BCB720: Introduction to Statistical Modeling

Fall 2020 Syllabus

Last Updated: 2020-07-22

Basic Information

Course identifiers: This document describes the syllabus for BCB720 in the Bioinformatics and Computational Biology Curriculum of BBSP.
Time: 3:00 – 4:15, Tue/Thu
Location: Via Zoom
Materials: All learning materials students will be posted on Sakai.
Restrictions: Class is limited to 30 students.

Instructors

Instructor: Prof William Valdar, Email: <william.valdar@unc.edu>, Web: <http://valdarlab.unc.edu>
Teaching Assistants: TBA
Student Services Manager: John Cornett <jcornett@email.unc.edu>

Course Description

This semester-long course 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 one semester. It covers probability, common distributions, Bayesian inference, maximum likelihood and frequentist inference, linear models, logistic regression, generalized and hierarchical linear models, and causal inference, plus, typically, additional topics from guest lecturers. The course makes use of the statistical programming language R, and all coursework is expected to be written using some combination of R and, either directly or indirectly, the document preparation language Latex, both of which are introduced in the course.

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, be comfortable with algebra, and somewhat familiar with matrix algebra. Specifically:
    - Essential calculus: functions (including inverse functions, exponential functions, logarithm functions), exponential and logarithm equations, graphing functions, limits, derivatives (including derivatives) of exponential and logarithm functions), chain rule, second/third derivatives, derivatives to find maximum and minimum values, integrals (definite and indefinite) and areas. [No trig needed!]
    - Essential other algebra: sums ($\Sigma_{i=1}^{n}$), products ($\prod_{i=1}^{n}$) and relations such as $\prod_{i=1}^{n} e^{x_i} = e^{\sum_{i=1}^{n} x_i}$, vectors, matrices, transpose, vector/matrix multiplication. [Note: understanding geometric interpretations of matrices/vectors is not necessary, just knowing how to do basic manipulation is enough.]
    - Helpful but not an essential pre-req: simple multivariate calculus (differentiating multiparameter functions), determinants, linear combinations, quadratic forms.

    - See the “Preparatory Exercises” listed on the course info website https://valdarlab.unc.edu/bcb720/ .
- Students are also expected to have some basic programming experience (including variables, “for loops”, etc). The course will make extensive use of the statistical package R and the math-friendly documentation language Latex (either directly or via Rmarkdown, Knitr, or similar), so familiarity with these will be helpful.
- Previous exposure to 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. Basic multidimensional analysis (PCA)

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. Homeworks should be written using Latex typesetting, either directly or through, eg, Rmarkdown, Knitr, etc.

Dates

Homework: Assignments will typically be distributed on Wednesdays, with a deadline for electronic submission on the Friday of the following week. Also, 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.

Drop date: Under normal circumstances, the latest date for dropping the course, or, for example, switching to auditor status, is sometime in October if using the web registration system or late November if going through John Cornett / official channels. Similarly, under normal circumstances, the last day to reduce course load in order to have tuition adjusted is early September. However, for this year, please check with the BCB student services manager John Cornett.

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 (see below), and completion of the anonymous evaluation accounts for the remaining 5%. There is **no final exam**.

Homeworks: 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. A homework that is handed in late, without prior agreement of the instructor, will have points in a manner described on the rubric of the homework sheet.

Grade conversion: Total course percentages will be mapped to HPLF course grades *based on* the following grade boundaries: H= 90+%, P=75+%, L=60+%, F=<60%. Specifically, in some years the instructor may use slightly adjusted boundaries if it is seems necessary to recalibrate against unintended changes in homework difficulty, etc. For students whose curriculum requires letter grades, grade boundaries will be based on the boundaries: A=90+%, B=80+%, C=70+%, D=60+%, F=<60%.

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.
- Electronic devices should be stowed away during class unless otherwise instructed. Most classes are pen-and-paper based.

Lecture and Homework Schedule (preliminary)

Key: WV=Will Valdar, CWC=Cavin Ward-Caviness, S=Survey

Week	Date	Lec #	Instr.	Description	HW
1	**Tue-11-Aug**	1 (C)	(TA)	Introduction to R and Latex	1
	Thu-13-Aug	2	WV	Set theory and probability	
2	**Tue-18-Aug**	3	WV	Conditional Probability	2
	Thu-20-Aug	4	WV	Distribution, Mass and Density functions	
3	**Tue-25-Aug**	5	WV	Expectation and Variance	3
	Thu-27-Aug	6	WV	Discrete distributions	
4	**Tue-01-Sep**	7	WV	Continuous distributions	4
	Thu-03-Sep	8	WV	Mixtures, Joint densities	
5	**Tue-08-Sep**	9	WV	Multivariate densities, likelihood	5
	Thu-10-Sep	10	WV	Bayesian inference	
6	**Tue-15-Sep**	11	WV	Estimation	6
	Thu-17-Sep	12	WV	Frequentist behavior	
7	**Tue-22-Sep**	13	WV	Confidence intervals	7
	Thu-24-Sep	14	WV	Hypothesis testing: concepts	
8	**Tue-29-Sep**	15	WV	Hypothesis testing: Wald, score, LRT	8
	Thu-01-Oct	16	WV	Power and false positive rate	
9	**Tue-06-Oct**	17	WV	FWER and FDR	9
	Thu-08-Oct	18	WV	Two-group tests: permutation and t-test	
10	**Tue-13-Oct**	19	WV	Linear models	10
	Thu-15-Oct	20	WV	Linear models: estimation	
11	**Tue-20-Oct**	21	WV	Linear models: testing	11
	Thu-22-Oct	22	WV	LM departures	
12	**Tue-27-Oct**	23	WV	Generalized linear models	12

	Thu-29-Oct	24	WV	[Overflow]	
13	**Tue-03-Nov**	25	WV	Linear mixed models	13
	Thu-05-Nov	26	WV	Hierarchical models and Bayesian regression	
14	**Tue-10-Nov**	27	CWC	PCA	S
	Thu-12-Nov	28	WV	Model selection	
15	**Tue-17-Nov**	29	WV	Causal inference and Experimental Design	

Syllabus Changes

The lead instructor reserves 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) -- *great for understanding linear models, generalized linear models, and estimation, but doesn't really cover hypothesis testing*
Wakefield (2013) Bayesian and Frequentist Regression Methods. Springer

Other suggested resources / further reading

Johnsen & Wichern (2004) "Applied Multivariate Statistical Analysis" -- *good intro to matrix algebra (chapter 2)* Valore (Alternate)
Gentle (2007) "Matrix Algebra: theory, computations, and applications in statistics" – SpringerLink
Harrell (2015) "Regression modeling strategies" -- *lots of good advice for applied work* - SpringerLink
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, from the 2019 end of class survey

From BCB students

"A magnificent but hard overview of all statistics."
"This course provides a strong foundation in probability and statistics for all students, with particular emphasis on applicability to biological experiments and data."
"A deep introduction to statistical modeling"
"A course for you to know how you may tailor an existing statistical technique instead of just blindly performing the statistics using Excel."
"Overview of statistical theory and implementation with a combination of frequentist and Bayesian statistics."
"Important concepts in statistics presented matter-of-fact-ly and much too quickly"
"Tons of information presented in a very compact format. Went in knowing nothing about statistics, came out feeling like I had a solid grasp on the subject. The homework was a massive time dump and a huge pain in the [censored], but really helped drive home the important concepts. Easily the most difficult and time consuming class I have ever taken. 10/10"
"The hardest and most time-consuming class of your career."
"This is a well-designed introductory statistics course that covers a wide range of topics using exceptionally high-quality materials."
"BCB 720 covers the underlying statistics and mathematics used in a variety of computational methods for biological research."
"Prepare to learn more than you ever thought you could, and gain insight into a field that is largely overlooked or misrepresented by many scientists. The homework assignments are time-consuming and at times burdensome, but at the end you realize that each piece served a purpose (as in a puzzle), and it comes together beautifully at the end."
"This is an intensive course, but you do get to grasp the overall concepts in statistics. This course will help you in understanding the statistical methods used in scientific literature."

From non-BCB students

"A very time consuming, but rewarding, introduction to statistics that can help you understand principles of statistics that will probably help you with your research."
"Requires a lot of effort and work but totally worth to understand statistical modelling"
"A general, theory-intensive statistics course that will make you a better scientist"
"You are gonna learn a lot about statistics and inference through daily life topics, and biology models. The assignments are intense overall and you will put all together: R coding, statistics theory, latex formating."
"A survey course that begins with predicting the outcome of a coin toss and ends with constructing linear models for a drug trial, this statistical modeling course gives you a quick-paced, in-depth introduction to ideas underpinning both classic and modern statistics. Over-simplified in both popular culture and by some professionals, the ideas of statistics are rich, complex, and grounded, and here you'll learn how to practice it with your eyes open."
"This class will be challenging, but you will learn a lot. It is a great class if you want to be exposed to a lot of different statistical concepts."

"BCB 720 is an Introduction to Statistics because of the foundational approach it takes to statistics, not because the material is at an introductory level. Students will learn about topics ranging from basic probability and density functions to estimation and linear modeling in a fast-paced classroom and through in-depth homework assignments."
"This is a comprehensive course on Statistics and how they can inform your certainty of events occurring or not occurring. You will learn how to apply data to distributions, infer what your data is indicative of and how to properly analyze the result of statistical tests."
"A brief but challenging course covering basics of statistical modeling. A lot of work but a good way to gain stats, R, and calculus skills all at once."
"A broad yet detailed overview of important statistical concepts you will almost definitely encounter."

Advice from students, from the 2019 end of class survey

From BCB students

"Do homework immediately."
"Generally, and especially if the student has a weaker background in mathematics, review calculus (limits, derivatives, and integrals) and linear algebra (matrix operations) before the course starts. A strong start there will make it easier to understand any new material presented in lecture. As people learn in different ways, also do not rely just on the lectures alone; supplement that material with textbooks and/or online resources."
"Go to office hours. Get started on the homework's early so you know what you have questions about. Make a study group and work on homework with other people in the class."
"It is not as intimidating as it looks. It is true that the workload may be a bit too much (an average of 6 hours for each assignment), but you can learn a lot from the homework."
"Be willing to put many many hours into the homework, understand that you won't comprehend everything on the lectures first read, read a lot of external sources when you're confused and the slides don't help. Go to office hourse, they're essentially mandatory."
"There is not much advice I can give. The homework takes a very long time. But you learn that on your own"
"Be prepared to spend 10-15 hours a week on homework."
"Don't take this course with any other time-consuming course and be prepared to work more than you did for any undergraduate course."
"Review lecture notes before attending classes. Review them again before working on assignments. Attend the office hours whenever possible."
"BCB 720 is worth it in the long-run. It takes years to fully understand the content. However, it is helpful to have somewhat of a clue when reading math papers or reading the documentation of some computational tool for biological research."
"-Start homework assignments fairly early, since most of them will take a long time to complete and you don't want to get stuck on something and not be able to go to office hours to work it out.
-Go to office hours and form study groups. You will save a lot of time if you work with people who will ask questions you have (or might have later on), and who may have strengths that you do not have but are helpful in completing assignments.
-Take in-class assignments seriously enough to formulate some kind of answer. It may be wrong (mine were on a few occasions), but it is helpful in tuning into the material and showing that you care. It took me a long time to realize this, but Will will not judge you if you are wrong."
"Be prepared to put (a lot of)time into the homework especially in the first month of the course. You need time to get used to writing in latex."
"Be prepared to spend 10-15 hours a week on homework."
"Brush up on calculus, R, and latex ahead of class if possible. However, you learn a lot as you go and can probably get by"

From non-BCB students

"This course does require comfort with somewhat complicated math and at least basic programming experience such that re-learning all of calculus or picking up R and LaTeX (if you don't already know them) won't

be the main challenge. The best advice I can give is to complete the homework as soon as it comes out whenever possible. Life happens - and doing the assignments right up against the deadline is bad for grades and learning. All that being said, if you really want the skills and knowledge that this course provides, you shouldn't be deterred by its difficulty. If you stick through it, you thank yourself once all is said and done."
"It is helpful to have buds to discuss topics for this class, so try to get to know and collaborate with your peers. Going to the TA's office hours is insightful."
"Read the lecture notes and go to office hours."
"My biggest piece of advice would be for them to make the most of the office hours. For the most part they are good at explaining and helping out and significantly cut down the time I spent struggling on the HW."
"Make use of and prioritize the office hours for each homework. Start the homework at least 3-4 days before the due date. Always take the in-class exercises seriously. Don't waste too much time formatting the homework. That time is better spent trying to understand the content and checking your work."
"Do not expect to get much done in your rotation/laboratory. Start early and often on the homework and use the 'print out' method of taking notes in class. No other note taking method worked for me."
"If you intend to take the course, do some refresher calculus, matrix algebra, and learn the log rules. That would have saved me so much time in the long run. Also, take advantage of the TAs. They were overall really great at helping out."
"Attend office hours as much as possible! Know that if you put more effort into the course and understanding concepts as you go, the more you'll get out of the course at the end. Expect to spend about 10-12 hours on homework each week, between actually answering questions and formatting. Talk to your mentor(s) or rotation advisors honestly about your course load. There is a lot of value in taking this course, but it is a serious effort. Thankfully my advisor also saw the value in this course, despite the time I spent away from lab."
"To ask questions, even if you feel like it's stupid, more than likely everyone else is confused as well. Start the HW before the week it's due. Take advantage of office hours! Use online resources as well for concepts you did not understand, but Will is usually pretty good at answering questions so reach out to him first."
"I will say be prepared. Read the lecture notes really carefully. Google is your best friend."