Imperial College London

Introduction to Statistical Thinking and Data Analysis

MSc in Epidemiology / Health Data Analytics
Autumn 2022

Module leads: Dr Jeff Eaton and Dr David Muller
(jeffrey.eaton@imperial.ac.uk and david.muller@imperial.ac.uk)
Teaching assistant: Bethan Cracknell-Daniels (bethan.cracknell-daniels19@imperial.ac.uk)

10 October 2022

Introduction to Statistical Thinking and Data Analysis

Statistics is the science of

- *collecting,*
- *summarizing,*
- *presenting, and*
- *interpreting data,*

and of using them to

- *estimate the magnitude of associations* and
- *test hypotheses.*

Kirkwood and Sterne, page 1.

Objectives

1. Understand the principles and interpretation of statistical inference, sampling from a population, confidence intervals, hypothesis testing.

2. Knowledge of the assumptions and appropriate application of statistical methods commonly used in epidemiological analyses.
   – T-tests, linear regression, logistic regression, survival analysis

3. Learn and apply the R language for data manipulation, visualization, and statistical analysis.

4. Gain experience manipulating and analyzing real-world data sets, and preparing, interpreting and communicating statistical analyses.

Objectives

Practice ***doing*** statistics

Course structure

- **Lectures:** Introduce theory and example of statistics concepts. Monday 10.45–12:30
  - Textbook: Kirkwood and Sterne *Essential Medical Statistics (2nd Edition)*

- **Problem set review sessions:** Monday 9:30-10:30
  - All students together in single room.

- **Applied statistics lab sessions:** Monday 13:30-15:30
  - Group work: 4-5 students
  - Presentations: 2 classrooms with ~ 30 students + 3 tutors

- **Small group tutorials:** Wednesday 9:30-11:00 (Epi) or Thursday 15:30-17:00 (HDA)
  - Groups of 4-5 students + 3 tutors

- **R courses**: independent study on *DataCamp*.

Imperial College London

ISTDA teaching team

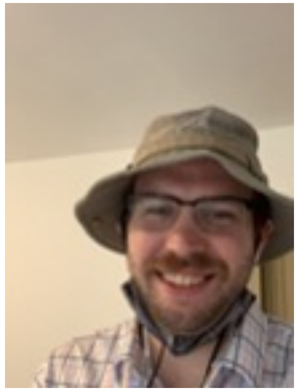
Jeff Eaton

David Muller

Lucas Cheng

Katherine Davis

Thomas Wright

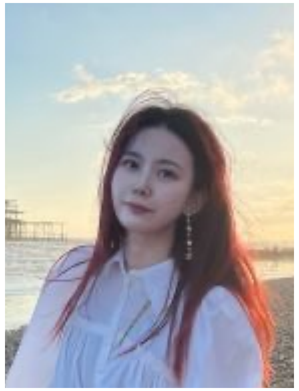


Victor Lhoste

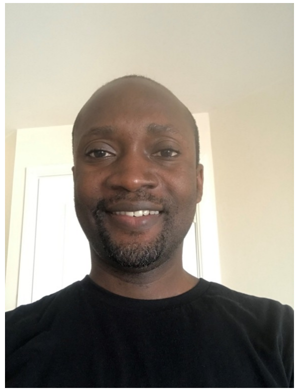
Bethan Cracknell Daniels

Panoraia Chortaria

Haowei Wang



Lanre Edun

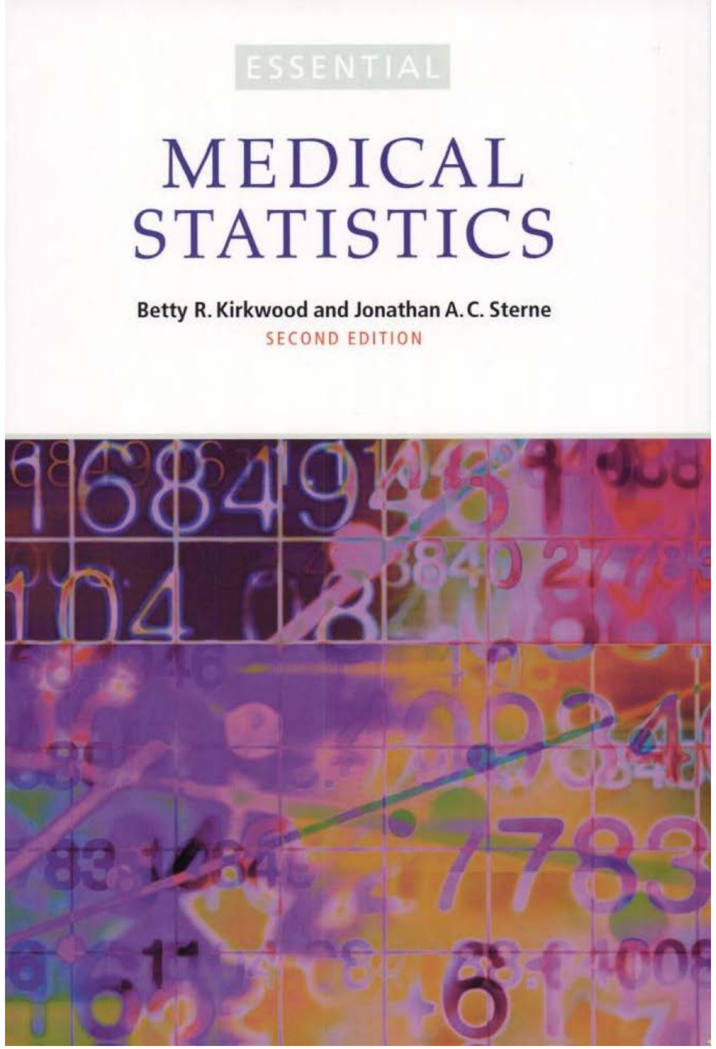


Imperial College London

Readings

- Textbook: Kirkwood and Sterne *Essential Medical Statistics (2nd Edition)*
  - Chapters assigned each week.
  - Electronic copy available from Imperial College London library.

- Supplementary readings in weeks 8–10 (see syllabus)



Lectures

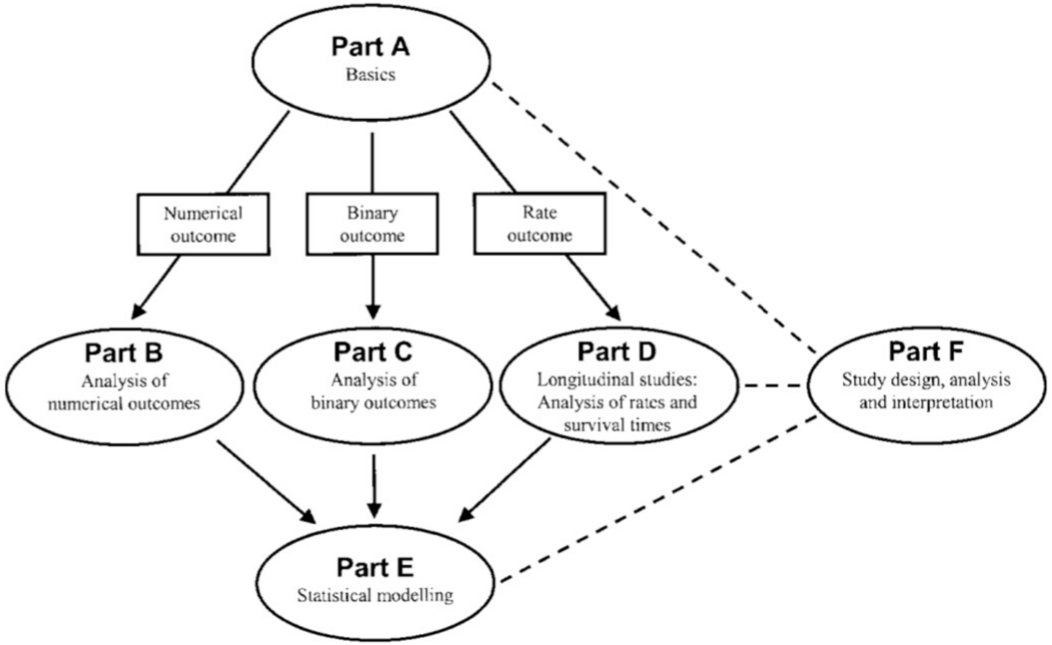

Part A
Basics
Numerical outcome
Binary outcome
Rate outcome
Part B
Analysis of numerical outcomes
Part C
Analysis of binary outcomes
Part D
Longitudinal studies: Analysis of rates and survival times
Part F
Study design, analysis and interpretation
Part E
Statistical modelling

Fig. 1.1 Organization of this book.

1. Principles of Inference, Sampling, Normal Distribution, Hypothesis Testing
2. Linear regression
3. Multiple linear regression and model building
4. Binary outcomes, comparing proportions, and chi-squared test
5. Logistic Regression
6. Poisson Regression
7. Survival analysis
8. Statistical modelling and maximum likelihood
9. Bayesian Inference, Missing data
10. Study design, Sample size calculation

Problem sets

- Weekly problem sets.
  - Work independently + in small group tutorial (Wednesday / Thursday).

- Discussed Monday 9:30–10:30.

- **A.** Consolidating key concepts from Lectures.
- **B.** Practice applying and reporting methods on actual datasets.
- **C.** Introduce more advanced statistical topics and approaches (simulation studies, robust standard errors, clustered data).

Applied Statistics Lab

- Practice *doing* statistics:
  - Data preparation and exploratory analysis
  - Developing an analysis plan.
  - Conducting analysis and interpreting results.
  - Presenting findings.

- Three group projects (4-5 persons) analysing a dataset to address a research question.
  - Continuous outcomes and linear regression,
  - Binary data and logistic regression, and
  - Longitudinal data and survival analysis.

- Culminating in 10-minute group presentation of findings (weeks 4, 7, 10).

Small group tutorials

- Assigned groups of 4-5 students, by course.

- Peer and tutored learning.
  - Epi: Wednesday 9:30 –11:00
  - HDA: Thursday 15:30 – 17:00

- Questions on lectures and readings
- Work on problem sets
- Help with R

Imperial College London

R programming

- **ISTDA is not an R course.** But focus on using R to conduct and communicate data and statistical analysis.

- DataCamp courses:
  - Introductory courses: **basic R / base R** — *recommended to complete before course start*
  - Intermediate courses: **tidyverse** — *not required for ISTDA, but recommended.*
  - Advanced courses: **R markdown, R programming, Git** — *outside scope of ISTDA, but key professional*

- Statistical analysis in R:
  - Examples of applying methods in lectures.
  - Practice applying statistical methods in weekly problem sets.
  - Applied Lab projects.
  - Peer and tutor support during small group tutorials.

DataCamp R courses

Introductory courses

- Introduction to R
- Intermediate R
- Data Visualization in R
- Introduction to Importing Data in R

Intermediate courses

- Introduction to the Tidyverse
- Data Manipulation with dplyr
- Joining Data with dplyr
- Cleaning Data in R
- Introduction to Data Visualization with ggplot2
- Intermediate to Data Visualization with ggplot2

Advanced courses

- Reporting with R Markdown
- Working with Dates and Times in R
- Introduction to Writing Functions in R
- Writing Efficient R Code
- Developing R packages
- Introduction to Git

DataCamp

Supplementary Content: Coursera

- Online module: *“Introduction to Statistics & Data Analysis in Public Health”* Online module
  - https://www.coursera.org/learn/introduction-statistics-data-analysis-public-health

- Content includes:
  - Video lectures (3-5 minutes)
  - Readings
  - Quizzes and formative assessments
  - R examples

- **Not required** for ISTDA
- Useful resource to consolidate learning through different modalities

coursera

Introduction to Statistical Thinking for Public Health

Assessments

- **Applied Statistics Lab Group presentations** (10%)
  - Three ten-minute group presentations, 3.3% each.
  - Weeks four (31 October), seven (21 November), and ten (12 December).
- **Statistical Theory and Practice Written Exam** (45%)
  - Knowledge and application of statistical principles and concepts.
  - Multiple choice and short answer; two hours.
  - Mock exam paper around Week 8.
- **Applied Statistics Mini-Project** (45%)
  - Paper reporting the results of an applied statistical analysis.
  - 2500 words in format of medical journal paper.

Communicating

- In-person sessions:
  - Lectures
  - Problem set review
  - Applied Stats Lab
  - Small-group tutorials

- Blackboard message board:
  - Questions on lectures
  - Questions and discussion on problem sets
  - Responses within 1-2 days; guide to prioritise problem set review sessions

- Microsoft Teams:
  - Applied Stats Lab Groups: chat, sharing files (Sharepoint, Office 365 online)
  - Small Group Tutorial: chat, sharing files

- Email:
  - Jeff Eaton: jeffrey.eaton@imperial.ac.uk
  - David Muller: david.muller@imperial.ac.uk
  - Bethan Cracknell Daniels: bethan.cracknell-daniels19@imperial.ac.uk

Applied Stats Lab: Room Assignments

G64				G65			
Group 1	Jingxian Huang	Group 4	Yuchen Xie	Group 7	Xihao Cao	Group 10	Ciara Hamilton
	Elena Venero Garcia		Elin Rowlands		Thomas Allwright		Alia Rafiq
	Michaelis Vasiliadis		Daniel Adams		Emily Knight		Huike Cheng
	Jian Chen		Lea Maria Khoueiry		Mehak Gurnani		Ria Sachdeva
	Seth Howes		Vaishnavi Shridar		Xheni Prebibaj		Helena Bicanic-Popovic
Group 2	Yiyang Shi	Group 5	Bing Chen	Group 8	Chiara Pligersdorffer	Group 11	Sandra Gudziunaite
	Emmanuelle Kern		Oliver Simmons		Angela Aumonier		Aditya Ramani
	Anu Bode-Favours		Daniel Huntley		Fiona Rice		Jaidip Gill
	Ka Ki Lui		Marina Berger		Nicole Cizauskas		Robert Campbell
	Siwei Wu		Wenjia Zhang		Yang Shen		
Group 3	Shuhui Li	Group 6	Mi Ma	Group 9	Mathias Brugel	Group 12	Juliet Arukwe
	Pin-Chun Wang		Harrison Goldspink		Abdul-Hakeem Khan		Gillian Sigle-Hall
	Cameron Appel		David Ensor		Gabrielle Provost		James Tait
	Kheerthiharan Saravanan		Megan Pete		Omar Eweis		Samuel Quill
	Sreenidhi Venkatesh		Wenqi Cho		Yuju Ahn		

Imperial College London

Tutorial Groups (Epi)

Tutors: Bethan Cracknell Daniels, Olanrewaju Edun, Haowei Wang

Wednesday 9:30–11:00

Group 1
Jingxian Huang
Yiyang Shi
Shuhui Li
Yuchen Xie
Bing Chen

Group 2
Mi Ma
Xihao Cao
Chiara Pligersdorffer
Mathias Brugel
Ciara Hamilton

Group 3
Sandra Gudziunaite
Juliet Arukwe
Elena Venero Garcia
Emmanuelle Kern
Pin-Chun Wang

Group 4
Elin Rowlands
Oliver Simmons
Harrison Goldspink
Thomas Allwright
Angela Aumonier

Group 5
Abdul-Hakeem Khan
Alia Rafiq
Aditya Ramani
Gillian Sigle-Hall

Tutorial Groups (HDA)

Tutors: Victor Lhoste, Thomas Wright, Panoraia Chortaria, Lucas Cheng

Thursday 15:30–17:00

Group 6
Michaelis Vasiliadis
Anu Bode-Favours
Cameron Appel
Daniel Adams
Daniel Huntley

Group 7
David Ensor
Emily Knight
Fiona Rice
Gabrielle Provost
Huike Cheng

Group 8
Jaidip Gill
James Tait
Jian Chen
Ka Ki Lui
Kheerthiharan Saravanan

Group 9
Lea Maria Khoueiry
Marina Berger
Megan Pete
Mehak Gurnani
Nicole Cizauskas

Group 10
Omar Eweis
Ria Sachdeva
Robert Campbell
Samuel Quill
Seth Howes

Group 11
Siwei Wu
Sreenidhi Venkatesh
Vaishnavi Shridar
Wenjia Zhang
Wenqi Cho

Group 12
Xheni Prebibaj
Yang Shen
Yuju Ahn
Helena Bicanic-Popovic

Any questions?