The University of North Carolina at Chapel Hill

Spring Semester 2017

BIOS 669-001

Working with Data in a Public Health Research Setting

I. Time and place — TuTh 9:30-10:45 PM, McGavran-Greenberg 2306

II. Instructor

Kathy Roggenkamp
Research Instructor and Manager of Statistical Computing
Dept. of Biostatistics and Collaborative Studies Coordinating Center
Office: Room 206, Suite 203, CVS Plaza, 137 E. Franklin St.
Phone: 919-966-5304
E-mail: kathy_roggenkamp@unc.edu
Office hours: By arrangement (after class will generally work, and we can always talk during the class period)

III. Textbook — None

IV. Course description

This three-credit course has a target audience of MS or MPH-seeking biostatistics students who are in their second semester of study. It aims to provide a conceptual foundation and practical training to these students in various aspects of working with data, since they will be using data from clinical trials or other public health research studies while in graduate school and after graduation. Topics include using SAS and SQL to transform data into structures useful for analysis, producing typical reports, working toward study data of high quality, and simulation with SAS. Prerequisite: BIOS 511, EPID 700, or permission of the instructor (basically, solid knowledge of SAS DATA step programming and familiarity with the SAS macro facility)

V. Course format

- Students will be expected to prepare for class by reviewing materials as specified by the instructor in the detailed course schedule. Such materials could include articles, instructor-provided notes, or prepared videos. All materials will be provided on Sakai.
- Since students will have reviewed relevant material before class, most class time will be spent working on problems.
- **To perform these activities, students will be required to bring a laptop or other portable device to class.** Students will be encouraged to use a BIOS 669-specific or other university-provided virtual session for running SAS and other software, or they can

use software on a UNC or BIOS cluster, or they can use software loaded locally on their device (regular SAS or SAS University Edition).

- Students are required to attend class and participate in class activities.
- Assignments will generally be turned in twice a week (electronically via Sakai). They will be evaluated promptly by me via comments and on a √-based scale (√ = meets expectations, √+ = exceeds expectations, √- = fails to meet expectations). See more details on grading at the end of this document. Assignments will be posted at 3 PM on the day preceding a class meeting day (Monday or Wednesday) and due at 7 AM on the day after class (Wednesday or Friday) – at least that is the posting schedule we will try to begin with. Assignments turned in late will receive a grade of √-. All assignments will be graded and returned via Sakai before the next class session (at least I will make every effort to do so, though one or two assignments might defeat me).
- There will be no written exams during the run of the course.
- Given the applied nature of this course, a final project of the student's choosing (with my approval) will take the place of a traditional final examination. The project will be due through Sakai electronic submission at the course's scheduled final examination time of 8:00 AM on Friday, May 5, 2017. We will meet as a class on the morning of May 5, and students who would like to make a ten-minute presentation based on their project will be welcome to do so – this could help your project and course grades. In terms of computing your course grade, the final project will be the equivalent of five regular assignments.
- Comments and contributions that will enhance the course in future years will be greatly appreciated.

VI. List of topics [1 class period per exercise unless otherwise noted]

- SAS refresher and an introduction to the METS clinical trial (METS data will be used for many course exercises) [2 exercises]
- PROC SQL, including an introduction to relational databases [6 exercises]
- Look-up tables [1 exercise]
- Analysis data sets and variables, including data cleaning, combining data, deriving and checking variables, producing needed data structures, and using an external macro to look at excluded observations [3 exercises over 5 class periods]
- PROC REPORT and general reporting concepts (including use of an external macro) [4 exercises]
- Metadata, including codebook production [3 exercises over 4 class periods]
- Calling R from SAS [1 exercise]
- An introduction to web scraping [1 exercise]
- Simulation in SAS, using both base SAS and IML [2 exercises]
- Data-driven programming [1 exercise]

- Speaker Virginia Pate on her programming for a typical big health data manuscript [1 class period, no exercise]

Grading in BIOS 669 for spring semester 2017

BIOS 669 is about learning and practicing techniques for handling data, mostly in SAS, beyond those SAS techniques covered in BIOS 511. You should be enrolled in the course because you want to learn by doing and possibly because you've heard from other students that the course is fun and useful. For most 669 assignments, there will be no absolute best answer in terms of technique. There will be many ways to approach most problems, and a goal of the assignments is for you to try things out in a low risk situation.

So, what about grading? My assumption in the following is that course grades mean this: H (or A) = clear excellence/exceeds expectations, P (or B) = entirely satisfactory/meets expectations, L (or C or D) = low pass/shows effort but fails to meet expectations, F = fail (consistently failing to turn in work or show effort). This grading plan is in line with the fact that the SPH grading system is designed so that the mode of the grading system is P.

On most 669 assignments, most of you will provide the "right answer" in terms of the SAS output. While not getting the "wrong answer" in terms of the SAS output is important, that will not be the main criterion for evaluating the quality of your work in this course. For getting the "right answer" (and sometimes even if you got the "wrong answer"), you will receive a check (√) on an assignment as long as I can tell that you put good effort into the work and made a reasonable attempt to apply techniques covered in the notes, videos, and readings to solve the assignment's problems. At the end of the day, a check would be equivalent to a P (for grad students) or a B (for undergrads) – you have met expectations.

For exceptionally nice work on an assignment (you exceeded my expectations), I will assign a check plus (√+) rather than a check. What might justify a √+? Basically, indications that you were highly engaged in the work and did a great job as a consequence. Such indications could include the following:

- Writing beautiful, elegant code
- Producing a beautiful table or graph
- Doing a task in multiple ways when that wasn't requested
- Using SAS or SQL options in particularly creative ways or using options that weren't covered in the notes – you might teach <u>me</u> something!
- Writing questions in your program comments that illustrate how thoughtfully you approached the work or making suggestions for how to improve the assignment in the future
- Something about your code makes me want to share it with the class as part of the posted solution (many of which show multiple ways to accomplish a task)
- Basically, things that made me smile or say "wow" as I was going through my stack of student work

I will also feel free to assign a check minus (√-) for work that shows a lack of effort and/or thoughtfulness (you failed to meet expectations).

From past experience in reviewing assignments for this course, I would estimate that 10-20% of work turned in would be √+ worthy, and not much work would call for a √-.

The current course schedule (which could change slightly) has you doing 24 assignments over the course of the semester, along with a final project (of your choice) that will be worth the equivalent of five regular assignments. Final projects will be evaluated on the basis of both ambition and execution as well as whether you presented your work to the class. As I compute grades, a √ will earn 2 points, a √- will earn 1 point, and a √+ will earn 3 points. When points are added up, I will look for a reasonable cut-point to distinguish outstanding performance from regular performance. I will also consider classroom participation – how much you contributed to making the class go well – in assigning final grades. **In the end, I expect about 30% of students to earn an H or A rather than a P or B** – and I hope and expect that no one will earn a lower grade, though I would not be reluctant to award a lower grade given a consistent lack of effort, as reflected by a final point total much lower than that of other students.

I am also aware of the ability to give undergraduate students grades of A-, A+, B-, B+, etc. and will consider awarding A+ for truly amazing performance by an undergraduate and B+ for undergraduate performance slightly below A qualification.

It is possible that I will post extra credit assignments to provide practice on topics that I couldn't fit into the run of the course. Students who complete these (if offered) will of course have a better chance of receiving an H or an A.

The spectrum of topics that one could cover in a course with the title *Working with Data in a Public Health Research Setting* is much, much broader than the list of topics currently on the course schedule. I myself am actively involved in learning R, Python, HTML/CSS, and JavaScript and using JavaScript-based libraries such as D3. I welcome discussions with students who are also involved in this wider sort of learning, whether through Coursera courses, Udacity courses, or independent study. Have you tried to do some web scraping with Python? Do you want to retrieve data from Twitter using an API? Are you interested in web-based data visualization? Let's talk about it! This sort of engagement can also affect your course grade.