Introduction to Statistical Thinking and Data Analysis

MSc in Epidemiology and MSc in Health Data Analytics, Imperial College London

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

Autumn 2022

Introduction

Welcome to *Introduction to Statistical Thinking and Data Analysis* (ISTDA). In this course you will develop the knowledge and skills required to conduct and interpret statistical analyses of epidemioglic and health data. The course will cover the types of data, study designs, and statistical tools suitable for the large majority of applied research and practice. You will use these statistical skills throughout your other course modules, your summer thesis projects, and your future careers.

By the end of this module, you will:

1. Understand the principles and interpretation of statistical inference, sampling from a population, confidence intervals, hypothesis testing.
2. Have knowledge of the assumptions and appropriate application of statistical methods commonly used in epidemiological analyses including t-tests, linear regression, logistic regression, survival analysis, and handling missing data.
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.

Course outline

In this course you will learn and practice key statistical methods for epidemiologic analysis through lecture, applied statistics group projects, and R software-based problem sets. Lectures and the textbook will introduce theory and examples of key statistical concepts. Weekly 'Applied Statistics Lab' sessions will provide practice in the application, interpretation, and presentation of data analysis and statistical findings through three group projects over the course of the term. Programming and statistical analysis using the R software programme will be developed through interactive online practice, problem sets, and applied lab projects. Weekly problem sets will be provided to practice and review concepts, application, and programming and revised together at the start of each week.

Statistical content

Statistical content will be delivered through weekly lectures and through the textbook.

The course textbook is:

- *Essential Medical Statistics (Second Edition)* by Betty R. Kirkwood and Jonathan A. C. Sterne

Course content will follow closely to the textbook and specific chapters are assigned corresponding to the lecture each week.

An electronic version of the textbook is available from the Imperial College London library via the following link: https://library-search.imperial.ac.uk/permalink/f/tlmp2g/44IMP_ALMA_DS5155792570001591.

Additional supplementary content is available from the Introduction to Statistics module developed by the Imperial School of Public Health in collaboration with Coursera: https://www.coursera.org/learn/introduction-statistics-data-analysis-public-health. Content on the Coursera platform consists of:

- Short three to five minute video lessons introducing statistical concepts and methods, prepared by Prof. Alex Bottle and Dr. Victoria Cornelius.
- Brief readings to extend key concepts.
- Quiz questions to consolidate learning of key concepts
- Practice data analysis tasks in the R software.

Coursera content is not required for this module, but are useful resources to consolidate learning through different modalities.

Statistical computing

The course will utilize the statistical software programme *R*, a free software environment for statistical computing and data visualisation. R runs on all major computer platforms (Windows / Mac / Unix). Download and install the most recent release of R for your computing platform from: https://cloud.r-project.org.

We recommend using the *RStudio* integrated development environment, a freely available software programme providing features to interact with R more efficiently. You will need to install both the R software and RStudio.

One of the attractive features of R is the large and enthusiastic user community and the large number of contributed extension packages. Most of the statistical methods covered in the this course are implemented in the standard R software (often referred to as *'base R'*), but packages extend R to implement the most cutting edge statistical methods and data analysis tools. In this course, we will particularly focus on learning a collection of packages referred to as the *tidyverse*. These packages provide powerful and efficient tools for data manipulation and visualisation, a large and important component of the applied statistical workflow.

***Introduction to Statistical Thinking and Data Analysis* is not an R computing course.** You will not be evaluated based on your R code or ability to complete tasks in R. But we will use R as a tool for conducting and communicating data and statistical analysis.

Course content will focus on how to use R to conduct and report statistical analysis. For learning and practising the core features of the R language and software, we recommend completing courses in the online learning platform *DataCamp* for learning and practising the core features and functionality of R.

The course instructors and tutors use R for their work on a daily basis and are available to assist and answer questions about using R ranging from basic tasks, conducting statistical analyses, or advanced applications.

Introductory topics

The following courses will introduce you to the key features of the R language and R software. This will cover the majority of essential R skills required to complete the ISTDA course, aside from specific statistical tools and methods learned through lectures and tutorial sheets.

- *Introduction to R*: https://www.datacamp.com/courses/free-introduction-to-r
- *Intermediate R*: https://www.datacamp.com/courses/intermediate-r
- *Data Visualization in R*: https://www.datacamp.com/courses/data-visualization-in-r
- *Introduction to Importing Data in R*: https://www.datacamp.com/courses/introduction-to-importing-data-in-r

Intermediate topics

These intermediate courses cover a more advanced set of R features, many focused on extension packages that are part of the *tidyverse*.

These are not required to be successful in the ISTDA course, but are a powerful set of tools for data manipulation and visualisation that you will likely find useful in the course, further study, and future projects. It is recommended that familiarise yourself with these tools through the following course.

- *Introduction to the Tidyverse*: https://www.datacamp.com/courses/introduction-to-the-tidyverse
- *Data Manipulatio with dplyr*: https://www.datacamp.com/courses/data-manipulation-with-dplyr
- *Joining Data with dplyr*: https://www.datacamp.com/courses/joining-data-with-dplyr
- *Cleaning Data in R*: https://www.datacamp.com/courses/data-cleaning-in-r
- *Introduction to Data Visualization with ggplot2*: https://www.datacamp.com/courses/introduction-to-data-visualization-with-ggplot2
- *Intermediate to Data Visualization with ggplot2*: https://www.datacamp.com/courses/intermediate-data-visualization-with-ggplot2

Advanced topics

The courses below introduce powerful R programming features and tools to support reproducible research, collaboration, and sharing. These skills are beyond the core statistical, data analysis, and communication content of the ISTDA course. But they are tools and workflows that will be an expected part of your statistics and data science toolbox in a professional environment.

- *Reporting with R Markdown*: https://www.datacamp.com/courses/reporting-with-rmarkdown
- *Working with Dates and Times in R*: https://www.datacamp.com/courses/working-with-dates-and-times-in-r
- *Introduction to Writing Functions in R*: https://www.datacamp.com/courses/introduction-to-writing-functions-in-r
- *Writing Efficient R Code*: https://www.datacamp.com/courses/writing-efficient-r-code
- *Developing R packages*: https://www.datacamp.com/courses/developing-r-packages
- *Introduction to Git*: https://www.datacamp.com/courses/introduction-to-git

In addition to online courses, the following books are useful for learning R computing including data manipulation, visualisation, and statistical modelling:

- *Hands-On Programming with R* by Garrett Grolemund: https://rstudio-education.github.io/hopr/
- *R for Data Science* by Garrett Grolemund and Hadley Wickham: https://r4ds.had.co.nz
- *The R Software: Fundamentals of Programming and Statistical Analysis* by Pierre Lafaye de Micheaux, Rémy Drouilhet, Benoit Liquet

The first two texts are fully and freely available online at the links above. Physical copies of the books are available for purchase if desired.

Lafaye de Micheaux, Drouilhet, and Liquet text is available as an e-book version from the Imperial College London library: https://library-search.imperial.ac.uk/permalink/f/tlmp2g/44IMP_ALMA_DS51105617330001591.

There are myriad other resources available online for learning and practicimg R. You are encouraged to explore them, and please share with your colleagues and tutors which materials you find most useful and effective.

Preparation

In advance of the course, we recommend the following preparation:

- Read *Part A: Basics* (Chapters 1-3) of *Essential Medical Statistics* by Kirkwood and Sterne.

- Install and become familiar with R and RStudio. The Appendices of *Hands-On Programming with R* by Garrett Grolemund describe how to install R and RStudio (Appendix A), installing and loading R packages (Appendix B), and loading and saving data in R (Appendix D).
- Complete introductory Data Camp courses / chapters:
  - *Introduction to R*: https://www.datacamp.com/courses/free-introduction-to-r
  - *Intermediate R*: https://www.datacamp.com/courses/intermediate-r
  - *Data Visualization in R*: https://www.datacamp.com/courses/data-visualization-in-r
  - "Chapter 1: Importing data from flat files with utils" from *Introduction to Importing Data in R*: https://campus.datacamp.com/courses/introduction-to-importing-data-in-r/importing-data-from-flat-files-with-utils

Course structure and components

There are four components for the ISTDA course (*all times London time zone*; GMT+1 to 30 October, GMT from 31 October to December):

- **Problem set review** sessions will be *Monday 9:30 to 10:30*.
- **Lectures** will be *Monday 10:45 to 12:30*.
- **Applied Statistics Lab** sessions will be be on *Monday 13:30 to 15:30*.
- **Small group tutorial** sessions will be on *Wednesday from 9:30 to 11:00* for MSc Epidemiology students and *Thursday from 15:30 to 17:00* for MSc Health Data Analytics.

Lectures

Monday morning lectures will be the primary venue for introducing the principles and interpretation of statistical methods and tools. Lecture content will follow closely to the content of the course textbook with examples and occsasional content drawn from other sources. You are recommended to read the relevant chapters of Kirkwood and Sterne *in advance* of the lecture and then likely revise with the textbook as you practice with tutorial sheets and applied statistics projects.

Lectures will be held on the St Mary's campus. Please see the timetable for the room for each week.

Lecture slides will be available online after the lecture along with datasets and R code for any examples presented in lecture. We will aim to record lectures and make them available for further review for **one week** following the lecture.

Problem Set Sessions

Problem sets will be provided following each lecture. Problem sets will focus on:

- Review and consolidate key concepts from lectures and readings.
- Practice applying and interpreting statistical tools from weekly lectures on example datasets.
- Introduce more advanced statistical concepts and approaches.

Problem sets are to be worked on independently or with peers over the course of the week, with opportunity to ask questions to course tutors during Wednesday or Thursday small group tutorials.

Problem sets from the previous week will be reviewed during review sessions each Monday morning from 9:30 to 10:30. These will be in the same room as the lectures.

There will be an online discussion board on Blackboard to ask questions on lectures, problem sets, or other course content. Questions on the message board will be answered within 2-3 days of posting, and be used to prioritise topics to review during Monday morning review sessions.

Applied Statistics Lab

Applied Statistics Lab sessions on Monday 13:30 to 15:30 are an opportunity to practice the activity that you will do day in and day out as an epidemiologist or biostatistician: dataset preparation and exploratory analysis, developing an analysis plan, conducting statistical analysis, and interpreting and reporting the results of statistical analyses.

You will complete three applied statistics group projects over the course of the term focused on practicing analysis and intepretation of common types of data and questions in epidemiologic and health data:

- Continuous outcomes and linear regression,
- Binary data and logistic regression, and
- Longitudinal data and survival analysis.

Each project will be conducted in groups of 4-5 peers over three weeks and culminate in a group presentation about your findings. Across each of the projects, you will practice applied statistics workflow including exploratory and descriptive data analysis and visualisation, developing an analysis plan to address your research question, carrying out and checking your analysis, and interpreting and reporting the conclusions of your analysis. Datasets will be actual datasets used to address real-world research questions, requiring data cleaning, decisions about inclusion/exclusion of cases, construction of appropriate metrics and indicators, and informed judgements about the construction and interpretation of variables and outcomes.

Applied Statistics Lab sessions will be held on St Mary's Hospital campus across three classrooms. Group and room assignments will be announced during the first week of the course.

Each lab group will be allocated a private channel on Microsoft Teams to communicate and work collaboratively on files.

Small Group Tutorials

Weekly small group tutorial sessions will provide opportunities for revision with peers and with a course tutor questions related to lectures, problem sheets, or R computing.

For MSc Epidemiology students, tutorial groups will meet on Wednesdays from 9:30 - 11:00 at St Mary's Hospital campus. For MSc Health Data Analystics students, tutorial groups will meet on Thursdays from 15:30 - 17:00 at St. Mary's Hospital campus.

Each tutorial group will also be allocated a channel on Microsoft Teams to exchange messages throughout the week outside of Wednesday or Friday sessions.

Assessments

There will be three modalities of assessment:

- Three **Applied Statistics Lab Group Presentations** will comprise 10% of the total course marks (3.3% each). Presentations will consist of reporting the results of three applied statistics group projects and are the primary opportunity to practice oral communication of statistical findings. Each group presentation will be 10 minutes occurring during weeks four, seven, and ten.
- A two hour **Statistical Theory and Practice Written Exam** on 12 January will comprise 45% of the course marks. The exam format will be multiple choice and short answer exam questions. The exam will assess knowledge and application of the statistical principles and concepts covered in the course. The exam will not assess R or statistical programming.
- The **Applied Statistics Mini-Project** will comprise 45% of the overall marks. The mini project instructions will be provided on 16 December 2022 and project papers will be due on 6 January 2023. For the exam you will be given a dataset and a research question. From this you will design an analysis

plan and conduct a statistical analysis to address the research question. The report will be a maximum of 2500 words in the format of a medical journal paper. This will be your primary opportunity to practice written communication of statistical findings.

Syllabus

Week 1 (10 October)

Lecture 1A: Principles of Inference, Sampling, Normal Distribution
Lecture 1B: Confidence intervals, Hypothesis testing, and p-values

- Learning objectives:
  - Understand the purpose and principles of statistical inference.
  - Calculate and interpret standard summary measures of a sample.
  - Define the normal probability distribution, why it arises, and why it is important.
  - Define the t-distribution and how to use it to calculate confidence intervals for a mean.
  - Understand the logic of hypothesis testing, including defining the null and alternative hypothesis.
  - Carry out standard hypothesis tests to compare two means.
  - Define and interpret a 'confidence interval'.
  - Calculate the area under the curve of the normal distribution.
- Reading:
  - Kirkwood and Sterne, Chapters 1-8

Applied Statistics Lab: Designing a Statistical Analysis

- Kirkwood and Sterne, Chapter 38 (*Strategies for analysis*)

Week 2 (17 October)

Lecture: Linear regression

- Learning objectives:
  - Describe simple linear regression and name the assumptions on which it is based.
  - Interpret linear regression coefficients, their confidence intervals and significance tests.
  - Fit linear regression models in R and check the assumptions of the regression model.
- Reading:
  - Kirkwood and Stern, Chapters 9-10, 12

Applied Statistics Lab: Project 1: Continuous Outcome

Week 3 (24 October)

Lecture: Multiple linear regression and model building

- Learning objectives:
  - Use multiple regression to describe, to adjust, and to predict.
  - Develop a model building strategy to identify which and how many variables to include in a model.
  - Interpret interactions and decide when to include interaction terms in multiple regression.
  - Know why and when to transform outcome and exposure variables in linear regression.
- Reading:
  - Kirkwood and Stern, Chapters 11-13

Applied Statistics Lab: Project 1: Continuous Outcome

Week 4 (31 October)

Lecture: Binary outcomes, comparing proportions, chi-squared tests

- Learning objectives:
  - Define and be able to identify binary data.
  - Understand and comput chi-squared tests for comparing proportions from population samples.
  - Calculate and interpret odds ratios.
- Reading:
  - Kirkwood and Stern, Chapters 14-17

Applied Statistics Lab: **Project 1 Group Presentation**

Week 5 (7 November)

Lecture: Logistic regression

- Learning objectives:
  - Apply and interpret logistic regression to estimate odds ratios, confidence intervals, and p-values.
  - Run multiple logistic regression and interpret output.
  - Evaluate model assumptions for multiple logistic regression.
- Reading:
  - Kirkwood and Stern, Chapters 19-21

Applied Statistics Lab: Project 2: Binary Outcome

Week 6 (14 November)

Lecture: Longitudinal data, Poisson regression

- Learning objectives:
  - Identify longitudinal data and estimate rates.
  - Understand the Poisson distribution and how it arises.
  - Use Poisson regression to estimate rate ratios.
- Reading:
  - Kirkwood and Stern, Chapters 22-25

Applied Statistics Lab: Project 2: Binary Outcome

Week 7 (21 November)

Lecture: Survival analysis and the Cox proportional hazards model

- Learning objectives:
  - Name the characteristic features of survival data, including censoring.
  - Interpret Kaplan-Meier survival curves and compare survival using log-rank tests.
  - Use R to produce Kaplan-Meier plots and descriptive statistics relating to survival analysis.
  - Name the assumptions and principles underlying Cox regression and interpret results.
  - Fit and check the assumptions of Cox regression models.
- Reading:
  - Kirkwood and Stern, Chapter 26-27

Applied Statistics Lab: **Project 2 Group Presentation**

Week 8 (28 November)

Lecture: Statistical modelling and maximum likelihood

- Learning objectives:
  - Understand the concept of likelihood of observed data.
  - Define a likelihood function.
  - Use a likelihood function to estimate parameters, p-values, and confidence intervals.
  - Recast regression models as maximum likelihood estimation.
- Reading:
  - Kirkwood and Stern, Chapter 28-29
  - Kirkwood and Stern, Chapters 30-31 (supplemental)

Applied Statistics Lab: Project 3: Survival Outcome

Week 9 (5 December)

Lecture: Bayesian Inference, Missing data

- Learning objectives:
  - Describe the difference between Bayesian and Frequentist inference approaches.
  - Identify mechanisms and assumptions for different types of missing data.
  - Describe the implications of different types of missing data on results of statistical analyses.
  - Determine suitable strategies to deal with missing data, including knowledge of when multiple imputation is appropriate.
- Reading:
  - Sterne J.A.C., White I.R, Carlin J.B, Spratt M, Royston P, Kenward MG et al. (2009) Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ; 338:b2393. https://www.bmj.com/content/338/bmj.b2393.long

Applied Statistics Lab: Project 3: Survival Outcome

Week 10 (12 December)

Lecture: Study design, Sample size calculation

- Learning objectives:
  - Determine the appropriate study design for a research question.
  - Choose an appropriate statistical method for a given study design.
  - Be aware of assumptions and limitations of each type of statistical analysis.
  - Understand principles and formulae for sample size calculation.
  - Define clustering in sample data, consequences for analysis, and how to address in sample size calculation and analysis.
- Reading:
  - Kirkwood and Sterne, Chapter 34-35

Applied Statistics Lab: **Project 3 Group Presentation**