

Practical Stats Newsletter for October 2006

In this newsletter:

1. Newsletter Archive now on the Practical Stats webpage
2. Computing the UCL95 for data with nondetects
3. Reference Links for computing the UCL95

1. Newsletter Archive now on the Practical Stats webpage

An archive of past Practical Stats newsletters is now available on our web site. Just click on the Newsletter Archive link in the left-hand column. Newsletters will be posted several months after they are sent to subscribers, so if you'd like to receive them in a more timely manner and have not yet done so, subscribe to receive our newsletters by email (sign up at www.practicalstats.com). If your email address changes, unsubscribe using the old address, and subscribe using the new one in order to stay current. In the next few newsletters we will be announcing a new course available in 2007 as well as more in our current "Statistics in Environmental Regulations" theme. Pass this newsletter along to others who may be interested. Remember, when you sign up, your email address goes nowhere else.

2. Computing the UCL95 for data with nondetects

Perhaps the most commonly-used statistic in environmental regulation is the UCL95, the upper one-sided 95% confidence limit (or bound) on the mean. It is a statistic derived from our observations. Assuming all of our sampling and measurement methods are appropriate, there is a 50% chance that the sample mean of our observations is lower than the true population mean. There is only a 5% chance that the UCL95 is lower than the true population mean. So 95% of the time, the UCL95 "covers" (is as high or higher than) the true population mean. The UCL95 is used as an upper limit on where the (unknown) true population mean is located.

Classical (parametric) computation of the UCL95 is a t-interval,
$$\bar{x} + t(0.95, n-1) * s/\sqrt{n}$$

where \bar{x} 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 we have a lot of 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 introduced in most statistics courses.

Statisticians call the argument for using this formula with non-normal data the Central Limit Theorem (CLT). In our Applied Environmental Statistics course we fully discuss when the CLT applies, and when it doesn't, for computing intervals such as the UCL95.

For data with nondetects, typical practice is to substitute one-half, or one over the square root of two, times the reporting limit for nondetects, and use the standard formula as if these estimates were real observations. Here at Practical Stats we have declared for years that this is a terrible procedure. Why it is so terrible is discussed in the textbook "Nondetects And Data Analysis" (NADA), as well as in the 2005 article "More Than Obvious" in Environmental Science and Technology (Helsel, 2005). There's a link to both of these below. If you would like to receive a pdf of the ES&T article, you can email Dennis Helsel at dhelsel[at]PracticalStats.com and he'll send you a copy. Change the [at] to @, of course.

Two recent reports have abandoned using 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 (also described in the NADA textbook). Note that they do not use Land's estimate, which has justifiably come under fire in recent evaluations of the UCL95. 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 the NADA 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.

The second report solidifies the reasons why. In a March 2006 report, Singh and others (2006) evaluated a large number of methods for computing the UCL95 with censored data. The study was commissioned by USEPA, and the methods found best will be incorporated into ProUCL, software for computing the UCL95 developed by Singh and others for the Superfund program. At issue is 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 above t-interval formula produced terrible results. 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."

The implications of their findings certainly extend to t-tests as well, as hypothesis tests are just the 'inverse' of confidence intervals. Instead of substitution, Singh et al. found that the 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, based on their findings). Probability plot (robust ROS) methods did not work as well as K-M - though still much better than substitution. 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. You can compute the UCL95 for nondetect data using the percentile bootstrap Kaplan-Meier method with the KMBMean macro, part of the free NADA macro collection for Minitab available on the Practical Stats web site since 2004. It is available there now. We fully cover its use in our NADA course (see web site for the next offering). You will be able to compute the UCL95 using all four variations of K-M with version 4.0 of ProUCL. That version will be available from a USEPA site in early 2007 (see below). It would also not surprise me if the other 3 recommended methods show up in the Minitab macro collection for NADA at some time in the future.

In another recent article in the SETAC (Society for Environmental Toxicology and Chemistry) journal, Sinha et al. (2006) found that the robust ROS method (they used the older name "Log probability regression") worked better than other methods they tested for computing the UCL95. They too evaluated coverage, and recommended the Chebyshev formula around the robust ROS estimate of the mean. They found it to be better than Aitchison's method, for example, one of the methods currently recommended in several environmental guidance documents. Helsel (2005) noted that Aitchison's method "is just substitution", and worked no better than substitution in simulation studies. Unfortunately, Sinha et al. did not include K-M in the methods they evaluated. If they had, they would no doubt have agreed with Singh and others (2006) findings that Kaplan-Meier provided better coverage than robust ROS, as stated above.

So, the bottom line: for computing a UCL95 of data with nondetects, avoid substituting arbitrary values. It simply doesn't work well. There are also better methods than Aitchison's (also called the D-LOG) method. MLE does not work well for skewed data and smaller sample sizes. Fortunately Kaplan-Meier, the standard method in medical statistics presented in the NADA textbook and taught in our NADA training course, works very well in a wide variety of situations. K-M gives estimates of the mean and quantiles, and can be used to estimate the UCL95 in one of four ways. All four of these works far better than substitution, and appear to work as well or better than robust ROS, the method we have provided software for over a number of years through the Practical Stats web site. To determine which of the four to use in given situations, read the Singh et al. (2006) report [link below]. The KMBMean macro from our package of NADA for Minitab macros computes the percentile bootstrap method. In early 2007 the free software package ProUCL version 4 (which incorporates K-M methods) will perform all 4 ways of computing the UCL95 with nondetect values.

There is ample evidence, and now/soon available free software, for regulatory agencies to finally abandon substitution as a recommended method for handling nondetects. Lets hope that settles this issue.

3. Reference Links for computing the UCL95

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 Dennis Helsel (see above).

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>
(link to the PDF file is half way down the page)

Sinha, Lambert and Trumbull (2006). Environmental Toxicology and Chemistry 25, p. 2533-2540.
<http://entc.allenpress.com/pdfserv/10.1897%2F05-548R.1>
Abstract is free. pdf requires subscription or payment.

ProUCL software (version 4 will include the new procedures for nondetects. Available early 2007):
<http://www.epa.gov/esd/tsc/software.htm>

Nondetects and Data Analysis textbook and Minitab macros
<http://www.practicalstats.com/nada>

'Til next time,

Practical Stats
<http://www.practicalstats.com>

-- Make sense of your data