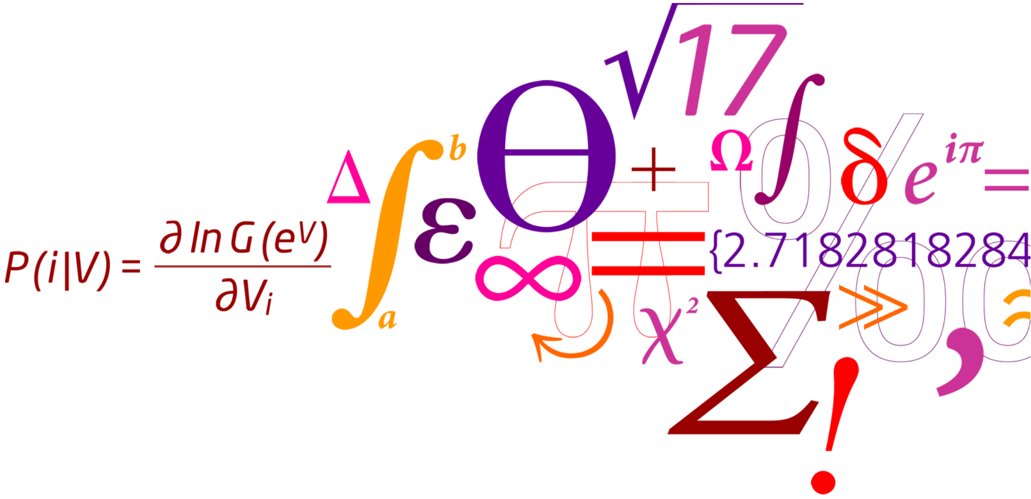
42588 – Data and data science

Week 1 – Introduction

31rst of January 2024



DTU Management Engineering
Department of Management Engineering

Today’s program

- The course
  - Purpose, learning objectives and content
  - Evaluation
  - Teachers and teaching philosophy
  - Practical stuff

- Data science and questions
  - PPDAC cycle

- Why do we look at data?

The course – purpose

- This course focuses on
  - how to collect and use data to answer questions about the world, and
  - the challenges that arise when we try to answer such questions, and
  - how to use data to make better decisions.

The course - learning objectives

A student who has met the objectives of the course will be able to:

- discuss the **relationship** between **problem**, **plan** and **data** to **answer questions**
- describe **data types** and ways to **collect data**
- discuss **data governance** and produce a **data management plan**
- discuss and apply **data processing tools** for imputation, sampling and weighting
- **summarise data** to answer relevant **questions**
- conduct **exploratory data analysis** and **visualisation**
- discuss differences between **various approaches** in **statistics/data science**
- apply **common techniques** to **analyse continuous data** and **interpret** the results
- apply **common techniques** to **analyse discrete data** and **interpret** the results
- present **focused** and **concise data analyses**

The course plan

Week	Date	Subject/Lecture	Literature	Exercises	Teachers
1	31/1	Introduction + questions and data	AoS chap. 3	Form groups + week 1 exercise	Stefan
2	7/2	Basics on data and variables	AoS chap. 1-2 (+ OM 1)	Project 1 – start	Stefan/Guest from Genmab
3	14/2	Surveys + data types + experimental data	Paper 1 (+ OM 2-5)	Project 1 – work	Sonja / Stefan
4	21/2	Governance + causality	Paper 2 + AoS chap. 4 (+ OM 6)	Project 1 – deadline	Hjalmar / Stefan
5	28/2	More on data, e.g. real-time data, online data	Paper 3 (+ OM 7-10)	Discuss data for project 2	Guido/ Stefan
6	6/3	Visualisation	Chap. 1,5,6,7,10,23, 24,29 in Wilke + (AM 1-2)	Integrated exercises + work on project 2	Mads
7	13/3	Spatial data	Chap. 1,14 in Gimonds	Week 7 exercises + work on project 2	Mads / Guest from Niras
8	20/3	Imputation/weighting/presentation proj. 2	Paper 4	First deadline of project 2 + Week 8 exercises	Mads
9	3/4	Data analytics I	ISL ch. 3 + paper 5	Work on project 3a	Stefan
10	10/4	Data analytics II	ISL ch. 6	Work on project 3a	Stefan
11	17/4	Data analytics III	ISL ch. 4	Work on project 3b	Stefan
12	24/4	Data analytics IV	TBD	Work on project 3b	Stefan
13	1/5	Summary and perspective	Paper 6	Project 3 – deadline	Stefan

Literature

- Course literature
  - Course notes
  - Chapters from Spiegelhalter (2019) The Art of Statistics – Learning from Data,
  - Chapters from James et al. (2015) Introduction to statistical learning,
  - Papers and book chapters

- Supplementary readings (optional):
  - Various papers
  - Chapters from Washington et al. (2011) *Statistical and Econometric Methods for Transportation Data Analysis*, CRC Press, Taylor & Francis Group

Projects (3-5 students) and oral exam (individually)

- Project 1
  - describe (fictitious) research question + data collection + data management plan
  - Hand-in 27/2 + feedback

- Project 2
  - Find data and work with processing + descriptives + visualisation
  - Presentation in week 8 + feedback

- Project 3
  - use data analytics to answer research question on data
  - Hand-in 30/4 + feedback

- Oral exam
  - Each student will have an individual exam mainly covering the three projects

The course - evaluation

- You will have to work on three projects
  - you will get feedback on each.
  - You will get feedback on the projects together with a judgement on a scale:
    - 0 - weak (not passed)
    - 1 - passed
    - 2 - OK
    - 3 - Good
    - 4 - Very good

- You will get a final overall grade on the 7-point scale based on the oral exam and the three projects.

The course - teachers

- Stefan Mabit, associate professor, room 195, building 358 (mail: smab@dtu.dk)
  - Research on statistical modelling of mobility, transport, behaviour, demand, e.g. mode choice, valuation, car demand, electrification.

- Mads Paulsen, assistant professor, room 111A, building 358 (mail: madsp@dtu.dk)
  - Research on flow, simulation of transport, route choice, assignment, and biking.

- Other lecturers
  - Sonja Haustein
  - Guido Cantelmo
  - Hjalmar Christiansen
  - Two guests

The course - participants

- Round of presentation (2 min per question)
  - What do you find interesting about your study (DSM)?
  - What has been the best course during your studies so far and why?
  - What do you like about studying at DTU?

The course – teaching philosophy

- Learning happens best through discussion/dialog.

- You are a small class, so feel free to ask any question during lectures/exercises.

- Teachers will help as much as possible but it is **you** who **have to do the work! (~9 hours per week)**

- **Remember the 3P: Prepare, participate, practice**

The course – practical stuff

- Recommended prerequisites
  - Basic statistics and introduction to machine learning

- Evaluation of the teaching
  - Ongoing feedback following the lectures
  - Midterm evaluation
  - Final official evaluation

- Teaching Wednesdays 13-17 in room 043/358
  - Lectures + exercises 13 - 16
  - Work on exercises + projects 15-17

What can you do to contribute to a good lecture session?

- Please arrive without disturbing if you are late.
- Please make sure that your mobile phone is switched off (or at least is in silent mode) and not visible during lectures.
- Your active participation in the session will benefit both your own learning outcome and that of others – including the lecturer!

Two examples of asking questions

- **Example 1** – chocolate and Nobel prizes

- The question is

Is there an effect from national chocolate consumption on the number of Nobel laureates from a country?

- The researchers therefore collected country-wise data on
  - National chocolate consumptions (kg/capita/year)
  - Number of Nobel Laureates (#/10 million capita)

Chocolate and Nobel prizes example

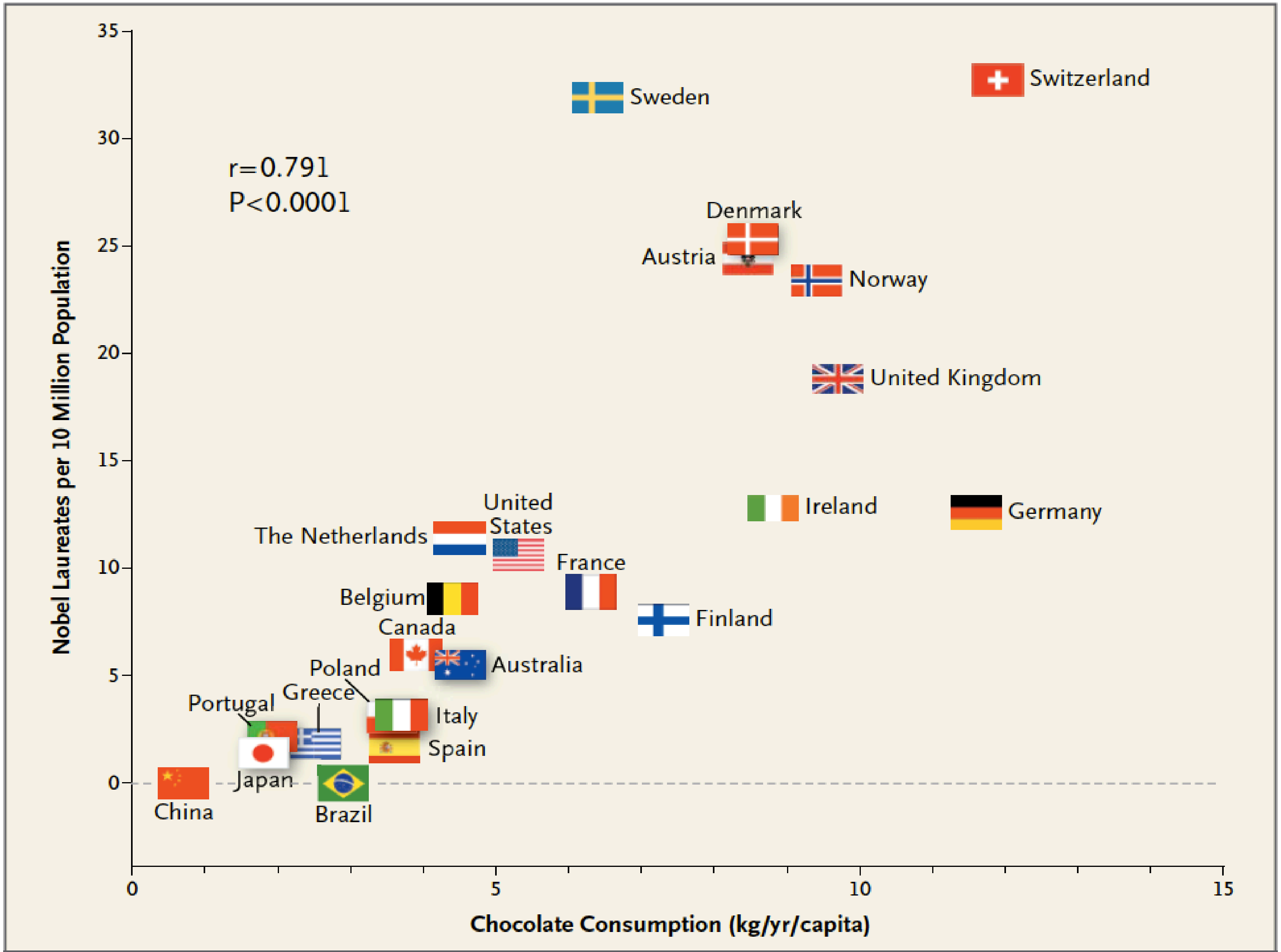


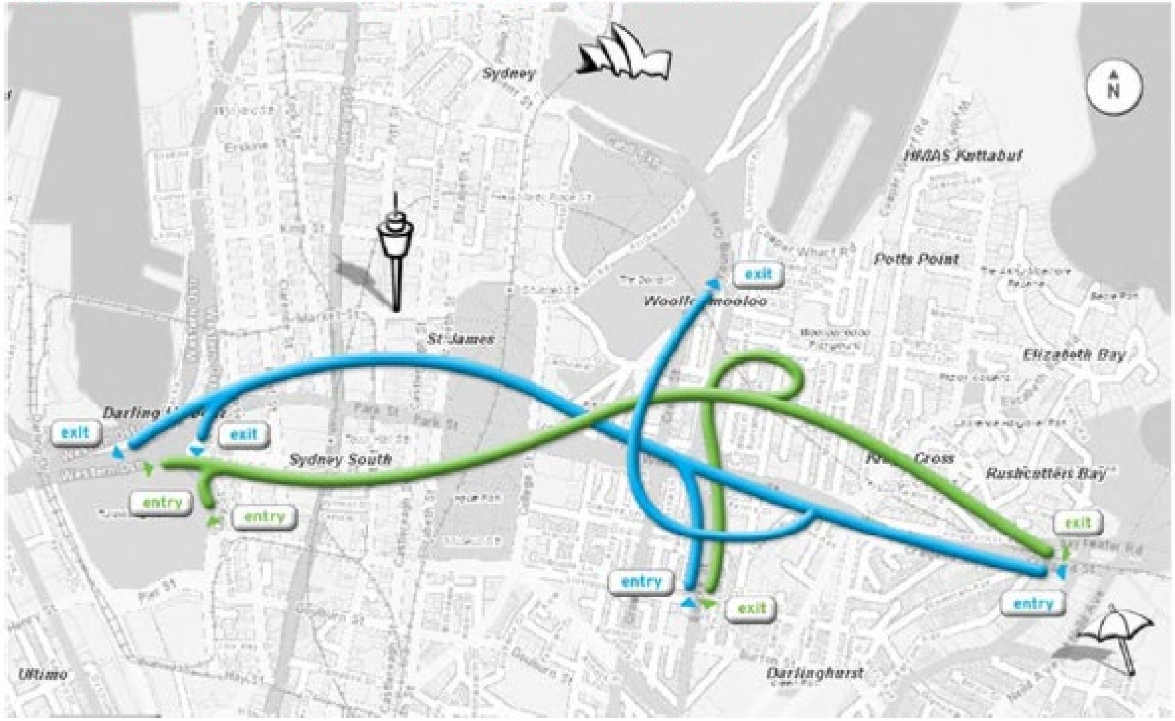
Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel

Example 2 – Sydney Cross City Tunnel

- In Sydney in 2005 a tunnel was completed below downtown at the cost of AUD 680 Million.

- An important questions is then: **What is the expected number of daily users?**

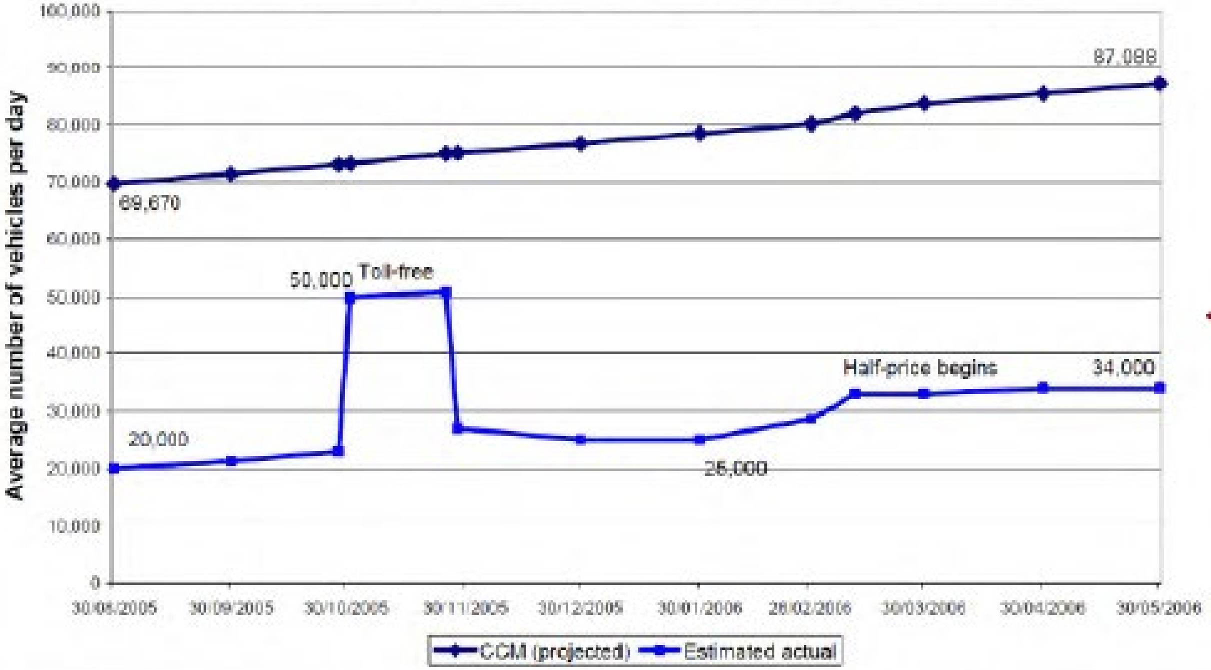
Sydney Cross City Tunnel (completed 2005, costs: $680mln)



When reality does not match the forecasts

Exhibit 2.5: Estimated actual patronage compared to CCM's projections - nine months

CCM predicted 87,088 vehicles/day by June 2006, and assumed 80% of that when the tunnel opened, which we calculate as around 70,000 vehicles/day. They then predict 1.3% p.a. growth to 99,967 by June 2016.



Source: Audit Office research. Information on CCM projected patronage obtained from RTA documents. Estimated actual patronage based on research plus CCM statements where available.

Bankrupt after 16 months

Break

Overall focus

- How to use **data** to
  - **answer questions**
  - **improve decision making**

- Both of these are aspects of **statistics** that are now included in **data science**

- Data science
  - Statistical science
  - Data literacy
  - Knowledge of subject matter
  - Data management
  - Programming and algorithm development

Data science, statistics and questions

- Data science includes statistical methods that can help
  - address questions based on data
  - judge whether results are the outcome of randomness or something systematic
  - make predictions

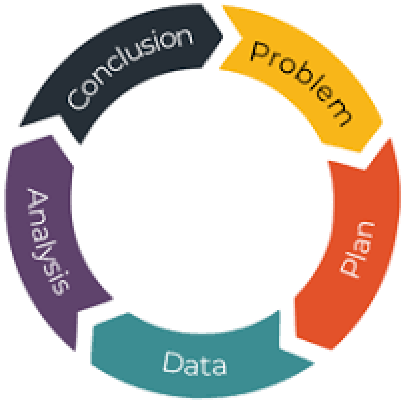
- **Statistical analysis** may be used to **investigate relationships,** e.g.
  - Are men more involved in car accidents than women?
  - Does an extra highway lane relieve congestion?

- **Statistical analyses** can also be used for **prediction** e.g.
  - What happens to the number of public transport passengers if prices drops with 40 %?
  - How does higher income affect car ownership/use?

The PPDAC cycle

- A useful way to organise a project/question related to data is the PPDAC cycle

    - Problem
    - Plan
    - Data
    - Analysis
    - Conclusion



- From https://dataschools.education/about-data-literacy/ppdac-the-data-problem-solving-cycle/

Problem

- Understanding and defining the problem/question.

- What to do to answer this problem?

- Identify the right question to solve
  - How much or how many?
  - Which category does this belong to?
  - Is this unusual?
  - Which is the best option to choose?

- Question:
  - How many trees are there in the world?

  - Discuss for 2 min with your neighbour how you would answer this question. Feel free to give an answer, ☺.

Plan

- Linking the question to data
  - What to measure?
  - How to measure?
  - When to measure?

- Putting it together - what is the study design?
  - What data are needed?
  - Where will the data come from?
  - Do you have access to some data or do you need new data to be collected?
  - Will there be sufficient data for robust analysis?
  - Where and how to store the data?
  - Ethical considerations

Exercise – problems and data

- Discuss with your neighbour (5 min).
  - What could be an interesting question relevant for the Danish society that you think could be answered with appropriate data?

- Consider as many questions below as possible
  - What data are needed?
  - Where will the data come from?
  - Do you have access to some data or do you need new data to be collected?
  - Will there be sufficient data for robust analysis?
  - Where and how to store the data?
  - Ethical considerations

Data

- Collect
  - Consider pilots for surveys
  - Assure quality in data collection
  - Store the data correctly and securely

- Management
  - Understand the data
  - Check the quality
  - Clean the data
  - Document the data

Data collection – example I

- Please take a piece of paper

- Please write
  - your height (cm)
  - your hand span of right hand (cm)
  - your gender
  - your commute distance to DTU in km

- Please return the paper with the data

Analysis

- Summary statistics
- Tables, graphs
- Visualisation
- (Hypothesis generation)
- Inference
- Modelling
- Prediction
- Validation

Conclusion and communication

- How does data answer the original question?

- Does it address the question partly or related questions that have appeared? Are there any aspects that have not been addressed?

- Discuss the results related to other evidence as well as the robustness of the results

- What are the conclusions?

- Actions – what should happen based on the conclusions?
- New ideas –what could be done differently next time?

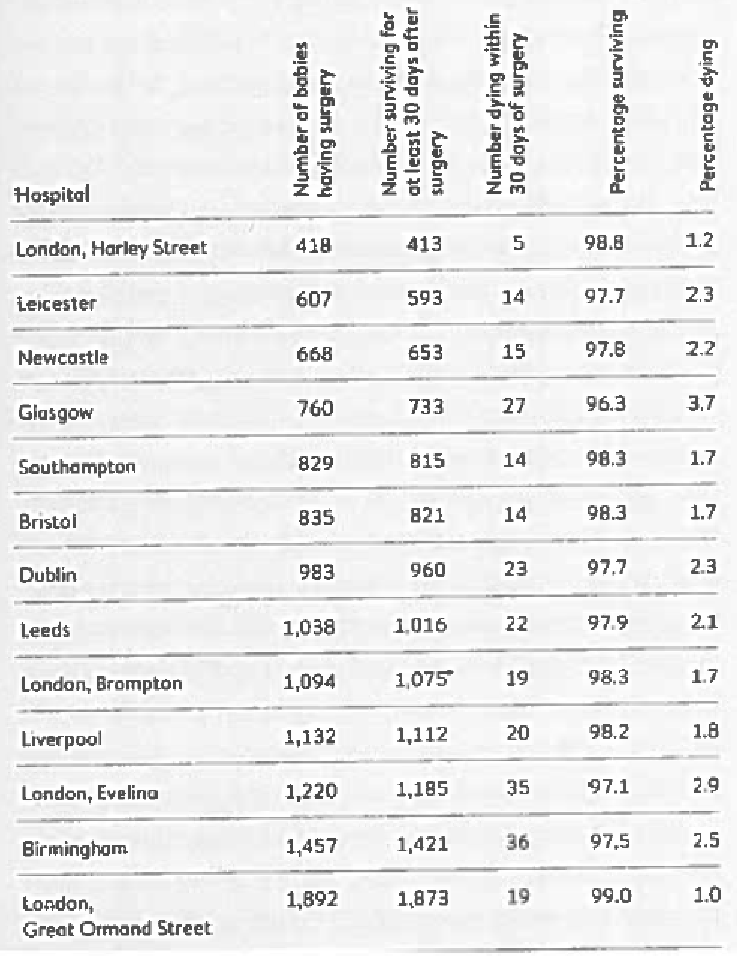
- Communication – what should be communicated, to whom, and how?
  - This is somewhat overlooked in statistics/data science courses

Communication

- In tables without a natural ordering, the ordering may influence interpretations – example from Spiegelhalter (2019).

- Here he could have ordered by survival rate. However, this would emphasize minor differences in rates that could be due to variation.

- So the choice was to order by size (number of surgeries)



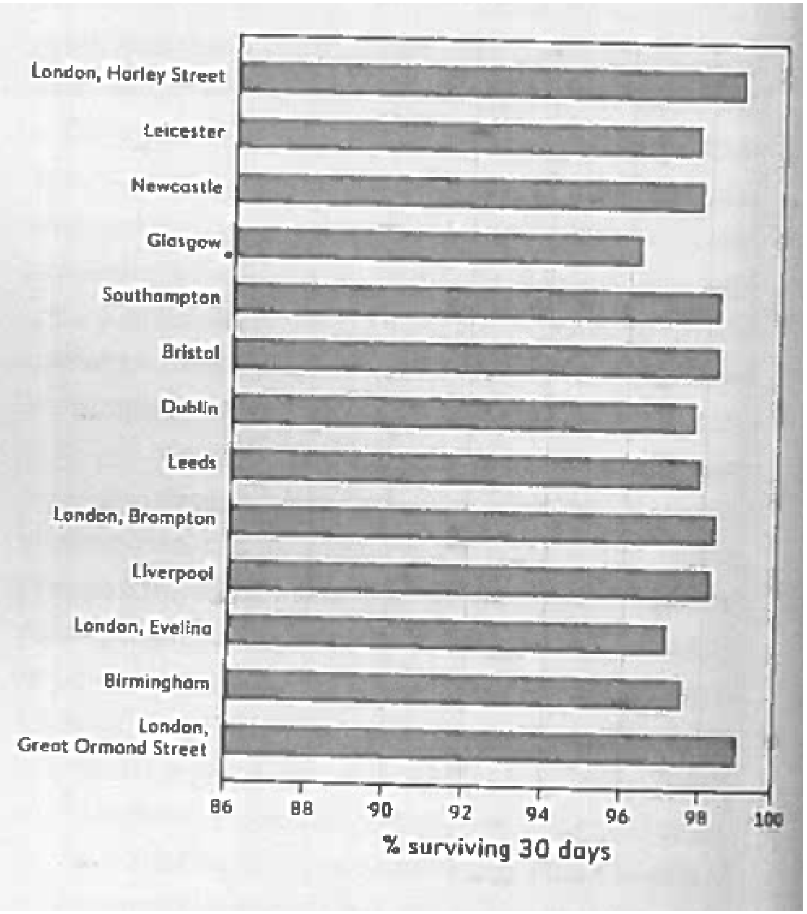
Hospital	Number of babies having surgery	Number surviving for at least 30 days after surgery	Number dying within 30 days of surgery	Percentage surviving	Percentage dying
London, Harley Street	418	413	5	98.8	1.2
Leicester	607	593	14	97.7	2.3
Newcastle	668	653	15	97.8	2.2
Glasgow	760	733	27	96.3	3.7
Southampton	829	815	14	98.3	1.7
Bristol	835	821	14	98.3	1.7
Dublin	983	960	23	97.7	2.3
Leeds	1,038	1,016	22	97.9	2.1
London, Brompton	1,094	1,075	19	98.3	1.7
Liverpool	1,132	1,112	20	98.2	1.8
London, Evelina	1,220	1,185	35	97.1	2.9
Birmingham	1,457	1,421	36	97.5	2.5
London, Great Ormond Street	1,892	1,873	19	99.0	1.0

Communication

- In diagrams/charts, the decision where to start the axis may influence interpretations – example from Spiegelhalter (2019).

- Consider if we had started at 0?

- Or, if we had started at 96?



Data science and ethics

- Statistics should bring
  - Clarity
  - Insight

- Statistics can be misused to promote specific opinions or attract attention

> “The only statistics you can trust are those you have falsified yourself”
>
> W. Churchill (maybe)

Break

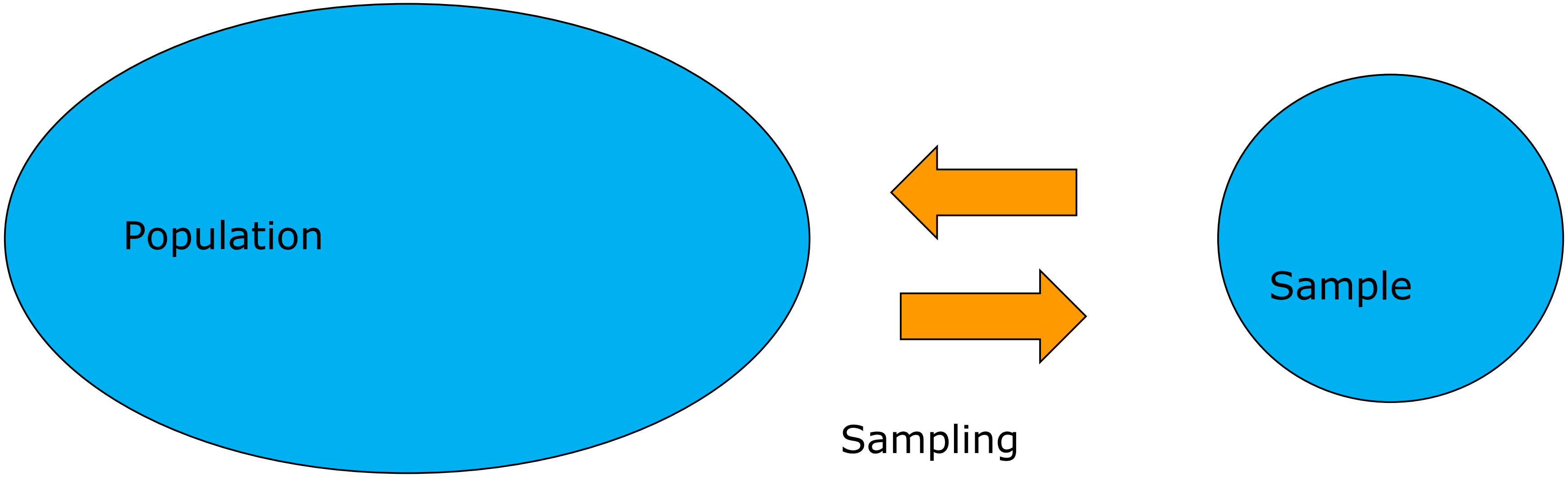
Why do we look at data?

- We might look at data to
  - Answer questions
  - Solve problems
  - Get insights
  - Make predictions
  - Monitor something

- But every time we answer a question, we need to consider the population of interest, i.e. consider the context of the question.

- Data is often a sample from some population of interest and **the relationship** between **sample** and **population** is **fundamental** for data science.

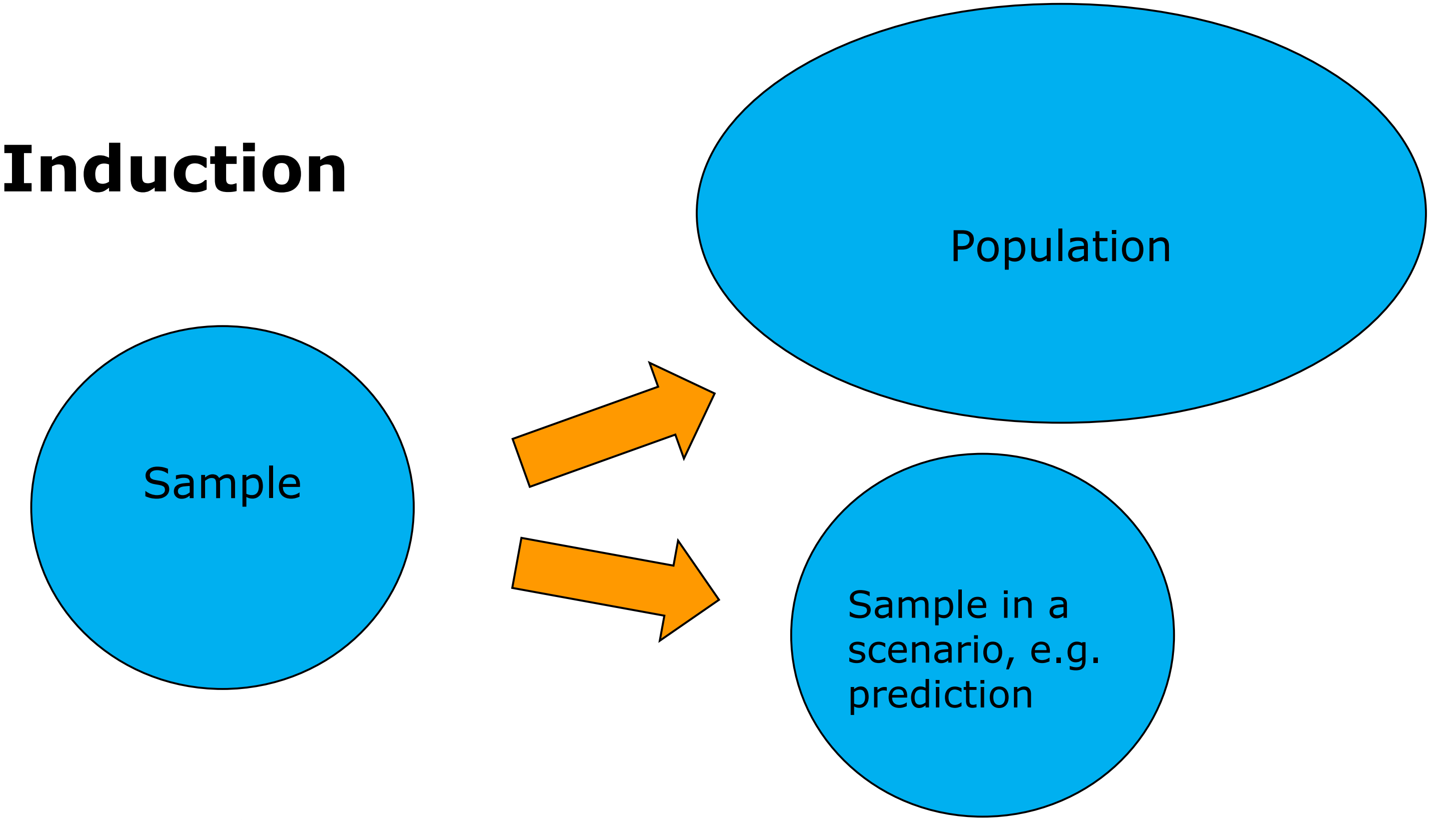
Statistics and knowledge



- **Deduction** is when we use knowledge about the population to describe the sample

- **Induction** is when we use knowledge about the sample to describe the population

Induction



- To get information about a population, we use a sample/data through
  - Descriptive statistics
  - Inference
  - Prediction

Sampling - example I

- How many of you like statistics?

- Discuss with you neighbour whether our class average is a good estimate of how many people “like statistics” in the Danish population (2 min).

Learning from data

- (1) Data -> (2) Sample -> (3) Study population -> (4) Target population
- Suppose we have a question, we would like to answer
- First, we need to find out what our target population is
- Second, we need to consider what part of the target population that could potentially be investigated
- Third, we need to consider how to sample from the study population
- Finally, we need to consider what data to collect from the sample to address the question

Learning from data

- 1) Data -> (2) Sample

- Here we need to be sure that what we measure in data for the sample are
  - Reliable, i.e. having low variability in repeated studies,
  - Valid, i.e. being a measure of what we actually want to measure, and hence avoid systematic biases

- Reliability
  - Answers to a survey should not depend on the mood of the respondent (unless part of the study) or the style of the interviewer or the framing of the survey

- Validity can be difficult
  - consider measurements of happiness or intelligence,
  - but also technical measurements, e.g. pollution from vehicles

More on reliability and validity

- Reliability is a big issue in social sciences, psychology, and bio/medical sciences where researchers have tried to reproduce earlier results.

- So when you redo measurements on the same or a similar sample, you expect similar results within uncertainty, and this does not seem to be the case.

- There are many papers on this. However, a glimpse at this is given in the paper “Statistik, politik og kunsten at reproducere”.

- Validity can also be problematic and includes many aspects like
  - How do we measure an aspect?
  - Surveys can be framed, e.g. Ryan Air had a survey where 92% of respondents were satisfied. A minor problem was that the only options for answering were “Excellent, very good, good, fair, OK”

Learning from data

- (2) Sample -> (3) Study population

- The question is whether the results we obtain from the sample represents results we would get from the population, i.e. do the study have **internal validity**.

- Two ways to obtain this
  - Random sampling, i.e. choosing element at random from the study population to include in the sample
  - Using the full population as sample, e.g. if you have a small population, or if you have access to the full population

- Random sampling may seem simple but can go wrong, e.g.
  - US presidential election 1936, Roosevelt vs. Landon, 61% vs. 37%
  - UK election 2015 predicted a Labour victory, Conservatives won
  - Vietnam war draft (1969), 26 days in December vs. 14 days in January

Types of sampling

- Random
  - You try to draw at random for study population

- Exogenous
  - You divide the study population into groups (strata) and try to draw at random within each strata, e.g. the sampling of DTUs travel behaviour survey (TU) stratifies on gender, age and geography.
  - If done correctly, it can reduce sampling costs

- Endogenous
  - You divide the study population into strata based on the endogenous variable of interest, e.g. to get information about electric vehicles we might sample 50% non-EV owners and 50% EV owners. This will create issues in estimation procedures that sometimes can be solved.

- Miscellaneous
  - All other sampling procedures.

Learning from data

- (3) Study population -> (4) Target population

- The question is whether we can actually study the target population.

- In some studies we can only study animals but humans are the target population.

- In studies where the target population is everybody living in Denmark, there could be issues with kids, elderly, people in prison.

- A main problem here is if the study population is now and the target population is the future or a scenario. This is problematic for many applications of statistics/data science, e.g. machine learning may generate brilliant models for a study population but fail to forecast changes a year ahead because this link is not considered.

- Fulfilling this is known as having **external validity.**

Sampling - example II

- Please indicate on the blackboard how many kids are in your family (your siblings + you).

- What is the class average?

- The average in Denmark (www.statistikbanken.dk) is
  - 1.77 kids per woman in 2000
  - 2.09 kids per family in 2022

- Discuss with you neighbour whether our class average is a good estimate of how many kids families have in Denmark (2 min).

What if we have the full population?

- On many occasions we are not in the stylised setting with a sample from an actual population.

- We can think of three types of populations
  - An actual population, e.g. a group of individuals
  - A virtual population, e.g. blood samples
  - A metaphorical population, e.g. if the sample is all countries in the world or everybody living in Denmark, we can think of the population as potential other versions of the world or Denmark.

- For all these different types of population, we can use the same methods from data science/statistics. But it is good knowledge to know what kind of population we are dealing with.

Feedback

- Final questions (2 min)

  1. What was the most interesting you learned during the lecture?

  2. What is your most important unanswered question based on the lecture?

Week 1 exercise

- Your should work together in groups. On the projects, you should be 3-5 students. But for this exercise you are welcome to be less if you prefer to work in smaller groups.

For next time

- Today is based on
  - The Art of Statistics chap. 3
- Other material
  - Statistik, politik og kunsten at reproducere (uploades til Learn)
- Read
  - The Art of Statistics chap. 1-2
  - Notes
- Form groups
- Work on the week 1 exercise.