Practical Stats Newsletter for January 2017

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

In this newsletter:
A. Practical Stats Courses
B. Statistics for "Small Data", Part 2. The UCL95
C. New Online Webinar

A. Practical Stats Courses

Our Applied Environmental Statistics online course is making its way to completion. There are 9 sections and I've finished 5. When all is ready, and I truly think that will be prior to the next newsletter in March, you'll find the course on our online training site: http://practicalstats.teachable.com/
I'll send out an announcement to this newsletter list when the course is ready for registration.

We also offer in-person training for groups you pull together. See http://practicalstats.com/training/ for details.

B. Statistics for "Small Data" Part 2. The UCL95

In October we looked at using exact tests rather than the more common "large-sample approximation" tests to wring the most out of hypothesis tests with small numbers of data. We also strongly recommended that with small data you never decide on which class of tests to use, parametric or nonparametric, based on a prior hypothesis test for normality. If you don't remember why, or didn't read our previous newsletter, go to the Newsletter Archive at http://www.practicalstats.com/news/bydate.html and take a look.

This month we answer the question "How do I compute a UCL95 with small data sets?" The UCL95 is a protective estimate of the mean of a data distribution. It is a value sufficiently high so that the true population mean, the mean out in the field (stream, aquifer, etc.) has only a 5% probability of being higher than it. The mean of a group of new observations (and not individual observations) should be compared to the UCL95 – see our February 2012 newsletter for more in-depth information, and computation for data with nondetects. With larger datasets (20 observations or more), bootstrapping will provide a good estimate of the mean regardless of shape of the data distribution. But what if you have fewer than 20 observations? Bootstrapping re-uses only the measured observations without adding a theoretical distribution to represent the shape of the data. With fewer than 20 observations you won't have a good representation of the parts or proportions of the distribution you haven't yet measured. Adding a 'model' by making a reasonable assumption about the shape of the distribution should improve estimates over those possible from only a handful of observations.

For 8 to 20 observations

Environmental data are often skewed due to the zero lower bound of data values. Unusual observations are therefore found on the high side, not often on the low side. Skewed distributions such as the lognormal, gamma, and Weibull are often useful to model the data shape. The decision of which distribution to use is best decided by using the shape of larger data sets for the same parameter, collected under similar conditions. As a second-best method, one could use a hypothesis test such as the Probability Plot Correlation Coefficient (PPCC) or Shapiro-Wilk tests to determine which distribution appears to best fit a small set (8 to 20) of observed data. The lower bound of 8 observations is based on simulations determining how well these tests can choose the distribution to model data shapes. Below eight, there's insufficient data to even guess what the distributional shape may be. It is also true that below eight observations, any method for computing the UCL95 is based on sufficiently few observations that the estimate is likely to be pretty far off the mark.

Below I'll use the Shapiro-Wilk test for guessing the best-fitting distribution, using the set of 12 downgradient concentrations from the remediation example in October's newsletter. The process:

1. Decide which distributions to consider. Here I'll use the normal, lognormal, and gamma distributions.
2. Test the goodness of fit for each of the distributions. The distribution with the highest (closest to 1) PPCC or Shapiro-Wilk W is the one that best fits the data. In R, using the EnvStats package (a very useful package, I highly recommend it) the commands are:

```
> downgradient <- as.numeric(c(0.390, 0.320, 0.300, 0.305, 0.205, 0.200,
0.195, 0.140, 0.145, 0.090, 0.046, 0.035))
> gofTest(downgradient, dist="norm")
> gofTest(downgradient, dist="lnorm")
> gofTest(downgradient, dist="gamma")
```

For the downgradient data, the Shapiro-Wilk W values are:

Normal	0.954
Lognormal	0.896
Gamma	0.934

The normal and gamma distributions fit best, because their statistics are closer to 1.0. The normal distribution has a secondary concern -- it may allow the lower end the distribution to go negative, an unrealistic situation for environmental variables that produces an estimate of the mean, and perhaps the UCL, which is too low. To test this, use the estimates of mean and standard deviation for the assumed normal distribution to compute whether 3 standard deviations below the mean is a negative number. If so, the lower end is indeed unrealistic. For the fitted normal to these data:

```
Estimated Parameter(s):          mean = 0.1975833
                                 sd   = 0.1132860
```

and 0.197-3*0.113 = –0.142. The low end of the distribution is indeed described as below zero. In this situation, choose instead the next best fitting distribution, here the gamma distribution.

```
> egammaAlt(downgradient, ci=TRUE,ci.type="upper")

Results of Distribution Parameter Estimation
--------------------------------------------
Assumed Distribution:            Gamma
Estimated Parameter(s):          mean = 0.1975833
                                 cv   = 0.6349288
Estimation Method:               MLE
Data:                            downgradient
Sample Size:                     12
Confidence Interval for:         mean
Confidence Interval Method:      Optimum Power Normal Approximation
                                 of Kulkarni & Powar (2010)
                                 using mle of 'shape'
Normal Transform Power:          0.246
Confidence Interval Type:        upper
Confidence Level:                95%
Confidence Interval:             LCL = 0.0000000
                                 UCL = 0.2822187
```

and the computed one-sided UCL95 equals 0.282.

For fewer than 8 observations
When there are fewer than 8 observations, whatever method or distribution you choose for computing the UCL95 is likely inaccurate. If you do know the shape of larger data sets for the same parameter under similar conditions, using that shape with the methods above for 8-20 observations is the best of the bad available options – collecting more data is the primary better option! Note that the computed UCL95 may be higher than all the observations currently measured, and for such small datasets this is a reasonable result.

Some guidance documents recommend using the maximum of the currently available data as the estimate of the UCL95 for small datasets. Is the maximum of <8 observations a helpful estimate of the UCL95? A sample mean concentration for skewed data is often around the 60th to 80th percentile, and its UCL95 will be higher. The maximum of 5 observations has a 33 percent chance of being below the 80th percentile, based on a binomial computation. Therefore the maximum of skewed datasets smaller than 8 observations is often too low to be equivalent to the UCL95, and more likely to be closer to the population mean. If you use the current maximum value, know that you are likely underestimating the UCL95 and so underestimating the risk that the true mean in the field is higher than your chosen criterion. There is likely a greater than 5% chance that the population mean exceeds your current maximum observation.

C. New Online Webinar

Our recent webinar "Introduction to Nondetects And Data Analysis" presented through the National Water Quality Monitoring Council is freely available for viewing on our training site -- http://practicalstats.teachable.com/
It covers

* the consequences of substituting one-half the reporting limit on data analysis (spoiler alert: its not good).
* what software is available for analysis of censored data.
* what nonparametric methods are available for estimating summary statistics, test for differences between groups, and for correlation/regression of censored data.
* what parametric methods are available for estimating summary statistics, test for differences between groups, and for correlation/regression of censored data.

Materials (pdf of the slides, a listing of Practical Stats newsletters on dealing with nondetects, and more) are also available online at http://www.practicalstats.com/training/

'Til next time,

Practical Stats
-- Make sense of your data