

An Intensive Course in Data Analysis with Statistical Case Studies

Professor John W. Emerson
Department of Statistics
Yale University

Organizer:

[Microsoft Statistics & Information Technology Laboratory of Peking University](http://iria.pku.edu.cn/PMSIT/Course07/)
[Beijing International Center for Mathematics Research, Peking University](http://iria.pku.edu.cn/PMSIT/Course07/)

Time: Aug. 6 – Aug. 31, 2007

Course homepage:

<http://iria.pku.edu.cn/PMSIT/Course07/>
<http://www.stat.yale.edu/~jay/China2007/>

Course Description

Statistical analysis of a variety of data problems including (but not limited to) an age discrimination lawsuit, the use of gambling point spreads in predicting outcomes of professional and college sports in the US, and the clustering of countries based on measures of environmental performance.

As we study these problems, you will develop skills useful in acquiring data directly from web pages, and assessing data quality. With the continued growth of Information Technology, more and more information is archived and available for study. However, quantity of data does not imply quality of data, and researchers must possess exceptional exploratory data analysis abilities to successfully confront and responsibly exploit these exciting new research opportunities. It is never wrong to do exploratory data analysis, and the application of statistical theory to real world problems in the absence of exploratory data analysis can have disastrous consequences.

The course will be computationally intensive; the most important prerequisite is a willingness to get your hands dirty working with real data sets. The second most important prerequisite is a willingness to consider statistics and data analysis at an intuitive, occasionally non-mathematical level, connected to real-world problems. The course will include an introduction to R.

In data analysis, I believe you learn as much (and sometimes more) when things "don't work" than when they go as planned. You have succeeded when you can figure out why something doesn't work (or why some analysis isn't appropriate) and deduce an appropriate course of action as a result. You must be willing to try out new things and to

make mistakes -- you can't break the computer. Seek to understand the mistakes, move onward, and you will be successful.

Course Preparation

Students are encouraged to try using R in advance: it is available (free of charge) from <http://www.r-project.org>, along with “An Introduction to R” – a helpful introduction to the language and how to use R for doing statistical analysis and graphics. However, the course will include formal instruction in the use of R for statistical computing.

Students with no prior coursework in probability or statistics are encouraged to prepare for the course using books such as “Introduction to the Practice of Statistics” by Moore and McCabe, or “Stats: Data and Models” by De Veaux, Velleman, and Bock. A more advanced mathematical preparation is also acceptable (but is not necessary). For example, “Mathematical Statistics and Data Analysis” by John Rice, or “An Introduction to Mathematical Statistics and its Applications” by Larsen and Marx contain material far beyond what is necessary course preparation.

If these books are not available to you, please ask your professors to recommend similar texts. The ones listed above are only provided as guidelines, and are popular in the US.

Bio

Professor Emerson was an undergraduate at Williams College, studying Mathematics and Economics. He was awarded the Donovan and Moody scholarships for graduate study at Oxford University, where he received a Master of Philosophy in Economics. As a doctoral student at Yale University, his development of mosaic plots in S-Plus won a Student Paper Competition of the American Statistical Association, and his implementation now appears in the core R distribution. His doctoral dissertation at Yale University was entitled “Asymptotic Admissibility and Bayesian Estimation.” Professor Emerson’s consulting activities range from criminal investigations for the FBI to clinical research at the Pfizer pharmaceutical company. He presented his recent work on the graphical exploration of spatially distributed time series and generalized pairs plots at the UseR! conference in Vienna, the Joint Statistical Meetings in Seattle, WA, and as the keynote speaker of the Connecticut Independent Teachers Association annual meeting. He provided essential supporting analysis for the Wall Street Journal’s revelation of corporate stock options backdating. His most recent publication, “Chance: On and Off the Ice” summarizes a problem with the new figure skating scoring system that he first described on the ABC World News Tonight during the Olympic Winter Games in Torino, Italy. Professor Emerson is the Director of Graduate Studies in the Department of Statistics at Yale, where his deep commitment to teaching, the fundamentals of data analysis, and the integration of teaching, research, and real-world problems are widely recognized.

Professor Emerson’s web site: <http://www.stat.yale.edu/~jay>