Analysis of Environmental Data With Nondetects

Statistical Methods for Censored Environmental Data

Dennis Helsel and Lopaka (Rob) Lee

August 2006

Continuing Education Workshop at the

Joint Statistical Meetings

American Statistical Association

Seattle, Washington

Based on the textbook

Nondetects And Data Analysis: Statistics for censored environmental data
By Dennis Helsel
Wiley, 2005

Workshop Schedule

Introduction

Current Practices for Censored Environmental Data

Plotting Data with Nondetects

- Censored boxplots, probability plots
- Survival curves, scatterplots

Review of use of the R statistical package and

- NADA for R

Estimating Descriptive Statistics

- Parametric method – Maximum Likelihood
- Robust ROS
- Nonparametric Method – Kaplan Meier
- Interval Estimates

Lunch

Two-Group Tests

- Generalized Wilcoxon test
- Maximum likelihood version of a t-test

Three or More Group tests

- Generalized Wilcoxon test
- Maximum likelihood version of ANOVA

Correlation

- Kendall's tau for censored data
- Likelihood correlation coefficient

Linear Regression

- Akritas-Theil-Sen line
- Censored regression by MLE

Review and Summary

Exercises using R

1. A Brief Overview of R and the NADA for R package

S-Plus and R are both implementations of the S-language. The S-language was originally developed by John Chambers and others at Bell Labs. S-Plus is a proprietary implementation currently owned by the Insightful Corporation. R is a completely free implementation that is developed by volunteers worldwide.

If you are completely new to R, the best place learn about it is the R Project website (http://www.r-project.org). The "Introduction to R" manual is free at the R Project website. Peter Dalgaard's "Introductory Statistics with R" is one of the best introductory texts. Also, Paul Murrell's "R Graphics" is the definitive source on manipulating graphics/plots in R. The R Project website maintains a listing of other books about R. Many books have been written for S-Plus and for the most part, they can be used with R.

The NADA for R package is a collection of datasets and S-language implementations of methods described in the book "Nondetects And Data Analysis: Statistics for censored environmental data" by Helsel, or the NADA book. The package is an official CRAN (Comprehensive R Archive Network) package, meaning that it can be installed automatically from within R. Assuming that you have a working copy of R, and a functioning Internet connection, the NADA for R package can be automatically installed with the following command:

```
> install.packages("NADA")
```

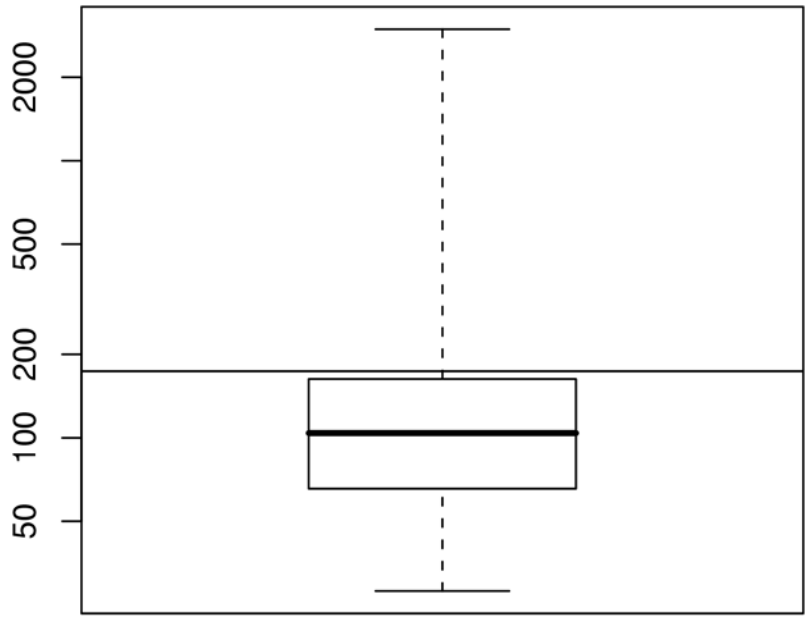
In writing the NADA for R package we have made a concerted effort to make function names and their usage contexts as consistent as possible. Almost all functions begin with the prefix "cen" -- for example, "cenfit", and "cenmle". Also, generic functions such as "mean", "quantile", and "plot" can be used with output objects from any of the NADA for R functions. This point will become clearer in the examples below.

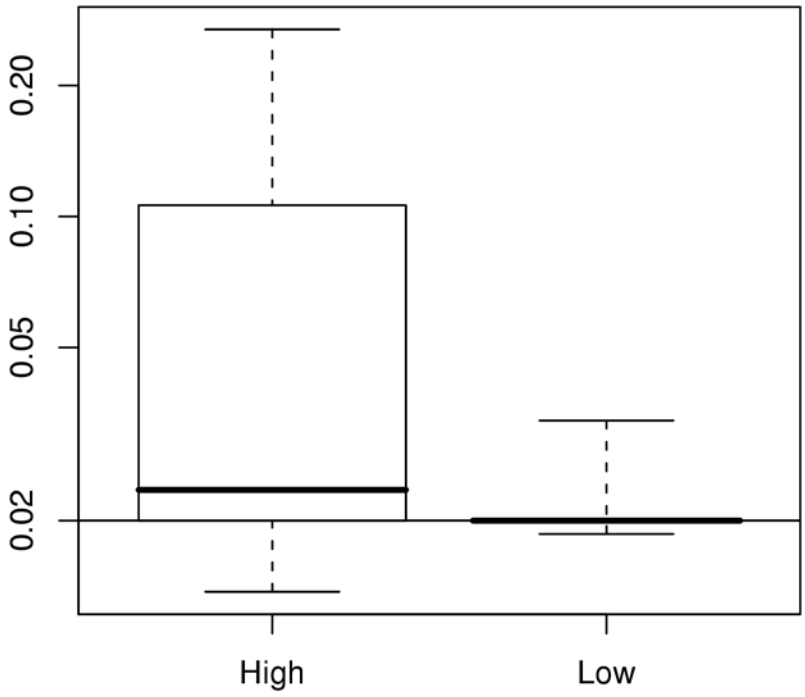
Keep in mind that both R and the NADA for R package are rapidly improving. What is missing today could be added very shortly. Additionally, development is completely open. If there is some functionality that is missing or something you've found to be wrong, you are welcomed to suggest, or contribute the code, for a solution – all contributors are properly recognized within the documentation. Email Rob Lee (rclee[atsign]usgs.gov) with suggestions for additions to the NADA for R package.

2. Plotting Data with Nondetects

Censored boxplot
A boxplot of the pyrene data (first graph) is not that helpful because the highest detection limit is quite high for this data. The horizonatal line is drawn at the highest detection limit in the data set. Percentiles above this line are unaffected by censoring. Percentiles below the line must be estimated, by Kaplan-Meier or ROS.

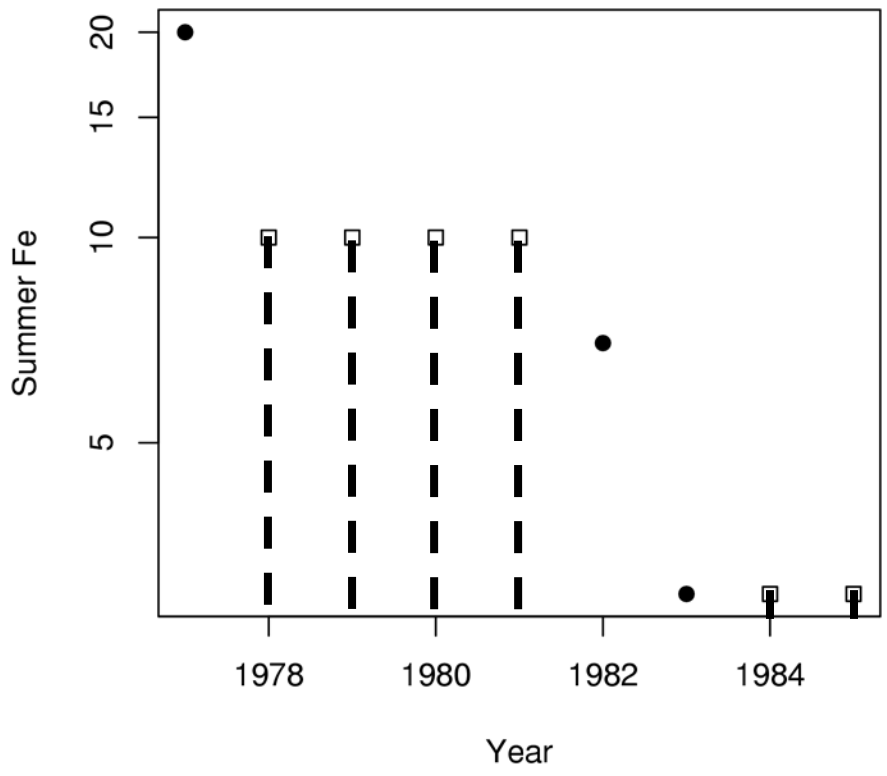
Boxplots for data with detection limits at lower percentiles are especially helpful when comparing among groups. This second boxplot effectively shows that the "High" group has generally higher values than the "Low" group.




0.20
0.10
0.05
0.02
High
Low

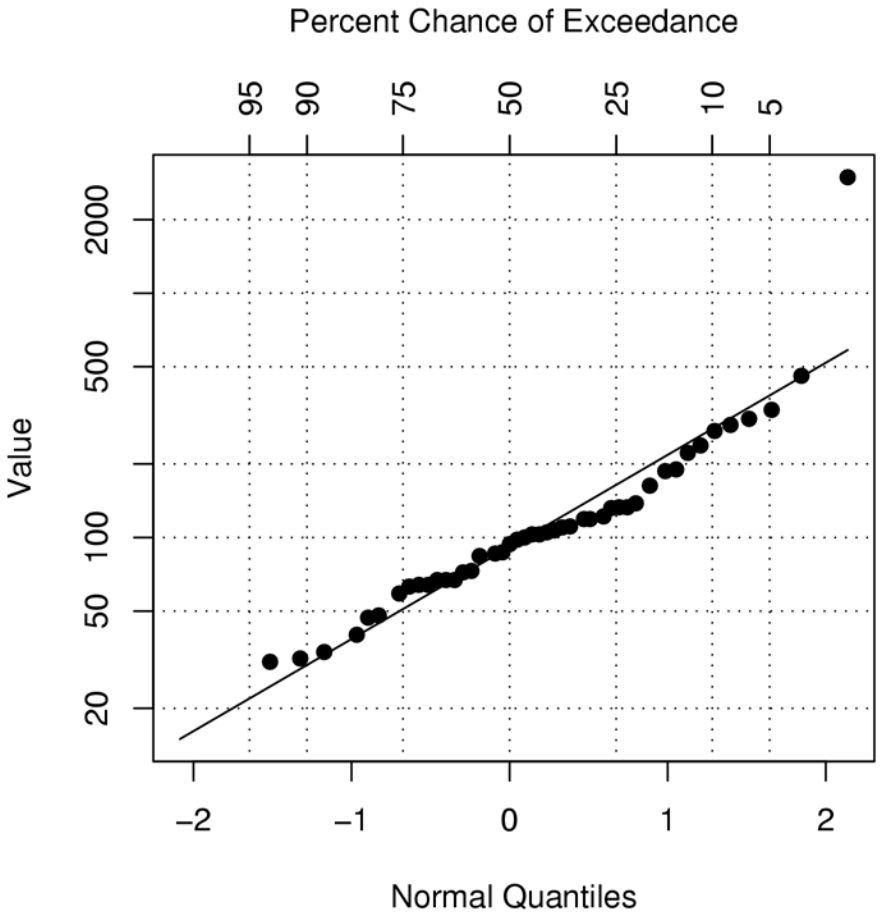
Censored scatterplot

Scatterplots are difficult to draw well with censored data, as the censored values are often drawn at the detection limit, or one-half the detection limit, or at zero. All such plots are misleading, because unique censored values are unknown. Instead, left-censored data can be plotted as intervals between zero and the detection limit for each observations. In this way, no false statements about where an individual value is located, or that all such observations are at the same value, are made. Below is a scatterplot for data with two detection limits, at 10 and 3:

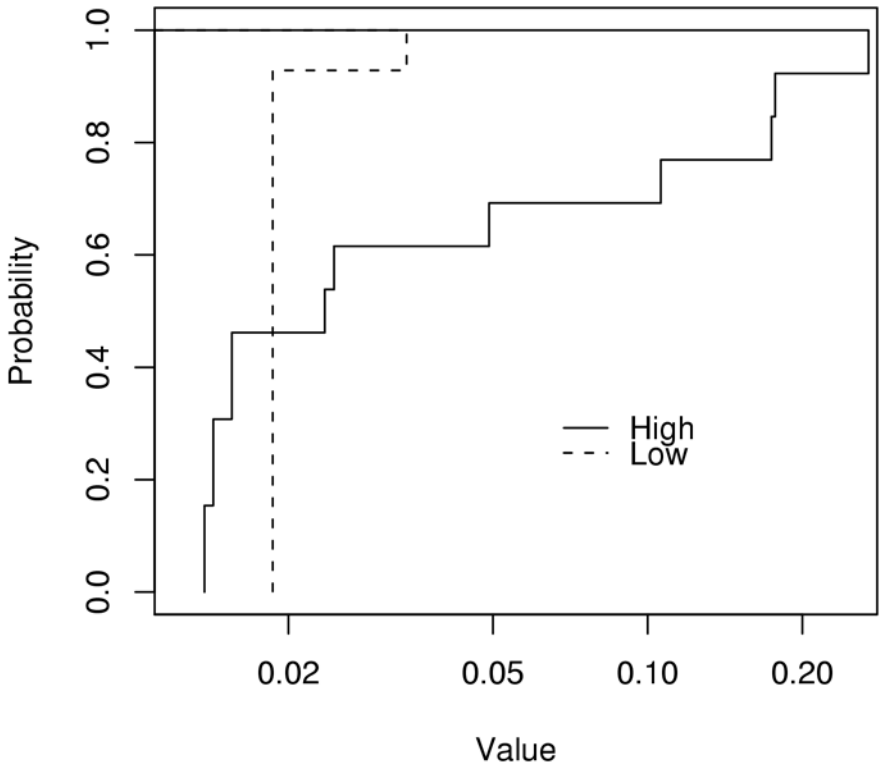


The next two plots are usually available with survival analysis software.

A probability plot can be used to evaluate whether data follow an assumed distribution. Here are the pyrene data compared to a lognormal distribution.



Survival function plots for one to several groups are available in NADA for R. Note that they appear "backwards" from survival function plots for right-censored time to event data, as these are left-censored values.



3. Estimating Summary Statistics

She (1997) is one of the few papers in the environmental sciences within the last decade to use survival analysis techniques. The organic contaminant pyrene was measured in benthic sediments of Puget Sound, Washington. Sampling locations were in areas of highest probable impact from discharged effluents. As we have seen, the Pyrene data set contains pyrene concentrations along with an indicator of whether each value is a detected observation, or a detection limit. There are 11 left-censored observations at 8 different detection limits out of a total of 56 observations.

Estimate the mean, median and standard deviation for these pyrene data using three methods:
1. Kaplan-Meier
2. ROS, and
3 maximum likelihood estimation (assume a lognormal distribution)

For MLE, make sure you observe a probability plot to see how well the data fit the assumed distribution.

The routines in NADA for R internally flip the original data by subtraction from a large constant, in order to produce right-censored values that can be input to survival analysis routines. If you are using other software, you will need to manually flip left-censored environmental data into right-censored values, and perform the operations on the transformed (flipped) values. Then sample estimates of location such as mean, median and percentiles must be re-transformed back into original units by subtracting the estimates from the same flipping constant. NADA for R does all of this automatically.

The ShePyrene data are distributed as a part of the NADA for R package. Import the data into your workspace and attach it to your search path with the following commands:

```
> library(NADA)
> data(ShePyrene)
> attach(ShePyrene)
> names(ShePyrene)
[1] "Pyrene"    "PyreneCen"
```

The ShePyrene dataset contains a numeric vector of observations (“Pyrene”), and a logical vector of censoring indicators (“PyreneCen”). The logical vector is a list of TRUE or FALSE values that indicate where the values in Pyrene are censored (TRUE), or uncensored (FALSE). All functions in the NADA package accept input in this format.

Exploring the data

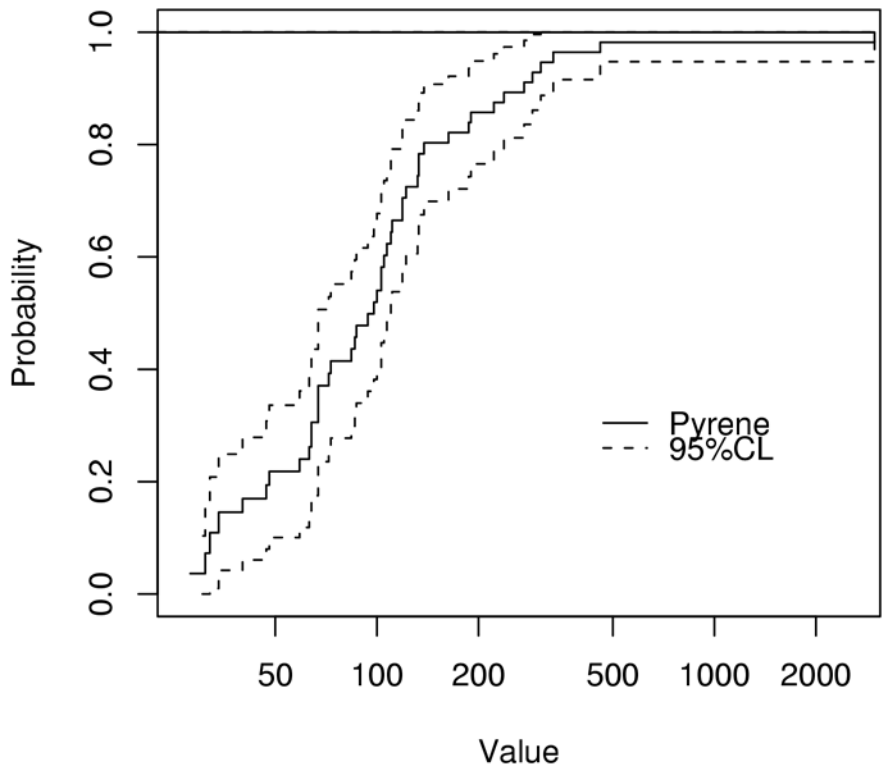
In R and S-Plus, the `summary` command is typically used to briefly describe the characteristics of the data. In the NADA for R package, the `censummary command fulfills the same role for censored data:`

```
> censummary(Pyrene, PyreneCen)
Summary:
      n   n.cen pct.cen     min     max
   56.0    11.0    19.6    28.0  2982.0

Thresholds and counts:
 28  35  58  86 117 122 163 174
  1   2   1   1   1   1   3   1

Uncensored observations between each threshold:
 28  35  58  86 117 122 163 174
  3   3  10  11   2   5   1  10

ROS probability of exceeding each threshold:
   28    35    58    86   117   122   163   174
0.963 0.852 0.778 0.555 0.333 0.292 0.196 0.179
```

Calculating Summary Statistics and Plotting Models

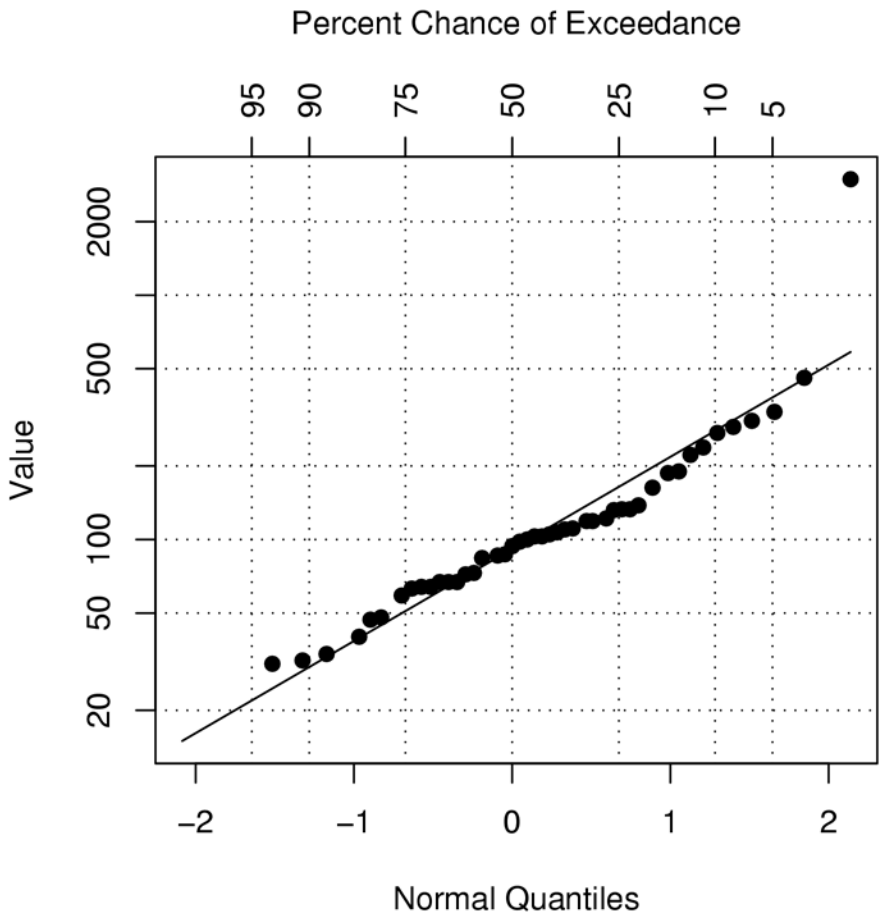
Kaplan-Meier. The `cenfit` command constructs a K-M model of the input data. Once a model is constructed (and assigned to a name), you can also plot it using the generic `plot` command. The plot of a K-M object is a step function (survival function plot) with confidence limits. The default confidence limit is 0.95. Later we will modify this.

```
> pykm = cenfit(Pyrene, PyreneCen)
> pykm
       n    n.cen   median     mean       sd
 56.0000  11.0000  98.0000 164.0945 389.5899
> plot(pykm)
```



ROS. The `cenros` command constructs ROS models. It works almost exactly like the `cenfit` command. A plot of an ROS object displays the detected (uncensored) observations and the probability plot - regression model.

```
> pyros = cenros(Pyrene, PyreneCen)
       n    n.cen   median     mean       sd
 56.0000  11.0000  90.5000 163.1531 393.1309
> plot(pyros)
```



From the probability plot, it looks like the data fit a lognormal distribution reasonably well. There is one large outlier, the uppermost point, which is far away from the other points and from the lognormal model. This may strongly influence the values for the mean and standard deviation from what they would be without that point. More information about the ROS regression, including a plot of residuals can be obtained using the generic "summary" command. Additionally, generic functions that typically work with linear regressions, such as "residuals", and "coef" also work with ROS objects.

Maximum Likelihood (MLE).

The cenmle command constructs MLE objects and works almost identically to the ROS and K-M function:

```
> pymle = cenmle(Pyrene, PyreneCen
> pymle

        n     n.cen    median      mean        sd
 56.00000  11.00000  91.64813 133.91419 142.66984
```

The MLE mean and standard deviation estimates are quite a bit lower than those produced by the other two methods. The issue is whether the one outlier should influence the result. If a simple arithmetic mean were computed, that point would strongly influence the result. K-M and ROS give it similar weight. As an optimization procedure, MLE apparently does not. The issue is whether this one point should have a strong influence on the estimates.

In summary, descriptive stats for the three methods are:

```
> censtats(Pyrene, PyreneCen)

       n    n.cen  pct.cen
56.00000 11.00000 19.64286

      median     mean       sd
K-M 98.00000 164.0945 389.5899
ROS 90.50000 163.1531 393.1309
MLE 91.64813 133.9142 142.6698
```

MLE is notorious for producing poor estimates of the standard deviation for 'small' datasets, where with this much skewness, small is probably fewer than 50 detected observations. This data set has 45 detected observations, so if they are considered to be close to a lognormal distribution, use of MLE might be fine. If there had been fewer observations we would certainly not prefer the MLE, and would use the ROS or Kaplan-Meier results instead.

4. Interval Estimates

For the ShePyrene data, compute a 2-sided 95% confidence interval around the mean concentration using Kaplan-Meier and MLE estimates of mean and standard errors.

Kaplan-Meier

Using cenfit to estimate summary statistics, a 95% confidence interval can often be printed out by the survival analysis routine. Even though the estimate for the mean is nonparametric, the confidence interval requires that the variation of possible estimates of the mean is normal for this confidence interval to be valid. The sample size is just over 50 but the data are skewed, so we expect that this normal approximation may not be very good, especially for the upper end.

The confidence limits for the K-M method (the cenfit function) can be adjusted using the conf.int parameter (the default is conf.int=0.95).

```
> pykm = cenfit(Pyrene, PyreneCen, conf.int=0.95)
> mean(pykm)
     mean        se   0.95LCL   0.95UCL
164.09450  52.06114  62.05655 266.13246
```

Note that once the confidence interval is specified, it is fixed for that model/object. For quantile estimates `conf.int` parameter is a true-false indicator that specifies if the confidence limits should be printed when returning the quantiles:

```
> quantile(pykm, conf.int=TRUE)
  quantile value     0.95LCL   0.95UCL
1     0.05    31   0.0000000 0.1032754
2     0.10    32   0.0000000 0.1607601
3     0.25    63   0.1184371 0.3615577
4     0.50    98   0.3607339 0.6366579
5     0.75   133   0.6284849 0.8603265
6     0.90   273   0.8118494 0.9738649
7     0.95   333   0.8874540 1.0000000
```

MLE:

The cenmle produces confidence intervals for the mean in a similar fashion to the previous functions:

```
> pymle = cenmle(Pyrene, PyreneCen, conf.int=0.95)
> mean(pymle)
       mean          se     0.95LCL     0.95UCL
133.9141886   0.1218482 102.5101028 174.9389516
```

A better method for computing confidence intervals and bounds for skewed data would be bootstrapping. This is not yet implemented within NADA for R. Bootstrapping of censored data estimates can be implemented using Minitab statistical software with NADA macros located at www.practicalstats.com/nada .

5. Two-Group Tests

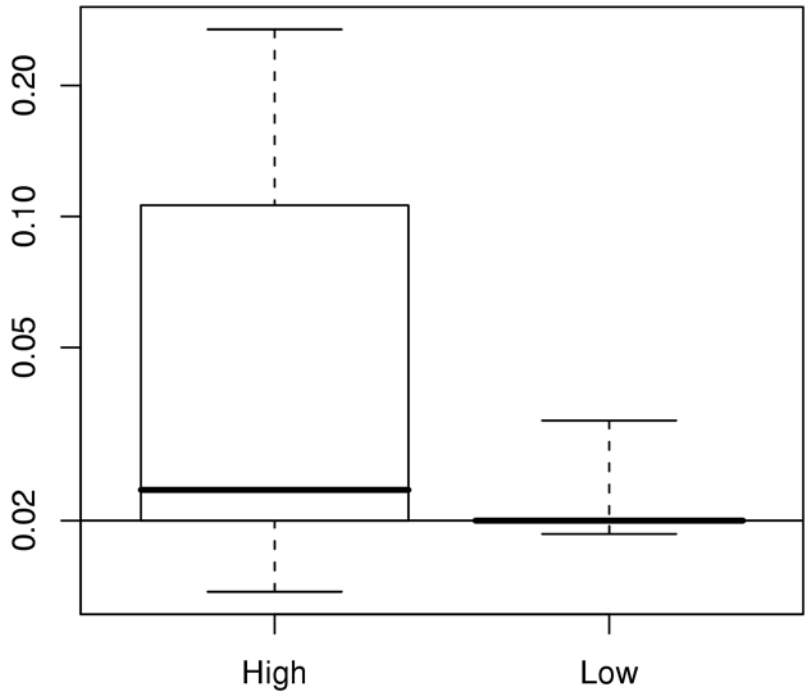
Golden and others (2003) measured lead concentrations in the bodies of black-crowned night herons before and after exposure to doses of lead nitrate in water. Of interest was whether lead would be incorporated into the feathers, organs and blood of the birds, and especially whether lead in feathers, which can be gathered without injuring the birds, is a good predictor of lead in the rest of the birds' bodies.

In worksheet Golden are the lead concentrations in blood for two groups, either birds with little or no exposure to lead (Low), or those who were exposed to much higher lead concentrations (High). Determine if the blood lead concentrations are higher in the High exposure group using the MLE version of a two-sample test, and using the nonparametric Peto-Prentice score test.

Also note that this should be a one-sided test. However, most survival analysis software for score tests do not allow a one-sided alternative, as the test mirrors the null and alternative hypotheses for an ANOVA type setup. P-values can be divided in half to produce one-sided p-values, after determining that the difference between groups is in the direction expected by the one-sided alternative.

Answer:
The boxplots show that the High-dose group certainly appears to have higher lead concentrations:

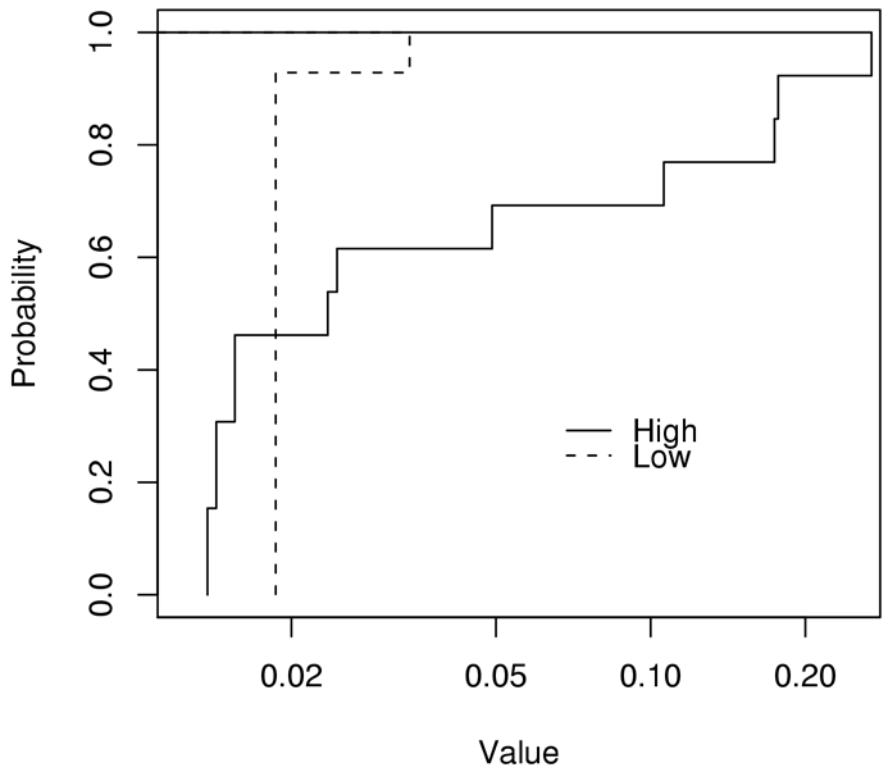
```
> data(Golden)
> attach(Golden)
> cenboxplot(Blood, BloodCen, DosageGroup)
```




```
> bloodPb = cenfit(Blood, BloodCen, DosageGroup)
                    n n.cen     median       mean          sd
DosageGroup=High 13     3 0.02352941 0.07003204 0.081487274
DosageGroup=Low  14    12 0.01864407 0.01973822 0.003945039
```

The K-M summary statistics produced by the cenfit command show that the group medians are almost identical. However, the standard deviations are quite different. A plot of the two survival curves show evidence that though the medians of the two groups are about the same (near the detection limit of 0.02), the upper halves of the two groups are quite different.

```
> plot(bloodPb)
```



Use the cendiff function to test for difference between the groups:

```
> cendiff(Blood, BloodCen, DosageGroup)
                     N Observed Expected (O-E)^2/E (O-E)^2/V
Dosage < 0.05=FALSE 13     7.20     4.47      1.67      4.75
Dosage < 0.05=TRUE  14     1.52     4.25      1.76      4.75

 Chisq= 4.8  on 1 degrees of freedom, p= 0.0293
```

The test statistic for the Peto-Prentice test has a two-sided p-value of 0.0293. The High group has higher values than the Low group, in the direction expected by our alternate hypothesis. Therefore to get a one-sided p-value we would cut the reported value in half, and say that there is evidence that the CDF of High > Low at p = 0.015.

MLE two-group test:

The `cenmle` function can be used to perform a parametric test of whether there is a significant difference in the mean of the lead concentrations in the two groups. Note that now the grouping variable (DosageGroup) must be specified as an explanatory variable.

```
> cenmle(Blood, BloodCen, DosageGroup)

                Value Std. Error       z        p
(Intercept)    -3.406      0.340 -10.003 1.48e-23
DosageGroupLow -1.885      0.647  -2.911 3.60e-03
Log(scale)      0.162      0.213   0.759 4.48e-01

Scale= 1.18

Log Normal distribution
Loglik(model)= 12.9   Loglik(intercept only)= 7.5
Loglik-r:  0.570889

Chisq= 10.65 on 1 degrees of freedom, p= 0.0011
Number of Newton-Raphson Iterations: 4
n = 27
```

The likelihood-ratio p-value of 0.0011 (a two-sided p-value) indicates that the means for the two groups are significantly different. The regression coefficient for DosageGroupLow (Value=-1.885) measures the magnitude of the difference between the Low and High groups. So in natural log units, the mean of the High group is 1.88 log units higher than the mean of the Low group. This translates into a ratio of their geometric means, so that the High group has a geometric mean (median) averaging e^1.88 = 6.55 times the Low group median.

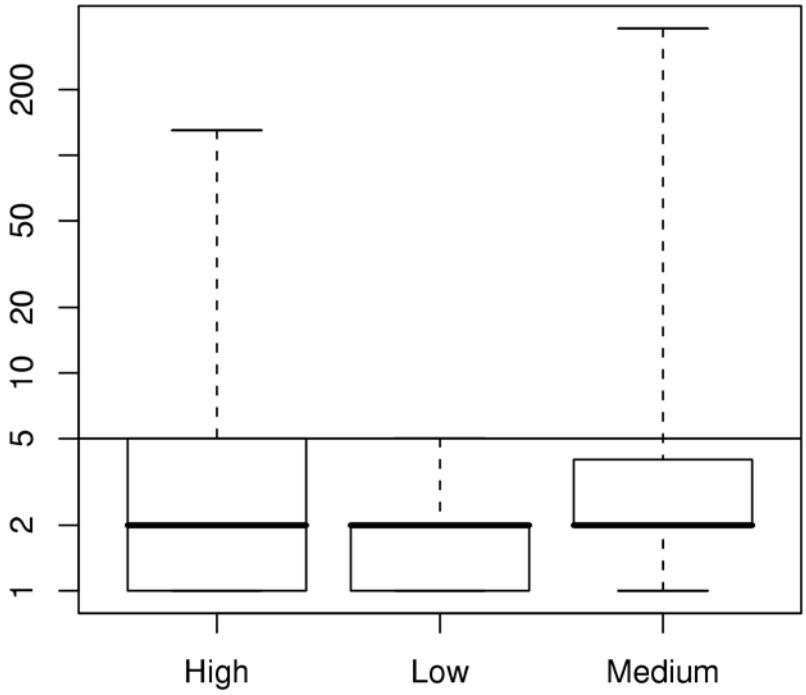
6. Comparing Three or more groups

TCE (an organic solvent and pollutant) concentrations in Long Island groundwaters were measured by Eckhardt et al (1989). The data are in TCE. Determine if the mean or median TCE concentration differs in groundwater under three land-use groups (Low, Medium, and High density residential areas). Use `cenmle` to run a parametric test on the means, and `cenfit` to run the nonparametric Peto-Prentice test of whether the distributions are identical or not.

Answer:

A censored boxplot shows that the Low density residential group has no detected concentrations, while the other two groups do.

```
> data(TCE)
> attach(TCE)
> cenboxplot(TCEConc, TCECen, Density)
```

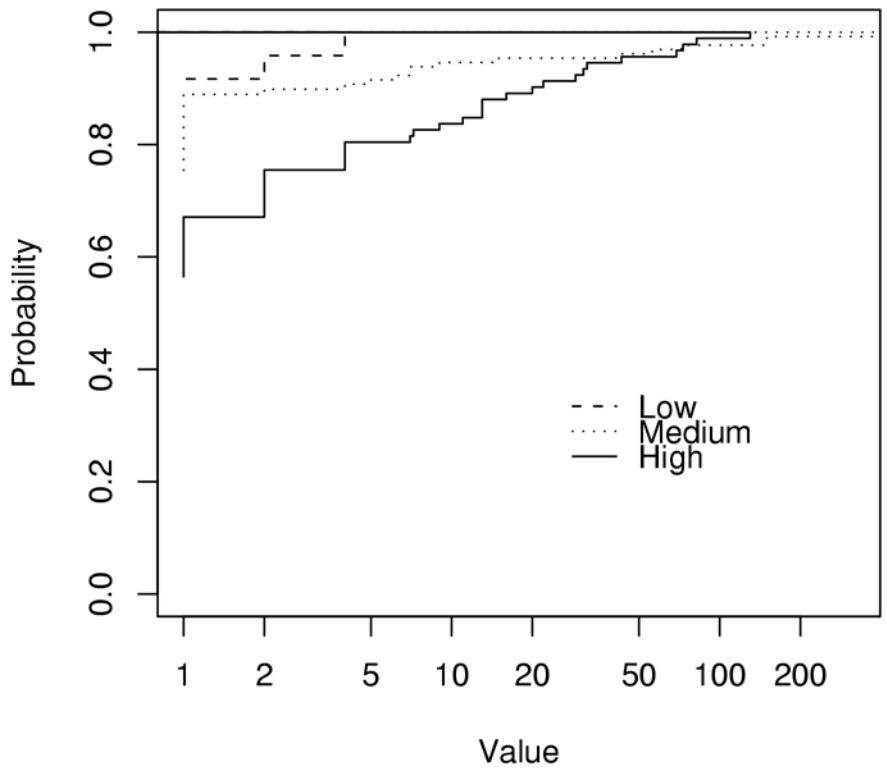


The `cenfit` command reports summary statistics for the three groups.

```
> cenfit(TCEConc, TCECen, Density)
                 n n.cen median     mean         sd
Density=High    92    58     NA 7.778019 19.4895231
Density=Low     25    23     NA 1.166667  0.6364688
Density=Medium 130   113     NA 7.867264 38.5791132
```

Plotting the cenfit object produces the left-censored survival function for each group:

```
> plot(cenfit(TCEConc, TCECen, Density))
```



The Peto-Prentice test is computed using the cendiff command:

```
> cendiff(TCEConc, TCECen, Density)

                 N Observed Expected (O-E)^2/E (O-E)^2/V
Density=High    92    30.45     18.2      8.26     15.65
Density=Low     25     1.73      5.7      2.76      3.62
Density=Medium 130    15.47     23.8      2.89      6.76

 Chisq= 16.3  on 2 degrees of freedom, p= 0.000295
```

Based on the two-sided p-value we would determine that there are significant differences between the CDFs of the three groups.

MLE:

For the parametric approach, a censored regression fitting parameters by MLE is performed with the group assignment as the explanatory variable. This is sometimes called a "factor", and internally the k groups are represented by (k-1) binary variables. Using the `cenmle` command:

```
> tcemle = cenmle(TCEConc, TCECen, Density)
> tcemle

               Value Std. Error     z        p
(Intercept)   -0.722      0.416 -1.73 8.28e-02
DensityLow    -3.060      1.138 -2.69 7.17e-03
DensityMedium -1.656      0.553 -2.99 2.76e-03
Log(scale)     1.048      0.111  9.41 4.76e-21

Scale= 2.85

Log Normal distribution
Loglik(model)= -308.7   Loglik(intercept only)= -316.4
Loglik-r:  0.2459125

Chisq= 15.41 on 2 degrees of freedom, p= 0.00045
Number of Newton-Raphson Iterations: 4
n = 247
```

The log-likelihood for this model (-308.7) is compared to the log-likelihood of the data without a group assignment (-316.4) and twice the difference compared to a chi-square distribution with (k-1) = 2 degrees of freedom: 2(-308.703 + 316.405) = 15.41 The resulting p-value of 0.0004 leads us to reject that the mean concentrations in all three groups are the same.

7. Correlation and Regression

The TCE data can also be related to the continuous explanatory variables Population Density, Depth to Water, and Percent Industrial land use – the data are found in the data set `TCEReg`. Note that the `TCEReg` dataset contains the same variable names as the `TCE` dataset. However, the `TCEReg` has a slightly different structure. Therefore, you must be careful to refer to the proper variables by either detaching/attaching the right datasets:

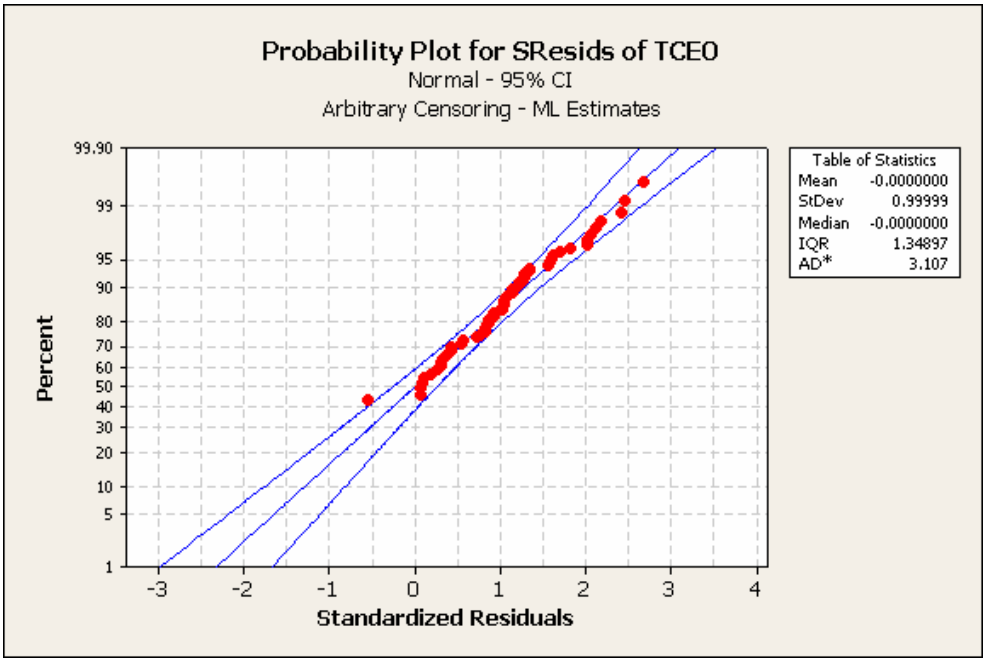
```
> detach(TCE)
> data(TCEReg)
> attach(TCEReg)
```

A. We would like the (parametric) likelihood correlation coefficient r and MLE regression equation for the best linear model between these three explanatory variables and TCE, testing to determine whether each variable is significant or not. Use the `cenreg` command

B. We would also like the value of Kendall's tau correlation coefficient and the Akritas-Thiel-Sen slope using the one explanatory variable Population Density, to determine whether there is a significant association between TCE concentration and Population Density without assuming a distribution for the residuals of the relationship. Use the `cenken` command.

Answer:

A. Parametric approach. First a normal distribution was assumed for the regression residuals, but the residuals plot was quite curved. Instead, a lognormal distribution is assumed for the residuals and the procedure run again. The residuals are essentially linear on the lognormal probability plot, so that a lognormal distribution is suitable for MLE regression of the TCE data.



Probability plot of residuals for MLE regression of the TCE data, assuming a lognormal distribution.

To compute the overall test for whether this three-variable model predicts TCE concentrations better than simply the mean concentration (the null model), the likelihood ratio test comparing the null model to the model with all 3 variables is performed using the summary function with the `cenreg` command:

```
> tcemle3 = cenreg(Cen(TCEConc, TCECen)~PopDensity+Depth+PctIndLU)
> tcemle3

               Value Std. Error      z        p
(Intercept) -2.88027    0.82355 -3.497 4.70e-04
PopDensity   0.25090    0.07452  3.367 7.60e-04
Depth       -0.00437    0.00233 -1.874 6.09e-02
PctIndLU     0.04065    0.05264  0.772 4.40e-01
Log(scale)   1.03378    0.11066  9.342 9.43e-21

Scale= 2.81

Log Normal distribution
Loglik(model)= -303   Loglik(intercept only)= -316
Loglik-r:  0.321

Chisq= 26.9 on 3 degrees of freedom, p= 6e-06
Number of Newton-Raphson Iterations: 4
n = 247
```

The overall log-likelihood test statistic G^2_0 is computed for the overall test as:

$$G^2_0 = 2\,[\ln L(\beta) - \ln L(0)] = 2\,[-302.931 - (-316.404)] = 26.9$$

Comparing 26.95 to a table of the chi-square distribution with k=3 degrees of freedom (3 explanatory variables), the resulting p-value equals < 0.0001, much less than the alpha of 0.05. So the three-variable model is considered better than no model at all.

Based on the partial log-likelihood tests, the best regression model for these data has Popden and Depth as explanatory variables.

```
> tcemle2 = cenmle(Cen(TCEConc, TCECen)~PopDensity+Depth)
> tcemle2

               Value Std. Error     z        p
(Intercept) -2.79067    0.81018 -3.44 5.72e-04
PopDensity   0.25959    0.07405  3.51 4.56e-04
Depth       -0.00434    0.00234 -1.85 6.37e-02
Log(scale)   1.03487    0.11068  9.35 8.78e-21

Scale= 2.81

Log Normal distribution
Loglik(model)= -303   Loglik(intercept only)= -316
Loglik-r:  0.318

Chisq= 26.4 on 2 degrees of freedom, p= 1.9e-06
Number of Newton-Raphson Iterations: 4
n = 247
```

The final model for explaining TCE concentrations is

$$\ln TCE = -2.79 + 0.260 * PopDensity - 0.004 * Depth$$

And the likelihood r correlation coefficient (Loglik-r) for this model is 0.32.

The MLE lognormal regression model is pictured as a curve in Figure 12.3 of the NADA textbook. Most of the data are nondetects, and plot on top of one another at the low end of the y-axis scale. A variable which would improve on this model is one that would explain the high concentrations occasionally seen at lower population densities.

B. Approach based on Kendall's tau.

Kendall's tau and associated Akritas-Thiel-Sen (ATS) line can be currently computed with only one explanatory variable. Population density (Popden) is used as the most significant of the three possible explanatory variables. The resulting output includes the slope which, when multiplied by Popden and subtracted from the TCE concentration data, results in residuals having a Kendall's tau correlation of zero:

```
> cenken(log(TCEConc), TCECen, PopDensity)
tau
[1] 0.1458477

slope
[1] 0.3835066

p
[1] 0.0003007718
```

A comparison of the MLE and ATS results is

Method	Slope	Intercept	p-value
MLE	0.309	-3.73	<0.001
ATS	0.383	NA	<0.001

Workshop and Materials Prepared by:

Dennis Helsel, U.S. Geological Survey

Lopaka Lee, U.S. Geological Survey

References:

Our papers/textbook on methods for censored environmental data

Lee, Lopaka, and Helsel, Dennis, (in press), Statistical analysis of water-quality data containing multiple detection limits II: S-language software for nonparametric distribution modeling and hypothesis testing. submitted to Computers & Geosciences

Lee, Lopaka, and Helsel, Dennis, 2005, Statistical analysis of water-quality data containing multiple detection limits: S-language software for regression on order statistics, Computers & Geosciences 31, 1241-1248

Helsel, D.R., 2005. ***Nondetects And Data Analysis: Statistics for censored environmental data***. John Wily and Sons, New York. 250 p.

Helsel, D.R., 2005. More Than Obvious: Better methods for interpreting nondetect data. Environmental Science and Technol. 39 (20), p. 419A–423A**.**

Helsel, Dennis R., 1990, Less Than Obvious: Statistical Treatment of Data Below the Detection Limit, Environmental Science and Technology 24(12), p. 1766-1774.

Helsel, Dennis R., and Cohn, Timothy A., 1988, Estimation of Descriptive Statistics for Multiply Censored Water Quality Data, Water Resources Research 24(12), p. 1997-2004, December 1988.

Helsel, Dennis R., and Gilliom, Robert J., 1986, Estimation of distributional parameters for censored trace-level water-quality data. II: Verification and Applications, Water Resources Research, v. 22, p. 147-155.

Gilliom, Robert J., and Helsel, Dennis R., 1986, Estimation of distributional parameters for censored trace-level water-quality data. I. Estimation Techniques, Water Resources Research, v. 22, 135-146.

Application of methods for censored data to environmental sciences:

Akritas, M.G., 1994, Statistical analysis of censored environmental data: Chapter 7 of the Handbook of Statistics, Volume 12, edited by G.P. Patil and C. R. Rao. North-Holland, Amsterdam.

Lee, L, and D. R. Helsel, 2005. Baseline models of trace elements in drinking water of the United States. Applied Geochemistry 20, 1560-1570.

She, N., 1997, Analyzing censored water quality data using a non-parametric approach: Journ American Water Resources Assoc. 33, 615-624.

Millard, S.P. and S. J. Deverel, 1988, Nonparametric statistical methods for comparing two sites based on data with multiple nondetect limits: Water Resources Research 24, 2087-2098.

References to data used in this workshop:

Eckhardt, D.A., W.J. Flipse and E.T. Oaksford, 1989, Relation between land use and ground-water quality in the upper glacial aquifer in Nassau and Suffolk Counties, Long Island NY: U.S. Geological Survey Water Resources Investigations Report 86-4142, 26 p.

Golden, N. H., B. A. Rattner, J. B. Cohen, D. J. Hoffman, E. Russek-Cohen, and M. A. Ottinger, 2003, Lead accumulation in feathers of nestling black-crowned night herons (Nycticorax nycticorax) experimentally treated in the field: Environmental Toxicology and Chemistry 22, 1517-1524.

She, N., 1997, Analyzing censored water quality data using a non-parametric approach: Journ American Water Resources Assoc. 33, 615-624.

For many other references to the work of others, see ***Nondetects And Data Analysis***.