

STAT 201: Statistical Inference for Data Science

Course Information:

- Time and Place
- Description
- Prerequisites for STAT 201
- Textbook
- Expanded Course Description
- Course Software Platforms
- Learning Outcomes
- Teaching Team
- Assessment
 - Course breakdown
 - Group project breakdown
- Schedule
- Policies:
 - Late/Absence
 - Free pass
 - Autograder Policy (Autograder-Policy)
 - Re-grading (Re-grading)
 - Device/Browser (Device/Browser)
 - Exams
 - Missed Final Exam
 - Academic Concession Policy
 - Academic Integrity
 - Plagiarism

Time and Place

- Mondays and Wednesdays 16:00 - 17:30 (Zoom link on canvas)
- Friday office hour (optional): 11:00 - 12:00 (Zoom link on canvas)

Description

Classical and simulation-based techniques for estimation and hypothesis testing, including inference for means and proportions. Emphasis on case studies and real data sets, as well as reproducible and transparent workflows when writing computer scripts for analysis and reports.

Prerequisites for STAT 201

- DSCI 100: Introduction to Data Science

Textbook

We are using an open source textbook available free on the web: ModernDive: Statistical Inference via Data Science (<https://moderndive.com/>) developed by Chester Ismay and Albert Y. Kim.

We may also make some references to OpenIntro Statistics (<https://leanpub.com/openintro-statistics>) by David Diez, Mine Cetinkaya-Rundel, and Christopher Barr which is available by donation.

Expanded Course Description

Data Science skills and tools have been increasingly in demand in the last few years. This course is designed as a method of furthering experience and gaining such skills in the specific area of applications of statistical inference. STAT 201 will explain key concepts through the analysis of data while exploring various and diverse case studies. Special attention will be given to writing computer scripts to create reproducible analyses. In this course, students will be introduced to classical results based on exact and approximate distributional theory in comparison with results derived from simulation-based or resampling approaches, such as permutation and bootstrapping. The resampling framework allows the course to go beyond inference for means and proportions to other parameters needed in practice. This course will be computationally heavy, therefore, students will become well versed in implementing these concepts by writing computer scripts suitable for modern data analysis (e.g. R). There will be an emphasis on reproducible workflows, clear communication of findings and a general understanding of key concepts decontextualized from specific applications.

Course Software Platforms

Students will learn to perform their analysis using the R programming language (<https://cran.r-project.org/>). Worksheets and tutorial problem sets as well as the final project analysis, development, and reports will be done using Jupyter Notebooks (<http://jupyter.org/>). Students will access the worksheets and tutorials in Jupyter Notebooks through Canvas. Students will require a laptop, chromebook or tablet in both lectures and tutorials. If a student does not have their own laptop or chromebook, students may be able to loan a laptop from the UBC library (<https://services.library.ubc.ca/computers-technology/technology-borrowing/>).

Learning Outcomes

By the end of the course, students will be able to:

- Describe real-world examples of questions that can be answered with the statistical inference methods presented in this course (e.g., estimation, hypothesis testing) and apply inference skills and concepts to answer such questions.
- Explain what random and representative samples are and how they can influence estimation.
- Write computer scripts to perform estimation and hypothesis testing via simulation-based inference approaches, as well as by applying results from exact and approximate distributional theory.
- Interpret and explain results from confidence intervals and hypothesis tests.
- Compare the application of simulation-based inference approaches with the application of exact and approximate distributional results.
- Effectively visualize point estimates and different measures of uncertainty (e.g., confidence intervals, standard errors) by writing computer scripts.
- Discuss the impact of type I & II errors as well as responsible use and reporting of p-values on hypothesis tests.
- Explain estimator bias and uncertainty, and write a computer script to calculate it.
- Discuss how an estimator's bias arises (e.g., sample bias, study design), and its implications in statistical inference.

- Perform all aspects of a statistical analysis (from data consumption to reporting) using reproducible and transparent computer scripts.

You can also check the learning outcomes for each week of the course here (<https://github.com/UBC-STAT/stat-201-student/blob/master/lecture-learning-outcomes.md>).

Teaching Team

Position	Name	e-mail
Instructor	Vincenzo Coia	vincen.coia@stat.ubc.ca (mailto:vincen.coia@stat.ubc.ca)
Instructor	Rodolfo Lourenzutti	lourenzutti@stat.ubc.ca (mailto:lourenzutti@stat.ubc.ca)
TA	Creagh Briercliffe	(use Piazza)
TA	Ela Bandari	(use Piazza)

Assessment

In each class (lecture and tutorial) there will be an assignment. **Lecture and tutorial worksheet due dates are posted on Canvas.** To open the assignment, click the link (e.g. `worksheet_01`) from Canvas. To submit your assignment, just make sure your work is saved (File -> Save and Checkpoint to be sure) **on our server** (i.e., using the link from Canvas) before the deadline. Our server will automatically snapshot at the due date/time.

Course breakdown

Deliverable	Percent Grade
Lecture worksheets	5
Tutorial problem sets	15
Group project	20
Two Midterms	30
Final exam	30

Test Dates:

- *Midterm 1*: Monday, Feb 8 at 4pm
- *Midterm 2*: Monday, Mar 15 at 4pm

Group project breakdown

Deliverable	Percent Grade
Proposal	3
Peer review	2
Final report	10

Deliverable	Percent Grade
Team work	5

Schedule

Lectures are held on Mondays. Tutorials are held on Wednesdays and build on the concepts learned in lecture. - **MD** refers to ModernDive (<https://moderndive.com/>) - **OIS** refers to OpenIntro Statistics (<https://www.openintro.org/download.php?file=os3&referrer=/stat/textbook.php>)

Week	Instructor	Primary TA	Topic	Readings
1 (https://github.com/UBC-STAT/stat-201-student/blob/master/lecture-learning-outcomes.md#week-1-introduction-to-statistical-inference-and-sampling) (Jan 11)	Vincenzo	Creagh	Introduction to Statistical Inference and Sampling	MD: Chapter 7 - 7.2.3 Sampling (https://moderndive.com/7-sampling.html)
2 (https://github.com/UBC-STAT/stat-201-student/blob/master/lecture-learning-outcomes.md#week-2-populations-and-sampling) (Jan 18)	Vincenzo	Ela	Populations and Sampling	MD Chapter 7.2.4 onwards (to the end of chapter 7) Sampling (https://moderndive.com/7-sampling.html#sampling-framework) and OIS: Chapter 3.4 Random Variables
3 (https://github.com/UBC-STAT/stat-201-student/blob/master/lecture-learning-outcomes.md#week-3-bootstrapping-and-its-relationship-to-the-sampling-distribution) (Jan 25)	Vincenzo	Creagh	Bootstrapping and its Relationship to the Sampling Distribution	MD: Chapter 8 - 8.2 Bootstrapping (https://moderndive.com/8-confidence-intervals.html) and MD: Chapter 8.7.1 Comparing bootstrap and sampling distributions (https://moderndive.com/8-confidence-intervals.html#bootstrap-vs-sampling)
4 (https://github.com/UBC-STAT/stat-201-student/blob/master/lecture-learning-outcomes.md#week-4-confidence-intervals-via-bootstrapping) (Feb 1)	Rodolfo	Ela	Confidence Intervals via Bootstrapping	MD: Chapter 8.3- 8.7 Confidence Intervals (https://moderndive.com/8-confidence-intervals.html#ci-build-up)

Week	Instructor	Primary TA	Topic	Readings
5 (https://github.com/UBC-STAT/stat-201-student/blob/master/lecture-learning-outcomes.md#week-5-mid-term-1-and-preparation-for-projects) (Feb 8)	Vincenzo	NA	Mid-term #1 and Preparation for Projects	
READING WEEK				
6 (https://github.com/UBC-STAT/stat-201-student/blob/master/lecture-learning-outcomes.md#week-6-hypothesis-testing-via-simulationrandomization) (Feb 22)	Rodolfo	Ela	Hypothesis Testing via Simulation/Randomization	MD: Chapter 9 - 9.4.1 Hypothesis Testing (https://moderndive.com/9-hypothesis-testing.html)
7 (https://github.com/UBC-STAT/stat-201-student/blob/master/lecture-learning-outcomes.md#week-7-confidence-intervals-of-means-and-proportions-based-on-the-assumption-of-normality-or-the-central-limit-theorem) (Mar 1)	Rodolfo	Creagh	Confidence Intervals Based on the Assumption of Normality or the Central Limit Theorem	OIS: Chapter 5.2 Confidence intervals for a proportion; MD: Chapter 7.5.2 Central Limit Theorem (https://moderndive.com/7-sampling.html#sampling-conclusion-central-limit-theorem); MD: Appendix A Normal Distribution (https://moderndive.com/A-appendixA.html#appendix-normal-curve); and MD: Chapter 8.7.2 Theory-based confidence intervals (https://moderndive.com/8-confidence-intervals.html#theory-ci)
8 (https://github.com/UBC-STAT/stat-201-student/blob/master/lecture-learning-outcomes.md#week-8-classical-tests-based-on-normal-and-t-distributions) (Mar 8)	Rodolfo	Ela	Classical Tests Based on Normal and t-Distributions	MD: Chapter 9.6.1 Theory-based hypothesis tests (https://moderndive.com/9-hypothesis-testing.html#theory-hypo)

Week	Instructor	Primary TA	Topic	Readings
9 (Mar 15)	NA	NA	Mid-term #2 and Introduction of Methods Based on Exact and Asymptotic Distributional Theory	OIS Chapter 3.1.2 Probability
10 (https://github.com/UBC-STAT/stat-201-student/blob/master/lecture-learning-outcomes.md#week-10-errors-in-inference) (Mar 22)	Rodolfo	Creagh	Errors in Inference	MD: Chapter 9.4.2 Types of errors (https://moderndive.com/9-hypothesis-testing.html#types-of-errors) and MD: Chapter 9.6.3 Problems with p-values (https://moderndive.com/9-hypothesis-testing.html#problems-with-p-values)
11 (https://github.com/UBC-STAT/stat-201-student/blob/master/lecture-learning-outcomes.md#week-11-beyond-two-group-comparisons) (Mar 29)	Vincenzo	TBD	Beyond two-group comparisons	
12 (April 5)	NA	NA	Project Week	
Exam period	NA	NA	Final Exam	

Policies

Late/Absence

Regular attendance to lecture and tutorials is expected of students. Students who are unavoidably absent because of illness or other reasons should inform the instructor(s) of the course as soon as possible, preferably, prior to the start of the lecture/tutorial. There will be no make-up quizzes (midterms). Students who miss quizzes 1 or 2 or an assignment and want to request an Academic Concession need to contact the Instructor as soon as possible and provide a self-declaration form (https://canvas.ubc.ca/courses/65765/files/12177943?module_item_id=2778150). Failing to present a declaration may result in a grade of zero.

Late lecture and tutorial worksheets will receive a grade of 0. For other assessments, a late submission is defined as any work submitted after the deadline. For a late submission, the student will receive a 50% deduction of their grade for the first occurrence. Hence a maximum attainable grade for the first piece of work submitted late is 50%. Any additional pieces of work that are submitted late will receive a grade of 0 for subsequent occurrences.

Free-pass

Due to the extreme challenging circumstances we are facing, we are going to drop the lowest grade worksheet and tutorial (one of each).

Autograder Policy

Many of the questions in assignments are graded automatically by software. The grading computer has exactly the same hardware setup as the server that students work on. No assignment, when completed, should take longer than 5 minutes to run on the server. The autograder will automatically stop (time out) for each student assignment after a maximum of 5 minutes; **any ungraded questions at that point will receive a score of 0.**

Furthermore, students are responsible for making sure their assignments are *reproducible*, and run from beginning to end on the autograding computer. In particular, **please ensure that any data that needs to be downloaded is done so by the assignment notebook with the correct filename to the correct folder.** A common mistake is to manually download data when working on the assignment, making the autograder unable to find the data and often resulting in an assignment grade of 0.

In short: whatever grade the autograder returns after 5 minutes (assuming the teaching team did not make an error) is the grade that will be assigned.

Re-grading

If you have concerns about the way your work was graded, please contact the TA who graded it within one week of having the grade returned to you. After this one-week window, we may deny your request for re-evaluation. Also, please keep in mind that your grade may go up or down as a result of re-grading.

Device/Browser

Students are responsible for using a device and browser compatible with all functionality of Canvas. Chrome or Firefox browsers are recommended; Safari has had issues with Canvas quizzes in the past.

Exams

In this course, two 45-minute mid-terms will be administered at the start of the lecture in Weeks 5 and 9. Besides, there will be a two-hour final exam. Notwithstanding the grading weights above, students must pass the final exam to pass the course. All the exams will be open-book; however, you are not allowed to communicate with another person during the exams.

Missed Final Exam

Students who miss the final exam must report to their faculty advising office within 72 hours of the missed exam, and must supply supporting documentation. Only your faculty advising office can grant deferred standing in a course. You must also notify your instructor prior to (if possible) or immediately after the exam. Your instructor will let you know when you are expected to write your deferred exam. Deferred exams will ONLY be provided to students who have applied for and received deferred standing from their faculty.

Academic Concession Policy

Please see UBC's concession policy (<http://www.calendar.ubc.ca/vancouver/index.cfm?tree=3,329,0,0>) for detailed information on dealing with missed coursework, quizzes, and exams under circumstances of an acute and unanticipated nature.

Academic Integrity

The academic enterprise is founded on honesty, civility, and integrity. As members of this enterprise, all students are expected to know, understand, and follow the codes of conduct regarding academic integrity. At the most basic level, this means submitting only original work done by you and acknowledging all sources of information or ideas and attributing them to others as required. This also means you should not cheat, copy, or mislead others about what is your work. Violations of academic integrity (i.e., misconduct) lead to the breakdown of the academic enterprise, and therefore serious consequences arise and harsh sanctions are imposed. For example, incidences of plagiarism or cheating may result in a mark of zero on the assignment or exam and more serious consequences may apply if the matter is referred to the President's Advisory Committee on Student Discipline. Careful records are kept in order to monitor and prevent recurrences.

A more detailed description of academic integrity, including the University's policies and procedures, may be found in the Academic Calendar at <http://calendar.ubc.ca/vancouver/index.cfm?tree=3,54,111,0> (<http://calendar.ubc.ca/vancouver/index.cfm?tree=3,54,111,0>).

Plagiarism

Students must correctly cite any code or text that has been authored by someone else or by the student themselves for other assignments. Cases of plagiarism may include, but are not limited to:

- the reproduction (copying and pasting) of code or text with none or minimal reformatting (e.g., changing the name of the variables)
- the translation of an algorithm or a script from a language to another
- the generation of code by automatic code-generation software

An "adequate acknowledgement" requires a detailed identification of the (parts of the) code or text reused and a full citation of the original source code that has been reused.

The above attribution policy applies only to assignments. **No code or text may be copied (with or without attribution) from any source during a quiz or exam. Answers must always be in your own words. At a minimum, copying will result in a grade of 0 for the related question.**

Repeated plagiarism of any form could result in larger penalties, including failure of the course.

Attribution

Parts of this syllabus (particularly the policies) have been copied and derived from the UBC MDS Policies (<https://ubc-mds.github.io/policies/>).