Practical Stats Newsletter for October 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 and Webinars
B. Bootstrapping
C. View Webinars Anytime, for Free

A. Practical Stats Courses and Webinars
On our online training site: http://practicalstats.teachable.com/

Our Nondetects And Data Analysis (NADA) course is available online. One year's access to the materials costs $695. It covers estimation of summary stats, regression, group testing, and even some multivariate methods, all without substituting fabricated numbers like ½ the detection limit. The R scripts included provide new functionality to make data analysis easier, and are a step forward from the NADA package in R.

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.

Our next webinar is
Never Worry About A Normal Distribution Again
on Oct 22, 2019 11:00 AM MDT
Permutation tests and bootstrap intervals avoid a normality assumption, returning accurate p-values and interval widths while being distribution-free. This webinar will describe how these methods work, where you can find them, and demonstrate their benefits over the older traditional methods such as t-tests and t-intervals.

Register at:
https://attendee.gotowebinar.com/register/7079679905772943628

B. Bootstrapping
Bootstrapping has a simple premise: the shape of the distribution of the data you've collected is the best estimator of the shape of the distribution of data out in the field from which you sampled. It avoids assuming a normal or other theoretical shape to compute measures of variability. Developed by Bradley Efron in 1977, the bootstrap has been used in many scientific fields to solve thorny issues that are hard to do mathematically. Indeed, Efron was awarded the International Prize in Statistics (the "Nobel Prize" for statistics) in 2018. Over 14,000 journal articles in Environmental Statistics, 9700 in Earth Sciences and over 44,000 articles in Agricultural and Biological Sciences have used the bootstrap (American Statistical Association, 2018). It is regularly used to estimate confidence intervals on the mean

(including data with nondetects) and weighted means, tolerance intervals, intervals around regression models, and much more. Here are three examples.

1. For molybdenum concentrations in a shallow aquifer, first a Shapiro-Wilk test shows that the data appear to be non-normal is shape:

```
> shapiro.test(Moly21$MOLY)

       Shapiro-Wilk normality test
data:  MOLY
W = 0.54395, p-value = 0.000005014
```

The 95% confidence interval on the mean, which when a t-interval is used assumes normality, is:

```
t.test(MOLY, alternative ='two.sided', conf.level=.95)

       One Sample t-test
data:  MOLY
t = 2.099, df = 15, p-value = 0.05315
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.0144863  1.8914863
sample estimates:
mean of x
   0.9385
```

The 95% confidence interval is (-0.014 to 1.89). The negative lower end is a good indication that the data don't follow a normal distribution and so another method must be used. Making no assumptions about the population distribution of MOLY, the bootstrap confidence interval is:

```
> Bootmean(MOLY, conf = 95, R = 10000)

Bootstrap Confidence Intervals of the Mean of MOLY
Using boot in R

BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 10000 bootstrap replicates
CALL :
boot.ci(boot.out = B, conf = Conf, type = TYPE)

Intervals :
Level     Percentile
95%   ( 0.2575,  1.9244 )
Calculations and Intervals on Original Scale
```

The 95% confidence interval on the mean is 0.2575 to 1.9244. The lower end is not negative, and is 0.27 ug/L higher than the t-interval's limit. The difference is shown in Figure 1 by the dashed red (t-interval) and blue (smoothed data) curves. Five percent of the data are below 0.257 but not below 0 (the blue smooth). The normal distribution used by the t-interval goes below zero, misfitting the data. This is fairly common with skewed data as the upper end extends out further from the mean than does the lower end, but the t-interval is symmetric by design around the mean. The t-interval's lower end is strongly affected by its larger upper end to provide symmetry, resulting in lower ends of t-intervals that are too low.

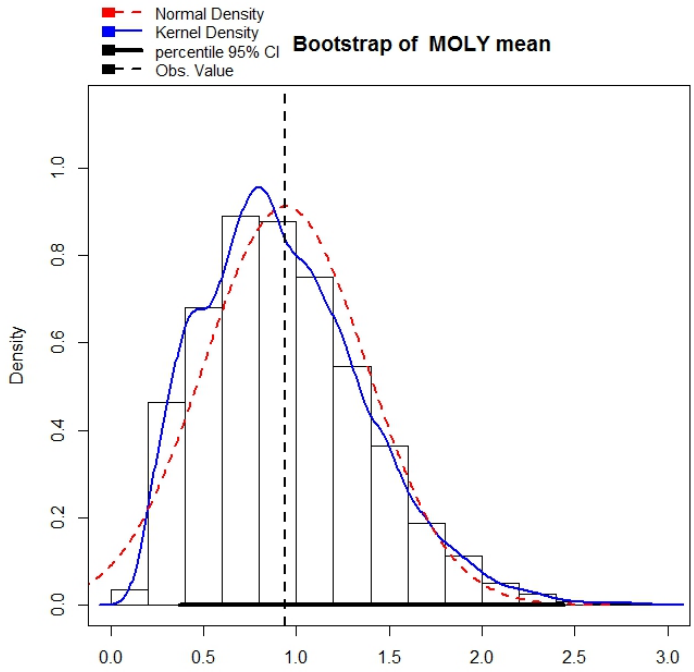

Figure 1. Data histogram with smooth (solid) and fitted normal distribution (dashed line)

2. A second use of bootstrapping includes computing confidence intervals around the Theil-Sen line, the line often used for trend analysis. Computing a distribution-free interval around the location of the line, considering variation in both slope and intercept, is difficult to do from an equation. It's not difficult with bootstrapping. The data and R code that produced Figure 2 are from the Unified Guidance (USEPA, 2009) authored by Kirk Cameron.

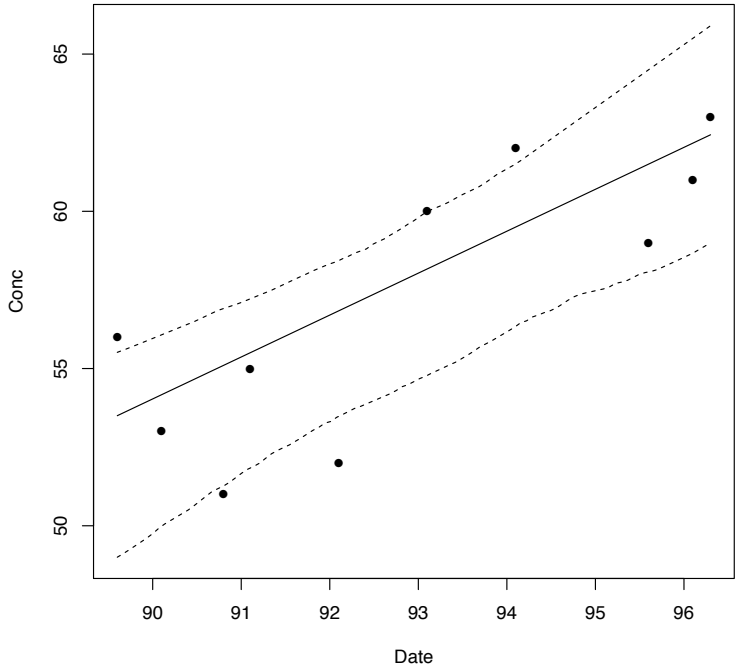

Figure 2. Theil-Sen trend line with bootstrap 90% confidence intervals

3. Using the EnvStats package of R, written by Steve Millard, a bootstrapped 95% upper confidence limit (UCL95) around the nonparametric Kaplan-Meier estimate of the mean of pyrene concentrations with 8 detection limits provides a completely distribution-free estimate of the UCL95 for data with nondetects:

```
> enparCensored(Pyrene, BDL.1, ci=TRUE, ci.type = "upper", ci.method="bootstrap",
n.bootstraps = 5000)

Results of Distribution Parameter Estimation
Based on Type I Censored Data
-------------------------------------------
Assumed Distribution:            None
Censoring Side:                  left
Censoring Level(s):               28  35  58  86 117 122 163 174
Estimated Parameter(s):          mean    = 164.09450
                                 sd      = 389.41379
                                 se.mean =  49.75292
Estimation Method:               Kaplan-Meier
Sample Size:                     56
Percent Censored:                19.64286%
Confidence Interval Method:      Bootstrap
Confidence Interval Type:        upper
Confidence Level:                95%
Confidence Interval:             Pct.LCL =   0.0000
                                 Pct.UCL = 262.8707
                                 BCa.LCL =   0.0000
                                 BCa.UCL = 259.0544
```

The 95% UCL is 262.87 or 259.05, depending on which bootstrap method is chosen.

Bootstrapping is widely applicable and distribution-free. It provides more accurate interval estimates when data are not normally distributed, the typical situation in environmental science. It provides estimates very similar to normal theory results when data do follow that distribution, so there is little penalty for using it. Are you still worrying about whether your data follow a normal distribution or not? Bootstrapping is presented in more detail in our Applied Environmental Statistics 1 online course.

C. View Webinars Anytime, for Free

Recordings of many of our webinars are freely available on our Online Training Site, http://practicalstats.teachable.com . Click on the "Show more courses" button to see more than the top three items, which are paid courses. There are now 9 webinars there for viewing:

1. Intro to Nondetects and Data Analysis
2. Fitting Distributions to Data with Nondetects
3. Testing Differences in Groups of Data With Multiple Detection Limits
4. The Mystery of Nondetects: How Censored Data Methods Work
5. Correlation and Regression for Data with Nondetects
6. Seven Perilous Errors in Environmental Statistics
7. Statistics For Managers
8. Intro to R
9. Forty Years of Water Quality Statistics: What's Changed, What Hasn't?

The first five discuss methods for data with nondetects from our NADA course. See our webinars page http://www.practicalstats.com/training/webinar.html for a short description of each. Both our courses and webinars can be streamed in your timezone, at any time of day you want to view them, on a computer or tablet or larger-format smartphones.

'Til next time,

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