Probability and Statistics

MATH40005, Spring 2023

Dean Bodenham

Imperial College London

What is Statistics?

Popular view of statistics

“There are three kinds of lies: lies, damned lies and statistics.”
- Anonymous

An expert's view of statistics

"Modern statistics, like telescopes, microscopes, x-rays, radar, and medical scans, enables us. . . to see through the mist and confusion of the world about us, to grasp the underlying reality."
- David Hand

(Source: *Statistics: A Very Short Introduction* by David Hand, Oxford University Press, 2008)

Why study statistics?

The foundation of experimental science:

Theory → experiments → data → **statistics** → conclusions

“Mathematics starts with the truth and then discovers the world; statistics starts with observing the world and then discovers the truth.”
- David Hand

- Physics
  - link: Statistics in the Discovery of a Higgs Boson by David van Dyk
- Medicine
  - Does a drug/treatment actually work?
  - Is there a link between a particular gene and a disease?

Modern applications

- Government
  - planning allocation of resources
- Manufacturing
  - quality control
- Medicine
  - Experimental design, clinical trials
- Banking
  - fraud detection, credit scoring
- Insurance
  - pensions, how to set premiums
- Finance
  - hedge funds, investments

The Joy of Stats

- Produced by Gapminder and BBC, starring **Hans Rosling**.
- Seems to be freely available on Gapminder website (only small window?):
  - https://www.gapminder.org/videos/the-joy-of-stats/
- Also available on Vimeo:
  - https://vimeo.com/18477762

Overview of module

Module intended learning outcomes

- Introduction to statistical “way of thinking”, and learn:
  - to perform several statistical analyses
  - the mathematical background behind statistical methods
  - how to reason in the face of uncertainty
  - given data, how to investigate relationships within the data
  - a few common pitfalls in statistics, and how to avoid them
- Be able to critically read and understand common statistical analyses
- Be able to perform basic statistical analyses in R

Overview of content: Weeks 1-4

- Exploratory Data Analysis
  - several methods for representing data visually
  - how to use these visualisations to interpret the data
  - topics will be spread out **over first five weeks**
- Central tendency and dispersion
  - the most common statistics and alternatives
  - bias and variance of estimators
  - confidence intervals for parameters
- Samples of normal random variables
- **Midterm in Week 6 on this material**

Overview of content: Weeks 5-8

- Hypothesis testing
- Covariance and correlation
- Introduction to statistical models
- Likelihood and maximum likelihood estimation
- Pitfalls in statistics (spread out over several weeks)
  - Correction for multiple hypothesis testing
  - Spurious correlations
  - Simpson's paradox

Overview of content: Weeks 9-11

- Simple linear regression
- Bayesian Inference
- The bootstrap
- Pitfalls in statistics
  - Anscombe's quartet

Meet the team

Spring 2023:

- Dr Dean Bodenham
- Qiquan Wang (Senior GTA) and team of GTAs

Lecture notes

- Released with gaps in exercises or proofs
- Some gaps filled in during lectures
- **Strongly encourage students to start with gapped version**
- Completed version is also available
- R code is included in several places in the notes

Recommended books

Although all the material is contained in the **notes and problem sheets**, students often request additional reading material. Here are a few recommendations:

- Bertsekas and Tsitsikils (2002) Introduction to probability, 2nd edition
- Casella and Berger (2002) Statistical inference, 2nd edition
- DeGroot and Schervish (2012) Probability and Statistics, 4th edition
- Evans and Rosenthal (2004) Probability and statistics: The science of uncertainty
- Hand (2008) Statistics: A Very Short Introduction

Lectures, problem sheets, problems classes and office hours

Lectures

- Usually two lectures per week, in the Clore
  - Tuesdays, 11.00-11.50
  - Fridays, 11.00-11.50
  - Exceptions: Week 1, and no lecture on Tuesday 14 February (midterm day)

Problem sheets and problems classes

- **Two types of problem sheet!**
- Problem-based learning classes on Thursdays, 15.00-15.50
  - Weeks 3, 4, 5, 7, 8, 9, 10; see Blackboard for schedule
  - Numbered 8-14 (Term 1 was 1-7)
- Problem sheets (self study) in same weeks as PBL classes
  - Weeks 2, 3, 4, 5, 7, 8, 9; see Blackboard for schedule
  - No PBL class/problem sheet in Week 6 because of midterm

Office hours

- Office hours
  - Tuesdays and Fridays, 13.00-13.50, in office Huxley 531.
- One-off computing office hour
  - Thursday 9 February, 13.00-13.50
  - Maths Learning Centre (Level 4 computing room)
  - At this point, students must be comfortable making an R markdown document which will be **required for the coursework**

Assessments

Assessments

- Blackboard Quiz released 31 January, due 2 February (0%)
- Midterm: Tuesday 14 February **(5%)**
- Coursework **requires R**: released Thu 23 February, due Thu 9 March **(7%)**
- Blackboard Quiz released 14 March, due 16 March (0%)
- **Please check Blackboard announcements**
- Any difficulty meeting assessment deadlines - please contact your personal tutor and the Senior Tutor
- May/June exam: half on Term 1, half on Term 2 **(70%)**
- **No calculators** for any written assessments.

The R programming language

R and RStudio

- R
  - programming language
  - many statistical functions and packages
  - one of the outcomes of this module is to become familiar with R
  - with R Markdown can be used to make reports
- RStudio
  - integrated development environment (IDE)
  - free to download
  - works on Windows, Mac, Linux
  - **Highly recommended**
- For installation instructions, see the guide on the Imperial Faculy of Natural Sciences website: https://imperial-fons-computing.github.io/rstudio.html
- Note that **R must be installed before installing RStudio**.

RStudio Cloud

- RStudio Cloud provides RStudio in a browser
- Most useful packages are already installed!
- A free account provides only 15 hours of server time per month
  - This should be more than enough for the module, but if you are doing additional computing in R, this will not be enough
- **Advice:** it is better to install R and RStudio on your device, but RStudio Cloud can be a backup option and a way to get started

Learning R

- Chapter 2 on Exploratory Data Analysis covers how to get started with basic plots, etc.
- Other sections in notes provide code for certain figures and analyses
- Problem sheets will gradually introduce R exercises
- For more background, the official Introduction to R notes are great
    - link: https://cran.r-project.org/doc/manuals/R-intro.pdf
- Goes over basic syntax and data structures, but also advanced topics
- Only need:
    - Chapters 1, 2, 6, 7, 9
    - Chapter 8 (Sections 8.1 and 8.2 only)
    - Chapter 10 (Sections 10.1, 10.2 and 10.3 only)
    - Other chapters are advanced and not necessary

Questions?

Let's get started. . .