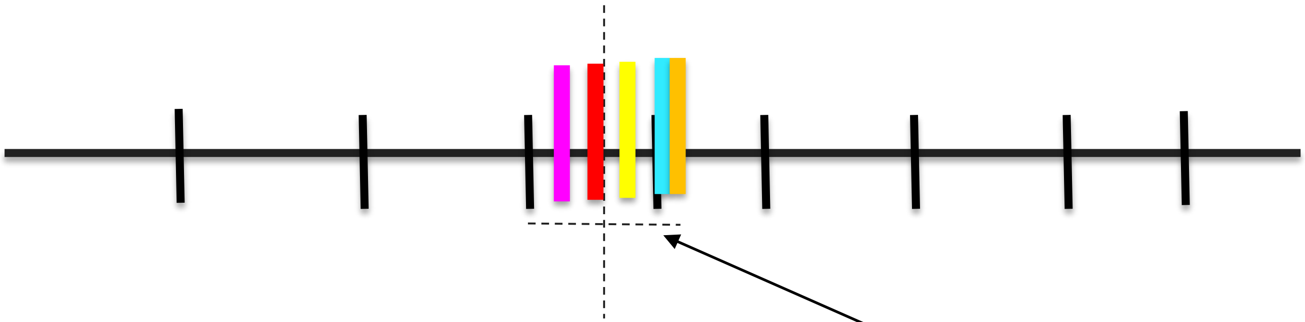
- The standard deviation **quantifies the variation within the data (each sample).**



Standard deviation: tell you how the **data** is distributed around the mean

The Standard Deviation vs The Standard Error

- The standard error quantifies the variation in the means between multiple set of samples



This tells you how the **means** are distributed

3. Hypothesis testing

What is hypothesis testing?

- Hypothesis is a theoretical statement (premise or claim) concerning a certain feature of the studied statistical population that we want to test (investigate)

  e.g. *‘prolonged exposure to loud noise increases systolic blood pressure in the statistical population’*

- Hypothesis testing is a procedure of testing whether sample data is consistent with statements (hypotheses) made about the statistical population.

- Research hypothesis =! Statistical hypothesis

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 33cl juices (or they say so..)

"I am convinced that shops are 'underpouring' its orange juices and they 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?

Steps of hypothesis testing

1. Formulate the statistical hypothesis for the test:
    a. State the null (H_0) and alternative (H_a) hypothesis.
2. Specify the level of significance (alpha, $\alpha = .05$)
3. Compute your test statistic and p-value
4. Make a statistical decision

1. Formulate statistical hypothesis

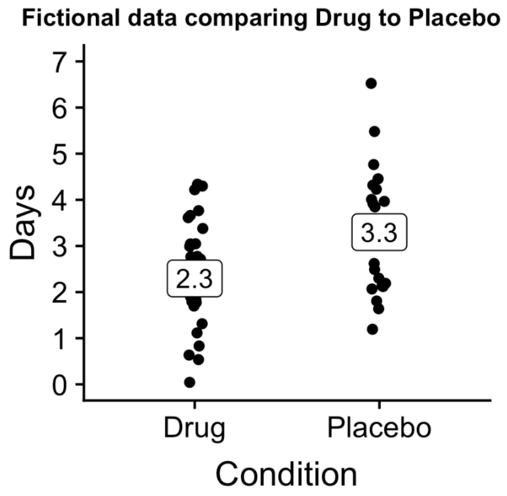
- There are two kinds of hypothesis:
  - H_0: a statement that usually claims zero effect (called "null hypothesis")
    - *E.g. "the mean age of female and males are **NOT** differents."*
  - H_a: a statement that actually want to test (called "alternative hypothesis")
    - *e.g." the mean age of females and males are differents."*

- We want to get answers to questions starting, typically like these:
  - *"Is there any differences between"*
  - *"Is there any relationship"*

1. 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).

Hypothesis	Description	Example
Null (H_0)	A proposed effect does not exist and there is NO variation	Drug and placebo have the same effect
Alternative (H_a)	A proposed effect does exist and there is variation	Drug and placebo do **NOT** have same effect



1. Formulate statistical hypothesis: Applied

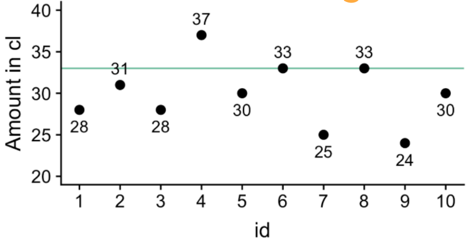
The mean amount poured in 33cl orange juice by the shops is ***not*** equal to 33cl:

H_0: $\mu = 33$, H_a: $\mu \neq 33$

Orange Data

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

Collected data 10 oranges from Shop



Possible outcomes for this test:

- Reject Null Hypothesis
- Don’t reject Null Hypothesis

1. Formulate statistical hypothesis

Potential Errors in hypothesis testing

- *Type I error (false positive):*
  - we reject the Ho although it was actually true
- *Type II error (false negative):*
  - we accept Ho, although it was actually false

	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

1. Formulate statistical hypothesis: Applied

- Null hypothesis is: *“You are not pregnant”* (commonly accepted as ‘*boring*’ result)

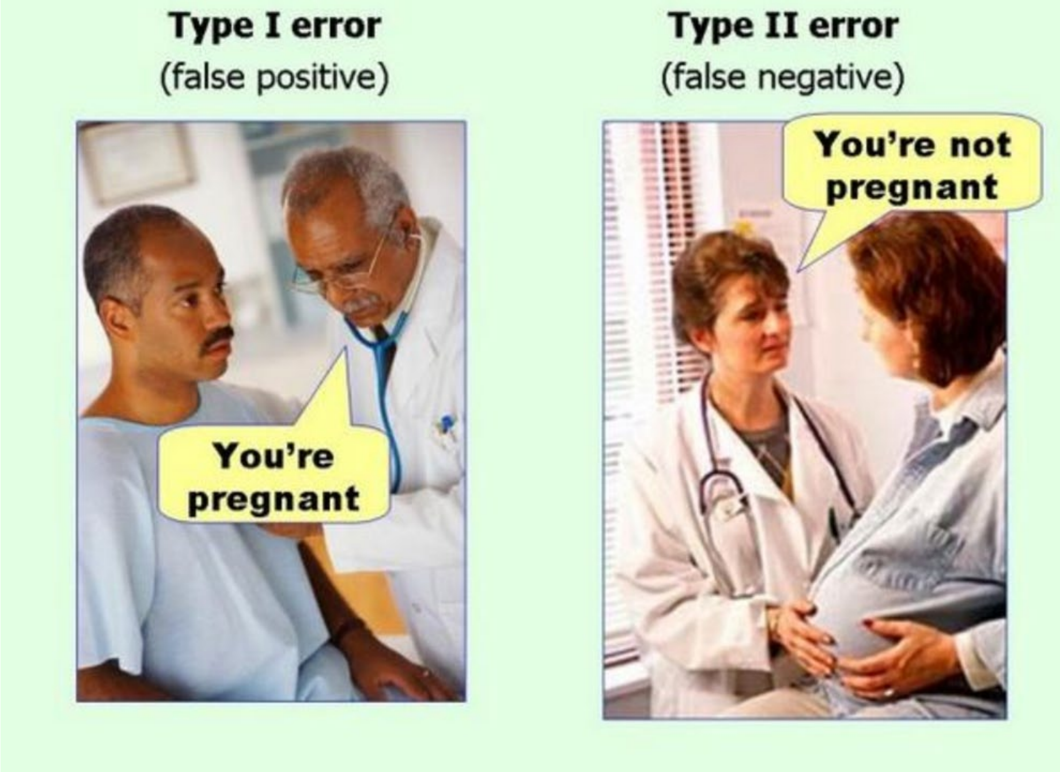
?

Type I error
(false positive)

Type II error
(false negative)

 c.utrillaguerrero@maastrichtuniversity.nl

1. Formulate statistical hypothesis: Applied



1. Formulate statistical hypothesis: Applied

Steps of hypothesis testing

1. Formulate the statistical hypothesis for the test:
   a. State the null (H_0) and alternative (H_a) hypothesis.
2. **Specify the level of significance (alpha, α = .05)**
3. Compute your test statistic and p-value
4. Make a statistical decision

2. Specify Significance level (alpha α)

- It sets the level of risk of being wrong
- It indicates the probability of rejecting the null hypothesis, when it is, in fact, true (error type I: “male patient is pregnant while he isn’t)
- It is an arbitrary and a priori declared probability threshold
- 5% is usually the highest significance level that researchers are willing to accept, though it can be less
- Commonly used to compare with p-values

2. Relationship between alpha and p-value

- p-values help you to reject or accept the null hypothesis
- If the p-value is small, it indicates the result was unlikely to have occurred by chance (results are significant)
- Large p-value means that results are not significant (i.e. sampling error)
- *E.g. a p value of 0.0254 is 2.45%. This means there is a 2.54% chances your results could be random. In contrary, if p-value is 0.9(90%) means your results have a 90% probability of being complete random.*
- Significance level (Alpha = α) is the threshold value that we measure p-values against

2. Relationship between alpha and p-value

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

Steps of hypothesis testing

1. Formulate the statistical hypothesis for the test:

   a. State the null (H_0) and alternative (H_a) hypothesis.

2. Specify the level of significance (alpha, α = .05)

3. **Compute your test statistic and p-value**

4. Make a statistical decision

3. 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	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

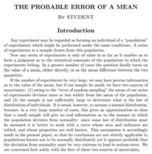
The t-statistic was developed and published in 1908 by William Gosset, a chemist/statistician working for the Guinness brewery in Dublin. Guinness employees were not allowed to publish under their own name, so Gosset published under the pseudonym "Student".

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, ...)
```



THE PROBABLE ERROR OF A MEAN

By STUDENT

Introduction

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).

3. Compute your test statistic and p-value

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)

Observed (Data)

$$t = \frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{N}}$$

Expected value under null hypothesis

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

3. Compute your test statistic and p-value

Equation for an independent samples* *t*-test

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

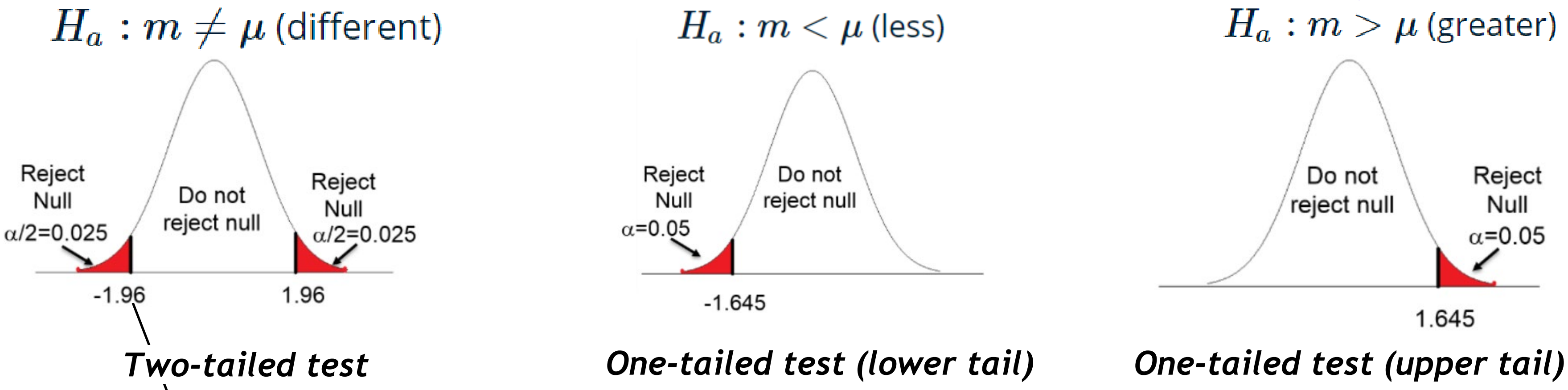
$$t = \frac{\bar{X}_1 - \bar{X}_2}{\mathrm{SE}}$$

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

SE = standard error of the sample
σ = sample standard deviation
n = number of samples

3. Compute your test statistic and p-value

Differences between one sided and two sided test alpha = 5%



Critical values = A value of our test statistic that marks the limits of our extreme values

Steps of hypothesis testing

1. Formulate the statistical hypothesis for the test:
    a. State the null (H_0) and alternative (H_a) hypothesis.
2. Specify the level of significance (alpha, $\alpha = .05$)
3. Compute your test statistic and p-value
4. **Make a statistical decision**

4. Make a statistical decision: Applied

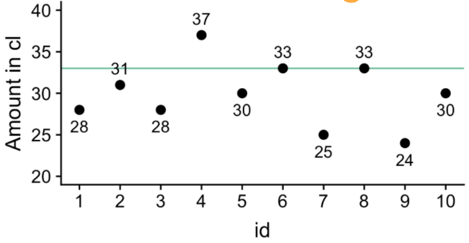
The mean amount poured in 33cl orange juice by the shops is *less* than 33cl:

H_0: μ = 33, H_a: μ < 33

Orange Data

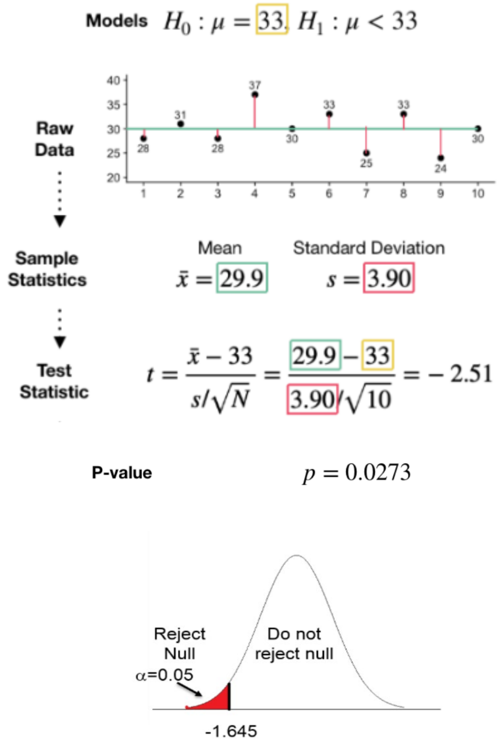
I ordered 10 oranges, and measured the exact amount in each cup, here the results:

Collected data 10 oranges from Shop



4. Make a statistical decision: Applied

Steps 1 through 4	Result
Null (H_0)	Mean is equal to 33
Alternative (H_a)	Mean is less 33
Level significance (α)	0.05 level
Test statistic	-2.51
Critical values	[-1.645]
P-value	0.0273
Conclusion	***p-value < .05 threshold, we conclude that the null hypothesis is likely to be wrong and they are pondering less than 33cl***



Is there any differences in the mean orange juice poundered in a glass of 33 cl 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

c.utrillaguerrero@maastrichtuniversity.nl

Independent sample t-test (Applied)

Steps 1 through 4	Result
Null (H_0)	Equal means (mean.M = mean.A)
Alternative (H_a)	Difference between them
Level significance (alpha)	0.05 level
Test statistic	0.51925
p-value	0.6099
Conclusion	***p - value > .05, as such we not reject Ho at 5% significance level.***

Two sample t-test in R

````
```{r}
# Define orange sample
orange_maastricht <- c(28,31,28,37,30,
             33,25,33,24,30)

orange_amsterdam <-  c(24,31,25,37,30,
             33,23,32,24,30)

# Conduct independent t-test
t.test(x = orange_maastricht,   # sample values Maastricht
       y = orange_amsterdam,    # sample values Amsterdam
       paired = FALSE, # different and independent observations from samples
       var.equal = TRUE) # we assume variance are equal thus, we run Student t.test

```
````

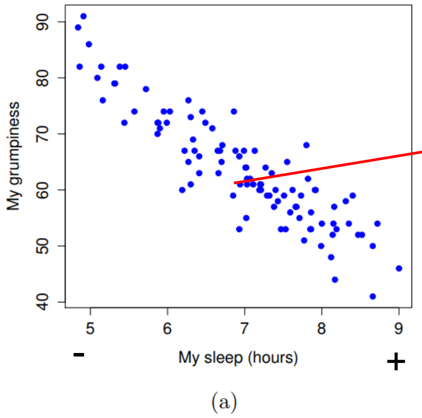
```
        Two Sample t-test

data:  orange_maastricht and orange_amsterdam
t = 0.51925, df = 18, p-value = 0.6099
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.046056  5.046056
sample estimates:
mean of x mean of y
     29.9      28.9
```

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

3. Correlations and covariance

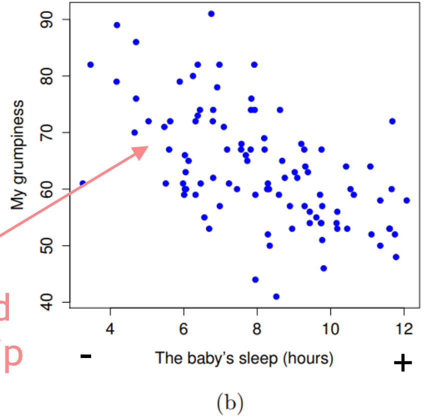
The strength and direction of a relationship

- 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*).



Dots concentrated around a line = strong relationship

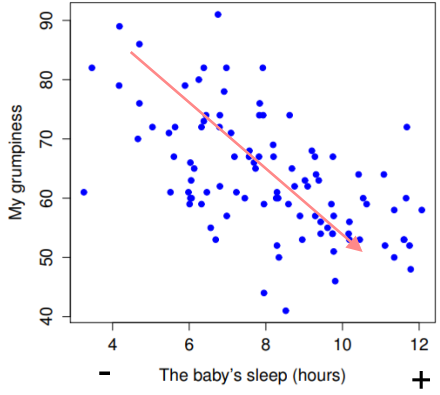
Dots widely spread = weak relationship



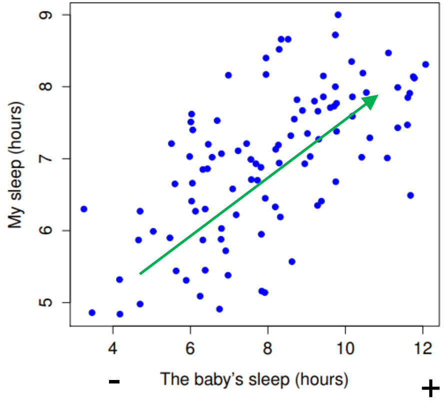
Scatterplots showing the relationship between my sleep hours and my grumpy mood (a) and the relationship between my niece sleep hours and my grumpy mood (b)

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)

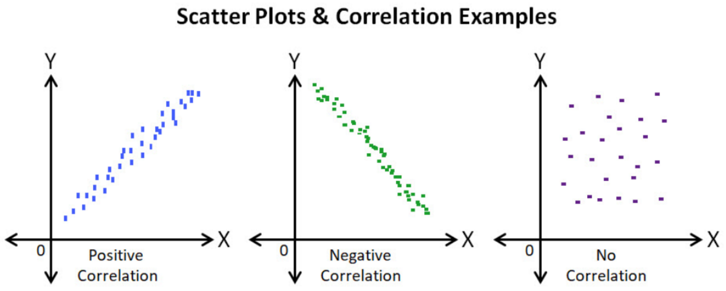


(b)

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:

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

- 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.

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

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

 c.utrillaguerrero@maastrichtuniversity.nl

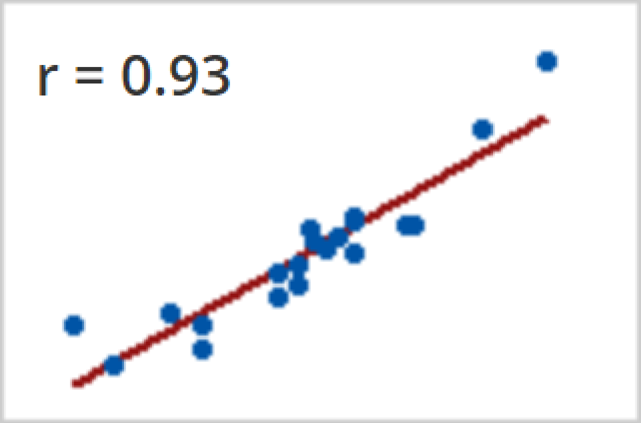
Interpreting the correlation

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

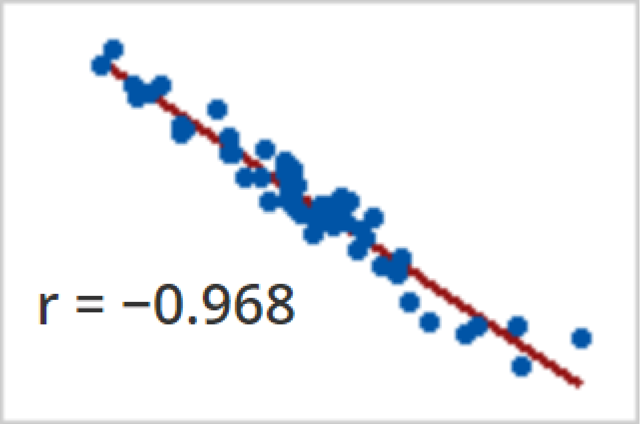
Correlation	Strength	Direction
-1.0 to -0.9	Very strong	Negative
-0.9 to -0.7	Strong	Negative
-0.7 to -0.4	Moderate	Negative
-0.4 to -0.2	Weak	Negative
-0.2 to 0	Negligible	Negative
0 to 0.2	Negligible	Positive
0.2 to 0.4	Weak	Positive
0.4 to 0.7	Moderate	Positive
0.7 to 0.9	Strong	Positive
0.9 to 1.0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



Large positive relationship

Interpreting the correlation

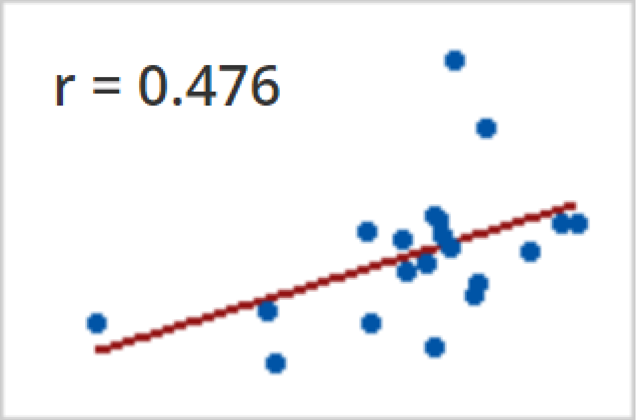
Correlation	Strength	Direction
-1.0 to -0.9	Very strong	Negative
-0.9 to -0.7	Strong	Negative
-0.7 to -0.4	Moderate	Negative
-0.4 to -0.2	Weak	Negative
-0.2 to 0	Negligible	Negative
0 to 0.2	Negligible	Positive
0.2 to 0.4	Weak	Positive
0.4 to 0.7	Moderate	Positive
0.7 to 0.9	Strong	Positive
0.9 to 1.0	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 is a numerical expression of the relationship between two 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
- Scatter plot are a method of visually represent this bivariate relationship.

IDS Summer School

29th of June - 2nd of July

Open to everyone

Online Course! Register Here