DR JUAN H KLOPPER
SCHOOL FOR DATA SCIENCE AND COMPUTATIONAL THINKING
STELLENBOSCH UNIVERSITY

Medical and Life Science Statistics using R

A course in statistics for professionals in healthcare and the life sciences using the R language for statistical computation

Introduction

This is a course that teaches the fundamentals of statistical analyses commonly used in healthcare and the life sciences. As a professional in these fields, we rely heavily on the published literature to inform our practice and to stay abreast of new findings. As such, it is of vital importance to be able to interpret the research questions, the study design, the methods employed to conduct the study, and the results. This requires a thorough understanding of the statistics.

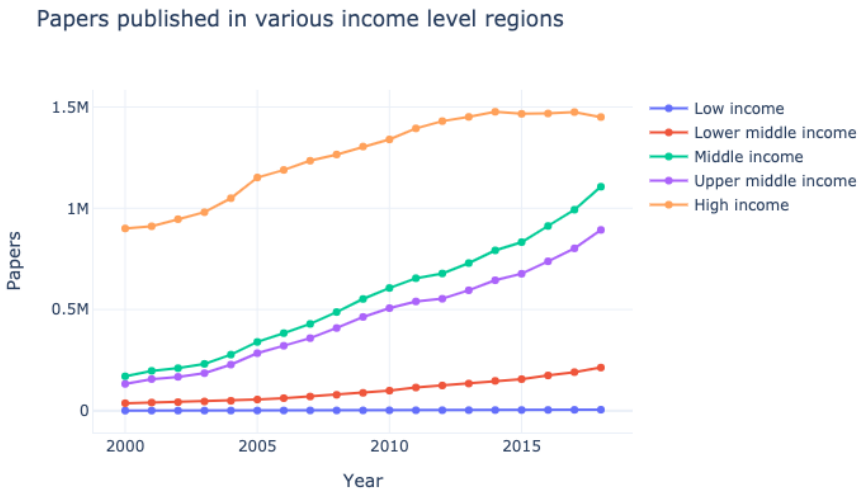
Many of us also have the desire to contribute to research. While it is possible to hand over the design of data capture tools and the analysis of the data to a third party, it takes away from the achievement if we do not have a proper understanding of the analysis of the data. It is all too easy to notice when a researcher presents their results at a meeting or conference, that they lack an understanding of fundamental statistics, having copied from the results of their data analyst or statistician.

While some are fortunate enough to become a part of established research groups, many of our colleagues are not in such privileged positions. While they may have the desire to contribute to research, they have to rely on themselves to bring this to fruition. Many research questions remain unanswered, simply awaiting the empowerment of individuals with the desire to do so.

The aim of this course and others like it, is simple then. It is to *empower*. The ability to understand statistical research analysis and the ability to perform your own statistical tests are very empowering indeed.

While there are many successes in research, we can do so much more. The graph below shows how much more there is still left to do.

It is through research, finding the answers to outstanding question and finding solutions to the problems facing humanity that we move forward. It is by empowering individuals with the desire to contribute, that we make this happen.



Target audience

This course is for any professional in healthcare and the life sciences who wishes to learn the fundamentals of statistical tests commonly used in research. The course develops an intuitive understanding of biostatistics, without the burden of mathematical rigor.

The concepts of statistical tests are developed using the leading open-source software tool for statistics analysis. The R language was developed especially for statistical analysis. With appropriate experience it becomes a very easy to use, yet powerful tool, for data analysis.

Outcomes

By successfully completing this course you will have a deep appreciation for, and understanding of, statistical analysis. This includes an understanding of common statistics such as *p* values, *t* test, confidence intervals, logistic regression, and many more.

At the end of this course, you will know about study design, randomization, data collection, summary statistics, and the creation of graphs and plots. You will know how to conduct the most commonly used statistical tests in the literature and understand how to interpret the results.

This is a course dense with information and knowledge and will require dedication. When successfully completed, you will be empowered to start a lifelong journey of learning and discovery. The abilities of the R language grow on a daily basis and will serve as your constant companion as you develop as a researcher.

Assessment

This course contains many problem sets. Learning how to conduct statistical tests is much like learning a new spoken language. It takes practice and dedicated effort. There are many opportunities throughout the course to showcase your understanding of the topics covered through the use of data analysis that you will have to perform yourself.

System requirements

The R language for statistical computation is software that can be downloaded and installed. A second software tool called RStudio, is a coding environment in which you write your R code and perform your analysis. It is very much like document writing software such as Microsoft Word or Google Docs.

This course requires access to a computer and access to the internet. While material will be available to guide you on installing the required software on your own computer, we will instead use RStudio in the cloud. All the documents and tools will already be created for you. You will only need to open a free account at https://rstudio.cloud . If you wish, you can however do your own installation and work from there.

Course syllabus

This course will cover the following 15 sections:

1. Welcome and introduction
    a. This section provides…
        i. Logistical detail about the course
        ii. Motivation for the course
        iii. A code of conduct
2. An introduction to R and RStudio
    a. This section develops…
        i. A familiarity with the RStudio coding environment
        ii. A familiarity of writing R code through simple arithmetic and the different types of data collections
3. Study types and randomisation
    a. This section provides…
        i. A classification and explanation of different study types
        ii. Information on randomisation in experimental studies (randomised trials)
4. An introduction to statistical terms
    a. This section develops…
        i. An understanding of different data types which forms the basis of deciding on which statistical tests to use
        ii. The concepts of the sample space of variables

iii. An understanding of how to collect data that is amenable to data analysis and that will answer research questions

5. Working with data (part I)
    a. The R language provides for many tools and methods to extract useful sections of data. This section develops…
        i. An understanding of importing data into RStudio
        ii. An understanding of the dimensions and type of data that is imported
        iii. An understanding of how to select specific sections of the data required for analysis
6. Working with data (part II)
    a. This section repeats some of the concepts in the previous section but uses the more advanced and much more powerful *tidyverse* principles. These principles are fairly new to R and have become the standard for working with data.
7. Descriptive statistics
    a. This section develops the following abilities…
        i. Conducting point estimate calculations such as the mean, median, and mode
        ii. Conducting measure of dispersion calculations such as variance, standard deviations, ranges, interquartile ranges, and quantiles
8. Comparative summary statistics
    a. While descriptive statistics gives an overall view of the data, it is important to compare the results between different groups.
    b. This section repeats the concepts for the previous section but shows how to divide the data into groups for comparison.
9. Visualising data
    a. Data visualisation is one of the most important aspects of understanding data.
    b. This section introduces common plots including…
        i. Scatter plots
        ii. Box-and-whisker plots
        iii. Bar charts
        iv. Histograms
10. Sampling and sampling distributions
    a. This section develops an intuitive understanding of inferential statistics. It teaches the fundamentals of patterns in data used to calculate p values and confidence intervals.
    b. It covers the topics of…
        i. The normal distribution
        ii. Sampling from a population
        iii. Sampling distributions…
            1. The *z* distribution
            2. The *t* distribution
            3. The χ^2 distribution
            4. The *F* distribution
11. Parametric comparison of means
    a. This important section develops…

        i. An understanding of inferential statistics
        ii. A final realisation of the concept of the *p* value
        iii. An understanding of Student's *t* test comparing two means
        iv. An understanding of analysis of variance (ANOVA) comparing more than two means
12. Linear models
    a. This section teaches how to conduct…
        i. Correlation tests
        ii. Linear regression
        iii. Logistic regression
13. Assumptions for the use of parametric tests
    a. While the section on the parametric comparison of means provided for analysis using common tests, these cannot always be used.
    b. This section showcases how to analyse data to check whether parametric tests can and should be used.
    c. It includes tests for…
        i. Normality
            1. Visual tests
            2. Statistical tests
        ii. Homogeneity of variance
        iii. Variable data types
14. Nonparametric tests
    a. When the tests in the previous sections show that parametric tests cannot be used, nonparametric alternatives are available.
    b. This section includes…
        i. The Mann-Whitney-U test
        ii. The Kruskal-Wallis test
        iii. Spearman correlation
        iv. Kendal correlation
15. Analysing categorical data
    a. This section showcases the common tests used for categorical variables and includes…
        i. The χ^2 test for proportions
        ii. The χ^2 test for independence
        iii. Fisher's exact test
        iv. The McNemar test

Duration and dates

This course will be presented over five days.

- Day 1
    - Sections 1, 2, 3, and 4
- Day 2
    - Sections 5 and 6
- Day 3

    - Sections 7, 8, and 9
- Day 4
    - Sections 10, 11, and 12
- Day 5
    - Sections 13, 14, and 15

Sessions

The five days will be divided as below. Adequate time will be provided for breaks. The course material will be available beforehand, and it is advised that preparation for the following day be done every evening.

Formal lectures will contain a live session and recorded video.

All assessments must be forwarded by 14H00 every day.

09h00-12h00
Formal lecture
12h00-14h00
Complete assessments
14h00-16h00
Discussion of assessments

Students are required to attend all lectures and submit assessments in order to receive the attendance certificate.