BCB720: Introduction to Statistical Modeling

Fall 2021 Syllabus

Last Updated: 2021-10-11

Basic Information

Course identifiers: This document describes the syllabus for BCB720 in the Bioinformatics and Computational Biology Curriculum of BBSP.
Time: 2:00 – 3:15, Tue/Thu
Location: Marsico 6004
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: Brooke Felsheim <felsheim@email.unc.edu>, Sool "Will" Lee <sool_lee@unc.edu>, Ellen Risemberg <erisembe@email.unc.edu>,
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 $\prod_{i=1}^{n} e^{x_i} = e^{\Sigma_{i=1}^{n} x_i}$, 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, C=need computer, S=Survey

Week	Date	Lec #	Instr.	Description	HW
1	**Thu-19-Aug**	1 (C)	(TA)	Introduction to R and Latex	
	Tue-24-Aug	2	WV	Set theory and probability	1
2	**Thu-26-Aug**	3	WV	Conditional Probability	
	Tue-31-Aug	4	WV	Distribution, Mass and Density functions	2
3	**Thu-02-Sep**	5	WV	Expectation and Variance	
	Tue-07-Sep	6	WV	Discrete distributions	3
4	**Thu-09-Sep**	7	WV	Continuous distributions	
	Tue-14-Sep	8	WV	Continuous and derived distributions	4
5	**Thu-16-Sep**	9	WV	Multivariate densities, likelihood, mixtures	
	Tue-21-Sep	10	WV	Bayesian inference	5
6	**Thu-23-Sep**	11	WV	Estimation	
	Tue-28-Sep	12	WV	Frequentist behavior	6
7	**Thu-30-Sep**	13	WV	Confidence intervals	
	Tue-05-Oct	14	WV	Hypothesis testing: concepts	7
8	**Thu-07-Oct**	15	WV	Hypothesis testing: Wald, score, LRT	
	Tue-12-Oct			UNC-CH WELLNESS DAY	8
9	**Thu-14-Oct**	16	WV	Power and false positive rate	
	Tue-19-Oct	17	WV	FWER and FDR	9
10	**Thu-21-Oct**			FALL BREAK	
	Tue-26-Oct	18	WV	Two-group tests: permutation and t-test	10
11	**Thu-28-Oct**	19	WV	Linear models	

	Tue-02-Nov	20	WV	Linear models: estimation	11
12	**Thu-04-Nov**	21	WV	Linear models: testing	
	Tue-09-Nov	22	WV	LM departures	12
13	**Thu-11-Nov**	23	WV	Generalized linear models	
	Tue-16-Nov	24	WV	Linear mixed models	13
14	**Thu-18-Nov**	25	WV	Hierarchical models and Bayesian regression	
	Tue-23-Nov	26	WV	Model selection	S
15	**Thu-25-Nov**			THANKSGIVING RECESS	
16	**Tue-30-Nov**	27	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 2020 end of class survey

From BCB students

"This course is an excellent introduction to probability and statistical modeling. It provides an overview of many important statistical topics and is presented in a way that is straightforward with achievable and rewarding challenges."

"It is a very useful course that lays foundation for future statistic-related research and career."

"This course is an extremely useful and in-depth introduction to several topics in probability, statistical modeling and statistical inference, contextualized with examples involving biological data."

"This course provides a comprehensive overview of statistical modeling in a way that is especially relevant to biological research."

"A difficult stats course covering a wide range of concepts and often focuses on the "why this works" in statistics."

"This class definitely will make you suffer and question your intelligence at times. However, in the end, you will be glad that you took this course. You will learn a lot."

"This is a course that provides tools for and understandings of statistical modeling based on foundational concepts of probability, calculus, and linear algebra."

"Introduction to statistical modeling, distributions, linear models. Background in statistics and R helpful but not required"

"Very time consuming course that might even cause you to mix up your existing knowledge of statistics. This course requires in depth background in R and very very basic mathematics."

"This is the most useful course in BCB and will be a great foundation for your research in the future."

"This course will take you from the fundamentals of probability and statistics to building statistical models from scratch. It's difficult in the beginning if you're not from a prob/stat background but you will gain a lot from start to finish if you make a serious effort to learn. A lot of this class is learning the language around statistics and what the jargon means mathematically."

"A very challenging course. Easily the toughest class I have ever taken. Most assignments are hard so make sure you have a strong math background before you take this course. Dr. Valdar and the TA's are a wonderful resource."

"An introduction to statistical topics in computational biology."

"This class is an introduction to many of the fundamental statistical methods that are used in modern research. It's a ton of work, but it's very eye opening."

From non-BCB students

"A challenging course which covers a wide variety of topics used in literature related to BCB topics. The course will provide enough support for these topics that students can have some confidence in their ability to further develop their understanding for their particular research needs."

"A broad but rigorous overview of the statistics know-how you need to analyze and understand data effectively."
"An enjoyable, if occasionally obscure, foray into mathematical statistics that leaves you well prepared to expand your prowess in research formats."
"This course is a hands-on introduction to core concepts in statistical modeling. The content is centered on biostatistics applications."
"An overview of statistic concepts and applications that provide students with analytic tools to interpreting data."
"An introduction to advanced statistical concepts that are useful in a lot of computational research."
"A wide survey of statistics, beginning with the basics and building on them and culminating into important applications of the field. Well suited to introduce students to a variety of statistical techniques, methods, and ideas while additionally providing resources for individual curiosity."
"A solid introduction course to statistical modeling. Comprehensive overview of statistical models, probability distributions, and hypothesis testing."
"This is a relatively hard course, those who seek to pursue biostatistics as their major should enroll."
"A shaky rope bridge from introductory statistics to generalized linear models, linear mixed models and advanced hypothesis testing."
"A good general survey of the major statistical topics involved in research."

Advice from students, from the 2020 end of class survey

From BCB students

"I would tell students taking the course to start the homework as soon as possible. Attempting the problems early on is the best way to get the most out of office hours. I would also suggest going to office hours even if you haven't started the homework or do not have questions, since other students may ask about problems/issues you may run in to. Since this semester was online, it was good to discuss problems with other students (not just the TA) during office hours, too."
"Go over the slides."
"Start the homework the week/weekend before it is due, and go to office hours having attempted each of the problems so you know what questions to ask."
"Go to office hours, and start homework early so you can identify what you might have questions about."
"Take the course a week at a time. Try not to get overwhelmed. The concepts and workload gets easier to manage as you get used to it."
"Realistically, start your homework on Monday and utilize all of the office hours to answer your questions. Also, form a homework group!"
"To echo what I read in the commentary about the course in this year's syllabus, office hours are essential, and getting an early start on the homeworks is extremely helpful for formulating questions. The assignments generally supplement the lectures extremely well, and if you're a person who requires applications to understand concepts, you will find that generally the formula this class is a winning one."
"Go to office hours, and start homeworks early!"
“Do not take this course!”
"Learn linear algebra and preview some knowledge of calculus."
"Prioritize your time to focus on this class for the first 4 or 5 weeks. The material gets easier after that point."
"Make sure you don't have a heavy course load when you sign up for this class. The assignments will easily take over 10-15hrs. Attend office hours and start homework early. Reviewing in class exercises also helps. it is a very challenging course and not for the faint hearted."
"Just leave enough time for yourself to do homework / understand lectures and the course isn't very difficult."
"Don't go into the class expecting a cakewalk"

From non-BCB students

"Set aside time, start early, attend office hours, find friends to work through the material with, attend all lectures possible, apply yourself or you will struggle."

"Make sure to keep up with the lectures. If there is something you don't understand, get it sorted out as soon as possible. Don't be shy - your classmates have the same questions as you! Office hours are a life-saver. Support and encourage your classmates."
"Hmm. Start early, even though I know you won't. Make friends who you can complain together with because parts will be really confusing and having classmates you can argue with will really help fix that. Also R coding in Latex is always going to add at least a few hours to your write up. Finally to drink water and get plenty of sleep."
"Start the homework early and look through the lecture slides thoroughly for help. Try not to zone out during class, especially if it's virtual again."
"I would recommend making a study group to bounce ideas off of. It is super helpful to talk through problems out loud. I also recommend using google as a resource to supplement concepts you don't quite understand. Also, it is better to ask questions sooner rather than later."
"Definitely start the homework assignments right away. Each one requires a lot of work and it will be very stressful if you wait until the last minute. Also, going to office hours is very beneficial but you will get the most out of it if you have already started the homework so you know what they're talking about."
"Do not procrastinate on the homeworks. Make sure you do the preparatory exercise to know what you might need to catch up on before the start of the course. Take time to learn Latex and R if you do not have prior experience."
"Start on the homework early. Make good use of reviewing the lectures. Also attending office hours is super helpful. Don't leave homework until the last minute as they are quite time-intensive. Take good notes on lectures and homeworks."
"I think paying attention to the class and try to understand the concept and read more on the class material or watch video on the same day or within a couple of days, else this is very volatile and you will have difficulty to understand the concept at a later time point."
"If you prepare for this class by taking four years of math classes and some stats it's not so hard and really enjoyable. You'll learn how to use the theoretical things you learned for real problems."
"Just brush up on some probability theories and get started early on homework."

Students' opinions of pre-requisites, from the 2020 end of class survey

"There are so many online resources for R, that I feel like no prior use of R was necessary. I last used R in high school and that seemed like adequate prior knowledge. As for mathematics, having had calculus 1 and 2 and brushing up on linear algebra/matrix operations seemed like all the math that was necessary. All other math topics were covered in the class.Already knowing Latex was extremely helpful. Though Latex is not too difficult, not knowing it would have greatly added to the stress of the weekly homework."
"Calculus I. Basic programming."
"Familiarity with calculus and matrix algebra. R is easy enough to pick up if you have a basic understanding of how to code, but I do think if a student has no exposure to programming (i.e. doesn't understand what a loop is, how variables work, etc.) they would have a hard time."
"I think a student should be familiar with at least one programming language and should have previously taken a calculus course (exposure to matrix algebra is also helpful but probably not necessary). If they've never used R or LaTeX there would probably be a bit more overhead at the beginning of the course."
"I don't think there should be prerequisites. I didn't have the math prerequisites and I faired fine. But I think either being prepared either on the math front or the coding front is valuable for not making students feel overwhelmed."
"General knowledge of algebra and calculus is required. General programming knowledge is required."
"I think generally the pre-requisites are reasonably well described as they currently are. Maybe a little more clarity on the point that this course will develop skills in R for data visualization and analysis if individuals do not have much experience. I was pretty intimidated by Latex and R when I first started the course, but I learned a lot and eventually picking up understandings of functions was generally painless."
"I honestly think what you have is fine. It might have been helfpul to know some of these things heading into the class but it is doable without more of a background and i think that this course is really valuable for people

who may not have the "required" background (like me). I definitely didn't ace every homework but I learned so much more than any other class in my graduate career."
"Expertise in R and Latex and recent course on basic math."
"I think mathematics should be ok.Programming of R should familiar with some loops, data tables, figure plots, and drawings. Don't worry, the homework will give you the hint and TA will teach you how to draw a figure. If you are not familiar with R but familiar with any other programming language that would be fine."
"Calc 1, and 2, and be familiar with basic r syntax and generally how a for loop works."
"need at least 2 semesters of calculus (esp differential equations and logarithms) and familiarity with R and latex. R experience is definitely useful. the course would have been twice as hard had I not known R."
"Calculus is required. I think some ability with programming is useful but I don't think knowledge of R is required."
"I think the only kind of prerequesite would be basic knowledge of calculus, linear algebra, and stats. I had no familiarity with R, and it was a breeze to pick up just for this class."

From non-BCB students

"At least comfortable with typical calculus level math (derivatives, integrals, etc) and at least an introductory level of programming experience. (even just something like SWIRL to learn the very basics of R). Mostly could learn the programming as you go but the math would be difficult to start from scratch."
"Willingness to learn and work hard is the best prerequisite. Calculus, basic R experience required. Previous statistics course quite helpful, but not required."
"Familiar with vector and matrix based programming and mathematics. A willingness to hurt your brain."
"Previous programming experience in any language, math up to linear algebra and calc 1 (probably not required, but helpful)."
"Prerequistites of linear algebra, calculus, and some type of programming course."
"Having some programming experience is definitely helpful but it doesn't need to be in any particular language/platform. You don't really need to know advanced calculus but it would definitely be helpful if you have taken some statistics classes before or even used advanced statistics in your research."
"Prerequisites might include good understanding of basic calculus techniques of the derivative and integral as well as solid algebraic skills. A very low level (if any) of understanding with regards to programming is needed."
"Basic R programming skills are helpful but not necessarily required. Solid prerequisite knowledge of calculus and matrix algebra are needed."
"Understanding of differential and integration and understanding of R and Latex"
"Calculus 1 and 2, some programming helpful but not necessary if you have time to teach yourself the basics."
"Background in probability theory and algebra and some experience working with R."