BCB720: Introduction to Statistical Modeling

Fall 2023 Syllabus

Last Updated: 2023-08-18

# 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](mailto:william.valdar@unc.edu)>, Web: <<http://valdarlab.unc.edu>>

**Teaching Assistants:** Sean Connelly <[sean\_connelly@med.unc.edu](mailto:sean_connelly@med.unc.edu)>, Nolan Hamilton <[nolan.hamilton@unc.edu](mailto:nolan.hamilton@unc.edu)>, Alec Lobanov <[lobanov@email.unc.edu](mailto:lobanov@email.unc.edu)>, Ryan Videgar-Laird <[ryanvl@unc.edu](mailto:ryanvl@unc.edu)>

**Student Services Manager:** John Cornett <[jcornett@email.unc.edu](mailto: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 (), products () and relations such as , 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.** | **Description** | **HW** |
| **1** | **Tue-22-Aug** | **Preliminaries** | 1 (C) | (TA) | Introduction to R and Latex |  |
|  | **Thu-24-Aug** | **Probability** | 2 | WV | Set theory and probability | 1 |
| **2** | **Tue-29-Aug** | 3 | WV | Conditional Probability |  |
|  | **Thu-31-Aug** | 4 | WV | Distribution, Mass and Density functions | 2 |
| **3** | **Tue-05-Sep** |  |  | WELLNESS DAY |  |
|  | **Thu-07-Sep** | 5 | WV | Expectation and Variance |  |
| **4** | **Tue-12-Sep** | 6 | WV | Discrete distributions | 3 |
|  | **Thu-14-Sep** | 7 | WV | Continuous distributions |  |
| **5** | **Tue-19-Sep** | 8 | WV | Continuous and derived distributions | 4 |
|  | **Thu-21-Sep** |  |  | BCB RETREAT |  |
| **6** | **Tue-26-Sep** | 9 | WV | Multivariate densities, likelihood, mixtures | 5 |
|  | **Thu-28-Sep** | **Statistical Inference** | 10 | WV | Bayesian inference |  |
| **7** | **Tue-03-Oct** | 11 | WV | Estimation | 6 |
|  | **Thu-05-Oct** | 12 | WV | Frequentist behavior |  |
| **8** | **Tue-10-Oct** | 13 | SUB/Vid | Confidence intervals | 7 |
|  | **Thu-12-Oct** | 14 | WV | Hypothesis testing: concepts |  |
| **9** | **Tue-17-Oct** | 15 | WV | Hypothesis testing: Wald, score, LRT | 8 |
|  | **Thu-19-Oct** |  |  | FALL BREAK |  |
| **10** | **Tue-24-Oct** | 16 | WV | Power and false positive rate |  |
|  | **Thu-26-Oct** | 17 | WV | FWER and FDR |  |
| **11** | **Tue-31-Oct** | 18 | WV | Two-group tests: permutation and t-test | 9 |
|  | **Thu-02-Nov** | **Modeling** | 19 | WV | Linear models |  |
| **12** | **Tue-07-Nov** | 20 | WV | Linear models: estimation | 10 |
|  | **Thu-09-Nov** | 21 | WV | Linear models: testing |  |
| **13** | **Tue-14-Nov** | 22 | WV | LM departures | 11 |
|  | **Thu-16-Nov** | 23 | WV | Generalized linear models |  |
| **14** | **Tue-21-Nov** | 24 | WV | Linear mixed models | 12 |
|  | **Thu-23-Nov** |  |  | THANKSGIVING RECESS |  |
| **15** | **Tue-28-Nov** | 25 | WV | Hierarchical models and Bayesian regression |  |
|  | **Thu-30-Nov** | 26 | WV | Model selection | S |
| **16** | **Tue-05-Dec** | 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](http://search.lib.unc.edu/search?R=UNCb5202600)

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)](http://www.valorebooks.com/textbooks/applied-multivariate-statistical-analysis-6th-edition/9780131877153)

Gentle (2007) “Matrix Algebra: theory, computations, and applications in statistics” –[SpringerLink](http://search.lib.unc.edu/search?R=UNCb5320513)

Harrell (2015) “Regression modeling strategies” -- *lots of good advice for applied work* - [SpringerLink](http://link.springer.com/book/10.1007/978-3-319-19425-7)

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](http://search.lib.unc.edu/search?R=UNCb5951811)

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)?***

* "A solid general introduction to most of the stats that you will ever need."
* "Trial by fire intro to all of the statistical theory you'd need to succeed in Dr. Valdar's lab. 20-40% will directly pertain to any given student's research. "
* "The course covers variety of topics in statistics which will help getting into more advanced statistics topics. If no statistics or math background, the class might consume too much time to fully understand and get work done on time."
* "Comprehensive course that will set you up with a good foundation for taking related / more advanced statistics courses. "
* "A broad and rigorous introduction to statistical tools that will be useful for biologists. This class reviews a wide range of topics related to statistical modeling, including probability theory, bayesian and frequentist methods, and linear modeling. "
* "A crash course on 'real' stats that is challenging, useful, overwhelming, and fair. It moves very quickly, but is not as bad as some horror stories suggest. "
* "This is a wide-reaching statistics course that will cover nearly anything you could think to need bioinformatically, at enough depth to be useful with enough theory to help you learn more on your own if you want to."
* "Be prepared"
* "A difficult course that was veery rewarding and helped me understand statistical concept and get much better at R, which has helped me in doing analysis of my own data."
* "This course serves as a foundation for statistical knowledge that you can use and/or build upon."
* "An overview of the theory and math behind frequently used statistical methods that you will certainly encounter in your work and reading."
* "A stats course taught for people trying to know enough to barely get by."
* "This class provides a systematic overview over statistical concepts, modeling and analysis that might be helpful to you in application of the techniques."
* "This is the statistics class for people that took the statistics for science majors class in undergrad. Learn the math behind all your ~favorite~ statistics concepts (probability, likelihood, and linear models) and develop a strong base for continuing your computational biology education at UNC. "
* "An exhaustive introduction to computational statistics, covering concepts that address many problems you will come to face in research."
* "A super condensed, jam-packed statistics course presenting a wide-range of topics suitable for any BCB student."
* "BCB 720 is a high-paced course that provides the statistical background necessary to understand the basics of model building in data science. "
* "This course teaches you the statistics that will actually be helpful to you as a graduate student.  It's fast paced and covers lots of material, but will make reading papers easier."
* "A broad yet in-depth introduction to statistics"
* "A phenomenal in-depth crash course on a wide-range of foundational statistical concepts. A stats class like no other."
* "If you want to be a statistical programmer, it's a good course."
* "This course is going to take more time than you expect. Unfortunately, most of the topics will also be irrelevant to your current research. To preserve your sanity, I would focus on only trying to learn the lectures which actually pertain to your research, and simply get through the others. "
* "A broad overview of probability, point estimation, hypothesis testing, and linear modeling relevant to biomedical research."
* "Introduction to probability and statistical modeling."

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

* "Learn LaTeX and refresh on calculus before the first homework."
* "Definitely go through the lecture materials multiple times if you struggle to understand a specific topic. Also, always use Google."
* "The "start the homework" early suggestion shouldn't be taken lightly; even if you understand all the course concepts and won't struggle with the homework, it still takes a large amount of time and you won't be expecting it, it will take you some time to learn how long it takes. "
* "Complete the preparatory exercises before the first class and get latex working. The first homework will be a lot more difficult if you don't. Be ready to spend 15+ hours on some of the first homeworks if you aren't already good at stats. The homework is less time-consuming as you go along. I think my PR was 2-3 hours. Find a study group early. You will get free friends and can help each other when you are stumped!”
* "You do not need any advanced math, but would greatly benefit from being well-versed in the basics. Go to office hours and form a study group, both will serve as sanity checks and help prevent wasted time. "
* "Be comfortable with algebra and calculus, preferably before you start but if you have to learn it on the fly it's doable. Be comfortable with base R, and keep a LaTeX reference handy. Just accept that the class requires a lot of time, and just get through it."
* "I would NOT recommend nursing students taking this course."
* "Definitely go to the TA office hours they were great and super helpful in explaining concepts or clarifying what the hw questions were looking for."
* "Do the prep assignment. Go to office hours. Make a study group."
* "Office hours are helpful. Pay attention and ask questions during class."
* "don't."
* "Start the assignments early, and think about what are the relationship between the simple guiding questions and the topics in class."
* "Form a study group and try to meet weekly, even when you think you fully understand the homework. If you get stuck on something don't spend too much time trying to figure it out on your own. Reach out to your peers/study group--chances are at least one person has an idea of how to at least start the problem. "
* "Refresh your calculus knowledge (derivatives, integrals, etc). This was the biggest hurdle I think. "
* "It is a lot of work, but it is absolutely worth it at the end."
* "Go to office hours and make study groups"
* "Work in groups and organize study sessions.  Take advantage of office hours.  Don't let yourself spend to long on one homework question."
* "Watch a few videos on probability density functions and the calculus behind them. Practice matrix algebra before you get to the 2nd half of the course. The homeworks take several hours to complete, so start them as soon as they're released."
* "The assignments are very time-consuming, so try to start them early. I also really benefited from watching the recorded lectures."
* "Be familiar with calculus and have a basic concept about linear models."
* "Learn from your peers"
* "Don't be intimidated by the comments on the syllabus from previous students.  I was quite stressed about the course beforehand due to these comments, but I've found that 720 has an unfair reputation among BCB students due to a vocal/dramatic minority of students with limited backgrounds in math/coding."
* "I would tell students to not give up and to not feel discouraged if the course seems impossible at the beginning. The probability, algebra, and calculus parts were very difficult but when we start talking about testing and modeling I actually started enjoying the course because I could easily see an application of what I was learning. I would also recommend they create study groups with the rest of the class on the 1st day. "

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

* Some background in calculus and linear algebra. Would be nice to have taken the lower level undergrad courses in these topics. Some experience in programming would be great. R is fairly easy to pick up."
* "Must be comfortable manipulating equations, basic calculus, and linear algebra. Should have worked in R before or have fairly extensive experience in other languages. "
* "Calculus 2, linear algebra, beginner level R."
* "Calculus and exposure to matrix algebra. Basic programming and exposure to R. "
* "Basic coding experience (a class or just worked with something involving coding before. Calculus 1. "
* "Be comfortable with calculus 1, exponents, logs, sums/products, and basic linear algebra.
* Be familiar with basics of R/Rmarkdown (or Quarto). The base library is basically all that is needed. "
* "This course requires college algebra and calculus 1. Linear algebra is preferred and a basic programming course may be helpful but is not necessary."
* "Some R programming, calculation, matrix are needed (please do not assume that those knowledge will be well covered in this course)."
* "I think calc1/2 is a good prereq math wise. Programming wise I probably needed more R skills than I had, but the class really helped me improve them"
* "I think basic calculus and algebra."
* "Some R, Some calculus. Some linear algebra is helpful."
* "Know so little that you just accept everything told in lectures but know R pretty well and latex semi-well."
* "The students should be at least familiar with the most basic R programming (creating vectors and data frames, ranking, reading files), intermediate knowledge of integral and differentiation, and can do basic matrix calculation."
* "Integrals, derivatives, basic probability, familiarity with a programming language (preferably R). For a more advanced understanding: partial derivatives and matrix algebra. "
* "Calculus, at least up to level II (NC school system) and Linear Algebra. This would cover most mathematical properties needed for this course. As for programming, intermediate R knowledge will be helpful."
* "Linear Algebra, Calculus I and II. Basic programming skills, no need for a specific language."
* "One basic introductory statistics course; calc 1 and 2; some sort of coding background"
* "Calculus experience necessary.  Some programming and linear algebra experience is helpful."
* "Calc I required, Calc II optional, but recommended. Intro level R. "
* "Calculus I and an introductory statistics class."
* "Have learnt R before."
* "Intro-level proficiency in R and a documentation language (latex, Rmd, etc.). Matrices, algebra, calculus, differential equations"
* "Calculus I and some coding experience, preferably in an interpreted programming language. Additional courses in Math/CS are obviously helpful, but I don't think they're necessary as formal prerequisites. "
* "I think R experience should be required."