Practical Stats Newsletter   for June, 2005

In this newsletter:
1. New NADA software freely available for download
2. Testing for normality with censored data
3. Course Schedule?

1. New NADA software freely available for download

Version 1.6 of the NADA macros for Minitab is now available at
http://www.practicalstats.com/nada

The new version includes a new macro, KMSTATS, which easily computes summary statistics for data with nondetects using the Kaplan-Meier method described in the NADA textbook. The macro "flips" the data, and "unflips" the results back to original units. Version 1.6 also includes an update of the KMCENS macro, adding optional multiple comparison tests following the Kruskal-Wallis test for censored data. The latter was by request - we're glad to get your input on what routines you would find most useful. Just email nada[at]practicalstats.com .

Also on the website is a link to the NADA for R software, to which updates have been made since the last newsletter. R is a free statistical package that runs on Macs, PCs and Unix machines. It is based on the S statistical language. R does not have a GUI interface, but is called by typing commands on a command line. But if you're fine with R, NADA for R provides an increasing number of the routines found in the NADA textbook.

2. Testing for normality with censored data

The topic in this month's newsletter is also a request from a reader. Though we can't answer every request, we're glad to get your questions on how to interpret environmental data. Some of your questions may end up here!

Censored data include some values known only to be below or above some threshold. The most common application in environmental data is to "nondetect" or "less-than" data. How can you test such data to see if they fit a standard distribution such as the normal?

Remeber that when data are not censored, here's the procedure:
1. Values are ranked from smallest to largest and assigned a rank of i = 1 to n. The observation with the smallest value receives a rank of 1, up to the largest equals n.
2. Ranks are turned into a percentile or cumulative frequency, a value between 0 and 1, using a plotting position formula. The most common formula for constructing a normal probability plot is the Blom formula. See Helsel and Hirsch (2002) http://pubs.water.usgs.gov/twri4a3 page 23 for a listing of commonly used plotting

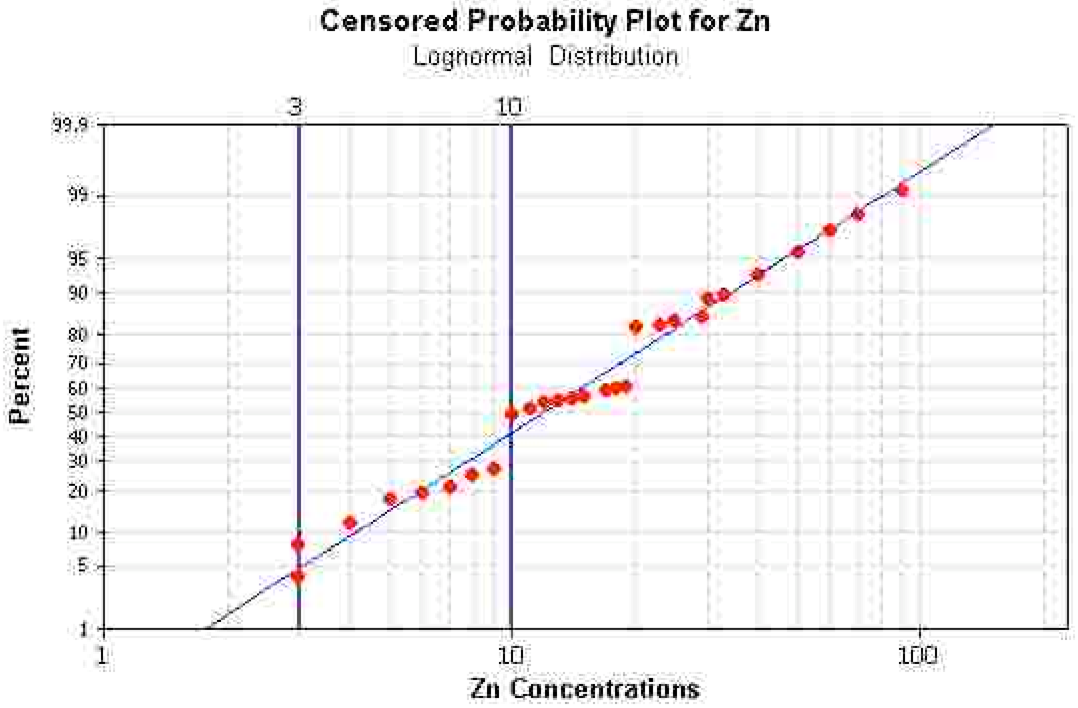
positions. They chose the Cunnane formula, and gave reasons why. However, for probability plots and associated tests, the Blom plotting position is most often used.
3. If cumulative frequency were plotted on a linear y scale, versus the data values on the x axis, you would have a cumulative distribution function (cdf), sometimes called an empirical cdf. This graph is also sometimes called a quantile plot. See figure 2.4 of H&H for an example. It is pretty standard for these plots that the percentiles go on the Y axis, and the data on the X axis.
4. Cdfs are transformed into a normal probability plot by "stretching out" the ends of the frequency axis, making it nonlinear. This is done mathematically by transforming the cumulative frequencies (0 to 1) into "normal quantiles" or "standard normal deviates", found by entering a table of the standard normal distribution. These values will go from approximately -3 to +3, with median in the middle at 0. Plotting normal quantiles versus the data produces a normal probability plot. Some people plot the normal quantiles on the X axis, others on the Y axis. There is no standardization. See figure 2.8 of H&H for an example. Some software will show the linear "normal quantiles" axis. Other software will show a nonlinear cumulative frequency or percentile scale. For an example of use of the nonlinear scale ("probability paper"), ee figure 2.9 of H&H.

Testing normlity when data are censored
When data are censored, the detected values can be plotted similarly to the process above. Nondetects are not plotted - there is no unique, single number known to represent a nondetect. It lies somewhere between 0 and the detection limit. However, this plot is not the same as a standard probability plot after deleting the nondetects. Instead, the values of the nondetects influence the percentiles of the detected observations, and so their normal quantiles. Detects will be shifted to the right on a probability plot, as compared to a plot where nondetects were deleted. This is easiest to see and understand when there is only one detection limit (DL).

For one DL
With one DL, a censored probability plot will be identical to drawing a probability plot where all nondetects are set to be the same value, anywhere between 0 and 0.99 times the DL. Draw the probability plot, then wipe off the nondetects from the plot - we don't really know which single data value to plot them at. So if there are 30% nondetects, the plotting position or percentile of the lowest detected observation will start at just above 0.30. The CROS macro can be used to create a probability plot for censored data. See Figure 5.5 of NADA, where the DL is drawn as a horizontal line across the plot.

For multiple DLs
With more than 1 DL, adjustments must be made in the percentiles of detects based on the numbers of data, both detects and nondetects, falling between the DLs. The CROS software will do this for you, and is available on the NADA website. It uses the Helsel-Cohn (1988) definitions of plotting positions, spreading nondetects evenly in probability space between the probabilities of 0 and the detection limit for that observation. The resulting percentiles for nondetects influence where the detects are plotted. See figure 5.6 of NADA, where there are three detected values plotted in-between the two detection limits. Commercial software for survival analysis will perform a similar procedure using

either Kaplan-Meier or Maximum Likelihood estimates of percentiles for the detected observations. In addition, it usually performs an official hypothesis test for normality. Minitab, for example, will draw a probability plot for censored data using its Reliability/Survival > Distribution Analysis menu items, and test using the Anderson-Darling test. For example, the plot below is a censored probability plot of Zn data with DLs at 3 and 10. The detected values are plotted to compare to a lognormal distribution.



3. Course Schedule?

We expect an open-enfollment version of both Applied Environmental Statistics and Less Than Obvious to be offered in late Fall or Winter. Right now our schedule is quite busy with courses directly taught to agencies, and the open-enrollment course scheduling is lagging. New courses will always be posted on the Practical Stats webpage, usually about 4 months before a course is offered. So keep checking. We'll keep trying on this end.

'Til next time,

Practical Stats
http://www.practicalstats.com

-- Make sense of your data