Advanced Statistics for Biological Sciences

BIOL422 | Winter | 2022

Instructor: Dr. Pedro Peres-Neto, Professor, Department of Biology

Please contact me via Moodle (using its messaging system). But, if Moodle is down and you have an emergency, then contact me at pedro.peres-neto@concordia.ca with BIOL-422 as the subject. For contacting TAs, see below.

Teaching strategy and access to course material and information

Remote teaching strategy until we return to in-person: A course in Biostatistics is well suited for remote teaching. For the time being, we will use a mix of Synchronous & Asynchronous learning for lectures. This means that some lectures will happen in real-time whereas others not. All lectures will be recorded and made available regardless whether they were synchronous or asynchronous by the regular lecture schedule. Tutorials will be synchronous online (and not recorded) but given that most of you you have basic knowledge in R, you can work on the tutorials on your own and ask questions in the Forum.

Access to course material: Most course material (lecture videos, slides, lecture notes, etc) will be posted every week in a WebBook that you can have access via Moodle or using its direct web link. The WebBook structures the entire course. Forums and assignments will be posted in Moodle as per Concordia policy.

.info = SUCCESS

CHECK MOODLE, OUR WebBook AND EMAILS ROUTINELY
PAY DETAILED ATTENTION TO INFORMATION

Lectures & Computer Labs are via Zoom until in-person is resumed
Zoom links will be posted in Moodle prior to the lecture or lab session

Lectures: Tuesday & Thursday 11:45am-1:00pm	Professor Pedro Peres-Neto (Instructor)
Lab tutorials	**TAs (doctoral students)**
Lab section 101 or 201: Tuesday 14:00-16:00	Aliénor Stahl
Lab section 102 or 202: Tuesday 16:15-18:15	Gabrielle Rimok

????? QUESTIONS ?????

[Moodle FORUM]: We expect that you ask general questions first using the Moodle Forum. Given that we are in a remote mode, we need (more than ever) to generate an environment of cooperation. Students should feel free to answer Forums to assist other students. The instructor will also answer questions via the Forum
[Instructor]: Office hours Tuesdays & Thursdays 10:30AM to 11:30AM - by appointment only; send message via Moodle.
[TA]: Each TA will set up their own schedule for office hours with students in their lab sections.

If you have questions that you rather not post in the Forum:
If related to lecture, course material or any other issue: contact the instructor via Moodle
If related to lab material: contact your TA via their email

Course Description: This course is designed to teach students modern statistical and data science tools that support answering research questions in biological sciences. Examples and applications will be drawn from a wide range of biological fields including cell biology, ecology, environmental sciences, epidemiology, genetics, molecular biology and genomics. Lectures will present and explain technical concepts within an applied context whereas computer labs will provide hands-on data analyses using the R software environment for statistical computing and graphics.

Objectives: Upon successful completion of the course, students will be able to: express scientific questions in a statistical manner; decide which techniques are better suited for different types of biological problems; report statistical results in an effective manner; adapt the knowledge and practice they learned to new biological questions. Formulas (formulae) are presented so that students gain intuition about their nature, but their memorization is not required in exams and in tutorials.

Lectures: The teaching strategy in lectures is to use multiple examples from different fields of biology so that students can gain experience on the technical and application aspects of a multitude of advanced statistical methods used in Biology. We also use simulated data to demonstrate some critical theoretical concepts before applying to real data.

Computer-based labs (tutorials): The application of statistical methods covered in lectures will be practiced using data examples extracted from existing sources (e.g., published studies, data repositories). Applications will be based on the software environment R for statistical computing and graphics.

Statistical software: We will use R, which is a free software environment for statistical computing and graphics. R has become the de facto standard platform for performing statistical analyses in biology. Knowledge of R has now become a skill required in the job markets of many disciplines, including Biology.

Assessment calendar in a glance (details are provided in the next pages)

January	February	March
11	10 **REPORT 1**	15 **ESSAY 4**
13	15	17 **REPORT 3**
18	17 **ESSAY 3**	22
20 **ESSAY 1**	22	24 **ESSAY 5**
25	24 **REPORT 2**	29
27	**March**	31 **REPORT 4**
February	01 mid-term break	**April**
01 submit review subject	03 mid-term break	05
03 **ESSAY 2**	08 **MIDTERM 1**	07 **MIDTERM 2**
08	10	12 **LITERATURE REVIEW**

EVALUATION IN A GLANCE
(details given below)

5 **SHORT ESSAYS** (3% each) = 15%

4 **REPORTS** (5% each) posted 2 weeks prior to the deadline = 20%

2 **MIDTERM Exams** (20% & 25%) = 45%;
Midterm 1 (75 minutes); midterm 2 (48 hours)

1 **LITERATURE REVIEW** = 20%

DEADLINES

You should also become familiar with the general Concordia calendar:
https://www.concordia.ca/students/undergraduate/undergraduate-academic-dates.html

Assessment details – READ WITH ATTENTION

(Five) SHORT ESSAYS (3% each x 5 =15%). Deadlines are at 6PM on the day indicated in the calendar (previous page). All 5 essays will be posted on Moodle in the first week of classes so that you can work on your own pace. They are to be sent in a word file to your TA. No excuse for delays other than medical or a short-term absence for medical accommodation (see below) is accepted.

(Four) REPORTS (5% each x 4 = 20%). Deadlines are at 6PM on the day indicated in the calendar (previous page). Report instructions will be posted in Moodle two weeks prior to their deadlines. Either existent data or synthetic data (i.e., made up by you) will be used and R code will have to be produced to analyze the generated data. Specific formats for the reports will be provided in the instructions. They should be sent in a word file (report) and csv file (data) to your TA. If an excused issue (example: illness) is incurred, an extension for the report may be granted by the instructor (Dr. Peres-Neto and not your TA). Such extensions are expected to be requested at least 24 hours before the report is due (TIP - avoid start working on the report too close to the deadline). You must send a written request to Dr. Peres-Neto via his Concordia email with a medical note attached; or a short-term absence for medical accommodation (see below). The following grading penalties will be applied to late reports without justification: 1 day or less - 10%; 1-2 days - 20%; 2-3 days - 35%; 3-4 days - 50%; 4-5 days - 70%; more than 5 days - 90%.

MIDTERM 1 (NOT OPEN BOOK) (20%). Exam 1 is on March 8. Students have a total of 75 minutes to complete. Exams contain multiple choice and short essay questions related to lectures and additional material distributed during the term. No excuse for not taking the exam other than medical is accepted. If you miss the midterm for medical reasons (including the short-term absence), the grade will be accumulated to your midterm 2.

MIDTERM 2 (OPEN BOOK but students can't consult anyone other than their own material) (25%). On April 7 (1pm), the exam will be posted on Moodle and students will have about 48 hours to solve it (from April 7 1pm to April 9 2pm). It will involve a mix of essay questions and statistical problems involving R code. The goal of this assignment is to make sure you are following the material closely during the session and paying attention to R coding and problem solving. The assignment is made so that you can realistically solve it within about 8 hours. **That said, this is only realistic if you prepare yourself in advance by studying and reviewing the tutorials prior to the assignment.**

LITERATURE REVIEW (20%). It involves a review of statistical tools applied to a particular field of your interest. The format of the review will be posted in Moodle. The review is due on April 12 (6pm) and should be sent in a word document by email to your TA. Students are required to send a brief description (maximum of 300 words via email to their TAs) by Feb 1, 6pm (1% of the review mark will be deduced if description is sent late).

SHORT-TERM ABSENCE FORM - The short-term absence form lets you submit your request for short-term medical accommodation without documentation like a medical note.

https://www.concordia.ca/students/absence-form.html

GRADING SCHEME: A+=91-100, A=85-90, A-=80-84, B+=77-79, B=73-76, B-=70-72, C+=67-69, C=63-66, C-=60-62, D+=57-59, D=53-56, D-=50-52, F<50.

TUTORIALS: No reports are required.

NOTE though that reports and midterm 2 are heavily based on tutorials and your knowledge of R. Therefore you should attend tutorials and work on them weekly.

MATERIAL USED in the COURSE and for STUDYING for EXAMS

MANDATORY: Lectures (videos), lecture notes (may contain more material then the videos), tutorials, material available in our WebBook.

Subjects covered in the course

Biological problems and associated data.

R for statistical computing.

Data structure and types of statistical variables.

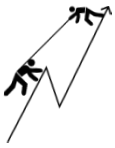
Field versus laboratory studies, experimental versus observational studies.

The concepts of probability, parameters and maximum likelihood; revisiting inferential statistics and statistical hypothesis testing.

Revisiting Analysis of Variance (ANOVA) – parametric and non-parametric.

Advanced Multiple testing and post-hoc analysis.

Multifactorial Analysis of Variance.



Analysis of Covariance (ANCOVA).

Fixed versus random factors: mixed model ANOVA.

Multiple regression and variation partitioning.

Generalized Linear Models (GLMs); spatial and phylogenetic autocorrelation: generalized least square solutions.

Multivariate analyses: introduction and the concept of latent variables and processes.

Multivariate inference: Multivariate Analysis of Variance (MANOVA) and Discriminant Function Analysis (DFA).

Multivariate analyses: Principal Component Analysis (PCA), Principal Coordinate Analysis (PCoA) and Correspondence Analysis (CA).

Multi-response multiple regression: Redundancy Analysis (RDA), relating species characteristics to their environments.

Cluster Analysis, Machine learning, Classification and Regression Tree (CART), and K means.

Advanced non-parametric inference: Monte Carlo testing and bootstrap.

RIGHTS AND RESPONSABILITIES – PLAGIARISM & ACADEMIC INTEGRITY

PLAGIARISM: The most common offense under the Academic Code of Conduct is plagiarism which the Code defines as "the presentation of the work of another person as one's own or without proper acknowledgement." This could be material copied word for word from books, journals, internet sites, professor's course notes, etc. It could be material that is paraphrased but closely resembles the original source. It could be the work of a fellow student, for example, an answer on a quiz, data for a lab report, a paper or assignment completed by another student. It might be a paper purchased through one of the many available sources. Plagiarism does not refer to words alone - it can also refer to copying images, graphs, tables, and ideas. "Presentation" is not limited to written work. It also includes oral presentations, computer assignments and artistic works. Finally, if you translate the work of another person into French or English and do not cite the source, this is also plagiarism. In simple words: DO NOT COPY, PARAPHRASE OR TRANSLATE ANYTHING FROM ANYWHERE WITHOUT SAYING FROM WHERE YOU OBTAINED IT!
Source: https://www.concordia.ca/students/academic-integrity.html

ACADEMIC INTEGRITY: What you can and can't do on assignments and exams? watch this Concordia video: https://www.concordia.ca/cunews/main/stories/back-to-school/video-what-is-academic-integrity.html