The Mystery of Nondetects

How Censored Data Methods Work

Dennis R. Helsel

PracticalStats.com

Objectives of the 'Mystery' webinar

1. To demonstrate how censored data methods work without substituting a value for nondetects
2. To motivate you to use censored data methods by increasing your understanding of them, and trust in them
3. To highlight one of the many aspects of the new online course Nondetects And Data Analysis now available at http://practicalstats.teachable.com

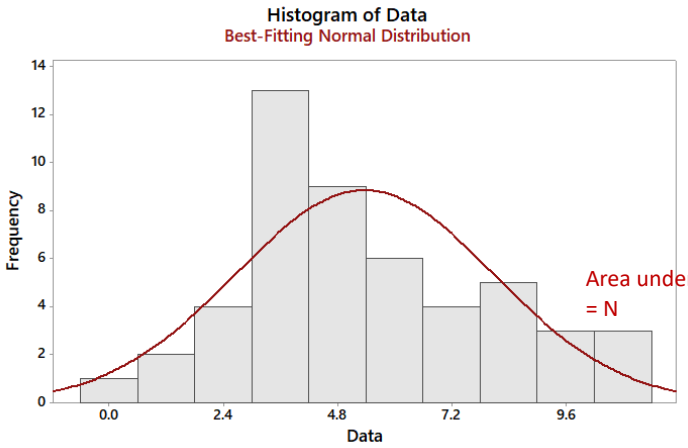
Outline: How Censored Data Methods Work

1. Maximum Likelihood (MLE) using pdfs
2. Kaplan Meier using cdfs
3. ROS using Q-Q plots
4. Testing with 1 DL using ranks and familiar methods
5. Testing with multiple DLs using scores

1. pdf: Probability Density Functions

The familiar "bell shaped curve" of the normal distribution

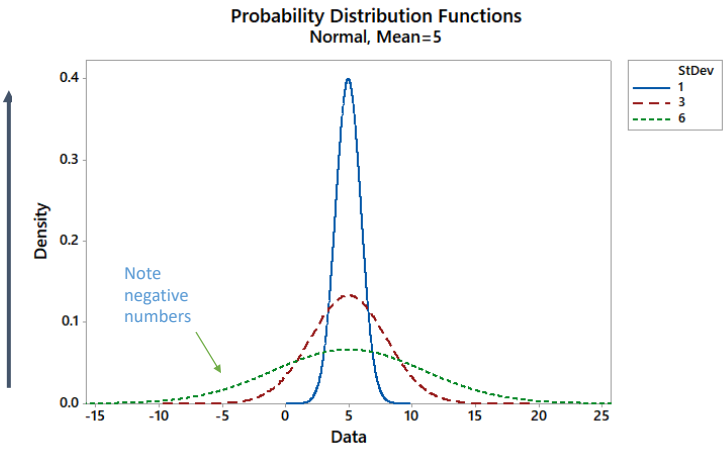
Frequency scale: Total = N. Or divide by N to get "density", the % of the observations. Total percentage (sum of area of the bars) =1



1. pdf: Probability Density Functions

Maximum Likelihood: Vary the mean and standard deviation to fit to the data

Density is the percent of the observations. Area under each curve = 1

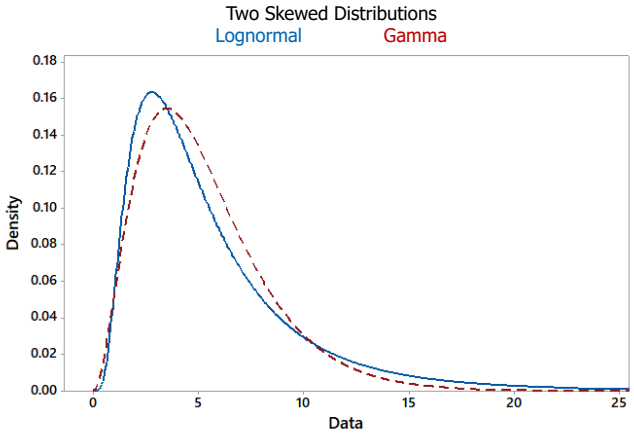


5

1. pdf: Probability Density Functions

Maximum Likelihood: You must decide which type of distribution to use

Density is the percent of the observations. The area under each curve = 1



Two Common Skewed Distributions: Lognormal and Gamma

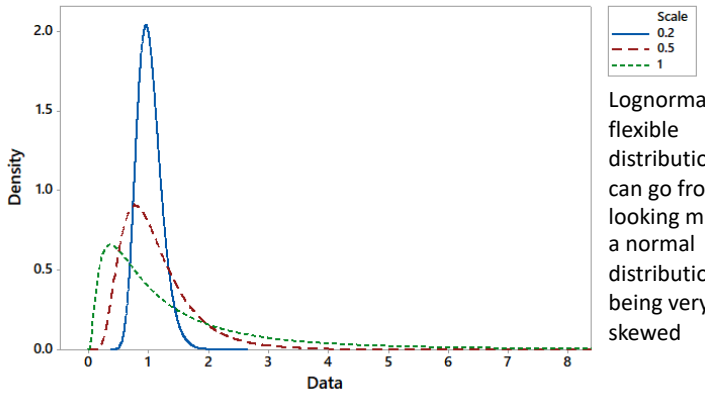
Lognormal has slightly higher probability of 'high outliers'

6

1. pdf: Probability Density Functions

Maximum Likelihood (MLE) optimizes the standard deviation to find the best fit to the data

Density is the percent of the observations. The area under each curve = 1

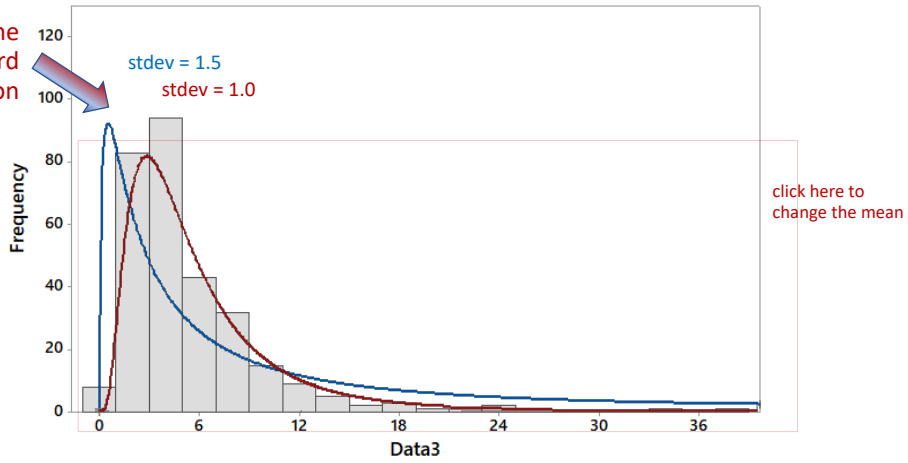


Lognormal is a flexible distribution that can go from looking much like a normal distribution to being very, very skewed

1. MLE: Fit distribution to data

log-likelihood = error of the fit. Optimization: changing the mean and std dev to find minimum error

Changing the standard deviation

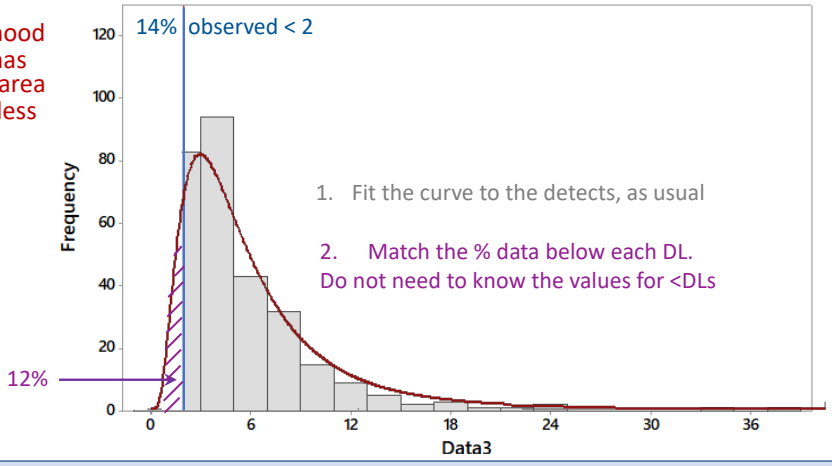


click here to change the mean

1. MLE: Fit distribution to censored data

Minimize the log-likelihood. For censored data it has two parts, one for detects and one for nondetects.
We don't have values for the lowest 14% of the data, only knowing that they are <2.

Maximum Likelihood (MLE) best fit has 12% of the total area under its curve less than 2.



1. MLE: Fit distribution to censored data

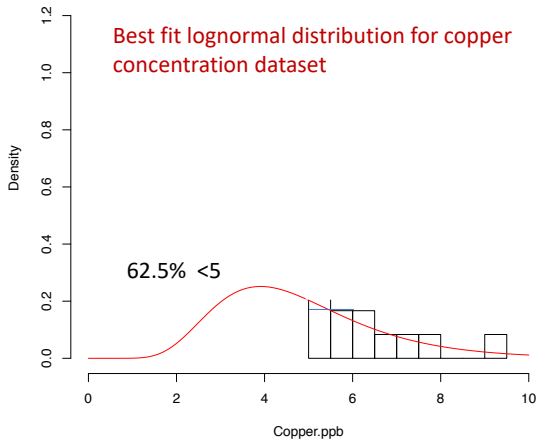
NADA package

```
> cenmle(Copper.ppb, Censored)
       n     n.cen    median      mean        sd
24.000000 15.000000  4.508118  4.841466  1.895949
```

EnvStats package

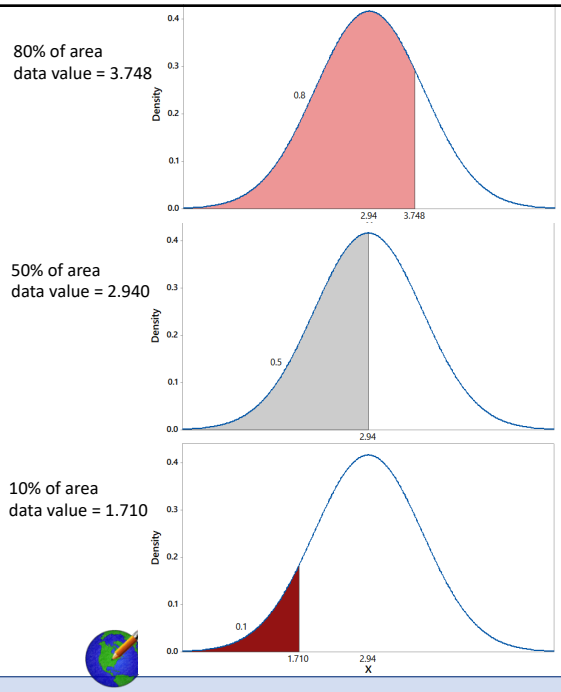
```
> elnormAltCensored(Copper.ppb, Censored, ci=TRUE, ci.type="upper")
Results of Distribution Parameter Estimation Based on Type I Censored Data
------------------------------------------
Assumed Distribution:            Lognormal
Censoring Level(s):              5
Estimated Parameter(s):          mean = 4.8414656
                                 cv   = 0.3916064
Estimation Method:               MLE
Sample Size:                     24
Percent Censored:                62.5%
Confidence Interval Method:      Profile Likelihood
Confidence Level:                95%
Confidence Interval:             UCL = 5.612114
> cu.dist <- elnormCensored(Copper.ppb, Censored, ci=TRUE, ci.type="upper")
> cu.param <- cu.dist$parameters
> hist(Copper.ppb, xlim = c(0, 10), prob = TRUE)
> curve(dlnorm(x, mean=cu.param[1], sd = cu.param[2]), add=TRUE, col = "red")
```

Best fit lognormal distribution for copper concentration dataset



1. MLE Summary

- Must assume a distribution
- No substituted values are used
- Nondetects affect the computations of mean, standard deviation, and percentiles through their observed percentage of values below each DL (the percentile probability of each DL)
- MLE works best with 50 or more observations, given the skewness of environmental data

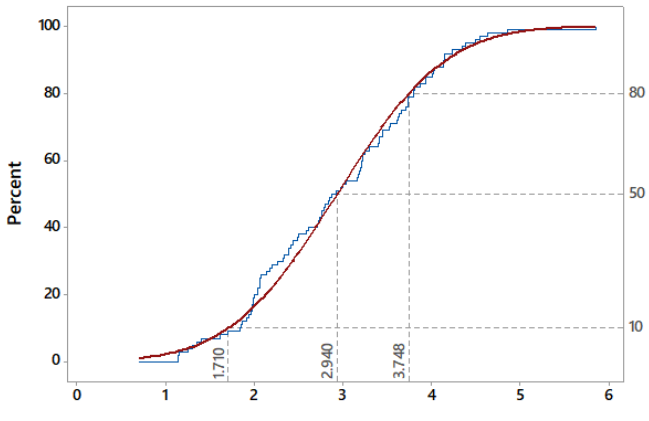


2. cdf: Cumulative Distribution Functions

A cdf simply accumulates the probabilities of a pdf

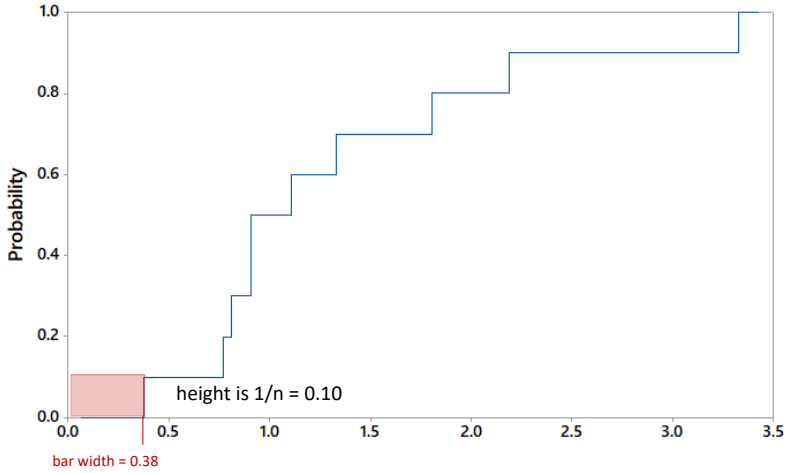
Step function plots the observed data

Smooth curve is the theoretical fitted distribution



2. cdfs and the Kaplan-Meier Method

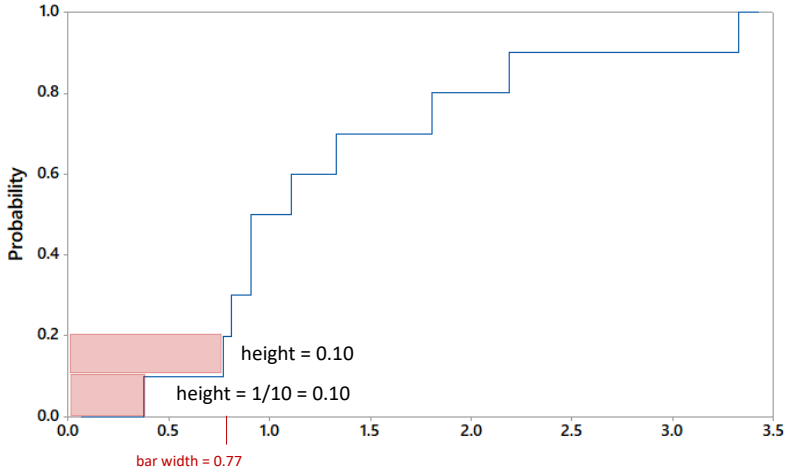
No NDs to start with. n=10
3.33
2.19
1.81
1.33
1.11
0.91
0.91
0.81
0.77
0.38



13

2. cdfs and the Kaplan-Meier Method

No NDs to start with. n=10
3.33
2.19
1.81
1.33
1.11
0.91
0.91
0.81
0.77
0.38

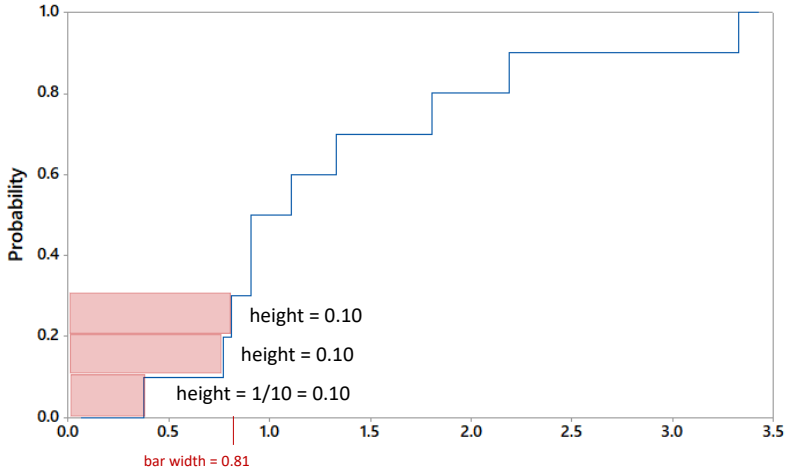


14

2. cdfs and the Kaplan-Meier Method

No NDs to start with. n=10
3.33
2.19
1.81
1.33
1.11
0.91
0.91
0.81
0.77
0.38

… and so on, until



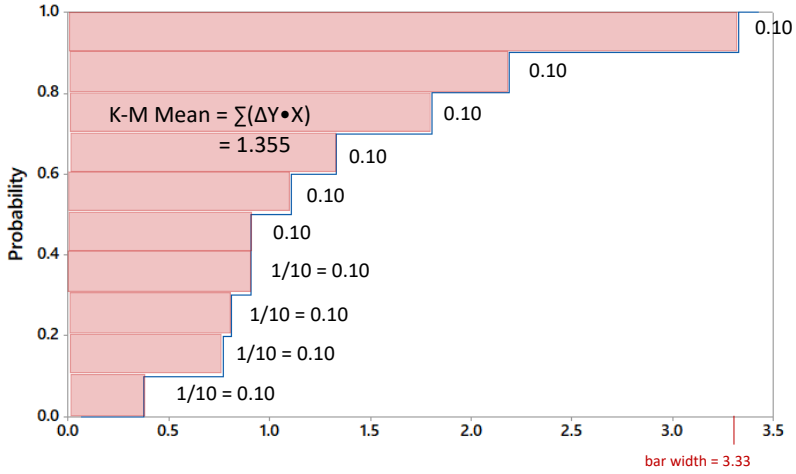
15

2. cdfs and the Kaplan-Meier Method

No NDs to start with. n=10
3.33
2.19
1.81
1.33
1.11
0.91
0.91
0.81
0.77
0.38

Mean = ∑data/10
= ∑(0.10•data values)
= ∑(ΔY•X)

The mean equals the sum of the height *width for the bars, or the area in color to the left of the cdf, down to x = 0



16

2. cdfs and the Kaplan-Meier Method

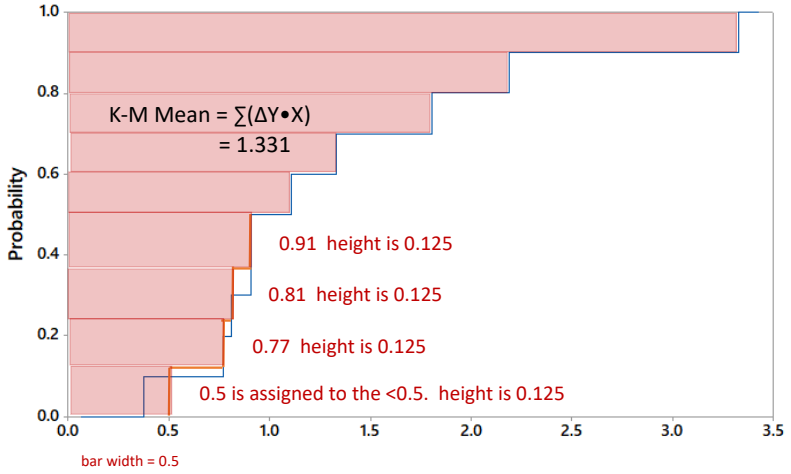
- Kaplan-Meier takes each nondetect and reassigns its probability to the detects that occur below it.
- This assumes the observed shape of the data below the highest nondetect is the best indicator of the shape of the data in that region
- Observations below the lowest DL are treated as detected values, keeping their probabilities. The DL value is usually assigned to the < lowest DL data.
- If there were only 1 DL, this means that Kaplan-Meier in essence substitutes the DL for nondetects. So it is less useful than other methods when there is only 1 DL.

2. cdfs and the Kaplan-Meier Method

Concentrations WITH NDs. n=10.

3.33
2.19
1.81
1.33
1.11
~~0.91~~ <1 no bar, redistributes this 0.10
0.91 + 0.025
0.81 + 0.025
0.77 + 0.025
~~0.38~~ <0.5 + 0.025

The mean still = ∑(ΔY•X), the area to the left of the adjusted cdf down to X = 0.



2. Kaplan-Meier Method

an example of when it doesn't work well
the Copper concentrations

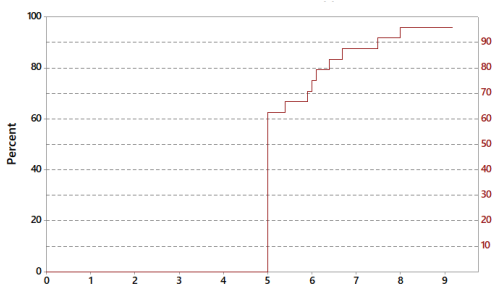
Using R (EnvStats package): Copper Background dataset

```
> enparCensored(Copper.ppb, Censored, ci=TRUE, ci.method =
"bootstrap", n.bootstraps = 5000)

Based on Type I Censored Data

----------------------------------------------

Censoring Level(s):              5    (only 1 DL)
Estimated Parameter(s):          mean    = 5.6750000
                                 sd      = 1.1177544
                                 se.mean = 0.1457466
Estimation Method:               Kaplan-Meier
Sample Size:                     24
Percent Censored:                62.5%
Confidence Interval Method:      Bootstrap
Median:  <5
```



There is no model for this nonparametric method for how data descend below the lowest detection limit. Kaplan-Meier assigns all of the probability for nondetects at the lowest DL to the DL itself. This produces an upward bias (we know they are <DL but are counted as at the DL). So with 1 DL the K-M estimate of mean is too high. With multiple DLs this isn't as much of an issue.

ROS uses a model for how data descend below the lowest DL. In return, you get a relatively unbiased estimate of the mean, and a numerical value for the median even when there are >50% NDs.

 19

2. Kaplan-Meier Summary

- No assumption of a distribution is needed
- No substituted values are used
- Nondetects affect the computations of mean, standard deviation, and percentiles through their observed percentage of values below each DL (the percentile probability of each DL)
- K-M works best with 2 or more detection limits (technically, it works well when the proportion of data below the lowest DL is small)

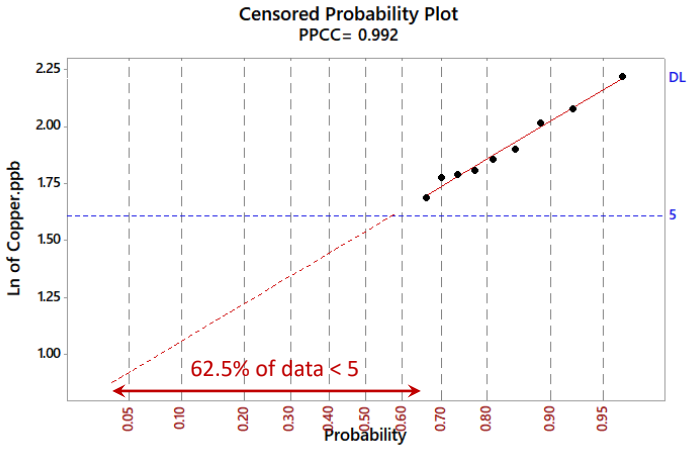
 20

3. Q-Q plot (probability plot)

Copper concentrations

Q-Q plots sometimes have a nonlinear scale of Probability (percentiles) for the normal distribution on one axis, as shown here. The mean is the Y axis value at x=0.50

More commonly, the distribution is represented on the x axis by the linear Normal Quantiles scale, where the mean = 0 and the scale is in standard deviations above and below the mean (see next slide)



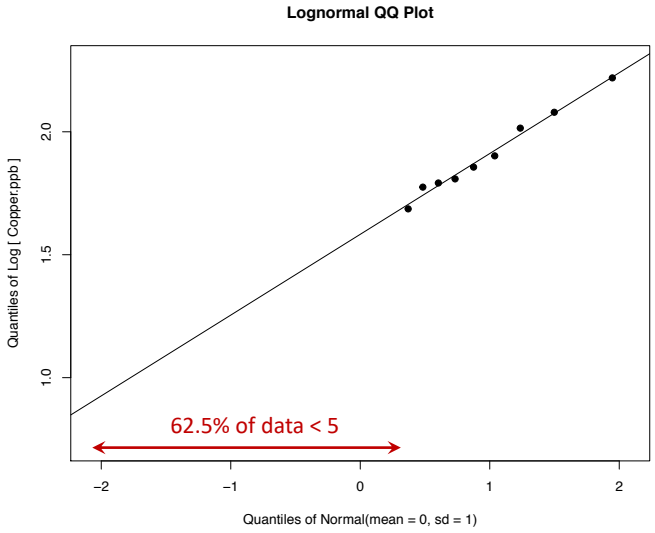
3. Censored Q-Q plot (probability plot)

Copper concentrations:

Q-Q plots represent the distribution as a straight line.

Detected obs are plotted as points. Censored data are not (you don't know a unique value for them)

The straight line is fit to both the detected observations and (by plotting them at their correct percentiles) the proportion of data below each DL (here 62.5% <5).

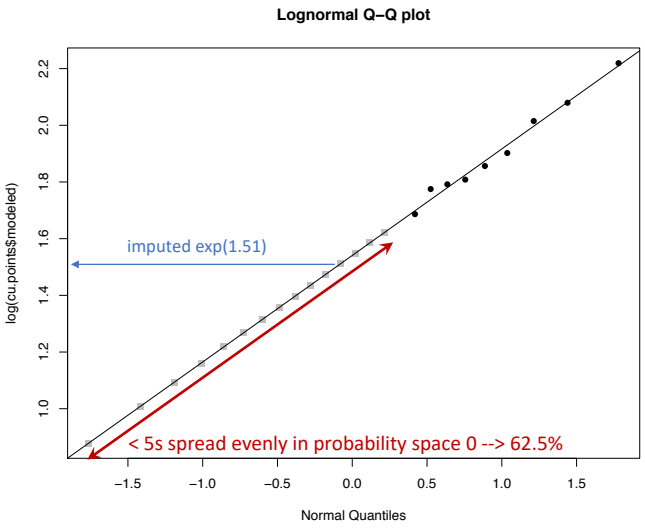


3. ROS using a Q-Q plot

Regression on Order Statistics (ROS):
Copper concentrations

Regression line from detected values is extended down to the low end of the distribution to model placeholders for nondetects. These are placed at the Probability / normal quantiles for the set of values <5, as the probabilities for nondetect data are corporately known, though their individual concentrations are not known

Placeholders are shown here as gray squares just to illustrate the process. These are spread evenly in probability space from 0 to 62.5%, NOT in concentration space from 0 to the DL. In that way they follow the distribution for the data. Y values are imputed from the lognormal model (one is at exp(1.51).



3. ROS: robust Regression on Order Statistics

Imputed

2.408 ← <5
2.740 ← <5
2.986 ← <5
3.193 ← <5
3.381 ← <5
3.555 ← <5
3.722 ← <5
3.885 ← <5
4.045 ← <5
4.206 ← <5
4.367 ← <5
4.533 ← imputed exp(1.51) <5
4.703 ← <5
4.879 ← <5
5.065 ← <5
& detects: 5.4 5.9 6.0 6.1 6.4 6.7 7.5 8.0 9.2

robust ROS: Nonparametric in regards to the detects
Parametric in regards to the nondetects

Note that it is possible to impute a value higher than the DL (5 for these data). This is not a problem -- true concentrations slightly higher than the DL have close to a 50% chance of being reported as <DL by the laboratory.

If there are no NDs, ROS gives the same mean and stdev as usual.

ROS summary statistics
mean (of imputed + detected data) = 4.95
stdev = 1.74

3. ROS: robust Regression on Order Statistics

Using R (NADA package)

```
> mean(cenros(Copper.ppb, Censored))
[1] 4.95288
```

Assumes lognormal distribution for nondetects

3. ROS Summary

- A distribution must be assumed for modeling the nondetect portion of the distribution. Lognormal is the most common.
- No substituted values are used
- Nondetects affect the computations of mean, standard deviation, and percentiles through their observed percentage of values below each DL (the percentile probability of each DL)
- ROS is the most generally applicable method of the 3. It works well in most situations

Comparing the three methods

	MLE	K-M	ROS
Gives the same mean as ∑/N when no NDs	No	Yes	Yes
Assumes a distribution?	Yes	No	Partial
Sensitivity	High	None	Low
Use with 1 or more DLs?	Yes	Only multiple DLs	Yes
Useful with small datasets?	No	Yes	Yes
Better than substitution?	If enough data to find correct distribution	Yes	Yes
Mean of copper data	4.841 (lognormal)	5.67 (1 DL: biased high)	4.95

Blue is generally better than red on this chart, indicating that ROS is the most generally applicable method.

 27

4. Testing Group Differences with 1 DL Using Ranks

Mean = ∑ obs /n . The mean is a standardized total.

(Percentile) probability = rank /n(+1) The probability is a standardized rank.

- Ranks are a measure of relative order in a dataset. Tests based on ranks are tests for difference in percentiles

	<5	<5	<5	<5	<5	<5	<5	6	8	11	17	20	22	
Ranks:	1	2	3	4	5	6	7	8	9	10	11	12	13	If we knew their values
Ranks:	4	4	4	4	4	4	4	8	9	10	11	12	13	Since we don't & so they are tied

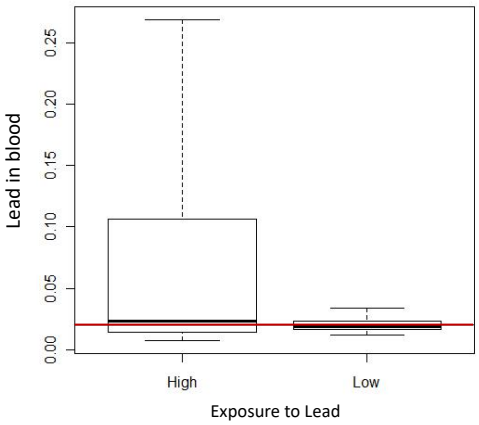
Why 4? It is the mean of the ranks 1 - 7, and so preserves their sum (their relative weight in the dataset)

 28

4. Testing Group Differences with 1 DL Using Ranks

- One detection limit at 0.02
- The % <0.02 is less in the High group (below 50%) than in the Low group (>50%)
- High group: fewer ND tied low ranks, plus the higher ranks for the detected data, result in a significant difference between the groups
- No substitution of values for nondetects required. Their info is shown by the difference in % NDs
- With all NDs represented by tied ranks, the usual Wilcoxon rank-sum test:

```
> wilcox.test(Golden$LT02 ~ DosageGroup)
```



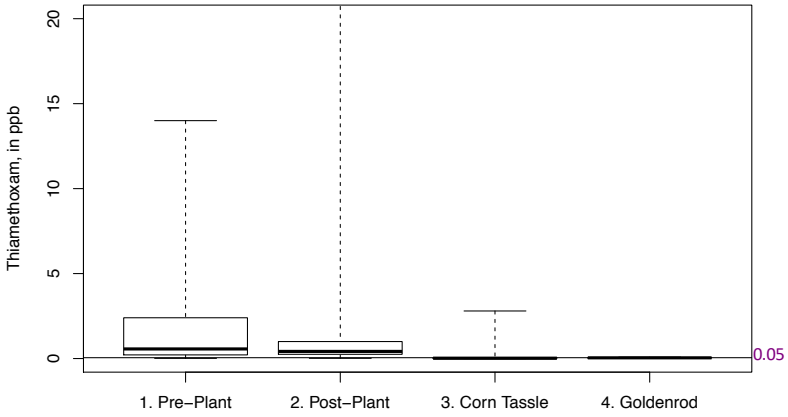
5. Testing with Multiple DLs using Scores

- Nonparametric test similar to the Wilcoxon rank-sum test
- Scores are ranks for the detected data, adjusted for censoring - similar to K-M percentiles
- R version is called the "Peto-Peto" test
- Similar tests called Peto-Prentice, HF1, Generalized Wilcoxon and Gehan
- Cendiff command in NADA for R

5. Testing with Multiple DLs using Scores

```
> cenboxplot (Thiamethoxam, as.logical(ThiaCens),
as.factor(SamplingEvent), ylab = "Thiamethoxam, in ppb", xlab =
"Sampling Event", log=FALSE, ylim =c(0,20))
```

command from the NADA package

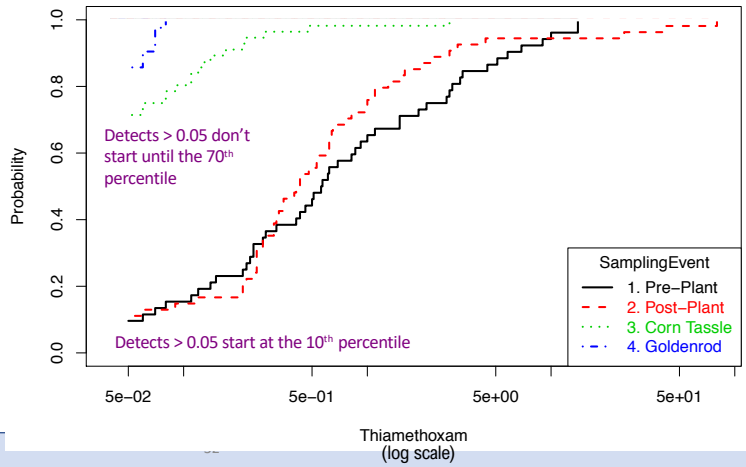


31

5. Testing with Multiple DLs using Scores

```
> plotcdf (Thiamethoxam, ThiaCens, SamplingEvent, logscale=TRUE)
```

- CDFs using Kaplan-Meier percentiles
- Higher quantiles for a given probability → cdf plots to the right.
- Pre- and Post- Plant are similar. All other comparisons appear different.



5. Testing with Multiple DLs using Scores

```
> cen1way (Thiamethoxam, ThiaCens, SamplingEvent)
     Oneway Peto-Peto test of CensData: Thiamethoxam   by Factor: SamplingEvent
      Chisq = 127   on 3 degrees of freedom     p = 2.35e-27

     Pairwise comparisons using Peto & Peto test
data:  CensData and Factor
               1. Pre-Plant 2. Post-Plant 3. Corn Tassle
2. Post-Plant  0.416        -             -
3. Corn Tassle 6.5e-15      6.5e-15       -
4. Goldenrod   6.5e-15      7.1e-15       0.055
```

For more info on this test, see our "Testing Groups for Data with Multiple DLs" webinar on practicalstats.teachable.com

Pre-A	Post-A	Corn Tassle B	Goldenrod B

Summary: The Mystery of Nondetects

- Censored data methods use the information in the detected values, plus in the % of data below each DL, to compute statistics and hypothesis tests
- No values for the NDs are needed
- ROS imputed values are "air pillows", not estimates of what was in the samples. Do not use individually, only collectively. Do not use in hypothesis tests or regression
- With 1 DL, familiar nonparametric tests like the rank-sum and Kruskal-Wallis tests can be used. All data - in either group - below the DL are tied with each other at the lowest rank
- Nonparametric hypothesis tests (Peto-Peto) can be computed using rank/score methods. They are appropriate for data with 1 or multiple DLs
- Find them in survival analysis or reliability analysis sections of stat software

Our Next Webinar

Tuesday July 16th 11 am Mountain time

* No webinar in June
- Topic TBD
- Sign up for our newsletter/announcement list to get the registration link emailed to you. Respond to the survey you'll get in a few minutes to opt into the list, or send email to ask@practicalstats.com
- Or check our webinars page periodically at http://practicalstats.com/training/webinar.html to register for it.

 35

This 'Mystery of Nondetects' webinar will be available Thursday for streaming

- at our Online Training Site

  http://practicalstats.teachable.com/

Let colleagues who missed it know about it.

 36

Thank you for attending

- Much of the material is based on the book Statistics for Censored Environmental Data Using Minitab and R, 2nd Edition by Dennis Helsel (2012).
- All of the material, plus much more, is now available in our online course Nondetects And Data Analysis, on our Training Site.
- All opinions are my own and do not represent those of anyone else you can think of.

- Questions?

Get in touch!

Dennis Helsel ask@practicalstats.com

Courses & free webinars at our Training Site: http://practicalstats.teachable.com