Practical Stats Newsletter for January 2009

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

In this newsletter:
1. Upcoming Courses
2. An Introduction to Effects of Serial Correlation
3. Upcoming Lectures

Happy New Year and a special welcome to those who have joined the newsletter distribution since our previous review of stat software last fall. You will find all our past newsletters on the Practical Stats newsletter site, http://practicalstats.com/news

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

Applied Environmental Statistics, our flagship one-week course to enable you to "make sense of your data", 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.

Nondetects And Data Analysis, the course that illustrates methods for correctly handling data with nondetects, will be held this summer. Negotiations are ongoing, but it appears this course and our multivariate course (see below) will again be held the same week, probably in Austin TX. Details and registration should be available soon online.

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. More detail on course content is on our website.

2. An Introduction to Effects of Serial Correlation
We are beginning to teach new courses to sponsoring organizations on the R statistical package, and on time series methods. Time series procedures have not been widely taught or applied in the environmental sciences outside of slowly-changing phenomena such as lake levels or 15-minute precip volumes. Today, however, some water quality measurements are beginning to be collected by automatic methods every 15 minutes or so. Data collected this frequently have a strong memory (correlation) from one measurement to another. This correlation violates the assumptions of hypothesis tests

and regression, where each observation is considered uncorrelated from (independent of) the next. Using all of the data in these procedures without modification results in p-values that are too small, so that tests declare differences when they should not and explanatory variables are considered significant when they should not. Recognition of this problem has spurred interest in the use of time series and similar methods.

The basic effect of violating the independence assumption is that the sample size N used in calculations is too large. There are N observations, but not N independent observations. A simple application is when calculating a confidence interval around the mean. The length of the interval is built upon the standard error of the mean. Because N is too large, the standard error (the standard deviation divided by the square root of N) is too small. The interval is therefore too short. Similarly, the standard error and its square, the error variance are too small when conducting hypothesis tests. These parameters are found in the denominator of all parametric hypothesis tests. If a t-test were employed to differentiate the means of two groups, the denominator of the test is too small due to serial correlation, and so the test statistic is too large. It rejects the null hypothesis and declares differences when those differences are not really there.

Two simple concepts can get you started in adjusting for serial correlation. First, measure it. Pair sequential observations (x1, x2) where x1 is the first observation in time and x2 the second. Similarly for (x2, x3) all the way down to (x n-1, xn), the last two observations in sequence. Compute correlation using these n-1 pairs and you have a measure of serial correlation. If there is a sequential dependence, the correlation coefficient will be significant.

Second, adjust simple hypothesis tests and/or sample sizes to their 'effective' values, the values representing the number of independent pieces of information rather than the number of (correlated) measurements. This idea is not new, and was presented for water resources data by Matalas and Langbein in a 1962 paper of the Journal of Geophysical Research (vol 67, no 9). The effective sample size Neff will be less than the observed sample size N when serial correlation is present. For the serial correlation coefficient r, the effective sample size is

Neff = N / { (1+r)/(1-r) – (2/N)*r*((1-r**N)/(1-r)**2}, where ** represents 'raised to the power'.

The value between the brackets will be greater than 1 when r is greater than 0. Using the effective sample size instead of the observed sample size is a simple way to adjust for serial correlation in hypothesis tests and confidence interval calculations. For regression, more formal time series methods are warranted.

3. Upcoming Lectures

Dennis Helsel is giving an introductory lecture, *Handling Nondetect Data Correctly*, at the following locations. The April SETAC meeting is open for outside attendees.

Jan 20, 2009 NY Dept. of Environmental Conservation, Albany NY.
Feb 20, 2009 South Florida Water Mgmt District, W. Palm Beach.

Apr 23, 2009 SETAC Rocky Mt Chapter and USEPA Region 8, 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