Hypothesis testing

INTRODUCTION TO STATISTICS

George Boorman
Curriculum Manager, DataCamp

datacamp

Why do we need to know about hypothesis testing?

- Hypothesis testing is used to compare populations
- Hypothesis testing is everywhere!
  - Can a change in price lead to increased revenue?
  - Will changing a website address result in increased traffic?
  - Is a medication effective in the treatment of a health condition?

[1] Image credit: https://unsplash.com/@towfiqu999999

The history of hypothesis testing

- Hypothesis testing dates back to the 1700s!

- Human sex ratio
  - More male births than female births

[1] Image credit: https://unsplash.com/@kellysikkema

Assume nothing!

- Start by assuming no difference exists
- This is called the ***null hypothesis***

Male versus female birth ratio

- **Null hypothesis:**
  - No difference in gender birth ratio between women who do and do not take vitamin C consumption
- **Alternative hypothesis:**
  - A **difference** exists in gender birth ratio between the two populations
  - **More** female births occur among women taking vitamin C supplements

Hypothesis testing workflow

- Define the target populations
  - Adult women taking or not taking vitamin C supplements
- Develop null and alternative hypotheses
  - Births are equally like to be male or female in both populations
  - More births are female among women taking vitamin C supplements
- Collect or access sample data
- Perform statistical tests on the sample data
- Draw conclusions about the population

How much data do we need?

- Central limit theorem
  - Mean male and female births gets closer to the population means as sample size increases
  - Time and resource intensive

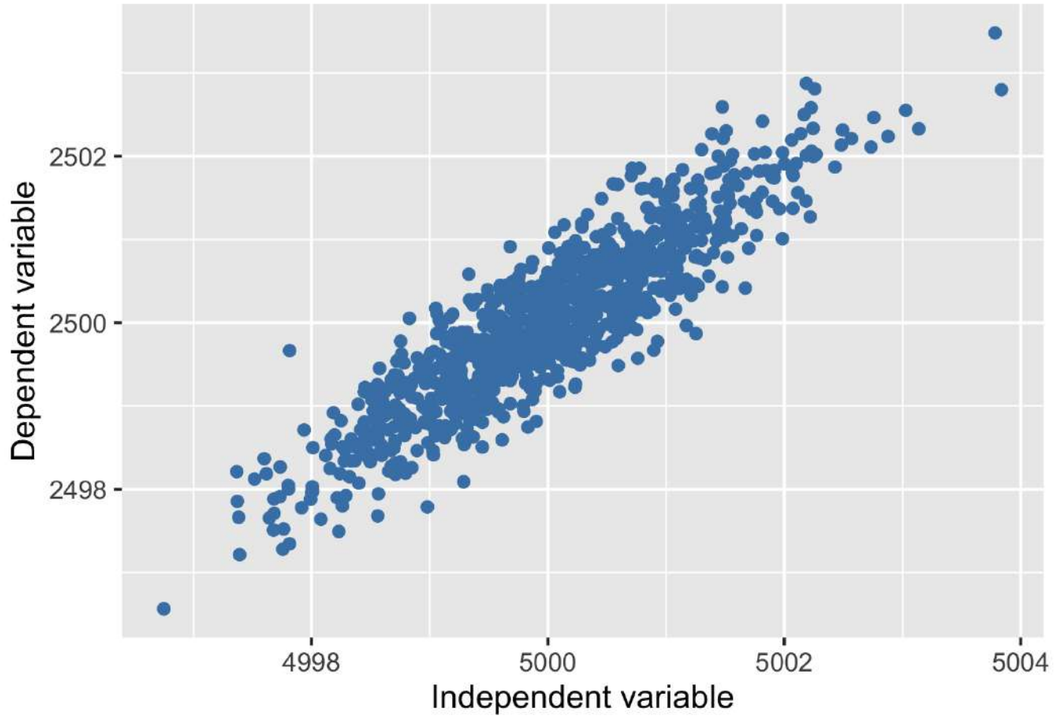
- Look at peer-reviewed research on similar hypothesis tests to decide on the sample size

[1] Image credit: https://unsplash.com/@jxnsartstudio

Independent and dependent variables

- **Independent variable:**
  - Unaffected by other data
  - Vitamin C supplementation
- **Dependent variable:**
  - Affected by other data
  - Birth gender ratio

- Commonly used to describe hypothesis test results



Let's practice!

INTRODUCTION TO STATISTICS

datacamp

Experiments

INTRODUCTION TO STATISTICS

George Boorman
Curriculum Manager, DataCamp

datacamp

Experiments, treatment, and control

- Experiments are a subset of hypothesis testing
  - Experiments are not just conducted in academia

Experiments aim to answer: *What is the effect of the treatment on the response?*

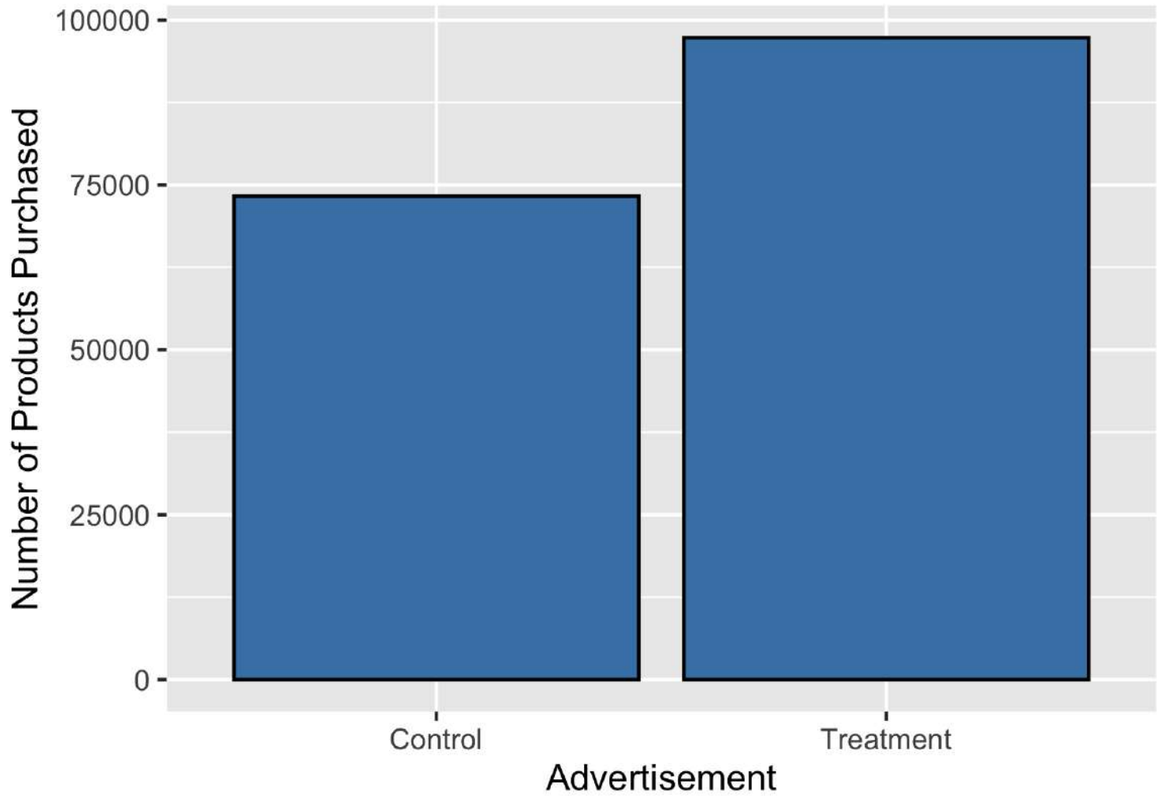
- Treatment: independent variable
- Response: dependent variable

[1] Image credit: https://unsplash.com/@nci

Advertising as a treatment

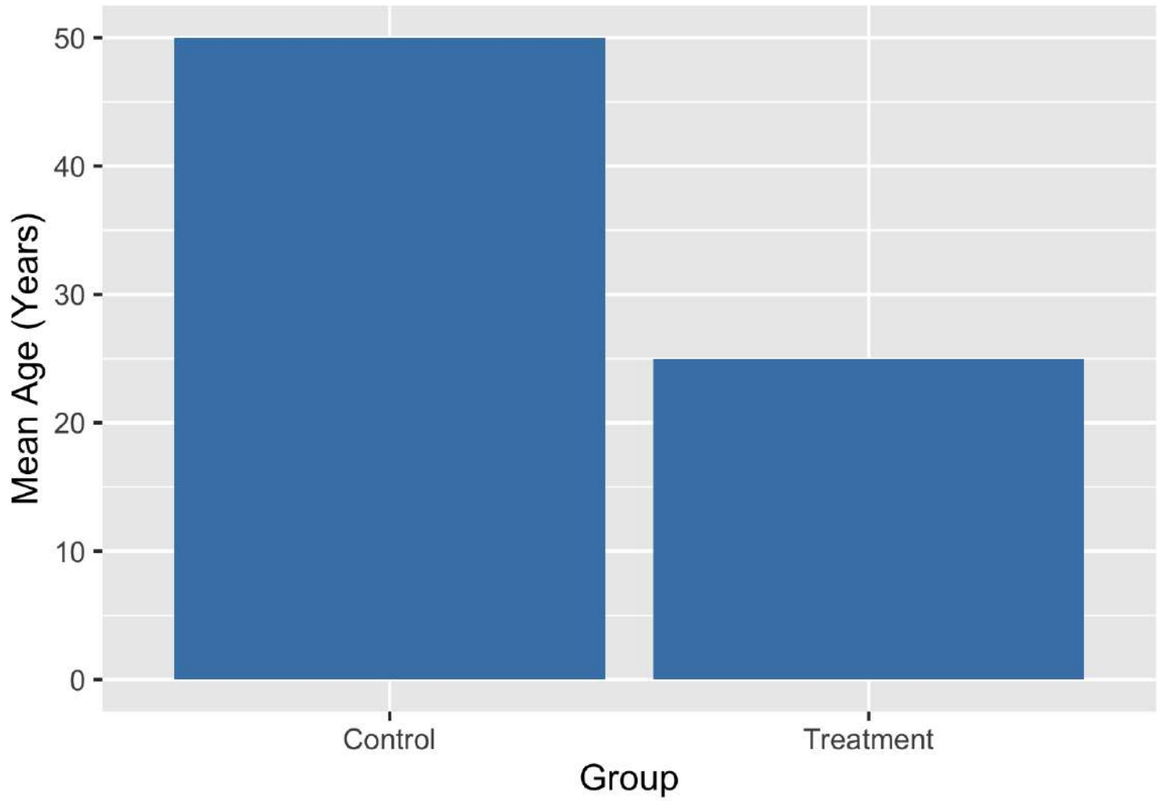
What is the effect of an advertisement on the number of products purchased?

- Treatment: advertisement
- Response: number of products purchased



Controlled experiments

- Participants are assigned to *either* the treatment group or the control group
  - ***Treatment group*** sees the advertisement
  - ***Control group*** does not see the advertisement
- Groups should be comparable to avoid introducing *bias*
- If groups are not comparable, this could lead to drawing incorrect conclusions



The gold standard of experiments

- **Randomization**
  - Participants are assigned to treatment/control *randomly*, not based on any other characteristics
  - Choosing randomly helps ensure that groups are comparable
  - Known as a **randomized controlled trial**

- **Blinding**
  - Participants will not know which group they're in
  - Participants receive a placebo, which resembles the treatment but has no effect
  - In clinical trials it is common to use a sugar pill

The gold standard of experiments

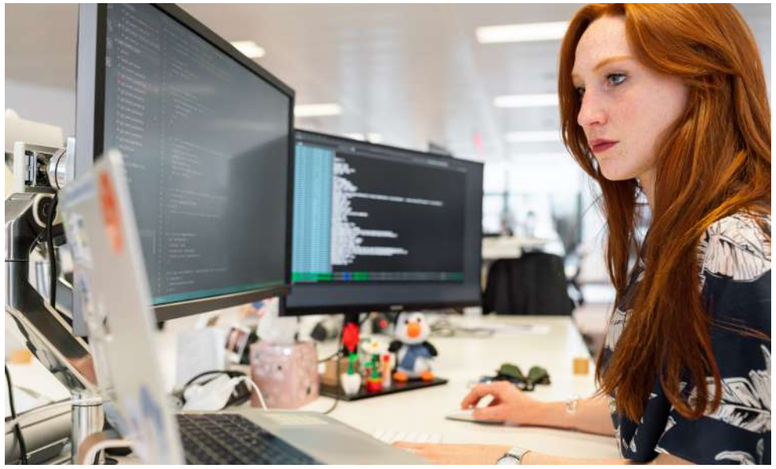
- **Double-blind randomized controlled trial**
  - Person administering the treatment/running the study doesn't know whether the treatment is real or a placebo
  - Prevents bias in the response and/or analysis of results

Fewer opportunities for bias = more reliable conclusion about causation

Randomized Controlled Trials vs. A/B testing

- **Randomized controlled trial**
  - Can be multiple treatment groups
  - Popular in science, clinical research

- **A/B testing**
  - Popular in marketing, engineering
  - Only split evenly into two groups



[1] Image credits: https://unsplash.com/@towfiqu999999; https://unsplash.com/@thisisengineering

Let's practice!

INTRODUCTION TO STATISTICS

datacamp

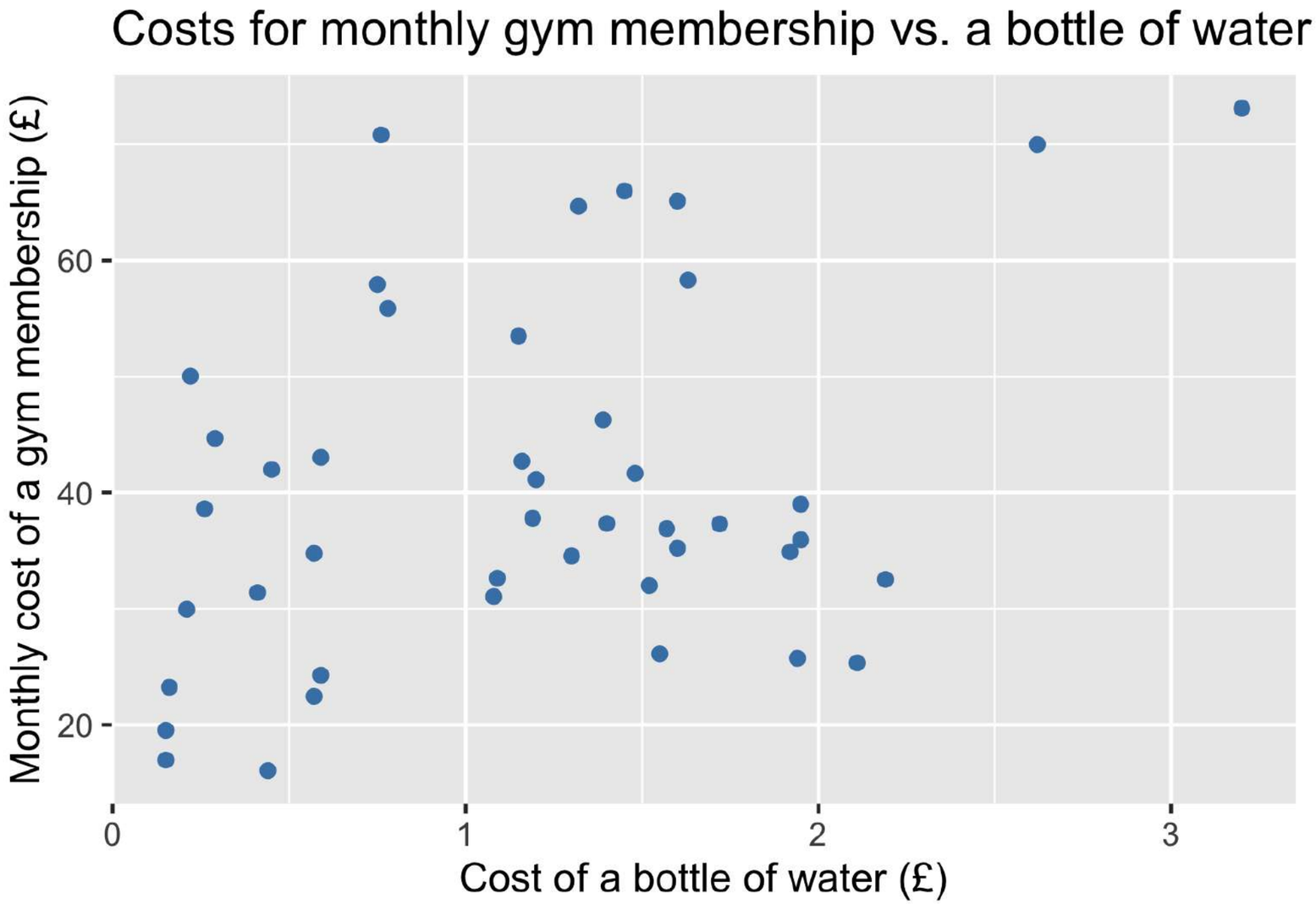
Correlation

INTRODUCTION TO STATISTICS

George Boorman
Curriculum Manager, DataCamp

datacamp

Relationships between two variables



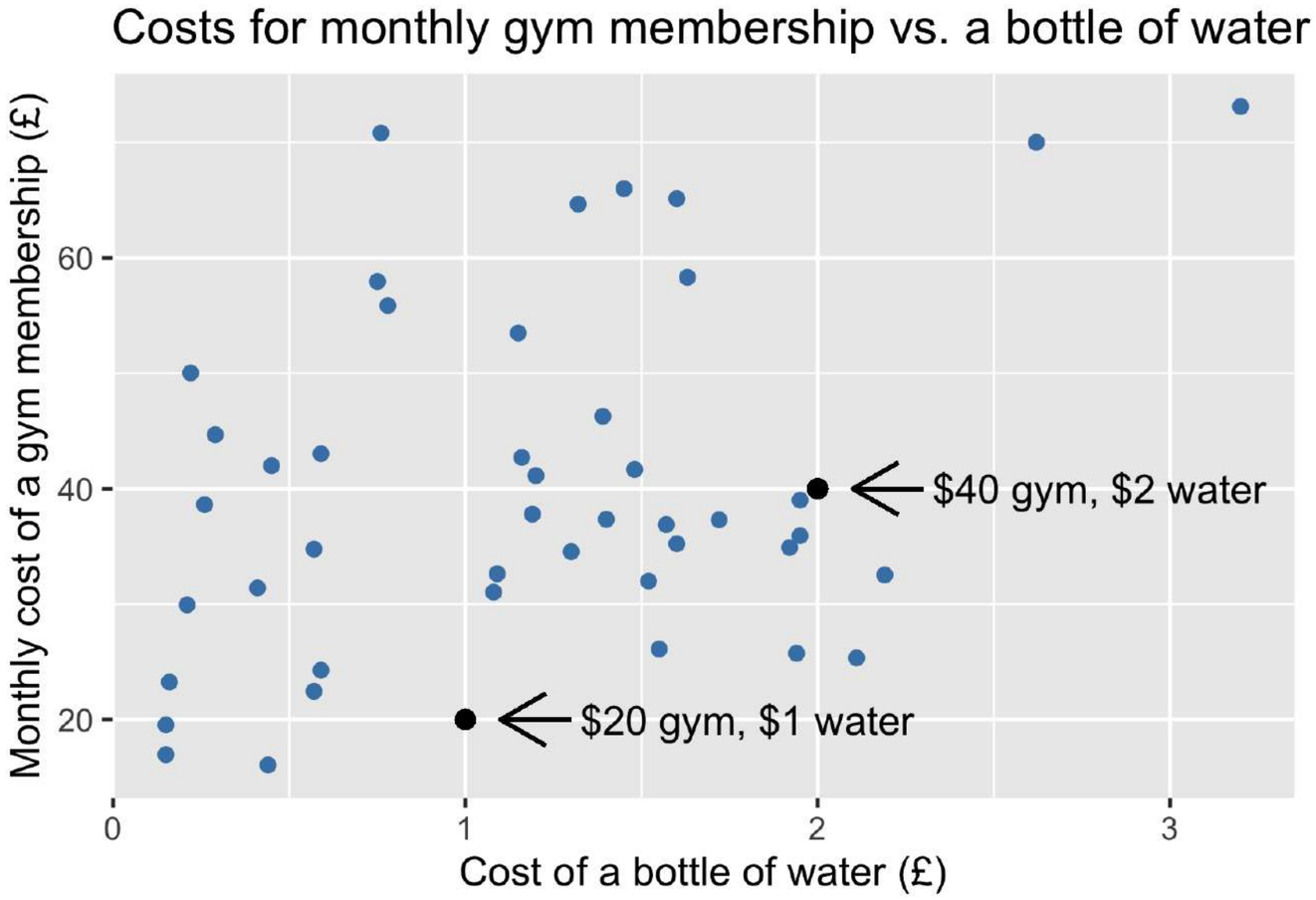
Pearson correlation coefficient

- Published by Karl Pearson in 1896!
- Quantifies the strength of a relationship between two variables
- Number between **minus one** and **one**
- Magnitude corresponds to strength of relationship
- Sign (+ or -) corresponds to direction of relationship

[1] https://royalsocietypublishing.org/doi/10.1098/rsta.1896.0007

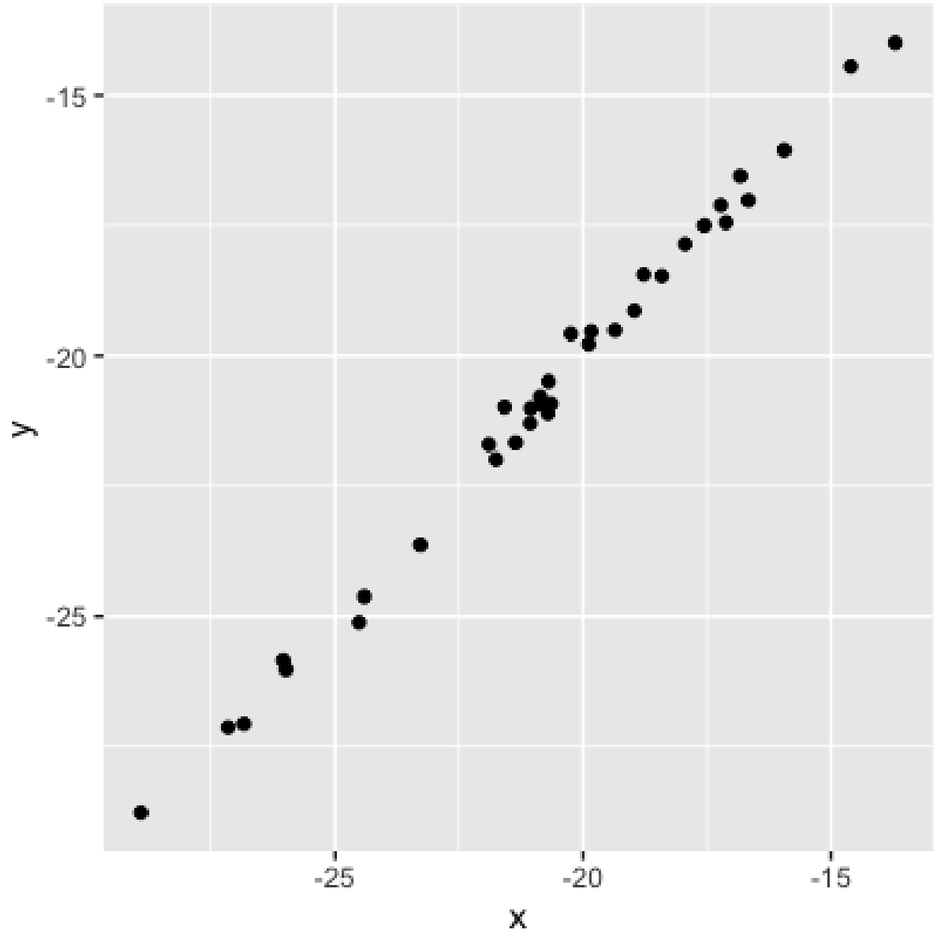
Linear relationships

- Linear = proportionate changes between dependent and independent variables



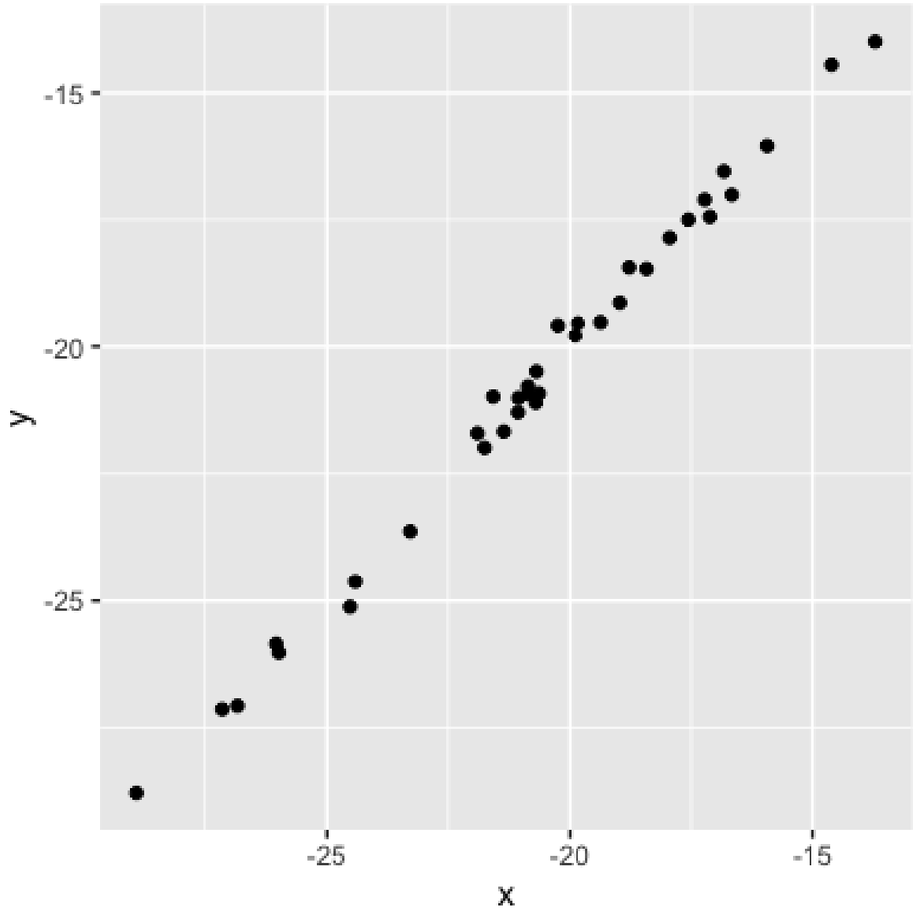
Values = strength of the relationship

0.99 (very strong relationship)

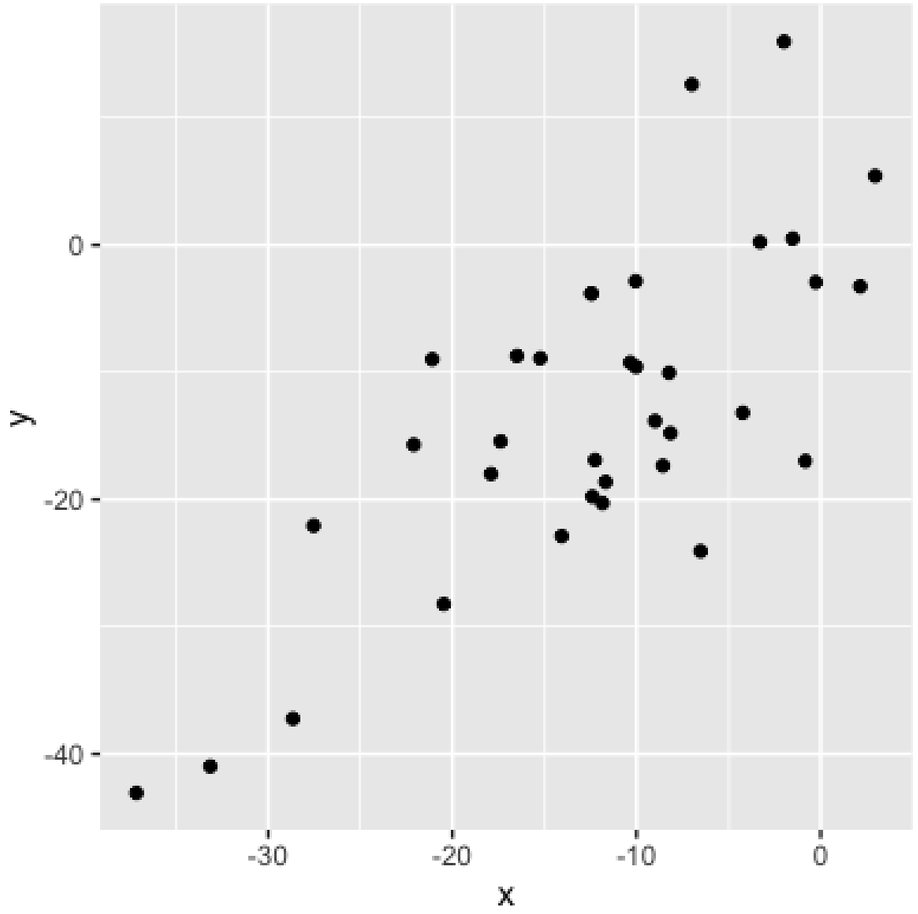


Values = strength of the relationship

0.99 (very strong relationship)

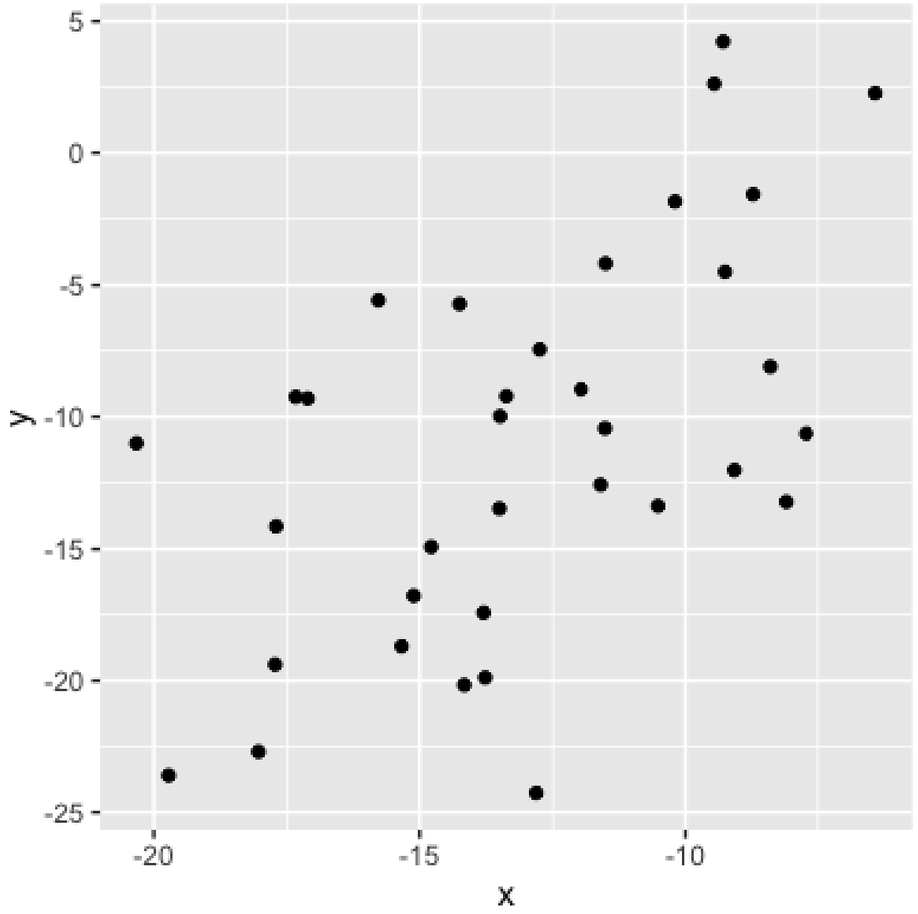


0.75 (strong relationship)



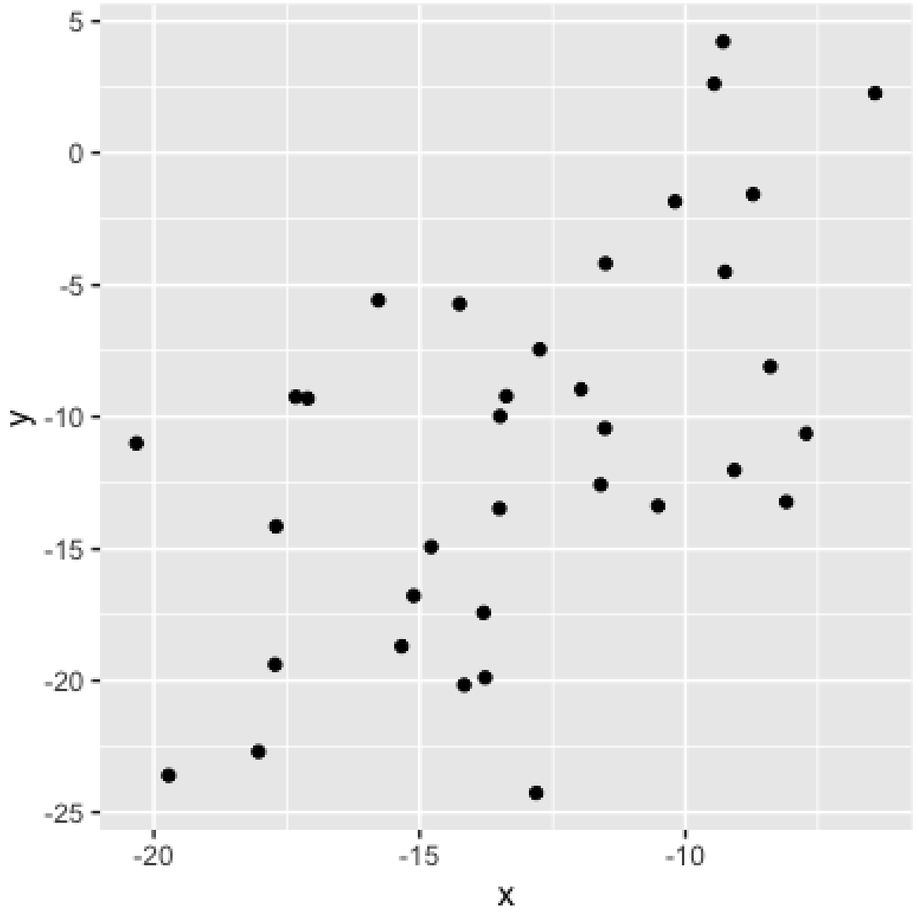
Values = strength of the relationship

0.56 (moderate relationship)

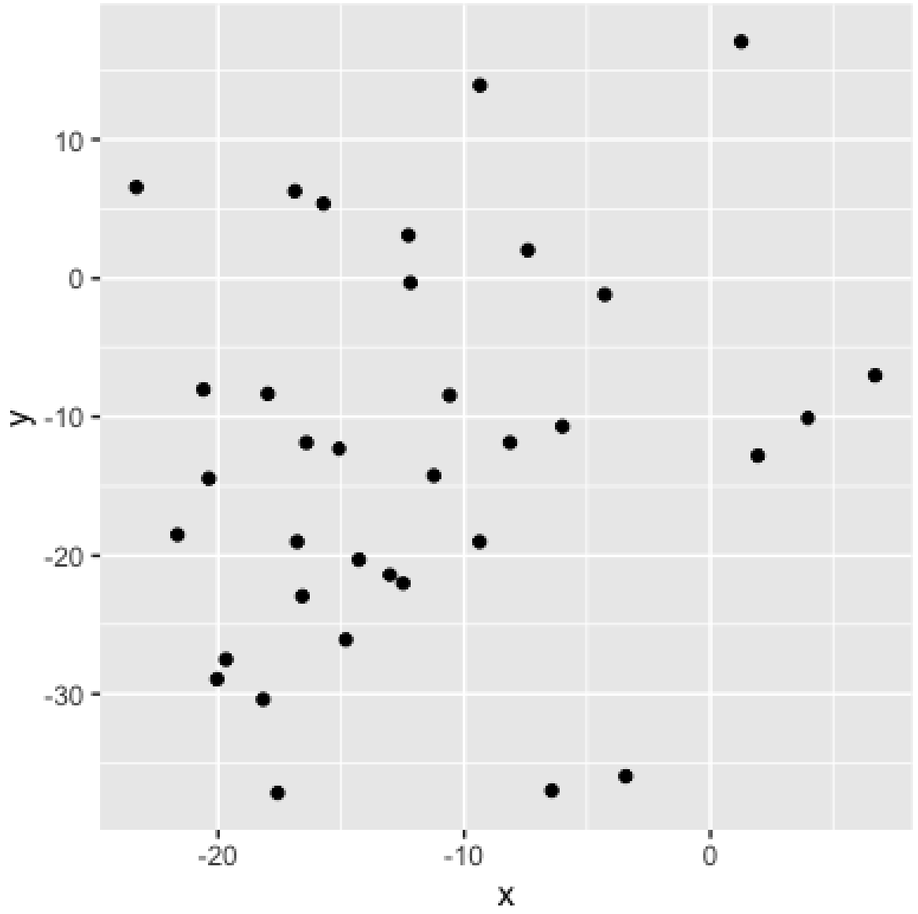


Values = strength of the relationship

0.56 (moderate relationship)

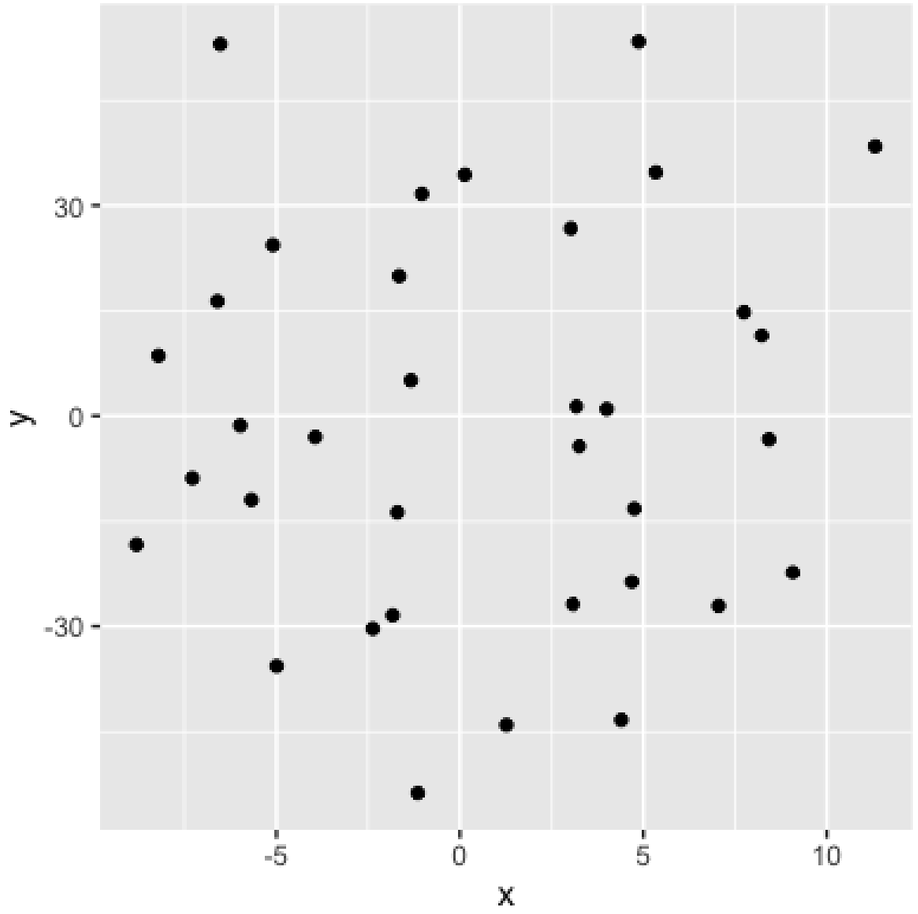


0.21 (weak relationship)



Values = strength of the relationship

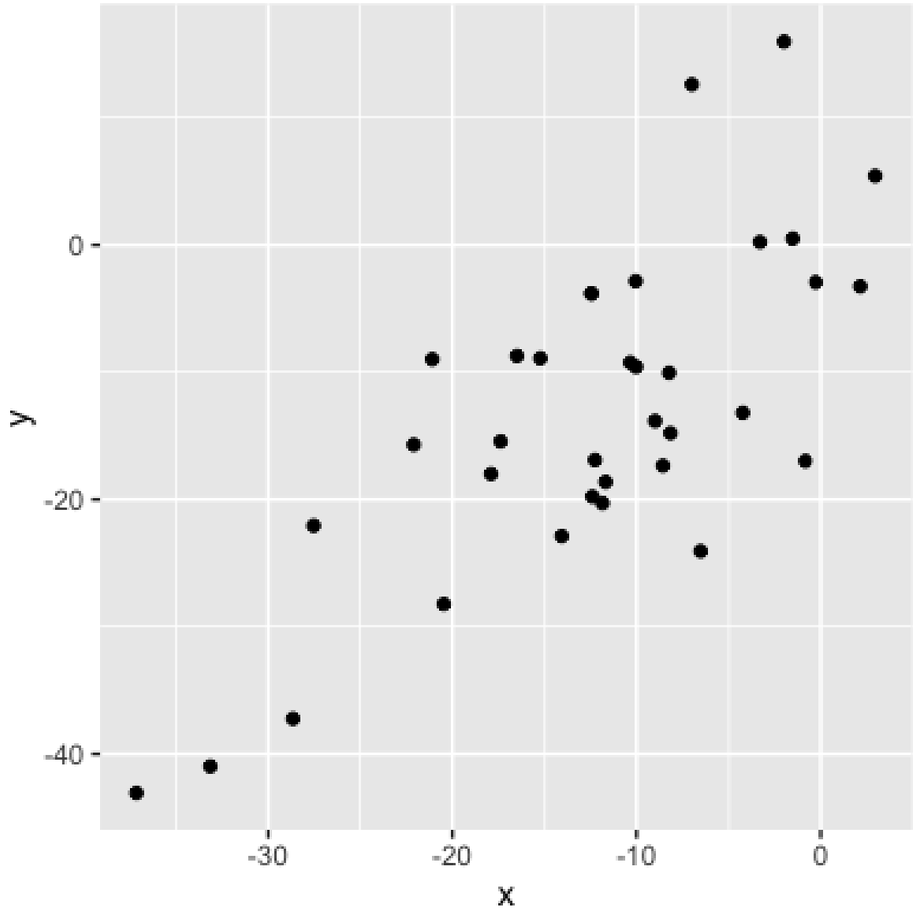
0.04 (no relationship)



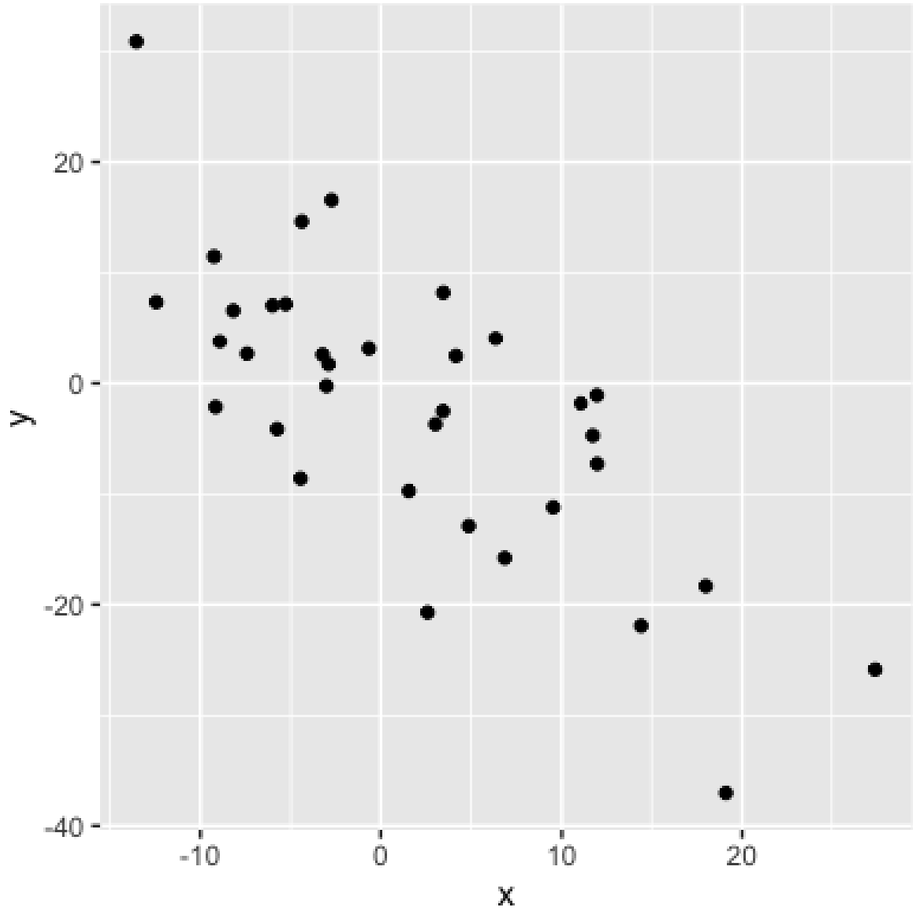
- Knowing the value of **x** doesn't tell us anything about **y**

Sign = direction

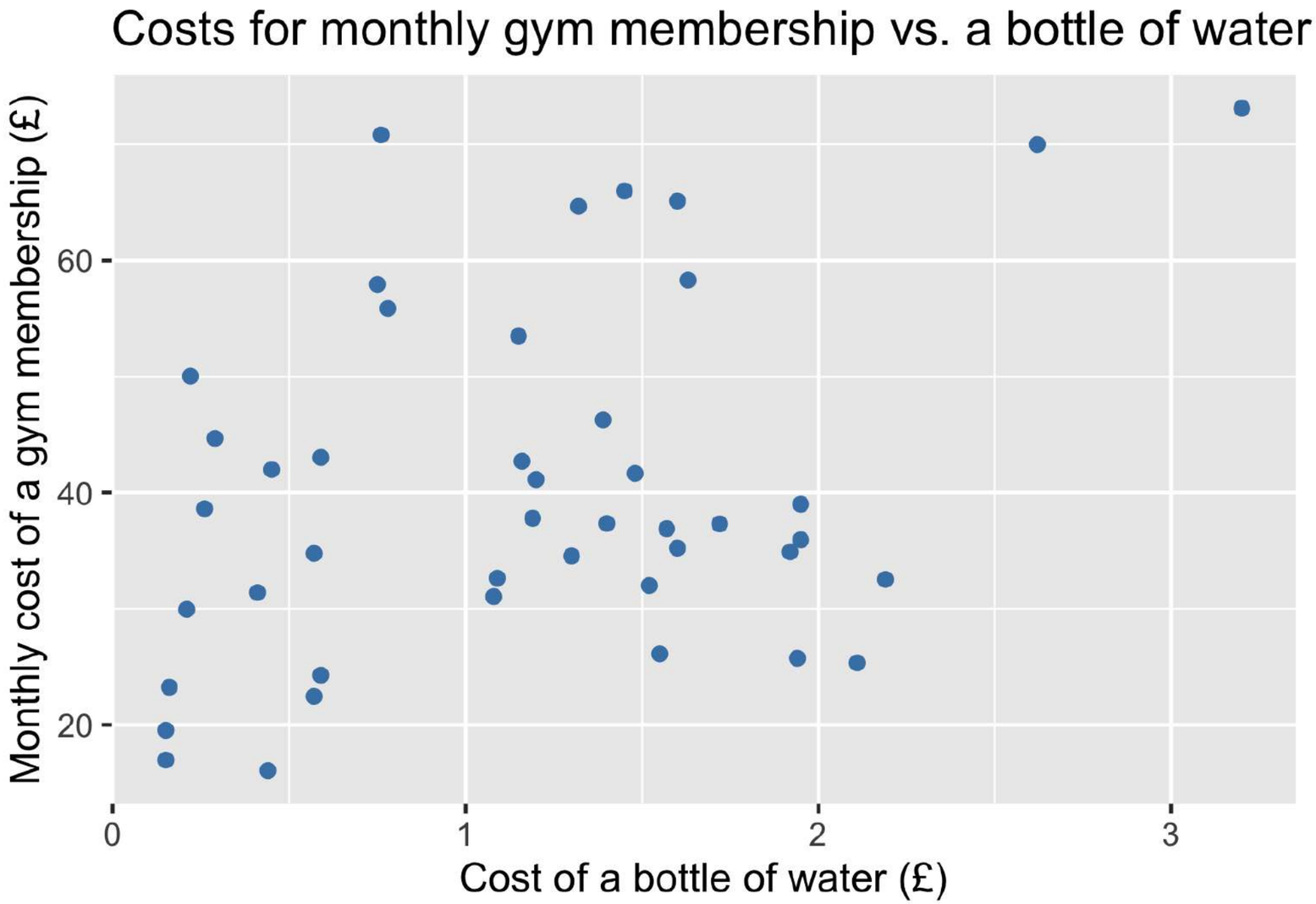
0.75: as x increases, y increases



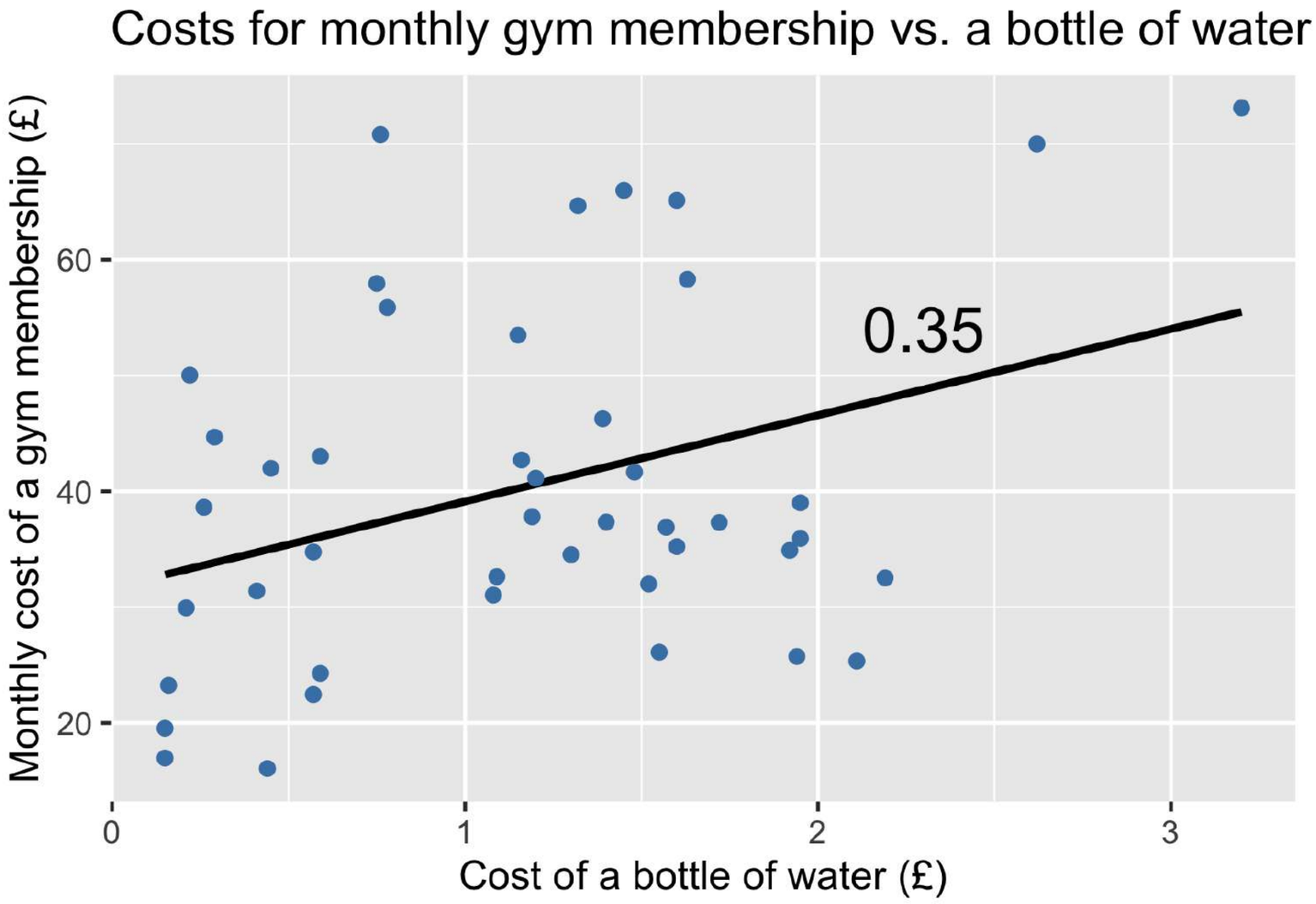
-0.75: as x increases, y decreases



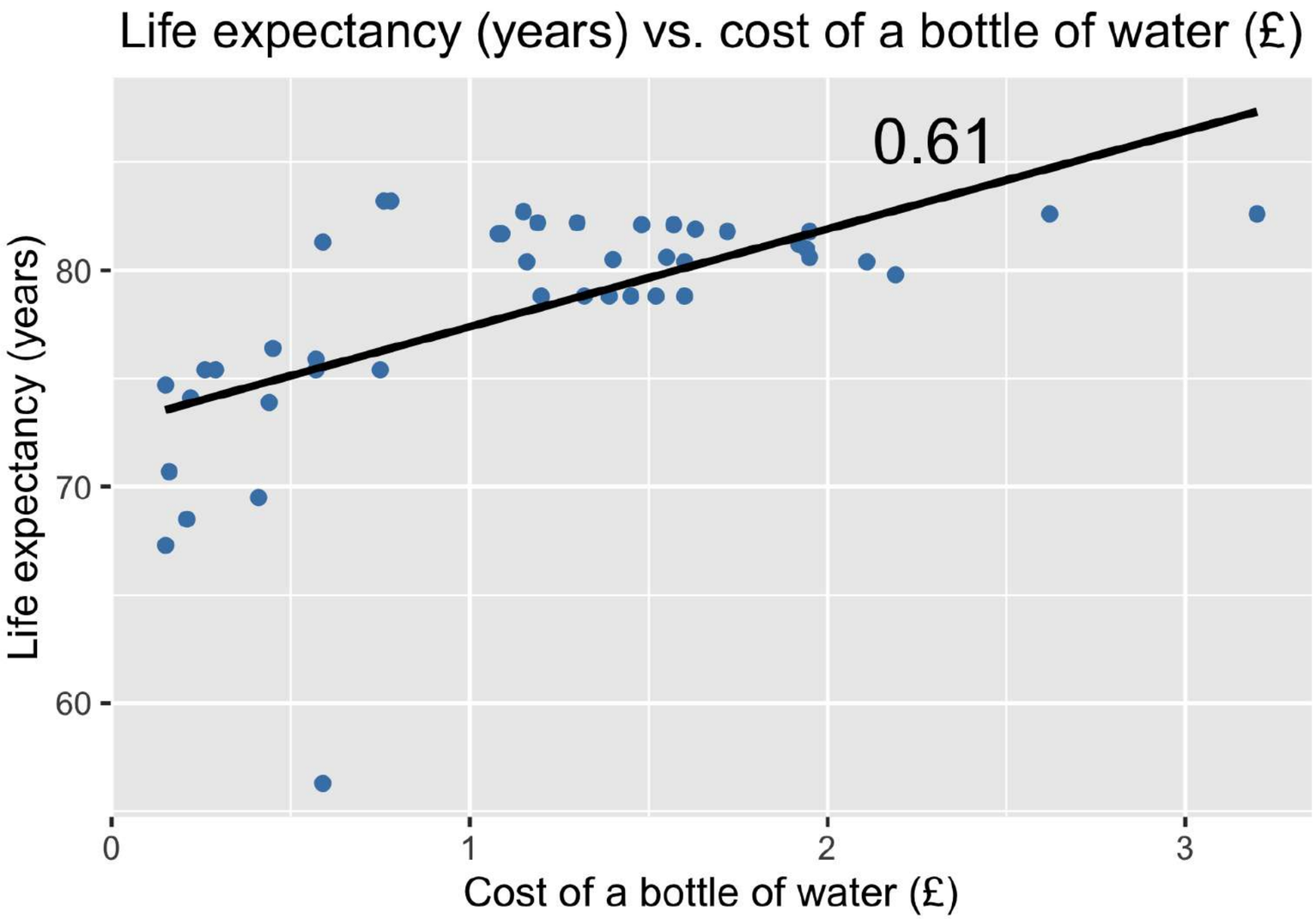
Gym costs vs. water costs



Adding a trendline



Life expectancy vs. cost of a bottle of water



Correlation does not equal causation

- Will increasing the cost of water result in an increase in life expectancy?

- Correlation does not equal **causation**

[1] Image credit: https://unsplash.com/@micheile; https://unsplash.com/@jon_chng

Confounding variables

- What else might be affecting life expectancy?
  - A bottle of water costs more in countries with strong economies
  - These countries generally offer access to high-quality healthcare
- The strength of the *economy* could be a **confounding variable**
  - A confounding variable is not measured, but may affect the relationship between our variables

Let's practice!

INTRODUCTION TO STATISTICS

datacamp

Interpreting hypothesis test results

INTRODUCTION TO STATISTICS

George Boorman
Curriculum Manager, DataCamp

datacamp

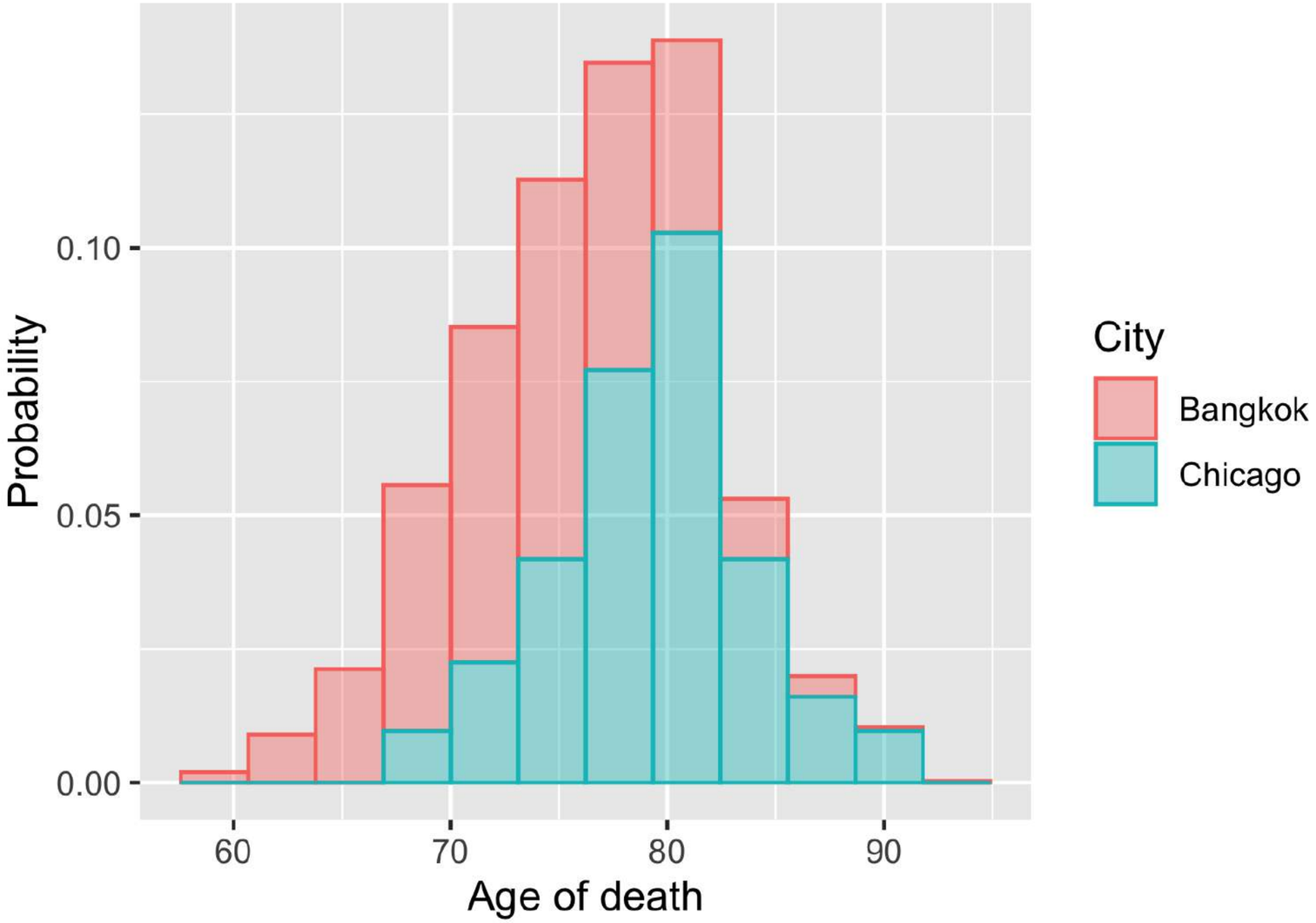
Life expectancy in Chicago vs. Bangkok

- **Null hypothesis:**
  - There is no difference in life expectancy between Chicago residents and Bangkok residents
- **Alternative hypothesis:**
  - Chicago residents have a longer life expectancy than Bangkok residents

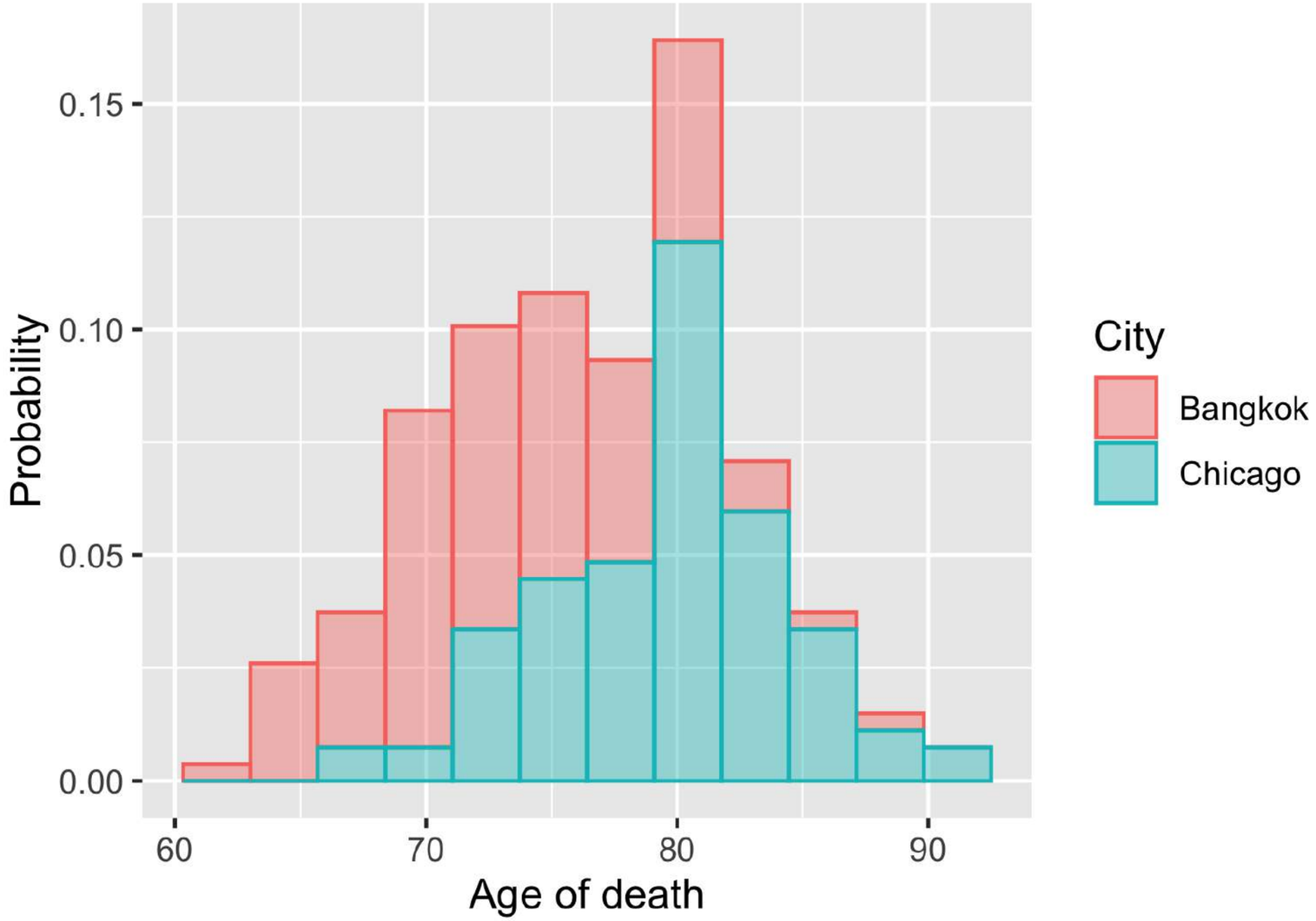
Chicago

Bangkok

Sampling distribution



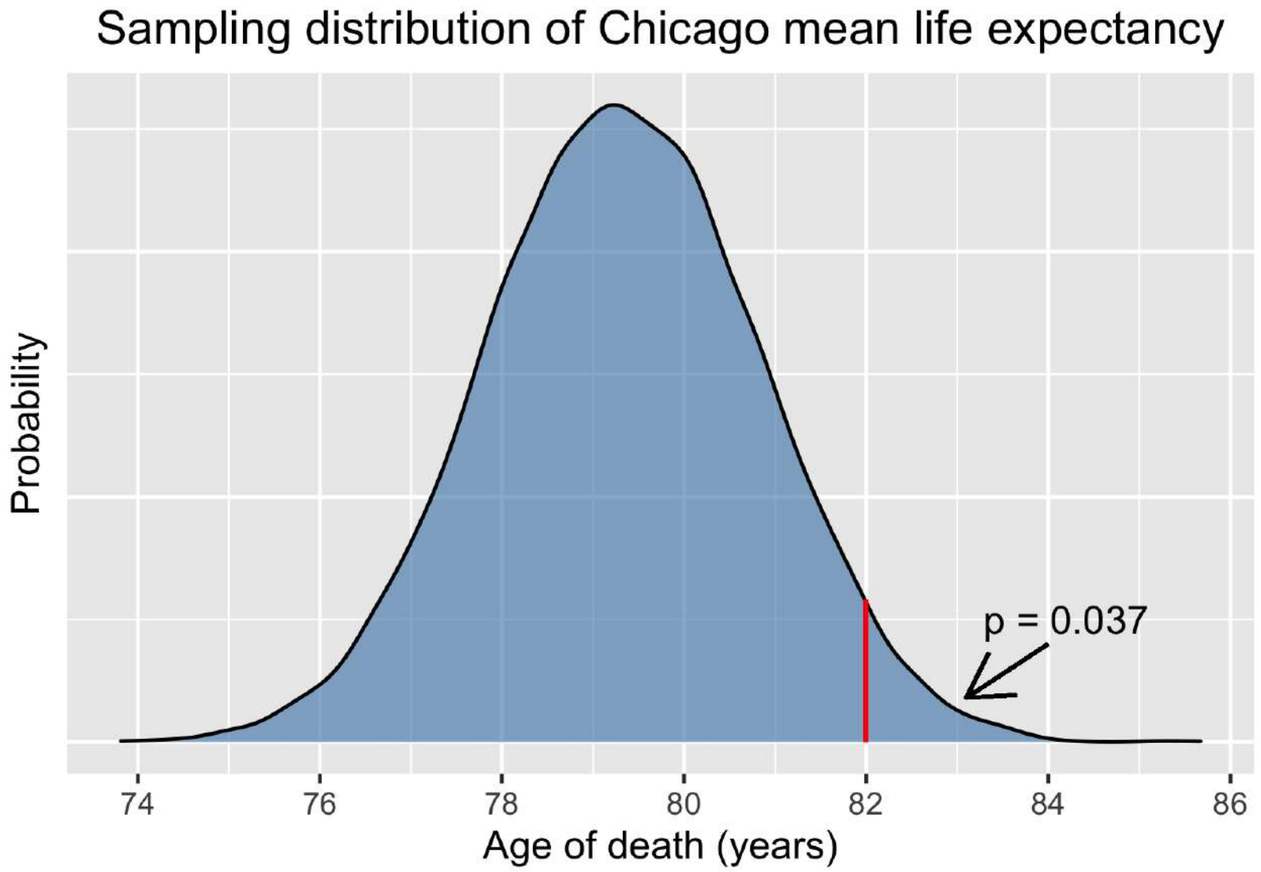
Different samples



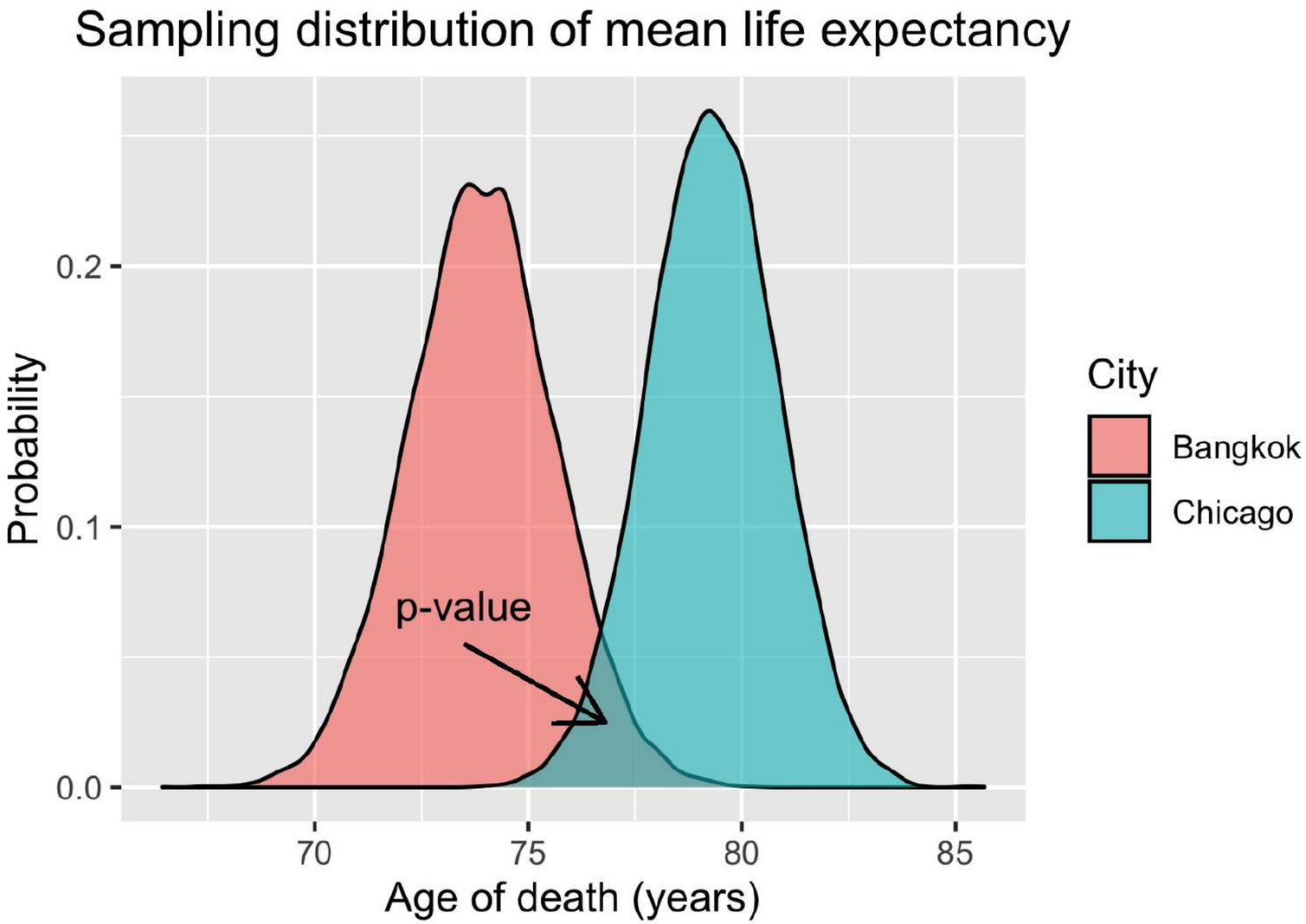
Sampling distribution of mean life expectancy

p-value

- p
  - Probability of achieving this result, assuming the null hypothesis is true



p-value


Sampling distribution of mean life expectancy
Probability
0.2
0.1
0.0
p-value
70
75
80
85
Age of death (years)
City
Bangkok
Chicago

Significance level (α)

- To reduce the risk of drawing a false conclusion:
  - Set a probability threshold for rejecting the null hypothesis
- Known as α or *significance level*
- Decided before data collection to minimize bias:
  - Otherwise they could choose a different α to serve their interests
- A typical threshold is 0.05
  - 5% chance of wrongly concluding that Chicago residents live longer than Bangkok residents
- If $p \leq \alpha$, reject the null hypothesis
- These results are said to be *statistically significant*

Type I/II error

	Null hypothesis is TRUE	Null hypothesis is FALSE
Reject null hypothesis	Type I Error	
Accept null hypotheis		

Type I/II error

	Null hypothesis is TRUE	Null hypothesis is FALSE
Reject null hypothesis	Type I Error	
Accept null hypotheis		Type II Error

Type I/II error

	Null hypothesis is TRUE	Null hypothesis is FALSE
Reject null hypothesis	Type I Error	
Accept null hypotheis	Correct conclusion	Type II Error

Type I/II error

	Null hypothesis is TRUE	Null hypothesis is FALSE
Reject null hypothesis	Type I Error	Correct conclusion
Accept null hypotheis	Correct conclusion	Type II Error

Drawing a conclusion

Let's practice!

INTRODUCTION TO STATISTICS

datacamp

Congratulations!

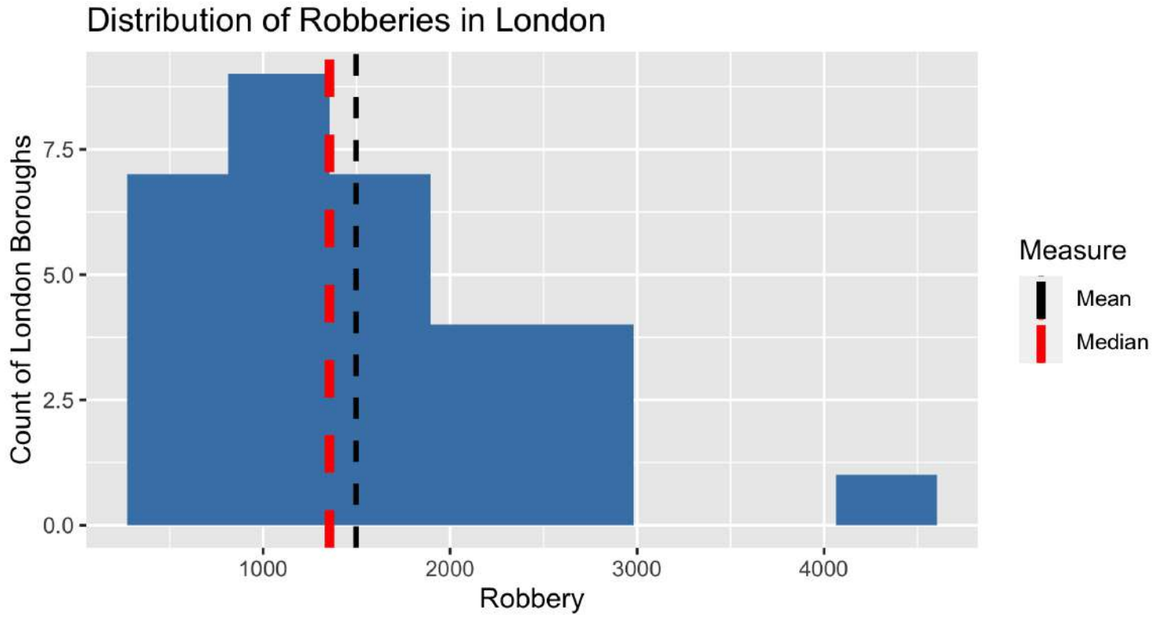
INTRODUCTION TO STATISTICS

George Boorman
Curriculum Manager, DataCamp

datacamp

What you've covered

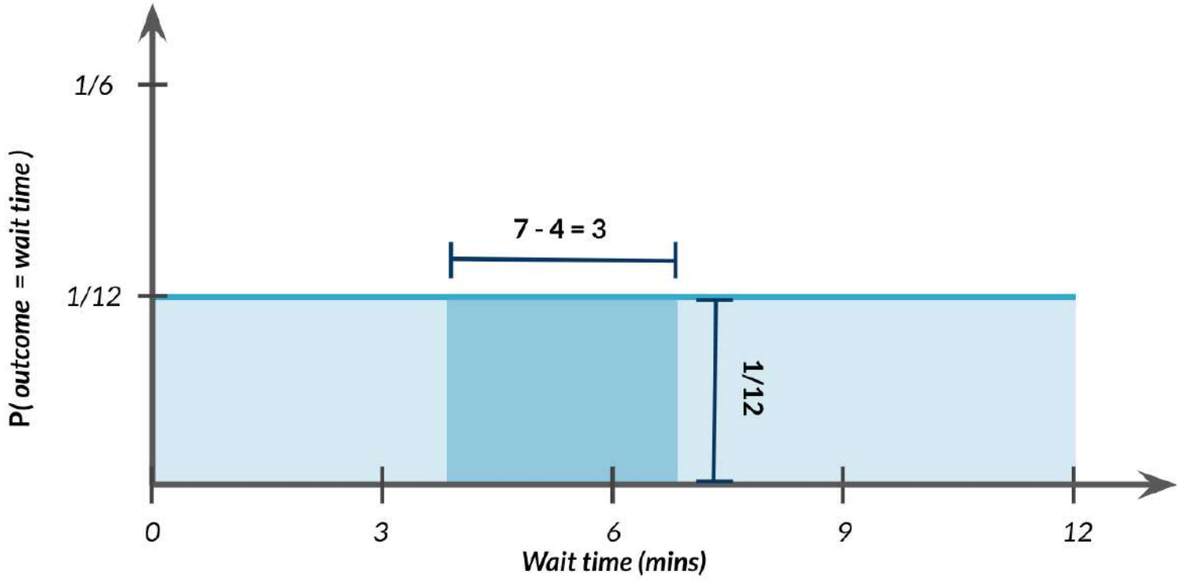
- Types of data
  - Ordinal, nominal, continuous, interval
- Descriptive and inferential statistics
- Measures of center
  - Mean, median, mode
- Measures of spread
  - Variance, standard deviation



What you've covered

- Probability

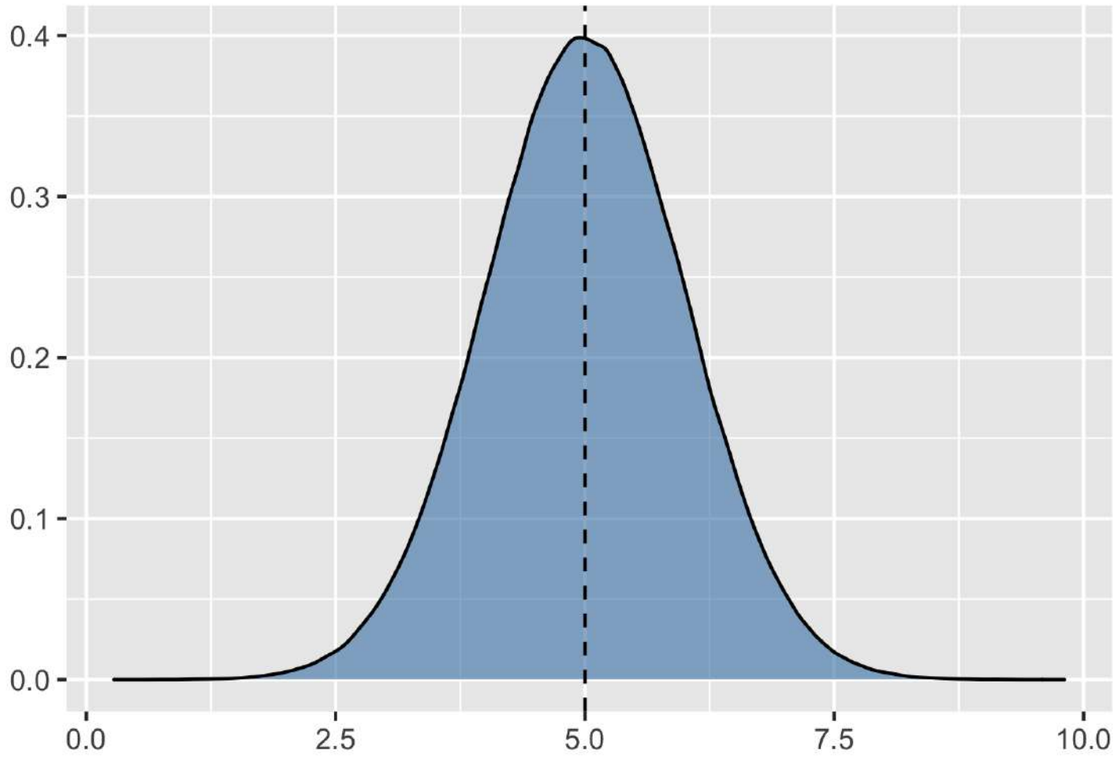
$$P(4 \leq \text{wait time} \leq 7) = 3 \times 1/12 = 3/12$$



- Conditional probability
- Discrete distributions
- Continuous distributions

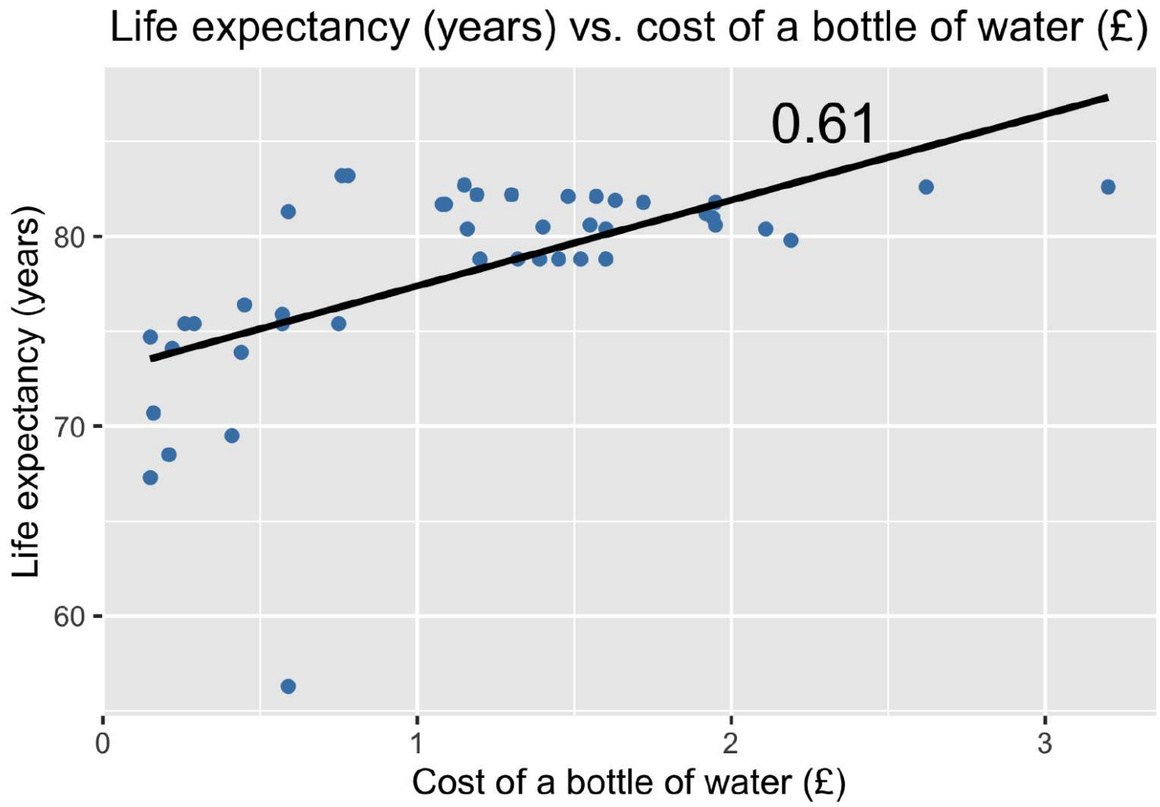
What you've covered

- The binomial distribution
- The normal distribution
- The Poisson distribution
- The central limit theorem



What you've covered

- Hypothesis testing
- Randomization, treatment, and control
- Correlation
- Interpreting hypothesis test results



Where to from here?

- Understanding Data Science
- Data Science for Business
- Understanding Machine Learning

Thank you!

INTRODUCTION TO STATISTICS