Introduction to Statistical Thinking and Data Analysis

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

Autumn 2019

Introduction

Welcome to *Introduction to Statistical Thinking and Data Analysis* (ISTDA). The purpose of this course is to 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 ubiquitously in 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 practicals. 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 lab tutorials. Weekly problem sets will be provided to practice and review concepts, application, and programming and revised together at the start of each week. Optional advanced mathematics review sessions will provide an opportunity to review key mathematical underpinnings for second-term advanced statistics elective modules.

Statistical content

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].

The textbook will be supplemented by the excellent series of 'Statistics Notes' authored by J. Martin Bland and Douglas G. Altman in the *British Medical Journal* between 1994 and 1999. Specific notes will be recommended accompanying relevant lecture material each week. A brief history of the 'Statistics Notes' series is available here from J Martin Bland's website at the University of York.

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 here: 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.

We will use three texts 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
- *Cookbook for R* by Winston Chang: http://cookbook-r.com

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

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 demonstrators 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).
- Read and work through Projects 1-3 of *Hands-On Programming with R* by Garrett Grolemund.

Course structure and timetable

There are four required and one optional classroom components for the ISTDA coures:

- **Tutorial Review** sessions will be *Monday mornings XX to XX* in room XX (MSc Epi) or XX (MSc HDA).
  - Dr. Jeff Eaton and TBC
- **Lectures** will be *Monday mornings XX to XX* in room XX.
  - Dr. Jeffrey Eaton and TBC
- **Applied Statistics Lab** sessions will be be on *Monday afternoons from XX to XX* in room XX (MSc Epi) or XX (MSc HDA).
  - Dr. Deborah Schneider-Luftman
- **R Statistical Computing** sessions will be on *Wednesday mornings from XX to XX* in room XX (MSc Epi) or XX (MSc HDA).
  - Dr. Juliette Unwin

- Optional **Advanced Math Refresher** sessions will be *Wednesday afternoons from XX to XX* in room XX.
  - *Dr. Barbara Bodinier?*

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.

Lecture slides will be available online after the lecture along with datasets and R code for any examples presented in lecture.

Applied Statistics Lab

Applied Statistics Lab sessions on Monday afternoons 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.

R statistical computing tutorials

Wednesday morning R statistical computing tutorial sessions will introduce the R software, tools for data manipulation and visualisation with R, and how to conduct statitistical methods decribed in the lectures using R. The first several weeks will entail lectures introducing features of R and reviewing key R tools required for lectures, tutorial sheets, and applied statistics lab sessions. Sessions will also provide an opportunity to ask course tutors questions about statistical or programming questions related to lectures, tutorial sheets, and applied statistics lab sessions.

Tutorial sheets

Tutorial sheets with practice problem sets will be provided at the start of each week to consolidate and practice the statistical concepts discussed in each lecture. 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 morning R tutorials. Problem sets from the previous week will be reviewed on Monday mornings immediately preceeding each lecture. Successful completion of problem sets will prepare you well for the statistical theory and practice written exam.

Maths review (optional)

Optional Maths Review sessions on Wednesday afternoons will review core mathematical used in advanced statistics.

JE: Marc, Barbara – please add brief summary of topics to be covered in maths review sessions.

These sessions are recommended for any students not recently familiar with these topics who are planning to proceed to mathematical modelling, advanced statistics (e.g. Bayesian statistics, spatial statistics), or machine learning modules in the second term.

Assessments

There will be three modalities of assessment:

- Three **Applied Statistics Lab Group Presentations** will comprise 20% of the total course marks (6.7% 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 approximately 15 minutes occurring during weeks four, seven, and ten.
- A one hour **Statistical Theory and Practice Written Exam** on will comprise 40% of the course marks. The exam format will be multiple choice and short answer exam with pen and paper. The exam will assess knowledge and application of the statistical principles and concepts covered in the course.
- The **Applied Statistics Mini-Project** will comprise 40% of the overall marks. The mini project instructions will be provided on and project papers will be due on . 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 3000 words in the format of a medical journal paper. This will be your primary opportunity to practice written communication of statistical findings.

Syllabus

JE: To add Bland and Altman notes and other relevant readings (e.g. good missing data review)

JE: Marc, Barbara – please add any readings or content relevant to Maths refresher sessions.

JE: Ettie – please fill in reading assignments for R content, or let's discuss and do together

Week 1 (7 October): Principles of statistical inference, sampling variability the normal distribution

- Kirkwood and Sterne, Chapters 1-5
- Kirkwood and Sterne, Chapter 38 (*Strategies for analysis*)
- Grolemund, Project 1.

Week 2 (14 October): Confidence intervals, hypothesis testing, and p-values

- Kirkwood and Stern, Chapter 6-8
- Grolemund, Projects 2-3

Week 3 (21 October): Analysis of variance and linear regression

- Kirkwood and Stern, Chapters 9-13
- Grolemund and Wickham, *TBC*

Week 4 (28 October): Binary outcomes, comparing proportions, chi-squared tests

- Kirkwood and Stern, Chapters 14-17
- *Applied Statistics Lab Group Presentation 1*

Week 5 (4 November): Logistic Regression

- Kirkwood and Stern, Chapters 18-21

Week 6 (11 November): Longitudinal data, Poisson Regression

- Kirkwood and Stern, Chapters 22-25

Week 7 (18 November): Survival analysis

- Kirkwood and Stern, Chapters 26-27
- *Applied Statistics Lab Group Presentation 2*

Week 8 (25 November): Statistical modelling, Maximum likelihood, Bayesian inference

- Kirkwood and Stern, Chapters 28-30, 33
- Supplemental reading: Kirkwood and Stern, Chapters 30-31

  JE: This week feels a bit full to give more than a cursory introduction to these topics. It would also be nice to at least cover clustered data a bit (Chapter 31).

  JE: The Bayesian chapter (Chapter 33) in Kirkwood and Sterne is a bit of a throw away. We ought to think about what the important take away for them is on this topic to prepare them for future modules and pull in some other content.

Week 9 (2 December): Model building, Missing data

- Carpenter JR, Kenward MG. *Missing data in randomised controlled trials— a practical guide* 2007. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.468.9391&rep=rep1&type=pdf

  JE: We might wish to replace this week with deeper elaboration on some of the previous week topics on model-based statistics.

  JE: I'm refusing to teach 'model selection' other than to say don't do it, hence 'model building'. This topic, and perhaps missing data, might be covered sufficiently in the applied stats lab sessions and the Epi module that we don't need a lecture on it.

Week 10 (9 December): Study design, Sample size calculation

- Kirkwood and Sterne, Chapter 34-35
- *Applied Statistics Lab Group Presentation 3*

  JE: These chapters double as good course and exam review because they