COURSE COLOUR CODE: GRAPE

MASTER OF BIOTECHNOLOGY PROGRAM

Compulsory Course Component

BTC1859H

DATA SCIENCE IN HEALTH I

Nicholas Mitsakakis

Summer Term, 2022

MASTER OF BIOTECHNOLOGY

UNIVERSITY OF TORONTO MISSISSAUGA

BTC1859H – Data Science in Health I

Course Outline (Summer, 2022)

Class Location:	Maanjiwe nendamowinan Building, Room 3160 (MN-3160)
Class Times:	Monday and Thursday, 6:00-9:00PM & Wednesday, 29-Jun in KN-L1215
Instructor:	**Nicholas Mitsakakis**
Office Location:	Online
Office Hours:	TBC
Contact:	n.mitsakakis@theta.utoronto.ca

Course Description

This course will introduce students to biostatistics and data science. This course is intended for both students new to the area and those with prior training.

The sessions will include lectures and hands-on tutorials that include real-time exercises. Statistical and data analysis methods covered will start with descriptive statistics and basic univariate tests and continue to more advanced regression models. It is key that students are able to identify which methods to apply to what kind of data set, the assumptions of the model and how to interpret the output. Special emphasis in the course will be placed on critical thinking around analytical methods to be used. Problem sets will be focused on the application of statistical modeling to the biological and health sciences. This may include laboratory or clinical data sets. Your defense of your analysis, as well as critiquing the work of others, will require you to draw upon some of your knowledge of biology and the health sciences.

A key component of the course will involve programing in R in order to conduct statistical analysis. Students will have both individual and team assignments to provide practice coding in R, one of the main languages used today in performing statistical analysis. Comfort with R will be helpful in learning other languages in the future in a statistical context. Off-the-shelf software, while more convenient, may not be available in the work environment you find yourself in and certain tests you may need, may not be available in any such software. Thus, learning to code is the best path forward for future practitioners of data science.

Course Objectives

At the conclusion of this course, students should be able to:

1) Identify different types of data and use the appropriate descriptive statistics
2) Identify and apply the correct statistical tests and methods for a given problem
3) Implement the above in the programming language R
4) Understand the assumptions behind the statistical tests

5) Critique the use of statistical tests and methods for a given data set and study objective
6) Understand the unique situations that arise in biology for statistical analysis, in terms of current practices, strengths and weaknesses.

Marking Scheme

The breakdown of the grade for the course will be as follows:

Class Participation	15%
Individual Homework Assignments	25%
Midterm Exam	15%
Team Project	20%
Final Exam	25%
TOTAL	**100%**

The following marking scheme is in effect for the Individual Homework Assignments:

Analysis	30%
Code	20%
Interpretation	30%
Writing and presentation	20%
TOTAL	**100%**

The following marking scheme applies for the Team Project:

Presentation	35%
Report	65%
TOTAL	**100%**

Marking scheme for the report is as follows:

Analysis	30%
Code	20%
Interpretation	25%
Writing and presentation	25%
TOTAL	**100%**

Participation & Online Conduct of Classes

Participation mark is based on attendance and the quality of class participation. The quality of class participation includes appropriate preparation for the material and insightful questions or comments in class discussion. Occasionally quizzes will be handed in and students will have to take them. Marks of these quizzes will count as part of the participation grade. Attendance will also factor into your participation grade.

Quizzes

Students are required to prepare properly for each class by carefully studying the pre-class reading material that is provided in the course syllabus. Quizzes will be occasionally given and students will need to take them; the marks will count as part of the participation grade.

Individual Homework Assignments

A number of assignments where students will need to apply some of the methods they will have learned. These assignments are assigned at the individual level and collaboration among the students as they work on these assignments is prohibited. Students will be given about 1 week to complete each assignment.

Team Project

Students will need to analyse a data set (case study), to create a report and to present the analysis in the class. All team members must have approximately equal presentation time. You should introduce the problem, your methods and analysis. The focus of your talk is on your analysis. If you cover material too quickly or you are cryptic, you will be penalised in your presentation score. Your team's ability to answer questions will be part of your presentation score.

Further Recommendations for Team Project Presentations

To increase illegibility font size should be no less than 18-point font (if using PowerPoint) and a limit of one figure per slide. To control presentation pace, at least one minute must be spent reviewing each slide, and it cannot be cluttered with content. References should be at the bottom of each slide, not at the end of the slide deck.

Midterm Exam

The final exam will be up to 1 hour in duration and will involve multiple choice and short answer questions. The midterm exam will cover material based on approximately the first half of the course. Material in the assigned readings, lectures and tutorials could be tested in the exam. The exam will be conducted online, with more details to follow.

Final Exam

The final exam will be 2 hours in duration and will involve short answer questions. Material in the assigned readings, student presentations, guest lecturers, tutorials and all other material covered in class are all relevant material that could be tested for on the exam. The exam will be conducted online, with more details to follow. **The date and time of the final exam will be Monday, 11-Jul at 6:00-8:00PM.**

Course Material

The materials in the slides and lecture notes will be the main resource for this course. The notes will be mainly referring to and using the following textbooks:

- Primary textbook: *Introductory Statistics with R*, Peter Dalgaard (2nd edition), Springer. This book will be the main textbook of the course and we will follow it closely. However, we will also cover some additional material and topics not covered in this book. The book can be found in an electronic format in the UofT Library. To be referred as "IntroStatsR" in the schedule below.
- Secondary textbook: *OpenIntro Statistics* (3rd Edition), David Diez, Christopher Barr, Mine Cetinkaya-Rundel. We will use this book to cover supplementary material not covered in the main textbook. The book can be freely accessed and downloaded online by https://www.openintro.org/stat/textbook.php?stat_book=os. To be referred as "OpenIntroStats" in the schedule below.

Software

The freely available statistical software R will be used for this course. Additionally the programming editor/environment RStudio will be used. Both R and RStudio can be downloaded freely from the Internet, from the following sites—

- R: https://www.r-project.org
- RStudio: http://www.rstudio.com

After installing R and RStudio, you should see icons in your Applications folder on OS X or All Apps in Windows 10. Starting RStudio (the editor) also starts an R session, so you do not need to run R directly.

The students are required to download and install both R and RStudio prior to the first lecture/tutorial.

Other Resources

IMI Health & Wellness Resources: IMI graduate students have access to a variety of health and wellness resources which we encourage you to use at any time. The IMI Embedded Counsellor is a dedicated counsellor, through the HCC, available to meet with IMI students directly. Call 905-828-5255, share that you are an IMI graduate student, and ask for an appointment. You may also access MySSP (open 24 hours), the Mental Health Wayfinder Tool, Good2Talk and the UTM Health and Counselling Centre at any time.

SCHEDULE OF ACTIVITIES

Unit	Date	Topic	Assignment
1	30-May	**Introductory material, data, study designs, sampling** Introduction to the course, description of structure, format, expectations etc. Introduction to data, ways to describe data, data collections principles, sampling, experimental vs. non-experimental studies, numerical data, categorical data	○ **Tutorial:** Introductory review of R, reviewing basic concepts, simple examples ○ **Reading:** IntroStatsR Ch 1-2; OpenIntroStats Ch 1.1-1.5
2	2-Jun	**Descriptive and exploratory statistics** Descriptive statistics, methods to describe a distribution vs. data (mean, variance/standard deviation, skewness), exploratory analysis, graphical methods (for continuous and categorical data)	○ **Tutorial:** exploratory statistics, plotting functions ○ **Reading:** IntroStatsR Ch 4; OpenIntroStats Ch 1.6-1.7
3	6-Jun	**Review of probability and distributions** Review of probability concepts, conditional probability, random variables, probability distributions (continuous, discrete), sampling from a distribution, normal, Bernoulli, binomial, Poisson; estimation, standard errors, confidence intervals	○ **Tutorial:** R functions for distributions ○ **Reading:** IntroStatsR Ch 3; OpenIntroStats Ch 2-3 ○ **Assignment 1** is handed out (see Assignment description for due date)
4	9-Jun	**Statistical tests for continuous data** Basic hypothesis testing, p-values; t-tests (one, two samples, one, two side), paired, unpaired, equal, non-equal variance, assumptions, non-parametric tests (Wilcoxon)	○ **Tutorial:** R functions for each test, examples, exercises ○ **Reading:** IntroStatsR Ch 5,7; OpenIntroStats Ch 4,5.1-5.3,5.5
5	13-Jun	**Statistical tests for categorical data** Contingency tables, chi-square tests, comparison of proportions, test of independence, Fisher's exact test, tests for count data, assumptions	○ **Tutorial:** R functions for each test, examples, exercises ○ **Reading:** IntroStatsR Ch 8; OpenIntroStats Ch 6 ○ **Assignment 2** is handed out (see Assignment description for due date)

Unit	Date	Topic	Assignment
6	16-Jun	**Correlation and linear regression** Review of correlation (Pearson, Spearman); linear regression; assumptions, mathematical foundation, OLS, remedies when assumptions do not hold; predictions; model fit; residuals; What to report	○ **Tutorial:** Examples of building a regression model; exercises ○ **Reading:** IntroStatsR Ch 6,11; OpenIntroStats Ch 7, 8.1-8.3 ○ **Assignment 3** is handed out (see Assignment description for due date)
7	20-Jun	**Data processing and cleaning** Principals of data science pipeline; Steps of preparing data for analysis; Data preprocessing; Data cleaning; Dealing with Missing data; Formatting; Transforming; Relevant R packages	○ **Midterm Exam** ○ **Tutorial:** Examples of coding such approaches; exercise of using these approaches ○ **Reading:** TBD
8	23-Jun	**Logistic Regression** Context and examples of use in healthcare; Mathematical overview of logistic regression and its assumptions; Common problems and deception in using this tool	○ **Tutorial:** Examples of logistic regression using R; Exercises; ○ **Reading:** IntroStatsR Ch 13; OpenIntroStats Ch 8.4-8.5 ○ **Assignment 4** is handed out (see Assignment for due date) ○ **Team Project:** data set and description is handed out
9	27-Jun	**Sample Size and Statistical Power** Examples of use in healthcare; Mathematical basis for determining appropriate sample size and its assumptions; Statistical power and its assumptions	○ **Tutorial:** Examples using existing R functions and packages; examples using coding from scratch; ○ **Reading:** IntroStatsR Ch 9; OpenIntroStats Ch 5.4 ○ **Assignment 5** is handed out (see Assignment description for due date)
10	29-Jun	**Topics in hypothesis testing and regression analysis** Analysis of Variance, Kruskal Walis, multiple regression, covariate adjustment, co-linearity in regression; multiple testing problem, Bonferroni correction, post-hoc tests ANOVA	○ **Tutorial:** Examples and exercises using R, practical problems ○ **Reading:** IntroStatsR Ch 7; OpenIntroStats Ch 5.5.2-5.5.5
11	4-Jul	**Variable and Model selection** Measures of goodness of fit; comparison of models; variable selection procedures	○ **Tutorial:** Examples of coding such approaches; exercise of using these approaches ○ **Reading:** OpenIntroStars Ch 8.2; additional material TBA
12	6-Jul	**Team Project Presentations** Team presentations of 12 minutes each; Question period following each presentations	**Team Project Report is due**

Procedures & Rules

MISSED TEST(S)/FINAL EXAM: A student that misses a test due to illness must submit a completed University of Toronto Student Medical Certificate (available at: http://www.utm.utoronto.ca/registrar/sites/files/registrar/public/shared/pdfs/medcert_web.pdf) to the Instructor or Program Office (KN-209). Only the University of Toronto Student Medical Certificate will be accepted in support of petitions that cite illness as the reason for the request. Documentation concerning physician examinations must show that the physician was consulted on the day of the test date or immediately after, i.e. the next day. A statement from a physician that merely confirms a report of illness and/or disability made by the student is not acceptable. Documentation citing non-essential, preplanned medical procedures will not be acceptable. All documents must be originals and must be presented in person with a valid UofT student card within 72 hours of missing the test. Beyond 72 hours from the test date, further documentation of continued illness or disability will be required from a physician.

A student that misses a test due to domestic tragedy, at the discretion of the instructor, must provide acceptable documentation validating the explanation for absence.
If a test is missed and the student does not provide acceptable documentation validating the explanation for absence, a grade of “0” may be assigned at the instructor’s discretion.

If a test is missed and validating documentation is accepted the students are expected to write a make-up test. Students must contact the instructor immediately by phone or email to make arrangements.

LATE ASSIGNMENTS: Late assignments will be assigned a late penalty. We recognize that there may be valid reasons for late assignments. In order for these reasons to be accepted without penalty, supporting documentation will often be required. Extensions are given at the discretion of the instructor and require supporting documentation. In all cases, be sure to contact the Instructor before the assignment due date. The Instructor will ask you to report in writing your reasons for lateness and to state a revised due date.

Unless there is a previous communication with the instructor and specific permission has been obtained, a deduction of 10% marks per day will be applied for late work, with a maximum of 4 days. After that submissions will not be acceptable.

Instructors may not grant extensions beyond SGS course deadlines. Such considerations must be negotiated between students, instructors, the program director and SGS.

ACADEMIC MISCONDUCT: Students should note that copying, plagiarizing, or other forms of academic misconduct will not be tolerated. Any student caught engaging in such activities will be subject to academic discipline ranging from a mark of zero on the assignment, test or examination to dismissal from the university as outlined in the School of Graduate Studies academic handbook. Any student abetting or otherwise assisting in such misconduct will also be subject to academic penalties.

Students agree that by taking this course all required papers may be subject to submission for textual similarity review to Turnitin.com for the detection of plagiarism. All submitted papers will be included as source documents in the Turnitin.com reference database solely for the purpose of detecting plagiarism of such papers. The terms that apply to the University's use of the Turnitin.com service are described on the Turnitin.com web site.

Communication

Logging in to your Quercus Course Website

Like many other courses, BTC1859H uses Quercus for its course website. To access the BTC1859H website, or any other Quercus-based course website, go to the UofT portal login page at: **https://q.utoronto.ca** and log in using your UTORid and password. Once you have logged in to the portal using your UTORid and password, look under the **Courses** menu item, where you'll find the link to the BTC1859H course website along with the link to all your other Quercus-based courses.

E-mail Communication with the Course Instructor

At times, the course instructor may decide to send out important course information by e-mail. To that end, all UofT students are required to have a valid UofT e-mail address. You are responsible for ensuring that your UofT e-mail address is set up AND properly entered in the ROSI system.

Forwarding your utoronto.ca e-mail to a Hotmail, Gmail, Yahoo or other type of e-mail account is not advisable. In some cases, messages from utoronto.ca addresses sent to Hotmail, Gmail or Yahoo accounts are filtered as junk mail, which means that e-mails from your course instructor may end up in your spam or junk mail folder.

You are responsible for:

1. Ensuring you have a valid UofT e-mail address, properly entered in the ROSI system
2. Checking your UofT e-mail account on a regular basis.