February 29, 2004

HES 703 / STAT 301 / STAT 501
Statistical Computing and Graphics

Frank E. Harrell, Jr. (f.harrell@vanderbilt.edu)
Professor of Biostatistics
Department of Biostatistics
Vanderbilt University School of Medicine

Class Web Page:
http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/StatCompCourse

Objectives : To be able to use a high-level object-oriented statistical computing language to

1. input data into a computer
2. perform statistical calculations
3. examine data
4. compute and display descriptive statistics
5. understand how commonly used graphical techniques can result in optical illusions and poor communication of information
6. learn how to construct graphics that can accurately display information in multiple variables
7. conduct exploratory data analysis

Secondary objectives are to learn the rudiments of LaTeX, a markup typesetting language for producing technical reports, and to use command languages for reproducible research and reporting.

Statistical Software: S-Plus 6.1 and R 1.5.1
Operating System: Microsoft Windows 95/98/NT/2000
S-Plus 6.1 or R 1.5.1 for Linux/UNIX
R 1.5.1 for MacOS X
Add-on S library: `Hmisc`

Course Overview: Statistical Computing and Graphics is a core course required of all students enrolled in the M.S. in Health Evaluation Sciences program in the Clinical Investigation, Epidemiology, and Health Services Research and Outcomes Evaluation tracks. It is also a required course (at least on a pass/fail basis) for all incoming Statistics graduate students beginning Fall, 2000. The course is also a course for undergraduates interested in statistics and other quantitative disciplines.

The course is based on the premises that (1) graphical presentation of information is fundamental to understanding data and presenting research findings and (2) modern computing tools are required for extracting information and drawing inferences from data and for constructing the best graphical displays. Such computing tools allow researchers to generate highly informative graphics in a flexible fashion with a minimum of programming. These tools also allow researchers having a grasp of statistical concepts to use modern computationally-intensive techniques to draw inferences without making as many strict assumptions (e.g., about data distributions) as in the past.

The course will utilize both lecture and interactive laboratory format. The former will be used to teach students about optical illusions and other problems with many traditional graphical displays, elements of graphical perception, the elements of good graphical display, and fundamentals of the S statistical computing language. During interactive labs students will learn how to use S to process, analyze, and graph data (both raw data as well as summary data). The course also teaches rudiments of LaTeX for compiling and typesetting scientific reports using a LaTeX web server.

Credits: 3 semester hours graduate credit for HES M.S. and STAT 501 students. The first hour of the Monday lab is mandatory for all students receiving 3 hours of credit. STAT 301 students have the option of taking the class for 2 hours of credit by not enrolling in the STAT 301 1-hour lab.

Labs: Labs available for doing homework:
Wilson 308 (Mon., Wed. noon-3p when class or lab not in session)
ACHS (Academic Computing Health Sciences) 8a-5p Mon-Fri (3rd floor Hospital West by DHES)
Health Sciences Library Learning Resources Center (1st floor of HSL) 7:30a-midnight Mon-Thurs, 7:30a-7p Fri, 9a-7p Sat, 9a-midnight Sun.

Obtaining Software: R for Windows, Linux, or MacOS X may be downloaded from `www.r-project.org`.[1]

Prerequisites: Ability to copy, move, and edit files on PCs. Knowledge of a programming language such as Basic, Python, Perl, C, C++, Fortran, or S is also required. ITC and the Health Sciences Library teach a variety of short courses which will provide spreadsheet or Internet skills which are also helpful. You must have had an introductory statistics course or be taking one concurrently.

Outline and Approximate Schedule:

Bold numbers to the right of topics indicate sequential lecture numbers. K&O denotes Krause and Olson. A&H denotes Alzola and Harrell. UG denotes the

[1] R is almost fully compatible with S and is free and open source, so it can be used after you leave UVa if your future employer does not wish to purchase S-Plus. R has only a rudimentary graphical user interface, has less file import/export capabilities, and fewer options for interfacing with Microsoft Office products than S-Plus has.

S-Plus User's Guide and PG denotes the S-Plus Programmer's Guide. The last two are for supplemental reading and are available as online manuals from the `Help` menu. Chapter or section numbers pertain to A&H if the source is not stated.

1. Overview of Programming and Command Languages **(1)**
   (a) Types of languages/system interfaces
      i. Commercial vs. open source
      ii. Commands vs. graphical user interfaces (GUIs)
      iii. Compiled vs. interpreted commands
      iv. Procedural vs. functional languages
   (b) Example command language: `graphviz` for drawing tree diagrams
   (c) Example text processing markup language: LaTeX
   (d) Using the Biostatistics LaTeX Web server to produce `PDF` reports
   (e) Statistical languages and packages **(2)**
   (f) Why S? (PG 1 first 2 pages)
   (g) History and background of S-Plus (K&O pp. 1-6)
2. Starting and Using S-Plus
   (a) Starting S-Plus (A&H 1.2.2, K&O 2.2)
   (b) The command window
   (c) Entering and saving commands using the script window (K&O 2.16, A&H 1.5, UG 8)
   (d) Starting R
3. S-Plus Statistical Computing and Graphics Language (K&O 3, A&H 1.4, UG 2,9, PG 1)
   (a) General rules (K&O 1.3)
   (b) S as a calculator: arithmetic operators
   (c) Assignment operator
   (d) Our first functions: `print` and `rm` (remove)
   (e) Making vectors: `c, rep, seq` functions
   (f) Grouping expressions
   (g) Logical values
   (h) Operating on vectors (also A&H 2.4)
   (i) Missing values (K&O 4.3)
4. Objects, Getting Help, Functions, Attributes, and Libraries (A&H 2, UG 9, PG 2) **(3)**

   (a) Objects (A&H 2.1, K&O 3.2)
   (b) Getting help (A&H 2.2, K&O 3.1.2)
   (c) Functions (K&O 4.2, A&H 2.3)
   (d) Subsetting vectors (A&H 2.4, K&O 3.4-3.6)
   (e) Matrices, lists, data frames (K&O 4.1, A&H 2.5, PG 3)
   (f) Attributes (A&H 2.6)
   (g) When to quote names (A&H 2.7)
   (h) Add-on function library: Hmisc (A&H 2.9)

5. Data in S-Plus (A&H 3) **(4)**
   (a) Importing data (K&O 2.3.1, 2.9, 11.2.2, A&H 3.1, UG 9, PG 4)
   (b) Adjustments to variables after input (A&H 3.2.3, 4.1.5, K&O 11.6)
   (c) Writing data, customized printing
   (d) Inspecting data (descriptive statistics, checking quality)

6. Operating in S-Plus (A&H 4)
   (a) Reading and writing data frames and variables: The search list and `attach` function (A&H 4.1.1, K&O 4.1.3)
   (b) Subsetting and sorting data frames (A&H 4.1.2)
   (c) Using the Hmisc `upData` function for changing data frames (A&H 4.1.5)
   (d) Functions for manipulating and summarizing data; by processing (A&H 4.2.1-2, K&O 7.1, PG 3)
   (e) Reference material for advanced data manipulation examples (A&H 4.2.5-8) **(5)**
   (f) Recoding variables and creating derived variables (A&H 4.3)
   (g) Simple missing value imputation (A&H 4.5)

7. Review of Data Frame Creation, Annotation, and Analysis (A&H 4.4)
   (a) Importing external data
   (b) Making global changes to a data frame
   (c) Changing variables within a data frame
   (d) Analyses of the entire data frame
   (e) Analysis of individual variables

8. Probability and Statistical Functions (A&H 5, UG 9) **(6)**
   (a) Statistical summaries (K&O 7.1)
   (b) Probability distributions (K&O 7.3)

   (c) Statistical tests (K&O 7.4)

9. Advance reading: Chapter 4 of Cleveland (except p. 255). After that read Cleveland chapters 1 and 2 and sections 3.1, 3.4-3.6

10. Making tables (A&H 6)

11. Constructing Graphics (A&H 10) **(7)**

    (a) Goals in communicating information (Cleveland chapter 4)

    (b) Pattern perception and estimation

    (c) Survey of findings about human perceptions of various aspects of graphics (angles, lengths, relative lengths, color, slope, Weber's law etc.)

    (d) Optical illusions and other problems with pop charts

    (e) Graphics disasters and some of Cleveland's remedies (Cleveland chapter 1 and 2 and sections 3.1, 3.4-3.6)

    (f) Summary of elements of good graphic construction: Constructing graphics to improve human perception & accuracy of information transfer

    (g) Graphical disasters

    (h) Graphical marvels

12. Presentation of Students' Examples of Good and Bad Graphics **(8)**

13. Graphics for One or Two Variables (A&H 11.3, UG 3-4) **(9)**

    (a) One-dimensional scatterplots (Cleveland p. 133)

    (b) Histograms and density plots (Cleveland 3.3 pp. 133-136)

    (c) Cumulative distribution plots (Hmisc `ecdf` function; see Rosner)

    (d) Box plots (Cleveland pp. 139-142)

    (e) Scatter diagrams with or without transformations of variables (Cleveland pp. 100-101, 3.1, 3.5, 4.8)

    (f) Reference lines

14. Conditioning and Graphics for Three or More Variables (UG 4) **(10)**

    (a) Dot plots (Cleveland 3.4, 4.1, 4.6, 4.9, pp. 267, 269)

    (b) Conditioning plots and the principle of small multiples (Cleveland pp. 114, 152-153, 167, 249-250, 267, Section 3.10, 3.11, 4.9)

    (c) 3-D plots

        i. Scatterplot matrices (Cleveland 3.9, K&O 7.7)

        ii. Varying symbol characteristics to represent more than two variables in a scattergram (Cleveland Figure I, 3.5, 3.10, 3.11, 3.13, 4.4)

    iii. Bubble plots

    iv. Thermometer plots for geographic displays (Cleveland pp. 240-241, Cleveland & McGill 5.3, 2^{nd} example from online help file for `symbols()`)

    v. 3-D plots for almost smooth surfaces

        A. 3-D scatterplot

        B. Perspective plots (K&O 7.2)

        C. Contour plots (K&O p. 7.2)

        D. Image plots (Cleveland Fig. II, pp. 211-212, 4.3, `image` function (K&O 7.2), two examples on web page)

    vi. Interactive and dynamic graphics

        A. Identifying points (A&H 11.2)

        B. 3-D perspective plots with rotation

        C. Brush and spin (Cleveland 3.12)

        D. "Live" graphics on web pages using Java graphlets from S-Plus 6.0 on Linux/UNIX (see web page)

(d) Trellis graphics (A&H 11.4, K&O 6) **(12)**

    i. `bwplot, densityplot, dotplot, histogram, splom, density, stripplot, xyplot` functions

    ii. Hmisc `xYplot` function for error bars and bands (A&H 11.4.1)

    iii. Hmisc `panel.bpplot` Trellis panel function for extended box plots (A&H Figure 11.11)

    iv. Plotting summary statistics using Hmisc `summarize` and `xYplot` functions and other Trellis graphics functions (A&H 11.4.3-4)

    v. Plotting error bars and bands using `xYplot` and `Dotplot` (A&H 11.4.1-2) **(13)**

    vi. Summary of Functions for Aggregating Data for Plotting (A&H 11.4.4)

15. More Statistical Calculations (A&H 5.3) (lab)

    (a) Basic power/sample size calculations

    (b) Computing confidence limits using commands & menus (CLs based on t and binomial distributions (`binconf` function))

    (c) Showing sampling distributions of sample means (central limit theorem)

16. Nonparametric Trend Lines (Cleveland pp. 18-19, 168-179, A&H 11.3)

    (a) Nonparametric regression fits (trends without linearity assumptions) using Hmisc `plsmo` function

    (b) Example of basic plotting commands and nonparametric smoother using 1996 Olympics medal counts

17. Review and Miscellaneous Functions (A&H 11-12, K&O 5, UG 9, PG 8) **(14)**
    (a) Graphics devices (A&H 12.2, K&O 5.2)
    (b) Putting S-Plus graphs into Microsoft Office documents (A&H 12.2.4)
    (c) Graphical snapshot of a dataset (review)
    (d) Graphical summary of missing data patterns (review)
    (e) Graphical summary of interrelationships of variables (review of Hmisc `varclus` function))
    (f) Basic scatterplots; scatterplot matrices
    (g) Turning tables into plots (review of `summary.formula`)
18. Case Study: Boston Housing Price Data

Texts:

Krause A, Olson M: *The Basics of S and S-Plus, Second Edition*, New York: Springer, 2000.
Harrell FE, Alzola CF: *An Introduction to S-Plus and to the Hmisc and Design Libraries*, 2002.
Cleveland W: *The Elements of Graphing Data*, Summit NJ: Hobart Press, 1994.
S-Plus 6 User's Guide, Seattle: Insightful Corp (available as an online manual).
S-Plus 6 Programmer's Guide, Seattle: Insightful Corp (available as an online manual).
Oetiker T: *The Not So Short Introduction to LaTeX*, available from `http://ctan.tug.org/tex-archive/info/lshort/english/lshort.pdf`.[2]

Especially Useful References:

Tufte ER: *The Visual Display of Quantitative Information*, Cheshire CT: Graphics Press, 1983.
Tufte ER: *Envisioning Information*, Cheshire CT: Graphics Press, 1990.
Tufte ER: *Visual Explanations*, Cheshire CT: Graphics Press, 1997.
Cleveland WS: *Visualizing Data*, Summit NJ: Hobart Press, 1993.
S-Plus Online Manuals. (Also see P. 4 of Alzola & Harrell)
Wilkinson L: *The Grammar of Graphics*, New York: Springer, 1999.
Wallgren A, Wallgren B, Persson R, Jorner U, Haaland, J: *Graphing Statistics & Data*, Thousand Oaks: SAGE Publications, 1996.

Datasets:

1. In S-Plus libraries

[2] There is a link to the LaTeX server from the class web page. This has a link to the Oetiker document as well as containing summaries of LaTeX usage pertinent to this course.

2. Other datasets from Rosner's *Fundamentals of Biostatistics, 5th edition.* ASCII files and data dictionaries are available at `http://biostat.mc.vanderbilt.edu/twiki/pub/Main/StatCompCourse/RosnerData.zip`.
3. Other datasets from web page

See `Rosner` and `Datasets` area on the web page. Several of the Rosner datasets have already been converted to S-PLUS data frames. These are under `dumpdata.sdd` under the `Rosner` area. You can use these for assignments in which you are not asked to create the data frames.

Class Announcements:

Class Discussion Group: This is an excellent way to post questions and answers because all postings and replies are "threaded" (categorized), so you can return to the discussion group weeks later and still benefit from seeing answers regarding a specific topic. The discussion group is an excellent way to keep in touch with the class and even more to ask and answer questions. I hope that all students will use it to

- ask or answer any question whatsoever related to group assignments
- ask or answer any logistical or purely technical questions related to individual work assignments
- ask or answer any questions about statistical computing and graphics concepts that are not directly related to a pending individual work assignment

Be **sure** to check existing topics for posting your message, to avoid creating any unnecessary new topics that will make it more difficult for others to navigate the discussion board.

Assignments: Computer problems and graphical critique/design. Assignments are usually due 7 days after they are assigned. Last day to turn in final project: 13 December. One letter grade is deducted per day late unless prior arrangements are made with the instructor or teaching assistant (these prior arrangements are allowed once or twice per semester for students with excellent attendance records otherwise). The vast majority of assignments may be done as group projects, with all group participants signing a single copy verifying their participation. The signature must be accompanied by the following statement.

> The undersigned students each participated meaningfully and significantly in this group assignment.

For select assignments and for the final project you must do all the work on your own, and sign the honor pledge. For these, and for the quizzes, signing the honor pledge means that you have not read solutions or answers from previous courses. These assignments will be designated as "independent work" or "projects." Projects count 2 × group homeworks. Final project counts 4× group homeworks.

Work must be as concise as clear communication will allow. You must use the LaTeX server to format all work you turn in except for quiz answers and the "graph finding" exercise. Print the resulting PDF document (or occasionally E-mail it to the teaching assistant if you have checked with him first).

If an assignment asks you to use menus to obtain output, do the following. When the output is a graph or a listing of results, turn in the graph or listing, preceeded by the names of the menus you used to obtain the results, if this was not already specified in the assignment. When the assignment asks you to obtain a result using a command, show the commands in what you turn in. When a simple command is used to obtain a standard graph (e.g., `ecdf(age)` or `hist.data.frame(mydataframe)`), you need not turn in the graphical output. You do not need to turn in anything when the problem asks you to merely manipulate the system (e.g., to navigate the help system). When in doubt about what should be turned in for a given assignment, ask the teaching assistant or instructor in advance.

You may obtain general assistance from the teaching assistant on the assignments. You may also obtain general assistance from the teaching assistant, instructor, and other students through the above E-mail group.

Exams: There will be quizzes almost weekly, counting equal to a group assignment. There will be no makeups for quizzes unless prior arrangements have been approved by the instructor. There may be a final and/or a midterm exam. A quiz lasts about 15 minutes and tests concepts and a few details about most important S language commands. The concepts from which quiz questions are drawn and which you are responsible to master are listed at `http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/StatCompConcepts` on the class Web page. It is exceptionally important to understand concepts taught in the course, as students not understanding these concepts are frequently unable to use the computer to analyze data in their other courses or theses, or in their final project for this course. The single lowest quiz grade for the semester will be ignored for each student.

Be sure that if you do not understand the concepts you take advantage of the Teaching Assistant's and Instructor's office hours, ask questions in class and lab, and that you make use of the discussion board. This discussion group is ideal for asking questions about concepts as well as about details.

Class participation: necessary

What to bring: Floppy or Zip disk to each class unless you are facile with the ITC `Home Directory` service.

Assistance: Questions for teaching assistant or instructor outside class - E-mail or phone for appointment but try to come to the lab help session or during open office hours. First try the E-mail group, so that other students can see your question and the answers you get from others.

Anonymous Feedback to Instructor: See web page.

Name (Last, First)	E-mail Address	Dept.	Regis– tered?	Computer Languages Known	Operating System

If your E-mail ends in `@virginia.edu` omit that part.
For Operating System enter e.g. W98,W2000,XP,Linux,Unix,Mac and if Mac specify which Mac OS (e.g. Mac OSX).