BCS1520: Statistics

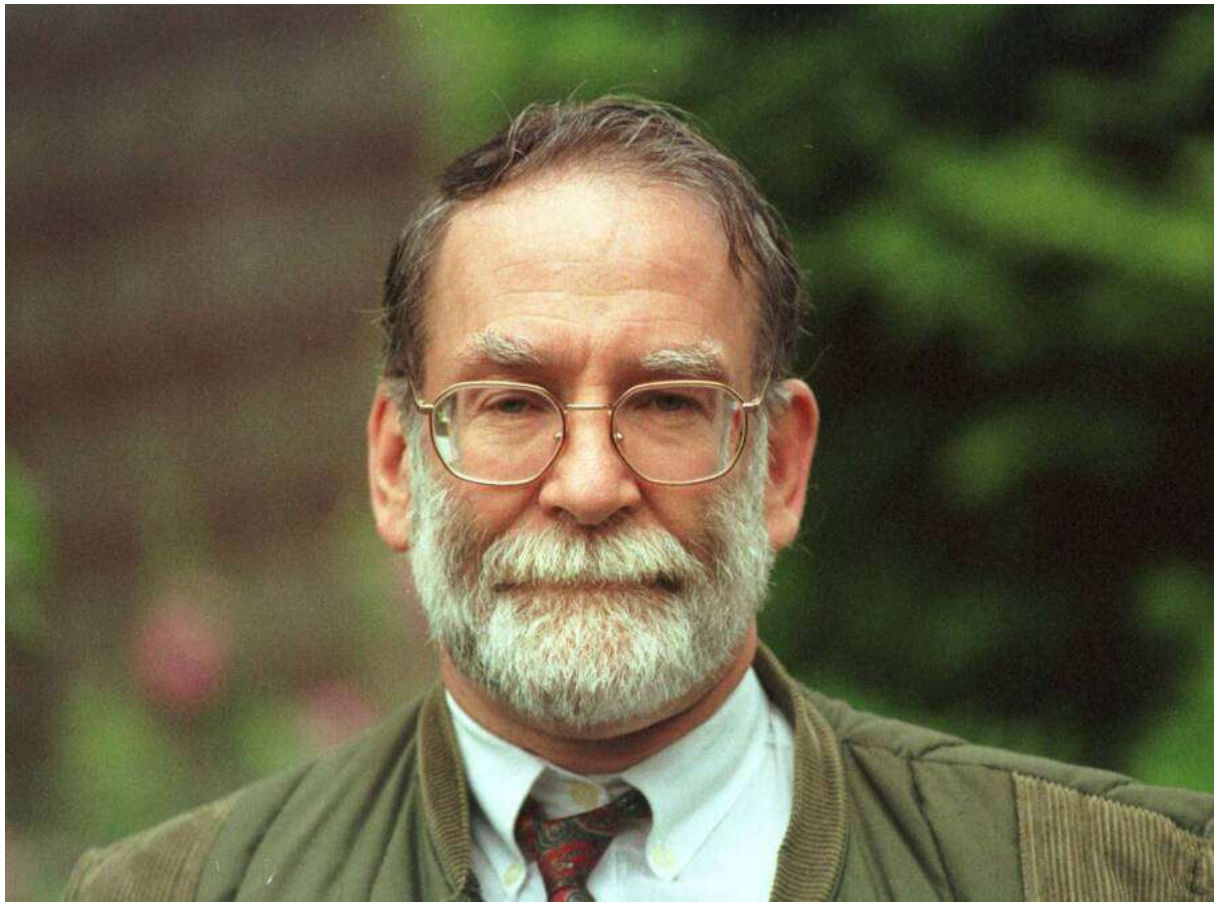
Lecture 01:
What is data science?

Anirudh Wodeyar

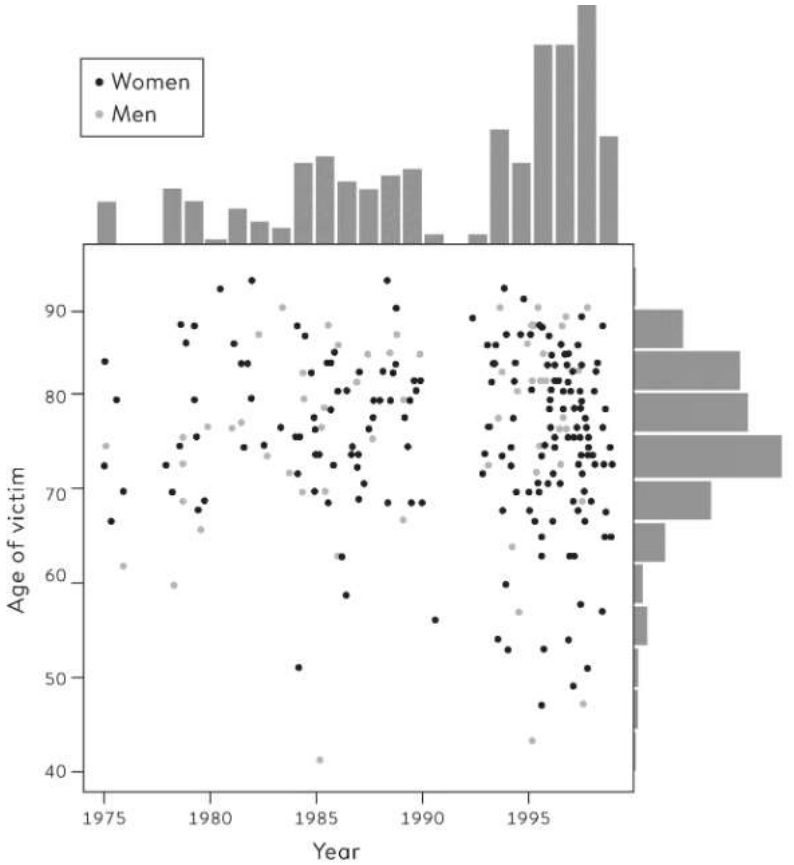
Maastricht University | Department of Advanced Computing Sciences

Using Data Science to judge a serial killer

- Harold Shipman: Britain's most prolific murderer
- Family doctor who killed ??? elderly patients



Using Data Science to judge a serial killer



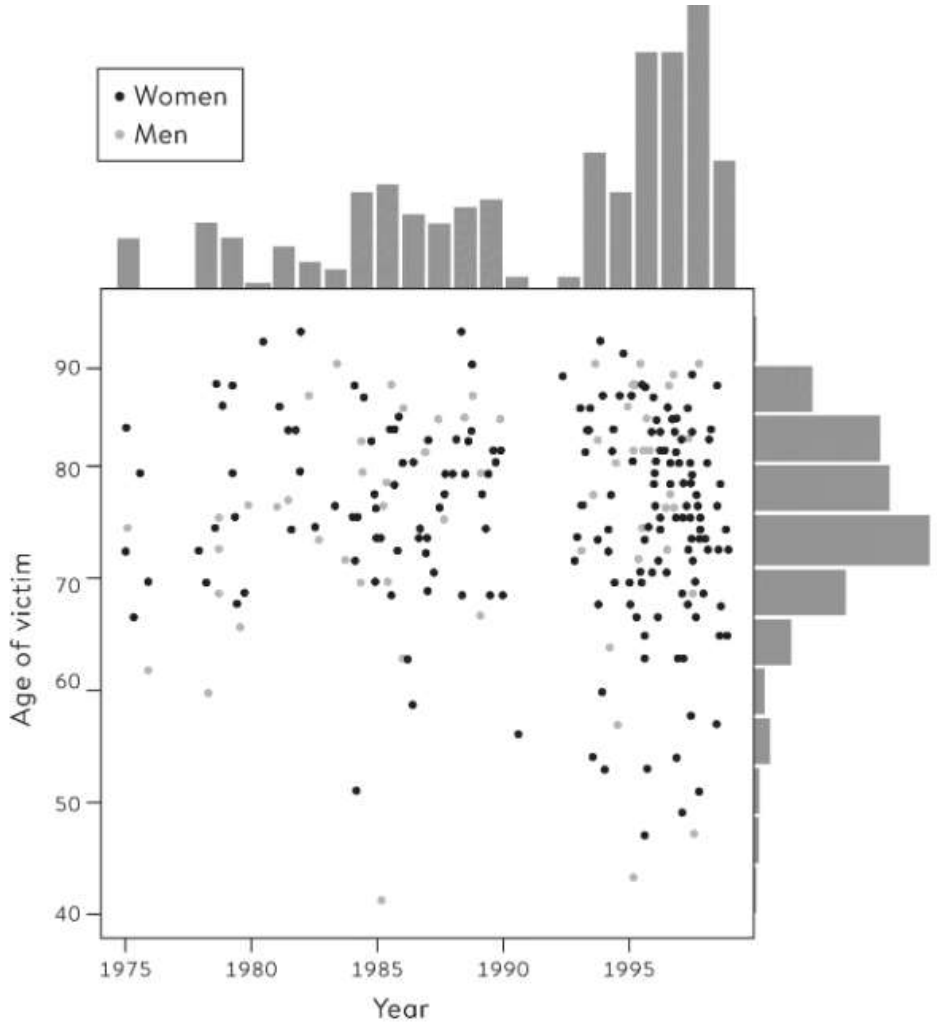
Using Data Science to judge a serial killer

What do we know about him?

- Family doctor
- Between 1975 to 1998 killed 215 of his elderly patients with an opiate overdose (maybe 45 more)
- Said nothing at his trial
- *Caught after he tried to forge a will of a patient*

Using Data Science to judge a serial killer

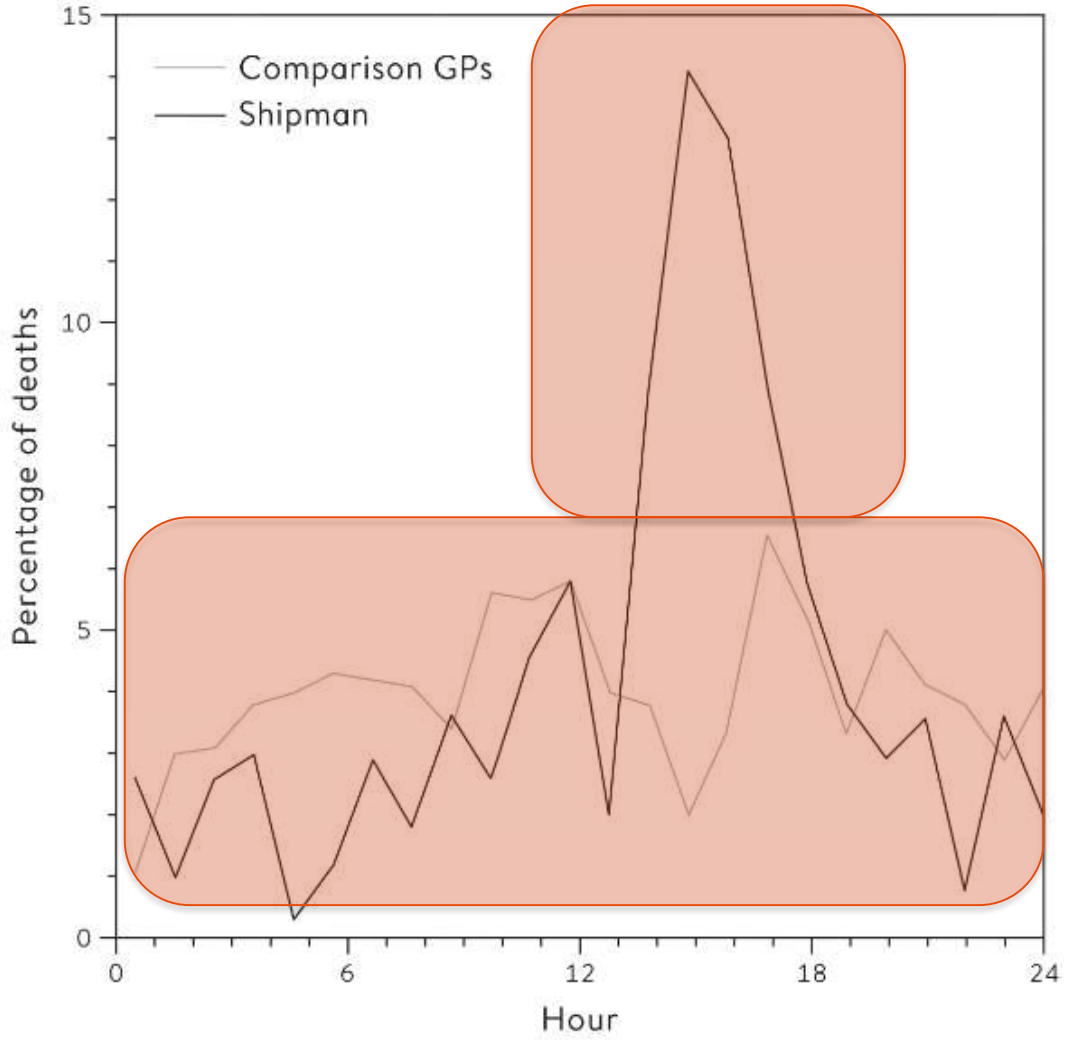
What were the characteristics of the people he murdered and when did they die?



Art of Statistics, David Spiegelhalter

Using Data Science to judge a serial killer

How did Shipman commit his murders?

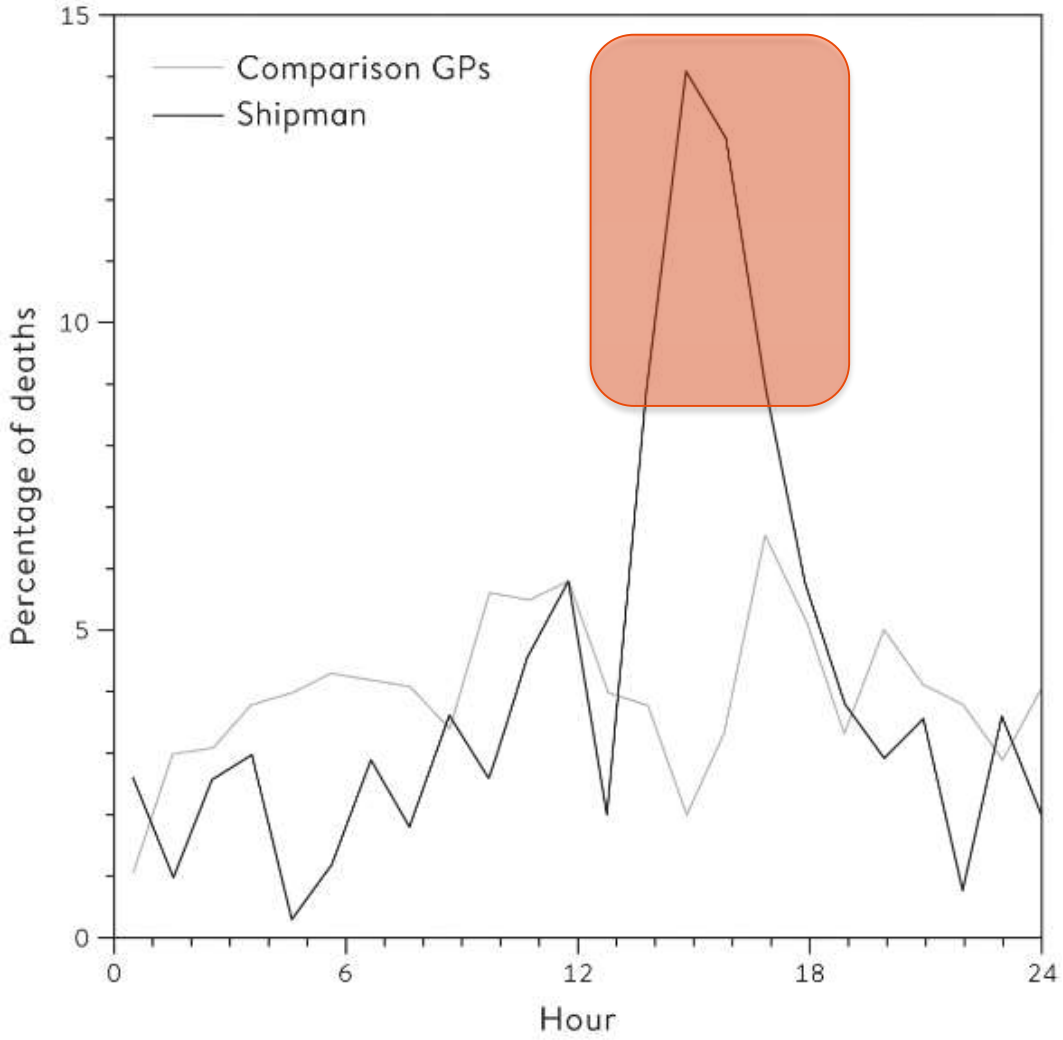


Art of Statistics, David Spiegelhalter

Using Data Science to judge a serial killer

Using Data Science to judge a serial killer

Dame Janet Smith: *"I still do feel it was unspeakably dreadful, just unspeakable and unthinkable and unimaginable that he should be going about day after day pretending to be this wonderfully caring doctor and having with him in his bag his lethal weapon … which he would just take out in the most matter-of-fact way."*



Art of Statistics, David Spiegelhalter

One branch of data science: Forensic statistics

- No models, no fancy math, just looking at data carefully, again, and again
- After critically assessing data, we need to plot it as well: how do we decide how/when to do that?
- Could also have used formal statistical analysis, but how?

The World Can Be Quantified

To understand the nature of Shipman's murders, we abstracted out the details, we turned people into numbers, each person into a dot (slightly terrifying, but in the end, useful).

Definition

Data science and statistics is the process of abstracting the world into numbers and then finding the patterns across those numbers using critical reasoning and mathematical tools.

The World Can Be Quantified

- As an example, consider the question: “How many computing machines are in this building?”
- How do we go about answering this question?
- First, we need to answer what is a computing machine? Is your phone a computer? A calculator?

The World Can Be Quantified

- Now, how can we quickly assess how many computers are in this building?
- How many lab rooms on this floor?
- Each student has a laptop. How many lecture halls?
- How many students in the study hall on the ground floor?
- How can we verify this?

The World Can Be Quantified

- Quantification can be messy and yield inaccurate numbers relative to our intended goals.
- What if we wanted to know how much an injection hurt and used a survey from 0 to 10?

The World Can Be Quantified… but this needn't be meaningful

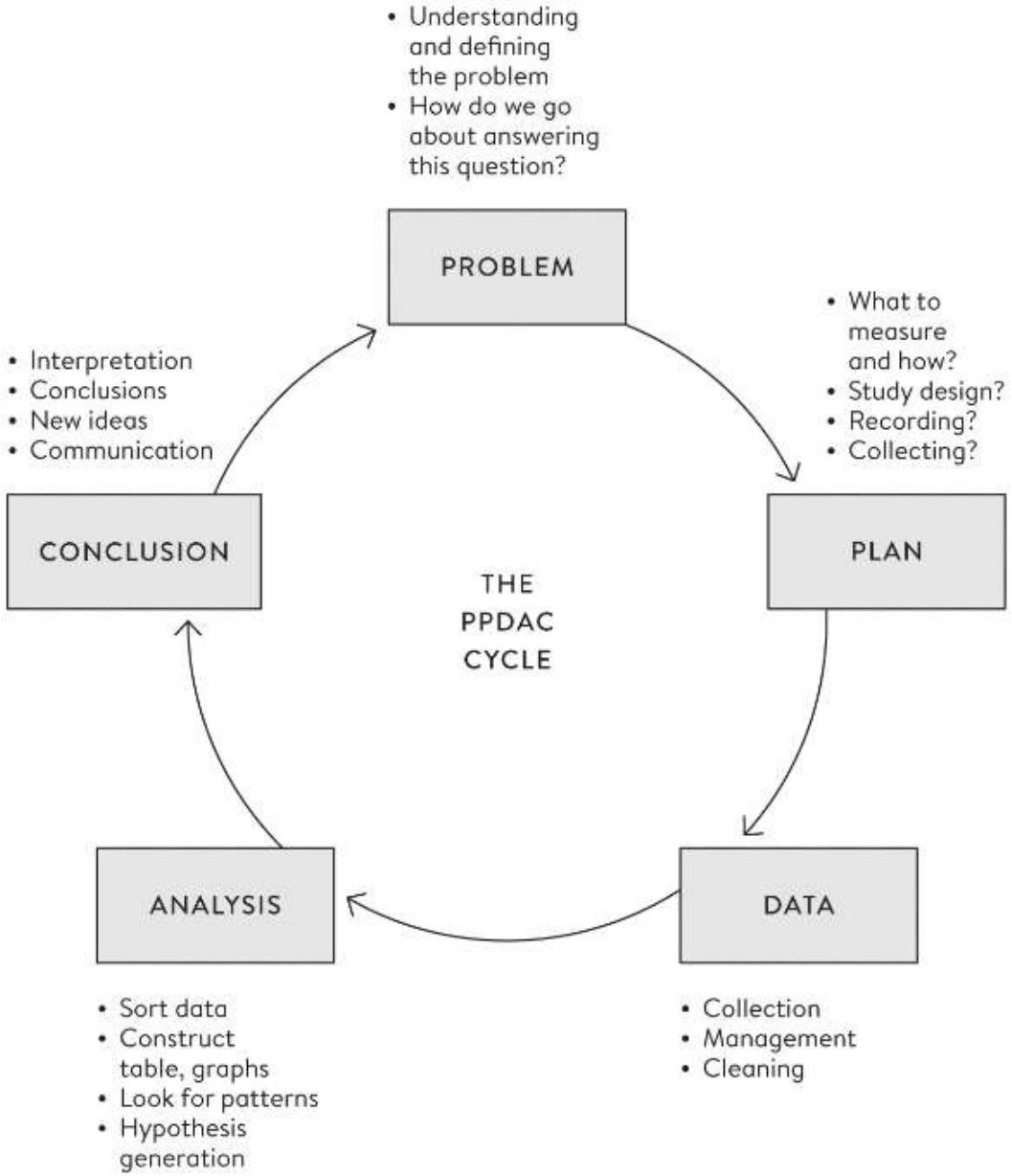
- Data science and its cousin - statistics - are always constructed based on judgements (e.g. what is a computing machine?).
- **Delusional to believe data is the real world.**
- But, we can still understand an approximated fact about the real world with data.
- Sometimes, that fact can be extrapolated.

The World Can Be Quantified… but this needn't be meaningful

Data has two main limitations:

- 1. Data is always an imperfect measure of what we are really interested in: so how can you improve the measure?
- 2. Anything we choose to measure will differ from place to place, person to person and time to time: so how can you tame or at least understand that variability?

Answering these two questions well and in a meaningful manner is the main goal of data science. How can we do that?


• Understanding and defining the problem
• How do we go about answering this question?
PROBLEM
• What to measure and how?
• Study design?
• Recording?
• Collecting?
PLAN
DATA
• Collection
• Management
• Cleaning
ANALYSIS
• Sort data
• Construct table, graphs
• Look for patterns
• Hypothesis generation
CONCLUSION
• Interpretation
• Conclusions
• New ideas
• Communication
THE PPDAC CYCLE

What are you hoping to get from this class?

Class Goals

- Achieve a basic level of data literacy:
  - the ability to understand the principles behind learning from data,
  - carry out basic data analyses,
  - understand what are good and bad visualizations, &
  - critique the quality of claims made based on data.

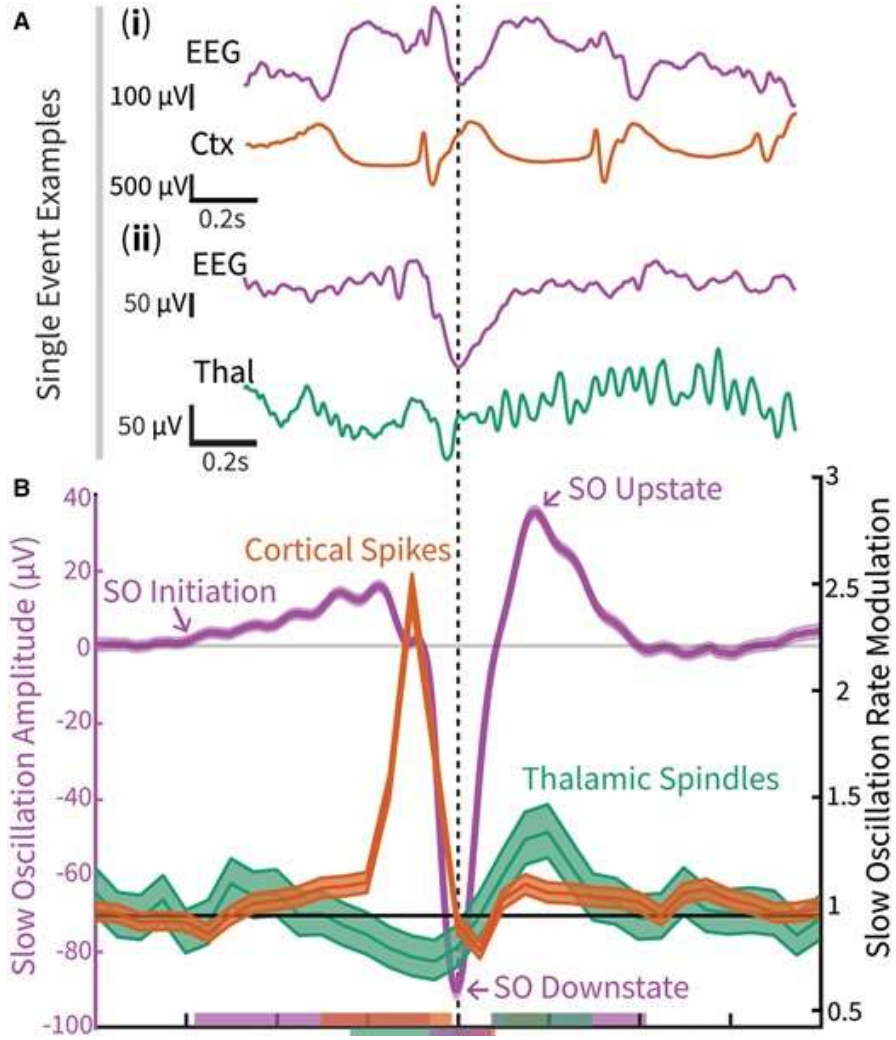
Course organization

About me

About me

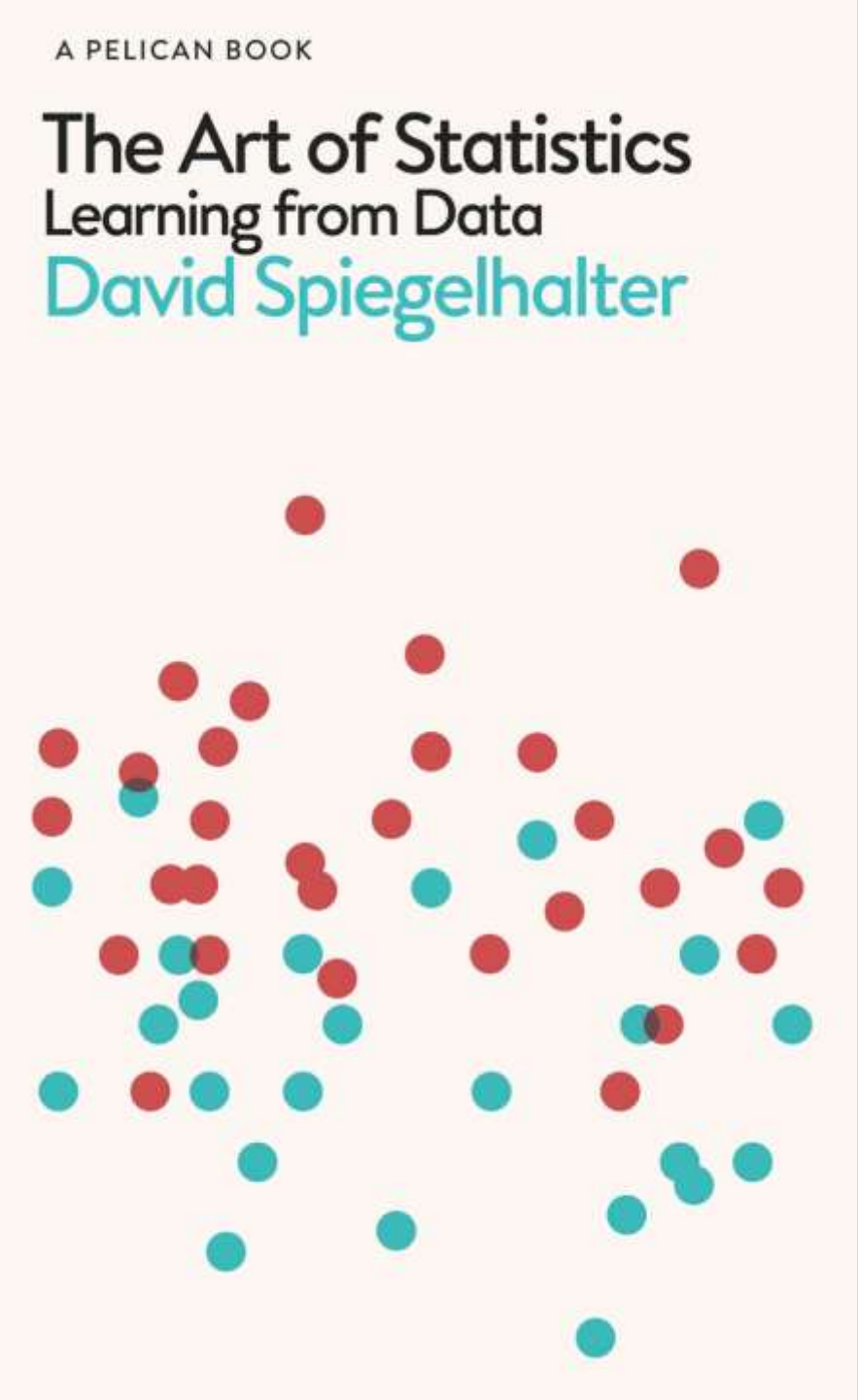
About me: Research

- Love thinking about how statistics can help us understand different scales of neural activity
- And how we can build tools that allow us to change our neural activity and from that, our minds



Textbook

Art of Statistics by David Spiegelhalter

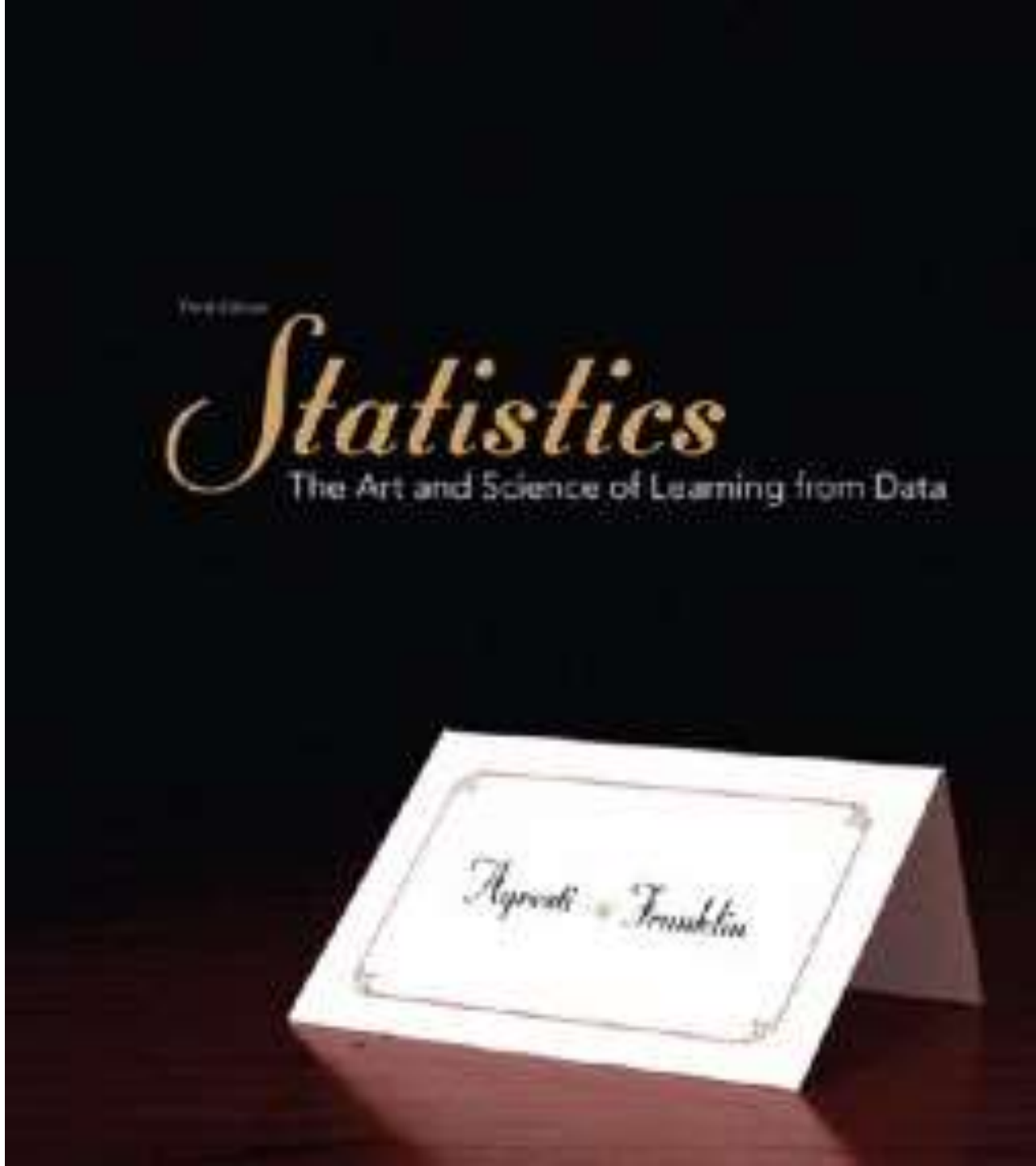


Maastricht University

Support Resource

Statistics
The Art and Science of Learning from Data

Agresti and Franklin
3rd Edition



Maastricht University

Instruction format

Read chapters

Lectures (Concept & Live analysis)

Labs

Organization of Lectures

- Introduce concepts
- Live analysis: examples of these concepts coded up in Python

Labs

- Processing data
- Analyzing data
- Visualizing data

Organization of Labs

- Labs are run in two groups

	Lecturer	Participants
Group A	Tim Dick	Last name starting with A – H
Group B	Niloufar Yousefimanesh	Last name starting with I – Q
Group C	Luuk Verblij	Last names starting from R - Z

Grading

The following components will be examined, and will count towards the final grade of the course:

- Written Exam (WE), graded 0.0-10.0, weight 100% (mandatory)

Assessment component	Form and extent (length)	Grading scheme	Form of the Resit
Written exam (mandatory)	2 hours	Points per exam question depend on difficulty and expected length.	Regular resit in the ongoing academic year

Exam

- Closed book exam
- **Duration:** 120 minutes, without breaks.
- **Allowed aids:** Pen, calculator from DACS allowed calculator list.
- Mock exam will be posted on CANVAS (time TBD)
  - Multiple Choice: related to discussed concepts
  - Statements: true/false + explanation
  - Data analysis problems: describe the steps to take

Communication

- Lectures
- Labs
- Canvas discussion board
- Please, do NOT send us e-mails with questions

Course Overview

Tentative Schedule

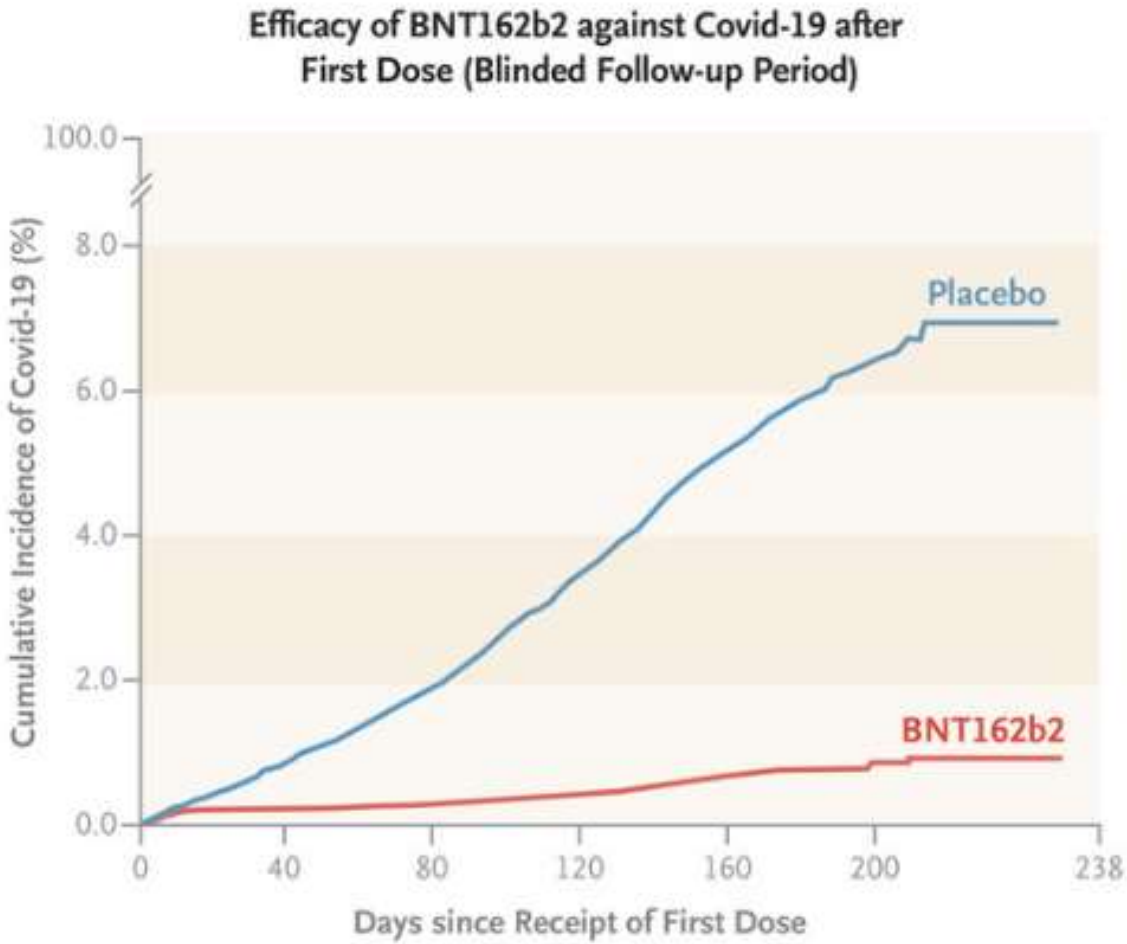
Wk	Activity	Topic and Materials	Date and Time	Location
1	Lecture 1	What is data science? Categorical data, data summaries and data visualization. Chapter 1 from AoS Chapter 3.1 and 2 from AoSoD (prioritize topics covered in class)	Mon 31/03/2024 08:30-10:30h	EPD150 Conference Hall
	Lecture 2	Data summaries and data visualization contd. Inductive Inference. Chapter 1.2, 2 and 6.2 from AoSoD Chapter 2 and "Bell-shaped Curve" in 3 from AoS	Thu 2/04/2024 11:00-13:00h	EPD150 Conference Hall
	Tutorial 1	Working with data in Python	Thu 3/04/2024 16:00: 18:00h	PHS1 [1] C0.020 [2] C0.004 [3] C0.016
2	Lecture 3	Machine Learning Chapter 6 from AoS	Mon 7/04/2024 8:30-10:30h	EPD150 Conference Hall
	Tutorial 2	Visualizing Data [Python]	Wed 09/04/2024 11:00-13:00h	PHS1 [1] C0.020 [2] C0.004 [3] C0.016
	Tutorial 3	Decision Trees [Python]	Thur 10/04/2024 16:00-18:00h	PHS1 [1] C0.020 [2] C0.004 [3] C0.016
3	Lecture 4	Randomized control trials and causality. Chapter 4 from AoSoD Chapter 4 from AoS	Mon 14/04/2024 8:30-10:30h	EPD150 Conference Hall
	Tutorial 4	Linear Regression [STRIKE!]	April 16	PHS1 [1] C0.020 [2] C0.004
	Lecture 5	Probability Theory Chapter 8 from AoS	Thurs 17/04/2024 16:00-18:00h	EPD150 Conference Hall

Tentative Schedule (continued)

4	Lecture 6	**Probability and Statistics 1** **Chapter 6 from AoSoD** **Chapter 8 from AoS**	**Wed 23/04/2024 11:00-13:00h**	**EPD150 Conference Hall**
	Tutorial 5	Probability theory	Wed 23/04/2024 13:30-15:30h	PHS1 [1] C0.020 [2] C0.004
	Tutorial 6	Bootstrapping for uncertainty, central limit theorem, probability distributions [Python]	Thu 24/04/2024 16:00-18:00h	PHS1 [1] C0.016 [2] C0.020
5	Lecture 7	Probability and Statistics 2. Chapter 7, 8 from AoSoD	Mon 28/04/2024 8:30-10:30h	EPD150 Conference Hall
	Tutorial 7	Confidence intervals and basic one-sample tests [Python].	Wed 30/04/2024 8:30-10:30h	PHS1 [1] C0.016 [2] C0.020
6	Lecture 8	Hypothesis Testing Chapter 10 from AoS Chapter 9 from AoSoD	Thurs 01/05/2024 16:00-18:00h	EPD150 Conference Hall
	Tutorial 8	Hypothesis Testing [Python].	Thu 07/05/2024 11:00-13:00h	PHS1 [1] C0.004 [2] C0.016
	Tutorial 9	Mock Exam	Thu 08/05/2024 16:00-18:00h	PHS1 [1] C0.016 [2] C0.020
	Tutorial 10	Review	Thu 12/05/2024 08:30-10:30h	PHS1 [1] C0.016 [2] C0.020

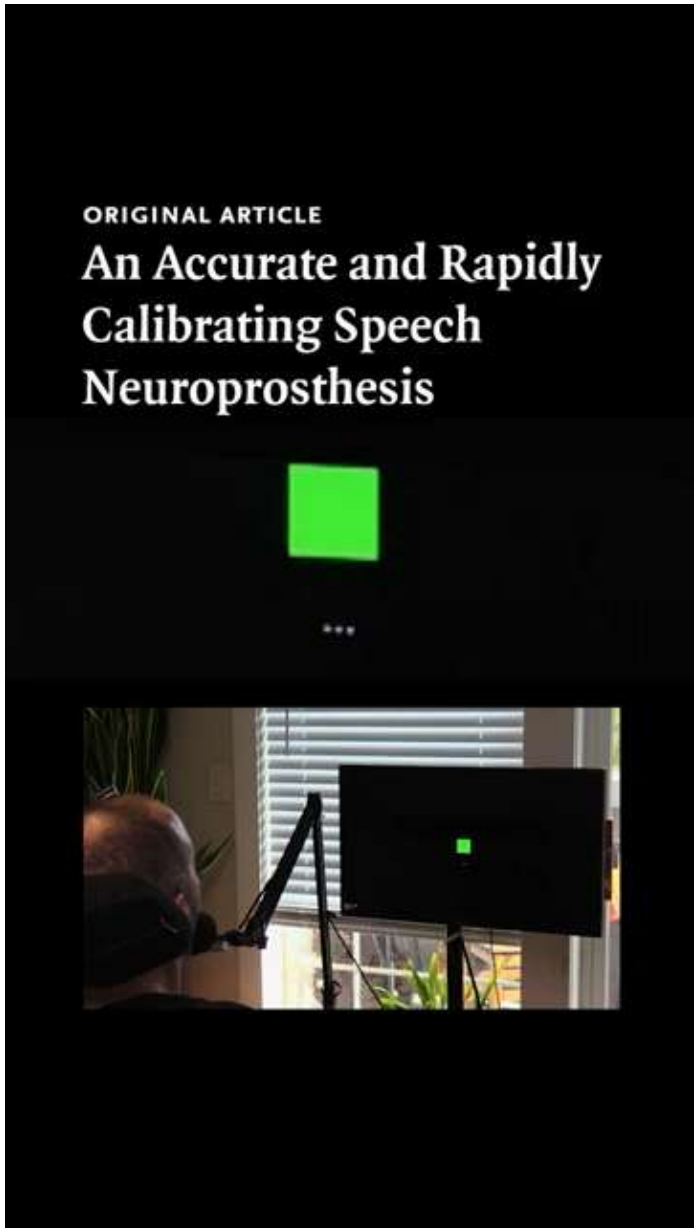
What is statistics good for?

- Vaccine and drug testing
- How can we know whether a given vaccine or drug actually does what we think it should do?



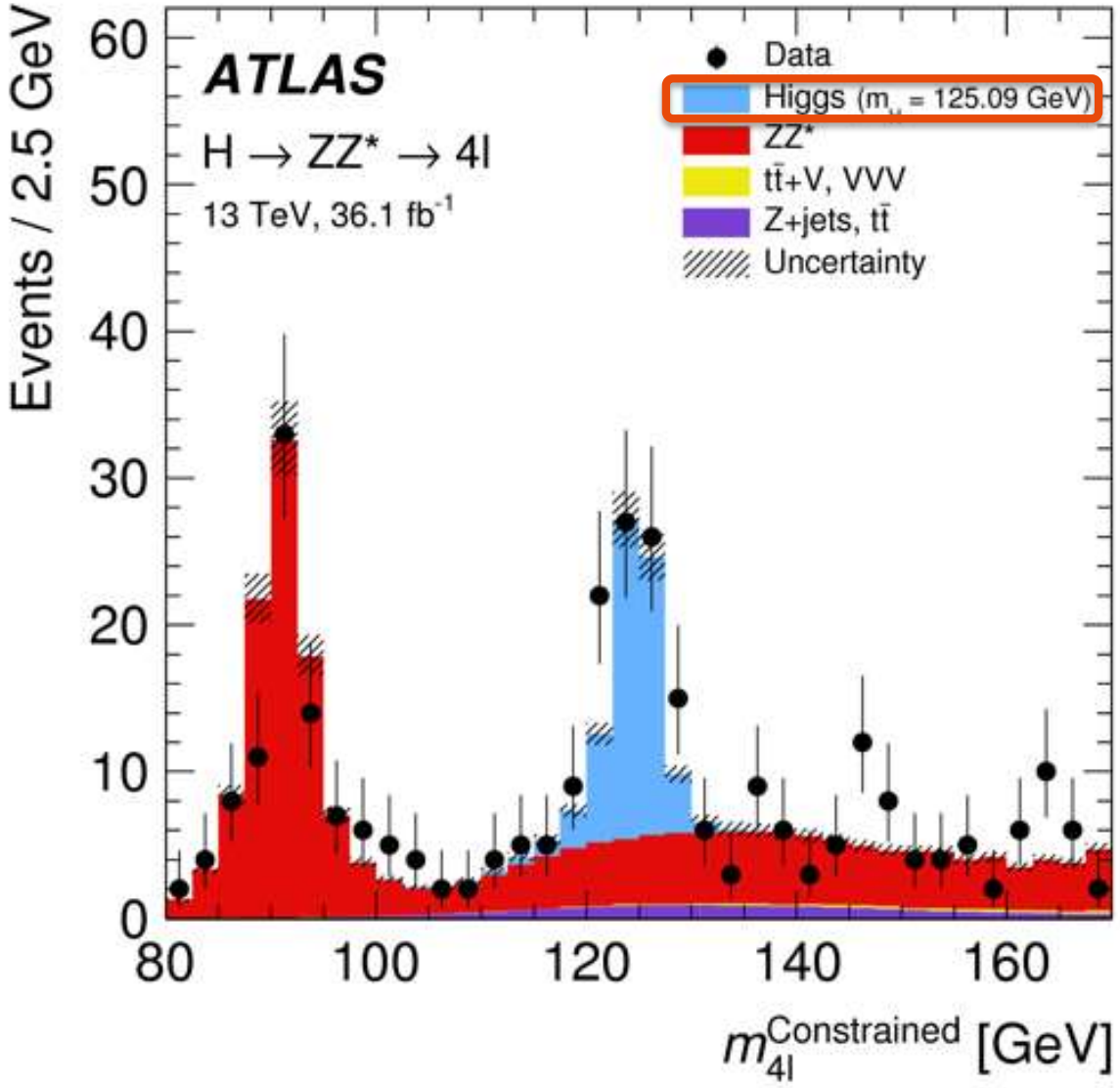
What is statistics good for?

- Helping people who have lost the ability to speak to be able to speak!
- This is a combination between statistics and its sibling fields machine learning and optimization.

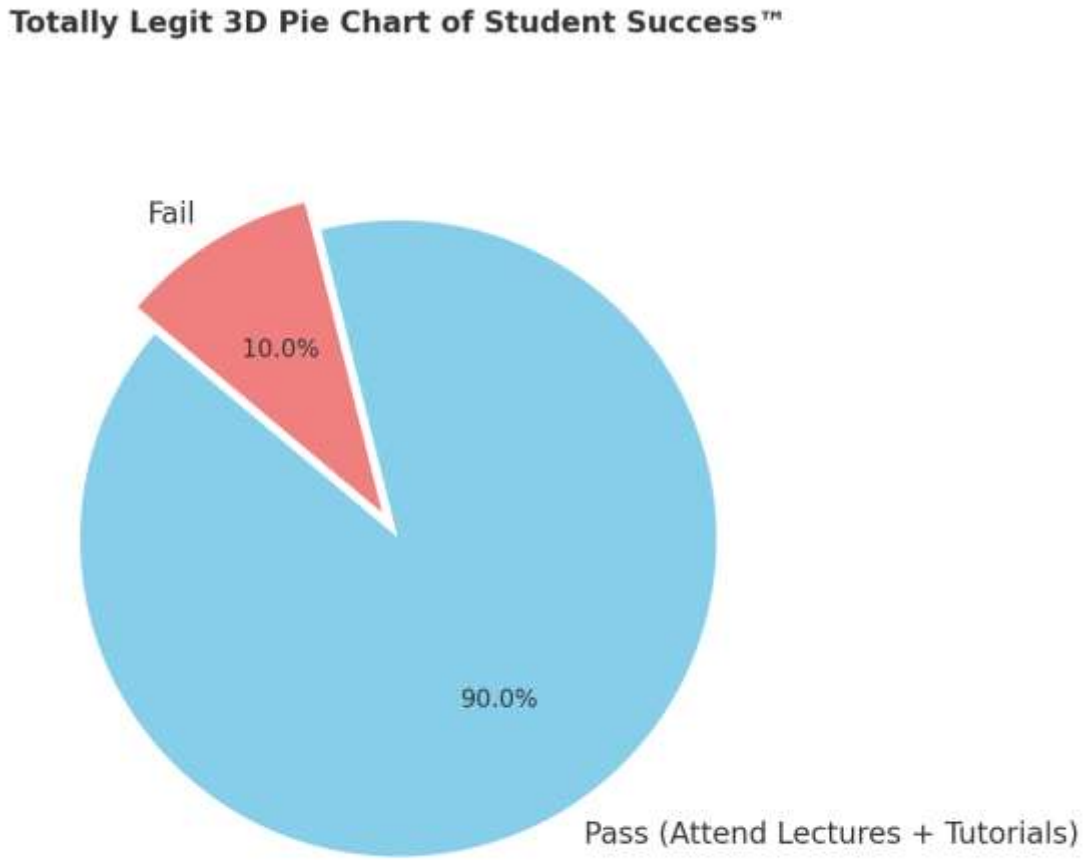


What is statistics good for?

- Identifying the fundamental particles of our world.
- The Higgs Boson's presence was confirmed with a statistical test.



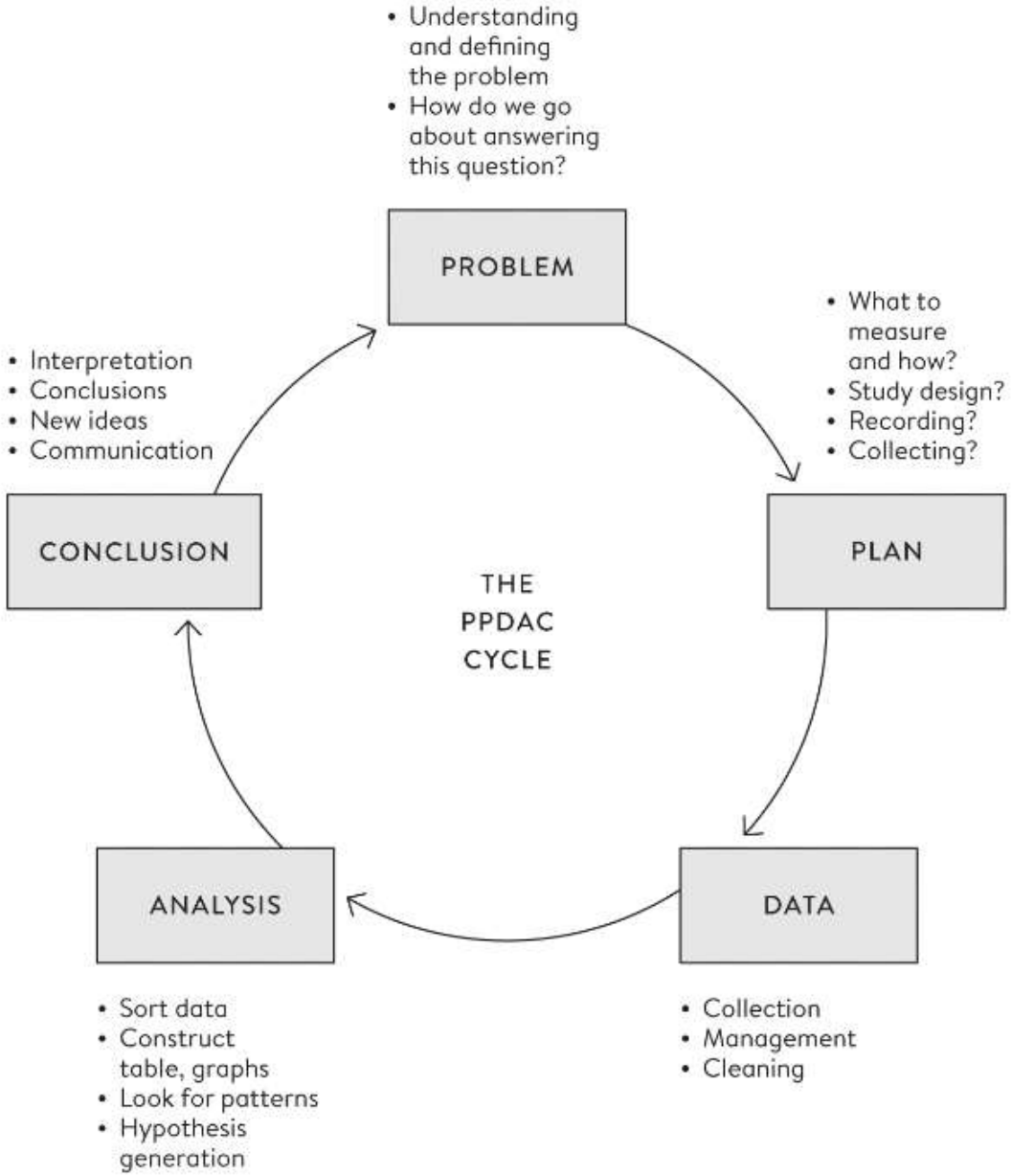
What is statistics good for?



Studying categorical data and percentages

What happened to children having heart surgery in Bristol between 1984 and 1995?

- Joshua was 16 months old and needed a critical heart surgery.
- He was at Bristol Royal Infirmary where there were stories about poor surgical survival rates, but his parents didn't know.
- There was even a last-minute meeting to consider canceling his surgery but parents weren't informed.
- Joshua passed away on the operating table.
- Our problem: Was this to be expected?


• Understanding and defining the problem
• How do we go about answering this question?
PROBLEM
• What to measure and how?
• Study design?
• Recording?
• Collecting?
PLAN
DATA
• Collection
• Management
• Cleaning
ANALYSIS
• Sort data
• Construct table, graphs
• Look for patterns
• Hypothesis generation
CONCLUSION
• Interpretation
• Conclusions
• New ideas
• Communication
THE PPDAC CYCLE

Data: Defining the variables.

- We have two variables (what's that?): child and death.
- How should we define them?
- **Definition**
- **A variable** is any measurement that can take on different values in different circumstances but is inter-related by the fact that it is the same *type* of measurement every time.
- DS and team defined an unsuccessful surgery as "**death within 1 month**".

Analysis: What were surgical outcome rates?

- Bringing data together from many sources can be hard.
- After integrating data across different sources:
  - Average risk across other hospitals suggested only 32 deaths should have happened for the number of surgeries performed.
  - There were, instead, **62 deaths** at Bristol. 30 more than expected.

Conclusion: Bristol had killer surgeons.

- Given how many more deaths were happening than expected, we can say with some certainty that Bristol did not have a good surgery department.
- BUT. There's always exceptions.
- This depends on the assumption that the cases that came through Bristol were the same as cases that came through other hospitals. What if the sicker kids tended to come to Bristol?
- We know this was not the case. However, this illustrates the cycle of PPDAC: We are now back at the **Problem** stage.

Categorical Data

Definition

Binary data is data that records whether individual events have happened or not (e.g. death vs no death), as these situations can take on only one of two values, they are binary (0/1). This is one form of categorical data.

For example: Assuming 100 children (<16 years old) passed through the Royal Bristol Infirmary, for each child we can mark them as 0 or 1 defining whether or not they passed away after surgery.

Categorical Data

Definition

Binary data can be summarized by a percentage or by two numbers: the total number of possible events and the sum of 1s (or equivalently the number of 0s).

For example: For the 100 children, we can describe the summary as 95 children survived out of 100 children or that there was a survival rate of 95%.

Note that already we have abstracted away a LOT of detail. That detail is usually what brings us back to the P of PPDAC.

Categorical Data

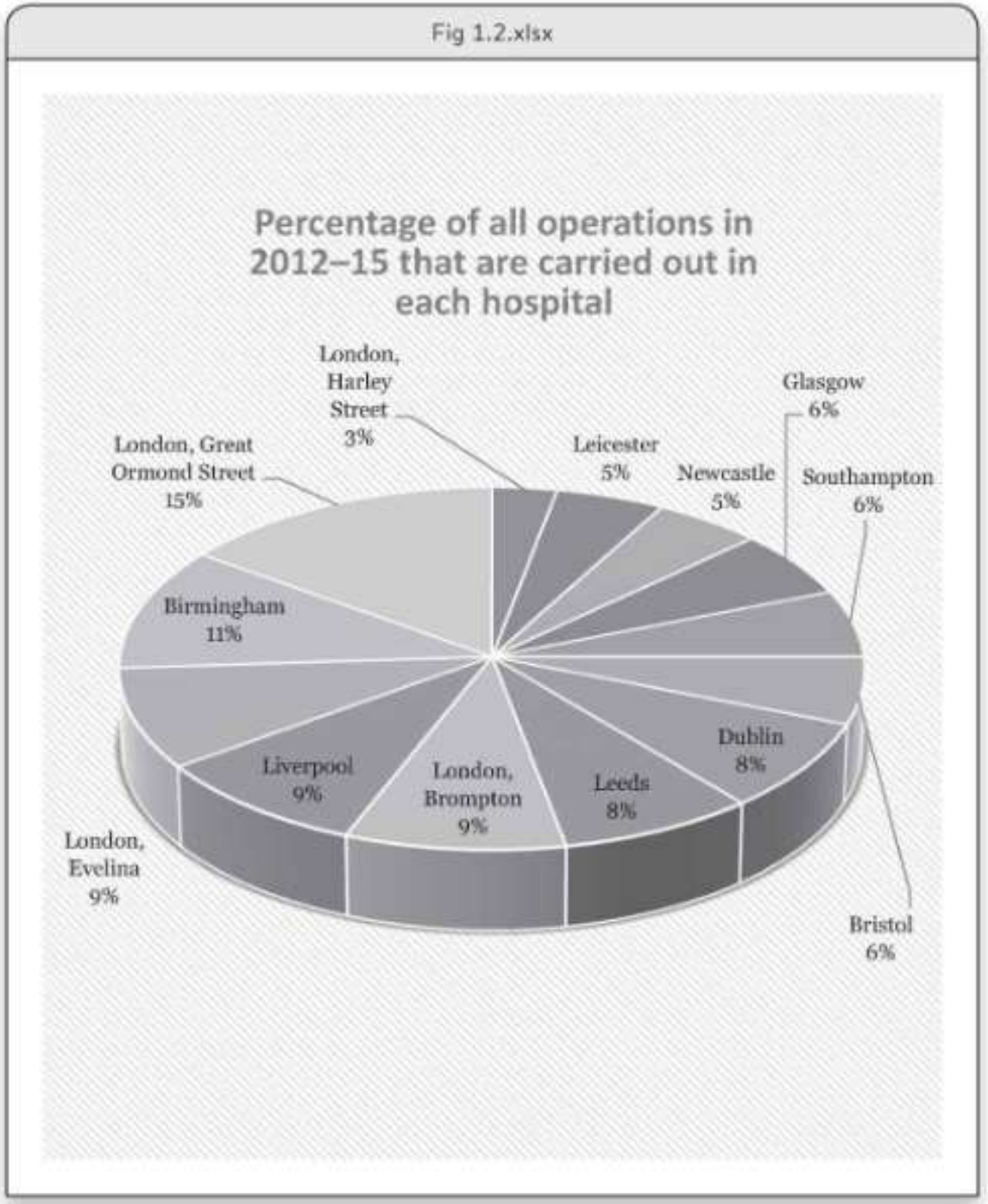
Definition

Categorical variables are measurements that can take on two or more categories (binary data is a form of categorical data).

- Unordered categories: Person's country of origin, color of a car or the hospital where an operation takes place (we can abstract them by assigning a random integer for each individual category).
- Ordered categories: Rank of military personnel, grade that you receive.
- Numbers that have been grouped: Levels of obesity (defined as ranges of BMI) or whether you are a child or an adult.

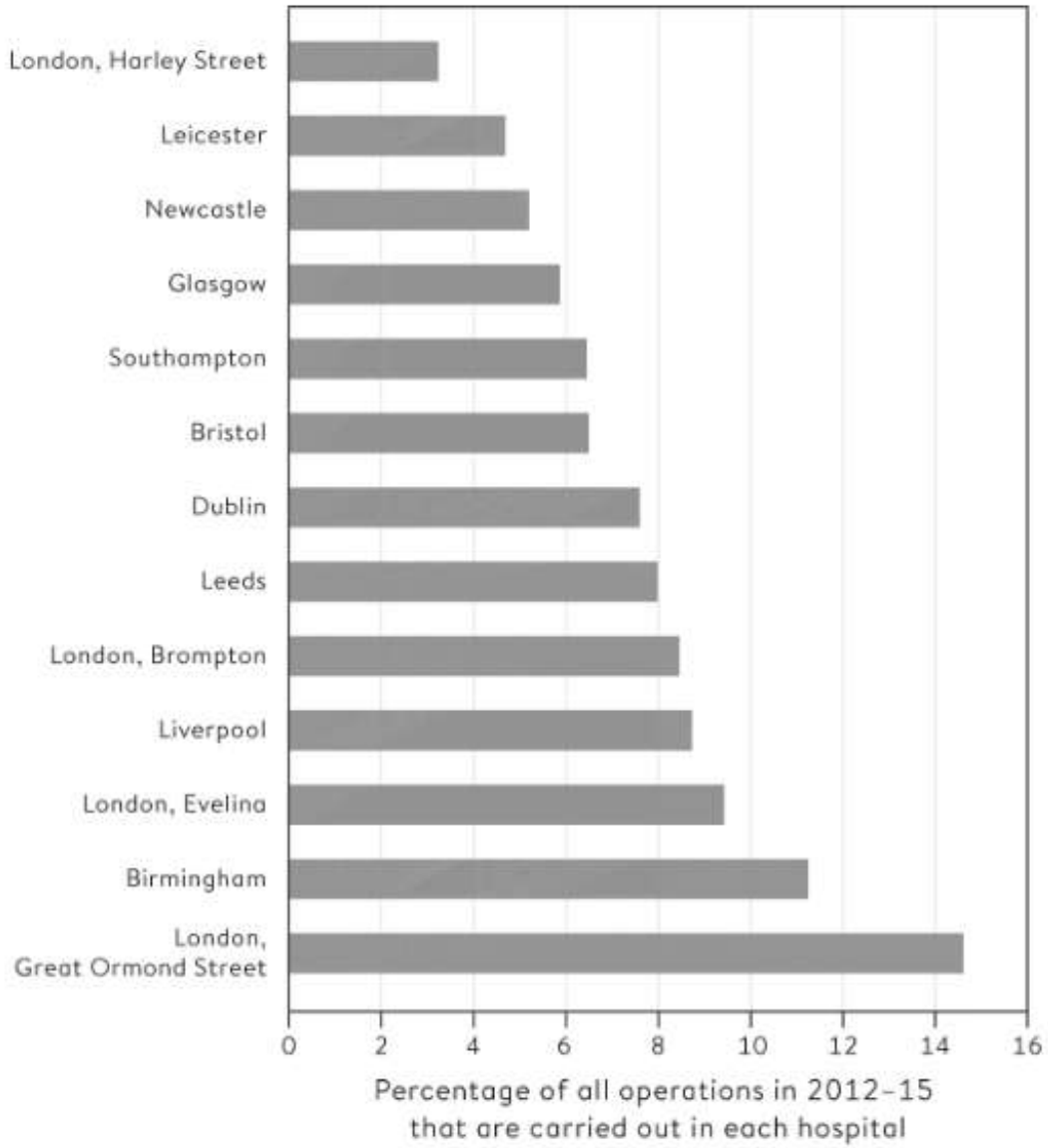
The urge to use pie charts for categorical data

- What were we trying to see with this plot?
- Why shouldn't we use this plot?



An alternative presentation

- Is this better?
- Why?



Comparing A Pair of Proportions

Daily Record

NEWS POLITICS FOOTBALL SPORT TV & CELEBS LIFE & STYLE MORE

BUY A PAPER ROAD RECORD PROPERTY MARKETPLACE COMPETITIONS HOROSCOPES DEALS FUNERAL NOTICES JOBS CROSSWORDS ADVERTISE WITH US NEWSLETTER SIGNUP

News › Health › Cancer

Bacon, ham and sausages have the same cancer risk as cigarettes warn experts

Processed meats are now on World Health Organisation list of substances as carcinogenic as tobacco

NEWS By Jane Kirby

12:02, 23 OCT 2015 UPDATED 20:27, 23 OCT 2015

Bookmark Comments 1

Comparing A Pair of Proportions

- Came from the International Agency for Research in Cancer putting red meat in same category as cigarettes and asbestos (a ‘Group 1 carcinogen’)
- “50g of processed meat associated with increased risk of bowel cancer of **18%**”
- What does that mean?

Comparing a Pair of Proportions

- The ”18%” increase is a relative risk.

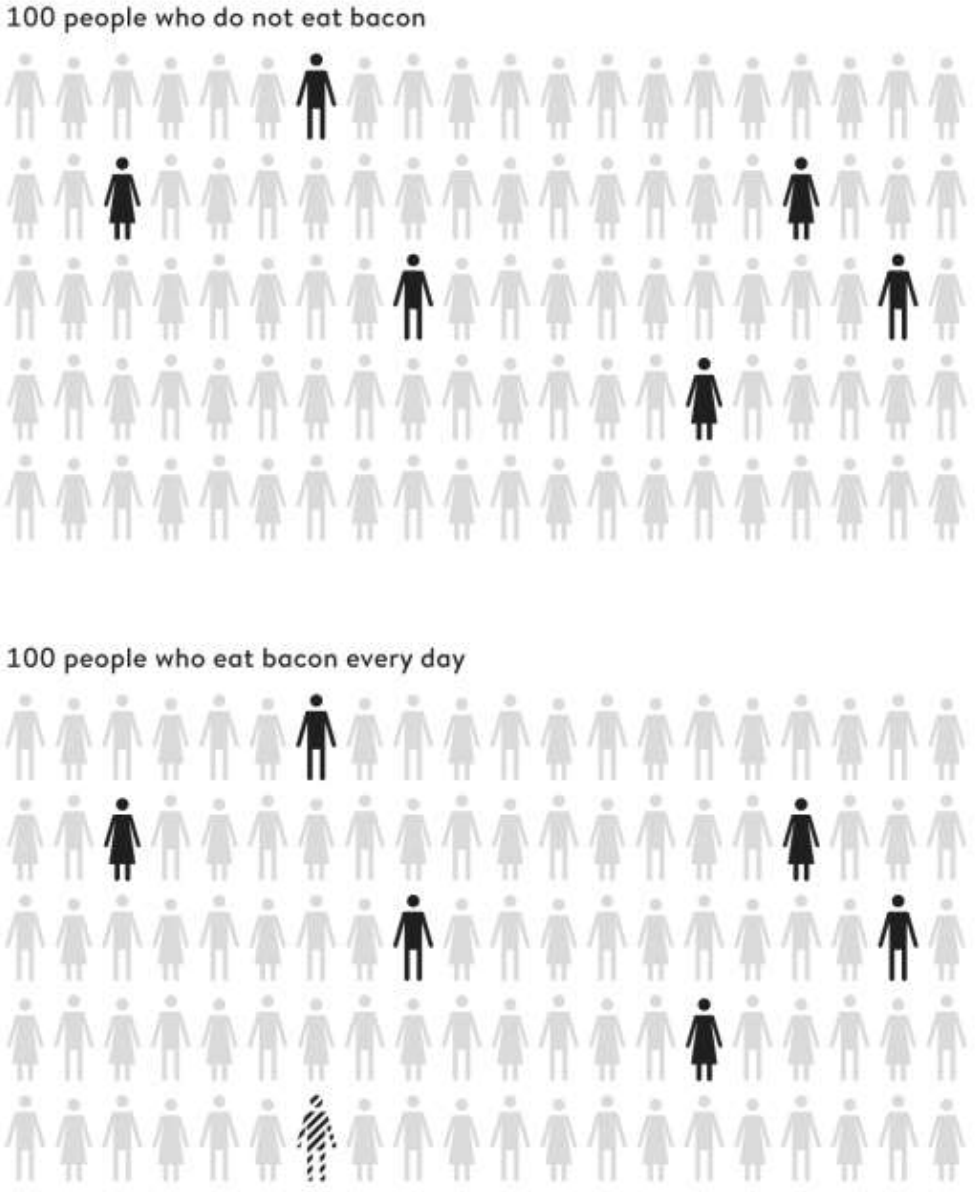
Definition

Relative risk: If the absolute risk among people who are exposed to something of interest is *p%*, and the absolute risk among people who are not exposed is *q%*, then the relative risk is *p/q*.

- What is the absolute risk? That is, how often do people eating meat get bowel cancer?
- ~7.1 out of 100 people get bowel cancer if they eat red meat (p=7.1%).
- ~6 out of 100 people get bowel cancer if they don’t eat red meat (q = 6.1%).
- (7.1/6) * 100 = 118.33% => 18% above the expected amount.

Comparing a Pair of Proportions

- But the e /100 and 7.1/100)
- There wa e risk from the expected



Definition	
Expected frequenci according to an assu	o occur in the future,

Comparing a Pair of Proportions

Method	Non-bacon eaters	Daily bacon eaters
Event rate	6%	7%
Expected frequency	6 out of 100	7 out of 100
	1 in 16	1 in 14
Odds	6/94	7/93
Comparative measures		
Absolute risk difference	1%, or 1 out of 100	
Relative risk	1.18, or an 18% increase	
'Number Needed to Treat'	100	
Odds ratio	(7/93) / (6/94) = 1.18	

Comparing a Pair of Proportions

Definition

Odds: For the bacon eaters the odds of bowel cancer at $^{p}/_{1-p}$ or (7/100)/(93/100) = 7/93,

	1 in 16	1 in 14
Odds	6/94	7/93

Definition

Odds ratio: Comparing two odds, like between bacon eaters and non-bacon eaters: $\frac{^{p}/_{1-p}}{^{q}/_{1-q}}$. *Q: When does this differ from the relative risk?*

Odds ratio	(7/93) / (6/94) = 1.18

Contingency Table

	Cancer	No Cancer	
No Red Meat	**6**	**94**	**100**

- Odds are (6/100)/(94/100) = 0.064 i.e. *a ratio of expected frequencies*.
- If we pretend that Cancer is Heads and No Cancer is Tails, we are checking how often we get cancer with the odds.
- We're unlikely to get heads (cancer) because of 6% expected frequency.

Contingency Table

	Cancer	No Cancer	
Red Meat	**7**	**93**	**100**
No Red Meat	**6**	**94**	**100**
	13	**187**	**200**

- Odds ratios tell us about how expected frequencies across one variable change due to a second variable.

Odds Ratio (OR) vs Relative Risk

OR = 1	**Relative Risk = 1**	**Implies that exposure doesn't affect *odds* of outcome (e.g. cancer).**
OR>1	**Relative Risk > 1**	**Implies exposure is associated with increased odds of the outcome.**
OR<1	**Relative Risk < 1**	**Implies exposure associated with lower odds of outcome.**

OR and Relative Risk are close when the risk/chance of heads is low.

OR = $\frac{p/1-p}{q/1-q}$. RR = $\frac{p}{q}$. We can see why.

Odds Ratio vs Relative Risk

	Statins	No Statins	
Muscle Pain	**87**	**85**	**172**
No Muscle Pain	**13**	**15**	**28**
	100	**100**	**200**

- OR = ??, so odds are slightly higher, but:
- Relative Risk = ??

Odds Ratio

Odds Ratio (OR) vs Relative Risk

- From cancer.ca:

> Smokers are about 20 times more likely to develop lung cancer than non-smokers. The longer a person smokes and the more cigarettes they smoke each day, the more their risk increases. Smokers are also at a higher risk if they're exposed to radon or certain chemicals in their home or workplace and continue to smoke.

- What kind of comparison is this?

Odds Ratio (OR) vs Relative Risk

Odds Ratio (OR) vs Relative Risk

- Calculate the odds ratio of smoking against non-smoking and relative risk for this case.