Practical Stats Newsletter for April 2009

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

In this newsletter:

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 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. We focus on newer, nonparametric methods for establishing links between variables. UMR will be held Aug 24-25, 2009 at the Hilton Garden Inn Downtown in Austin TX. Register online at our New 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.

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

2. The Sign Test for tied data, including nondetects

One of the simplest and oldest tests in statistics for paired data is the sign test (*see our March 2005 newsletter for a discussion of tests for paired data*). The difference between

two numbers within a row is computed, where the two columns represent different groups. The proportion of increases and decreases are counted, with the expectation that if levels of the data in the two groups is the same, about half of the pairs will show an increase (the sign of the difference is a plus), and about half of the pairs will show a decrease (the sign is a minus). If the percent of pluses or minuses is significantly different from 50%, the levels of the data in the two groups are judged to differ.

The sign test suffers from an assumption that ties do not occur – the differences are never zero. For environmental data, this is unrealistic. Nondetects often produce a large number of ties, i.e. a <1 being compared to a <1, for example. Counts of organisms are discrete, and so may have the same values per quadrant of earth, etc. For a small number of ties, the conventional adaptation is to ignore (delete) the ties, counting only the pluses and minuses. This 'quick fix' often leads to a rejection of the null hypothesis when it should not, as ties are evidence for the similarity of the two groups.

Several journal articles since 1990 have discussed adaptations of the sign test to handle ties. One that is easy to implement is the "Modified sign test" of Fong et al. (2003, The American Statistician, v 57, 237-240). This version of the sign test is used for nondetects in the textbook *Nondetects And Data Analysis*, the textbook for our training course of the same name. A macro to compute it (the csign macro) is available on the Practical Stats website, as part of the NADA macros for Minitab.
http://www.practicalstats.com/nada/nada/downloads.html
While presented as a macro for data with nondetects, the macro can be used for any data with ties, including counts of organisms. Simply declare all of the data in both columns to be detected values.

Here's an example of use of the modified sign test. Two sites are compared, an up(gradient or stream) site that is likely clean, and a down(gradient or stream) site that may be contaminated. The zeros may represent "<1" of a concentration, or might be 0 counts of a nasty organism that we hope is not often present.

	Up	Down
1	0	0
2	0	0
3	0	0
4	0	0
5	0	0
6	0	2
7	0	4
8	2	5
9	2	7
10	3	5
11	6	6

We want to compare the null hypothesis of "for half of the pairs we'd expect Up>Down, and half Up<Down" indicating that the two groups have similar levels, to the alternative

of "Down looks contaminated", or Down>Up more than 50% of the time. If the percent of pairs where Down>Up is quite larger than 50%, we will reject the null hypothesis of no contamination at Down and go with the alternative, that there is contamination at Down.

The traditional sign test looks at this and says "ties don't count", so we have 5 non-tied cases (rows 6-10). Of these 5, all are higher at Down. 5 pluses out of 5 cases seems rather strong evidence to reject similarity, and the test finds a p-value of 0.03. The standard sign test rejects the null hypothesis, declaring that contamination has occurred.

But whoa! We dropped out 6 of the 11 cases when they didn't show an increase. There isn't any reason to do that, except that the traditional sign test assumed that ties wouldn't occur. In reality, 5 out of the 11 observations show an increase, but 6 do not. We could go to the other extreme and say that zeros are the same as decreases, and compute the standard sign test with 5 pluses and 6 "minuses" or no-increases. This p-value is 0.726, indicating little evidence that Down is greater than Up. But this again doesn't agree with reality, these are zeros, no change, not decreases. So the best p-value for reflecting the presence of ties must be somewhere between these extremes.

Details on the modified sign test are found in the *NADA* textbook, and of course in the Fong et al. article presenting the test. In essence, the p-value of 0.726 for obtaining 5 increases out of 11 cases is adjusted downwards, counting the zeros as zeros instead of minuses. The tie-corrected p-value equals 0.375, reflecting the pattern for all 11 observations rather than just the 5 nonzero differences.

The modified sign test should be in the toolbox of any environmental scientist that deals with nondetects, or counts of organisms. It is one of the many procedures we discuss in the upcoming NADA class this summer in Austin, TX. Using the correct test makes a difference. Knowing about correct procedures is essential to doing a good job. Can you afford to not upgrade your skills?

3. The Cost of Upgrading Skills versus Not Upgrading

By looking for efficiencies, Practical Stats has kept the cost of our training courses stable for years. Tuition for our Applied Environmental Statistics course was $1495 in 1999, and is $1395 now ten years later, as one example. We've continually added new methods as statisticians have developed them that improve your ability to analyze environmental data. The current economy is tough, so changing to meet the challenges of the future means upgrading the skills of yourself and/or your employees. Training dollars are tight, but we're fighting back. We've placed our courses this year in locations where you can come and stay without renting a car, downtown locations where food and hotels are adjacent to the course venues. We've looked for locations where costs are lower than larger cities, yet access by road and airports is relatively easy.

We've also done two specific things to further help with costs. First, we've reduced the registration cost for our August 09 ***Untangling Multivariate Statistics*** two-day course by $100 over last year. It will likely return to its normal, higher price next year.

Second, we're adding an optional half-day seminar on using R statistical software at the end of our AES course in Tacoma, Washington. Any person registered for AES can stay Friday afternoon and learn the basics of the free R statistical system. R is the international standard for statistics, was developed with the intent of being freely available, and runs on Windows, Macintosh and Linux operating systems. See our Summer 2006 newsletter (*06Summer_R.pdf*) on the Practical Stats website for more information on R. Or better yet, register and join us in Tacoma this September!

'Til next time,

Practical Stats
-- Make sense of your data