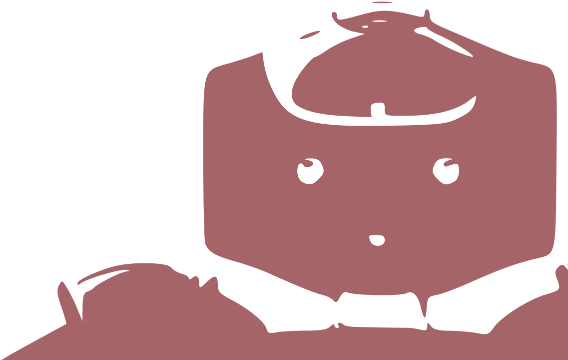
You can download the sources of this presentation here:
github.com/severin-lemaignan/lecture-hri-data-analysis

Data Analysis for HRI



Séverin Lemaignan

Bristol Robotics Lab
University of the West of England

IN THIS LECTURE

- Two questions to answer:

IN THIS LECTURE

- Two questions to answer:

  Are my groups different?

IN THIS LECTURE

- Two questions to answer:

  Are my groups different?
  Does a specific variable explain the difference?

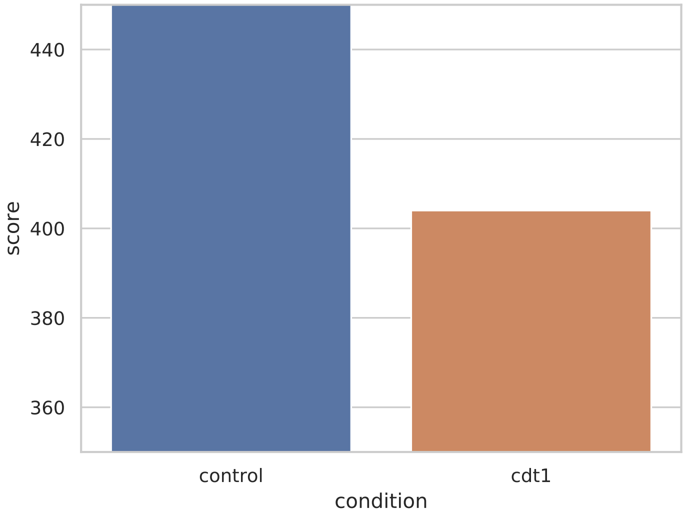
IN THIS LECTURE

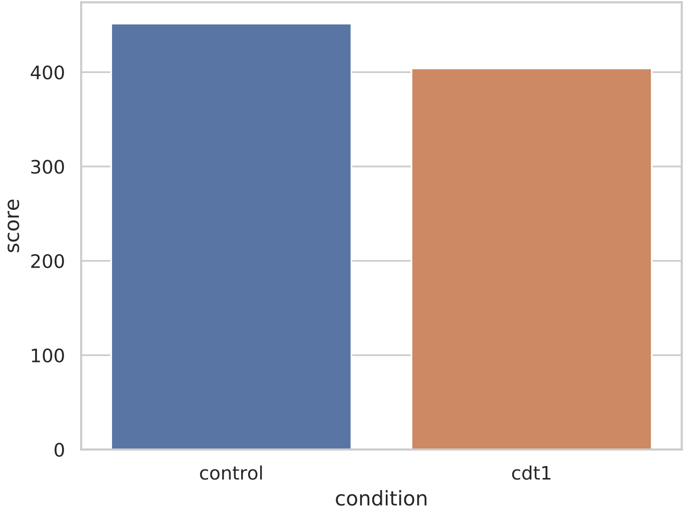
- Two questions to answer:
  - Are my groups different?
  - Does a specific variable explain the difference?
- Hands-on data analysis with Python!

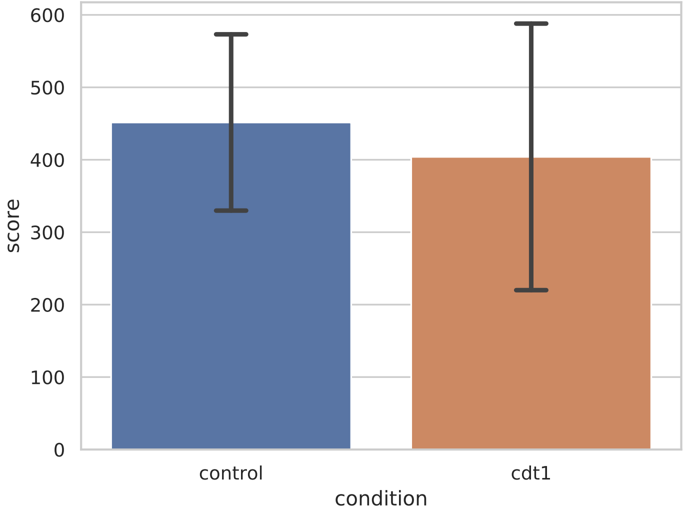
ARE MY TWO GROUPS DIFFERENT?

A DATASET

pptID	age	condition	score	heartrate
1	22	control	643	76
2	26	cdt1	234	72
3	24	control	356	73
4	24	cdt1	587	75
5	29	cdt1	561	75
6	31	control	544	75
7	20	control	470	74
8	23	cdt1	212	72
9	23	control	388	73
10	22	cdt1	201	72
11	28	control	278	72
12	29	cdt1	599	76
13	27	control	366	73
14	21	cdt1	597	75
15	22	cdt1	571	75
16	30	control	554	75


440
420
400
380
360
score
control
cdt1
condition


400
300
200
100
0
score
control
cdt1
condition

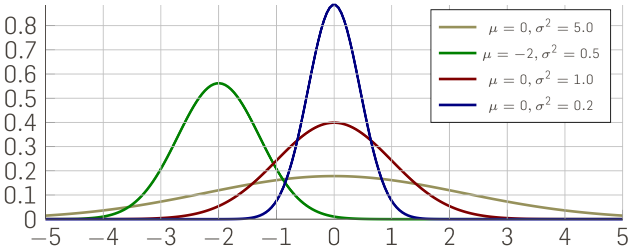

600
500
400
300
200
100
0
score
control
cdt1
condition

Is there a difference?

- Is the distribution the same?
- How big the difference? → **effect size**
- Could chance explain that difference?

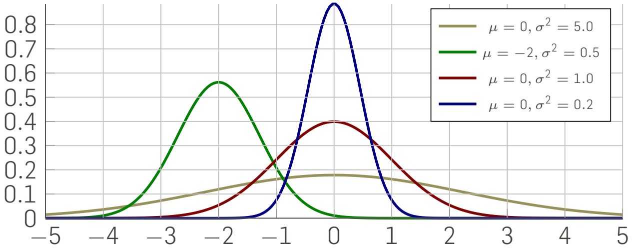
IS THE DISTRIBUTION THE SAME?

Data often (but not always!) follows a normal (or Gaussian) distribution. Two parameters: **mean μ and variance σ^2**.



IS THE DISTRIBUTION THE SAME?

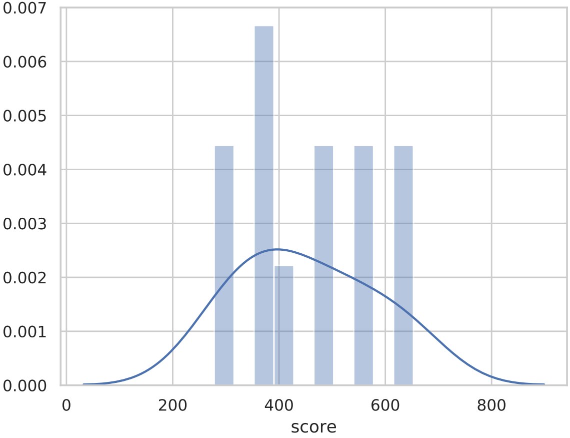
Data often (but not always!) follows a normal (or Gaussian) distribution. Two parameters: **mean μ and variance σ^2**.



Many statistical tests only work if the underlying data follows a normal distribution – so-called **parametric tests**.

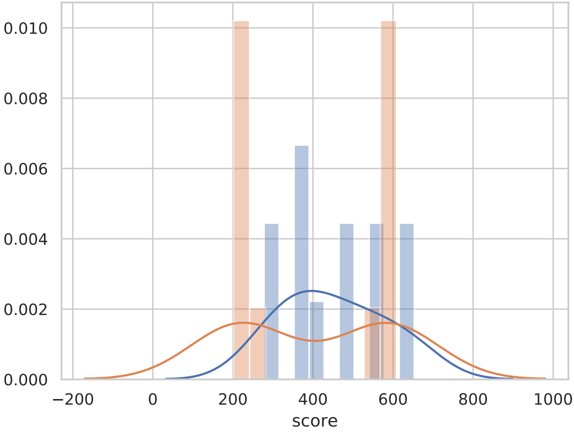
You need to check that your data is normally distributed first! (for instance, by plotting it)

COMPARE DISTRIBUTIONS (HISTOGRAMS, DENSITY)

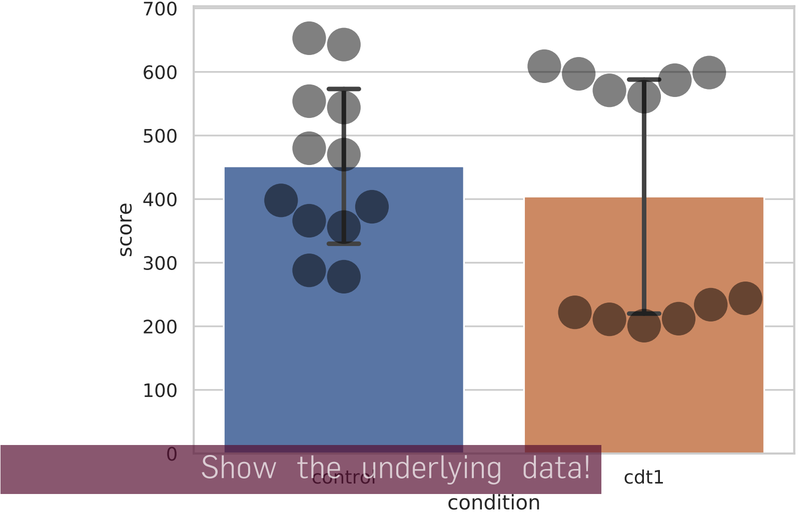


Control group

COMPARE DISTRIBUTIONS (HISTOGRAMS, DENSITY)

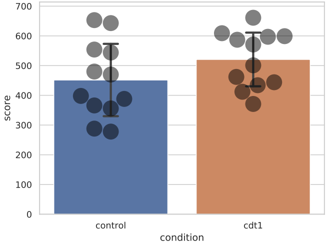


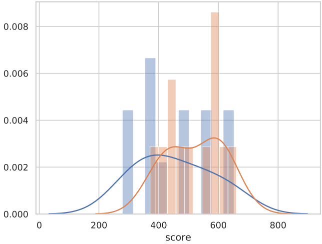
Control + condition group → beware the bimodal distribution!



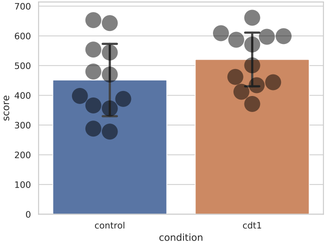
700
600
500
400
300
200
100
0
score
control
cdt1
condition
Show the underlying data!

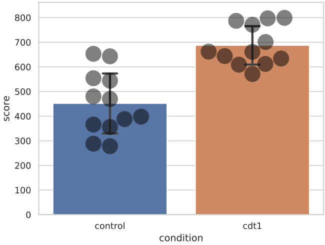
TWO ADDITIONAL DATASETS

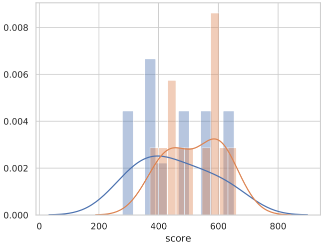


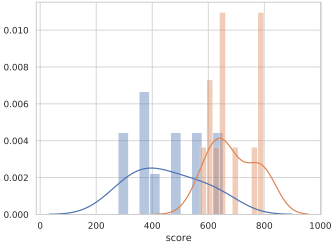


TWO ADDITIONAL DATASETS

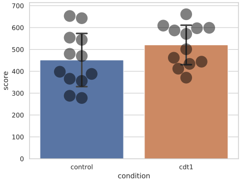






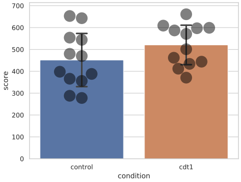


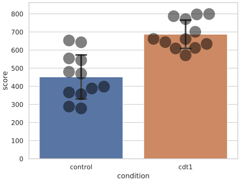
HOW BIG IS THE DIFFERENCE?



	mean	std
cdt1	516.5	85.3
control	451.5	127.1
$\mu_1 - \mu_2$	69.2	

HOW BIG IS THE DIFFERENCE?

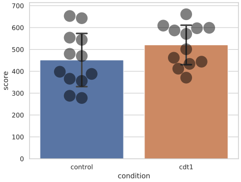




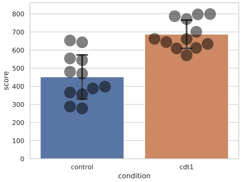
	mean	std
cdt1	516.5	85.3
control	451.5	127.1
$\mu_1 - \mu_2$	69.2	

	mean	std
cdt1	687.3	81.5
control	451.5	127.1
$\mu_1 - \mu_2$	235.8	

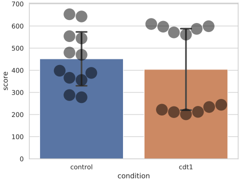
HOW BIG IS THE DIFFERENCE?



	mean	std
cdt1	516.5	85.3
control	451.5	127.1
$\mu_1 - \mu_2$	69.2	

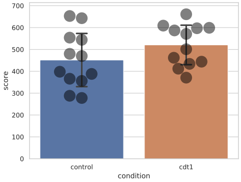


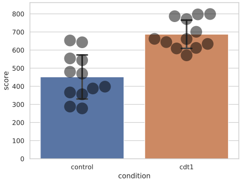
	mean	std
cdt1	687.3	81.5
control	451.5	127.1
$\mu_1 - \mu_2$	235.8	

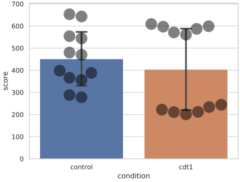


	mean	std
cdt1	404.0	192.2
control	451.5	127.1
$\mu_1 - \mu_2$	47.5	

HOW BIG IS THE DIFFERENCE?







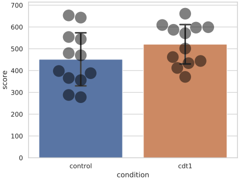
	mean	std
cdt1	516.5	85.3
control	451.5	127.1
$\mu_1 - \mu_2$	69.2	

	mean	std
cdt1	687.3	81.5
control	451.5	127.1
$\mu_1 - \mu_2$	235.8	

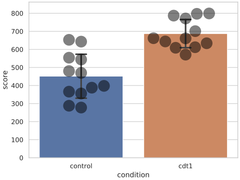
	mean	std
cdt1	404.0	192.2
control	451.5	127.1
$\mu_1 - \mu_2$	47.5	

does not account for the variance in the dataset

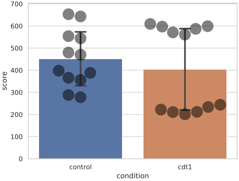
HOW BIG IS THE DIFFERENCE?



	mean	std
cdt1	516.5	85.3
control	451.5	127.1
$\mu_1 - \mu_2$	69.2	
$\frac{\mu_1 - \mu_2}{\sigma}$	0.62	

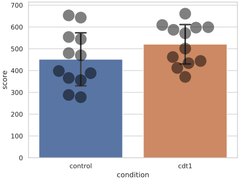


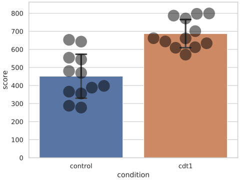
	mean	std
cdt1	687.3	81.5
control	451.5	127.1
$\mu_1 - \mu_2$	235.8	
$\frac{\mu_1 - \mu_2}{\sigma}$	2.21	

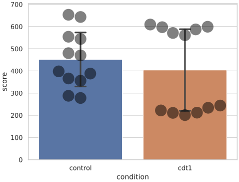


	mean	std
cdt1	404.0	192.2
control	451.5	127.1
$\mu_1 - \mu_2$	47.5	
$\frac{\mu_1 - \mu_2}{\sigma}$	0.29	

HOW BIG IS THE DIFFERENCE?







	mean	std
cdt1	516.5	85.3
control	451.5	127.1
$\mu_1 - \mu_2$	69.2	
$\frac{\mu_1 - \mu_2}{\sigma}$	0.62	

	mean	std
cdt1	687.3	81.5
control	451.5	127.1
$\mu_1 - \mu_2$	235.8	
$\frac{\mu_1 - \mu_2}{\sigma}$	2.21	

	mean	std
cdt1	404.0	192.2
control	451.5	127.1
$\mu_1 - \mu_2$	47.5	
$\frac{\mu_1 - \mu_2}{\sigma}$	0.29	

A common measure of effect size: **Cohen's *d*** $= \frac{\mu_1 - \mu_2}{\sigma}$

$\rightarrow$ Interactive visualisation and interpretation of Cohen's ***d***

DIFFERENCE DUE TO CHANCE?

A statistical hypothesis test makes an assumption about the outcome, called the **null hypothesis**.

Our *null hypothesis* is that there is no difference between the means of our two populations.

DIFFERENCE DUE TO CHANCE?

A statistical hypothesis test makes an assumption about the outcome, called the **null hypothesis**.

Our *null hypothesis* is that there is no difference between the means of our two populations.

p-value: probability of observing the result *given that the null hypothesis is true*.

$\Rightarrow$ **Meaning of a low p-value?**

DIFFERENCE DUE TO CHANCE?

A statistical hypothesis test makes an assumption about the outcome, called the **null hypothesis**.

Our ***null hypothesis*** is that there is no difference between the means of our two populations.

p-value: probability of observing the result ***given that the null hypothesis is true***.

$\Rightarrow$ **Meaning of a low *p*-value?**

To interpret ***p***, you need to choose a ***significance level*** α.
For instance, 10% (0.1), 5% (0.05), 2% (0.02)...

$p = 0.05$

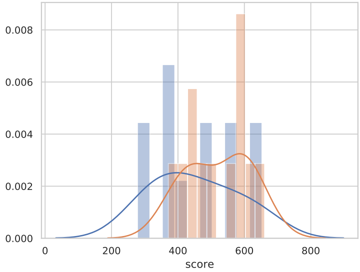
'There's only 5% of chance of observing these distributions if my null hypothesis is true (ie, no difference between my groups).'

HOW TO CALCULATE *P*?

- If parametric data, **Student's *t*-test**

HOW TO CALCULATE *P*?

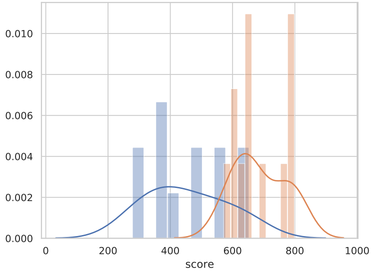
- If parametric data, **Student's *t*-test**



t statistic	-1.51
p	0.155

HOW TO CALCULATE *P*?

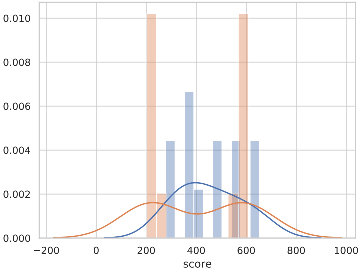
- If parametric data, **Student's *t*-test**



t statistic	-5.41
p	< 0.001

HOW TO CALCULATE *P*?

- If parametric data, **Student's *t*-test**



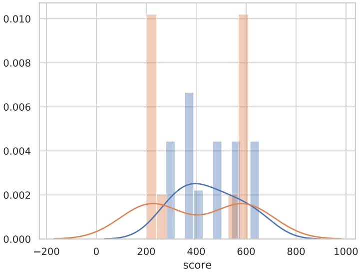
t statistic	0.71
p	0.48

HOW TO CALCULATE *P*?

- If parametric data, **Student's *t*-test**
- If non-parametric data, **Mann-Whitney *U*-test**

HOW TO CALCULATE *P*?

- If parametric data, **Student's *t*-test**
- If non-parametric data, **Mann-Whitney *U*-test**



U statistic	46.0
p	0.07

See Wikipedia page for examples and interpreation of ***U***

IMPACT OF *N*?

What is the impact of the sample size *n* on *p*?

IMPACT OF *N*?

What is the impact of the sample size *n* on *p*?

The higher ***n***, the more unlikely the difference is due to chance

$$\nearrow n \Rightarrow \searrow p$$

IMPACT OF *N*?

What is the impact of the sample size n on p?

The higher ***n***, the more unlikely the difference is due to chance

$$\nearrow n \Rightarrow \searrow p$$

BE CAREFUL WITH "STATISTICALLY SIGNIFICANT"!

gender	iq
male	76.51
male	76.53
female	76.66
female	76.65
female	76.64
female	76.63
male	76.54
female	76.64
male	76.51
female	76.60
female	76.63
male	76.52
female	76.64
male	76.51
female	76.60
female	76.63

BE CAREFUL WITH "STATISTICALLY SIGNIFICANT"!

t statistic	12.52
p	< 0.001
Mean female	76.64
Mean male	76.54

$$M_{female} > M_{male}, p < 0.001$$

BE CAREFUL WITH "STATISTICALLY SIGNIFICANT"!

t statistic	12.52
p	< 0.001
Mean female	76.64
Mean male	76.54

$$M_{female} > M_{male}, p < 0.001$$

Girls are more intelligent! We knew it!

BE CAREFUL WITH "STATISTICALLY SIGNIFICANT"!

t statistic	12.52
p	< 0.001
Mean female	76.64
Mean male	76.54

$$M_{female} > M_{male}, p < 0.001$$

Girls are more intelligent! We knew it!

...wait... how big is our effect?

$M_{female} - M_{male} = 0.1$ on a scale of 100??

BE CAREFUL WITH "STATISTICALLY SIGNIFICANT"!

t statistic	12.52
p	< 0.001
Mean female	76.64
Mean male	76.54

$$M_{female} > M_{male}, p < 0.001$$

Girls are more intelligent! We knew it!

...wait... how big is our effect?

$M_{female} - M_{male} = 0.1$ on a scale of 100??

Cohen's *d*

$d = \frac{\mu_1 - \mu_2}{\sigma} = 4.12 \Rightarrow$ high, because σ very low

STATISTICAL POWER ANALYSIS

Statistical power

The statistical power of a hypothesis test is the probability of detecting an effect, if there is a true effect present to detect.

or:

Statistical power

The statistical power of the test is the probability that the test correctly rejects a ***false*** null hypothesis.

STATISTICAL POWER ANALYSIS

Types of errors

- **Type I error**: Reject the null hypothesis when there is in fact no significant effect ***(too optimistic!)***
- **Type II error**: Not reject the null hypothesis when there is a significant effect ***(too pessimistic!)***

STATISTICAL POWER ANALYSIS

Types of errors

- **Type I error**: Reject the null hypothesis when there is in fact no significant effect *(too optimistic!)*
- **Type II error**: Not reject the null hypothesis when there is a significant effect *(too pessimistic!)*

Power = 1 - Type II Error

STATISTICAL POWER ANALYSIS

A puzzle with four pieces:

- **Effect size**
- **Sample size**
- **Significance** (chance of Type I error – found inexistant effect)
- **Statistical power** (1 - chance of Type II error – missed the effect)

EXAMPLE: POWER ANALYSIS OF STUDENT'S *T*-TEST

- **Effect size**: Cohen's $d > 0.8$
- **Significance**: 5%
- **Statistical power**: 80%
- **Sample size?**

EXAMPLE: POWER ANALYSIS OF STUDENT'S *T*-TEST

- **Effect size**: Cohen's d > 0.8
- **Significance**: 5%
- **Statistical power**: 80%
- **Sample size?**

Using for instance Python's `statsmodels.stats.power.TTestIndPower`, we can compute that $n = 25.5$ (per condition)

EXAMPLE: POWER ANALYSIS OF STUDENT'S *T*-TEST

- **Effect size**: Cohen's $d > 0.8$
- **Significance**: 5%
- **Statistical power**: 80%
- **Sample size?**

Using for instance Python's `statsmodels.stats.power.TTestIndPower`, we can compute that $n = 25.5$ (per condition)

A good read on statistical power analysis:

A Gentle Introduction to Statistical Power and Power Analysis in Python

ARE MY GROUPS DIFFERENT? SUMMARY

- 2 groups, independent measures, normal distribution: **Independent *t*-test**

ARE MY GROUPS DIFFERENT? SUMMARY

- 2 groups, independent measures, normal distribution: **Independent *t*-test**
- 2 groups, dependent measures, normal distribution: **Paired *t*-test** (for instance, conditions are within-subject)

ARE MY GROUPS DIFFERENT? SUMMARY

- 2 groups, independent measures, normal distribution: **Independent *t*-test**
- 2 groups, dependent measures, normal distribution: **Paired *t*-test** (for instance, conditions are within-subject)
- 2 groups, non-parametric: **Mann-Whitney U** (and **Wilcoxon signed-rank test** for paired samples)

ARE MY GROUPS DIFFERENT? SUMMARY

- 2 groups, independent measures, normal distribution: **Independent *t*-test**
- 2 groups, dependent measures, normal distribution: **Paired *t*-test** (for instance, conditions are within-subject)
- 2 groups, non-parametric: **Mann-Whitney U** (and **Wilcoxon signed-rank test** for paired samples)
- Three or more groups: **ANOVA** (analysis of variance)

ARE MY GROUPS DIFFERENT? SUMMARY

- 2 groups, independent measures, normal distribution: **Independent *t*-test**
- 2 groups, dependent measures, normal distribution: **Paired *t*-test** (for instance, conditions are within-subject)
- 2 groups, non-parametric: **Mann-Whitney U** (and **Wilcoxon signed-rank test** for paired samples)
- Three or more groups: **ANOVA** (analysis of variance)

ARE MY GROUPS DIFFERENT? SUMMARY

- 2 groups, independent measures, normal distribution: **Independent *t*-test**
- 2 groups, dependent measures, normal distribution: **Paired *t*-test** (for instance, conditions are within-subject)
- 2 groups, non-parametric: **Mann-Whitney U** (and **Wilcoxon signed-rank test** for paired samples)
- Three or more groups: **ANOVA** (analysis of variance)

Always report an **effect size** (for instance, **Cohen's *d***)

ARE MY GROUPS DIFFERENT? SUMMARY

- 2 groups, independent measures, normal distribution: **Independent *t*-test**
- 2 groups, dependent measures, normal distribution: **Paired *t*-test** (for instance, conditions are within-subject)
- 2 groups, non-parametric: **Mann-Whitney U** (and **Wilcoxon signed-rank test** for paired samples)
- Three or more groups: **ANOVA** (analysis of variance)

Always report an **effect size** (for instance, **Cohen's *d***)

Keep a close eye on your data distributions (**plot them**)

DOES ONE VARIABLE EXPLAIN THE DIFFERENCE?

OUR DATASET

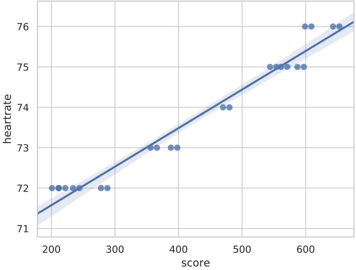
pptID	age	condition	score	heartrate
1	22	control	643	76
2	26	cdt1	234	72
3	24	control	356	73
4	24	cdt1	587	75
5	29	cdt1	561	75
6	31	control	544	75
7	20	control	470	74
8	23	cdt1	212	72
9	23	control	388	73
10	22	cdt1	201	72
11	28	control	278	72
12	29	cdt1	599	76
13	27	control	366	73
14	21	cdt1	597	75
15	22	cdt1	571	75
16	30	control	554	75

ASSOCIATION

What is the degree of association between two variables?

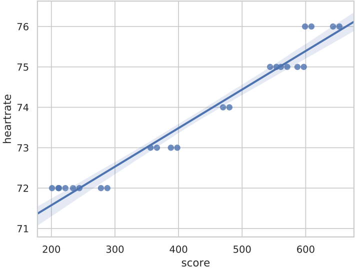
$\rightarrow$ main tool: correlation

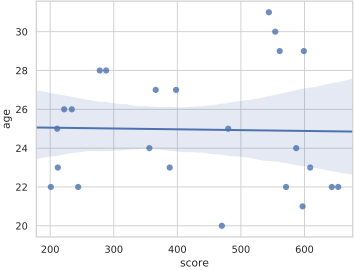
PEARSON CORRELATION



Pearson's correlation	
ρ	0.98
p	< 0.001

PEARSON CORRELATION

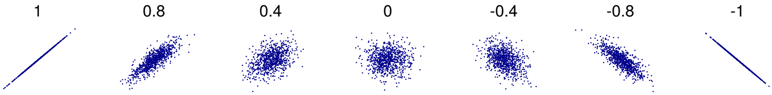




Pearson's correlation	
ρ	0.98
p	< 0.001

Pearson's correlation	
ρ	-0.022
p	0.92

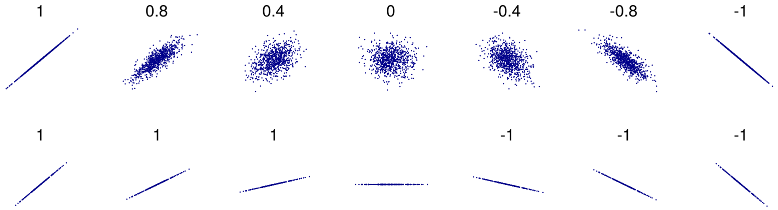
INTERPRETATION OF ρ



ρ reflects the degree of linearity and direction

Source: Wikipedia

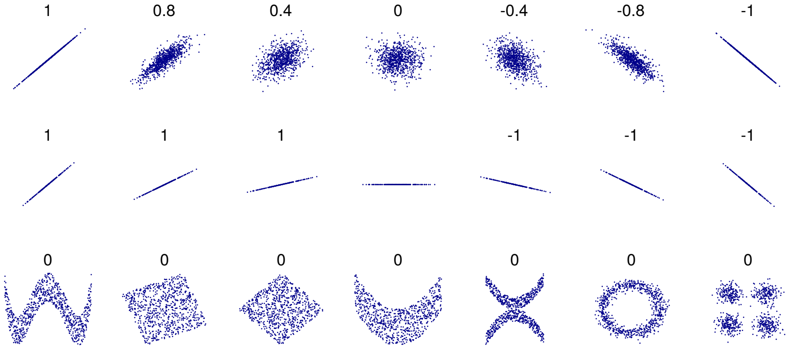
INTERPRETATION OF ρ



ρ does not reflect the slope of the regression line

Source: Wikipedia

INTERPRETATION OF ρ



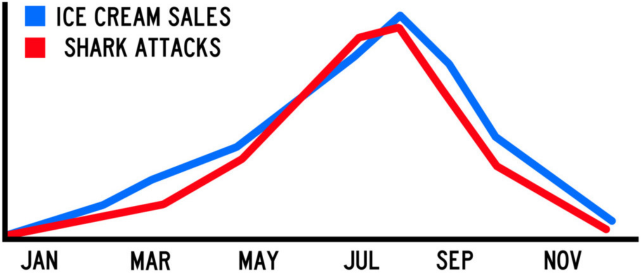
ρ does not capture non-linear interactions

Source: Wikipedia

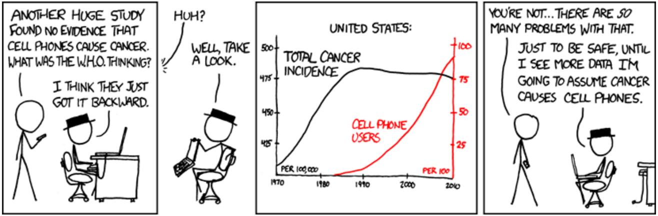
OTHER MEASURES OF ASSOCIATION

- Non-parametric ordinal data: **Spearman rank correlation**
- Association between categorical data (for instance, relationship between 'gender' and 'preferred style of cuisine'): **Pearson's Chi-Square** χ^2

CORRELATION IS NOT CAUSATION

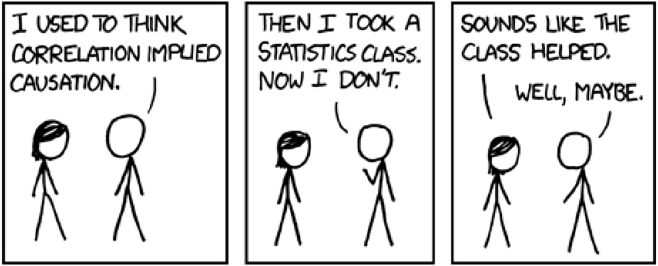


CORRELATION IS NOT CAUSATION



Source: XKCD

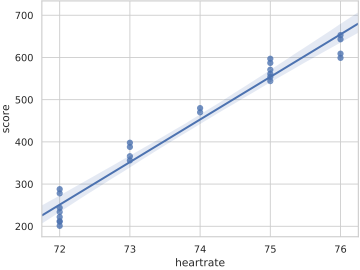
CORRELATION IS NOT CAUSATION



Source: XKCD

CORRELATION IS NOT CAUSATION

Be careful when writing something like:



“the significant positive correlation between the heart rate and the score shows that you need to have a fast heart to win”

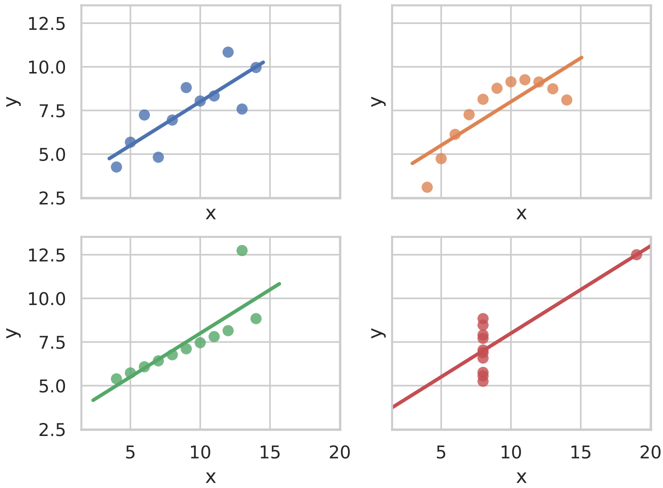
TO CONCLUDE: ANSCOMBE'S QUARTET

I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

TO CONCLUDE: ANSCOMBE'S QUARTET

Property	**Value**
Mean of x	9
Sample variance of x	11
Mean of y	7.50
Sample variance of y	4.125
Correlation between x and y	0.816
Linear regression line	y = 3.00 + 0.500x
Coefficient of determination of the linear regression	0.67

TO CONCLUDE: ANSCOMBE'S QUARTET



IN PRACTICE

THE TOOLS

Data analysis tools:

- R: **www.r-project.org**
- Python's Pandas: **pandas.pydata.org**

THE TOOLS

Data analysis tools:

- R: **www.r-project.org**
- Python's Pandas: **pandas.pydata.org**

Jupyter notebooks are a great way of creating an interactive, easy-to-follow, data analysis.

(SIDE NOTE ON PYTHON FOR DATA ANALYSIS)

Python is the leading language in data analysis/data mining/machine learning. **Learn it!**

(SIDE NOTE ON PYTHON FOR DATA ANALYSIS)

Python is the leading language in data analysis/data mining/machine learning. **Learn it!**

Large set of tools $\Rightarrow$ the SciPy landscape can be confusing at first:

(SIDE NOTE ON PYTHON FOR DATA ANALYSIS)

Python is the leading language in data analysis/data mining/machine learning. **Learn it!**

Large set of tools ⇒ the SciPy landscape can be confusing at first:

- `numpy`, `scipy`: the 'math' core

(SIDE NOTE ON PYTHON FOR DATA ANALYSIS)

Python is the leading language in data analysis/data mining/machine learning. **Learn it!**

Large set of tools ⇒ the SciPy landscape can be confusing at first:

- `numpy`, `scipy`: the 'math' core
- `ipython`, `Jupyter notebook`: interactive Python

(SIDE NOTE ON PYTHON FOR DATA ANALYSIS)

Python is the leading language in data analysis/data mining/machine learning. **Learn it!**

Large set of tools ⇒ the SciPy landscape can be confusing at first:

- `numpy`, `scipy`: the 'math' core
- `ipython`, `Jupyter notebook`: interactive Python
- `matplotlib`, `seaborn`: data visualisation (including plotting)

(SIDE NOTE ON PYTHON FOR DATA ANALYSIS)

Python is the leading language in data analysis/data mining/machine learning. **Learn it!**

Large set of tools ⇒ the SciPy landscape can be confusing at first:

- `numpy`, `scipy`: the 'math' core
- `ipython`, `Jupyter notebook`: interactive Python
- `matplotlib`, `seaborn`: data visualisation (including plotting)
- `pandas`, `statsmodels`: stats, data analysis (modelled after R)

(SIDE NOTE ON PYTHON FOR DATA ANALYSIS)

Python is the leading language in data analysis/data mining/machine learning. **Learn it!**

Large set of tools ⇒ the SciPy landscape can be confusing at first:

- `numpy`, `scipy`: the 'math' core
- `ipython`, `Jupyter notebook`: interactive Python
- `matplotlib`, `seaborn`: data visualisation (including plotting)
- `pandas`, `statsmodels`: stats, data analysis (modelled after R)
- `scikit-learn` (along with specialist ML libraries: TensorFlow, pyTorch): machine learning

(SIDE NOTE ON PYTHON FOR DATA ANALYSIS)

Python is the leading language in data analysis/data mining/machine learning. **Learn it!**

Large set of tools ⇒ the SciPy landscape can be confusing at first:

- `numpy`, `scipy`: the 'math' core
- `ipython`, `Jupyter notebook`: interactive Python
- `matplotlib`, `seaborn`: data visualisation (including plotting)
- `pandas`, `statsmodels`: stats, data analysis (modelled after R)
- `scikit-learn` (along with specialist ML libraries: TensorFlow, pyTorch): machine learning
- `anaconda` (and a few other): Python distribution for scientific computing

Let’s give it a go!