UNSW SYDNEY

UNSW SCIENCE

SCHOOL OF MATHS AND STATISTICS

MATH5945

CATEGORICAL DATA ANALYSIS

Term 3, 2019

MATH5945 – Course Outline

Information about the course

Course Authority: Professor Jake Olivier
E-mail: j.olivier@unsw.edu.au
Office: RC-2051

Consultation: Please use e-mail to arrange an appointment.

Credit, Prerequisites, Exclusions:

This course counts for 6 Units of Credit (6UOC). This course is an optional component of the postgraduate coursework Masters programs of the Department of Statistics. Once you have been admitted to the postgraduate program of the Department of Statistics, there are no further prerequisites.

Lectures: There will be a 3-hour lecture block per week during weeks 1-9. Lectures will be Tuesdays 11-2pm in RC-1043.

Labs: In addition to formal lecture material and discussion of problems, we will have computer laboratory work during which theory and methods will be put into practice on various data sets. Computer labs will be available online. The labs will be Tuesdays 3-4pm in RC-G012C.

Moodle: Further information, lecture slides, and other material will be provided via Moodle.

Course aims

This course will give you a solid methodological background in Categorical Data Analysis as a backbone of Applied Statistics. You will learn the theoretical foundations of the most commonly applied techniques in analysing data where the response outcomes are categories. The accompanying explanatory variables may also be categorical or be continuous.

You will learn how to analyse dependencies in various contingency tables. You will learn, in detail, the methodology of the generalised linear model. Within this framework, you will study logistic regression, Poisson regression, logit and log-linear models and the analysis of categorised time-to-event data. You will apply the Generalized Likelihood Ratio testing methodology for choosing the "most suitable" model within a hierarchical set of models.

The classical logistic regression models will be extended to cover polytomous responses (multinomial regression). SAS and R-based computing will feature prominently. At the end of the course you should be able to use all of the above techniques

in your work as an applied statistician, for practical analysis of real datasets containing categorical data.

Relation to other statistics courses

The course is useful in its own right for any applied statistician. It will also be helpful if you study other courses in the Master of Statistics Program such as MATH5826 (Statistical Methods in Epidemiology), MATH5906 (Design and Analysis of Clinical Trials), MATH5916 (Survival Analysis), MATH5895 (Nonparametric Analysis), and MATH5806 (Applied Regression Analysis).

Student Learning Outcomes

- Understand inference for a single proportion and analyse data summarised in contingency tables.
- Able to use the general terminology, notation and concepts in the theory, methods and applications of Categorical Data Analysis. This also includes understanding the different sampling aspects and the relationships between the sampling schemes.
- Able to apply in a rigorous way various aspects of inference for log-linear models and for logistic regression, Poisson regression and logit models. This also includes the ability to do model comparisons and to apply model choice strategies within a given hierarchical set of models.
- Able to analyse categorised Time-to-event Data.
- Able to combine summary statistics from multiple studies in a meta-analysis, assess the presence of publication bias and estimate heterogeneity between studies.
- Able to write simple SAS instructions about data input and output and to code your own simple categorical data analysis procedure using SAS.
- Able to use the common SAS procedures for categorical data analysis such as FREQ, LOGISTIC, CATMOD, and GENMOD. Apply these to the analysis of datasets, interpret the results and draw conclusions. On a few occasions, SAS procedures will be supplemented with `R` functions such as `chisq.test`, `fisher.test` and `glm` as well as functions found in the `MASS` and `metafor` packages.

Relation to graduate attributes

These outcomes are closely related to the graduate attributes "Research, inquiry and analytical thinking abilities" and "Information literacy" (through the computing component of the course).

Teaching strategies underpinning the course

Lecture notes provide a brief reference source for this course. New ideas and skills are first introduced and demonstrated in lectures, then students develop these skills by applying them to specific tasks in additional problems and formal assessment through compulsory assignments and the final exam. Computing skills are developed and practiced in computer labs.

Rationale for learning and teaching strategies

I believe that effective learning is best supported by a climate of inquiry, in which students are actively engaged in the learning process. Hence, besides giving solid methodological background, this course is structured with a strong emphasis on discussing problems and solutions during labs, tutorials and assignments, and students are expected to devote the majority of their class and study time to the solving of such tasks. Questions will be expected and asked during lectures.

Effective learning is achieved when students attend all classes and have prepared effectively for classes by reading through previous lecture notes.

Assessment

Assessment in this course will use problem-solving tasks of a similar form to those practiced in lectures, tutorials and labs, to encourage the development of the core analytical and computing skills underpinning this course and the development of analytical thinking.

Assessment

Assessment task	%	Available	Due
Assignment 1	15	Week 2	Week 4
Assignment 2	15	Week 5	Week 7
Assignment 3	15	Week 8	Week 10
Final exam	55	N/A	TBA

In all assessments, marks will be awarded for correct working and appropriate explanations and not just the final answer.

Assignments

Rationale: Assignments will give students the opportunity to try their hand at more difficult problems requiring more than one line of argument and also introduce them to aspects of the subject which are not explicitly covered in lectures.

Assignments must be YOUR OWN WORK or severe penalties will be incurred.

You should consult the University web page on plagiarism

Late assignments will not be accepted.

Examination

Duration: Two hours. This will involve both theoretical and computational types of questions.

Rationale: The final examination will assess student mastery of the material covered in the lectures and labs.

Further details about the final examination will be available in class closer to exam time.

Additional resources and support

Lab Exercises

Additional exercises will be worked in the lectures. Some problems will be given out for YOU to work alone to enhance mastery of the course. These will not be marked as opposed to the Assignment questions which will be marked. Lab instructions will be provided on Moodle. More precise information will be given in lectures and on Moodle announcements.

Lecture notes

Lecture notes will be provided on Moodle.

Textbooks

Most of the course material can be in the lecture notes. This is a list of additional resources you may find useful.

Agresti A. (2012) *An introduction to categorical data analysis*, 3rd Edition. Wiley.

Dobson AJ, Barnett AG. (2008) *An introduction to generalized linear models*, 3rd edition. CRC Press.

Stokes, M.E., Davis, C.S., Koch, G.G., *Categorical data analysis using SAS*, SAS Press (2012).

SAS manual: procedures FREQ, LOGISTIC, CATMOD, GENMOD, GLIMMIX, SGPLOT

Moodle

Most course materials and computer lab materials will be available on Moodle and you should check regularly for updates. However, some materials may be handed out as a hard copy only.

Computer laboratories

Computer laboratories (RC-M020 and RC-G012) are open 9-5 Monday-Friday on teaching days. RC-M020 has extended teaching hours (usually 8:30am-9pm Monday-Friday, and 9am-5pm Monday-Friday on non-teaching weeks).

Course Evaluation and Development

The School of Mathematics and Statistics evaluates each course each time it is run. We carefully consider the student responses and their implications for course development. It is common practice to discuss informally with students how the course and their mastery of it are progressing.

Administrative matters

Additional Assessment

Information regarding additional assessments are available via the School of Mathematics and Statistics web page at
`http://www.maths.unsw.edu.au/currentstudents/additional-assessment`.

School Rules and Regulations

Fuller details of the general rules regarding attendance, release of marks, special consideration etc are available via the School of Mathematics and Statistics web

page at
http://www.maths.unsw.edu.au/currentstudents/student-services.

Plagiarism and academic honesty

Plagiarism is the presentation of the thoughts or work of another as one's own. Issues you must be aware of regarding plagiarism and the university's policies on academic honesty and plagiarism can be found at
http://www.lc.unsw.edu.au/academic-integrity-plagiarism
and
https://www.gs.unsw.edu.au/policy/documents/studentcodepolicy.pdf

Detailed course schedule

It is intended that the following topics will be covered in the given order. Any variation from this will be indicated by the lecturer. The skeleton lecture notes are the most direct way to learn about the topics. Some additional links to useful information from textbooks are mentioned in the last column of the table below.

Week	Topic	Online	Assessment
1	Statistical inference for a single proportion, 2×2 tables, ordered $2 \times k$ tables		
2	Combining 2×2 tables, Confounding and interactions, McNemar's test, Simpson's paradox and causal inference	Computer lab 1	Assignment 1 available
3	More discrete distributions, exponential family		
4	Generalised linear models (GLM), maximum likelihood		Assignment 1 due 5:00pm
5	Log-linear models, hierarchical/nested models and nesting, model selection	Computer lab 2	Assignment 2 available
6	Logistic regression, Polytomous regression, Conditional logistic regression	Computer lab 3	
7	Logit models and their relation to log-linear models, overdispersion and negative binomial regression, zero-inflated models	Computer lab 4	Assignment 2 due 5:00pm
8	Analysis of discrete categorised time-to event data	Computer lab 5	Assignment 3 available
9	Meta-analysis of count data, Generalised linear mixed models, random effects logistic regression	Computer lab 6	
10			Assignment 3 due 5:00pm