BIO STATISTICS COLLABORATION OF AUSTRALIA

Study Guide

Data Management and Statistical Computing (DMC)

Semester 2, 2024

Prepared by:

Dr Louise Marquart-Wilson
School of Public Health
The University of Queensland

THE UNIVERSITY OF QUEENSLAND

THE UNIVERSITY OF ADELAIDE AUSTRALIA
SUB CRUCE LUMEN

Contents

Data Management and Statistical Computing (DMC)
Semester 2, 2024

Contact details

Dr Michael Waller	**Dr Louise Marquart-Wilson**
School of Public Health Faculty of Medicine The University of Queensland Herston Road, Herston, QLD 4006	School of Public Health Faculty of Medicine The University of Queensland Herston Road, Herston, QLD 4006
Office phone: (07) 3365-5552	*Office phone*: (07) 3346-4687
Email: m.waller@uq.edu.au	*Email:* l.marquart@uq.edu.au

If you have any general BCA queries, please contact: Jaqē Vaughan or Emily Higginson at the BCA Coordinating Office on 02 9562 5076/54 or email bca@sydney.edu.au

Background

The aim of this unit is to provide students with the knowledge and skills required to undertake moderate to high-level data manipulation and management in preparation for statistical analysis of data typically arising in health and medical research.

Context within the program

DMC provides an introduction to R and Stata. These two software packages have been chosen as learning two statistical packages can provide appropriate skills and understanding of statistical programming to aid with learning additional software in the future; and secondly, the availability and accessibility of software for students. R software is a free and extremely flexible software and is increasingly used in practice, and Stata software is a flexible and powerful software with both inbuilt and user-developed packages. Although Stata requires a paid license to access it, the price point is more reasonable than some other software options. This course provides the foundational knowledge on data management and statistical computing required to undertake the coursework for all units in the BCA program and represents the basic level of computing proficiency expected of a practising biostatistician.

Competent usage of more than one statistical package to perform common tasks is essential, as in recent years new statistical methods emerge on different software platforms with increasing frequency. This aspect is also reflected in the delivery of the BCA units that follow DMC, where the methods covered therein have been developed in R but not in Stata, or vice versa.

Prerequisites
None.

Co-requisites
None.

Unit summary

This unit introduces the software packages Stata and R, with the aim of providing a foundation to build upon in further studies and biostatistical career. The aim of this unit is to provide students with the knowledge and skills required to undertake moderate to high level data manipulation and management in preparation for statistical analysis of data typically arising in health and medical research.

The unit is delivered through the Canvas eLearning platform at the University of Sydney.

Unit content will be uploaded to the Canvas e-learning platform in PDF format, including course modules notes, exercises, assignments, coursework solutions and supplementary material (except readings, which for this unit will be mostly found on the recommended textbooks or in publications available through University libraries).

Regular discussions on unit material will take place on Canvas' Discussion Board. Tutorial sessions and zoom drop-in sessions will be held during the semester. Participation in the tutorials is highly encouraged and recommended, and details on schedule and content will be posted on Canvas.

Students are expected to engage with the Canvas eLearning platform often to ensure that they are aware of any notifications relating to the unit.

Workload requirements

The expected workload for this unit is 10-12 hours per week on average, consisting of guided readings, discussion posts, independent study, and completion of assessment tasks. Please note that for students that are new the programming and using statistical software packages, that there may be a higher workload, particularly at the beginning of the semester when the software packages are being introduced. Additional time will be required to become familiar with the software package and also the different coding languages.

Learning Outcomes

At the completion of this unit students should be able to:

1. Be able to undertake data manipulation and management using two major statistical software packages (Stata and R);

2. Be able to appropriately display and summarise data using statistical software;
3. Understand how to check and clean data;
4. Be able to link data files through unique and non-unique identifiers;
5. Have fundamental programming skills for efficient use of statistical software;
6. Understand key principles of confidentiality and privacy in data storage, management and analysis.

Unit content

The unit is divided into 3 modules, summarised in more detail below. Each module will involve approximately 4 weeks of study and includes the following material:

Module 1: Importing and exporting data; recoding and formatting data; labelling variables and values; use of date data, displaying and summarising data. Construction of suitable programming scripts to reproduce results.

Module 2: Graphs, Data management and Statistical Quality Assurance Methods. Includes advanced graphics for production of publication-quality graphs.

Module 3: More Advanced Statistical Computing: Using functions to generate new variables, appending, merging and transposing data; programming skills including macros, loops, user-defined functions and programs.

Study materials for all Modules are downloadable from the Canvas unit site.
Canvas is administered by the BCA Office at the University of Sydney. A Unikey will be provided for you to access Canvas. If you are a continuing student, please use the login details you used previously. For any issues with access, please contact the BCA office at bca@sydney.edu.au.

Assignments and supplementary material, such as datasets, will be available within each Assignment item. Please note that we are not able to post copies of copyright material (journal articles and book extracts)—for these you will have to rely on your home university's library.

Recommended approaches to study

Students should work through each module systematically, following the module notes and any readings referred to, and working through the accompanying exercises. *You will learn a lot more efficiently if you tackle the exercises systematically as you work through the notes.* You should also work through all the computational examples in the notes for yourself on your own computer.

Outline solutions to the exercises in each module (except those to be submitted for assessment, as described below) will be posted online at the midway point of the allocated time for the module. This is intended to encourage you to attempt the

exercises independently (or via the eLearning site), and yet not make you wait too long to see the sketch solutions.

Make the most of this unit by engaging with coordinators and fellow students on the Discussion Board and in Tutorials. These are safe spaces to discuss the course material and related ideas and students are encouraged to make the most of them by engaging in respectful discussion.

Method of communication with coordinator(s)

Questions about administrative aspects or course content can be emailed to the coordinator at m.waller@uq.edu.au or l.marquart@uq.edu.au. Please use "DMC:" in the Subject line of your email to assist in keeping track of our email messages. Coordinator/s will be available to answer questions related to the module notes and practical exercises, and to address any other issues that require clarification.

Please note that instructors are not necessarily available every day of the week and you should expect that it may take a day or so to respond to questions (possibly longer over weekends, public holidays and during breaks).

We strongly recommend that you post content-related questions to the Discussion Board in the unit site. Questions about Assignments should be directed to the coordinator in the first instance to avoid any Academic Honestly issues.

Module descriptions

Each of the three modules for this unit are divided into part A and B (apart from a brief addendum for Module 2 referred to as "Module 2C"). Modules run for a total of 4 weeks each and are scheduled to begin on a Monday.

The due date for submission of required assignments from each module is 11:59pm on the due date as indicated below. Please note that this is the local time in Brisbane (AEST), as this is the delivering institution this semester.

Below is an outline of the study modules, followed by a timetable and assessment description table.

Module 1:

- Reading in and importing datasets in various file formats into R and Stata
- Exporting datasets in Stata, R and other formats
- Exploring datasets descriptively
- Investigating potential data management issues
- Documentation and reproducibility of data manipulations
- Working with dates

Module 2:

- Planning data collection and database design
- Data cleaning and monitoring
- Preparing datasets for analysis
- Investigating relationships between variables
- Missing data basics
- Graphical displays for descriptive analysis

Module 3:

- Data manipulations and wrangling with numerical and string functions
- Merging, appending and reshaping datasets
- Local and global macros, scalars
- Loops
- Creating custom programs
- Saving estimation results for later use

Unit schedule

Semester 2, 2024 starts on Monday the 29th of July, 2024.

Week	Week commencing	Module	Assessment
1	29 July 2024	Module 1A	
2	5 August 2024	Module 1A	
3	12 August 2024	Module 1B	**Assignment 1 available, 16 August**
4	19 August 2024	Module 1B	
5	26 August 2024	Module 2A	
6	2 September 2024	Module 2A	**Assignment 1 due, 2 September**
7	9 September 2024	Module 2B	**Assignment 2 available, 13 September**
8	16 September 2024	Modules 2B & 2C	
Mid-semester break (23 – 27 September)			
9	30 September 2024	Module 3A	**Assignment 2 due, 30 September**
10	7 October 2024	Module 3A	
11	14 October 2024	Module 3B	**Assignment 3 available, 18 October**
12	21 October 2024	Module 3B	
13	28 October 2024		
	4 November 2024		**Assignment 3 due, 4 November**

Assessment

Assessment for DMC includes 3 written assignments worth either 30% or 35% each, as per table below, and will be made available as per timetable above.

Assessments are **due by 11:59pm on the stated day**.

Assessment name	Assessment type	Coverage	Learning objectives	Weight
Assignment 1	Assignment	Module 1	1,2,3	30%
Assignment 2	Assignment	Module 2	1,2,3,5	35%
Assignment 3	Assignment	Modules 1,2 & 3	1,2,3,4,5,6	35%

In general, you are required to submit work typed in Word or similar. We strongly recommend you become familiar with equation typesetting software such as Microsoft's Equation Editor for algebraic work. You may submit neatly handwritten work, however please note that marks will potentially be lost if the solution cannot be understood by the markers due to unclear or illegible writing. Handwritten work should be scanned and collated into a single pdf file and submitted via the eLearning site. See the BCA Assessment Guide for guidelines on acceptable standards for assessable work.

Students are encouraged to discuss relevant topics in the Discussion Board. However, please avoid posting questions relating directly to assessable material. These should be emailed to the Unit Coordinator in the first instance. *Explicit solutions to assessable exercises should not be posted for others to use.* Each student's submitted work must be clearly their own, with anything derived from other students' discussion contributions clearly attributed to the source.

Submission and academic honesty policy

All assessment material should be submitted via the relevant Assessment module in Canvas unless otherwise advised. Turnitin plagiarism detection is applied to all submissions. For detailed information, please see the BCA Assessment Guide, which includes links to the Academic Honesty policies at member universities. Please familiarise yourself with the procedures and policies at your home university. You will need to indicate your compliance with the plagiarism guidelines and policy at your home university.

A special note regarding "contract cheating" sites: Unfortunately, there have been instances in the past of students using such websites to post assignment questions and receive solutions (usually for a fee). We have arrangements with these sites to identify the student posting questions or accessing the solutions, and such students will be referred to and face disciplinary processes at their home university.

Use of ChatGPT or other generative AI tools in assessment tasks
The assessment tasks in this Unit have been designed to be challenging, authentic and complex. Although individual assessment components may provide specific guidance regarding the use of generative AI tools (e.g., ChatGPT), successful completion of these components will require students to critically engage in specific contexts and tasks for which artificial intelligence will provide only limited support and guidance. In all cases, a failure to reference the use of generative AI may constitute student misconduct under the Student Code of Conduct of your University of enrolment. To successfully complete assessment tasks, students will be required to demonstrate detailed comprehension of their written submission independent of AI tools.

Late submission and extension procedure
The standard BCA policy for late penalties for submitted work is a 5% deduction from the earned mark for each day the assessment is late, up to a maximum of 10 days (including weekends and public holidays). Extensions are possible, but these need to be applied for (by email) as early as possible. The Unit Coordinator can approve extensions up to three days; for extensions beyond three days, you must apply to your home university, using their standard procedures.

Learning resources

Access to the following textbooks is recommended, as these texts contain the further reading material referenced in the Module Notes for the Course:

Stata
Juul S, Frydenberg M. An Introduction to Stata for Health Researchers, 5th ed. Stata Press, 2021. To purchase: https://www.stata.com/bookstore/introduction-stata-health-researchers/

R
Wickham H, Cetinkaya-Rundel M. and Grolemund G. R for Data Science (Second Edition). O'Reilly 2023 (freely available online at https://r4ds.hadley.nz/)

Your University Library may have an ebook (Full Text Online) version of the Juul and Frydenberg text; the Wickham, Cetinkaya-Rundel and Grolemund text is freely available at the web link provided. If you have any issues accessing these texts please contact me.

Software requirements and assumed knowledge

No previous computing or programming knowledge is assumed for this course.

However, as pointed out previously, access to the following software packages should be organised ahead of the start of the course:

- **Stata**: version 14 or later (Latest version is v18). Please check with your Program Coordinator whether free licences are available through your host

university. Should you require to purchase a licence, please see BCA Software Guide for more detail.

- **R**: (Latest version is 4.3.1) The Comprehensive R Archive Network (r-project.org)
- **RStudio IDE**: RStudio Desktop - Posit

This is a practical unit designed to develop computing and programming skills in Stata and R; delays in gaining access to the software may impact your ability to complete the unit.

For help with R, please see Learning R in the Student Resources site. For help with Stata, please see Introduction to Stata in the Student Resources site. If you have not yet organised access to these packages, you should do so as soon as possible. This unit requires regular use of the relevant software; delays in gaining access to these packages may impact your ability to complete the course. Information on how to download R and RStudio, and access Stata can be found in the BCA Software Guide.

Required mathematical background

Mathematical proficiency at pre-university level (basic algebra and statistics, such as familiarity with percentiles, interquartile range, standard deviation, minimum, maximum, mean and median).

Feedback

Our feedback to you:

The types of feedback you can expect to receive in this unit are:

- Formal individual feedback on submitted exercises in assignments
- Responses to questions posted on Discussion Board and in Tutorials

Your feedback to us:

One of the formal ways students provide feedback on teaching and their learning experience is through the BCA student evaluation survey at the end of each semester. The feedback is anonymous and provides the BCA with evidence of aspects that students are satisfied with and areas for improvement.

Unit changes, including response to recent student evaluation

DMC was last delivered in Semester 1, 2024. No major unit changes were implemented.

Acknowledgments

In Semester 1, 2020, the R notes were redeveloped by Dr Jennie Louise at The University of Adelaide, and further changes were implemented including incorporation of online tutorials and video content. Further changes were implemented in Semester 2, 2023, including additional modification of the course notes.