

Practical Stats Newsletter for June 2009

Subscribe and Unsubscribe: <http://practicalstats.com/news>

All of our past newsletters: <http://practicalstats.com/news>

In this newsletter:

1. Upcoming Courses
2. Visualization of Multivariate Data
3. New Time Series class for frequently-collected data

1. Upcoming Courses

We are offering three open-enrollment courses in August and September 2009.

Untangling Multivariate Relationships is our 2-day course covering the multivariate methods of primary interest to environmental science, focusing on seeing, testing, and modeling multivariate relationships. Relationships between chemical and biological measures are used as examples. Newer, nonparametric methods for establishing links between variables supplant traditional methods dependent on multivariate normality. UMR will be held Aug 24-25, 2009 at the Hilton Garden Inn Downtown in Austin TX. Register online at our Upcoming Classes page:
http://www.practicalstats.com/new_classes/classes.html.

Nondetects And Data Analysis, the course that illustrates methods for correctly handling data with nondetects, will be held August 26-27 (just following the UMR course) at the Hilton Garden Inn Downtown in Austin, TX. Register for both and save on travel dollars! Online registration is available at
http://www.practicalstats.com/new_classes/classes.html

New topics include how to sum a series of components to get a total when some components are nondetects. Stop substituting one-half the detection limit, and make sense of data with nondetects.

Applied Environmental Statistics, our flagship 4.5 day overview of statistical methods for environmental sciences, will be held in Tacoma, Washington at the University of Washington–Tacoma on September 21-25. Topics include how hypothesis tests work, how to build regression models, load estimation and trend analysis. New topics include how to perform permutation tests and bootstrapping, both used to avoid traditional assumptions of a normal distribution that are questionable with environmental data.

** Also included, a free (optional) half-day introduction to using R statistical software, for enrolled students only.

You can always find a complete course listing at

http://www.practicalstats.com/new_classes/classes.html.

2. Visualization of Multivariate Data

Visualizing multivariate data is a challenge – we live and work in 3-dimensional space. Yet tables of many variables side-by-side provides little insight into what patterns our data contain. Methods such as multidimensional scaling [see the April 07 Practical Stats newsletter] and PCA score plots collapse and present some relationships in two-dimensional, static plots. But faced with possible relationships between 4 or more variables, what patterns escape this process and go unnoticed?

In recent years dynamic graphics have been used to better picture relationships among variables in more than 2 dimensions. Axes are rotated and interchanged in order to see linear and nonlinear dependencies, outliers, and other important data characteristics. Brushing (selecting points with the mouse and coloring them to highlight their position) can track important groups or single values across views and even across separate plots. While some commercial statistics software offers a few of these capabilities, a more comprehensive package is the free GGobi software (www.ggobi.org).

GGobi is an open-source visualization program for exploring high-dimensional data. It runs on Macs, Windows and Linux operating systems. We use it in our UMR class alongside more traditional software to check the results of clustering, PCA and other multivariate methods. Relationships among variables are illustrated as the data cloud rotates in multiple dimensions. Data 'tours' swap variables in and out to determine which axes present the most dramatic information. If you would like a demonstration of how GGobi can be used with your data, download the QuickTime movie "Teaching with ggobi" at <http://www.ggobi.org/docs/>. This movie shows how to see the results of a factor analysis, PCA, or MLE model across more than two dimensions. It demonstrates brushing techniques to observe how positions of data in two dimensions change as values of a third variable are varied from low to high. The video also shows a clustering technique called 'self-organizing maps' for consistent and logical classifications over five dimensions. A second movie at the same webpage, named "Tours", illustrates the use of GGobi's controls for finding the maximum separation between groups, a visual discrimination that is often not viewable in the series of 2-d scatterplots possible between original variables. Dynamic visualization, where data are rotating in 3-dimensional space while axis variables are alternated in order to inspect relationships among all variables, is a powerful tool for finding or verifying relationships among several environmental variables. Download and use GGobi to explore its capabilities, using your own data. Or better yet, attend our UMR course in Austin TX this August to see how relationships among multiple chemical, physical and biological variables can be untangled and clarified with this and other interesting software.

3. New Time Series class for frequently-collected data

Some scientific data today are measured electronically at closely-spaced intervals in time, such as every 15 minutes or more frequently. Automatic recorders in streams and estuaries measure values only minutes apart from one another; remotely-sensed data provide an essentially continuous stream of values. Data this close together in time generally violate the independence assumption of standard statistical procedures – the

assumption that one value provides no information on the magnitude of the next observation. The consequences are that standard statistical methods such as hypothesis tests and regression produce invalid results, and bad models [see the January 09 newsletter on serial correlation for more info].

This winter Practical Stats will begin offering a new two-day Time Series Methods course focusing on hypothesis tests and regression models for data measured frequently in time. The goal is that students would be able to produce better models, better tests, and better forecasts after taking this course. Topics covered include:

- * How to determine whether serial correlation is present
- * Effects of serial correlation on hypothesis tests and regression models
- * Testing correlated data using the effective sample size
- * Regression models of correlated data – Trends, Seasonal components, Step changes
- * How to recognize autoregressive, moving average, and other time series models
- * Forecasting (water-quality) variables for serially-correlated data
- * Relating 2 or more time series (leading and lagging variables)
- * Line of organic correlation for estimating percentiles of forecasted data
- * Comparing serially-correlated data to standards
- * Bootstrap methods for time series data

Look for times and locations later this year on our website, at http://www.practicalstats.com/new_classes/classes.html

We hope to offer the first class in a unique and spectacular setting. Stay tuned.

'Til next time,

Practical Stats

-- Make sense of your data