The University of North Carolina at Chapel Hill

Spring Semester 2018

BIOS 669-001

Working with Data in a Public Health Research Setting

I. Time and place — TuTh 9:30-10:45 PM, McGavran-Greenberg 2308

II. Instructor

Kathy Roggenkamp
Research Instructor and Manager of Statistical Computing
Dept. of Biostatistics and Collaborative Studies Coordinating Center
Office: at Carolina Square: 123 W. Franklin St., Bldg. C, Suite 450
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 elective 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, using metadata, and simulation with SAS. The course also offers an introduction to regular expressions, experience with web scraping, and exposure to cross-language tools such as Jupyter notebooks. 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) unless we decide on a different schedule. 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 4, 2018. We will meet as a class on the morning of May 4, and students who would like to make a 5 to 10 minute presentation based on their project will be welcome to do so – this can only 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]
- Data cleaning [1 exercise]
- Analysis data sets and variables, including combining data, deriving and checking variables, producing needed data structures, and using an external macro to look at excluded observations [2 exercises and one small group activity over 4 class periods]
- An introduction to regular expressions [1 exercise]
- Calling R from SAS [1 exercise]
- An introduction to web scraping [1 exercise]
- PROC REPORT and general reporting concepts (including use of an external macro) [4 exercises]
- Metadata, including codebook production [3 exercises over 4 class periods]

- Simulation in SAS, using both base SAS and IML [2 exercises]
- 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 2018

My primary goal for BIOS 669 is that you learn a lot about useful techniques for handling data, mostly in SAS, beyond techniques covered in BIOS 511. I hope that learning a lot is your primary goal as well. The "computer lab" nature of the course is intended to facilitate learning by doing and to allow me to spend most of my teaching time on being your consultant and providing feedback on your work rather than on preparing and giving lectures. I also think that learning is facilitated in low risk situations – "let me try this and see if it works" - so maintaining a feeling of low risk is another one of my goals.

Given my desire for a learning environment rather than a judgmental "get the right answer" environment, I would love for BIOS 669 to be a pass/fail course. Alas, that is not a possibility. While my focus in reviewing your work will be on providing feedback, I will also have to rate it to a certain degree so that I can offer you a fair final grade (that is, it would be frowned upon if I gave every student the same final grade).

I estimate that about 1/3 of class members will earn an H (or A) as their final grade. I take seriously that an H (or A) is given for truly outstanding performance that often exceeds my expectations. Normal though good performance will receive the modal grade of P. Here is the approach you should take to maximize your chances of getting an H (or A).

- Exhibit thoughtfulness and engagement in your approach to the class and the assignments. If you do not seem to be taking the course very seriously, I doubt that you will earn an H (or an A or A-).
- Consistently do any optional/extra credit problems
- Do an outstanding final project (reasonably ambitious, well-executed), including making a presentation to the class during our final exam session
- Often going along with the above: Consistently writing elegant, well-structured code with lots of white space (indention, blank lines) to put your code in its best light – such code will be easily understood by a person as well as correctly interpreted by a computer

These don't guarantee an H (or A), but you are not likely to get an H (or A) if you don't do at least the first three of these.

In concrete terms, I will base your final grade on a cumulative sum of points assigned to your work on all course assignments, where the final project counts like five regular assignments. Each individual assignment will be given a check plus (3 points), a check (2 points), or a check minus (1 point), and these points are what I will add up in the end.

- Check plus – you did an exceptional job – clear excellence, you exceeded my expectations
- Check – you did a good/reasonable job – entirely satisfactory, meets my expectations

- Check minus – rarely given but reflects disappointing effort or failure to understand the assignment when you had many chances for understanding – will also be given for handing in an assignment late
- Not turning in an assignment will of course result in 0 points for that assignment

Over the course of the semester, you are likely to earn a mixture of checks and check plusses, and hopefully no check minuses or 0's. Since your final project is the last of your work that I will see, it is likely to have a big effect on your overall course grade.

Possibly you are interested in some reasons why I might assign a check plus for an assignment. The following could be factors in addition to the "elegant code" and "do optional questions" items mentioned above:

- Producing an especially beautiful table or graph
- Doing a task in multiple ways when that wasn't requested
- Using software in particularly creative ways or using options that weren't covered in the notes – you might teach me something!
- Making suggestions for how to improve an 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)
- Making me say "wow" as I'm going through my stack of student work

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 √-.

When points are added up at the end of the course, 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.

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 used Anaconda Navigator to run Python through Spyder or to have access to Jupyter notebooks? 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.