Course Syllabus and Information
Biostatistics 201B

Course Times and Locations

This course meets Monday, Wednesday and Friday mornings from 9:00-9:50 in CHS 43-105A. There is also a TA-led discussion section each week on Friday from 10:00-10:50 a.m. in 43-105A and 3 lab sections which meet in the computer classroom, CHS A1-241, Thursdays from 10:00-10:50 and 12:00-12:50 and Fridays from 11:00-11:50. If you have a scheduling problem with your current lab session, please let me know as soon as possible. In general you may attend more than one lab provided there is sufficient space. However, please attend your assigned lab initially so we can get a reasonable headcount.

Instructor: Catherine Sugar

Office: CHS 51-236C
Phone: (310) 794-1078
FAX: (310) 267-2113
E-mail Address: csugar@ucla.edu

Office hours: My official office hours are Mondays from 10:00-11:00 after class and Fridays from 1:00-2:00. However, I am always happy to answer questions via e-mail or to meet with you at times outside of office hours. Because of my other obligations this quarter it would be helpful if you e-mail ahead to make sure I am free if you want to see me at a time outside normal office hours.

Teaching Assistant:

We are fortunate to have Nada Abdalla, whom most of you will remember from 201A, as our teaching assistant this quarter. Nada will be running the labs and discussion sections, holding office hours, assisting you with the computer packages and homework and of course helping me out with the grading. Her information is as follows:

Office Location: CHS A1-279 (Biostatistics Consulting Lab)
Office Hours: Thursdays, 11:00-12:00.
E-mail Address: n_a_abdalla@yahoo.com

Text and Prerequisites

There is no required textbook for the class; all the materials you will officially need, including my scanned lecture notes and handouts, will be available on the class web site. However it is frequently useful to have additional references. The *Primer of Applied Regression & Analysis of Variance* (either the second or third edition) by Stanton A. Glantz and Bryan K. Slinker which served as the optional text for Biostatistics 201A last quarter has sections on a number of the 201B topics as well. It is available at the bookstore and copies will also be on reserve in the Biomed Library. However, some of the material we will cover this quarter is more advanced or specialized and I will provide links to additional useful texts and papers on these topics on the class website throughout the quarter. Many of these are already posted if you want to start doing some reading. The prerequisite for this course is Biostatistics 201A or another equivalent graduate level course in applied data analysis and regression modeling. This prerequisite is very serious; the class will begin where

201A left off last quarter and will build heavily on the linear modeling framework. If you are unsure whether you have the necessary background for this course please contact me right away.

Course Objectives

Statistical modeling, which deals with mathematically describing the relationship between a target variable and a set of predictors based on data, is a central part of many research projects. Biostatistics 201A, the first course in the graduate Applied Regression sequence, focuses on the standard multiple linear regression model. This model is very flexible and can be used successfully in many applications; however it makes a number of core assumptions about distributions (of variables, error terms and parameter estimates), independence of observations, the forms of the relationships among the variables and the like which will not always hold. In Biostatistics 201B we will learn how to modify or extend the linear regression model in a variety of ways. Major topic areas include

- **Non-parametric and Semi-parametric Methods:** These techniques are useful when you do not know or do not wish to make assumptions about the underlying distributions of the key terms involved in your analyses. Examples include classical techniques such as the Spearman rank correlation, Wilcoxon rank-sum and signed-rank tests, and the Kruskal-Wallis test (which generalize correlation, t-tests and ANOVA respectively) as well as more modern techniques such as smoothing and simulation-based methods, including permutation tests and the bootstrap.

- **Generalized Linear Models:** The standard multiple regression model assumes that the average value of a (continuous) outcome variable can be written as a linear combination of predictor variables plus normal error. If the outcome variable is not continuous (e.g. in the case of categorical or count data), the distribution of the errors is not normal, or the relationship of the predictors to the outcome is not linear on the original scale, then the model will need to be modified. The generalized linear model framework handles all of these possibilities and includes as examples logistic regression (standard, ordinal, generalized ordinal, nested), multinomial regression, probit regression, Poisson regression, negative binomial regression and many others.

- **Repeated Measures and Mixed Models:** In many studies one has multiple measurements on each subject or experimental unit, either under different test conditions or at multiple time points. Measurements within subjects tend to be correlated, violating the assumption of independent errors in the standard regression model. This can be handled by directly modeling the structure of the relationship between the measurements or by introducing "random effects" to capture the within subject variability. Terms you will hear for such approaches include repeated measures ANOVA/regression, generalized linear mixed models, multi-level models and hierarchical models. We will learn the basics of dealing with such data, although complete coverage is well beyond the scope of this course.

- **Unevenly Sampled, Partially Observed or Missing Data:** Ideally one's data consist of a representative random sample from the population of interest and each subject has equally reliable values for each variable. However in practice this is not always the case. Weighting can be used to adjust for differing degrees of uncertainty (non-constant variance) in the measurements; propensity score methods can be used to help reduce selection bias; survival analysis and related techniques (Kaplan-Meier curves, Cox regression, etc.) are used to handle censored data (situations in which there is only partial information about a variable for some subjects); and imputation methods (mean or regression substitution, hot-deck techniques, multiple imputation) can be used to handle missing values, making it possible to include subjects with incomplete data in standard analyses. This course will provide a basic introduction to these topics, adding additional detail as time permits and student interest dictates.

This course is primarily designed for masters and doctoral students in fields outside of biostatistics or for those intending to be consulting biostatisticians and will have a heavy emphasis on practical applications as opposed to theoretical development. A detailed outline of the lecture topics will be provided in a separate

course schedule handout. The associated ASPH (Associated Schools of Public Health) learning objectives and competencies are listed at the end of this syllabus.

Computing

Because the focus of this class is data analysis and statistical modeling, we will make considerible use of statistical computing programs. Through the labs and discussions we will continue to build your skills in STATA and SAS, two of the most widely used packages. In general we will allow you to use the package of your choice although sometimes one package will handle a particular technique more easily or provide more detailed output and will thus be recommended. The necessary instructions in both packages will be provided as part of the homework assignments. STATA and SAS are both available on the computers in Technology and Learning Center (TLC) lab in the Biomed Library and the CLiCC lab, or you can purchase a copy for use on your own computer. For additional campus locations and other statistical resources see the links on the class web page. The is also a useful set of tutorials on the Institute for Digitial Research and Education (IDRE, formerly UCLA Academic Technology Services) website, accessible at either http://www.ats.ucla.edu/stat or http://www.idre.ucla.edu/resources/stats.

Handouts and the Class Web Site

Our class web site is **http://csugar.bol.ucla.edu/Courses/201bwinter2017.html**. I will post all course materials including the assignments, solutions, lecture notes, practice exams, class notices, etc. on this site, so make sure you check it regularly.

Homework

I consider homework to be the single most important component of this class. It is extremely hard to learn statistics, or enjoy it, without working through practical examples. Assignments will be made and turned in roughly every one to two weeks. They will consist of a combination of mathematical problems and mini data analysis projects. **The write-up is as important as getting the "right" answer.** Your homework should always include English explanations of what you are doing, why you are doing it, and what the analysis allows you to conclude. If you do not do so, you WILL lose points! An attempt will be made to choose problems and examples from various areas of public health, medicine and current events to make the class more interesting and relevant.

Homeworks will generally be due on either Mondays or Wednesdays to allow people time to work through the problems after attending the corresponding labs and discussion sections. The homeworks will nominally be due in class but may be turned in to Nada's TA folder in the Biostatistics Department Office, CHS 51-254, up until 4:00 pm on the due date. Other than this grace period, no late homework will be accepted except in extraordinary circumstances or with my prior approval. You may drop your lowest homework score. Solutions to warm-up problems will be posted on the web when the assignment is handed out and solutions to the turn-in problems will be added after the assignments are handed in. Graded assignments will be returned in class and/or your lab session as soon as possible after the due date–in most cases in about a week. Concerns about grading should be reported directly to me. Changes in scores will not be made more than two weeks after an assignment or exam has been returned. Unclaimed assignments will be kept in a box in my office up until grades have been submitted at the end of the quarter and then discarded.

Group Project

In addition to the homework assignments, there will be a final project consisting of an in depth analysis of a real data set. The projects will be done in groups of 3-4 rather than individually and each group will select its own data set. The data set will need to be suitable for illustrating one or more of the major analytic techniques we will learn over the course of the quarter. Detailed instructions will be handed out later when

we have learned more about the core methods but it is a good idea to start thinking now about possible data sets. The project will be due in the final week of the quarter.

Academic Integrity

Academic integrity is an important part of university life, and will be taken seriously in this class. You may work on the homework assignments with other students. In fact, interaction with your classmates is strongly encouraged. **However, each student must write up each assignment on their own and in their own words.** There is to be no collaboration during examinations.

Students with Disabilities

Students needing academic accommodations based on a disability should contact the UCLA Center for Accessible Education (formerly the Office for Students with Disabilities) by calling (310) 825-1501 or going to A255 Murphy Hall. Please contact the CAE as early in the quarter as possible, preferably within the first two weeks, so that I can coordinate with them to make any necessary arrangements. For more information visit the CAE website at http://www.cae.ucla.edu.

Exams

There will be a midterm and a final exam in this course:

Midterm: Friday, February 17th, during lecture/discussion: 9:00-10:50 a.m. (110 minutes)
Final Exam: Thursday, March 23rd, 11:30 a.m.-2:30 p.m. (180 minutes), location TBA

The exams are closed book, closed notes. However, you may bring 2 sheets of paper (8.5 by 11) with formulas and notes on both sides to the midterm, and your midterm notes plus two additional sheets to the final. You should also bring a calculator and writing instrument. Any additional materials, including numerical tables, will be provided unless specified in class prior to the exam. In general I do not give late, early, or repeat exams except where required by university policy. However, I know that students in this class sometimes get the chance to attend a major conference or give a research presentation during the quarter. If such an opportunity comes up and it conflicts with an exam please let me know as early as possible and I will try to work out some sort of arrangement with you.

Grades

Grades will be based on:

20% Homework and mini data analysis projects. (There will be roughly 5-6 graded assignments, weighted equally unless specifically noted otherwise, and you may drop your lowest score.)
25% Midterm
20% Group Project
35% Final Exam

If it is to your benefit, you may drop your midterm score to 10% and increase the weight of the final exam to 50%. You will automatically be given the optimal score; there are no decisions to make in advance. It is also worthwhile to participate in class. Although it is not part of the numerical grading formula, I do take effort and participation into account when detemining the grade cutoffs.

Learning Objectives and Competencies

Learning Objectives	**Association of Schools of Public Health (ASPH) Competencies**
To understand and apply a wide variety of statistical modeling techniques as a means of addressing scientific and public heatlh issues.	A.4. Distinguish among the different measurement scales and the implications for selection of statistical methods to be used based on these distictions. A.6. Apply common statistical methods for inference. A.7. Apply descriptive and inferential methodologies according to the type of study design for answering a particular research question.
To interpret the results of such models appropriately.	A.9. Interpret results of statistical analyses found in public heatlh studies.
To understand the assumptions underlying major statistical techniques such as generalized linear models, use diagnostic techniques to assess model fit, and apply corrective actions or alternative techniques as needed.	A.3. Describe preferred methodological alternatives to commonly used statistical methods when assumptions are not met.
To gain competency in using software to run a wide range of statistical models.	F.8. Use information technology to access, evaluate, and interpret public heatlh data.
To refine professional communication and collaborative skills	A.10. Develop written and oral presentations based on statistical analyses for both public health professionals and educated lay audiences.