Practical Stats Newsletter for July 2012

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 Events
2. The UCL95 for data with nondetects
3. Ten Years of Practical News

1. Upcoming Events

Webinars: Get the content without the travel! Multiple people can be at one site for one registration fee.

Aug 20: Hypothesis Tests for Data With Nondetects. $250
Sept 17: Regression and Trend Analysis for Data With Nondetects. $250
Oct 15: NEW ProUCL: the Unofficial Users Guide. $100

In-person courses

Applied Environmental Statistics
Statistics, down to earth
December 3-7, 2012. Location to be announced soon.

You can register for courses and webinars through our "Upcoming Classes" page at
http://www.practicalstats.com/new_classes/classes.html

2. The UCL95 for data with nondetects

[This is an update of our October 2006 newsletter.]

The UCL95 is the upper one-sided confidence bound on the mean, and is frequently used as a regulatory limit in the United States. Assuming all of our sampling and measurement methods are appropriate, there is a 50% chance that the sample mean of our measurements is lower, and 50% chance that it is higher, than the true population mean in the aquifer, airshed, or other real world location. There is only a 5% chance that the true population mean exceeds the higher UCL95. The UCL95 is therefore used as an upper limit on where the (unknown) true population mean is located. The UCL95 "covers" (is as high or higher than) the true population mean with 95% probability.

Classical (parametric) computation of the UCL95 uses a t-interval,

$$\text{xbar} + t_{(0.95,\, n-1)} * \text{s/sqrt(n)}$$

where xbar is the sample mean, s is the sample standard deviation, and n is the number of observations. However, this formula assumes either that the data follow a normal distribution, or that there is a lot of observed data. The minimum sample size needed to use this formula with non-normal data increases as the skewness of the data increases.

The sample size needed also is higher when computing a one-sided interval (such as the UCL95) rather than the two-sided intervals described in most statistics courses. For small environmental datasets that are skewed, the formula does not provide adequate results.

For censored data (data with nondetects), users often substitute one-half, or one over the square root of two, times the reporting limit for nondetects, and use the standard t formula as if these substitutions were real observations. The distance that the UCL95 is above the sample mean is driven by the standard deviation (s). When a constant value is substituted for all nondetects, estimates for the standard deviation are notoriously poor. The standard deviation is usually under- (though occasionally over-) estimated, and therefore so is the UCL95.

Two guidance documents linked to Federal agencies have abandoned the use of substitution in the regulatory process when computing the UCL95. "Statistical Methods and Software for the Analysis of Occupational Exposure Data with Non-Detectable Values" by Frome and Wambach is Oak Ridge National Laboratory's 2005 report ORNL/TM-2005/52. Regulations of compounds toxic to humans assume as standard practice that exposures follow a lognormal distribution. The UCL95 is used as the upper bound on what value the mean exposure might be, comparing its value to legal standards to determine compliance. At issue in this report is what to do with nondetect measurements. They recommend computing the mean using maximum likelihood, not substitution, when data appear lognormal. The confidence bound is then computed using Cox's method, described in *Statistics for Censored Environmental Data* (Helsel, 2012). If data do not appear to follow a lognormal distribution, Frome and Wambach recommend using the Kaplan-Meier (K-M) method, also called the Product-Limit Estimator. K-M is the standard method for computing statistics with censored data in medical statistics, and is described fully in Helsel's 2012 textbook. The important thing to note here is that a major environmental regulatory manual has chosen NOT to recommend substitution as a viable procedure for incorporating nondetects.

Singh and others (2006) evaluated many methods for computing the UCL95 with censored data in a report for USEPA. At issue was coverage -- does the computed UCL95 "cover" (is it equal to or greater than) the population mean with 95% probability? They found that substituting half the detection limit (DL/2) and using the t-interval formula did not cover the true mean 95% of the time. One quote that stands out to us: "The DL/2 (t) UCL method does not provide adequate coverage (for any distribution and sample size) for the population mean, even for censoring levels as low as 10%, 15%. This is contrary to the conjecture and assertion (e.g. EPA (2000)) often made that the DL/2 method can be used for lower (<20%) censoring levels." Instead of substitution, Singh et al. found that nonparametric Kaplan-Meier (K-M) methods consistently produced the best estimates of the UCL95. Maximum likelihood methods did not provide good coverage for smaller sample sizes or for highly skewed data (so K-M would be better than the lognormal MLE recommendation of Frome and Wambach in these instances). Probability plot (robust ROS) methods did not work as well as K-M - though there was a programming error in the Singh implementation of ROS that may have degraded its

results. The authors tested several ways to compute the confidence bound around the K-M estimate of mean, and found four to work well: percentile bootstrap, bias-corrected percentile bootstrap, the t formula (using K-M estimates of mean and standard deviation), and the Chebyshev formula. The best performance among these four changed with data characteristics -- read their report to fine-tune when to use each of them. A free Minitab macro to compute the percentile bootstrap Kaplan-Meier procedure, the BootKM macro, is available in the NADA for Mtb collection available on the Practical Stats web site. All four variations of K-M intervals are found in ProUCL v.4 (website listed below).

One of the data sets used in the textbook *Statistics for Censored Environmental Data* (Helsel, 2012) is the Pyrene data set. Taken from a journal article measuring pyrene in sediments, of the 56 observations there are 11 nondetects at 8 different reporting limits. The data are represented by two columns in the 'indicator column' format: one column contains detected concentrations and reporting limit values. The second column indicates which is censored, using 0 and 1. Output from ProUCL provides the following results:

```
Mean of Detected Data:  190.1
SD of Detected Data: 435
Mean of Detected (logarithms): 4.711
SD of Detected (logarithms): 0.805
DL/2 Substitution Mean: 163
DL/2 Substitution SD:  393.2
95% DL/2 t UCL:  250.9
MLE normal Mean:  (N/A, negative)
Log ROS Mean:  163.2
Log ROS SD:  393.1
Log ROS UCL95 (t):  251.1
.....and many more.
```

```
Log ROS pctl bootstrap UCL95:  266.4
Log ROS BCA Bootstrap UCL95:  332.4
Log ROS 95% H-UCL:  170.4
Gamma Mean:  161.7
Gamma SD:  394.4
95% Gamma Approximate UCL:  257.4
KM Mean:  164.2
KM SD:  389.4
95% KM t UCL:  252.3
95% KM pctile bootstrap UCL:  264.9
95% KM Chebyshev UCL:  393.7
```

Two things to take away from the ProUCL output.
1) there are many ways to compute a UCL for data with nondetects!
2) before deciding which to use, you need to know what these methods are doing, and which provide good answers over a variety of situations.

In general, the KM method with either the t UCL or bootstrap UCL provide realistic answers across a variety of shapes of data (lognormal, etc.) without substituting arbitrary values like DL/2 for nondetects. In the above example, these are the bounds at 252.3 and 264.9.

More detail on how to use the methods in ProUCL, and which not to use, will be given in our October 15th webinar, ***ProUCL: The Unofficial Users Guide.*** See the Upcoming Courses page on our website for more information and to register.

Reference Links for computing the UCL95 for data with nondetects

Helsel (2005), More Than Obvious (why not to use substitution)
http://pubs.acs.org/subscribe/journals/esthag-a/39/i20/toc/toc_i20.html
Abstract is free. For pdf, email us at ask@practicalstats.com

Frome and Wambach (2005), Statistical Methods and Software for the Analysis of Occupational Exposure Data with Non-Detectable Values. Oak Ridge National Laboratory.
http://www.ornl.gov/~webworks/cppr/y2005/rpt/124028.pdf
Singh et al (2006), On the Computation of a 95% Upper Confidence Limit of the Unknown Population Mean Based Upon Data Sets with Below Detection Limit Observations.
http://www.epa.gov/esd/tsc/issue.htm
ProUCL version 4, available at:
http://www.epa.gov/esd/tsc/software.htm
Statistics for Censored Environmental Data textbook, Minitab macros and R routines:
http://www.practicalstats.com/nada/

3. Ten Years of Practical News

Our newsletters began ten years ago this summer, released on a quarterly basis. We've released six per year since 2009. This issue updates a topic first addressed in our October 2006 newsletter, and many of the older newsletters are still timely and of interest today. A number of the older issues address methods for censored data (data with nondetects). Others address equivalence tests, capabilities of statistical software (prices listed are outdated), and online sources of statistical information. We'll be updating the latter two topics in the coming months. Celebrate the anniversary of our newsletters --

1. sign up to receive them directly if you're not already doing so.
2. suggest new topics you'd like to see addressed (email ask@practicalstats.com with your suggestions).
3. download and read the issues relevant to your work.

The archive is at http://practicalstats..com/news/

'Til next time,

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