BCB720: Introduction to Statistical Modeling

Fall 2015 Syllabus

Last Updated: 2015-06-17

Basic Information

Course identifiers: This document describes the syllabus for BCB720 in the Bioinformatics and Computational Biology Curriculum of BBSP.

Time: 9:30am - 10.45am, Tue/Thu

Location: GMB 1007

Materials: All learning materials will be posted on Sakai.

Restrictions: Class is limited to 25 students.

Instructors

Lead instructor: Prof William Valdar, Room 5113,120 Mason Farm Road, Genetic Medicine Building, Campus Box # 7264, Chapel Hill NC 27599. Tel: +1 919 843 2833. Email: <william.valdar@unc.edu> Web: <http://valdarlab.unc.edu>

Co-instructor: Prof Ethan Lange, Room 5111, 120 Mason Farm Road, Genetic Medicine Building, Campus Box # 7264, Chapel Hill NC 27599. Tel: +1 919 966 3356. Email: <ethan_lange@med.unc.edu>. Web: <http://genetics.unc.edu/faculty/ethan-lange>

Teaching Assistants: TBA. Office hours for the TA will be arranged and posted at the beginning of the first class and updated as necessary.

Course Description

This module introduces foundational statistical concepts and models that motivate a wide range of analytic methods in bioinformatics, statistical genetics, statistical genomics, and related fields. It is an intensive course, packing a year's worth of probability and statistics into 2/3 of a semester. It covers probability, common distributions, Bayesian inference, maximum likelihood and frequentist inference, linear models, generalized and hierarchical linear models, and causal inference.

Target Audience

This course is targeted at graduate students in BBSP with either a quantitative background or strong quantitative interests who would like to understand and/or develop statistical methods for analyzing complex biological/biomedical data. In particular, it is intended to provide a spring-board for BBSP who would subsequently like to take graduate-level statistical courses elsewhere on campus.

Course Pre-requisites

Students are expected to know single-variable calculus (differentiation and integration in 1 dimension), be familiar with matrix algebra and have some programming experience. The course will include some material on partial differentiation of multiparameter functions, and use the statistical package R extensively; familiarity with these will be an advantage. Introductory statistics may or may not be an advantage (depending on how it was taught), but is not assumed.

Restrictions

The course is open to all graduate students of the Biological and Biomedical Sciences Program (BBSP) at UNC Chapel Hill. Other students, staff, or faculty may attend for credit, on an auditor basis or informally **only if**

- They have prior permission from the lead instructor, and
- There is space: that is, if they are not taking up a spot that would be otherwise used by a non-auditing (ie, full credit) BBSP student.

Moreover, graduate students from the Department of Biostatistics (BIOS) or the Department of Statistics and Operations Research (STOR) may audit only, and may not receive credit for this course.

Course Goals and Key Learning Objectives

1. Probability and distributions
2. Properties of random variables
3. Bayesian and frequentist approaches to statistical inference
4. Hypothesis testing
5. Linear models
6. Generalized linear models
7. Hierarchical/mixed models
8. Experimental design
9. Basic multivariate analysis
10. Simulation and method evaluation

Course Requirements

To obtain full credit, students must attend at least 80% of the lectures, complete all homeworks, and achieve at least a passing overall grade.

Dates

Homework assignments will typically be distributed on Wednesdays, with a deadline for electronic submission on the Friday of the following week. Anonymous student evaluations, required for 5% of the course marks, will be distributed for completion on Sakai within approximately a week of course completion. Students will have a week to complete the student evaluation.

Grades

Grades for the course (F,L,P,H) will be based on performance in the homeworks and on completion of the course evaluation. Specifically, the homeworks collectively account for 95% of the course marks, and completion of the anonymous evaluation accounts for the remaining 5%. Each homework will include multiple questions each providing a stated maximum number of points. The total number of points achieved by a student divided by the total possible will be scaled to the range 0 to 95 and used as the percentage of the grade arising from coursework. There is **no final exam**.

Course Policies

Students must attend the entire duration of at least 80% of the lectures unless they have permission of the lead instructor to do otherwise. Students are expected to be prompt, polite, collaborative when (and only when) asked, and to answer questions in class. Failure to hand in a homework on time without reasonable justification (eg, sickness) will result in automatic loss of 10% of that homework's maximum allowable points for each day over the deadline.

Time Table

Key: (C) = Students should bring (or be prepared to share) a laptop

Week	Date	Instructor	Lecture	Description	Homework
1	Tue, Aug 18	(TA)	1 (C)	Introduction to R	Homework 1 (WV)
	Thu, Aug 20	WV	2	Set theory and probability	
2	Tue, Aug 25	WV	3	Conditional Probability	Homework 2 (WV)
	Thu, Aug 27	WV	4	Distribution, Mass and Density functions	
3	Tue, Sep 01	WV	5	Expectation and Variance	Homework 3 (WV)
	Thu, Sep 03	WV	6	Discrete distributions	
4	Tue, Sep 08	WV	7	Continuous distributions	Homework 4 (WV)
	Thu, Sep 10	WV	8	Bayesian inference	
5	Tue, Sep 15	WV	9	Estimation	Homework 5 (WV)
	Thu, Sep 17	EL	10	MLEs, bias	
6	Tue, Sep 22	EL	11	Confidence intervals	
	Thu, Sep 24	EL	12	Hypothesis testing	Homework 6 (EL)
7	Tue, Sep 29	EL	13	Introduction to regression	
	Thu, Oct 01	EL	14	Multiple regression/ANOVA	
8	Tue, Oct 06	EL	15	Multiple regression in class examples	
	Thu, Oct 08	EL	16	Introduction to mixed models	Homework 7 (EL)
9	Tue, Oct 13	(TA)	17	Logistic regression	
	Thu, Oct 15			FALL BREAK	
10	Tue, Oct 20	WV	18	Bayesian regression	Homework 8 (WV)
	Thu, Oct 22	WV	19 (C)	Frequentist regression	
11	Tue, Oct 27	WV	20 (C)	Decisions about modeling	Homework 9 (WV)
	Thu, Oct 29	WV	21	Hierarchical and penalized regression	
12	Tue, Nov 03	WV	22	Causal Inference	Homework 10
	Thu, Nov 05	WV	23	Multivariate statistics	
13	Tue, Nov 10	WV	24	Graphical models	Homework 11
	Thu, Nov 12	WV	25	Handling missing data	
14	Tue, Nov 17	WV	26	Evaluating methods by simulation	Homework 12
	Thu, Nov 19	WV	27	Experimental design	
15	Tue, Nov 24	WV/EL/TBD	28	Case studies / Guest lecture	Homework 13
	Thu, Nov 26			THANKSGIVING RECESS	
16	Tue, Dec 01	WV/EL/TBD	29	Case studies / Guest lecture	Homework 14
	Thu, Dec 03	WV/EL/TBD	30	Case studies / Guest lecture	

Syllabus Changes

The lead and/or co-instructors reserve to right to make changes to the syllabus, including homework due dates.

Course Resources – preliminary list

There is no course textbook as such because no textbook seems to cover all the material in this course. Some textbooks that may be useful for supplemental reading are given below. However, be prepared to try a few books before finding one that is a good fit for you; a cheap way of doing this is to sample books that are freely available electronically at UNC (http://search.lib.unc.edu/search.jsp). Also, use web resources such as Wikipedia.

1st half of the course:

Westfall & Henning (2013) "Understanding Advanced Statistical Methods" – *chatty, popular with some students*

Casella & Berger (2002) "Statistical Inference" – *less chatty, more rigorous/mathematical*

DeGroot & Schervish (2011) "Probability and Statistics" – *less chatty, more rigorous/mathematical, tries to strike a balance between Bayesian and frequentist perspectives.*

Dekking, Kraailkamp, Lopuhaa, Meester (2007) "A modern Introduction to Probability and Statistics: Understanding Why and How" – *more gentle intro,* SpringerLink

Wasserman (2009) "All of Statistics" – *was recommended in previous years, but found by some to be a bit terse*

2nd half of the course:

Gelman & Hill (2007) -- good for understanding linear models and estimation, but not hypothesis testing

Also useful:

Christensen (2011) "Plane answers to complex questions: the theory of linear models" – *rather mathematical,* SpringerLink

Gentle (2007) "Matrix Algebra: theory, computations, and applications in statistics" – SpringerLink

Johnsen & Wichern (2004) "Applied Multivariate Statistical Analysis" -- *good intro to matrix algebra (chapter 2)*

Venables & Ripley (2002) "Modern Applied Statistics with S" -- *very terse but comprehensive on R (available free online)*

More basic than this course, but still useful:

Verzani (2004) "Using R for introductory statistics" -- *friendly chatty book on R*

Dalgaard (2008) "Introductory statistics with R" – *freely available via UNC's* SpringerLink

More references (eg, for specific subjects) will be given during and at the end of the course. Students are encouraged to ask the instructors for recommendations for books/resources on specific subjects or books/resources aimed at different levels.

Honor Code:

Students may collaborate in class, but each student's homework should be their own. In completing the homework, however, students are nonetheless encouraged to consult the lecture notes, online material, books and any other "passive" sources. They may discuss general strategies and concepts with their classmates and with the TA, and may ask the TA for clarification about the content of questions. The TA may provide guidance as to where they might be able to find example material that addresses problems similar (but not identical) to those posed in the homework

Comments and advice from previous years' students

Description of the course

"BCB 720 is an accelerated and concise overview of probability and statistics from both a frequentist and Bayesian perspective. This course introduces students to probability theory, probability distributions, hypothesis testing, and linear modeling."

"… This course is essentially 2 undergraduate courses in frequentist prob and stat coupled with Bayesian inference."

"This class is very challenging and time-consuming. It will cover more traditional topics in probability and frequentist stats all the way through Bayesian regression and causal inference; all topics that may come up in research."

"Strenuous course which outlines many of the fundamental elements of statistics used in common biological problems with large datasets. "

"… heavy workload but the material does come up in research and in other classes therefore it can be very valuable."

"… Look forward to spending … hours on problem sets. However, the course is very beneficial and I would recommend to everyone."

"… It goes into a bunch of different areas that can help provide a springboard and understanding to be able to take a further and deeper class in an area"

"I cannot even convey how much I loved this class. I really wish it could be a whole semester. The homework was incredibly interesting and the questions were some of the most thought provoking … I have ever been asked…"

"A really hard crash-course in probability and statistics for modern-day bioinformatics."

Advice

"Read a lot!! This is deep material and important material. Consult outside resources, look at examples, work through examples, read papers with applications- anything possible to make concepts tangible and intuitive. Buy the Casella and Berger book and read the chapters several times. Look at homework shortly after it has been assigned, let your brain work on it and come back to it a few days later."

"Really focus on the first few hw's because they are foundational."

"Do not wait until the last minute to start the homework assignments. They will always take longer than you think they will. Also, use the TA's office hours–I usually didn't, but when I did, it was very helpful, and I probably should have gone more."

"Get the book early on, and read along with the sections covered in class. Start homework early, go to review sessions and don't be afraid to ask questions when you really don't understand in class. If you really don't understand in class, you really won't when you start doing the homework. Be cautious about taking demanding classes in other departments at the same time."

"Practice R early on"

"Start the homework early (do a little bit each day). Also go to TA Office Hours."

“Get very familiar with calculus before the class even starts and write down every verbal explanation of the concepts during lecture because you will have a very hard time going back into the notes to figure out how to do the homework.”

“Plan ahead. The homeworks take time but if you do some reach night it becomes much much more manageable. Also remember that a lot of your learning and reasoning come from doing your assignments”

“Take advantage of office hours. Think about what the formulas represent/model. Start the homework immediately if you can. Learn R if you have time. Learn LaTeX so that you don’t have to waste part of your morning every Friday scanning homeworks at the library.”