Statistics 503 - Exploratory Methods and Data Mining  
Syllabus Spring 2004

Course Description: Approaches to finding the unexpected in data: data mining, pattern recognition and understanding. Emphasis is on data-centered, non-inferential statistics, for large or high-dimensional data, and topical problems. Simple graphical methods, as well as classical and computer-intensive methods applied in an exploratory manner.

Meeting times: MWF 2:10-3:00pm  
Location: Physics 0058  
Instructor: Dr Dianne Cook, 325 Snedecor Hall, 515 294 8865, dicook@iastate.edu  
Office hours: Tu 10:00-10:50am, W 10:30-11:30am or by appointment  
Class web site: http://www.public.iastate.edu/~dicook/stat503.html  
Class notes: Provided by the instructor online, and photocopied articles as appropriate  

Course outline: This will be a data-centered course, with segments of the semester focusing on particular data sets, each getting more complex as the semester progresses. The first week will focus on an introduction to data mining, and software.

Course Objectives: This course is designed to provide students with the experience of working with statistical methods for analyzing new, complex data, and arriving at reliable and detailed summaries of the data. Students will learn to identify a scientific problem underlying selected data examples and address this problem in the analysis and conclusion. The internet will be used to scavenge for software to conduct analyses and literature related to the scientific problem.

Computing: Software used will be R (www.R-project.org) and GGobi (www.ggobi.org). Both are free software packages. You should download and install a copy of each to a computer that you regularly use. The PC versions are easy to install. R is a command line, scripting data analysis language, so there may be a steep learning curve to using it, but the gain in flexibility for data analysis is worth it.
Course Grades: Each person will be a part of a work group of up to 4 people. Then for each module there will be an additional, similar data set to explore and report on. This will be done as a group. There will be approximately 4 of these assignments throughout the semester. Each will be worth 10pts.

There will be 4 short homeworks in between assignments which will test specific skills using the computer based on textbook exercises. These will be done individually. Each will be worth 5pts.

There will be two in-class exams, to be done individually. These will test (1) ability to explain different methods, including what types of things you can learn from data using the method, (2) finding faults with using particular methods, (3) recognizing mis-diagnoses or unjustifiable conclusions. Each will be worth 10pts.

There will be one project, which will involve pulling some data off the internet, or using your own data, determining appropriate questions, analysis of the data and reporting results. This will be completed with your working group. It will be worth 15pts.

This class works better if students actively participate in class, asking questions, volunteering answers to questions and making suggestions for the data analyses. There is 5 pts allocated to encourage this type of activity.

The total score will be out of 100 computed from assignments (40pts), homeworks (20pts), exams (20pts), project (15pts), class participation (5pts).