Methods of Applied Statistics

University of Toronto
Department of Statistical Sciences
STA442H1S Winter 2023

	LEC 0101/2001:	LEC0201:
Class Day/Time:	M 10am-12pm, W 11am-12pm ET in PBB250	M 3pm-5pm, W 3pm-4pm ET in SF1105

COURSE OVERVIEW

Course Description: Advanced topics in statistics and data analysis with emphasis on applications. Study Designs, introduction to generalized linear models, graphical methods, additional topics such as random effects models, designed experiments, model selection, analysis of censored data, introduced as needed in the context of case studies. The course will also introduce applications of modern machine learning methods. Statistical software `R` will be used throughout and will be required for the completion of various assessments during the term. The development of strong written communication skills will be emphasized.

Learning Outcomes: Learning Outcomes: By the end of this course, all students should have a solid understanding of application of modern statistical methods. Students should be prepared to show their understanding of the above through

1. application of methods through problem-solving questions;
2. description and explanation of concepts relating to the mathematical theory of statistical models;
3. recognizing the importance of assumptions and limitations of statistical methods.
4. understanding the proper use more advanced methods for data analysis and their limitations in various scenarios.

Pre-requisites: Pre-requisites are **strictly enforced by the department, not the instructor.** If you do not have the equivalent pre-requisites, you will be un-enrolled from the course. The pre-requisites are STA303H1/(STAC67H3, STAC51H3); CSC108H1/CSC110Y1/CSC120H1/CSC148H1/ CSCA08H3/ CSCA48H3/CSCA20H3/CSC108H5/CSC148H5

COURSE MATERIALS

Course Content: We have a Quercus course page for both sections of this course. All lecture slides, any recordings and materials will be posted on this Quercus course page. Further, any important announcements will also be posted in Quercus. Please make sure to check it regularly.

Textbooks: There are no required text books for this course. The lecture contents will be based on the contents of the following books;

1. Agresti, Alan. *Categorical data analysis*. John Wiley & Sons, 2012.
2. McCullagh, Peter, and John A. Nelder. *Generalized linear models*. Monographs on Statistics and Applied Probability 37. Second Edition. Chapman and Hall.
3. Lawless, Jerald F. *Statistical models and methods for lifetime data*. John Wiley & Sons, 2011.
4. Gelman, Andrew, and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006.

5. James, Gareth, et al. *An introduction to statistical learning.* New York: springer, 2013.

6. Hastie, Trevor, et al. *The elements of statistical learning: data mining, inference, and prediction.* New York: springer, 2009.

7. Diggle, Peter, et al. *Analysis of longitudinal data.* Oxford university press, 2002.

However, these books are optional and you don't need to buy. Rather the focus should be on the lecture materials.

Statistical Software: We will be using the R Statistical Software for performing statistical analyses in this course. R is a free software that can either be downloaded onto your personal computer or used in a cloud environment. We encourage all students to use RStudio through the JupyterHub for University of Toronto. This will allow you to login with your official UofT credentials and use RStudio without the need for a local installation and can be run on any device that has access to an internet connection. More information about using RStudio in JupyterHub will be provided early in the term. R code shown in class will be available on the course page and, along with any additional resources, should be sufficient to complete any assessment involving data analysis.

COURSE COMPONENTS

Lectures: Lectures will be conducted in person. Slides will be available after the class. Class time each week will comprise of a combination of lecturing, and code-along sessions. Where possible, you are encouraged to bring a laptop or tablet to follow along with the code.

Office Hours: Instructor and TAs will hold office hours in a combination of online and in-person formats. The office hour schedule and mode of delivery will be posted on Quercus once finalized. It is recommended that you visit office hours whenever you have a question about the material. It is always important to have material clarified as quickly as possible. Don't wait until the last minute to ask your questions!

Piazza: We will be using the Piazza as an online discussion forum, which can be accessed through the Quercus course page. **All questions about course material should be posted here** 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.

COMMUNICATION

How your instructor will communicate with you: All communication will be made through Quercus announcements or during lectures. Please ensure that you check Quercus regularly so you don't miss anything important.

Where to send content questions: We will be using the Piazza to collect student questions regarding course content, assignments, etc. All questions should be posted here.

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 the content board so 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, include your full name and student number in the text. Send all course related emails to sta442@utoronto.ca. Please allow up to 48-96 hours for a reply. Emails will not be monitored on evenings and weekends.

A note on email and discussion board etiquette: Please make sure that you communicate politely and respectfully with all members of the teaching team and your fellow classmates. Written communica-

tions can sometimes take a tone other than what was intended (e.g. can come off as dismissive, rude or insulting), so make sure you re-read or read out loud your email/post before sending it to make sure it has the tone you intended. For more tips on respectful communication, see professional communication tips. Piazza is a teaching and learning tool and therefore should only be used as such. Any posts that detract from the learning goal of the board will be removed to keep the board a safe space.

GRADING SCHEME

All the students will be evaluated in the following way:

Assessment	Date	Weight
Assignment 1	February 10	14%
Term Test	February 28	25%
Assignment 2	March 10	14%
Assignment 3	April 7	14%
Final Exam	April 11-28	33%

Please note that the last day to drop the course without penalty is March 19, 2023.

EVALUATION BREAKDOWN

Assignment: You will be given three assignment in the term. The purpose of these assignments is to develop your understanding of various statistical methods which will be discussed throughout the course. This will be useful for developing data analysis skills as well as to develop practical understanding of the methods taught in the class. The assignments will have a heavy focus on the use of statistical software (R specifically), and will involve applying the methods learned during lecture to a data set. The format of the assignments will be as follows:

1. use the methods taught in lecture to perform a small data analysis.
2. simulate unique datasets and writing your own functions instead of built in R functions.
3. solve some mathematical problems and explain the procedure

Term Test: The term test will be conducted in person during the scheduled Monday's class time for both the sections (see top of page 1). The test will be approximately 2 hours long. More details will be communicated closer to the test date. The test will cover material from Weeks 1-6.

Final Exam: The details about the final exam will be provided during the last week lectures. For the final exam we will be following standard University of Toronto Schedule. the final exam will be 3 hours in duration and will be scheduled by the Faculty of Arts and Science during the final assessment period.

LATE ASSESSMENT AND EXTENSION REQUEST POLICY

The assessment deadlines may change from the ones stated in the syllabus depending on how the lecture progresses. However, once the deadline(s) has been announced, the students need to submit the assignments

by the deadline. Students will be able to still submit the assignments up to 5 days after the deadline, however, each additional day will be accounted for 20% penalty.
Extreme Situations/Prolonged Illness Extensions: Should a student be experiencing a prolonged illness or other situation that prevents them from turning in their work by the deadline, they should **immediately contact their instructor and College Registrar** to inform them of their situation. They should also submit an Absence Declaration form on ACORN that lists every day during which they were incapacitated and unable to work. Accommodations or further extensions will not be considered without a completed declaration, and will only be considered for extreme circumstances.

Accessibility-Related Extension Requests: Students registered with Accessibility Services should notify the instructor as soon as possible if additional time is needed on assessments that are eligible for extensions. Please **notify the instructor by email of your situation and cc your accessibility advisor** in the process. The instructor will work with the accessibility advisor to determine an appropriate extension for your situation.

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.

Missed Assignment: Missing assessments will receive a 0.

Missed Term Test: If a student is experiencing a serious personal illness or emergency on the date of the test, the student **must declare their absence on ACORN and notify the teaching team via email no later than one week after the date of the test**. A make-up test will then be scheduled at a date and time determined by the instructor. **The format of the make-up is at the discretion of the instructor and may not resemble the format of the original (e.g. an oral exam).**

REGRADE REQUESTS

Regrade requests will be accepted for all assessments. Regrade requests must provide a justification for where there exists a grading error and/or how the work meets the grading rubric. These justifications must further be backed up with concrete references to the course material. All regrade requests will be accepted through a form available on the Quercus course page and will be accepted no later than one week after the grade for that assessment is released. **No regrade requests will be accepted by email or after the 1 week deadline.** The instructor further reserves the right to re-evaluate the assessment in its entirety (i.e. grades can go up, down, or remain unchanged). Please allow a few weeks for regrade requests to be processed by the instructor.

INTELLECTUAL PROPERTY

Course materials provided on Quercus, such as lecture slides, assessments, videos and solutions are the intellectual property of your instructor and are for the use of students currently enrolled in this course only. Synchronous sessions will be recorded and be made available to other students enrolled in the course. **Providing course materials to any person or company outside of the course is unauthorized use and violates copyright**.

ACADEMIC INTEGRITY

The University treats cases of plagiarism and cheating very seriously. It is the students' responsibility for

knowing the content of the University of Toronto's Code of Behaviour on Academic Matters. All suspected cases of academic dishonesty will be investigated following procedures outlined in the above document. If you have questions or concerns about what constitutes appropriate academic behaviour or appropriate research and citation methods, you are expected to seek out additional information on academic integrity from your instructor or from other institutional resources (see http://academicintegrity.utoronto.ca/). Here are a few guidelines regarding academic integrity:

- Being dishonest when reporting an illness or personal emergency to get an extension or accommodation is an academic offence.
- You may consult class notes/lecture slides during assessments, however sharing or discussing questions or answers with other students is an academic offence.
- Students must complete all assessments individually. Working together is not allowed unless otherwise specified.
- Paying anyone else to complete your assessments for you is academic misconduct.
- Completing assessments for another student is academic misconduct.
- Sharing your answers/work/code with others is academic misconduct.
- Using sources external to the course (anything not on Quercus) on an assessment is an academic offence.
- All work that you submit must be your own! You must not copy mathematical derivations, computer output and input, or written answers, etc. from anyone or anywhere else. Unacknowledged copying or unauthorised collaboration will lead to severe disciplinary action, beginning with an automatic grade of zero for all involved and escalating from there. Please read the UofT Policy on Cheating and Plagiarism, and don't plagiarise.

ACCESSIBILITY NEEDS

The University of Toronto offers academic accommodations for students with disabilities. If you require accommodations, or have any accessibility concerns about the course, the classroom, or course materials, please contact Accessibility Services as soon as possible: accessibility.services@utoronto.ca or http://accessibility.utoronto.ca.

TENTATIVE SCHEDULE OF TOPICS

Below is a tentative schedule of topics to be covered in class. The schedule is subject to change and modification.

Week (Dates)	Content
1 (Jan. 9-13)	**Introduction To the Course:** syllabus overview, review of statistical materials, importance of clear and reproducible communication, subjectivity of statistical tools, good data exploration, good communication practices, intro to JupyterHub and RMarkdown. Review of likelihoods and linear models
2 (Jan. 16-20)	**Study Designs and Types of Data:** Experimental and observational study designs. Contingency tables and confounding. Inference on categorical data. Motivation for Generalized Linear Models (GLM).
3 (Jan. 23-27)	**GLM:** Types of GLM. Interpretation of coefficients. Exponential family and iteratively reweighted least squares (IRLS)
4 (Jan. 30 - Feb 3)	**Prediction vs Inference:** Diagnostics of GLMs. Classification and discrimination. ROC curve and AUC. Cross-Validation and Bootstrap
5 (Feb. 6-10)	**Beyond Linear Models:** Tree Based Models. Penalized regression (LASSO, Ridge, Elastic-net). Splines and Generalized Additive Models (GAM). **Assignment 1 Due**
6 (Feb. 13-17)	**Survival Models:** Accelerated Failure Time (AFT) models. The Cox-proportional hazards model.
Feb. 21-24	READING WEEK
7 (Feb. 28 - Mar 3)	**Term Test**. Diagnostics of Survival Models. Martingales.
8 (Mar. 6-10)	**Black Box Models:** Ensemble trees: Random Forest and Boosting. Introduction to XGboost. Introduction to Artificial Neural Networks. Error Back Propagation. **Assignment 2 Due**.
9 (Mar. 13-17)	**Correlated Data:** Linear and Generalized Linear Mixed Models. Generalized Estimating Equations
10 (Mar. 20-24)	**MCMC:** Bayesian Inference and hierarchical models. Markov Chain Monte Carlo algorithms.
11 (Mar. 27-31)	**Bayesian Modeling:** Introduction to Stan. Practical Use of Bayesian Data Analysis.
12 (Apr. 3-7)	**Causality vs Prediction:** Causal Models. Use of Machine Learning in Causal inference. Explaining ML methods. Permutation based importance, LIME and Shap. **Assignment 3 Due**
Apr 11-28	Final assessment period