BCB720: Introduction to Statistical Modeling

Fall 2022 Syllabus

Last Updated: 2022-07-26

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 Canvas
Restrictions: Class is limited to 30 students.

Instructors

Instructor: Prof William Valdar, Email: <william.valdar@unc.edu>, Web: <http://valdarlab.unc.edu>
Teaching Assistants: Jess Byun, Gilbert Giri, Kalika Kamat, Dhuvi Karthikeyan.
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 springboard 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 be 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, C=need computer, S=Survey

Week	Date	Lec #	Instr.	Description	HW
1	**Tue-16-Aug**	1 (C)	(TA)	Introduction to R and Latex	
	Thu-18-Aug	2	WV	Set theory and probability	1
2	**Tue-23-Aug**	3	WV	Conditional Probability	
	Thu-25-Aug	4	WV	Distribution, Mass and Density functions	2
3	**Tue-30-Aug**	5	WV	Expectation and Variance	
	Thu-01-Sep	6	WV	Discrete distributions	3
4	**Tue-06-Sep**			WELLNESS DAY	
	Thu-08-Sep	7	WV	Continuous distributions	4
5	**Tue-13-Sep**	8	WV	Continuous and derived distributions	
	Thu-15-Sep			BCB RETREAT	
6	**Tue-20-Sep**	9	WV	Multivariate densities, likelihood, mixtures	
	Thu-22-Sep	10	WV	Bayesian inference	5
7	**Tue-27-Sep**	11	WV	Estimation	
	Thu-29-Sep	12	WV	Frequentist behavior	6
8	**Tue-04-Oct**	13	WV	Confidence intervals	
	Thu-06-Oct	14	WV	Hypothesis testing: concepts	7
9	**Tue-11-Oct**	15	WV	Hypothesis testing: Wald, score, LRT	
	Thu-13-Oct	16	WV	Power and false positive rate	8
10	**Tue-18-Oct**	17	WV	FWER and FDR	
	Thu-20-Oct			FALL BREAK	
11	**Tue-25-Oct**	18	WV	Two-group tests: permutation and t-test	
	Thu-27-Oct	19	WV	Linear models	9
12	**Tue-01-Nov**	20	WV	Linear models: estimation	

	Thu-03-Nov	21	WV	Linear models: testing	10
13	**Tue-08-Nov**	22	WV	LM departures	
	Thu-10-Nov	23	WV	Generalized linear models	11
14	**Tue-15-Nov**	24	WV	Linear mixed models	
	Thu-17-Nov	25	WV	Hierarchical models and Bayesian regression	12
15	**Tue-22-Nov**	26	WV	Model selection	
	Thu-24-Nov			THANKSGIVING RECESS	S
16	**Tue-29-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 last year's course survey

Q. In one or two sentences, how would you describe this course to next year's incoming students (eg, as would be written in a quick course description)?

- "This course delivers a strong medley of both applied and theoretical statistics, exposing the learner to a steady foundation of the modeling techniques most common to the field of biomedical research."
- "This class is a rigorous and concentrated course which provides a foundation of biomedical statistics up to linear modeling."
- "A tour-de-force of the foundations of modern statistics."
- "This is an extremely useful but challenging course that will feel like it's the worst class in the world in the middle of the semester. By the end, you'll be thankful you took it."
- "A throw-you-in-the-deep-end style introduction to grad school level statistics."
- "BCB 720 is a course that teaches a comprehensive set of statistical topics that are essential for research in biological and biomedical sciences."
- "This course provides a solid broad yet deep statistics background and help change your way of thinking towards events happening around you in general and science in specific."
- "Everything about statistics you need to know as biologist."
- "This is simultaneously the best and worst class you will take! It's a lot of work but absolutely worth it."
- "This course presents an overview of statistical modeling from its foundations rooted in probabilities to the derivation of linear models."
- "It was hard."
- "A fairly comprehensive and extremely useful course about variuos statistical methods that you will likely come across in the future"
- "The course introduces you to many statistical concepts. Use lectures and assignments to improve understanding."
- "On paper looks like a course to weed out students, but really is a very useful tool"
- "BCB 720 covers a wide range of statistical concepts to provide students with a solid statistical foundation for future bioinformatics/biostatistics research."
- "It is extremely detailed and difficult."
- "Fundamentals of statistical models with special emphasis on underlying driving factors of how and why mathematical models vary to capture estimates of processes."
- "This is a very useful class that lays a good foundation for your future research."

Q. What advice (if any) would you give to students taking the course next year?

- "Make sure you understand calculus from day 1. Don't try to re-learn/learn it on the fly throughout the semester."

- "The pre-class worksheet is helpful, but be sure that you don't just work through it, but really understand the topics, they will come in handy later in the class."
- "Form a study group as soon as possible. It can feel overwhelming but talking through the topics with your peers is the best way to make it feel manageable."
- "It's a doozy, but it's so worth it if you give it the time it deserves. Start early on the problem sets, especially in the beginning!"
- "Take advantage of office hours, don't procrastinate on the homework, and form study groups."
- "Don't get too lost in the details of the lectures. If you're struggling in class, focus on learning what you need to do the homework (you don't need to understand every single detail to do the homework). When doing homework, start at the slides. Oftentimes, hints or pieces of code you will need are built into them. Go to office hours - it will make homework feel much more approachable. Know you aren't the only one struggling, these assignments are hard!"
- "Make sure to go over the math refresher / review PDF file before the first class."
- "I would like to tell you that when I decided to take this course, I was worried I would fail it and the main reason behind this feeling was the feedback of the other students. However, it is not as bad as it sounds (my stats background isn't strong). The recipe to success in this course is: 1. Engage in the lecture, 2. Attend all the office hours and take notes in them. 3. Start the HW early. 4. Work hard."
- "Don't listen to what people have said in the past. The class is completely doable, just make sure you use the resources you have adequately and focus on starting early."
- "Attend all TA sessions and start the homework early!
- Knit your HW in parts to make it easier to troubleshoot errors."
- "Review probabilities and basic stats."
- "I think study group would be very helpful."
- "Start the homeworks early!"
- "Start homework assignments early."
- "Attend every class and office hours ask questions and don't stress."
- "1. Office hours, go to office hours, all of them. 2. Annotate the lecture notes. This isn't a class you want to be trying to jot notes down in a notebook for. 3. Never procrastinate on a homework. 4. Will is very approachable, if you are struggling reach out."
- "I would say to attend office hours right from the start of the course."
- "Be secure in your calc!!!"
- "Learn R before class if you haven't used R before. Do a lot of practice in R before taking this course. Start homework early and go to TA office hours every week."

Q. In one or two sentences, describe what you think should be the stated pre-requisites for the level of mathematics and programming.

- "A strong grasp of calculus I and probably II is required. My greatest struggles during this class were a direct result of limited calculus knowledge. I was more comfortable with the actual programming aspect of assignments and enjoyed doing them. I have a decent background in programming in R which certainly helped. I imagine it would be a bit of a learning curve otherwise, but I couldn't say exactly how much knowledge here is required to succeed."
- "I think that this class would be very difficult without programming experience, especially given that the homeworks must be completed in LaTex, which takes an exceptionally long time if you are not familiar with it."
- "Undergraduate level calculus, discrete math, and probability. Intro-level programming."
- "up to linear algebra and able to do the exercises in the R for Data Science book by Hadley Wickham."
- "R experience will be assumed. If you have zero experience (like I did) it's doable, but start learning ASAP, or pray Will/TAs offer some more in-depth tutorials next year!"

- "At least 1 semester of programming in a scripting language, calculus 1, and calculus 2."
- "Calculus 1, Introduction to R programming"
- "Don't think there should be pre-requisites. I technically had all pre-requisites many years ago and never used them until this class. If they can pick up math concepts, they should be allowed to take the course. Perhaps students can try their hand at a quiz instead?"
- "Calculus and R programming"
- "It might be easy if you have a statistic background, but someone who just likes me who doesn't have experience will be hard. so, watch YouTube or Kahn academy or Coursera etc."
- "Calculus and an introductory programming class should be required"
- "Calculus and introductory programming."
- "Some sort of stats class, I didn't take any and it shows"
- "Calc 1 and 2, linear algebra, some form of programming experience."
- "I think it should be basic derivatives and matrix multiplication"
- "comfort with calc, practice with programming, survey of statistics"
- "Intermediate level of mathematics but intensive experience with programming."