Practical Stats Newsletter for August 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
C. Our Newsletter Schedule

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
Untangling Multivariate Relationships
Permutation Tests and Bootstrapping

B. Comparing Data With Nondetects to a Standard
One of the topics we cover in both the Applied Environmental Statistics course and the Nondetects and Data Analysis course is how to compare the mean of a column of data to a single number such as a water quality standard or environmental health criterion. This topic didn't appear in the *Statistics for Censored Environmental Data using Minitab and R* textbook (the book on how to interpret data with nondetects, published by Wiley in 2012). Here is a short version of what we cover in the Nondetects online course.

Comparing a column of data to a single value is called a one-sample test in statistics û the column of numbers is a sample from the population of interest. The mean is the most common parameter that standards are compared to. A one-sample test on the mean is just a restatement of a confidence interval on the mean. The parametric test without nondetects that assumes data follow a normal distribution is a t-test, a restatement of the parametric confidence interval on the mean computed using a t-statistic. If the test is set up so that the null hypothesis is that the mean is at or below the standard, and the alternative hypothesis is that the mean exceeds the standard, the test can be restated as a computation of the lower confidence limit on the mean (LCL). If not only the mean exceeds the standard, but the 95% LCL also exceeds the standard, the corresponding hypothesis test has a p-value that is lower than 0.05 (1 û 0.95, the confidence coefficient of the LCL).

An LCL for data with nondetects can be computed using the parametric method of maximum likelihood. Below we compute the 95% LCL and its corresponding p-value for the test on the mean. A distribution-free test of whether the mean exceeds the standard can also be computed by bootstrapping the LCL û you'll need to see the April 2018 newsletter for examples of bootstrapping intervals for censored data.

Software for censored data such as the NADA package for R or routines for survival (sometimes also called reliability) analysis in commercial software compute LCLs by maximum likelihood. Getting a p-value back for the corresponding test is possible as well, with varying levels of difficulty. With most commercial software its just point and click. If you'd like an example of computing the LCL and test using the commercial package Minitab, send an email to ask@practicalstats.com and we'll gladly send it to you.

Below I compute the LCL and the test using the NADA package in R. R and its packages are free. Their hidden cost is the additional difficulty it sometimes takes to compute the things you want. These LCLs and tests with censored data are a good example of that.

The example dataset Atrazine comes with the NADA package, and contains atrazine concentrations in groundwater during June and September. To run this example, install and load the NADA package in R. Then use the following R commands, the lines preceded by the > symbol.

1. Load and attach the dataset.

```
> data(Atrazine)
> attach(Atrazine)
> head(Atrazine)
  Atra Month AtraCen
1 0.38  June   FALSE
2 0.04  June   FALSE
3 0.01  June    TRUE
4 0.03  June   FALSE
5 0.03  June   FALSE
6 0.05  June   FALSE
```

`AtraCen` is the censoring indicator. When `AtraCen` is FALSE, the Atra concentration is a detected value. When `AtraCen` is TRUE, the concentration is a nondetect censored at the limit in the `Atra` column. For example, row 1 is a detected concentration of 0.38, and row 3 is a concentration of <0.01. There are 48 rows, 24 collected in June and 24 collected in September.

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

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

3. Compute the mean and standard deviation of the concentrations, using maximum likelihood and assuming a normal (gaussian) distribution, using the `cenmle` command. The mean is the value for the intercept (this is a regression with no explanatory variables). The standard deviation of concentrations is called the 'scale' on the output.

```
> A2 <-cenmle(Cen(Atra,AtraCen), dist="gaussian", conf.int=0.95)
> summary(A2)
             Value Std. Error      z        p
(Intercept)  0.0423     0.0154   2.75 5.95e-03
Log(scale)  -2.5857     0.1444 -17.91 9.53e-72

Scale = 0.0753
```

```
Gaussian distribution
Loglik(model)= -13.4   Loglik(intercept only)= -13.4
Loglik-r:  0

Number of Newton-Raphson Iterations: 5
n = 24
```

The mean equals 0.0423 and the standard deviation equals 0.0753. These values are not the result of substituting any numbers for the 9 nondetect values. See the textbook mentioned above for a description of how maximum likelihood works. The p-value of 5.95 x 10^-3 is a test of whether the intercept equals zero, which isn't of interest to us. Ignore it, it is not what you're looking for.

4. Compute the 95% two-sided confidence interval on the mean.

```
> mean(A2)
      mean         se    0.95LCL    0.95UCL
0.04231219 0.01538353 0.01216102 0.07246336
```

The standard error of the mean is the second parameter computed, followed by the two-sided interval endpoints. However, the LCL printed is not the one-sided 95%LCL we are looking for, which would have 5% error below it. It is a two-sided LCL with only 2.5% error below it. In many R commands you can specify to compute a one-sided LCL, but not here.

5. Compute a 95% one-sided LCL by computing the two-sided 90% LCL. There will be 5% error below that two-sided LCL, so that value is also the one-sided 95% LCL. Saving it as an object (A5) will allow us to use the computed mean and standard error of the mean to compute a p-value.

```
> A4 <-cenmle(Cen(Atra,AtraCen), dist="gaussian", conf.int=0.90)
> A5 <- mean(A4)
> A5
      mean         se     0.9LCL     0.9UCL
0.04231219 0.01538353 0.01700853 0.06761585
```

The one-sided 95% LCL equals 0.017. The LCL represents the low end of where the true population mean might lie, and it exceeds the standard. When the LCL is above the regulatory standard, the mean is significantly greater than the standard at an alpha of 0.05. The significance level of 95% or 0.95 for the lower confidence limit corresponds to a 1-0.95 = 0.05 cutoff for alpha.

6. Obtain a p-value for the test that the mean is greater than the standard.
First, I'll use the one-sided 95%LCL of 0.017 as our 'standard'. Because it is a 95% LCL it is right at the alpha equals 0.05 level. Use the pnorm command, entering the standard, then the mean of observed data and the standard error of the mean (0.01538 above). The result is the pvalue.

```
> pval = pnorm(0.017,mean=A5[1],sd=A5[2])
> pval
[1] 0.04994281
```

which is right at 0.05 as expected.

Second, I'll use a value above the LCL, say 0.02 as the standard. Since the LCL dips below the standard the mean will not be significantly greater than the standard at an alpha of 0.05 so the p-value will be higher than 0.05.

```
> pval = pnorm(0.02,mean=A5[1],sd=A5[2])
> pval
[1] 0.07347425
```

Third, use a value below the LCL as the standard, say 0.01. Since the LCL lies above the standard, the mean is significantly higher than the standard at an alpha of 0.05 so the p-value will be smaller than 0.05.

```
> pval = pnorm(0.010,mean=A5[1],sd=A5[2])
> pval
[1] 0.01784506
```

These three `pnorm` commands demonstrate the relationship between the LCL and the p-value of a test. Of course, you'll only have one standard to enter, so you'll just need to enter it once into the `pnorm` command to get its computed p-value. Even then, given all of this trouble you might just be tempted to purchase commercial statistics software so that you can point and click and pull down menus to get the answer.

C. Our Newsletter Schedule

Our newsletter has been published six times per year since 2009 and four times per year before that, back to 2003. You'll find all of the ones that are still relevant on our News Archive page at http://practicalstats.com/news/archive.html .
However, I'm cruising toward retirement and so may miss a few here and there. You may have noticed we didn't get one out in June. I'll let everyone know when the final newsletter hits the street sometime in 2019. After that, I'll still be running our online courses, and anyone taking one of those will be able to get help from me with all of the course material. Otherwise, I hope to be doing other things than fulltime statistics.

'Til next time,

Practical Stats
-- Make sense of your data