# Introduction to Statistical Thinking and Data Analysis

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

*Module Leader: Jeff Eaton (jeffrey.eaton@imperial.ac.uk)*

*Autumn 2019*

## Timetable

| Week | Date | Time | Type | Topic | Lecturer | Room (Epi / HDA) |
|---|---|---|---|---|---|---|
| Week 1 | 7 October | 10:00 - 11:00 | Lecture | Principles of Inference, Sampling, Normal Distribution | Jeff Eaton | G64 |
| | | 11:15 - 12:15 | Lecture | Statistics in Epidemiology and Public Health | Prof Christl Donnelly | G64 |
| | | 14:00 - 15:30 | Lab | Statistical analysis plan; Project 1 introduction | Pazoki / Schneider-Luftman | MSc / G64 |
| | 9 October | 9:30 - 11:00 | Practical | R: Project 1 – R basics | Juliette Unwin | G64 |
| | | 14:00 - 15:30 | Lecture* | Calculus I: Derivatives | James Truscott | MSc Room |
| Week 2 | 14 October | 10:00 - 11:00 | Tutorial | Problem Sheet 1 Review | Jeff Eaton | Bannister |
| | | 11:15 - 12:15 | Lecture | Confidence Intervals, Hypothesis Testing | Jeff Eaton | Bannister |
| | | 14:00 - 15:30 | Lab | Project 1: continuous outcome | Pazoki / Schneider-Luftman | **MSc / Hynds** |
| | 16 October | 9:30 - 11:00 | Practical | R: Project 2 – Manipulating R Objects | Juliette Unwin | **MSc Room** |
| | | 14:00 - 15:30 | Lecture* | Calculus II: Integrals | James Truscott | MSc Room |
| Week 3 | 21 October | 10:00 - 11:00 | Tutorial | Problem Sheet 2 Review | Jeff Eaton | Rothschild |
| | | 11:15 - 12:15 | Lecture | Analysis of Variance, Linear Regression | Jeff Eaton | Rothschild |
| | | 14:00 - 15:30 | Lab | Project 1: continuous outcome | Pazoki / Schneider-Luftman | MSc / G64 |
| | 23 October | 9:30 - 11:00 | Practical | R: Data Visualisation | Juliette Unwin | G64 |
| | | 14:00 - 15:30 | Lecture* | Calculus III: Differential Equations | James Truscott | MSc Room |
| Week 4 | 28 October | 10:00 - 11:00 | Tutorial | Problem Sheet 3 Review | Jeff Eaton | Rothschild |
| | | 11:15 - 12:15 | Lecture | Binary Outcomes, Comparing Proportions | Jeff Eaton | Rothschild |
| | | 14:00 - 15:30 | Lab | **Project 1 presentation** | Pazoki / Schneider-Luftman | MSc / G64 |
| | 30 October | 9:30 - 11:00 | Practical | R: Loading and Formatting Data | Juliette Unwin | G64 |
| | | 14:00 - 15:30 | Lecture* | Linear Algebra I: Introduction to matrices and matrix calculus | Barbara Bodinier | MSc Room |
| Week 5 | 4 November | 10:00 - 11:00 | Tutorial | Problem Sheet 2 Review | Fred Piel | G64 |
| | | 11:15 - 12:15 | Lecture | Logistic Regression | Fred Piel | G64 |
| | | 14:00 - 15:30 | Lab | Project 2: Binary outcome | Pazoki / Schneider-Luftman | MSc / G64 |
| | 6 November | 9:30 - 11:00 | Practical | R: Dealing with Data I | Juliette Unwin | G64 |
| | | 14:00 - 15:30 | Lecture* | Linear Algebra II: Vector spaces and systems of equations | Barbara Bodinier | MSc Room |
| Week 6 | 11 November | 10:00 - 11:00 | Tutorial | Problem Sheet 5 Review | Fred Piel/David Muller | G64 |
| | | 11:15 - 12:15 | Lecture | Longitudinal Data, Poisson Regression | Fred Piel/David Muller | G64 |
| | | 14:00 - 15:30 | Lab | Project 2: Binary outcome | Pazoki / Schneider-Luftman | MSc / G64 |
| | 13 November | 9:30 - 11:00 | Practical | R: Dealing With Data II | Juliette Unwin | G64 |
| | | 14:00 - 15:30 | Lecture* | Linear Algebra III: Mathematical framework of linear regression | Barbara Bodinier | MSc Room |
| Week 7 | 18 November | 10:00 - 11:00 | Tutorial | Problem Sheet 6 Review | David Muller | Rothschild |
| | | 11:15 - 12:15 | Lecture | Survival Analysis | David Muller | Rothschild |
| | | 14:00 - 15:30 | Lab | **Project 2 presentation** | Pazoki / Schneider-Luftman | MSc / G64 |
| | **21 November** | 9:30 - 11:00 | Tutorial Q&A | Problem sheet Q&A (No R session) | Course tutors | G62 |
| | | 14:00 - 15:30 | Lecture* | Linear Algebra IV: Eigenvalues and eigenvectors | Barbara Bodinier | G62 |
| Week 8 | 25 November | 10:00 - 11:00 | Tutorial | Problem Sheet 7 Review | Jeff Eaton | Rothschild |
| | | 11:15 - 12:15 | Lecture | Statistical Modelling, Maximum Likelihood | Jeff Eaton | Rothschild |
| | | 14:00 - 15:30 | Lab | Project 3: Survival outcome | Pazoki / Schneider-Luftman | MSc / G64 |
| | 27 November | 9:30 - 11:00 | Practical | R: Good Programming Practice | Juliette Unwin | G64 |
| | | 14:00 - 15:30 | Lecture* | Linear Algebra V: Principal Component Analysis | Barbara Bodinier | MSc Room |

## Timetable

| Week | Date | Time | Type | Topic | Lecturer | Room (Epi / HDA) |
|---|---|---|---|---|---|---|
| Week 9 | 2 December | 10:00 - 11:00 | Tutorial | Problem Sheet 8 Review | Jeff Eaton | Rothschild |
| | | 11:15 - 12:15 | Lecture | Bayesian Inference, Missing Data | Jeff Eaton | Rothschild |
| | | 14:00 - 15:30 | Lab | Project 3: Binary outcome | Pazoki / Schneider-Luftman | MSc / G64 |
| | 4 December | 9:30 - 11:00 | Practical | R: Advanced Programming | Juliette Unwin | G64 |
| | | 14:00 - 15:30 | Lecture* | Python I: Introduction to Python | Gianluca Campanella | MSc Room |
| Week 10 | 9 December | 10:00 - 11:00 | Tutorial | Problem Sheet 9 Review | Jeff Eaton | Bannister |
| | | 11:15 - 12:15 | Lecture | Study design, Sample Size Calculation | Jeff Eaton | Bannister |
| | | 14:00 - 15:30 | Lab | **Project 3 presentation** | Pazoki / Schneider-Luftman | MSc / G64 |
| | 11 December | 9:30 - 11:00 | Practical | R: Revision | Juliette Unwin | G64 |
| | | 14:00 - 15:30 | Lecture* | Python II: Python computational using pyspark | Gianluca Campanella | MSc Room |

* Indicates optional Maths Refresher lectures

## 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 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.

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
- *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 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 components

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

- **Tutorial Review** sessions will be *Monday mornings 10:00 to 11:00* in room G64, Bannister Lecture Theatre, or Rothschild Lecture Theatre (see timetable).
  - Dr. Jeff Eaton, Dr. Fred Piel, Dr. David Muller
- **Lectures** will be *Monday mornings 11:15 to 12:15* in room G64, Bannister Lecture Theatre, or Rothschild Lecture Theatre (see timetable).
  - Dr. Jeffrey Eaton, Dr. Fred Piel, Dr. David Muller
- **Applied Statistics Lab** sessions will be be on *Monday afternoons from 14:00 to 15:30* in the MSc Room (MSc Epi) or G64 (MSc HDA).
  - Dr. Raha Pazoki, Oliver Eales, Katherine Davis (MSc Room)
  - Dr. Deborah Schneider-Luftman, Dr. Matthew Thomas, Sumali Bajaj (G64)
- **R Statistical Computing** sessions will be on *Wednesday mornings from 9:30 to 11:00* in room G64.
  - Dr. Juliette Unwin

- Optional **Advanced Math Refresher** sessions will be *Wednesday afternoons from 13:30 to 15:30* in the MSc Room.
  - Dr. James Truscott, Dr. Alice Ledda, Barbara Bodinier, Dr. Gianluca Campanella

### 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 sessions (optional)

Optional Maths Review sessions on Wednesday afternoons (2-3:30 pm) will review core mathematical used in advanced statistics. Sessions will cover the following topics:

- Weeks 1 through 3: Calculus (Derivatives, Integrals, Differential Equations).
- Weeks 4 through 8: Linear Algebra.
- Weeks 9 through 10: Introduction to Python.

These topics are relevant to statistical modelling and will provide the students with key concepts to better understand the mathematical framework underlying the models, their assumptions, and limitations. The sessions will involve a mixture of lecture and interactive problem sets to practice concepts.

These sessions are strongly recommended for any students 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 10 minutes occurring during weeks four, seven, and ten.
- A one hour **Statistical Theory and Practice Written Exam** on 7 January 2020 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 30 December 2019 and project papers will be due on 10 January 2020. 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

### Week 1 (7 October)

*Lecture:* Principles of statistical inference, sampling variability the normal distribution.

- 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.
  - Calculate the area under the curve of the normal distribution.
- Reading:
  - Kirkwood and Sterne, Chapters 1-5
  - Altman DG, Bland JM. Uncertainty and sampling error. 2014; 349: g7064. https://www.bmj.com/content/349/bmj.g7064
  - Altman DG, Bland JM. (1994) Quartiles, quintiles, centiles, and other quantiles. BMJ 309, 996. https://www.bmj.com/content/309/6960/996
  - Altman DG, Bland JM. (1995) The normal distribution. BMJ, 310, 298. https://www.bmj.com/content/310/6975/298

  - Altman DG, Bland JM. (2005) Standard deviations and standard errors. 331, 903. https://www.bmj.com/content/331/7521/903

*Special Lecture:* The Role of Statistics in Epidemiology and Public Health

- *Professor Christl Donnelly CBE FMedSci FRS*

*Applied Statistics Lab:* Reading a Journal Paper, Designing a Statistical Analysis

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

*Statistical Computing:* Project 1 – R Basics

- Grolemund, Project 1 (Chapters 1-2 in book 1 - 3 online).

*Maths Refresher (optional):* Calculus I: Derivatives

**Week 2 (14 October)**

*Lecture:* Confidence intervals, hypothesis testing, and p-values

- Learning objectives:
  - Define and interpret a 'confidence interval'.
  - 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.
  - Define 'Type I' and 'Type II' errors and statistical power.
  - Carry out standard hypothesis tests (1-sample t-test, 2-sample t-test, paired t-test).
  - Select an appropriate hypothesis test for a given research question.
- Reading:
  - Kirkwood and Stern, Chapter 6-8
  - Altman DG, Bland JM. (2005) Standard deviations and standard errors. BMJ 331, 903. https://www.bmj.com/content/331/7521/903.full.print
  - Bland JM, Altman DG. (1994) One- and two-sided tests of significance. BMJ 309, 248. https://www.bmj.com/content/309/6949/248

*Applied Statistics Lab:* Project 1: Continuous Outcome

*Statistical Computing:* Project 2 – Manipulating R Objects

- Reading: Grolemund, Project 2 (Chapters 3 - 6 in book 4 - 8 online)

*Maths Refresher (optional):* Calculus II: Integrals

**Week 3 (21 October)**

*Lecture:* Analysis of variance and 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.
  - Explain how multiple regression can be used to describe, to adjust, and to predict
- Reading:
  - Kirkwood and Stern, Chapters 9-13
  - Bland JM, Altman DG. (1994) Regression towards the mean. 308, 1499. https://www.bmj.com/content/308/6942/1499
  - Bland JM, Altman DG. (1994) Some examples of regression towards the mean. 309, 780. https://www.bmj.com/content/309/6957/780

*Applied Statistics Lab:* Project 1: Continuous Outcome

*Statistical Computing:* Data visualisation

- Reading: Grolemund and Wickham, Chapter 1 in book 3 online

*Maths Refresher (optional):* Calculus III: Differential Equations

**Week 4 (28 October)**

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

- Learning objectives:
  - Define and be able to identify binary data.
  - Understand and compute chi-suqared tests for comparing proportions from population samples.
  - Calculate and interpret odds ratios.
- Reading:
  - Kirkwood and Stern, Chapters 14-17
  - Bland JM, Altman DG. (2000) The odds ratio. 320, 1468. http://www.bmj.com/cgi/content/full/320/7247/1468

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

*Statistical Computing:* Loading and formatting data

- Reading: Grolemund and Wickham, Chapters 7, 8 and 13 in book 10, 11, 16 online

*Maths Refresher (optional):* Linear Algebra I: Introduction to matrices and basic matrix calculus

**Week 5 (4 November)**

*Lecture:* Logistic Regression

- Learning objectives:
  - Use logistic regression to estimate odds ratios, confidence intervals, and p-values.
  - Find and interpret simple odds ratios, and odds ratios adjusted for other variables, based on logistic regression coefficients.
  - Interpret Calculate odds based on results reported from logistic regression.
  - Fit logistic regression models using R.
- Reading:
  - Kirkwood and Stern, Chapters 18-21

*Applied Statistics Lab:* Project 2: Binary Outcome

*Statistical Computing:* Dealing with Data I

- Reading: Grolemund and Wickham, Chapter 9 in book 12 online

*Maths Refresher (optional):* Linear Algebra II: Vector spaces, linear mappings and systems of linear equations

**Week 6 (11 November)**

*Lecture:* Longitudinal data, Poisson Regression

- Learning objectives:
  -
  -
- Reading:
  - Kirkwood and Stern, Chapters 22-25

  - Altman DG, Bland JM. (1998) Time to event (survival) data. 317, 468-469. https://www.bmj.com/content/317/7156/468.1

*Applied Statistics Lab:* Project 2: Binary Outcome

*Statistical Computing:* Dealing with Data II

- Reading: Grolemund and Wickham, Chapter 10 in book 13 online

*Maths Refresher (optional):* Linear Algebra III: Mathematical framework of the linear regression

**Week 7 (18 November)**

*Lecture:* Survival analysis

- Learning objectives:
  - Name the characteristic features of survival data, including censoring.
  - Interpret Kaplan-Meier survival curves and compare survival using log-rank tests.
  - Name the assumptions and principals underlying Cox regression and interpret results.
  - Use R to produce Kaplan-Meier plots and descriptive statistics relating to survival analysis.
  - Use R to fit and check the assumptions of Cox regression models.
- Reading:
  - Kirkwood and Stern, Chapters 26-27
  - Bland JM, Altman DG. (1998) Survival probabilities (the Kaplan-Meier method). 317, 1572. https://www.bmj.com/content/317/7172/1572.full

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

*No Statistical Computing Session this week.*

*Maths Refresher (optional):* Matrix Algebra IV: Eigenvalues and eigenvectors

**Week 8 (25 November)**

Statistical modelling and Maximum likelihood

- Learning objectives:
  - Understand the concept of likelihood of observed data.
  - Recast regression models as maximum likelihood estimation.
  - Use likelihood ratios to construct confidence intervals or compare two groups.
- Reading:
  - Kirkwood and Stern, Chapters 28-29
  - Kirkwood and Stern, Chapters 30-31 (supplemental)

*Applied Statistics Lab:* Project 3: Survival Outcome

*Statistical Computing:* Good programming practises

- Grolemund and Wickham, Chapter 15 in book 19 online

*Maths Refresher (optional):* Linear Algebra V: Principal Component Analysis

**Week 9 (2 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:
  - Kirkwood and Stern: Chapter 33
  - 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
  - Bland JM, Altman DG. (1998) Bayesians and frequentists. 317, 1151. https://www.bmj.com/content/317/7166/1151.1
  - Altman D.G, Bland J.M. (2007) Missing data. BMJ 334, 424. https://www.bmj.com/content/334/7590/424

*Applied Statistics Lab:* Project 3: Survival Outcome

*Statistical Computing:* Advanced R Programming

- Reading: Grolemund, Project 3 (Chapters 7 - 10 in book, 9-12 online).

**Maths Refresher (optional):** Python I: Introduction to Python

**Week 10 (9 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**

*Statistical Computing:* Revision

*Maths Refresher (optional):* Python II: Python computational using pyspark