Department of Statistics
University of Wisconsin, Madison
Fall Semester, 2010

Stat 471: Statistical Computing

Class hours:	11:00-11:50am MWF
Classroom:	133 SMI
Instructor:	Douglas Bates
Office:	1245C MSC
Phone:	(608) 263 8531
Email:	`bates@stat.wisc.edu`.
Course website:	`http://www.stat.wisc.edu/~st471-1`
Office hours:	M: 4-4:40pm; R: 12:00-1:00pm.
Course email list:	stat471-1-f10@lists.wisc.edu

Description: The aim of this course is to enable the student to use the R language and environment for effective graphical presentation of data, for creating and presenting simulation results, for fitting statistical models to data and for making inferences through maximum likelihood or by Bayesian methods using Markov Chain Monte Carlo (MCMC). The homework assignments and suggested exercises will provide the opportunity to implement many of these methods using real or simulated data.

Prerequisites: Linear algebra, introductory statistics & probability.

References: All of the books listed below are on reserve at the Wendt library. These are suggested references, not required texts.

- *R Programming for Bioinformatics*, Gentleman (CRC Press, 2008).
- *Software for Data Analysis: Programming in R*, Chambers (Springer, 2008).
- *ggplot2: Elegant Graphics for Data Analysis* Wickham (Springer, 2009).
- *Lattice: Multivariate Data Visualization with R* Sarkar (Springer, 2008).

The last two references describe advanced graphics systems for R. We will use *ggplot2* more than *lattice*.

Grading. Grading will be composed of following components:

- **Homework assignments: 25%.** Bi-weekly homework assignments will include a lab component, where you will typically be required to access or simulate data and perform some task such as plotting the data or fitting a model to the data. Homework assignments will be submitted to the Learn@UW dropbox.
- **Midterm exam I: 25%.** An in-class midterm exam is scheduled on October 11, 2010.
- **Midterm exam II: 25%.** An in-class exam is tentatively scheduled on November 15, 2009.
- **Final project: 25%.** The final project will include designing, coding, testing and documenting a piece of statistical software, which typically will be in the form of an R package. Other ways of organizing and documenting your software can be used but please check with me beforehand. In particular, I need to be able to run the software to be able to evaluate it so check with me before using proprietary software. For

example, if you want to use *Matlab* I would suggest *octave* instead as I don't have convenient access to *Matlab*.

In order to avoid leaving things to the last minute, everyone should submit a project proposal by October 27th, 2010. The proposal should include a brief description of the data you plan to analyze and your tentative plan of analysis.

Final software packages are to be submitted on or before December 20, 2010.

You may submit a group project for groups of 2 or 3 students. Naturally a group project is expected to be more comprehensive than an individual project. All students in the group will receive the same grade. Please inform me if you plan a group project

You are not allowed to collaborate on the exams. You can share ideas on homework problem solutions but all the work presented on solutions must absolutely be yours. For example, it is acceptable to ask for assistance and clarification on the Learn@UW discussion forums but it is not acceptable to ask for someone else to provide you with their code.

If you discuss your solutions with a classmate (or someone else), you should mention that in your write-up.

You are allowed one letter-sized page of notes (you can use both sides if you wish) in each of the midterm exams. Questions on the midterms will be directed towards deciding what code snippets should be expected to provide, writing some code snippets for a particular task, and describing in more general terms how to produce a certain type of graph, simulation or analysis. More mathematical questions regarding numerical linear algebra may occur on the second midterm.

Tentative syllabus

Week	**Date**	**Topic**
1	Sep 3	Course orientation, installing and using R
2	Sep 8, 10	data organization, ggplot2 for plotting densities
3	Sep 13	plotting pmf's, cdf's and quantile functions
	Sep 15	simulating samples from densities, sample statistics
	Sep 17	comparing speeds of simulation methods, lab
4	Sep 20	plotting results, Q-Q plots, empirical densities
	Sep 22	fitting linear models, formula/data specification, extractors
	Sep 24	lab session
5	Sep 27	linear models (cont'd), simulation of lm fits
	Sep 29	analysis of variance models
	Oct 1	lab session
6	Oct 4	numerical linear algebra
	Oct 6	QR, Cholesky and singular value decompositions
	Oct 8	lab session and review for midterm 1
7	Oct 11	**midterm 1**
	Oct 13	discussion of and solutions to midterm 1
	Oct 15	lab session
8	Oct 18	generalized linear models, specification
	Oct 20	GLMs, evaluation and interpretation, IRLS algorithm
	Oct 22	lab session
9-16		to be determined