BCB720: Introduction to Statistical Modeling

Fall 2018 Syllabus

Last Updated: 2019-03-07

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: Room 2004, Marsico Hall

Materials: All learning materials will be posted on Sakai.

Restrictions: Class is limited to 30 students.

Instructors

Instructor: Prof William Valdar, Room 5113, 120 Mason Farm Road, Genetic Medicine Building, Chapel Hill. Email: <william.valdar@unc.edu> Web: <http://valdarlab.unc.edu>

Teaching Assistants: Dayne Filer <dlfiler@email.unc.edu>, Cody Herron <jcherron@unc.edu>, Andrew Hinton <andrew84@email.unc.edu>, Siyao Liu <siyao@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 to at least Calc I (differentiation and integration in 1 dimension), be comfortable with algebra, somewhat familiar with matrix algebra, and have some programming experience. 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); familiarity with these will be an advantage. Introductory statistics may or may not be an advantage (depending on how it was

taught), but is not assumed. The course will include some material on partial differentiation of multiparameter functions, and so familiarity with this will help also.

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, clustering)

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: The latest date for dropping the course, or, for example, switching to auditor status, is **October 16, 2018** if using the web registration system or **November 21, 2018** if going through John Cornett / official channels. Related: The last day to reduce course load in order to have tuition adjusted is **September 4, 2018.** Details at https://registrar.unc.edu/events/.

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.

Time Table (preliminary)

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

Week	Date	Lec #	Instr.	Description	HW
1	Tue-21-Aug	1 (C)	(TA)	Introduction to R and Latex	1
	Thu-23-Aug	2	WV	Set theory and probability	
2	Tue-28-Aug	3	WV	Conditional Probability	2
	Thu-30-Aug	4	WV	Distribution, Mass and Density functions	
3	Tue-04-Sep	5	WV	Expectation and Variance	3
	Thu-06-Sep	6	WV	Discrete distributions	
4	Tue-11-Sep	7	WV	Continuous distributions	4
	Thu-13-Sep			CANCELLED DUE TO WEATHER	
5	Tue-18-Sep	8	WV	Mixtures and transformations	
	Thu-20-Sep	9	WV	Bayesian inference	
6	Tue-25-Sep	10	WV	Estimation	5
	Thu-27-Sep	11	WV	Frequentist behavior	
7	Tue-02-Oct	12	WV	Confidence intervals	6
	Thu-04-Oct	13	WV	Hypothesis testing: concepts	
8	Tue-09-Oct	14	WV	Hypothesis testing: Wald, score, LRT	7
	Thu-11-Oct			READING DAY	
9	Tue-16-Oct	15	WV	Power and multiple testing	
	Thu-18-Oct			FALL BREAK	
10	Tue-23-Oct	16	WV	FWER and FDR	
	Thu-25-Oct	17	WV	Two-group tests: permutation and t-test	

11	**Tue-30-Oct**	18	WV	Linear models	
	Thu-01-Nov	19	WV	Linear models: estimation	8
12	**Tue-06-Nov**			CANCELLED DUE TO WATER CUTOFF	
	Thu-08-Nov	20	WV	Linear models: testing	
13	**Tue-13-Nov**	21	WV	Regression on non-normal data	9
	Thu-15-Nov	22	WV	Generalized linear models	
14	**Tue-20-Nov**	23	WV	Causal inference and experimental design	
	Thu-22-Nov			THANKSGIVING RECESS	
15	**Tue-27-Nov**	24	WV	Linear mixed models	10
	Thu-29-Nov	25	WV	Bayesian regression	
16	**Tue-04-Dec**	26	WV	Model selection	

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

Brown (2014) "Linear Models in Matrix Form: A Hands-On Approach for the Behavioral Sciences" SpringerLink

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

Harrell (2015) "Regression modeling strategies" -- *lots of good advice for applied work* - SpringerLink

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

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 BCB students

"This course was a good introduction to statistical methods that will be helpful for anyone in computational biology and anyone analyzing scientific data. It seems like a lot of work but you should be rewarded by learning a lot of new and useful concepts."

"An introduction to current statistical methods and their underlying theory. The statistics course you know you should take but really don't want to."

"A great introduction to Bayesian and Frequentist statistics."

"BCB 720 is a comprehensive overview of statistical concepts as applied in biomedical research."

"This is an introductory statistics course focusing on basic probability theory, statistical principles, and modeling with a bit of advanced flavors. Course is largely self contained with no/ little prior knowledge assumed."

"Hard but gratifying."

"Comprehensive statistics course, background in statistics and linear algebra required"

"An advanced introduction to the basic concepts of statistics. Not only theory, but not only applications. A thorough survey of statistical methods."

"Introduction to probability and statistics with a focus on linear models. "

From non-BCB students

“The time commitment needed for homework is sizeable - ideally start early.”

“This is an essential course for anyone interested in computational and statistical sciences. If you put in the effort, you are guaranteed to learn a lot of useful skills and intuitions about statistics–one of the best courses you will take.”

“This course seeks to bridge a much-needed gap between elemental statistics and advanced quantitative analysis of data. Its focus is on the practical application of statistical modeling concepts to better understanding biomedical phenomena.”

“Statistics course on how to think about data. More theoretical than applied. Difficult.”

“This course provides a good primer for those interested in using statistical methods in their future research.”

“The course is n broad overview statistical concepts that will be time intensive and challenging but will pay off, especially if you need to use models in your research.”

“Technically an ‘intro’ course, but really a deep dive into the mathematical underpinnings of statistical modeling (using R, sometimes). Hard. Lots of work. Not for the faint of heart.”

Advice

From BCB students

Start doing homeworks in advance

Don’t take any other time consuming courses with this course. Homeworks 1,2, and 5 are the most difficult and they get easier with time.

Review your calculus. Try to apply the things you learn to your research rotation. Join a study group, and discuss the questions. Also, talk about potential uses for the things you’re learning.

Three things: 1. Learn R beforehand and write out every homework in LaTeX: its time-consuming but it helps. 2. STUDY GROUPS are life! 3. Be wary of SIDS...

Brush up all requisite calculus that you will need before hand/early on

I would say read the problem sets when they become available. I would take the time to review the lecture notes soon after class. I would encourage the students to establish study groups, and make sure to go to office hours.

Better to have some pre-knowledge of statistics before taking this class.

I found printing out the slides to be very helpful. I also found comfort that the class becomes more practical and less theoretical as the semester progresses.

I strongly recommend incoming students to preview the lecture notes before class. Also, I would read the recommended textbooks or look for supplemental materials online to help understanding the lectures.

Homework assignments are critical and can be time consuming. Plan ahead and DO NOT get behind! Do not underestimate the rigor of this course based on previous statistics courses, chances are there will be things in this course that you will either learn for the first time or things that will be treated with more rigor than in any previous exposure. This isn’t your typical statistics course memorizing formulas and applying, it is a broad, well developed, rigorous theoretical treatment of major topics in statistical modeling.

Brush up on your basic algebra, matrix math, and learn to code in R before the class.

From non-BCB students

Take your time with the homework, because a lot of the value is found in the homework. Start early and go to TA office hours. They are definitely helpful, and starting early helps you think about the material more deeply.

Study groups are your friend! Formal study groups or having a buddy to discuss things with can really help you be sure you're grasping the content and questions well

They say from the get-go to start homework early and don't put them off but I'll say again; start homework early and don't put them off! Give yourself a few hours to type everything up in Latex, too. (And if you see a problem on SIDS, start looking into it ASAP!) Try and look into the homework as soon as possible, even if you don't start the work yet; it can help when you're in lecture if you know what you do and don't understand, and it gives you a chance to ask Will questions after class or go to TA office hours (which were always super helpful).

Work in groups. Don't have the expectation of retaining everything covered in much depth.

Read the homework sets as soon as they come out. Half the battle is understanding what the questions are actually asking you. Brush up on your algebra. The lectures are interesting but often not that helpful in completing the assignments.

It's a lot of work, but certainly worth the effort.

Don't take two other courses at the same time as this one (like I did).

Start the homework early just so you have enough time to deal with difficult questions.

"Make sure you have a level of programming skill such that even if you don't know all of the details of using R or LaTeX, you know you to ask and find answers to questions about them."

"Reviewing the lecture material before it's presented helps start the thought process that is required to get the most out of class time. Allocate considerable time to wrestling with homework problems, consulting multiple online resources to reason your way through."

"Do a lot of reading/studying/googling/watching videos on topics."

"Everything builds so make sure you have a solid understanding of the foundations from the first few lectures."

"I would suggest that students not be intimidated if they're new to statistics or have taken a long break from math courses. Utilize the resources that the course professors suggest and use Google. Don't expect to understand everything the first time around (or maybe even the second or third time), but try to grasp the bigger picture. Search for other explanations or definitions for terms or concepts presented in the class – it can be helpful to hear things presented in multiple ways. Don't be afraid to ask questions."

"Spend time trying to understand the concepts before working on the HW (re-read the lecture notes, find online tutorials) to be more prepared. Start the HW early so you can think about the tough questions for multiple days and go to office hours with specific questions. Go to office hours."

"Don't be afraid to answer questions in class. It is better to try and get the answer wrong than to sit there and not try. You will get out of this course what you put into it. It is very unlikely you will ever have the ability to learn this material in this way again, so take advantage of the opportunity if you think this will be relevant to you now or in the future. Review single variable calculus differentiation/integration, summation algebra, and matrix algebra at the beginning of the course. Give yourself several days to complete the homework assignments. It will be worth your time. Use internet resources/books from different places. Sometimes reading information explained in multiple ways can help clarify something confusing."

"Keep up with the readings"

“Review the lecture slides before working on the assignments”

“During the linear models part try using Gelman and Hill. It has a very didactic way of teaching, especially for people interested in applied statistics.”

“Be prepared to put in work. This was easily the hardest, most time intensive course that I have ever taken. Be warned it is a lot more work than anticipated.”

Selected comments and advice from previous years’ students

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

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

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

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

“Make sure that you start the homework’s earlier in the week and be prepared to spend a good hour or two each night, unless you want to spend a day out of the lab on Friday frantically trying to finish it.”

“1. Start homeworks early so you know what to expect. 2. You won’t actually know what to expect, because some innocent looking questions will take you hours.”

“I have never experienced more rigorous expectations out of a class with ”Introductory” in its title. Do not shrug off the importance the syllabus places on time investment. There are no tests in this class, so you should expect your knowledge to be challenged to a similar, if not higher standard on homework assignments. Take the time to learn R basics in the beginning of the course before it becomes increasingly complex. Set aside large blocks of time to tackle the homework. You will not be able to complete them the night before they are due and expect a passing grade. Also, take advantage of outside resources: TA office hours and outside reading helped my understanding tremendously.”

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

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

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

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

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