Practical Stats Newsletter for February 2009

Subscribe and Unsubscribe: http://practicalstats.com/news
All of our past newsletters: http://practicalstats.com/news

In this newsletter:
1. Upcoming Courses
2. PCA, Factor Analysis and Multiple Regression
3. Some New Courses In the Pipe

1. Upcoming Courses
Only a few days left to register at the discounted price for our upcoming Applied Environmental Statistics course. The price goes up by $100 on Feb 16th. AES is our flagship, one-week course that enables you to "make sense of your data". It will be held on the campus of the Colorado School of Mines in Golden, CO on March 2-6, 2009. Be sure to check our course website for a complete course outline, and to register online for the course. Registration is $1395 if you register before Feb 16th, so don't delay. A 10% discount is also available for multiple registrations on the same credit card. Topics added over the last few years include bootstrapping and permutation tests, both modern methods to avoid assumptions of normality when it is not valid, or when data sets are small enough that it cannot be assumed.

Summer Courses scheduled
Nondetects And Data Analysis, the course that illustrates methods for correctly handling data with nondetects, will be held August 26-27 at the Hilton Garden Inn Downtown in Austin, TX. Online registration is available at
http://www.practicalstats.com/new_classes/classes.html.
Stop substituting one-half the detection limit, and make sense of data with nondetects. New topics include how to sum a series of components to get a total when some components are nondetects.

Untangling Multivariate Relationships is our 2-day course covering the multivariate methods of primary interest to environmental science, focusing on what each method is designed to do, when to use them, and when not to. Methods that foster interpretation of relationships between chemical and biological measures are highlighted. UMR will be held Aug 24-25, 2009 at the Hilton Garden Inn Downtown in Austin TX, just prior to the nondetects course. Register online at our New Classes page (URL above and below).

You can always find a complete course listing at
http://www.practicalstats.com/new_classes/classes.html.

2. PCA, Factor Analysis and Multiple Regression
PCA is not just peanuts, but it has sickened or otherwise kept some scientists at a distance. One of the primary methods discussed in our Untangling Multivariate Relationships course, Principal Component Analysis (PCA) is a useful technique for

finding major axes of direction through a multivariate maze of data. It is sometimes confused with Factor Analysis, and its axis through multidimensional space is an alternative to another line generator, Multiple Regression. How do these three methods differ?

The first principal component is the axis that shoots through a data cloud by minimizing the distances between each point and the axis, where those distances are measured perpendicular radially around the axis. Points to the left of the line will have distances ("errors") going off to the left, and points to the right will have distances going off to the right. The result is an axis that describes the "longest direction" of the data. If the data were cigar-shaped, the first principal component would be the axis right down the center of the length of the cigar.

In contrast, multiple regression considers one direction and one variable as special. This is the variable plotted on the Y axis. All "errors" are computed in the Y direction. The multiple regression line with be the line that minimizes the squared distances in the Y direction between the line and the points. Because these distances are in a different direction than those used by PCA, the resulting regression line also differs from the first PCA axis. When the purpose is to predict values of Y, or for some other reason the Y axis is preeminent, then using multiple regression makes sense. When all variables are created equal, the PCA axis is the appropriate choice.

Factor Analysis (FA) is often confused with PCA in both books and software. FA has gotten a lot of bad press over the years because people have not understood what it can do, and often try to make it do something it cannot. FA creates new axes that follow, if possible, clusters of original variables. FA axes follow clusters of variables, while PCA axes follow the data. Think of the cigar-shaped data cloud again, this time oriented at 45 degrees between two axes, X1 and X2. Suppose there are two variables that generally go in the direction of X1, both describing the wetness of the climate (say precip volume and number of wet days in the previous month). Also suppose there are three variables going in the general direction of X2, all describing the abundance and diversity of vegetation found at the numerous sites. The PCA first component will still go through the center of the data, at 45 degrees between X1 and X2. Factor Analysis will result in two factors, one generally heading in the direction of X1 and interpreted as a wetness factor, and the second generally heading in the X2 direction and interpreted as a vegetation factor.

PCA, Factor Analysis and Multiple Regression have different goals, and result in axes going in different directions within multidimensional space. Keeping track of the goals of each will help you decide which to use in a given situation. Of course, much more info on them is presented in our Untangling Multivariate Relationships course this August.

3. Some New Courses In the Pipe

We've taught two new courses at several locations recently, and will likely take one or both public in late 2009 or early 2010.

a) "Time Series and Forecasting" applies methods for serially-correlated data to the frequent (15 minute intervals, for example) observations automatically measured and stored by modern electronic recorders. Methods for correctly adjusting for the lack of independence (see the 09_Jan Practical Stats newsletter) in hypothesis tests and regression models are featured. Also featured are potential pitfalls in using these data to predict values as surrogates for other more expensive, less-frequently measured data.

b) "Environmental Stats using R" teaches much of the material in our AES course using the free R statistical software system. R is the most comprehensive and up-to-date statistical software available, and runs on PCs, Macintosh and Linux operating systems. See our Summer 2006 newsletter for more detail. In the past it has been difficult for environmental scientists to learn yet another programming language. This course shows how R can be mastered with much less pain.

If you have interest in having either of these two new courses held at your site, email us: http://www.practicalstats.com/Contact/contactus/index.php

BTW, our traveling introductory lecture, *Handling Nondetect Data Correctly*, will be presented Apr 23, 2009 at the SETAC Rocky Mt Chapter meeting, held at the USEPA Region 8 offices in Denver CO. For more information, see the chapter's website at: http://www.setac.org/rmrc/ .

'Til next time,

Practical Stats
-- Make sense of your data