BCB720: Introduction to Statistical Modeling

Fall 2024 Syllabus

Last Updated: 2024-08-29

Basic Information

Course identifiers: This document describes the syllabus for BCB720 in the Bioinformatics and Computational Biology Curriculum of BBSP.
Time: 11:00 – 12:15, Tue/Thu
Location: Marsico 2004 (most dates) & Mary Ellen Jones 3116 (9/24 + 9/26 only)
Materials: All learning materials students will be posted on Canvas
Restrictions: Class is limited to 30 students.

Instructors

Instructor: Will Valdar, Email: <william.valdar@unc.edu>, Web: <http://valdarlab.unc.edu/teaching/>
Teaching Assistants: Luvna Dhawka <ldhawka@email.unc.edu>, Emma Klein <emrklein@email.unc.edu>, Kushal Koirala <kkoirala@email.unc.edu>, Katelyn Mcinerney <kamciner@email.unc.edu>, Zhang, Yiyan Zhang <yiyan@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 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: Calc I & II. 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!] *Suggested reading: there are many high school resources and videos explaining calculus at different levels. For a gentle introduction to calculus using R, students might consider Pfaff (2023) "Applied Calculus with R" by Springer, which is freely downloadable through UNC's SpringerLink (eg, search on https://hsl.lib.unc.edu).*

- 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	Section	Lec #	Instr.	Room	Description	HW
1	**Tue-20-Aug**	**Preliminaries**	1 (C)	(TA)	MH2004	Introduction to R and Latex	
	Thu-22-Aug		2	WV	MH2004	Set theory and probability	1
2	**Tue-27-Aug**		3	WV	MH2004	Conditional Probability	
	Thu-29-Aug		4	WV	MH2004	Distribution, Mass and Density functions	2
3	**Tue-03-Sep**					WELLNESS DAY	
	Thu-05-Sep	**Probability**	5	WV	MH2004	Expectation and Variance	3
4	**Tue-10-Sep**		6	WV	MH2004	Discrete distributions	
	Thu-12-Sep		7	WV	MH2004	Continuous distributions	4
5	**Tue-17-Sep**		8	WV	MH2004	Continuous and derived distributions	
	Thu-19-Sep		9	WV	MH2004	Multivariate densities, likelihood, mixtures	5
6	**Tue-24-Sep**		10	WV	*MEJ3116*	Bayesian inference	
	Thu-26-Sep		11	WV	*MEJ3116*	Estimation	6
7	**Tue-01-Oct**					BCB/GMB/Genetics Retreat	
	Thu-03-Oct		12	Sub/Vid	MH2004	Frequentist behavior	7
8	**Tue-08-Oct**		13	WV	MH2004	Confidence intervals	
	Thu-10-Oct	**Statistical Inference**	14	WV	MH2004	Hypothesis testing: concepts	8
9	**Tue-15-Oct**		15	WV	MH2004	Hypothesis testing: Wald, score, LRT	
	Thu-17-Oct					FALL BREAK	
10	**Tue-22-Oct**		16	WV	MH2004	Power and false positive rate	
	Thu-24-Oct		17	WV	MH2004	FWER and FDR	9
11	**Tue-29-Oct**		18	WV	*MEJ3116*	Two-group tests: permutation and t-test	
	Thu-31-Oct		19	WV	*MEJ3116*	Linear models	10
12	**Tue-05-Nov**		20	WV	MH2004	Linear models: estimation	
	Thu-07-Nov		21	WV	MH2004	Linear models: testing	11
13	**Tue-12-Nov**		22	WV	MH2004	LM departures	
	Thu-14-Nov		23	WV	MH2004	Generalized linear models	12
14	**Tue-19-Nov**	**Modeling**	24	WV	MH2004	Linear mixed models	
	Thu-21-Nov		25	WV	MH2004	Hierarchical models and Bayesian regression	S
15	**Tue-26-Nov**		26	WV	MH2004	Model selection	
	Thu-28-Nov					THANKSGIVING RECESS	
16	**Tue-03-Dec**		27	WV	MH2004	Causal inference and Experimental Design	

Syllabus Changes

The 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) "Data Analysis Using Regression and Multilevel/Hierarchical Models" -- *great for understanding linear models, generalized linear models, and estimation, but doesn't really cover hypothesis testing*
Wakefield (2013) "Bayesian and Frequentist Regression Methods". SpringerLink

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 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.

Use of Generative AI: You are not allowed to use Generative AI to create answers to these questions or to in any way solve the problems posed in this homework. Do not paste any these questions into an AI program. The purpose of these exercises is to train your mind to understand, digest, and problem solve; using such aids to do any of the thinking for you negates that purpose and will be a violation of the honor code. You may use AI tools to check your grammar or to help you understand concepts related to the material. Please look out for any updates regarding this policy

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)?

- Difficult. Be prepared to learn! But very helpful!"
- "BCB 720 is a comprehensive stats course covering basic R programming, Bayesian and Frequentist perspectives, MLE, and Linear/Generalized linear models. It takes a ground up approach sequentially building from basic theory up to models used in actual publications."
- "This is a great course to acquire the required theoretical knowledge to perform data analysis from various data, be it genetic data or everything else."
- "This class is a deep dive into how common statistical methods work at a base mathematical level and when/how to use them. It is a great place to start before taking more advanced statistics courses."
- "An incredibly fast paced course covering the basics of most of the field of statistics. "
- "Fast paced and rigorous course applying mathematical concepts to derive and explain statistical methodologies."
- "This is a fast-paced introduction to a lot of the statistics concepts you'll need as someone entering the computational biology field. "
- "A very important course that allows a better understanding of biostats. For a person that has a more biological background and not a mathematical background, it is essential to learn all the concepts discussed in this class. "
- "A comprehensive statistics course that covers most of the statistical topics one would encounter in the biological sciences, taught by a very nice professor who could come across as intimidating because of his 30-sided die."
- "Not hard as it sounds or explained by anyone but don't take it lightly. Assignments are really helpful in understanding the concept but it demands commitment and hours."
- "I would describe the class as very essential for computational analysis in this field. Nearly every aspect of scientific research requires understanding of statistic, probability, and modeling. This class gives an advanced introduction into the field and gets you prepared for the future."
- "A statistical course heavily based in calculus. This class also teaches LaTex and R statistic basics."
- "Good foundational course for statistics, gives a solid breadth of knowledge where you could can build upon as needed in your future research. "
- "You should take it even you are not seeking for the BCB track. "
- "This course has heavy mathematical and programming components and requires you to devote a significant amount of time to the course weekly. "
- "A crash course in stats that will equip with the knowledge to go independently learn more about stats if you desire. "
- "I love this course. It's very systematic. In almost every lecture, I can always learn something that have puzzled me before. "
- "This course will be super helpful, and you will learn lots about probability, modeling, and statistical inference. Do your best with assignments, use the lecture slides, and go to office hours to ask questions!!"
- "This is a full scope course of the uses and applications of statistical modeling for real work examples. This course covers the computation and theory of the most use statistical practices. "
- "A good course to get your foot wet with a wide range of statistical concepts."
- "This course is an introduction to all of statistics. The goal is to give you enough background in any one topic that you have your foot in the door to deeper level classes and so that you have an idea of the basis for a lot of the tests you're running."
- "Comprehensive review of many statistical concepts utilized in bioinformatics. "
- "Incoming students should know that this course is challenging, and they need to give their best effort."

- "Fast paced, for students who are mentally ready to have a busy semester and do some self study for all HWs as well."
- "You will learn a lot, and have a lot of HW, so plan accordingly. Also, the professor is super chilled, so be cool to him too."
- "A dense, fast-paced introduction to statistics for biological research. You will likely learn helpful information that applies to your research project."
- "Fundamental course, suitable for people who don't have any statistical background."

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

- "Don't take this class if you have a heavy course load "
- "Start early on the homeworks! And refresh your calculus/R skills."
- "This course is very hard. However, everything is settled to help you succeed: instructor, TAs, peers, etc. The two most important things are dedication and discipline. "
- "Review basic calc concepts before starting the course if possible. Homework often take several days to complete, start them early! Try your best in in-class exercises, it's an excellent way to check your understanding and ask questions while actively learning the material."
- "I am going to be brutally honest: If you are a second year, don't take this course. There is too much going on in the fall of second year to be able to take this course and actually have the mental capacity to absorb any of it. "
- "Make sure your class schedule is fairly light. Form a study group to discuss the homework and topics in class, as someone may have more math/statistics background than you."
- "Don't psych yourself out. This class has a reputation among BCB but it's quite fair. You can totally do it. Make sure you brush up on R and calculus. Above all else, take advantage of TA office hours!"
- "Pay a lot of attention in class and ask questions when something is not clear and always go to office hours for all HW."
- "Brush up on your R the summer before taking this course. Go to TA office hours - they are very helpful with tackling the homework. Prepare to spend a lot of time on the homework, so start working on them early."
- "The course isn't hard or difficult as we had been told by our seniors but underestimating the course would be a big mistake. The longer theoretical hours test your concentration level at it best. The assignment could be difficult but with the help of TAs and group discussion most of them are easy to solve. Be prepared to invest most of your weekend solving the assignment."
- "I would tell that to prepare a day or two in advance for the class session so that they can ask questions to the professor. Also, to attend TA study sessions and ask for help from peers."
- "Go to office hours."
- "Do assignments as earlier as possible. Preview for the lectures if possible."
- "Familiarize yourself with R PRIOR to starting the course. While you technically don't need to, it's a lot more difficult to learn to program in R and the course content simultaneously."
- "This class has a intense reputation, but if you're willing to put in the work (ie go to class, do the hw, attend office hours if needed, and potentially spend a little time after class reviewing the lectures) it actually is very manageable. In my opinion, it is a well taught and structured course that has a lot of useful information to learn. The beginning definitely feels overwhelming and you'll spend a lot more time on things to start with, but as you learn to speak the language of the class it becomes much more easy. Office hours are an excellent resource. "
- "I didn't use R markdown because it's too slow on my laptop. If you want to write in a style like the homework shows, try LaTeX. You can even embed R into your TeX software. Then you would be able to show the R code as well as the results in your pdf. Remember to setup the LaTeX+R environment before 1st homework. "

- "Seriously ignore what people say about BCB 720. It is simply a normal class, you will be fine. Just do your best, go to office hours, don't be afraid to ask questions in class, and befriend your classmates!"
- "Use the lectures as a main study material - they are very comprehensive and provide a great connection for all the statistic material expected in the course. However, the lectures are not functional without attending class."
- "Don't be afraid. It is not that bad."
- "Start early on the homeworks, join study groups, and review the material where it's confusing."
- "Review your math skills and just pay attention. "
- "Develop strong interest in statistics and get familiar with R programming."
- "1. Have a basic foundation ready of stats and R; 2. Prepare yourself mentally that it is going to too much work and if you are at a beginner stage, you have to work extra hard; 3. Go to TA hours; 4. Discuss concepts with classmates; 5. Attend every lecture even though it might be overwhelming at times"
- "You will have a lot of work to do, plan accordingly."
- "Make sure to brush up on calculus and linear algebra before starting the course if it's been a while since you've taken those classes. Get started on the homework early so you can plan to go to office hours if necessary. Teach and learn from your classmates; they are a great resource."
- "Learn more R."

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

- "Calculus and linear algebra "
- "Calculus 1/Intro Programming"
- "Perhaps adding some tutorials or online classes."
- "Calc I, Calc II (really just basic integrals), very intro level R coding/exposure"
- "I think at least an undergraduate level calculus class should be required. I am not sure about programming because I came in already knowing how to program. "
- "Experience in Calculus and Linear algebra are required as well as some level of R literacy."
- "Mathematics required: basic calculus (high school/college calculus). Programming required: I think you should have some idea of how to program in at least one language, because if you can pick up one, then you can pick up R. "
- "I think that the person has to have had contact with calculus and algebra but it is not necessary to have those skills in an advanced way since during the course you have time to develop those. Programming skills are required but also the course and office hours allow you to learn a lot during the course period."
- "Some background with using R and some exposure to latex. Definitely need to have background in calculus and some linear algebra."
- "A basic brush up with probability, cdf and pmf is good start. Rest of the lectures build upon previous concept so easy to understand. The basis programming should be fine to start the assignment. Apart from assignment 1 and 2, most of them could be solved using R- markdown."
- "1 semester of discrete math; 1 year of calculus (differential and integral); 1 year of statistics and probability; 1 year of programming in any language."
- "You should be comfortable using R before taking this class and should have taken calculus and statistics of some sort."
- "One semester of college calculous and familiarity with R. "
- "Understanding the rules of calculus and Linear Algebra. Perhaps some basic knowledge of probability."
- "Single-Variable Calculus; Familiarity with matrices and matrix operations; R programming"

- "Familiarity with calculus, linear algebra, and R. The more familiarity the more intuitive things will feel at the start, the less familiarity the more work you'll have to put in to make up for the lack of intuition (but this is definitely manageable)."
- "Should have known how to write in Latex or markdown. R is definitely required. I don't think it requires too much math. "
- "It would be an asset to be familiar with calculus + matrix algebra, coding in R, and working with LaTeX."
- "I think a basic level of coding required and Calc 1 or 2. "
- "Calculus I and maybe some matrix algebra? Familiar with basic R and latex?"
- "Have taken calculus 1 (and 2); Have some exposure to linear algebra/matrices; Be comfortable programming in R and Latex"
- "Pre-Req math : Calc and College level Stats; Pre-Req CS: Intermediate level in R"
- "Statistics and R programming will be a key advantage for students who are considering taking this class."
- "a general idea of R, basic mathematics taken in college."
- "Basic calculus is all you need, exposure to matrix algebra is a bonus. Some stats coding would be good, just so it does not become overwhelming."
- "Calculus I is necessary, and Calc II & linear algebra are ideal. Having a background in R is really helpful. I would have struggled a lot more if I didn't already have experience with the language."
- "No pre-requisites."

Q. Which books or other resources (if any) did you find helpful for this course?

- "none "
- "I only used lecture notes."
- "Harvard online statistics courses. All of statistics (book)"
- "My undergrad statistics textbook: "Probability and Statistics for Engineering and the Sciences" by Jay L Devore"
- "I did not use any books, I did however reference Wikipedia if I ever had a question"
- "I did not use any other resources besides TA OH and talking to other students in the course"
- "The lecture slides were the resources that I mostly used."
- "I did not use any of the books. I mostly used the lectures and the help of the TAs, along with Google search results and occasionally ChatGPT to explain concepts with more examples, especially for the application of the different distributions and tests we learned in class, "
- "I didn't follow any course book for the course."
- "Khan Academy "
- "The TAs were undoubtedly the best resource for this class."
- "BIOS 660 and BIOS 661. "
- “N/A”
- "A lot of googling was helpful. There are some good statistics videos on youtube that breakdown some of the topics, though there is not one consistent resource I used regularly outside of the course. (most of what I needed came from the lectures) "
- "I used Wikipedia to easily find the PDFs and parameters of distributions. Other than that, none. "
- "Latex cheatsheet."
- "To be honest, the most useful resource I used were my notes from my Mathematical Statistics II course, but that's not super useful for other students. "
- "I didn't use any."
- "The recommended textbooks are very helpful."
- "Papers from other universities which were created for students. "

- "I did not rely on the texts."
- "Understanding Advanced Statistical Mehtods (Peter H. Westfall and Kevin S.S. Henning), various websites online, Github/Stack Exchange."
- "No need."