

# Introductory Data Analysis

**John W. Emerson**  
**Yale University**  
**Spring 2004**

## Objectives and Prerequisites

This course provides an introduction to probability and statistics with strong emphasis on data analysis. Topics include numerical and graphical summaries of data, probability, hypothesis testing, confidence intervals, counts and tables, analysis of variance, and regression. Upon completion of this course, students should be able to think critically about data, to use graphical and numerical summaries, to apply standard statistical inference procedures and to draw conclusions from such analyses. This course will be computationally, not mathematically, intensive. There are no prerequisites other than a willingness to get your hands dirty using the computer.

## Practical Information

- Although my name is John, I would prefer that you call me Jay, Mr. Emerson, Dr. Emerson, or Professor Emerson. No extra credit points are based on this choice.
- E-mail is the best way to reach me: [john.emerson@yale.edu](mailto:john.emerson@yale.edu).
- Office: I'm in 24 Hillhouse Room 208, phone 432-0638. Office Hours Monday 10-noon and by appointment.
- Lectures: Monday and Wednesday, 1-2:15 and 2:30-3:45 in the StatLab. Please do not switch sections once the course begins; seating in the StatLab is limited.
- Review Sessions: as announced, sometimes in the StatLab, sometimes in the classroom of the Department of Statistics, 24 Hillhouse Avenue (the pink house with green shutters across from the Health Center).
- TFs: Daniela Cojocaru and Susan Perrone.
- Weekly homework assignments must be turned in to me, **at the beginning of class**, on the assigned dates.
- StatLab Information and Hours: in particular, note that students may obtain a free copy of S-Plus. A free software package called R may be [downloaded](#) from the web, runs on all platforms (Mac/PC/Linux), and is very similar (though not identical) to S-Plus. Software information and hours of operation are posted on the web site: <http://statlab.stat.yale.edu>. Consultants are regularly available for help using the software packages. However, the consultants are not TFs for the course; please be considerate.

## **Textbooks**

The required textbook is by Ramsey and Schafer (2002), *The Statistical Sleuth: A Course in Methods of Data Analysis*, 2<sup>nd</sup> edition. It is available at Barnes & Noble (the Yale Bookstore). There appear to be a limited number of copies. You may be able to get the book quickly (and possibly for less \$) from Amazon.Com or Half.Com, but make sure you get the correct edition.

Please note: this is a thick, heavy book at a somewhat higher level than this course. We will not cover the entire book, and emphasis will be placed on intuition, common sense, and practice working with real data. Don't be intimidated by the mathematical sections in the book.

For a much lighter (and fun) reference on introductory statistical concepts (not data analysis), you might consider a book by Gonick and Smith (1993), *The Cartoon Guide to Statistics*.

Useful S-Plus references include *The Basics of S-Plus*, *An Introduction to S and S-Plus*, and *Modern Applied Statistics with S*.

## **Homework**

Homework will be assigned and collected every Wednesday; the first assignment will be due on Wednesday, January 21. To succeed in this course, you must stay on top of the material. I will keep the Classes server (<http://classes.yale.edu>) completely updated with respect to the assigned homework. Problem sets will be collected at the **beginning** of class.

## **Exams**

There are no exams in the traditional sense. However, I reserve the right to administer a "final exercise" during the scheduled exam period. If I decide to try this, I will be flexible and will permit students to complete this exercise either (a) during the scheduled exam period, (b) on an evening during the last week of classes, or (c) at a selected time during reading week. Details will be finalized before Spring Break.

## **Computing**

Computing facilities are available to students at the StatLab. This is a computationally intensive course, and you will be required to experiment and follow examples given in class. Unedited computer output will not be accepted; you must carefully document and explain your work. You will not be required to do original "computer programming," so to speak; my examples in class will serve as a basis for any required analysis.

## **Grades**

The “final exercise” (if given) will be worth the equivalent of two homework assignments.

## **Data**

The web contains a wealth of data. Some sources provide “clean” data sets ready for use. Others contain raw data that could be compiled into new data sets. I will work on maintaining a list of good sources of data, at:

<http://research.yale.edu/statistics/datasets.html>.