Practical Stats Newsletter for Spring, 2003

It snowed here in Colorado again last week. But a spring snow -- 6 inches. During the winter we had over 4 feet of snow in two days! So with 6 inches, it must be spring. Over the winter I've been working on a book on treatment of censored data, data with 'less-thans'. Below is an example of the issues involved. The book should be out at the end of 2003.

In this newsletter:
1. Upcoming course schedules
2. Why substituting one-half for less-thans is a really bad idea.
3. Updates on Excel as a statistics package

1. Course Schedules

There are still openings for the Applied Environmental Statistics course in Golden, Colorado this June. Our apologies to those of you who were planning on attending the Sacramento course in late April. It set the record for the quickest yet to fill up. Colorado in June is beautiful. Hiking, camping, and no snow until you get up to Rocky Mountain National Park.

If you've been waiting for the "Less Than Obvious" course on handling nondetects to be taught again, look for an announcement soon. The course content has been undergoing a complete remake -- there's a lot of new methods out there that are being incorporated into the course and accompanying textbook. Expect a class in late 2003, along with the release of the book. If you would like a course taught in your area, and know that at least 10 students are available, email us at mailto:ask[at]practicalstats.com

Course listings are always current on our website. Let folks know that the June course in Golden still has openings.

2. Why substituting one-half for less-thans is a really bad idea.

Consider a comparison of two groups of data, one a possibly-contaminated test group and the other a control group. Are concentrations in the test group the same, or higher (one-sided alternative), than in the control group?

The classic approach for this design is the two-sample t-test. If data distributions do not follow a normal distribution, the nonparametric Mann-Whitney (also called Wilcoxon rank-sum) test should be run instead. With either test, a roadblock looms in the data of Table 1.1 – there are values below detection limits. Several detection limits.

Control Group		Test Group	
<1	<1	<2	<5
<1	<1	<2	<5
<1	<1	3.3	<5

<1	4.1	3.4	<5
1.0	7.0	<2	4.7
1.8	7.5	12.2	<5
2.2	15.4	<5	22.5
<2		6.6	

Table 1.1 Contaminant concentrations in a test and a control group, with multiple detection limits.

The most common method in environmental studies for dealing with such data is to substitute one-half the detection limit. Procedures manuals for at least two Federal agencies recommend this practice. Substitution of one-half dl in the above data results in the data of Table 1.2, and a Mann-Whitney test p-value of 0.015. The equivalence of the groups is rejected, and the two sets of data are declared different. Expensive remediation actions might be mandated for the elevated levels in the test group. Soil is ripped up. Industrial equipment is modified. Wells are closed.

Control Group		Test Group	
0.5	0.5	1.0	2.5
0.5	0.5	1.0	2.5
0.5	0.5	3.3	2.5
0.5	4.1	3.4	2.5
1.0	7.0	1.0	4.7
1.8	7.5	12.2	2.5
2.2	15.4	2.5	22.5
1.0		6.6	

Table 1.2 Contaminant concentrations in a test and a control group after substituting one-half the detection limit for nondetects

Now let's pull back a curtain. These data were not field data, but were computer generated. By generating data, the true situation is known. The data from both groups were generated from the same distribution – there is actually NO difference in their mean or median levels. Any reasonable method for analyzing these data should find no difference in the two groups. A Mann-Whitney test on the original (uncensored) data had a p-value = 0.43 prior to recoding some data with the detection limits of Table 1.1.

The fundamental problem with substituting one-half, or any other function of the detection limit, is in the statement that something is known (the values for nondetects) that really is not known. The test procedure takes our word that the value is really 0.5 times the detection limit, not some other value below the limit. The true value may have been anywhere below the detection limit, as far as we know. To compound this problem, the value substituted is not a function of anything known about the media sampled (the organism, water, or soil). It is a function of the precision, or lack of it, in the laboratory. It is a function of the process used by that laboratory, and laboratories differ considerably in how they compute detection limits. It may be a function of time, or of the dilution of the samples, or of the standard practices of the analyst, or of other conditions in the laboratory process. Using substitution can easily impose an artificial signal that

originally was not there. With substitution, the pattern of detection limits may impose a signal having nothing to do with the values in the samples themselves. This is especially a problem with multiple detection limits.

An error can also be easily made in the opposite direction with substitution. It is trivial to generate data whose levels originally DO differ between groups, but after censoring and substitution a t-test or Mann-Whitney test fails to find any evidence of difference. The result is not just a wrong conclusion by an hypothesis test. In the real world, contamination goes unnoticed. Remediation goes undone. Public health is unknowingly threatened.

Substitution is never necessary. It is especially bad when multiple detection limits are present. There are better ways. Valid methods exist for computing summary statistics, hypothesis tests, correlation and regression for censored data, even data with multiple detection limits. These methods have been used in disciplines other than environmental science for years. For example, a nonparametric score test produces a p-value of 0.47 for the censored Table 1.1 data. No substitution is involved, and multiple detection limits are handled with ease. I'll tell you more about these type of tests in upcoming newsletters. And of course in full form in the book and course by the end of this year.

3. Updates on Excel as a statistics package

Update I -- Excel.
In our Fall 2002 newsletter we spent a great deal of space discussing whether Excel was an adequate statistics package. We trust you found it helpful. If you missed it, the information is on the PracticalStats webpage. Our short answer to the question above was "Unless you have very simple needs, no it is not".

Since that time others have joined the discussion.

The American Statistical Association set out guidelines in late 2002 for the teaching and use of statistics. Within their recommendations document at
http://www.amstat.org/education/ASAendorsement.html
they make this statement: " Efficient computing tools are essential for statistical research, consulting, and teaching. Generic packages such as Excel are not sufficient even for the teaching of statistics, let alone for research and consulting."

A pretty definitive statement.

The authors of a 1999 review of Excel as a statistical tool came out with an update in December 2002, just after our Fall newsletter. They found that little had changed in the interim in the capabilities of Excel. It corroborates the findings in our summary, and is found in the article:

On the accuracy of statistical procedures in Microsoft Excel 2000 and Excel XP

B.D. McCullough and B. Wilson, (2002), Computational Statistics & Data Analysis, 40, pp 713 - 721.

Update II -- Statistical Methods in Water Resources.
The Helsel and Hirsch textbook (used in our Applied Environmental Statistics course) recently hit 60,000 downloads. And it is now out in a new version, version 1.1 . The only changes from the version available September 2002 to April 2003 is that a few errors were corrected. They're small enough that most of you would never find them. But if you want the latest and greatest, the new version is on the USGS website, http://water.usgs.gov/pubs/twri/twri4a3/

The errors (in a few formulae) are not worth reprinting the book if you've previously printed a paper version, but if you work from the digital version, grab it.

Til next time,

Practical Stats
http://www.practicalstats.com
-- Make sense of your data