STAT 305/605: Intro to **Biostatistical Methods** for Health Sciences ("Intro to Biostat")

Syllabus

Fall, 2021

Instructor: Tom Loughin (Call me "Tom")

Contact: Office: NA (All course elements are online) Phone: NA E-mail: tloughin@sfu.ca or through Canvas (for private personal matters only; not for general questions about course or content)

Class Time: All course elements are remote. Lectures are recorded, and you view lecture materials when you want to.

Questions: Instead of office hours, I will hold online live question-and-answer (QA) sessions Mondays, 10:00-11:30 and Wednesdays 16:00-17:00
The QA sessions will be held on Zoom, in a room address posted on Canvas. They are intended to address questions about course and assignment (statistical) content. All students are welcome to attend, and questions will be answered publicly. I will try to remember to record the sessions and post them for later viewing.
If you have a more personal matter to discuss, please contact me privately by e-mail. However, please note that I *cannot* answer course content or homework questions over e-mail. There are too may students to make that possible. I will simply not respond to these questions.

Discussion: I have set up two discussion boards in Canvas.
One will mainly be for *statistics* questions regarding clarifications of notes or assignments. TAs (and I) will check the board periodically, but I also welcome replies from other students. Participation is voluntary and not graded. Answers are generally public.
The other is for *logistics* questions about course procedures and conduct, not relating to statistics. Again, these will be public.

Book: *Principles of Biostatistics, Second Edition* by Marcello Pagano and Kimberlee Gauvreau (2018). CRC Press. (I will call this "PB" for short.)

Prerequisite: One of STAT 201, STAT 203, STAT 205, STAT 270, or BUEC 232.
This course is "second course in statistics." That means you are expected to have had one previous course in statistics, whether it is STAT or a stat-like course in another department. Basically, I expect you to know basic probability, the normal distribution, the t distribution, hypothesis tests and confidence intervals. I will quickly review these concepts at the start of the term.

Computing: The analyses in this course will become too complex to carry out by hand. We will use **R** for all statistical computations in this class. While the course is not about R programming, it will be necessary to learn enough R to demonstrate proficiency in the course material. Homeworks will require R, and questions on tests may use R output and ask minor questions about code. I will provide example programs for each different kind of analysis we learn. Students who have never used it before should begin getting familiar with it *immediately*. (I suggest RStudio as a front-end for R. Students not wanting to deal with a free software download could use Jupyter notebooks through https://sfu.syzygy.ca/. See the STAT Workshop Canvas Page and Assignment 0 for more details on R.)

Logistics: Each week I will post on Canvas some lecture notes in pdf and links to recorded lectures that explain these notes. The lectures will describe and explain the statistical analysis methods, and there will be a few short examples with bits of R code. The lecture notes will have exercises at the end. The exercises relating to the lectures covered each week will be due on Friday of the following week, so there will be a predictable assignment schedule.

Workshop: The STAT Workshop is a drop-in service staffed by TAs who can help you with understanding course material, making R work, figuring out how to solve homeworks, and preparing for exams. The same TAs mark assignments and answer questions on the discussion board. **The Workshop can be accessed using the STAT Workshop button on our Canvas page.**

GRADING and POLICIES

- Exams: There will be two midterms and a final. The midterm dates are **Monday, October 18, and Monday, November 22**, during our nominally scheduled class time from 12:30-14:20. The date for the final will be announced later. All tests are open book, open note.

- Missed Tests: Students who miss an exam due to some emergency must notify me as soon as you know that you will miss the assessment. This usually means before the exam or due time. If you wait until later tell me that you missed something, I will expect some evidence that you were so incapacitated that you could not possibly have used your cell phone to send an e-mail.

  - I don't drop midterms and reweight the final.

    * Some students play a game where they create excuses to skip most of the work from a term and load the weight onto the final exam.

    * In 2016, three students out of 76 missed my first midterm. All three gave me a vague doctor's note, two from the same doctor, and all notes were received 2-5 days after the midterm. Those same three students were the only three students who had not yet turned in a homework. Since I am a statistician, I can tell you that the probability that exactly these three students out of 76 would be the three who missed the exam is 1/70,300 (i.e., p-value is 0.00001 for a test of the null hypothesis that students' previous homework status was not related to missing the midterm). Not likely in a real world.
  - This often fails miserably—it's hard to learn an entire course in a week.
  - Instead, I fill in a predicted value for missed midterms based on *all* other course assessments. This is statistically a better estimate of performance than using only the final.
    * Both methods replace a midterm score with some number that wasn't observed. This is called IMPUTATION and is very commonly done in many biostatistical data sets.
    * If the final exam has a much lower average than the midterm (which it often does with me), using final exam performance penalizes the student who missed the midterm.
    * I will assign an "imputed value" to you based on how well you do on the rest of the marked elements in the class. If you are a median student on average, you will be assigned the median score from the missing element. If you are a top student in other elements, you will get a top score. If you are a bottom student, you will get a bottom score. Thus, your final exam will be worth no more than anyone else's.

- Homework: There will be graded weekly homework assignments, due each week on Fridays. Assignments will consist mostly of applications, many of which will involve R. We will use Crowdmark and Canvas for submitting homework assignments. Instructions will be given later when the first assignment is due.
  - Homeworks are where most of the real learning in the class is likely to take place. I can tell you something, but you will understand it better when you *do* it. Therefore, homeworks must be completed individually.

- Collaboration: Discussions among peers about ideas learned in class are highly encouraged. This often helps in the learning process. Discussion of assignments, however, must be done carefully, because assignments count toward a student's mark. **The work turned in by each student should represent that individual's understanding, ideas, and creativity. By submitting an assignment for marking, you assert that the work contained therein is fundamentally your own.** Specifically:
  - Your R code may be based on course templates or on passive examples available publicly on the internet. R programs may *not* be provided to you by another student or a tutor or a TA or any other person. You won't learn anything if you don't do the coding yourself. I will try to make it mostly painless, but for the

occasional problems, STAT Workshop TAs can help you to get syntax right in an occasional line of code. USE THEM!!!

  - Students may discuss techniques and approaches to solving problems with classmates (only after trying to solve it alone, of course). Students should not get answers to problems from another source, within or outside the class. Again, TAs can help.
  - Students may help one another with specific R coding issues. But large tracts of individuals' code should not be identical, even when it is based on the same class templates. Any coding assignment involves choices with many different options. I expect students' R code to be different enough—and not in trivial ways like object naming—that I can believe that each student put their own thought into it.
  - Violations of this policy constitute plagiarism, which is covered by university policy: https://www.sfu.ca/students/academicintegrity/what-is-it.html
  - I pursue cases of plagiarism vigourously. Cheating is a crappy way to live your life, and it's not fair to the vast majority students who are honest and work hard.
  - Note that anyone who provides answers to another student is also guilty of academic dishonesty, and will be held accountable for these actions.
  - **Please contact me if you have any questions on what is allowed before it becomes a problem!**

- Late Homework: Homework is considered late if it is not received before the specified due time. Homework handed in later is subject to a penalty of 20% of the homework's value for each day it is late. Homeworks will not be accepted after they are marked and returned to the class.

  - I always make allowances for emergencies. Talk to me as soon as possible if a problem arises. Expect to offer some evidence.
  - I will also usually give someone a day or two break if they are just swamped at the moment or an assignment is taking longer than expected. My patience will wear thin, however, if a student asks for multiple breaks. That signals poor time management or refusal to prioritize the course and get assignments done on time.
  - For ongoing illnesses and other longer-term issues, talk to me. We will work something out. Again, expect to offer some evidence.

- Marking Errors: TAs try to be consistent during marking but sometimes mistakes still happen. Also, students sometimes disagree with the proportion of marks granted for partially correct solutions. Dealing with weekly appeals for marking changes is time consuming and often pays little in return. I therefore want to substantially reduce the effort that you, the TA's, and I must put into marking appeals. **As an allowance for mistakes that happen over the course of the term, all students in the course will receive 2% bonus marks added in at the end of the term. In exchange, students will not appeal marking on assignments and midterms.**

(Exceptions will be made for errors in computing the final score on the assignment).

  - If a student truly feels that a huge error has been made, they may forfeit these 2 free percentage points and file an appeal through the STAT Workshop website. The 2 points will NOT be returned to the student, regardless of the outcome of the appeal.

- Scoring (This is subject to change):

  Assignments 15%
  Midterm1 25%
  Midterm2 25%
  Final 35%

- Grades will be assigned as follows:

  - STAT 305:
    A+: 95.00+ %
    A: 90.00–94.99%
    A-: 85.00–89.99%
    B+: 80.00–84.99%
    B: 75.00–79.99%
    B-: 70.00–74.99%
    C+: 66.00–69.99%
    C: 62.00–65.99%
    C-: 58.00–61.99%
    D: 50.00–57.99%
    F: 0–49.99%
  - STAT 605:
    A+: 95.00+ %
    A: 90.00–94.99%
    A-: 85.00–89.99%
    B+: 80.00–84.99%
    B: 75.00–79.99%
    B-: 70.00–74.99%
    C+: 66.00–69.99%
    C: 62.00–65.99%
    C-: 58.00–61.99%
    F: 0–57.99%

  Students who wish to receive a grade for the class must be enrolled in time to be placed on the class roster. Under no circumstances will credit for the class be given to a student who is not enrolled.

ANTICIPATED COURSE OUTLINE (SUBJECT TO CHANGE!)

1. Review (2 weeks)
   (a) Basic concepts of probability
   (b) Distributions, statistics, sampling distributions
   (c) Hypothesis testing, estimation and confidence intervals for means and proportions.
2. Elementary Analysis of Categorical Data (3 weeks)
   (a) 2×2 tables, diagnostic testing, sensitivity and specificity, the relative risk and the odds ratio.
   (b) Contingency Tables, Chi-square and Fisher's test,
   (c) Larger tables, multiple 2x2 tables, Simpson's paradox, Mantel- Haenszel method.
3. Regression Analysis for Numeric Responses (5 weeks)
   (a) Correlation, simple linear regression, estimation and testing for regression coefficients, evaluation of the model.
   (b) Multiple linear regression, inference for regression coefficients, model comparison, confounding and interaction
   (c) Indicator variables, interactions, model selection, prediction, model assumptions and checking.
4. Regression Analysis for Categorical Responses (2 weeks)
   (a) Logistic regression, Odds ratios, inference for regression coefficients
   (b) Model assumptions, case-control studies.
5. Survival Analysis (1 week)
   (a) Life tables, censoring, Kaplan-Meier method, log-rank test.

Learning Outcomes

Upon successful completion of this course, you will:

1. understand the concept of a statistical model and how such models correspond to specific hypotheses or questions,
2. be able to interpret the results of an analysis in relation to the original questions or hypotheses that motivated the analysis,
3. be familiar with data analysis methods commonly used in health sciences and understand the basic limitations of competing methods,
4. be able to use programs in R to complete a variety of basic statistical analyses,

5. understand and be able to critique the analysis methods described in published health research papers,

6. be able to communicate effectively with statistical consultants.