Practical Stats Newsletter for November 2015

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

In this newsletter:
1. 2016 Training
2. How to Compute Stats with Both Greater Thans and Less Thans
3. Updates for Minitab macros

1. 2016 Training

In-person courses:

Permutation Tests
January 11-12, 2016 $995 through Dec. 18, $1095 after.
Golden, Colorado
Permutation test procedures replace parametric tests like t-tests and ANOVA. Learn about these new, important methods for environmental statistics.
http://practicalstats.com/training/

Untangling Multivariate Relationships
January 13-14, 2016 $995 through Dec. 18, $1095 after.
Golden, Colorado
Untangle information in the pattern of chemicals and community structures. Multivariate methods for ecology, hydrology, geology, and other 'ologies.
http://practicalstats.com/training/

Or register for both courses. $1790 through Dec 18, 2015. $1990 after.

Applied Environmental Statistics
St. Paul, MN Feb 1-5, 2016 $950 (it's a bargain! Limited seats available.)
Organized by the Univ. of Minnesota Extension
https://drive.google.com/file/d/0B2ukVHiz4qEKYkN5Y2l3QXNqam8/view?pli=1

Webinars:
Nondetects And Data Analysis
Spring 2016. A series of 4 webinars. Schedule coming soon.

2. How to Compute Percentiles with Both Greater Thans and Less Thans
Censored data are values known only to be either below (nondetects) or above ('too numerous to count') a threshold. A single number is not available. What statistical methods exist for data that contains both types of censoring in the same dataset--sometimes called "doubly censored"?

Coliform bacteria data in surface water can be doubly censored. Values such as <1 and >2400 are both present in this example. The question is "How should percentiles -- median or the 75th percentile -- be computed with such data?" In the survival analysis sections of commercial software, both parametric and nonparametric methods are available to do that. For the parametric method, first find a reasonable fitting distribution to use as the model. The usual parametric method to compute characteristics is Maximum Likelihood Estimation or MLE, but there is also one using regression on probability plots. The nonparametric method is called Turnbull, and is an extension of the Kaplan-Meier procedure. See Helsel (2012) for more detail on these methods, or sign up for our Spring 2016 webinar series. Here we perform MLE using Minitab® as representative of commercial software capabilities. The survival package in R also has these capabilities.

The data were put into the interval-censored format (see Helsel, 2012) -- two columns are used to represent each value, the low end and high end of values bracketing each observation. Nondetects such as <1 have a value of 0 for the low end and the detection limit for the upper end. Greater-thans such as >2400 had a value of 2400 for the low end and '*', the missing value indicator, for the upper end (there is no upper limit estimated). Detected values of 25 had the value of 25 in both columns. Data in this format can be input to either MLE or Turnbull methods.

Data are input in their original units. Using 0 as the low endpoint of nondetects is fine even for distributions like the lognormal that cannot incorporate true zeros – the interval does not include the lower endpoint value, only approaches it. However, the example data had several "detected zeros", where zero was in both the lower and upper columns. The scientist believed they saw no colonies in the dish. Logs cannot be taken of "detected zeros", but can be incorporated into the analysis in at least three ways. The first (and best in my opinion) is to call them <1s. This simply acknowledges that coliforms could be present in the original sample but at a level difficult to observe with the small aliquot tested. The best fitting distribution is then selected as the model. A second method is to only consider distributions that can incorporate zeros. Power transformations such as the cube or fourth root have the benefit that their transform of a zero value is still zero. Invertebrate biologists often use the fourth root (data^1/4), but the root that most closely approximates a normal distribution should be used. This would be an excellent method (assuming the root-transformed data looked approximately like a normal distribution) if current coliform regulations did not specify the use of logarithms. Third, a three-parameter lognormal distribution could be used to model the data. This distribution adds a constant threshold to the data prior to computing logs, so that y=log(x-t), where t is the threshold. If fitted with a statistics package, the threshold is adjusted to best fit the data. Scientists often arbitrarily add 1 prior to taking logs (so t = -1). This may not be the best fitting value, and using different constants will produce different results. Letting the data establish the best fitting constant is a far preferable method to always using t=-1. We'll show the results of methods 1 to 3 below.

Method 1. Zeros as <1s.
Distributions were tested to see which best fit the data shape. Using methods for censored data insure that nondetetcts and greater-thans are included in the fitting procedure. Figure 1 shows the fit of data to four common distributions. The Weibull and lognormal distributions, both skewed distributions, fit best as shown by the straight pattern on their probability plots and lowest Anderson-Darling statistics. Censored values on both ends are not plotted as individual points, but they are used to define the percentile position in the dataset for plotting on the graph. For example, in each plot the lowest dot is at a coliform value of just below 1 near the 10^{th} percentile. This is because 225 of the 1028 values (22%) are nondetects, some of which have detection limits higher than, and some less or equal to, the lowest detected value. We select the lognormal distribution here because it has the best fit (lowest AD statistic), and also because the USEPA recreational bathing beach criteria specify the use of the geometric mean in their regulation. Note that these plots do not substitute a value for nondetects or greater thans. They simply count the proportion of each in determining percentiles for the entire data set, and then plot the detected observations at their computed percentiles.

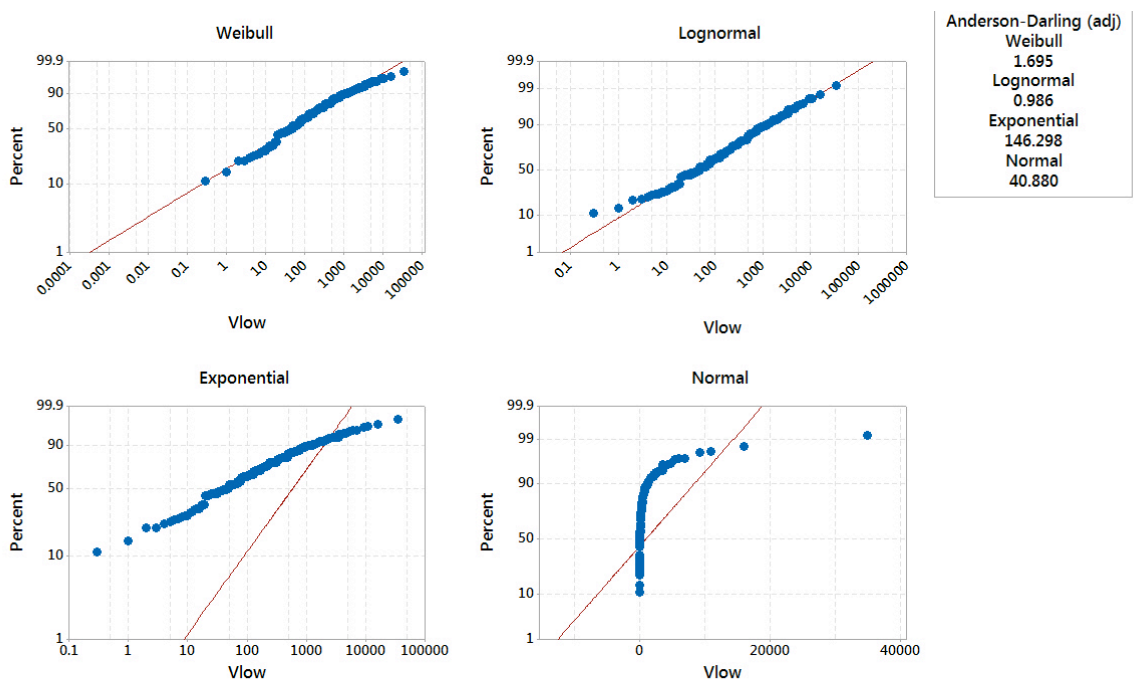

Figure 1. Fit of data to four common distributions

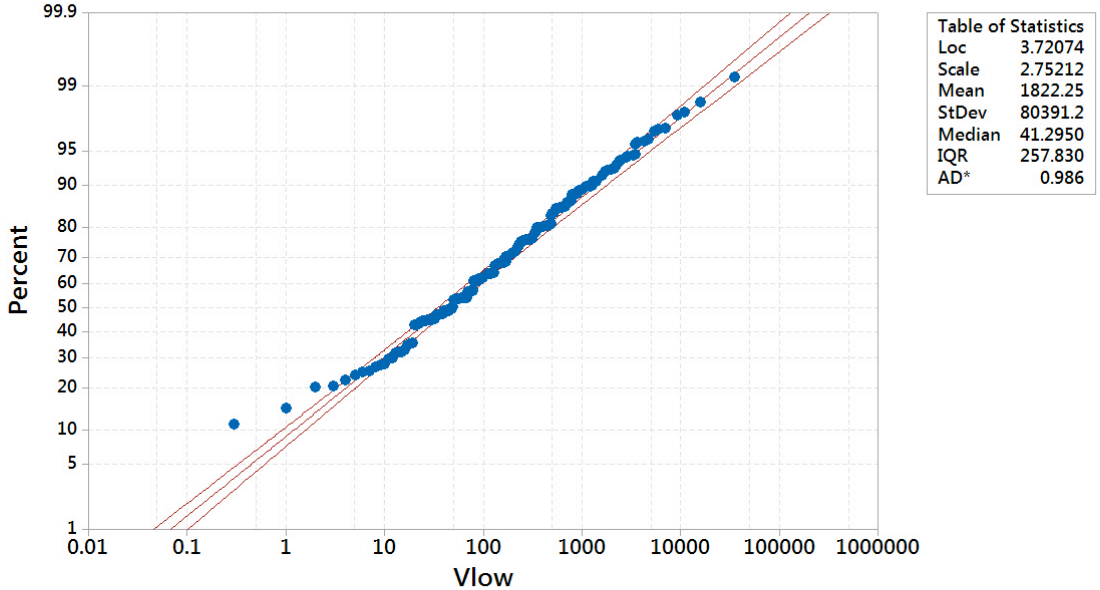

Figure 2. Lognormal Probability Plot for the Coliform Data

The listed median of 41.295 is the geometric mean. It is the mean of the logarithms (3.72) transformed back to original units (here in natural logs), estimating the median of data in original units. The software also provides a large table of percentiles, along with their 95% confidence intervals:

Table of Percentiles

Percent	Percentile	Standard Error	95.0% Normal CI Lower	Upper
1	0.0684499	0.0139093	0.0459627	0.101939
2	0.144941	0.0268636	0.100793	0.208427
5	0.446598	0.0712489	0.326677	0.610541
10	1.21380	0.167606	0.925996	1.59105
20	4.07347	0.468935	3.25069	5.10452
25	6.45247	0.694765	5.22484	7.96853
30	9.75253	0.992775	7.98854	11.9060
40	20.5631	1.92442	17.1170	24.7030
50	41.2950	3.68884	34.6625	49.1965
60	82.9288	7.36965	69.6725	98.7073
70	174.855	16.1969	145.824	209.664
75	264.283	25.4887	218.763	319.273
80	418.629	42.6760	342.812	511.214
90	1404.91	170.405	1107.65	1781.94
98	11765.3	1946.50	8506.99	16271.6
99	24912.7	4552.02	17413.7	35641.2
99.9	203911	47618.6	129021	322269

The nonparametric Turnbull method will similarly produce estimates, without assuming the data follow a lognormal or other distributional shape.

Method 2. Transforming data to normality with a power function such as the fourth-root. Raising data to a power less than 1 will transform right-skewed data to something more like a normal distribution, as does the log transformation. The primary benefit of a power transform is that zero values remain zero -- logarithms cannot be computed for a value of zero. Transforming the low and high columns for censored data by the fourth root, for example, a value of <10 becomes <(10^0.25), or <1.78. A censored normal distribution is then fit to the transformed values. Choose the power that most closely produces values that can be fit well by a normal distribution.

The fourth root is often used by biologists, so we start there. The transformed values are not fit very well by a normal distribution (Figure 3). As the exponent of the transform gets closer to zero, the result is more like a log transformation. Using the tenth root (x^1/10), a normal distribution is more closely approximated (Figure 4). The percentiles of the tenth-root transformed data are computed and re-transformed by raising them to a power of 10, resulting in the estimates in original units, below.

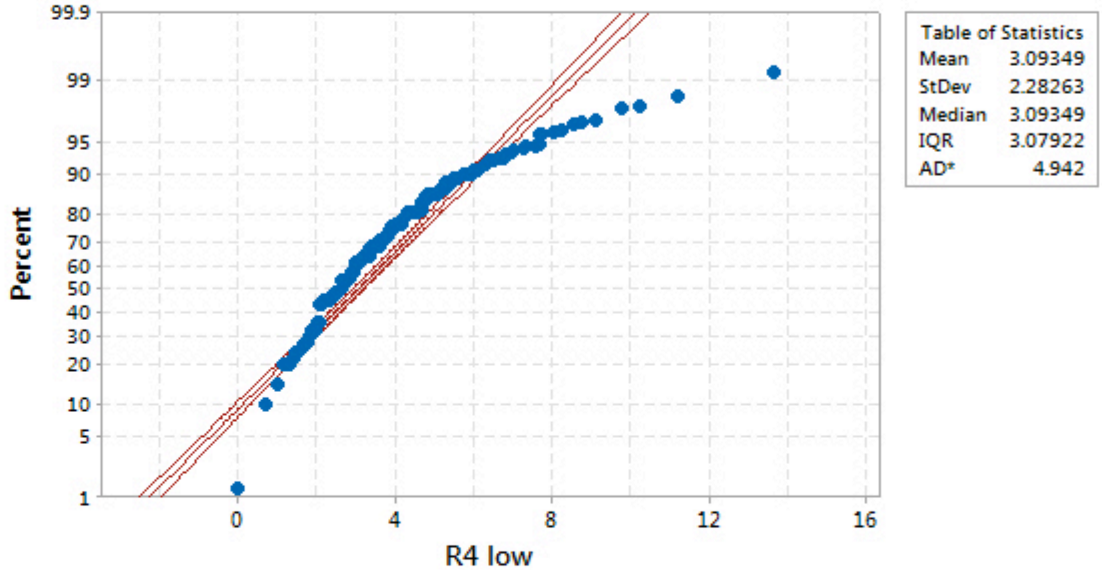

Figure 3. Fit of the fourth-root transform to a normal distribution

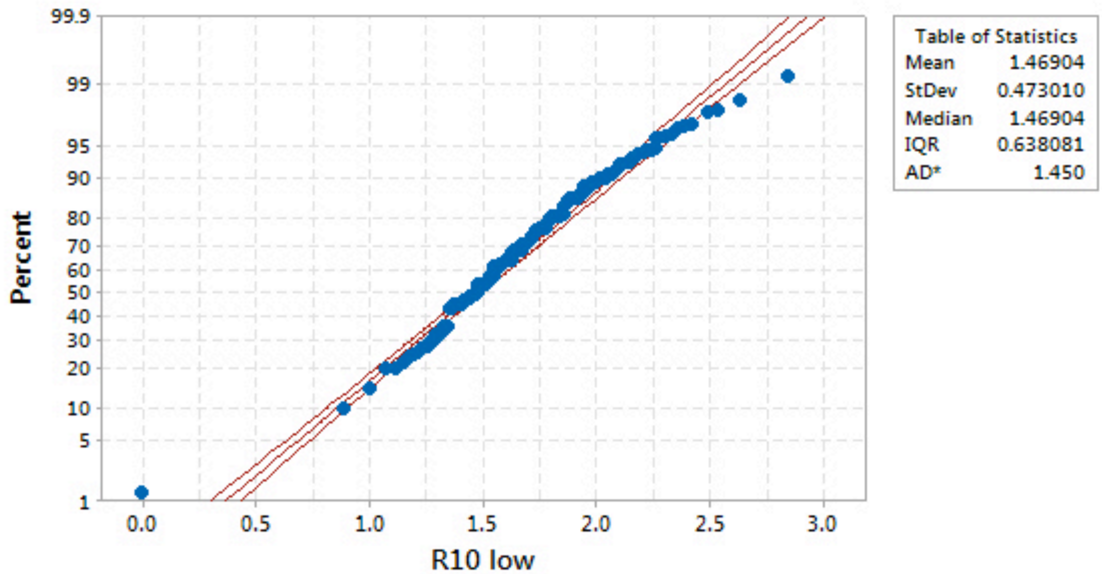

Figure 4. Fit of the tenth-root transform to a normal distribution

Estimated percentiles in original units using a tenth-root transformation

Percent	Percentile	Lower CI	Upper CI
1.0	0.0000	0.0000	0.0002
2.0	0.0009	0.0003	0.0029
5.0	0.0248	0.0114	0.0511
10.0	0.2288	0.1339	0.3803
20.0	1.9847	1.3858	2.8071
25.0	4.0457	2.9574	5.4820
30.0	7.3645	5.5680	9.6669
40.0	19.989	15.820	25.1218
50.0	46.811	38.093	57.2854
60.0	102.544	84.787	123.5798
70.0	222.808	185.477	266.7712
75.0	334.102	278.202	399.9154
80.0	514.952	428.011	617.4756
90.0	1481.37	1217.00	1796.3537
95.0	3282.29	2660.87	4031.4070
98.0	7494.90	5979.35	9347.8186
99.0	12542.0	9900.6	15801.5754
99.9	46751.4	35903.1	60465.5588

The "detected zeros" plot in Figure 4 as an outlier at zero, and may strongly influence the estimate of a mean. The biggest downside of using the power transform procedure is that computation of a mean in the original units is difficult. You would need the equation for that power (here the 10^{th} root) that modifies the re-transformed median to a value for the mean. Or, the 'smearing estimator' could be used (Helsel and Hirsch, 2002). Means do not translate across scales, while percentiles in transformed units can be directly retransformed back to original units.

Method 3. Using a 3-parameter lognormal and leaving zeros as zeros.
As shown below, the 3-parameter lognormal or other skewed distribution fits the data well (small A-D statistic), and gives similar results to Method 1.

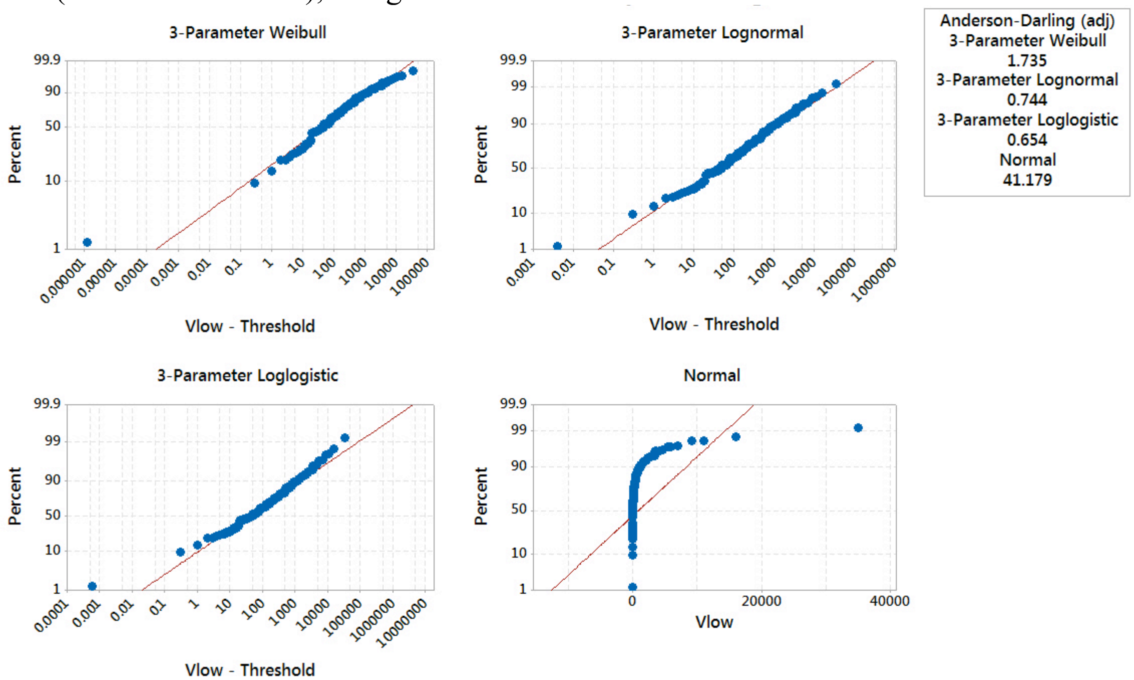

Figure 5. Fit of data to four 3-parameter distributions

The estimates of percentiles with the 3-parameter lognormal are quite similar to Method 1 where zeros were treated as <1s and a standard 2-parameter lognormal was fit to the data. An abbreviated table for the 3-parameter model gives results for several percentiles:

Table of (selective) Percentiles

Percent	Percentile	Standard Error	95.0% Normal CI Lower	Upper
10	0.890139	0.149142	0.597826	1.18245
25	5.28919	0.666681	3.98252	6.59586
50	38.1749	3.70755	30.9082	45.4415
75	275.375	28.5325	219.452	331.298
90	1630.25	222.189	1194.77	2065.73
95	4725.63	766.211	3223.88	6227.38
99	34792.0	7531.71	20030.1	49553.9
99.9	326076	91874.9	146004	506147

The problem with simply adding a 1 to data for the 3-parameter threshold, as is commonly done by some biologists, is that the choice is arbitrarily made. It will likely not fit the data as well as the Method 3 solution, and so produce less accurate estimates of means and percentiles. Let the data select the threshold instead.

Conclusion: the three methods produce estimates of medians that are within the 95% confidence intervals of the other methods. Methods 1 and 3 estimates of medians are geometric means, meeting regulatory guidelines. The power transform (Method 2) estimate of median does not use logarithms, so is not a geometric mean.

Method:	1. 0s as <s	2. Power transform	3. 3-param dist
Estimate of Median	41.2950	46.811	38.1749

Using one of the three methods shown here to compute descriptive statistics for doubly-censored data should be the norm in environmental analysis. Perhaps someday they will. Learn more about methods for censored data in our upcoming 2016 webinars.

3. Updates for Minitab macros

We provide free Minitab macros online for methods in the textbook *Statistics for Censored Environmental Data using Minitab and R* (the NADA macros for censored data), and for routines for general environmental statistics to students who take our Applied Environmental Statistics course. Minitab altered some of its commands in its latest version (17.2), changing how they work from that in 17.1, which disrupted the use of a few of our macros. If you are using our trend analysis macros from the Applied Environmental Statistics course with Minitab 17.2, email us at ask@practicalstats.com (stating the time and place you took our AES class) and we'll send you the macros that produce correct plots with 17.2. The freely-available NADA macros on our Downloads web page
http://practicalstats.com/downloads/
work with both 17.1 and 17.2 of Minitab. We regret the hassles, but changes to commands like this usually occur only in major revisions of software. We updated our macros when version 17 was released. We were surprised when these changes were included in this 'minor' recent 17.2 update of Minitab.

'Til next time,

Practical Stats
-- Make sense of your data