Practical Stats Newsletter for January 2019

Subscribe and unsubscribe: http://practicalstats.com/news
Archive of past newsletters
http://practicalstats.com/news/archive.html

In this newsletter:
A. Practical Stats Courses
B. Plotting Q-Q plots with nondetects. Why, how, and how NOT to plot them.
C. Free Webinars and U-Turns

A. Practical Stats Courses Our online training site: http://practicalstats.teachable.com/
Our self-paced Applied Environmental Statistics course is available in two parts, each $650 USD for a 1-year access for one person. Or get both courses together (equivalent to our week-long in-person course) in a bundle for $1200 USD. See our online training site at the link above.

B. Plotting Q-Q plots with nondetects. Why, how, and how NOT to plot them.
Q-Q (or probability) plots illustrate whether data follow a specific distribution. Why should you care? Distributions model how often data of specific values might be expected, and future departures from expectations indicate a change has occurred. Especially with small (<20 observations) data sets, a model is needed to better estimate parameters such as percentiles, the mean and UCL95.

In our upcoming webinar (see section C) I'll describe in detail the process of fitting distributions to data with nondetects. Here I'll demonstrate a visual aid, the Q-Q plot. You should also read about the basics of Q-Q plots for censored data presented in our archived Jan 2013 newsletter.
http://practicalstats.com/news/archive.html

First, the correct way to draw Q-Q plots of data with nondetects -- plot the detected values while adjusting their percentiles for the presence of nondetects. Below is an example for atrazine data censored at one detection limit (0.1 ug/L). Twenty-nine percent of the observations are reported as <0.1. The process for more than one detection limit is similar and described in the textbook *Statistics for Censored Environmental Data Using Minitab and R* (2012).

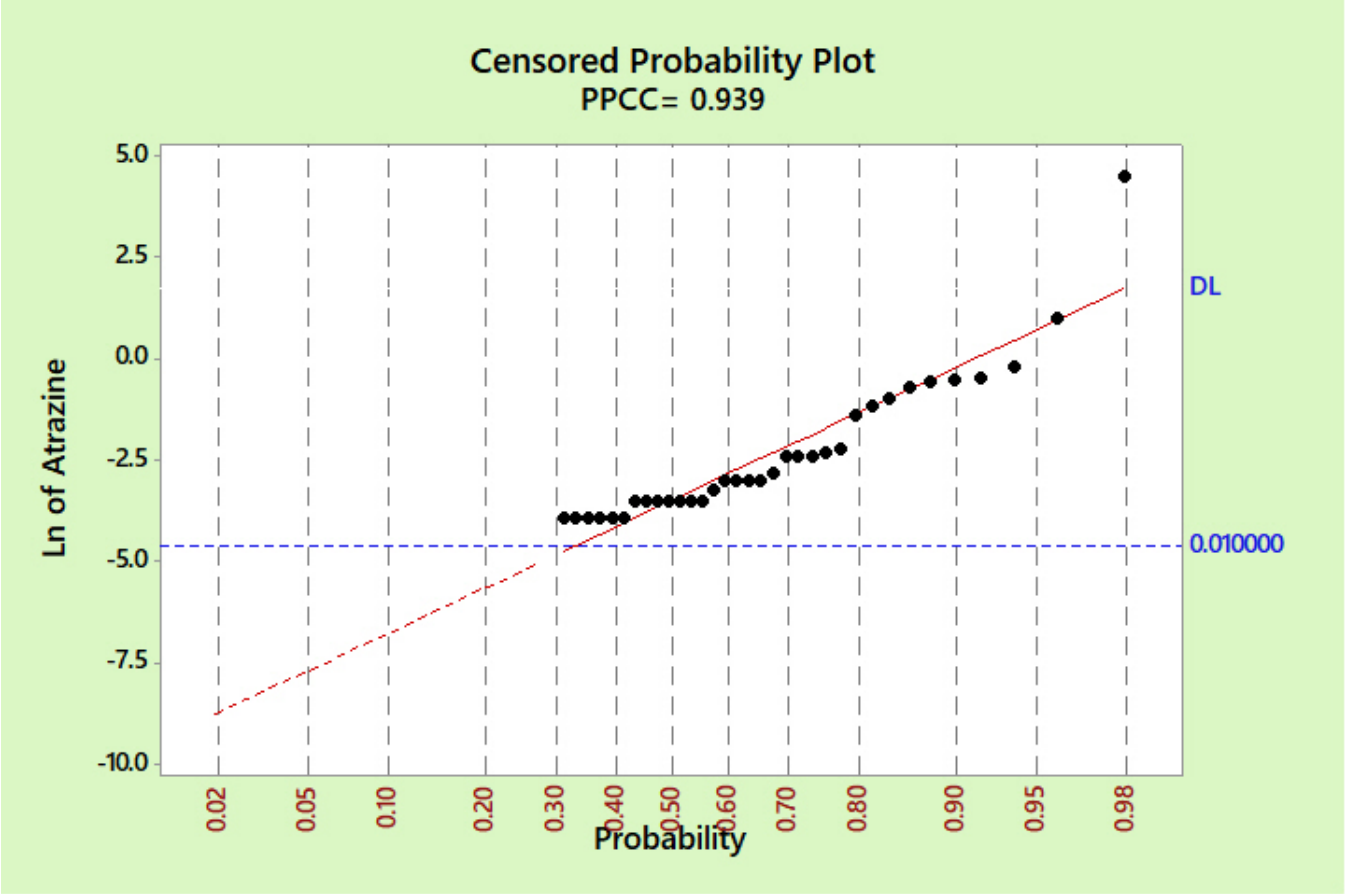

Figure 1. Q-Q plot of atrazine concentrations, compared to a lognormal distribution (shown as a straight line).

Note that no points are plotted until the 0.30 quantile (30th percentile). The probability of being at or below the lowest detected value of 0.02 is 30% -- 0.02 is the 30th percentile, just where it would be if measurable concentrations for the nondetects were available and those values plotted on the plot. However, since we don't have unique single values for the nondetects we have no correct way to plot them, so they remain off the graph. The shape of the data is correct because the points are plotted at their correct probabilities. It appears the data are reasonably fit by a lognormal distribution (PPCC is close to 1 and data are approximately linear), with one outlier at the upper end that should receive more attention. This is the method used in survival analysis routines of statistics software. Estimates of the mean, the UCL95 and percentiles for these data assuming a lognormal distribution should be accurate.

In contrast, here are three ways to NOT correctly draw a Q-Q plot or fit a distribution to these data:

1. Delete nondetects. This distorts the shape of the distribution by miscalculating the probabilities for all percentiles. Decisions about whether these data fit a particular distribution are very likely to be incorrect. A 'detects only' Q-Q plot is shown in Figure 2. Note that the lowest detect of 0.02 ug/L is at a probability of 2 percent, much further to the left and lower than it should be. Also note the data appear curved and the reported P-value is small – after arbitrarily cutting off the lowest 30% of data, the remaining 70% of data are not fit well by a lognormal distribution. That's not a comparison that makes much sense.

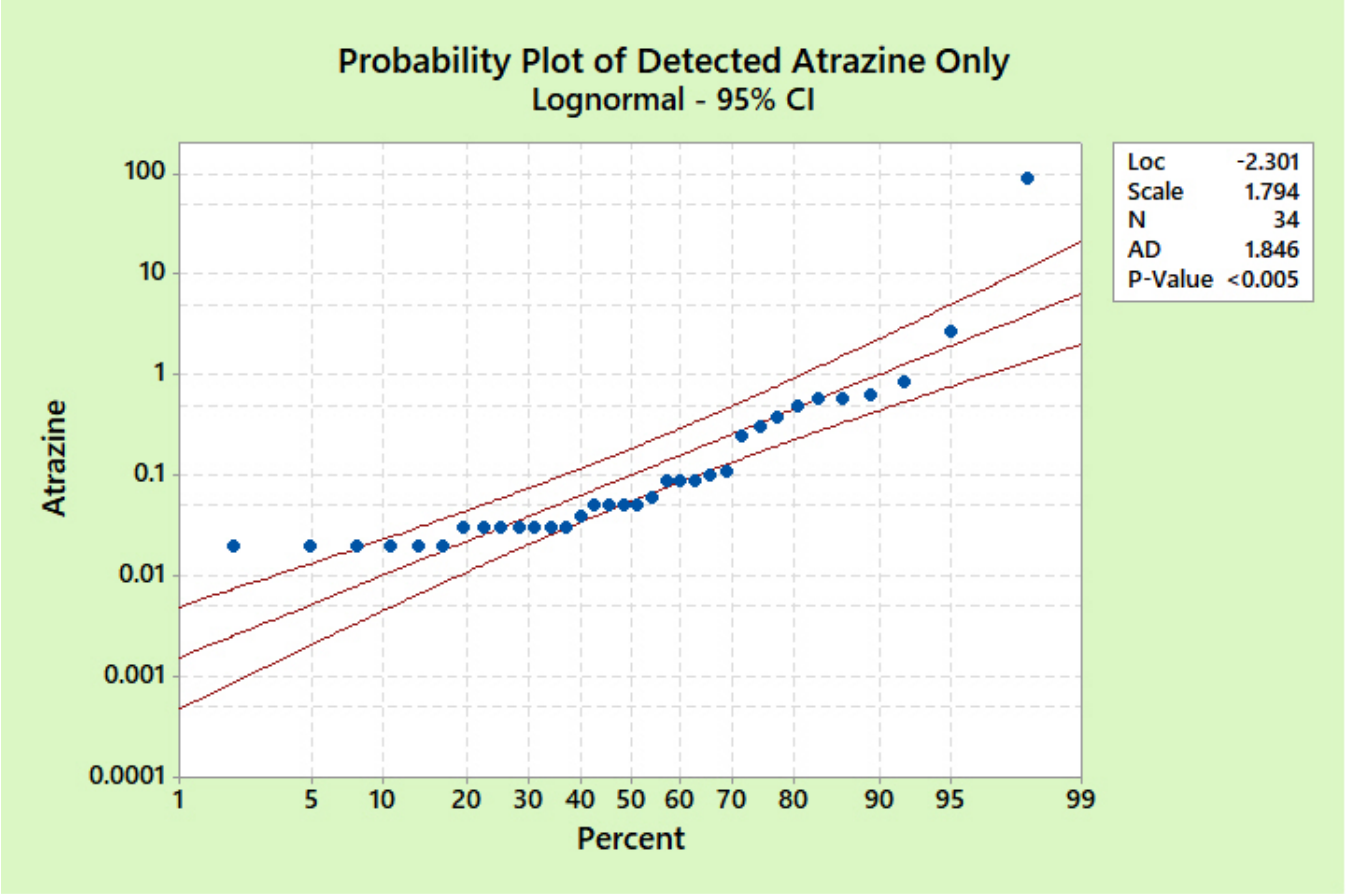

Figure 2. Lognormal Q-Q plot of only detected concentrations. The lower 30% of data have been inappropriately deleted.

2. Substitute the DL (ignore the censoring indicator). This misrepresents the shape of the data distribution for everything below the highest detection limit. On Figure 3 the lowest 30% of the distribution (the <0.01 values) plot as a straight line. Both the visual image and any statistics of the fit, such as the Anderson-Darling (AD) or PPCC statistic – are incorrect.

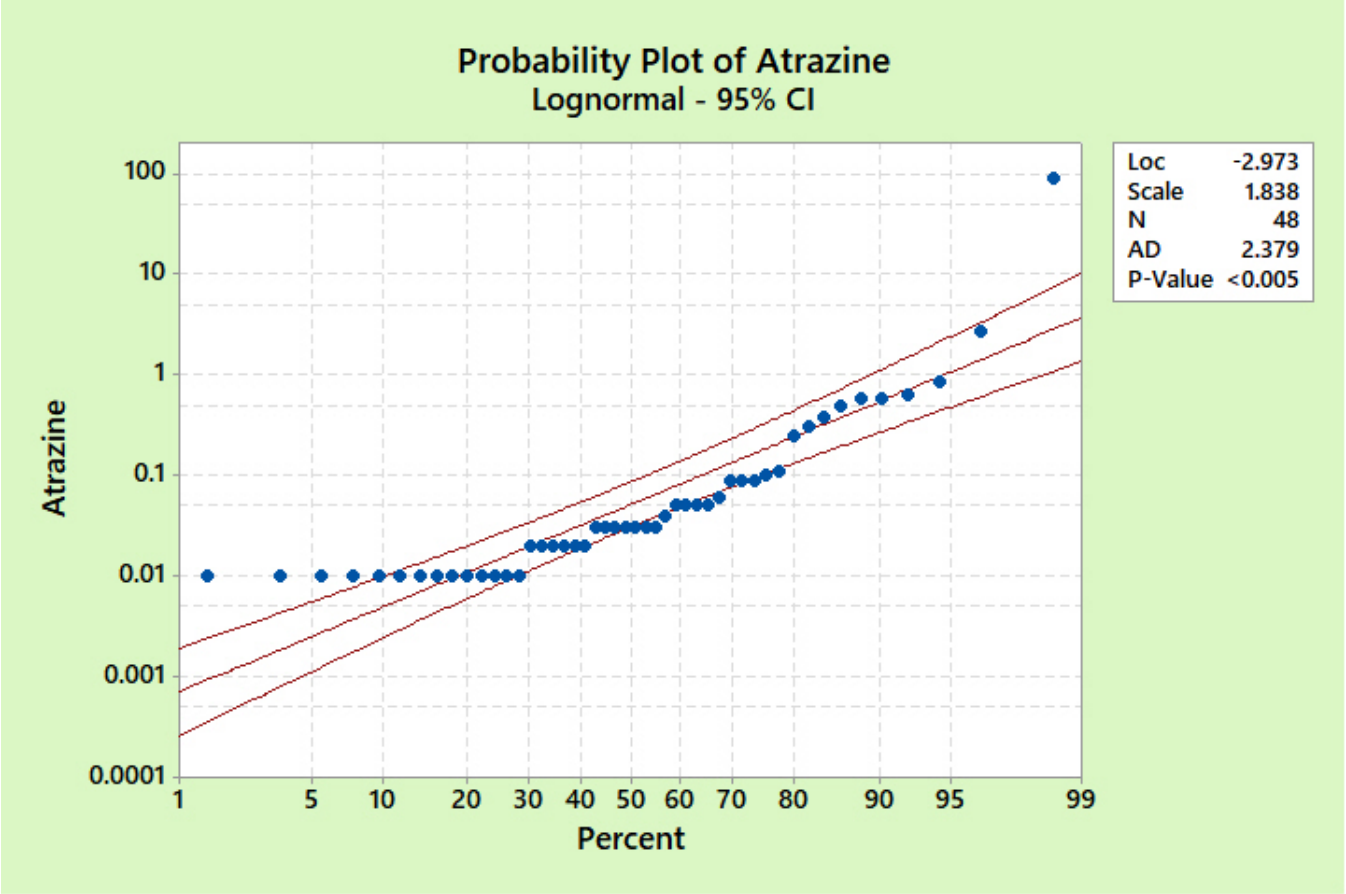

Figure 3. Lognormal Q-Q plot of atrazine concentrations, setting all values below 0.01 equal to the detection limit of 0.01.

3. Multiplying the DL by a fraction (like ½ DL). This has the same problem as #2. The plot will look similar to Figure 3 – the straight line of points representing the lowest 30% of data will just be at a lower concentration, at 0.005 if DL/2 is used. The inaccuracy of shape is the same, resulting in an inaccurate judgment of whether the data fit the distribution, and inaccurate estimates of a mean and UCL95.

Q-Q plots comparing data with NDs to multiple distributions are available in the 'survival analysis' sections of statistics software. Figure 4 shows how well four distributions fit the atrazine data. The lognormal distribution fits better than the other distributions because the lognormal PPCC is highest.

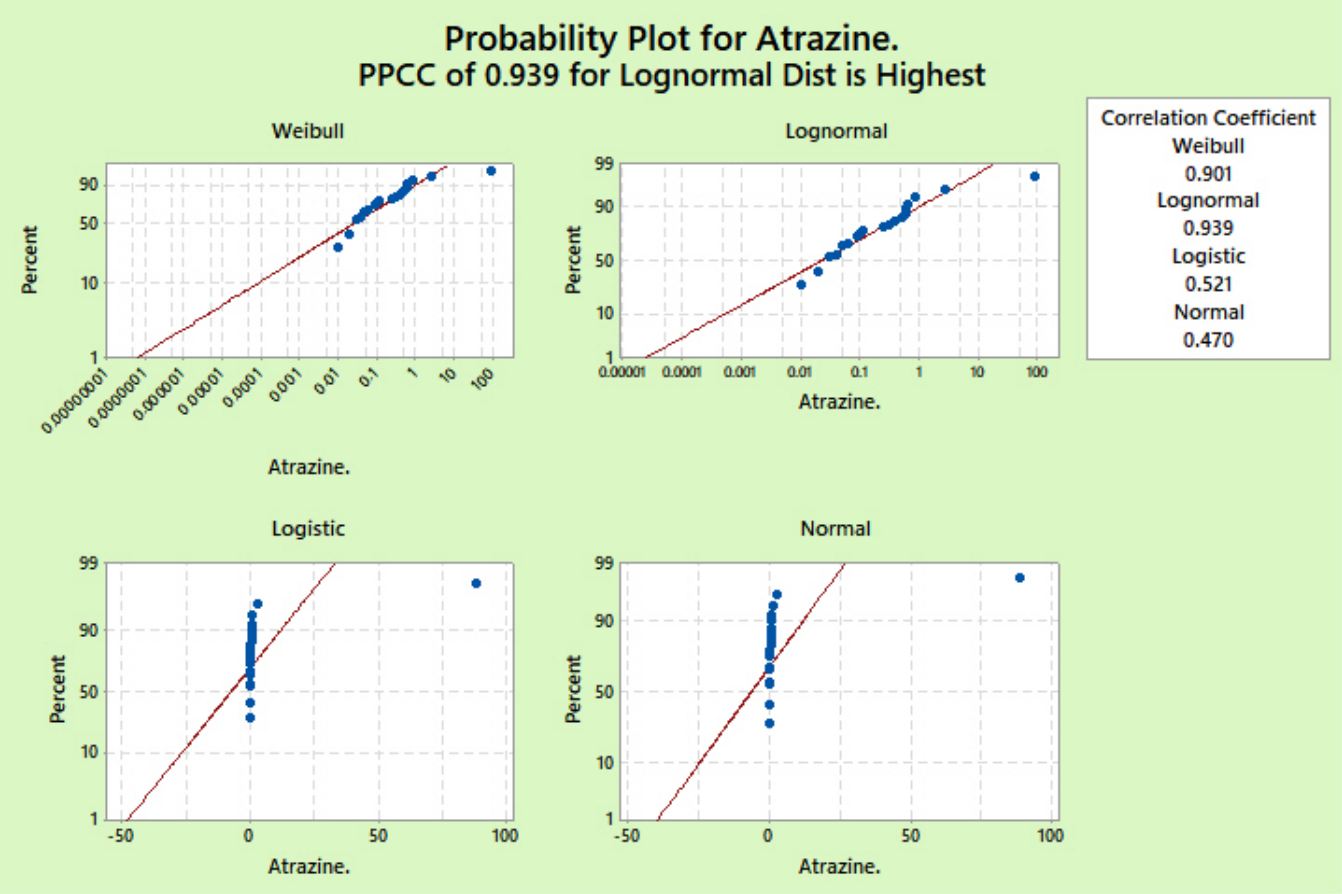

Figure 4. Comparing the fit of four distributions to the atrazine data. Note that the axes are reversed from those of Figures 1 to 3, with the atrazine concentration here plotted on the horizontal axis.

For more information about fitting distributions to data with nondetects, attend our free webinar on February 19th (see below). You might also benefit from reading our October 2018 newsletter, where we tested the fit of several distributions and chose the best one to use in computing a UCL95 for data with NDs.

C. Free Webinars and U-Turns

I've made a U-turn and am working hard to put our Nondetects And Data Analysis course online. The topic has been my primary research field, and I feel an urge to make it available. My apologies to those who have been waiting for other courses, but the NADA class will be the next one online. Right now, its scheduled to be available by the end of March, though it may be sooner. For the next few months I'll be coordinating the newsletter and webinar topics (see below) to go along with the nondetects theme. As the first down payment, for those of you using Minitab for data analysis, version 5.0 of my package of NADA macros for Minitab, adapted for Minitab v.18, is now available for download on our website.

In 2019 I'll be conducting live (and free) webinars approximately monthly. The first is "Fitting Distributions to Data With Nondetects" to broadcast on Tuesday Feb 19th at 1 pm Eastern, 10 am Pacific time. If you're in a different time zone or for whatever reason can't attend live, the recording will be available on demand. Future webinar topics and dates will be announced on our website, and in this

newsletter, which may be coming out more often than every two months (another U-turn). Hope you can join us!

'Til next time,

Dennis Helsel
Practical Stats LLC
-- Make sense of your data