Practical Stats Newsletter for May 2010

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

In this newsletter:
1. Upcoming Courses
2. Simple Techniques for Data with Nondetects
3. New Minitab macros

1. Upcoming Courses
Applied Environmental Statistics
Sept. 13-17, 2010 $1395 registration before August 30, 2010
$1495 after August 30th.
held at the Hampton Inn Tropicana
4975 S. Dean Martin Dr.
Las Vegas, Nevada 89118-1656

We continue to try and minimize costs for students. We looked across the country for a location with easy access by air and the lowest airfares, hotel and food costs. The best we could find was Las Vegas. The Hampton Inn is giving students a $79 per night rate. With these low costs, and the likelihood that we'll need to increase course registration costs next year, this is probably the AES course to attend if you want to keep costs low. Our registration fees have remained stable for over 10 years. You'll notice other technical training costs are much higher -- for example, Minitab offers a week-long software training course for around $2500. Sign up now to insure a seat, and afterwards tell your office staff and contacts about it. Last September's AES course in Tacoma was full.

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

2. Simple Techniques for Data with Nondetects
Nondetects And Data Analysis (both the course and textbook) spend considerable time on survival analysis methods applied to censored environmental data. In spite of the many documented cases of errors arising from substituting values for nondetects (see http://dx.doi.org/10.1093/annhyg/mep092 for a new explanation of the dangers), substitution remains popular because it is so easy. Yet there are two other easy solutions that do not introduce the errors inherent in substitution.

-- *For computing descriptive statistics*
What is a measure of the center for the following dataset?
<5 <5 8 15 19 24 27 33 41

To compute the mean we would need to use a survival analysis procedure like maximum likelihood estimation (MLE) to avoid the 'invasive data' problem of substitution. But there are two easier solutions.
a) use the median. The median of these 9 values is the 5th observation from the bottom, or 19.
b) use the percent of data above the detection limit. There are 7/9, or 78% of the data above 5.

And for this second dataset with two detection limits,
<5 <5 8 15 <20 24 27 33 41
the two options remain the same. The median is <20. And there are 4/9 or 44% of the observations above 20.

-- *In general*
We can generalize these two procedures to other statistical applications. First, we can use methods based on percentiles, or ranks. The rank i divided by (n+1) is the position for the percentile of that observation. The 6th ranked observation (24) above, for example, is at (6/10) or the 60th percentile. Methods based on ranks such as nonparametric hypothesis tests are procedures analyzing the percentiles of the data.

Second, we can treat the data as being either above or below the (highest) detection limit, and interpret the proportion of data falling above the limit. Binomial procedures allow us to discern changes in such proportions.

-- *For hypothesis testing*
Nonparametric hypothesis tests are based on percentiles, or ranks. To compare two sets of data, the Mann-Whitney (also called rank-sum) test can always be used without requiring substitution. The test determines whether the cdfs (the set of percentiles) in the two data sets are similar, or different. If multiple detection limits are present, all data below the highest limit are coded as being tied in order to use the simple version of this test. For example, the two data sets:
<5 <5 8 15 <20 24 27 33 41
and
<5 <5 <5 <5 6 9 10 12 16 21

are re-coded to
<20 <20 <20 <20 <20 24 27 33 41
and
<20 <20 <20 <20 <20 <20 <20 <20 <20 21

and their ranks are:
7.5 7.5 7.5 7.5 7.5 16 17 18 19
and
7.5 7.5 7.5 7.5 7.5 7.5 7.5 7.5 7.5 15

The identical procedure would precede the Kruskal-Wallis test when there are three or more groups. This simple method is far superior to substitution because no artificial pattern alien to the original data is placed into the data, as it is with substitution. Unlike substitution followed by t-tests or ANOVA, if differences are found by the Mann-Whitney or Kruskal-Wallis tests, they can be believed. If patterns are obscured by re-censoring at the highest limit, more complicated survival analysis methods are available. But we're trying here to keep it simple.

The second test procedure is the binomial-based 'test of proportions' or contingency table test. Here the proportions of data above the (highest) detection limit are tested for similarity or difference. Data are coded into two groups, above and below the highest detection limit. For the above data, the first group has 4/9 or 44% of values above 20, while the second group has only 1/10 or 10% above the value of 20. The test determines whether these two percentages are significantly different.

-- *For correlation and regression*
Nonparametric correlation coefficients Kendall's tau and Spearman's rho may be computed on data with nondetects, without substitution. Kendall's tau easily handles data with multiple detection limits, though the software is not usually written to do so. The nonparametric Theil-Sen line (used for the Mann-Kendall trend test, for example) may end up with a "<value" slope, however, so for a simple regression approach, the binary approach might be better. There is a multiple-detection limit version of Theil-Sen called the ATS line which avoids computing a less-than for the slope if 'keeping it simple' is inadequate.

One approach to perform regression with a binary Y variable is called logistic regression. Here the probability of being in the higher category, say the probability of recording a detected value, is predicted from one or more explanatory variables. Interpretation of the results is very similar to ordinary regression.

So all in all, these simple methods do not require substitution, can be computed with standard statistical software, and avoid the pitfalls of 'invasive data' that result from fabricating data by substitution. If you can't justify going to the more complicated procedures in *Nondetects And Data Analysis* that handle nondetects at multiple levels, these simpler methods should meet your requirements.

3. New Minitab macros
Version 2.8 of our NADA for Minitab macros are now available on our website.
http://practicalstats.com/NADA/downloads.html

These macros have been tested and run smoothly in the new version 16 of Minitab, as well as version 15. The macros that perform bootstrapping for confidence intervals of the mean and median of data with nondetects have been improved and renamed. A new macro has been added that performs bootstrapping for Regression on Order Statistics (ROS). For the first time, confidence intervals on the mean, including the UCL95, are

easily available using ROS. We expect this to be the last update of the NADA for Minitab macros before the second edition of the *Nondetects And Data Analysis* textbook arrives at the start of 2011.

'Til next time,

Practical Stats (Dennis Helsel)
-- Make sense of your data