Practical Data Management and Statistical Computing with SAS (Biostats 691F)

Class Sessions Fall 2015: Tuesday/Thursday 11:30 AM-12:45 PM

Professor: Penny Pekow, PhD
ppekow@schoolph.umass.edu

Classroom: Arnold 120

Webpage: http://people.umass.edu/bepi691f

Course Description

Research in the biomedical and social sciences commonly involves extensive data acquisition. Data may be acquired from a variety of sources: collected from a single survey, or multiple sources of data and types of measures on subjects including repeated measurements over time; and may include acquisition of previously collected data already digitized. Adequate monitoring of data collection and management has tremendous impact on the quality of information collected and computerized. When using data acquired in digital format, understanding the structure and coding of the data is imperative. Although research projects vary hugely by overall scope, research questions and study design, data from such projects require good management to produce meaningful descriptive and analytic results.

This course presents basic data management principles and practice, with the objective of developing a working knowledge of micro-computer software to use for managing, documenting and describing data.

This course is micro-computer based, making use of Windows-compatible micro-computers (PC’s). MAC users must operate windows emulator software in order to run some of the software taught, in particular, SAS. Software used in the course includes:

- SAS
- MS ACCESS
- MS Word
- MS Excel
- Adobe Acrobat

and others.
The course is recommended for first year masters and doctoral students who will participate in or conduct research involving collection and management of data with the ultimate goal of summarization and analysis.

- Although some statistical ideas will be discussed in the course, the focus of the course is not on statistical analysis.
- Similarly, the course will not discuss study design or development of questions for questionnaires (e.g., we don't discuss validity, reliability).

The course is “hands on”. There is the expectation that the student will not only learn what should be done, but also be able to do it in an efficient manner.

Topics covered include:

- **Design of data collection forms** (paper or on-screen) in a format that
  - enhances completeness and accuracy of information collected
  - facilitates efficient, accurate data entry
  - facilitates subsequent analysis

- **Getting data into the computer**
  - database design
    - *with statistical summarization as goal of data storage*
  - data entry

- **Documentation and monitoring of**
  - Data collection
  - Data entry

- **Documentation of data elements**

- **Processing and cleaning data (using SAS)**
  - Checking for completeness and accuracy
  - Combining files and manipulating data
  - Creating new variables

- **Documentation of data processing/cleaning**
  - Tracking changes in data values
  - Reporting on data processing

- **Sampling from a database**

- **Developing a Descriptive Summary Report**

- **Documentation of data analyses and reports**

- **Data security / Privacy Protection**
  - Data storage
  - De-identification of data
Prerequisites:

- Previous or concurrent enrollment in Biostats 540 – *Introduction to Biostatistics*, or the equivalent
- Basic working familiarity with word processing

Overview

Scope of the Course

Data management is the design and execution of an organized plan for the collection of data, its computer entry, editing, summarization and analysis, *along with clear documentation of the process*. This course has 2 main topic areas:

1. data management
   and
2. statistical computing.

1. Managing data for research studies

By *research studies* I mean studies that are unique and of limited duration such as:

- Experimental studies involving some limited number of subjects
- Surveys
- Multi-faceted longitudinal studies
- Observational or case/control studies with possible follow-up

We do not discuss management of account files, inventories, banking, purchasing, tele-marketing, search and retrieval applications, multiple-user databases, reference or medical record databases.

*Our focus is on development and maintenance of databases designed for specific projects with focused analytic goals* as opposed to long term storage and retrieval of information on individual cases.

2. Statistical Computing includes

1. Computation and manipulation of data as originally entered in the computer to produce data in a format for interpretable summarization, analysis and presentation.

2. Use of software tools for computation and tabular and graphic display of data, presentation of summary statistics, and preparation of data to perform complex analyses.
Use of Microcomputers for Managing Data

Modern microcomputers have changed the ways data are now collected, entered, stored, analyzed, shared and reported. New software and hardware tools continue appear at a rapid pace – this field is constantly changing. A multitude of both hardware and software options are available to choose among for various research projects.

Hardware and software selection are often made based upon such factors as storage capacity, speed, security features, analytic power, graphical capabilities, size of files, and financial resources available. While this is not a course on micro-computers, it is important that you are familiar with basic terminology so that you can understand the hardware requirements of the software you plan to use.

Software applications for data management and analysis differ chiefly in the ease of data entry, mechanisms for editing and accessing data, data security, and in their statistical and report-writing capabilities. Some programs are more quickly learned than others (with menu-driven point and click options) while others may require some programming (writing statements in ‘code’) and take more time and effort. Most have both options available to the sophisticated user. Additionally, some software programs can perform a wide range of functions, while others are designed with more limited specialized capabilities necessitating data transfer between software packages.

We will be learning a variety of software packages for different stages of data management and documentation. Learning to work with data on a variety of platforms will make it easier to learn new techniques as they come along. Systems, hardware and software are continually updated; the ability to learn and adapt to new computing environments is essential.

This course is IBM-compatible PC focused. I recommend that students using MACs install software that allows operation of Windows (e.g., Parallels, VMware or VirtualBox) and work in a PC environment for the course – or else work in campus computer labs. Some of the software we use has MAC-compatible versions, others, such as SAS and MS Access are PC-compatible only, but will run on a MAC operating as a PC with a Windows emulator.
Course Expectations

1. Class attendance / participation.
2. Group project participation.
3. Completion of weekly assignments.
4. Completion of a take-home final exam.

Grading Policy

Grading in the course will be based on classroom and group project participation, assignments, and a final exam. The weighting of these components is as follows:

Weighting of Grading Components

15% Class participation
45% Assignments
20% Group Project
20% Final Exam (Take Home)

Homework Policy:

Assignments are due in class on the assigned date. Usually assignments will be posted on the course website on a Thursday, and due the following Thursday. As a general rule, late assignments will receive a score of 0, and will not be accepted. *I expect printed copies of assignments to be turned in at the end of class. I will not accept email submission of an assignment without prior special arrangement and a really good reason.*

Discussion of assignments with classmates is encouraged, but unless specifically noted, completed assignments should involve largely independent work. If assignments are suspected to have been copied, a score of 0 will be given to all apparently "copied" and "copied from" assignments. **IT IS OK TO GET HELP** from classmates, other students in the computer lab, myself or the TA when you are stuck on some software problem. When in doubt about using other's work, ask for permission, and cite the work copied. **I encourage you to learn from each other.** However, you won't learn this material without hands-on experience – **you need to do your own work.**

Computer Manual Access

Online reference materials are available for most software packages, and links will be available on the course website. You may eventually choose to purchase one or more software manuals or reference books, but that is not required.
Course Text: There is no required text, but there are several books that you may wish to consider purchasing. You may choose to wait, and purchase these later on in the semester if you find you refer to them frequently.

Course Notes: The notes come in 2 parts and are required reading for the course. They are available as course packets for purchase.

Part I: Introduction to Data Management Using Micro-computers

1. Introduction to the Computer: Hardware and Software
2. Data Management Systems
3. Introduction to Case Studies
4. Data Collection Instruments
5. Documentation and Monitoring of Data Processing
6. Database Development
7. Quality Control and Documentation of your Data Management System
8. Data Security: Privacy and Protection of Data

Part II: Introduction to SAS for Data Management

1. Introduction to SAS – The Basics
2. Introduction to the DATA step
3. Using the DATA step:
   Subsetting, Concatenating, Merging Files;
   Creating new variables using conditional statements, arrays, do-end loops
4. More special features:
   PUT statements, using ODS, SAS Functions
5. Procedures for Data Descriptions:
   print, means, summary, univariate, tabulate, freq, forms, report, chart, plot
6. Procedures for Managing Data:
   sort, append, transpose, datasets, import, export, surveyselect
7. Introduction to Macros in SAS
**Recommended Reading** (find 1 or 2 that you like) – Note that some of these are now available in ebook editions also. Check online for details.


*The Little SAS Book: A Primer*, Delwiche, L. and Slaughter, S., SAS Press


*SAS 9.x Language Reference: Concepts, Volumes 1 and 2*

*Base SAS 9.x Procedures Guide, Volumes 1-2*

*SAS 9.x Language Reference: Dictionary, Volumes 1-4*

For the latest on what is available check the SAS publication page at: [http://support.sas.com/publishing/index.html](http://support.sas.com/publishing/index.html)


**Check out SAS online documentation and publishing for additional resources at:**

[http://support.sas.com/publishing/index.html](http://support.sas.com/publishing/index.html)
Computing Resources

Computers:

All students will be expected to use microcomputers – an important aspect of this course. We will make extensive use of SAS and other software. Computer access and all software used in the course are available on the Department of Public Health Computer Lab computers (Arnold 413). In addition, most of the software used is available on other University Computer Labs across campus. Information on availability of computers in classrooms and the Learning commons is available at: https://etna.oit.umass.edu/public/calendar

Additionally, software may be purchased or licensed for home use. Some software packages have a short-term (~1 month) free trial download that may be adequate for course use.

Computer Use/Access – Email:

All students must obtain an account from OIT and establish a UMass email address for this course. Information is available at: http://www.it.umass.edu/accounts/activate-your-account

The OIT account allows campus computer lab use as well as use of computers in the Department of Public Health (DPH) Micro Computer Resources rooms (Arnold 413). With an OIT account you will receive

- a UMass email address
- a Udrive address where you may store files for easy access from any lab on campus as well as off campus through the web

The SPHHS Computer lab is restricted for use by students taking graduate level courses in the Department of Public Health. Hours for the computer lab in Arnold House are normal building hours with the room lock combination provided in the course. Evening and weekend access to microcomputers in Arnold House is possible with the purchase of an Arnold Building key.

Obtaining an Arnold Building Key

Arnold House is open from 8-5, Monday through Friday. To access the building on evenings or weekends, a key is needed. To obtain a key, see Deb Osowski in 415 Arnold. A key deposit is required.
Software Purchase

This course will make extensive use of SAS (Statistical Analysis System) software. It is available in all OIT computer labs and in the Arnold House computer labs also.

If you desire a copy of SAS for your own computer, software is distributed via download, and is not purchased, but is licensed for 1-year periods. The student cost is $125 (initial year), and $45 (renewal years). SAS Software is available in the Lederle Lowrise Research Building, Room A119 (8:30AM-4:45PM). Again, you are not purchasing a copy of the software to own, but rather a license to use SAS software for a specified period, with annual renewal available for a fee.

System Requirements for SAS 9.4

Hardware and Software Requirements for SAS 9.4 can be found at http://www.umass.edu/statdata/software/sas/system_req.htm

SAS does not operate on MAC systems. To run SAS on a MAC you must use a Windows emulator such as Parallels or VMware. VirtualBox is free software to run Windows on a MAC, but then requires purchase and installation of Windows.

Other software used in the course may be available through OIT, from the University Stores, or may be obtained online. For information on licensing or purchase options on software check the OIT software page at: http://www.oit.umass.edu/software/supported_software

We will also be using Microsoft Access, Excel and Word. While you may choose to use other word processing or spreadsheet software, these are available in computer labs across campus, and are the standard in use in many companies. Course examples will be primarily in these packages.