BIO STATISTICS COLLABORATION OF AUSTRALIA

Study Guide

Data Management and Statistical Computing (DMC)

Semester 2, 2022

Prepared by:

David Fitzgerald
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, 2022

Contact details

David Fitzgerald

School of Public Health
Faculty of Medicine
The University of Queensland
266 Herston Rd, Herston, QLD, 4006

(07) 3365 5514

d.fitzgerald@sph.uq.edu.au

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

Welcome Letter

Welcome to Data Management and Statistical Computing (DMC). In this unit we will develop statistical computing skills essential for managing and analysing data in health and medicine. This course provides an introduction to R and Stata, with the aim of giving you a foundation to build upon in your further studies and in your biostatistical career. This unit is delivered through the eLearning site at the University of Sydney. All course content other than readings (discussed below) will be uploaded to eLearning, including assignments and supplementary material. Discussions of material will take place on the Discussion Board. There is currently an Introductions thread on Discussion Board; please use this thread to introduce yourself to the rest of the class. This unit requires access to two statistical software packages: R and Stata (detailed shortly). You should organise access to these as soon as possible.

If you have any questions or issues, please contact me by email at the address above. I hope you enjoy the course!

David Fitzgerald
July 2022

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

In this course we will develop statistical computing skills essential for managing and analysing data in health and medicine. This course provides an introduction to R and Stata, with the aim of giving you a foundation to build upon in your further studies and in your biostatistical career.

There are no pre or co-requisites.

Unit summary

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.

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.

Learning Outcomes

At the completion of this unit students should:

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 generally includes the following material:

• Module 1: The basics. Importing and exporting data; recoding and formatting data; labelling variables and values; use of date data, displaying and summarising data.

• 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 loops, arguments and programs/macros.

Study materials for all Modules are downloadable from the eLearning unit site. Assignments and supplementary material, such as datasets will be posted to the unit site. Please note that we will not automatically post out copies of the course notes to all students. If you do wish to have a hard copy of the course notes please request these from the course coordinator.

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.

You are encouraged to post any content-related questions to eLearning, whether they relate directly to a given exercise, or are a request for clarification or further explanation of an area in the notes. The Discussion Board and Tutorials 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. 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 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

As described above, there are 3 modules in this course; each module has been divided into Part A and Part B, each scheduled over a fortnight. Each module sub-section is scheduled to begin on a Monday and conclude on the Sunday of the following week. The due date for submission of required assignments from each module is 2:00 pm (Sydney Time) on the due date. Please note this will change with daylight saving time that NSW, Victoria, ACT and Tasmania use during the warmer months.

Module 1:
The Basics. Importing and exporting data; recoding and formatting data; labelling variables and values; use of date data, displaying and summarising data.
Module 2:
Graphs, Data management and Quality Assessment: Includes advanced graphics for production of publication-quality graphs.
Module 3:
Data Management: Using functions to generate new variables; appending, merging and transposing data; programming skills including loops, arguments and programs/macros.

Unit schedule

Semester 2, 2022 starts on Monday 25 July

Week	Week commencing	Module	Assessment
1	25th July	1A	
2	1 August	1A	
3	8 August	1B	Assignment 1 available 8th August
4	15 August	1B	
5	22 August	2A	**Assignment 1 Due 25th August**
6	29 August	2A	
7	5 September	2B	Assignment 2 available 5th September
8	12 September	2B	
9	19 September	3A	**Assignment 2 Due 22nd September**
Mid Semester Break			
10	3 October	3A	
11	10 October	3B	Assignment 3 Available 10th October
12	17 October	3B	
			Assignment 3 Due 27th October

Assessment

Assessment includes 3 written assignments worth 30 or 35% each. All assignments will be posted on eLearning 2.5 weeks before the due date and must be submitted by 2pm.

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

See the BCA Assessment Guide for guidelines on acceptable standards for assessable work.

We encourage students to discuss the relevant parts of the notes among themselves, via eLearning. However, *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.

Individual feedback will be provided to each student; model solutions will also be provided once all marked assignments have been returned. Summary statistics on results for the entire class will also be provided. Assignments should be submitted via the assignment submission tool on eLearning; if you experience difficulties with this submission method, assignments can be submitted via email.

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.

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

Text Books

It is recommended that you have access to the following textbooks:

- Jull S, Frydenberg M. An Introduction to Stata for Health Researchers, 4th ed. Stata Press, 2014.
- Wickham H, Grolemund G. R for Data Science. O'Reilly 2017. Dalgaard, P. (available online https://r4ds.had.co.nz/)

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

Readings

In addition to the textbooks, various other materials may be set as required or supplementary readings in each module. These cannot be uploaded to eLearning but you should be able to access them through your university's library; further assistance in accessing readings will be given during the course if necessary.

Software requirements and assumed knowledge

You should have access to the following software packages:
• Stata version 12 or later (the latest version is v17)
• R version R64 3.4.2 or later (the latest version is 4.0.2)
• RStudio version 1.3 or later (the latest version is 1.4)

If you have not yet organised access to these packages, you should do so as soon as possible. This is a practical course which 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 Textbook and Software Guide.

For help with R, please see Learning R in the Student Resources site.

Feedback

Our feedback to you:

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

- Formal individual feedback on submitted assignments
- Responses to questions posted on the 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 2022. In Semester 1, 2020, the R notes were redeveloped by Dr Jennie Louise, and further changes were implemented including incorporation of online tutorials and video content. Further changes have been implemented in Semester 1, 2022, including additional modification of the course notes.