Practical Stats Newsletter for January 2013

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

In this newsletter:
1. Upcoming Events
2. Graphing Censored Data
3. The Unofficial Users Guide to ProUCL4 (e-book) is now available

1. Upcoming Events

The 2013 Practical Stats course and webinar schedule is now available on our webpage. This year we'll present two webinar series, the Nondetects And Data Analysis series, and the Applied Environmental Statistics (AES) series. You may register for an entire series, or individual webinars. The AES series replaces one of the two in-person AES classes we have taught each year with a webinar format containing much of the material from our one-week AES course. An in-person AES course will also be held during fall 2013 in the Denver CO area. Other in-person courses will include Time Series and Forecasting and Untangling Multivariate Relationships, both in summer 2013 and located in the Olympia/Tacoma/Seattle WA area.

Our first two webinars:

Monday Feb 25. Urban Legends in Environmental Statistics $50
There are seven common errors made in environmental data analysis. Learn why parametric methods rarely have more power than nonparametric tests; why t-tests on logarithms do not tell you whether one group's mean is higher than another; why outlier tests cannot tell you whether data are 'wrong'; and the inside stories of four other common misconceptions in environmental statistics.

Monday March 25. Invasive Data: Why Not Substitute ½ the Detection Limit? $50
Measurements of trace chemicals in environmental media (water, air, soils, biota) frequently result in values reported only as less than the laboratory reporting limit ("less-thans", "nondetects", and "qualified values"). The most commonly-used method for incorporating nondetects is to substitute one-half the reporting limit and continue as usual. This may obscure patterns and trends that are present, or create those that are not present in the original data. It is fraught with error. Learn why, and what you can do instead.

To register and for more information, see our Training page at
http://www.practicalstats.com/training/

2. Graphing Censored Data

This month's topic is another request from newsletter subscribers. It has its own chapter in *Statistics for Censored Environmental Data using Minitab and R* (Helsel, 2012). http://www.wiley.com/WileyCDA/WileyTitle/productCd-EHEP002278.html
It is difficult to discuss graphs without showing more of them than can be done here, so I encourage you to take a look at the full chapter. You may also be interested in our June 2005 newsletter (available online at our 'News' tab) where we discussed the related topic of how to perform a goodness of fit (normality or Q-Q) test for censored data.

Censored data include values known only to be below or above a threshold, such as "nondetects" in environmental studies. When graphing data that include nondetects, here are the basic ideas:
a) nondetect values are not plotted as individual points, but they influence where detected observations occur on the plot.
b) plots for censored data are never identical to just deleting the nondetects and then drawing the plot. Such plots will always be wrong, and should be avoided.
c) plots where nondetects are shown at an individual value -- one-half the detection limit (DL), etc.-- are always wrong, and should be avoided.

To illustrate how the process works, consider a Q-Q plot, where data are compared to an assumed distribution such as the normal or lognormal distribution. 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 probability or cumulative frequency, a value between 0 and 1, using a plotting position formula. The standard formula for constructing a normal probability plot is the Blom formula (Helsel and Hirsch, 2002 – "H&H"), probability = (i-0.375)/(n+0.25).
3. Observed data values are plotted on one axis, and "normal quantiles" (also called "standard normal deviates") or probabilities plotted on the other. Normal quantiles are on a linear axis and range from approximately -3 to +3, with the mean/median in the middle at zero. Probabilities, if shown instead, are on a nonlinear axis. In the old days this was called "probability paper". Here is an example of observed data, their Blom probabilities and normal quantiles.

Data	Probability	Normal Quantile
0.427	0.068	-1.494
0.719	0.176	-0.932
0.969	0.284	-0.572
1.373	0.392	-0.274
2.749	0.500	0.000
3.496	0.608	0.274
5.831	0.716	0.572
6.083	0.824	0.932
11.798	0.932	1.494

When data are censored, the detected values can be plotted using the same process. A <2 is below a detected 5, for example, and so is ranked below the detected observation. Nondetects are not plotted, but they influence the probabilities of the detected observations. If the first 4 observations from the above data had been censored, there would be 4/9 = 44% nondetects, so the probability of the lowest detected observation will be above 0.44. For n=9 it is at 0.50.

Data	Probability	Normal Quantile
<2	--	not plotted
<2	--	not plotted
<2	--	not plotted
<2	--	not plotted
2.749	0.500	0.000
3.496	0.608	0.274
5.831	0.716	0.572
6.083	0.824	0.932
11.798	0.932	1.494

With more than 1 DL, adjustments must be made in the probabilities of detects based on the numbers of data, both detects and nondetects, falling between and below the DLs. Software computes these probabilities using either regression on order statistics (ROS, sometimes called "least squares"), Kaplan-Meier or maximum likelihood methods. Suppose there are two additional observations added to the above data, both of which are <5. Below are their probabilities computed using the ROS method.

Data	Probability	Normal Quantile
<2	--	not plotted
<2	--	not plotted
<2	--	not plotted
<2	--	not plotted
2.749	0.566	0.165
3.496	0.646	0.376
<5	--	not plotted
<5	--	not plotted
5.831	0.795	0.825
6.083	0.864	1.097
11.798	0.932	1.489

Note that the <5s are known to be below the three highest observations. The Blom probabilities for the top three out of 11 observations would have been 0.77, 0.86 and 0.94. Their ROS probabilities are similar, showing that the presence of nondetects below have been accounted for in the same way as if they had been lower detected values. The probabilities for the lower two detected observations also increase by the addition of the two <5s, as there is some probability that these data will be below the lower detects. See Helsel (2012) for a more detailed discussion of how ROS and the other methods work.

Censored boxplots display these data by plotting the portions of the diagram above the highest detection limit. For the above data including the <5s, the top portion of the box in Figure 1 is visible because the 75th percentile (probability = 0.75) is higher than the highest DL of 5. The center median line of the box is not visible, indicating that more than 50% of observations are below 5.

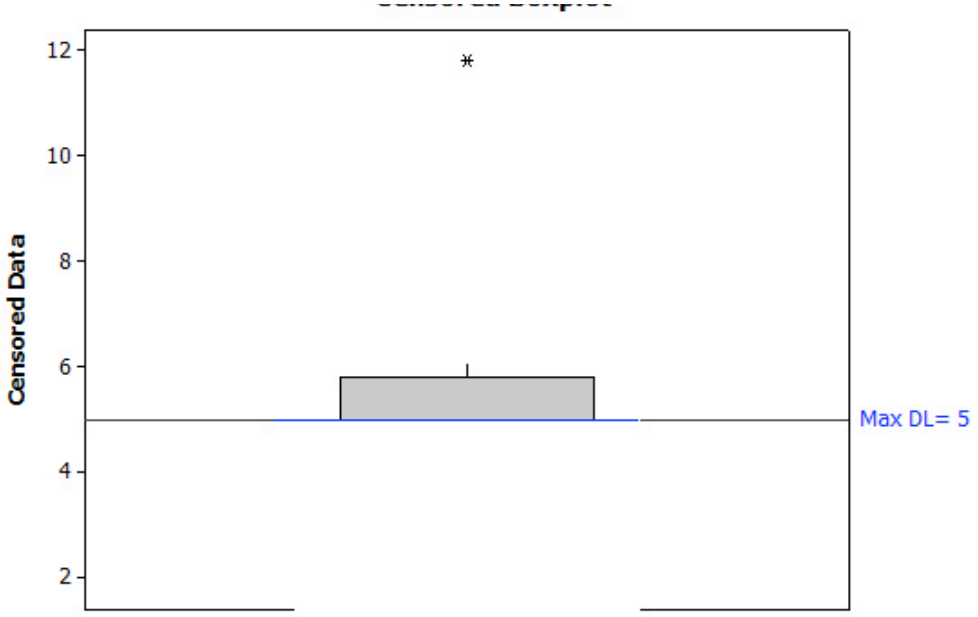

Figure 1. Boxplot of data with two detection limits

Censored probability or Q-Q plots (Figure 2) display each detected observation as a point, at probabilities adjusted for the presence of nondetects. This gives more detail than a censored boxplot.

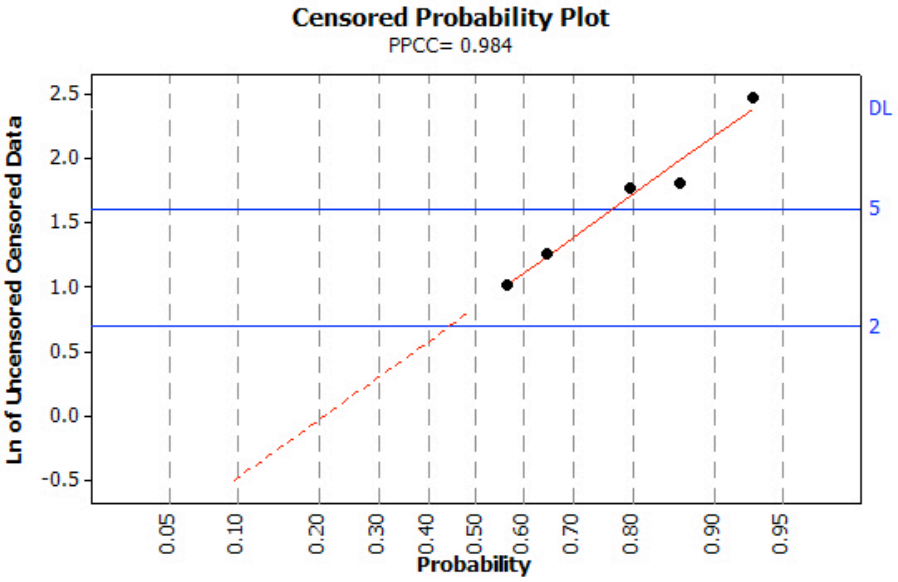

Figure 2. Q-Q plot of data with two detection limits

Scatterplots take a different approach for censored data. Nondetects are shown as the interval between which the possible values lie. A <5, for example, can be plotted as the interval [0,5] using a dashed line. All observations, both detects and nondetects, are plotted on a censored scatterplot. Detected observations are plotted as usual, with a single point. Using the above data set as the Y variable, a censored scatterplot is

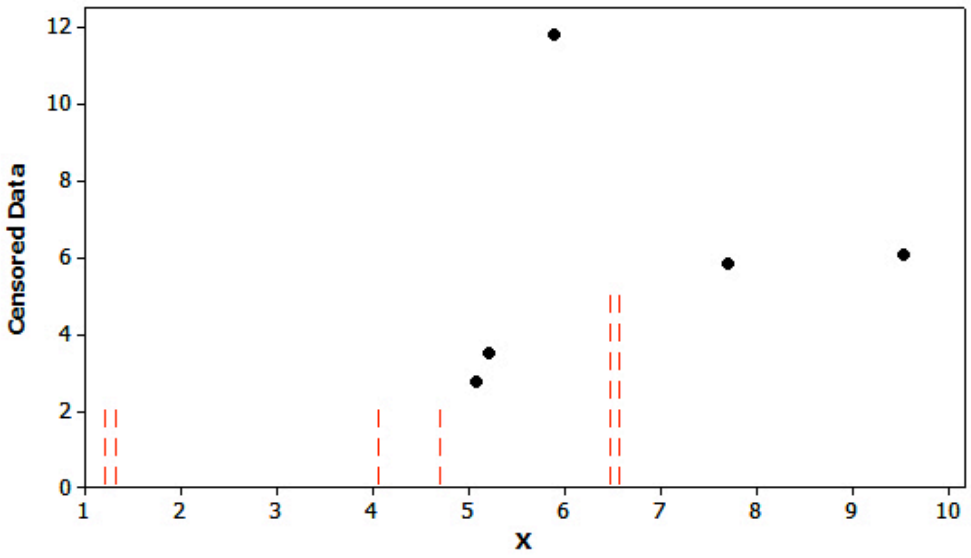

Figure 3. Scatterplot of data with two detection limits

Note that plotting nondetects at a single point, say at DL/2, would give a false image of the data. It is not known whether the <2 values are higher or lower than the <5 values, yet this would be indicated if half the DL were used. It is almost certain that all four <2s are not identical, yet they would appear so if half the DL were used. Substitution with a scatterplot will result in a false image of the data.

Much more detail on graphing data with nondetects is given in chapter 5 of Helsel (2012). The main principles to retain, however, are

1. Do not simply delete nondetects.
2. Do not plot nondetects at some fraction of their detection limits.
3. Plot detected values at positions adjusted for the presence of nondetects.

3. The Unofficial Users Guide to ProUCL4 (e-book) is now available

Our new e-book, *The Unofficial Users Guide to ProUCL4* is now available from the Kindle store at:
http://www.amazon.com/Unofficial-Users-Guide-ProUCL4-ebook/dp/B00AU8C1QI/

It can be read on Kindle, iOS, Macintosh, Windows, and Android platforms (computers, smartphones, tablets). Yes, this includes iPad. UUG4 is a concise, easy to read guide to USEPA's free software for estimating the UCL95, the statistic used in many regulatory applications. UUG4 is like a “home inspector” when purchasing a home. It looks for areas that are, and are not, performing well. It informs the reader of both so that you are

able to best use ProUCL4 with your data. To make this task intuitive, the book uses the ‘stoplight’ format, grading each menu option as either green (Go: in good working order), yellow (Caution), or red (Stop: not in working order). The most common grade for menu options in ProUCL4 is yellow, indicating that users will need a guide such as this one to get the most from their version 4 software, and avoid problem areas. UUG4 enables you to make sense of ProUCL4's output, and of your data.

'Til next time,

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