Scientific methods and data ethics

Josef Wilzén and Hector Rodriguez-Deniz

Division of Statistics and Machine Learning
Dept. of Computer and Information Science
Linköping University

2023-08-23

Outline

Summary

What is science?

Models

Scientific Revolutions

Ethics in Science and Statistics

Summary

Take home message:

- Be critical! Question it!
  - data, plots, graphs, tables
- Science is hard
- Correlation does not imply causation
- Who is behind the results? What is the agenda?
  - money, power, prestige, reputation
- Don't do bad thing with data: Ethics matter
  - "With big data comes big responsibility"

Intro

This lecture

- science = “all science”, not just “natural science”
- A smorgasbord of different topics

Intro

Apartment prices in Linköping

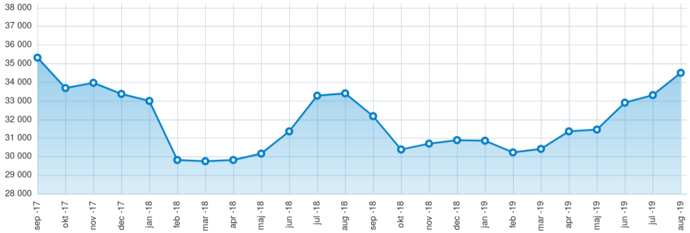

Figure: From Svensk Mäklarstatistik

Intro

Apartment prices in Linköping

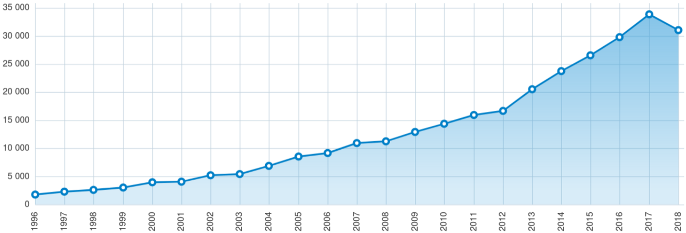

Figure: From Svensk Mäklarstatistik

Example

Consider the observed dataset

$$x = \begin{pmatrix} 0 & 1 & 2 & 3 \end{pmatrix} \qquad y = \begin{pmatrix} 2 & 3 & 4 & 5 \end{pmatrix}$$

Problem: We want to understand the relation between x and y. The assumption $y = f(x)$ is given. Which function f should be used? Why?

→ Science to the rescue!

What is science?

- "Large amount of relatively secure knowledge"
- Results – description of facts or explanation of dependencies
  - Results are often published in scientific journals or books
  - Peer review
- A process – the methods and activities that lead to the results
  - E.g. Experiments, computations, theorize

What is science?

Science vs. pseudoscience

- Methods are established within the scientific community
- Scientific methods have clear and explicit rules and procedures
- Replication is important: details matter
  - Methods that are arbitrary and cannot be repeated are not scientific
- Science should be as objective as possible
  - Researcher bias should be reduced
- Scientific knowledge is created within the scientific community
- All new results should be related to existing knowledge in the field

Categories

- Nomothetic (general) studies
  - General laws
- Idiographic (specific) studies
  - Describes specific objects and processes
  - What happened at the Battle of Hastings?
- Formal science
  - The study of constructed and formal systems
  - Logic, mathematics, statistics
- Empirical science
  - The study of objects and processes in the “real world”
  - Medicine, history, economics

Scientific explanations

- Deductive explanations
  - Based on a number of premises
  - With help of the premises, conclusions are deduced with the help of logic
  - If the premises are true, then the conclusions are also true
- Probabilistic explanations
  - No general law. Premises that have a (high) probability to be true (or to happen)
  - The "probabilistic conclusions" are not true in a formal way, but may be probable
  - Probability theory and statistical inference are used formalize the probabilistic explanations

Philosophy of Statistics

- Probability:
  - Relative frequencies in series of events
  - Degree of belief, epistemic
- Statistical inference
  - Frequentist statistics
  - Bayesian statistics
- Philosophy of Statistics: https://plato.stanford.edu/entries/statistics/

Models

- A model can be compared with a map
- What kind of map is important for the following persons?
  - taxi driver
  - orienteerer
  - epidemiologist

Models

- "All models are wrong. Some are useful." - George E.P. Box
- Science often speaks of models
- Model: a representation of a process or a system
  - Important features are a part of it (often emphasized)
  - Other features are not included
- Historically mechanical models important
- Nowadays: Theoretical or mathematical models are much more important.

Models

- Purpose of scientific models:
  - Understand, define, quantify, visualize or simulate a process or a system
  - Common approach when working with complex problems
  - Calculations and predictions becomes easier
- A theory is always a model: makes a phenomenon understandable
  - All models are not theories, e.g. mechanical model
- Causal models
  - Describes the causal mechanisms of a system.
  - Causal diagram is a directed graph that displays causal relationships
  - Confounding factor: important in statistics
    - Correlation does not imply causation

Scientific Revolutions

- Research is often a cumulative process
  - Continuous revision of old knowledge
- Sometimes there are revolutions: Old theories are rejected and replaced with new ones
  - Chemical revolution: Lavoisier
  - Scientific theory of evolution: Darwin
  - Theory of relativity: Einstein
  - Quantum mechanics
  - Convolutional neural network and deep learning within image classification (2012)

Scientific Revolutions

- Thomas Kuhn: “The structure of Scientific Revolutions”, 1962
  - Normal science →revolution and crisis →Normal science
- Paradigm = central hypotheses
  - researchers are laying a puzzle
  - After a while: to many pieces do not fit
  - A revolution happens when the central hypotheses are rejected: A new puzzle start
- It can be hard in practice to define and observe scientific revolutions.
  - When does a hypothesis become a paradigm?
- Paradigms can be subjects of “religious” belief

Ethics in science

- Ethics or moral philosophy:
  - Deals with what is right or wrong
  - How to act?
- Research ethics: How to handle moral issues that arise during or as a result of research activities

Ethics in science

Bad examples from history:

- Nazi human experimentation
  - A large number of prisoners were forced to participate, the experiments typically lead to death, trauma, permanent disability etc.
  - Lead to the Nuremberg Code after the Nuremberg trials
- Tuskegee syphilis experiment (US, 1932-1972)
  - African-American men were used to see the effect of untreated syphilis infection, without consent of the participants
- Vipeholm experiments Vipeholm experiments (Sweden, 1945-1955):
  - Intellectually disabled were fed with sweets in order to provoke dental caries, the aim was to determine the role of of carbohydrates

Ethics in science

- A researcher's work is regulated by rules and regulations
- Researcher's own ethical responsibility that
  - Research has good quality
  - Is morally acceptable
- Professional Ethics
  - Research activity is driven by a number of implicit and explicit norms that decide what good science is. Ex. Helsinki Declaration
  - Follow national and local rules: issues like concerning discrimination, harassment and humiliation, gifts to the researcher
  - Field specific codes of ethics: Ethical code

Ethics in Statistics

- American Statistical Association (USA): “Ethical guidelines for statistical practice”
- Royal Statistical Society (UK): ”Code of conduct”
- International Statistical Institute: ”Declaration of Professional Ethics”
- Swedish Statistical Society: “Svenska statistikfrämjandets etiska kod för statistiker och statistisk verksamhet”

Ethics in Statistics

“Ethical Guidelines for Statistical Practice”

- Professional Integrity and Accountability
- Integrity of data and methods
- Responsibilities to Science/Public/Funder/Client
- Responsibilities to Research Subjects
- Responsibilities to Research Team Colleagues
- Responsibilities to Other Statisticians or Statistics Practitioner
- Responsibilities Regarding Allegations of Misconduct
- Responsibilities of Employers, Including Organizations, Individuals, Attorneys, or Other Clients Employing Statistical Practitioners

Ethics in Statistics

Handling of data has a special role in statistics and data science.

- Special care must be taken when data is collected, stored and used for statistics and machine learning.
- General Data Protection Regulation (GDPR)

GDPR

Look here and here.

- Scope: “The General Data Protection Regulation exists to protect individuals’ fundamental rights and freedoms, in particular their right to protection of their personal data.”
- GDPR:
  - EU law on data protection
  - Regulate the use of personal data
  - Called “Dataskyddsförordningen” in Sweden

Ethics in Big Data

“In today’s most common digital business model, consumers pay for ’free’ products with their personal data.”

from: Big data, artificial intelligence, machine learning and data protection

Ethics in Big Data

A few starting principles

1. Ownership: Individuals own their own data
2. Transaction Transparency: The use of the data should be transparent
3. Consent: informed and explicitly expressed consent is needed to use the data
4. Privacy
5. Currency: Individuals should be aware of financial transactions resulting from the use of their personal data and the scale of these transactions
6. Openness: Aggregate data sets should be freely available

Ethics in Big Data

5 Principles for Big Data Ethics from “Towards Data Science”:

- Private customer data and identity should remain private: private data obtained from a person with their consent should not be exposed for use by other businesses or individuals with any traces to their identity.
- Shared private information should be treated confidentially: Third party companies share sensitive data — medical, financial or locational — and need to have restrictions on whether and how that information can be shared further.
- Customers should have a transparent view of how our data is being used or sold, and the ability to manage the flow of their private information across massive, third-party analytical systems.

Ref [here]

Ethics in Big Data

5 Principles for Big Data Ethics from "Towards Data Science":

- Big Data should not interfere with human will
- Big Data should not institutionalize unfair biases like racism or sexism. Machine learning algorithms can absorb unconscious biases in a population and amplify them via training samples.
  - Example: Microsoft's Twitter chatbot "Tay": the robot began releasing racist and sexually-charged messages

Ref [here]

Facebook–Cambridge Analytica data scandal

- Facebook–Cambridge Analytica data scandal: Big political scandal in early 2018
- The company Cambridge Analytica had collected personal data from millions of peoples' Facebook profiles
  - without consent
  - used it for political advertising purposes.
- Whistle-blower: Christopher Wylie

Facebook–Cambridge Analytica data scandal

- Aleksandr Kogan researcher at Cambridge University created an app
    - "This Is Your Digital Life"
- Several hundred thousands of Facebook users gave consent to be part of the survey only for academic use.
- Facebook's design allowed data to be collected from the social network of the participants
    - This allowed Cambridge Analytica to collect data from up to 87 million users

Facebook–Cambridge Analytica data scandal

Cambridge Analytica used the data to

- Create psychographic profiles of Facebook users
- Profiles used to choose advertisement that most effectively persuade specific groups of persons
- Used in political campaigns with the aim to affect elections, examples
  - 2016 United States presidential election
  - 2016 United Kingdom European Union membership referendum
  - Many other countries and elections

Discussion

- What is ethical to do with user data on social media platforms?
- "Personal data as gold": companies using data as main source of profit, what is good ethics in such business? Do the users understand what their data are used for?
- What responsibilities does a machine learner or data scientist working for a social media company have?
  - What to do if your boss asks you to do something that maybe feels wrong? Eg. collect or analyze personal data when it is unclear if consent is given
- Is it right to produce a machine learning system (in a democratic country) and then sell the system to totalitarian regime, who wants to use the system control its citizens?

References I

Books:

- Ladyman, James, Understanding Philosophy of Science, Routledge, London, 2002
- Dagfinn Föllesdal, Lars Wallöe, Jon Elster, Argumentationsteori, språk och vetenskapsfilosofi, Thales, Stockholm, 2001
- Data Ethics – The New Competitive Advantage, [link]

References II

Links

- Stanford Encyclopedia of Philosophy [here]
  - Philosophy of Statistics [here]
  - Scientific Method [here]
  - Science and Pseudo-Science [here]
  - Scientific Progress [here], Scientific Revolutions [here]
  - The Problem of Induction [here]
  - Bayes' Theorem [here]
- Probability and Induction [here]
- CODEX website - rules and guidelines for research [here]
- Big data, artificial intelligence, machine learning and data protection [here]
- DataEthics: [dataethics.eu]