LSM3257

Module Guide

AY22/23; Sem 2

Updated 4 Jan 2023

Coordinator: Ian Z.W. Chan | DBS | NUS

About the module

Data analysis has always been crucial for research but is also becoming increasingly important in many jobs and industries. In this module, we **teach you to analyse data in a practical manner**. We start from how to identify what kind of data you have, give you tools to help you decide what analysis to use, and show you how to do the analysis in R. You will have to learn some statistical concepts, but there will be very little math (pinky promise!).

With most people, mentioning "statistics" elicits a visceral, horrified fight or flight response. We hope to change this through this module! **It will be challenging** (many are learning two new and difficult fields—statistics and programming—at the same time) but it will also be **rewarding and beneficial**. By the time you finish this module, you will be well on your way to becoming a **Stats Guru**!

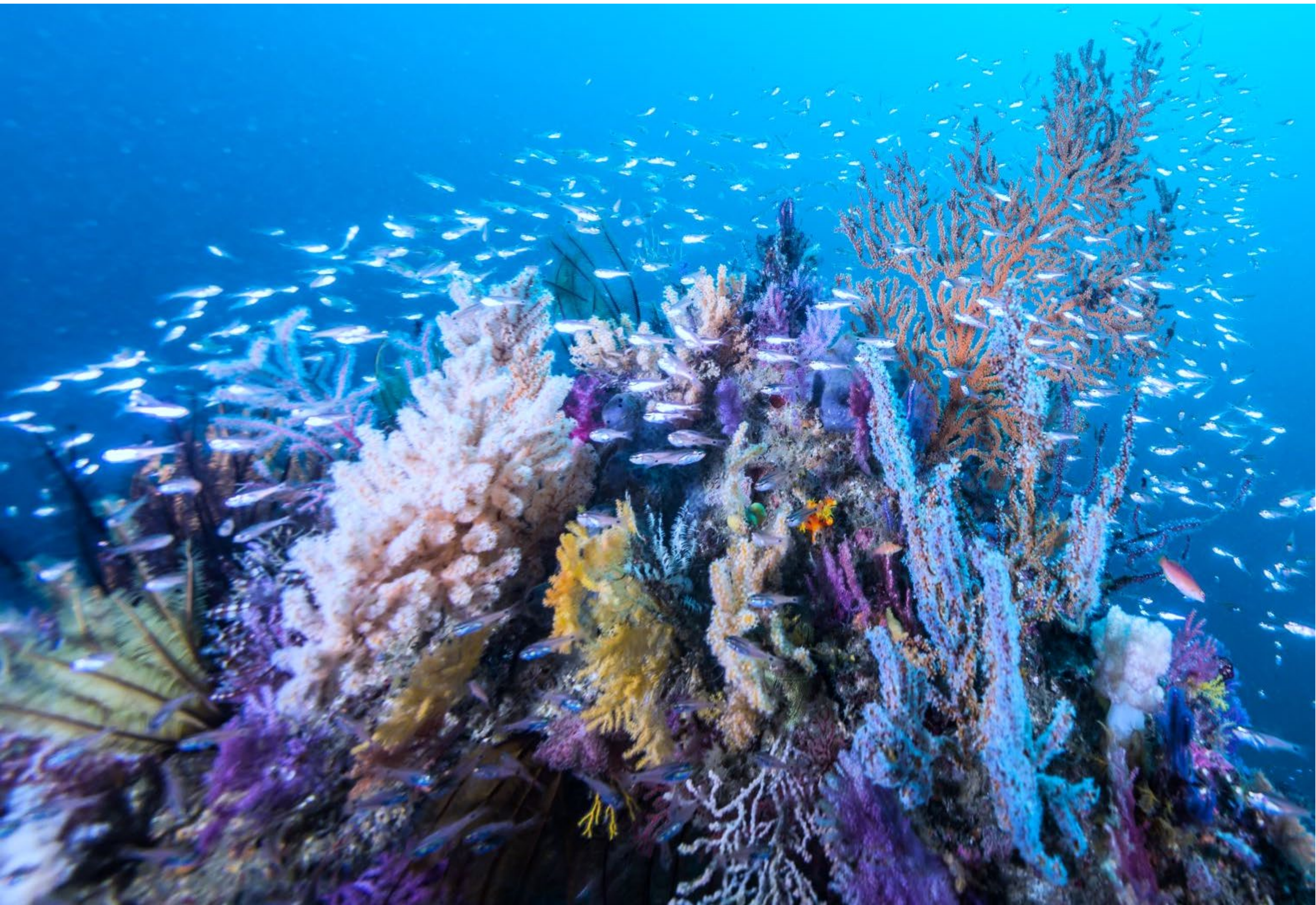
Looking forward to a great semester together!

Your teaching team:


Dr Ian Chan


Ma Jinqi


Su Tingting



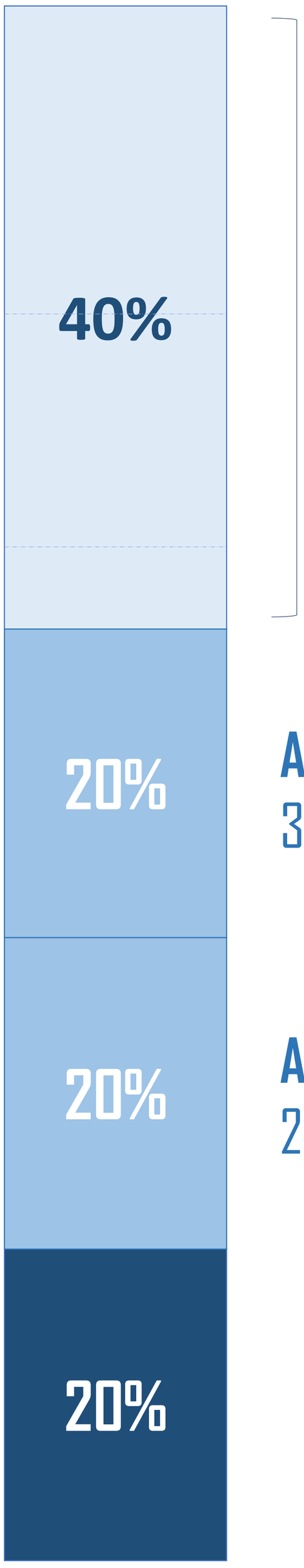
Assesments and Due Dates

Report (Individual): 20%
28 Apr 2023, 2359h

Presentation (Group): 15%
3, 10 Apr 2023†
(Submit slides by 2359h the day before your presentation)

Proposal (Group): 5%
12 Feb 2023, 2359h

Hands-On Exercises (HEx)
Throughout the course, see Schedule and Readings for specific deadlines



Project

Analysis Challenge 2
31 Mar 2023, 2359h

Analysis Challenge 1
24 Feb 2023, 2359h

† Different groups will present on different days. We will allocate slots during the Semester.

<u>Late submission policy</u>: Every day late = minus 10% of the assignment's grade

P.S. If you have problems, contact me early to let me know. I am constrained by University policy but I **<u>want</u>** to help.

Hands-On Exercises (HEx)

You will be given exercises to **attempt in** class and can continue to work on them after the lecture. **You will be required to submit the exercises on Canvas** before the next class. These will be **marked only based on whether you have genuinely attempted the questions** and not whether you get them correct.

WHY? These exercises are designed to help you **learn by doing**! They will **give you an idea of how well you have understood the lectures** and help you **practise what you have learnt in class** during the week so you come to the next class ready to progress some more.

Analysis Challenges (AC)

You will be given two take-home assignments similar to (but more challenging than) the Hands-On Exercises and asked to **perform analysis using sample datasets**. It will be **open book** (including the internet, to be as authentic as possible to the real-world data analyses you will need to do in future) but please **do NOT discuss your answers with one another**.

Just as "there are many ways to skin a cat", there are many ways to write code in R. So as long as your code can run to produce the correct result or analysis, you will be marked correct. You will also NOT be double-penalised for wrong answers (I will explain more in class).

AC1 will cover the material from Weeks 1–6, and AC2 will mainly cover material from Weeks 7 onwards. The questions will be released on Canvas.

WHY? These tests help you to gauge your ability to **perform authentic data analyses** using the frameworks and R code which you have learnt.

Project

In **groups of 3–4** …

Step 1: Find a dataset of your choice. This could be your own work (data you previously collected), data from your supervisor/colleagues if you are doing a project in a lab (e.g. ask your colleagues for datasets which have not yet been analysed) or data from international agencies or published papers in open repositories (e.g. datadryad.org, singstat.gov.sg; there are plenty out there).

Step 2: Formulate a research question. Look at the variables in your dataset and think about what question(s) they can be used to answer. Create a short (≤ 100 characters) and informative title.

Step 3: Write a short proposal in the form of an Abstract (≤ 300 words) describing the background, research question, data to be used and preliminary ideas for analysis. The structure of the abstract would be similar to the abstract of a research article but without results or conclusions. See the sample provided. The proposal should be **submitted as a Group**. I will give you feedback on this and you are free to modify your research question after my feedback.

WHY? This assignment is designed to give you the opportunity to apply the frameworks and R code which you have learnt to **perform authentic data analysis**, **interpret your analyses** and **present your findings** both orally and in a written form—real-world skills which you will need to employ in your work and research.

Step 4: Deliver a 10-minute presentation followed by Q&A. The presentation will consist of: Introduction describing the background to the research and question (1 to 2 slides), M&Ms detailing the dataset(s) used and statistical analyses performed, and Results (with figures). The class will ask questions and give suggestions regarding your analysis. See the sample provided. The presentation will be **delivered as a Group**. Submit your slides the day before your presentation slot.

Step 5: Write a report to **be submitted Individually**. It should consist of the following sections:
(i) Title (≤ 100 characters);
(ii) Abstract (≤ 150 words);
(iii) Introduction (≤ 600 words): background information on what has been done before, what has not been done, the research question that you are answering and why the question is important;
(iv) Methods divided into 2 subsections:
    (a) Data Collection describing how the dataset was collected;
    (b) Data Analysis;
(v) Results; and

(Methods + Results: ≤ 1500 words)

(vi) Discussion (≤ 600 words): to describe the implications of your results and future work;
(vii) References: pick any format (e.g. APA) and use it consistently;
(viii) Figures and Tables, integrated within the article text;
(ix) Supporting Information with figures and tables not typically included in an article (e.g. diagnostic plots) and the R code used for the data management and analysis. This can be submitted as an Appendix to the report or in separate files.

My door and email are always open for consultation!

Project (cont'd)

Project Proposal

CRITERION	SCORE		
A. Research Question	**0**	**5**	**10**
	Not present - Research question is not present - Motivations are not explained		Clearly stated - Research question is identified - Motivations are clearly explained
B. Proposed Statistical Analysis	**0**	**5**	**10**
	Inappropriate - Proposed analysis is not appropriate for the dataset and/or research question		Appropriate - Proposed analysis is appropriate for the dataset and/or research question
C. Writing	**0**	**5**	**10**
	Poorly-planned and Error-filled - Replete with typos and grammatical errors - Exceeds the prescribed word limits		Well-planned and Error-free - Checked for typos and well-written - Adheres to the prescribed word limits

Project Presentation

CRITERION	SCORE			
A. Analysis	**0**	**5**		**10**
	Inaccurate & Inappropriate - Dataset is inappropriate for the research question - Analyses performed are inappropriate for the dataset - Analyses are executed incorrectly - Analysis results are interpreted inappropriately	Mostly Inaccurate & Inappropriate	Mostly Accurate & Appropriate	Appropriate & Accurate - Dataset is appropriate for the research question - Analyses performed are appropriate for dataset - Analysis is executed correctly - Analysis results are interpreted appropriately
B. Presentation Visuals	**0**	**5**		**10**
	Ineffective - Figures do not support points raised - Figures are too small - Slides are too wordy and/or cluttered	Somewhat Effective		Effective - Figures support the points raised - Figures are large and all text is legible - White space is used well
C. Delivery	**0**	**5**		**10**
	Ineffective - Points are unconvincing and hard to follow - Unsure with material and slides, e.g. a lot of "umm" - Presentation overshoots the time limit given - Questions are not addressed, e.g. answers are off-topic or excessively long	Somewhat Effective		Effective - Points are well-supported and communicated simply & clearly - Polished and smooth - Presentation finishes within the time limit given - Questions are addressed well
D. Q&A Participation	**No**	Asked other groups at least one question during the presentation		**Yes**

Optional (for Proposal and Presentation): Team-mate Grading	- Team-based work can be a thorny issue so if you have any problems you cannot resolve, do approach me as early as you can for me to mediate. I would like to work out some arrangement where everyone wins. As a last resort, if things really don't work out, we can do team-mate grading for the whole team so long as at least two team members email me to request for it.

Project Report

CRITERION	SCORE		
A. Research Question	**0**	**5**	**10**
	Inappropriate & Unclear - Motivations behind the research question are not present or clear - Research question is not present or not clearly explained	Somewhat Appropriate & Clear	Appropriate & Clear - Motivations behind the research question are clearly explained - Research question is clearly explained and justified from the Background
B. Statistical Analysis	**0**	**5**	**10**
	Inappropriate & Incorrect - Dataset is inappropriate for the research question - Analyses performed are inappropriate for the dataset - Analyses are executed incorrectly - Analysis is too simple/complex for the dataset and research question	Partially Appropriate & Correct	Appropriate & Correct - Dataset is appropriate for the research question - Analyses performed are appropriate for the dataset - Analyses are executed correctly - Analysis is suitably complex for the dataset and research question
C. Interpretation & Display Items	**0**	**5**	**10**
	Inappropriate & Inadequate - Analysis results are not linked back to the research question - Analyses are interpreted inappropriately, e.g. the wrong conclusions are drawn or the results are over-interpreted - Display items (figures and tables) are absent or do not support the points being raised	Somewhat Appropriate & Adequate	Appropriate & Adequate - Analysis results are linked back to the research question - Analyses are interpreted appropriately and conclusions are justified - Display items (figures and tables) support the points being raised and help the reader to understand the conclusions more intuitively
D. Writing	**0**	**5**	**10**
	Poorly-planned and Error-filled - Replete with typos and grammatical errors - Lines of thought are difficult to follow - Arguments are unconvincing - Exceeds the prescribed word limits	Some Planning but with some Errors	Well-planned and Error-free - Checked for typos and well-written - Lines of thought are clearly explained - Arguments are convincing - Adheres to the prescribed word limits

Schedule and Readings

Classes on **Mondays, 8am-12nn** at **S1A-02-17**

Important Note: please bring your own laptop with R and R Studio installed for this class. I will provide enough power sockets so that everyone will be able to charge their laptop.

Wk	Date	Lectures / **Activities** / *Assignments*	**Readings**
1	9 Jan 23	Lecture 1: Basic Data Handling **Project Preparation** ***HEx 1 (due 15 Jan 2359h)***	R Book: Ch 1 Intro to R: Ch 2 & 9
2	16 Jan 23	Lecture 2: Data Exploration & Visualisation ***HEx 2 (due 25 Jan 2359h)***	R Book: Ch 3.2, 4.1 to 4.4, 5.2, 5.6 & 5.8 Intro to R: Section 6, 7 & 12
3	23 Jan 23	No Class: Public Holiday	
4	30 Jan 23	Lecture 3: Basic Statistical Concepts & Tests ***HEx 3 (due 5 Feb 2359h)***	R Book: Ch 8 Watch: Probability Distributions, Parametric vs. Non-Parametric Tests, Power Analysis
5	6 Feb 23	Lecture 4: Regression ***Project Proposal (due 12 Feb 2359h)***	R Book: Ch 10 Intro to R: Section 11.1–11.5
6	13 Feb 23	Lecture 5: ANOVA & ANCOVA	R Book: Ch 11 & 12
		Recess Week ***Analysis Challenge 1 (due 24 Feb 2359h)***	
7	27 Feb 23	Lecture 6: GLS & LME ***HEx 6 (due 5 Mar 2359h)***	R Book: Ch 19.1–19.9
8	6 Mar 23	Lecture 7: GLM ***HEx 7 (due 12 Mar 2359h)***	R Book: Ch 13.1 to 13.8, 14.1 & 16.1 Intro to R: Section 11.6
9	13 Mar 23	Lecture 8: GLMM & Survival Analysis ***HEx 8 (due 19 Mar 2359h)***	R Book: Ch 29 & 19.10
10	20 Mar 23	Lecture 9: Multivariate Stats, GAM & Bayesian Stats	R Book: Ch 18, 22 & 25 Watch: Multivariate Statistics, Bayesian Stats A & B
11	27 Mar 23	No Class: **Project Consultations** ***Analysis Challenge 2 (due 31 Mar 2359h)***	-
12	3 Apr 23	***Slides (due 2 Apr 2359h)*** ***Project Presentation Seminar***	-
13	10 Apr 23	***Slides (due 2 Apr 2359h)*** ***Project Presentation Seminar***	-
		Reading and Exam Weeks ***Project Report (due 28 Apr 2359h)***	

Course textbooks

Readings are supplementary, not compulsory.

Extremely highly recommended:

1) **R Book** Crawley (2012). The R Book, 2nd Edition. John Wiley & Sons, Ltd. E-book available through NUS Library.

Additional:

2) **Intro to R** Venables et al. (2021). An Introduction to R. R Foundation.
Free download: https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf