Practical Stats Newsletter for October 2018

Subscribe and unsubscribe: http://practicalstats.com/news
Archive of past newsletters
http://practicalstats.com/news/archive.html

In this newsletter:
A. Practical Stats Courses
B. Comparing Data With Nondetects to a Standard, Part 2. or
"Why would you ever want to use a 't-test' for data with nondetects?"
C. Free webinars

A. Practical Stats Courses
Our self-paced Applied Environmental Statistics course is available in two parts on our online training site: http://practicalstats.teachable.com/
The two courses separately are each $650 USD for a 1-year access for one person. Or get both courses together (equivalent to our week-long course) in a bundle for $1200 USD.

Online classes coming soon to the training site:
Nondetects And Data Analysis Permutation Tests and Bootstrapping
Untangling Multivariate Relationships

B. Comparing Data With Nondetects to a Standard, Part 2. or
"Why would you ever want to use a 't-test' for data with nondetects?"

I had gotten several recent requests asking how to compare data with nondetects to a water quality standard, usually phrased something like "how do I run the t-test for data with nondetects?" I answered that question in the Aug 2018 newsletter by performing a test incorporating nondetects using maximum likelihood (ML), while assuming a normal distribution. This test has the same limitations as a t-test, and therefore was not the best method that could be done. It was just the method that some of you had asked for.

When a ML method assumes a normal distribution, there are two limitations. First, the data should be similar in shape to a normal distribution. Data with nondetects almost never are. Below we find that the Atrazine concentrations used in the August newsletter aren't modeled well by a normal distribution. Therefore the ML test assuming a normal distribution in the August newsletter will likely lose power, and p-values be too high, because the data don't fit the required distributional model.

The second limitation is that the lower end of a fitted normal distribution often goes below zero when there are nondetects. This is unrealistic for data that physically can't be negative, and may produce poor estimates of means, confidence limits, and other parameters. So the August newsletter procedure has severe limitations for data with nondetects.

What then can be done? Use a better-fitting distribution to model the data. Commonly used skewed distributions include the lognormal and gamma distributions. Then bootstrap a set of means that serve as the null hypothesis, and determine the probability of equaling or exceeding the observed sample

mean. That is the p-value for the test. If you haven't heard of bootstrap or permutation tests before, our Permutation and Bootstrapping online course will be available on our Training Center by the end of November. Applying these tests to censored data is done in our Nondetects And Data Analysis course, also coming soon on our Training Center. Or go to our News Archive at
http://practicalstats.com/news/archive.html
and click the `Perm Tests` button. You'll see several newsletters discussing them, and may get more out of this newsletter by reading the archived newsletters first.

The process to compute the LCL and associated p-value is:
1. Load the dataset, available after loading the NADA package. See last month's newsletter for definitions of variables.

```
> data(Atrazine)
```

2. Isolate the twenty-four June concentrations and save them as the dataset "atra". Attach to atra.

```
> atra <- Atrazine[Atrazine$Month =="June",]
> attach(atra)
```

3. Test which skewed distribution best fits your data. Load the EnvStats package. Then…

```
> boxcoxCensored(Atra, AtraCen, lambda=seq(0,1,0.1))

Results of Box-Cox Transformation
Based on Type I Censored Data
---------------------------------
Objective Name:                  PPCC
Data:                            Atra
Censoring Variable:              AtraCen
Censoring Side:                  left
Censoring Level(s):              0.01
Sample Size:                     24
Percent Censored:                37.5%

 lambda      PPCC
    0.0 0.9604439
    0.1 0.9484589
    0.2 0.9340850
    0.3 0.9175783
    0.4 0.8992978
    0.5 0.8796770
    0.6 0.8591889
    0.7 0.8383084
    0.8 0.8174798
    0.9 0.7970917
    1.0 0.7774619
```

Values for lambda are power coefficients, representing a data transformation of x^lambda where x is the column of data values. A lambda of 0.0 represents the lognormal distribution. A lambda of 0.3 approximately represents the gamma distribution. A lambda of 1 represents the normal distribution. The highest PPCC statistic is for data closest to a straight line on a probability plot, and so represents the distribution that best fits the data. The lognormal is the highest of the three with a PPCC of 0.96; the gamma is second-best at 0.917 and the normal is not surprisingly lowest at a PPCC of 0.777. We therefore use the lognormal to model the data distribution.

Note that you can get a nice graph of the three distributional fits with a command in the fitdistrplus package. I won't go into details here, but you'll see below that the data (step function) are fit best by the lognormal distribution. The command is:

```
cdfcompcens(list(cdflogn, cdfgamma, cdfnorm),legendtext=c("lognormal","gamma", "normal"))
```

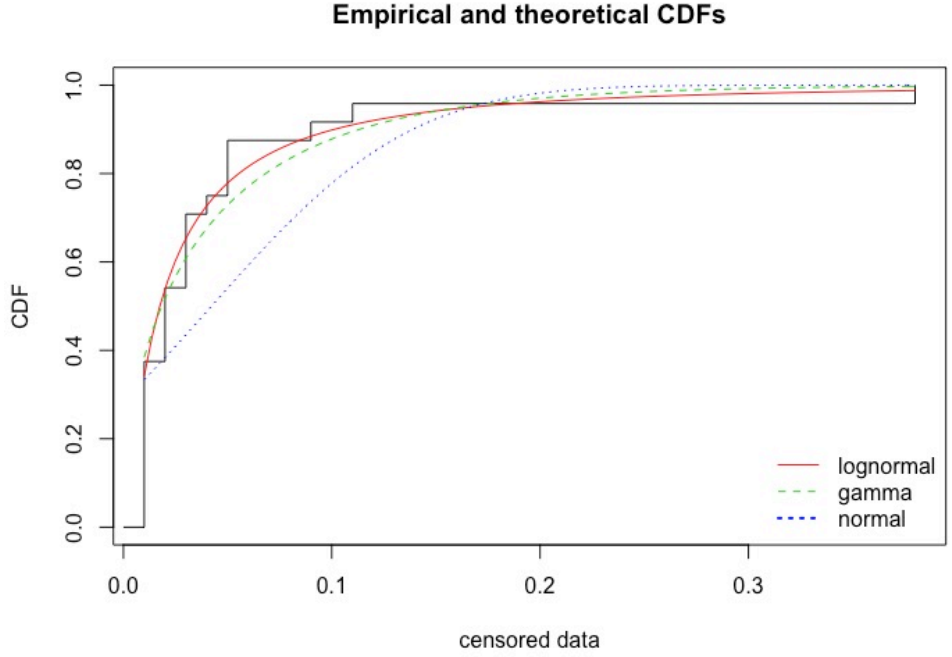

Figure 1. Three distributions (smooth curves) fit to the atrazine data (step function)

4. Compute the sample mean and LCL by bootstrapping the best-fitting distribution. The lognormal command below displays the mean and LCL in original units, not logarithms.

```
> elnormAltCensored(Atra, AtraCen, ci=TRUE, ci.type = "lower", ci.method = "bootstrap",
n.bootstraps = 10000)

Results of Distribution Parameter Estimation
Based on Type I Censored Data
-------------------------------------------
Assumed Distribution:            Lognormal
Censoring Side:                  left
Censoring Level(s):              0.01
Estimated Parameter(s):          mean = 0.04471125
                                 cv   = 2.35625345
Estimation Method:               MLE
Data:                            Atra
Censoring Variable:              AtraCen
Sample Size:                     24
Percent Censored:                37.5%

Confidence Interval for:         mean
Confidence Interval Method:      Bootstrap
Number of Bootstraps:            10000
Number of Bootstrap Samples
With No Censored Values:         0
Number of Times Bootstrap
Repeated Because Too Few
```

```
Uncensored Observations:         0
Confidence Interval Type:        lower
Confidence Level:                95%

Confidence Interval:             Pct.LCL = 0.02428512
                                 Pct.UCL =        Inf
                                 BCa.LCL = 0.02336678
                                 BCa.UCL =        Inf
```

The 95% LCL is 0.024, so if the standard were at 0.02 we would expect the mean 0f 0.0447 to be found significantly higher than the standard – the p-value for the test would be below 0.05. In a similar fashion, if the gamma distribution were the best fit, its command would be:

```
> egammaAltCensored(Atra, AtraCen, ci=TRUE, ci.type = "lower", ci.method = "bootstrap",
n.bootstraps = 10000)
```

5. Compute again just to save the mean and coefficient of variation (CV) of the observed data.

```
lognorm.mle <- elnormAltCensored(Atra, AtraCen)
  sample.mean <- lognorm.mle$parameters[1]
  sample.cv <- lognorm.mle$parameters[2]
```

6. Bootstrap from a lognormal distribution with the standard as the mean and the data's sample coefficient of variation.

Repeatedly generate the same number of observations as in the original data from the lognormal distribution at mean = 0.02 and save their means. This provides a picture of the null hypothesis when the mean is equal to the standard value. The p-value for a test of the null hypothesis that the mean atrazine is equal to (or less than) the standard is the proportion of the generated means equal to or above the observed mean of 0.0447. It answers the question "how likely is it to get a mean of 0.0447 when the true mean is at the standard value?"

```
> comp.mean = 0
> for (i in 1:10000) {
+     logdat.mc <- rlnormAlt(length(Atra), mean=0.02, cv = sample.cv)
+     comp.mean[i] <- mean(logdat.mc)
+   }
> pval <- sum(as.integer(comp.mean >= sample.mean))/10000
> cat("p-value for H0: the mean does not exceed", 0.02, "=", pval)

p-value for H0: the mean does not exceed 0.02 = 0.0174
```

Conclusion: The mean atrazine concentration exceeds the standard of 0.02, as shown by the p-value of 0.017, assuming a lognormal distribution. A plot is always helpful to visualize this:

```
> hist(comp.mean, main = "Histogram of bootstrapped means when true mean = 0.02", xlab =
"bootstrap means")
> abline(v=sample.mean, col="blue", lty = 2)
> text(sample.mean, 4500, labels = titl, pos=4, col="blue")
> text(0.08, 500, labels = ptext, col="red")
```

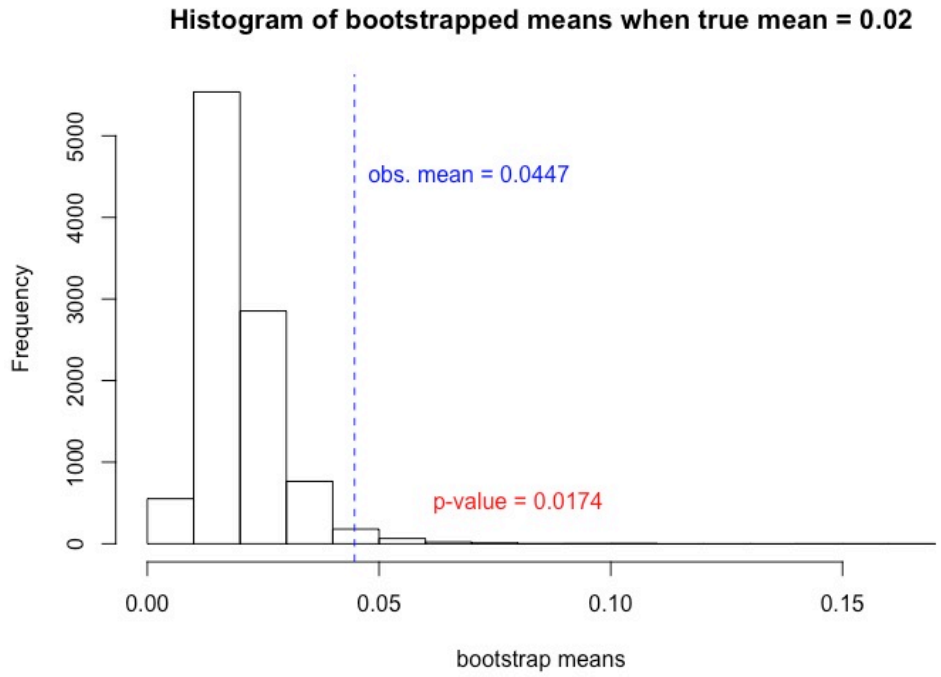

Figure 2. Histogram of 10,000 bootstrapped mean concentrations for a true mean concentration set at 0.02. The dashed blue line shows the value of the observed mean of 0.0447.

As with all bootstrap methods, your p-value may be slightly different. However, if many (such as 10,000) repetitions are used, the difference in p-values between runs will be quite small. And by the way, the misfitted normal distribution from August's newsletter did not find this significant difference from a standard of 0.02. Its p-value was 0.07. That is the loss of power that occurs when applying a method that assumes a normal distribution to non-normal, especially skewed, data.

C. Free webinars

There are now four free webinars available for you to listen to on the Practical Stats training site. The most recent is "Intro to R", an introduction to using R software for those who want help getting started with R. The other webinars you'll find there (by clicking the "View all courses" button) are:

An Introduction to Nondetects And Data Analysis
7 Perilous Errors in Environmental Statistics
Stats for Managers

'Til next time,

Practical Stats
-- Make sense of your data