

Methods of Data Analysis II

University of Toronto
Department of Statistical Sciences
STA303H1S Summer 2024

LEC 0101

Instructor: George Stefan
Email: sta303@course.utoronto.ca
Class time/location: MW 1:00–4:00 PM in BA 1160
Office hours: W 4:00–5:00 PM in BA 1160

COURSE OVERVIEW

Course Description: STA303H1 is designed as a follow-up to STA302H1 will primarily focus on extending the linear model to account for categorical outcomes. Topics will include inference for two-way contingency tables; inference for generalized linear models (GLMs); logistic regression; Poisson regression and log-linear models; overdispersion and quasi-likelihood. Later in the course, we will briefly introduce methods for correlated data, such as generalized linear mixed models (GLMMs) and generalized estimating equations (GEEs). Statistical analysis will be conducted using R. Students will be expected to simulate datasets using R, as well as interpret R code and output on assignments and exams.

Content and emphasis of the course is defined by means of all lecture material. Lecture notes will be posted but may not include all examples and details covered during the lectures. Students will also be graded on participation via online poll questions. It is therefore very important to attend all lectures, as there is normally no simple way to make up for missed lectures, aside from perhaps obtaining another student's notes. There will also be many lecture examples using statistical software R.

Prerequisites: STA302H1 or STAC67H3 or STA302H5. **Exclusion:** STAC51H3

Prerequisites are **strictly enforced by the department, not the instructor**. If you do not have the equivalent prerequisites, you will be un-enrolled from the course.

COURSE MATERIALS

Course Content: Lecture slides, R code, and announcements will be posted via Quercus. Please make sure to check Quercus (and your email) regularly so that you do not miss anything. Delivery of this course is in-person; lectures will not be recorded and students are expected to be available to attend all lectures.

Textbook: There are no required textbooks. All evaluations will be conducted based on lecture materials and assignments. However, there are some useful books which can be used as references. The material for the course will mainly be inspired by the first two textbooks listed; these are also a good source of supplementary practice exercises.

1. A. Agresti. *Categorical Data Analysis* (3rd ed.) Wiley, 2012.
2. J. Faraway. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models* (2nd ed.) Chapman & Hall/CRC, 2016.
3. D. Clayton & M. Hills. *Statistical Models in Epidemiology*. Oxford University Press (OUP), 2013.
4. P. McCullagh & J.A. Nelder. *Generalized Linear Models* (2nd ed.) Chapman & Hall/CRC, 1989.
5. P. Diggle, P. Heagerty, K. Liang, S. Zeger. *Analysis of Longitudinal Data* (2nd ed.) OUP, 2013.

Statistical Software: We will be using RStudio to perform statistical analyses. R is a free software that can either be downloaded onto your personal computer or used in the cloud. You may access the U of T cloud version [here](#). Please note that I will be using the most recent version 4.4.1 during my lectures but the cloud version has only been updated to 4.3.2. If you choose to work with R on your personal computer, then installation will be a two-step process:

1. The base R framework is available for download [here](#) for Windows, Mac and Linux operating systems.
2. Next, RStudio is a good integrated development environment to R (makes it simpler to work in R) and can also be downloaded for free [here](#).

If you wish to use R Markdown to compile PDF documents on your own computer, you will have to also install (and update) [MiKTeX](#). This is highly recommended for your assignments in this course, as it will make it much easier to integrate code and R output into your submissions.

COURSE COMPONENTS

Lectures: Lectures will be conducted in-person in BA 1160. Slides and R code will be posted in advance of each lecture but may not contain all details and worked-out examples; thus, students are strongly encouraged to attend class. Lectures will occasionally include poll questions (via Poll Everywhere) which will act as a check on understanding material as we progress. **Students should bring a device which has internet access.** You are also encouraged to bring a laptop or tablet to follow along with live coding sessions.

Office Hours: The instructor will hold a weekly in-person office hour and each of the four TAs will hold a weekly online office hour. The TA office hour schedule will be posted on Quercus once finalized. You are welcome—and in fact, encouraged—to visit office hours if you have any questions about the material. Since this is a summer course, the material can quickly become overwhelming if you fall behind, so I would strongly advise that you do not leave questions until the last minute.

COMMUNICATION

Piazza: We will be using Piazza as an online discussion forum, which can be accessed via the Quercus course page. You can also find our class page [here](#). Students can post anonymously to classmates on Piazza, but the identity of the author of all posts is viewable by instructors and TAs. **All questions about course material should be posted on Piazza or asked during TA/instructor office hours.** The instructor and TAs will monitor the board and will help answer questions but students are encouraged to answer posts and help their fellow classmates.

All posts and conduct on Piazza must remain professional. Posts regarding personal matters such as inquiries about grades, absences, regrade requests, etc. should be communicated via email and NOT be posted on Piazza. Piazza is intended for students to receive support regarding course information and content and thus should be an overall positive and professional environment. Postings that do not align with this environment will be removed.

When to email the instructor: The instructor will only respond to emails of a private or sensitive nature. If you email the instructor with content-related questions, you will be asked to repost your question on Piazza so that the answer may benefit all students. Should you need to email the instructor about a sensitive or personal nature, please use your official “mail.utoronto.ca” email, and include your full name and student number in the text. All course-related emails should be sent to sta303@course.utoronto.ca. Please allow up to 48-96 hours for a reply. Emails will not be monitored on evenings and weekends.

GRADING SCHEME

All students will be evaluated based on the maximum of the following two schemes:

Assessment	Due date/occurring	Grade %	
		Scheme 1	Scheme 2
Participation	Ongoing	5%	5%
Assignment 1	July 18 at 11:59 PM	10%	10%
Midterm	July 22 from 1:00–4:00 PM in EX 310	25%	0%
Assignment 2	August 12 at 11:59 PM	10%	10%
Final Exam	Sometime during August 15-23 (TBD)	50%	75%

Please note that the last day to drop the course without penalty is July 29, 2024.

EVALUATION BREAKDOWN

Assignments: You will be assigned two assignments during the term. The purpose of these assignments is to develop your theoretical and data analysis skills. The assignments will have a strong focus on the use of R, and may involve applying the methods learned during lectures to a dataset. Assignment 1 will be distributed on July 8 and may cover material up to the fourth lecture (July 15). Assignment 2 will be distributed on July 31 and will cover material up to the ninth lecture (August 7). Late submissions will receive an automatic 20% penalty. Solutions will be posted 24 hours after the deadline has passed; **therefore, if the assignment has not been submitted by this time, you will receive a zero.**

Participation: There will be poll questions occasionally interspersed throughout the lectures, which may assess conceptual knowledge or involve quick calculations. They are meant to serve as a knowledge check to ensure that we are all on the same page. In some cases, I will take up the question in class. In order to receive the 5% participation grade, you must record an answer for at least 70% of the poll questions. **Your answer does not have to be correct to receive credit.** The number of poll questions per lecture may vary, so you are encouraged to attend as often as you are able. You will be registered with your U of T email before the first lecture on July 3. Please ensure that you check your email for login instructions.

Midterm: The midterm on July 22 will take place in EX 310 during our usual lecture time and will be based on the content of the first five lectures (up to and including July 17) and Assignment 1. There will be a mix of theory and applied questions. More information on the test (format, number of questions, etc.) will be provided during class, closer to the date.

Final Exam: The final exam will be three hours and will take place sometime during the Faculty of Arts & Science (FAS) exam period, which spans August 15-23. The date will be announced via Quercus once it has been settled. The exam is cumulative (all eleven lectures and both assignments will be covered) and the focus on the material will be uniform over the entire course.

Students should bring a non-programmable calculator for the midterm and the final exam. No aids will be permitted, but the exams will include a formula sheet (which will be posted on Quercus in advance). The formula sheet will include formulas which you may or may not need and are not necessarily an indication of what will appear on the exams.

MISSED ASSESSMENT POLICY

If you experience a prolonged absence due to illness or emergency that prevents you from completing any number of assessments, please contact your College Registrar as soon as possible so that any necessary arrangements can be made. If you do not hand in either of the two assignments, you will receive a zero. **If you miss the midterm you will automatically be graded under Scheme 2 described above**, i.e. the midterm will be worth 0% and the final exam will be worth 75%. You do not need to notify the instructor if you miss the midterm. Keep in mind that the final exam covers nearly twice the amount of material compared to the midterm; it is therefore inadvisable to miss the midterm for any trivial reason, and you have to assume this risk at your own discretion. No further accommodations will be provided.

If you are not able to write your final exam at the scheduled time or if you miss a final exam for reasons outside your control, you may submit a deferred exam petition, which is a request to write your exam at a later time. Please see the [Faculty of Arts and Science Deferred Exam policy](#) for more information.

REGRADE REQUESTS

Mistakes occasionally happen when marking. If you feel there is an issue with the marking of the assignments or midterm, you may request that it be re-marked. The course re-mark policy exists to correct mistakes, and any request should clearly identify the error (for example, a question that was not marked, or a total incorrectly calculated). Requests to correct such mistakes must be sent by email to sta303@course.utoronto.ca. For consideration, any email for a re-mark request:

- Must not be sent within the first 24 hours of the release of the assessment grade;
- Must be received within two weeks of the date that the marks for the assessment became available;
- Must include “STA303 Regrade Request [Assessment Name]” in the subject line of the email;
- Must include your full name and student number;
- Must give a specific, clear, and concise reason for each request, referring to a possible error or omission by the marker. Re-mark requests without a specific reason will not be accepted.

Please note that your entire test/assignment may be re-marked when submitting a remarking request. It is possible that a remark request will result in a lower mark. For the final exam, the re-mark process will be handled by the Faculty of Arts and Science.

ABSENCE DECLARATION

You may use the [ACORN Absence Declaration Tool](#) to support your request for academic consideration in your courses. The tool can be used to declare an absence once per academic term (e.g., the fall term) for a maximum period of seven (7) consecutive calendar days. The ACORN Absence Declaration Tool cannot be used to seek academic consideration for any matters that requires a petition such as missing a final exam or final assessment.

INTELLECTUAL PROPERTY

All course materials are copyrighted. If they are from the textbook, the copyright belongs to the textbook publisher. If they are provided by an instructor (for example, lecture notes, computer code, assignments, tests, solutions) the copyright belongs to the instructor. Distributing materials online or sharing them with anyone in any way is a copyright violation and, in some situations, an academic offence.

ACADEMIC INTEGRITY

Academic integrity is fundamental to learning and scholarship at the University of Toronto. Participating honestly, respectfully, responsibly, and fairly in this academic community ensures that the University of Toronto degree that you earn will be valued as a true indication of your individual academic achievement, and will continue to receive the respect and recognition it deserves. Familiarize yourself with the University of Toronto's Code of Behaviour on Academic Matters available [here](#).

Discussion about lecture materials, textbook concepts and course concepts with your classmates and the teaching team is encouraged, but **it is expected that you work independently on all assessments**. Please note, you may not submit for credit any work that was completed by someone else. This includes, but is not limited to, partially or fully completed code, written answers, answers to problems, communication of solutions, and plagiarism. In particular, you are expected to complete and submit independent work for all assignments and exams. You may discuss lecture materials and general course concepts, but it is expected that you work individually and independently through all STA303 assessments. You may use code provided by your STA303 instructors or TAs without providing a citation. If you use code from any other source, you must provide the source. To protect yourself from potential academic integrity offences, do not share your code and written submissions anywhere (including on social media sites). Discussion or sharing of test questions and/or solutions with others during (or after) the tests is not permitted.

Academic offenses will be taken very seriously and dealt with accordingly. If you have any questions about what is or is not permitted in this course, please do not hesitate to contact your instructor via email or by visiting office hours.

POLICY ON GENERATIVE AI

Students may not use artificial intelligence tools for taking in-person assessments, i.e. the midterm and the final exam. However, these tools may be used to aid in understanding and completing the assignments. Keep in mind that your exams will in part assess your knowledge of the assignment content, so regardless of how you choose to complete them, you should ensure that you understand the material.

ACCESSIBILITY NEEDS

The University of Toronto is committed to accessibility. Students with diverse learning styles and needs are welcome. If you require accommodations for a disability/health condition, or have any accessibility concerns about the course, the classroom, or course materials, please feel free to approach me and/or Accessibility Services as soon as possible: accessibility.services@utoronto.ca or <https://studentlife.utoronto.ca/task/register-with-accessibility-services/>.

CLASS SCHEDULE – TENTATIVE

This is the tentative outline for Summer 2024. Topics may be reduced or additional topics may be added at course instructor's discretion.

Lecture	Content
1 (July 3)	Introduction. Review of discrete distributions, moment-generating functions, CLT, maximum likelihood. Inference for proportions. Score, Wald, likelihood ratio-based confidence intervals and hypothesis tests.
2 (July 8)	Study designs, contingency tables, risk difference and risk ratio. Odds ratio, rate ratio, delta method. Test of independence.
3 (July 10)	Small-sample inference. Fisher's exact test, confounding and interaction. Review of linear regression. Logistic and probit regression.
4 (July 15)	Introduction to generalized linear models (GLMs). Exponential family and iteratively reweighted least squares (IRLS).
5 (July 17)	Poisson regression. Overdispersion and negative binomial regression. Offset terms in GLM. Models for binomial, multinomial and ordinal responses.
July 18	Assignment 1 due.
6 (July 22)	Midterm
7 (July 24)	Log-linear models. Diagnostics and goodness-of-fit for binary response. Stepwise variable selection. Classification and discrimination for logistic regression. Model diagnostics. Cross-validation, ROC curve and bootstrap method.
8 (July 29)	LASSO, ridge and elastic-net. Matched case-control and conditional logistic regression. Overview of methods for dependent data. Deadline to drop course without penalty.
9 (July 31)	Quasi-likelihood and quasi-binomial regression. Linear mixed-effects models (LMM).
10 (August 7)	More on linear mixed-effects models. Repeated measures and longitudinal data.
11 (August 12)	Generalized linear mixed models (GLMM) and generalized additive models (GAMs). Assignment 2 due.
12 (August 13)	Overflow/review lecture.
August 15-23	Final exam period