Nondetects And Data Analysis:
Estimating Descriptive Statistics

Dennis R. Helsel, Ph.D
Practical Stats

Estimating Descriptive Statistics

Statistics using percentiles

– median, IQR

Survival analysis methods

– Statistics using ROS, Kaplan-Meier, Maximum likelihood estimation

2

Estimating Descriptive Statistics

1. Statistics using percentiles

Multiple DLs -- the Oahu data set

censor at highest RL, compute median, percentiles

0.5	0.5	0.5	0.6	0.7	0.7	<0.9	0.9
<1.0	<1.0	<1.0	<1.0	1.5	1.7	<2.0	<2.0
<2.0	<2.0	<2.0	<2.0	<2.0	<2.0	2.8	3.2

<2.0	<2.0	<2.0	<2.0	<2.0	<2.0	<2.0	<2.0
<2.0	<2.0	<2.0	<2.0	<2.0	<2.0	<2.0	<2.0
<2.0	<2.0	<2.0	<2.0	<2.0	<2.0	2.8	3.2

Median = <2.0 IQR = <2.0

Simple 1 DL data set

with one DL use the data as is

<10	<10	<10	<10	<10	<10
<10	12	15	19	21	21
24	26	28	32	39	43

n = 18. Median is the value at the 19/2 = 9.5th position.

Median = (15+19)/2 = 17

IQR = 26.5 - <10 = [>16.5 to 26.5]

3

2a. Survival analysis -- MLE

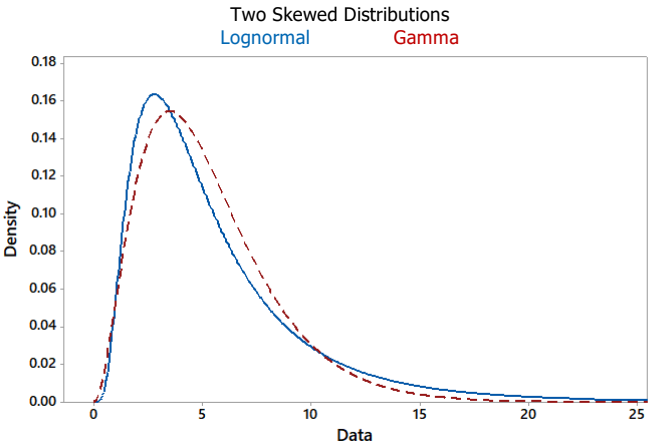
- Maximum Likelihood Estimation -- parametric method.
- You designate which distribution to use (Lognormal is most common for environmental data)
- Finds estimates of mean and std dev, or slope and intercept of regression, that are the most likely to have produced the data you have observed – using both the detected values and the proportion of data below each detection limit.

4

First task: Choose the distribution

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

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



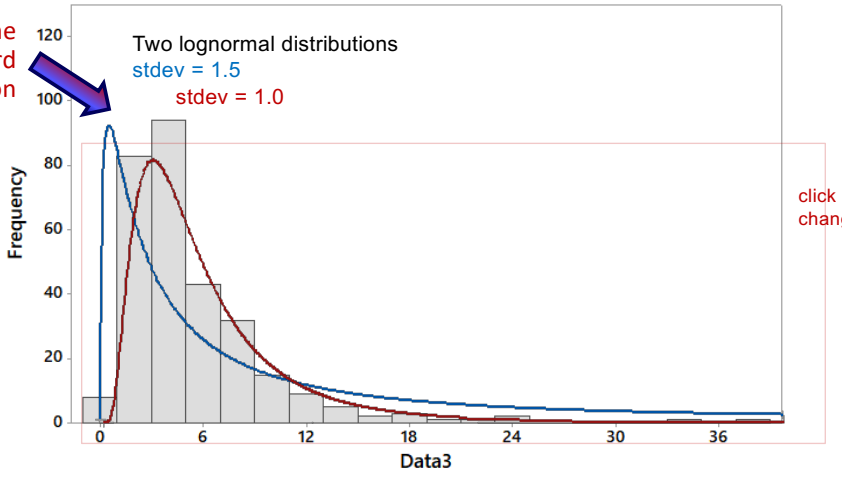
From Section 4, choose the distribution based on PPCC (max) or BIC (min)

5

MLE chooses the best fit

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

Changing the standard deviation

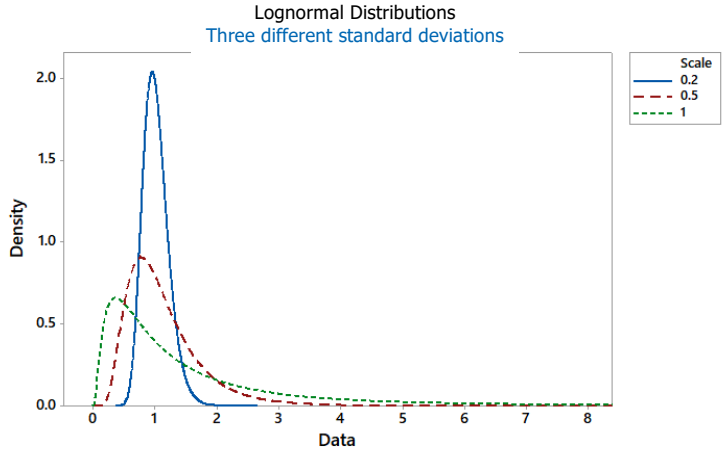


click here to change the mean

6

MLE chooses the best fit standard deviation

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



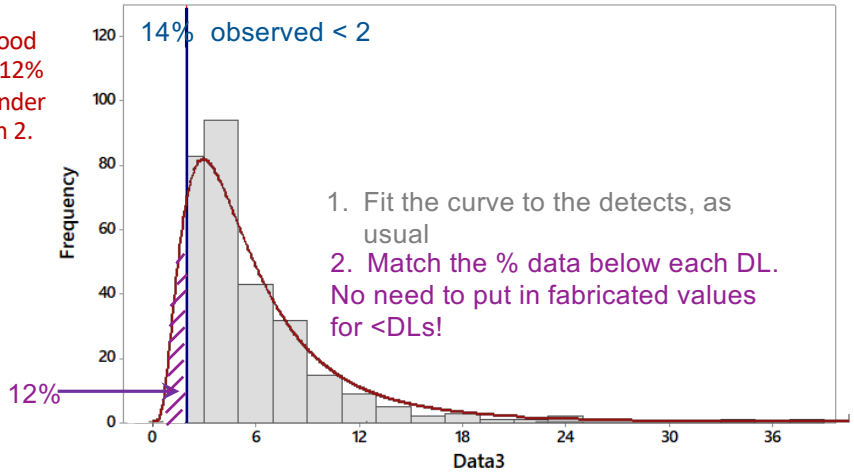
Lognormal is a flexible distribution that can go from looking much like a normal distribution to being very, very skewed

7

MLE fits 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. Fit the curve to the detects, as usual
2. Match the % data below each DL. No need to put in fabricated values for <DLs!

8

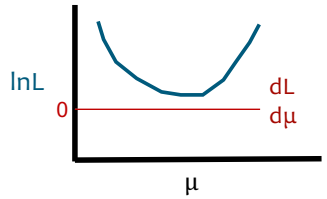
How MLE Works

The log-likelihood function lnL(μ, σ)

L(μ, σ) = piece for detects + piece for %nondetects

Minimize lnL by changing mu and sigma to get the minimum lnL. In an ideal world, lnL = 0

$$\frac{d(\ln L[\mu,\sigma])}{d\mu} = 0$$



1. MLE: Fit distribution to censored data

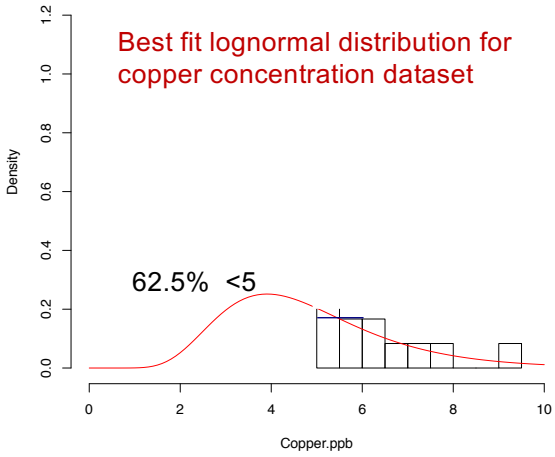
NADA package

```
> 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



10

2a. MLE Summary

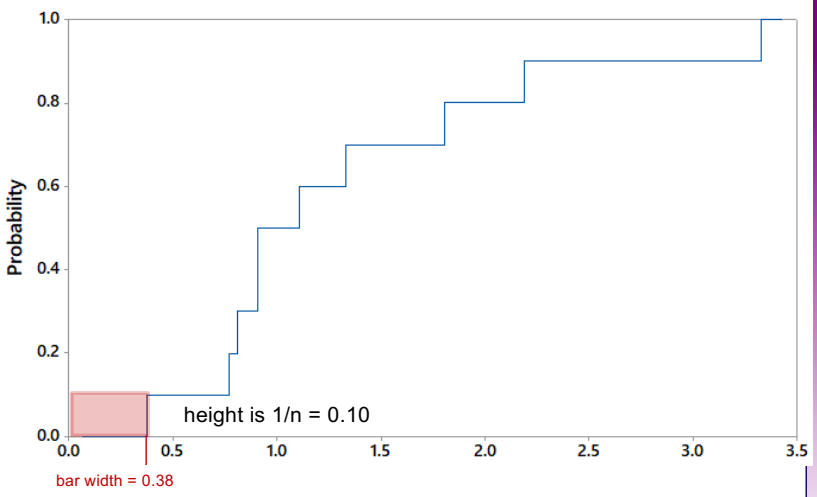
- Must assume a distribution
- No fabricated values are used
- Nondetects affect the computations of mean, standard deviation, and percentiles through their observed percentage of values below each DL (the probability ≤ each DL)
- MLE works best with 50 or more observations, given the skewness of environmental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

11

2b. cdfs and the Kaplan-Meier Method

No NDs to start with. n=10

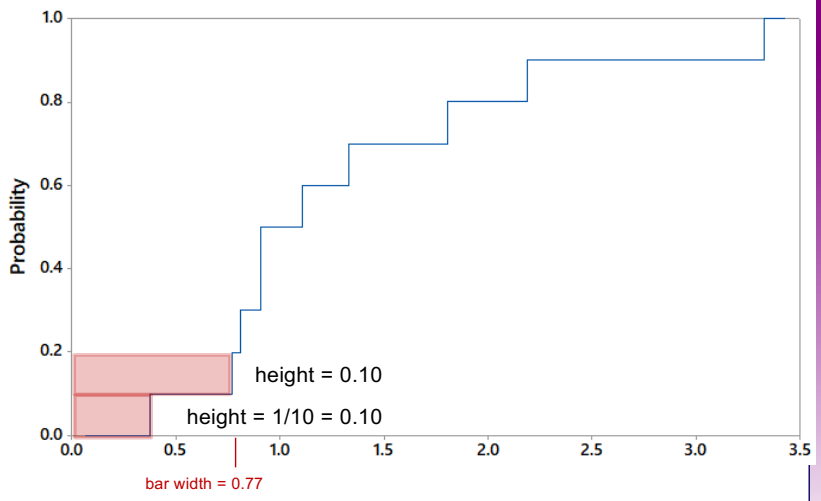
3.33
2.19
1.81
1.33
1.11
0.91
0.91
0.81
0.77
0.38



12

2b. 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



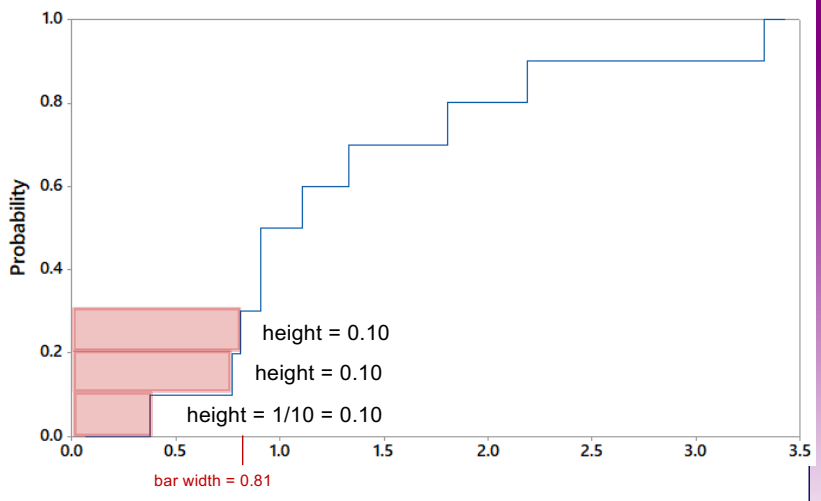
13

© PracticalStats.com

2b. 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



14

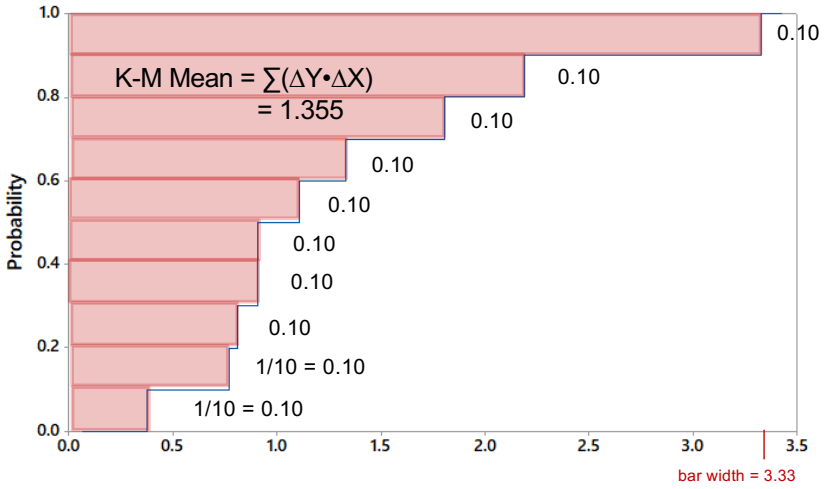
2b. cdfs and the Kaplan-Meier Method

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•X_i)
= ∑(ΔY•X_i)

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



15

© PracticalStats.com

2b. cdfs and the Kaplan-Meier Method

It takes each nondetect and reassigns its probability to the detects that occur below it. This assumes the observed shape of the data below a detection limit is the best indicator of the shape of the data in that region

Observations below the lowest DL are treated as detected values at the DL, keeping their probabilities. Kaplan-Meier is a nonparametric procedure. No model of how to extrapolate below the lowest DL

If there were only 1 DL, this means that Kaplan-Meier in essence substitutes the DL for nondetects. So it is less useful than ROS when there is only 1 DL. K-M works well for multiple DLs, especially when there is not a good fit to a specific distribution, or when there is a small proportion of data below the lowest DL.

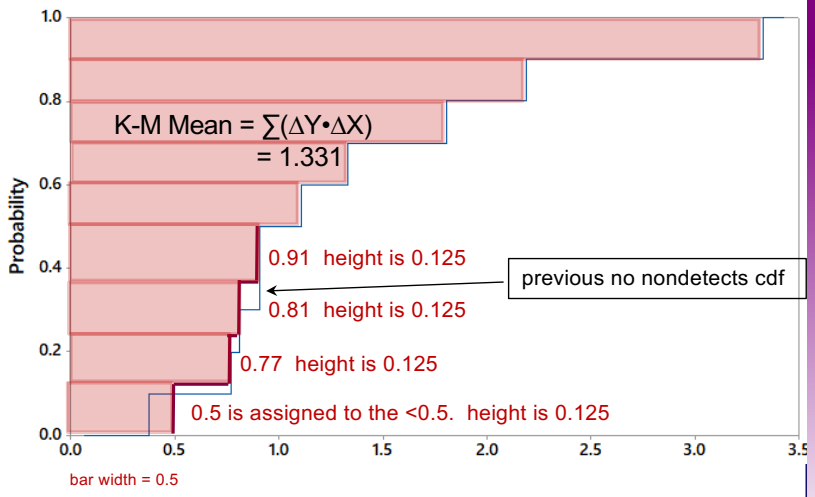
16

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 = $\sum(\Delta Y \bullet X_i)$, the area to the left of the adjusted cdf down to X = 0.



17

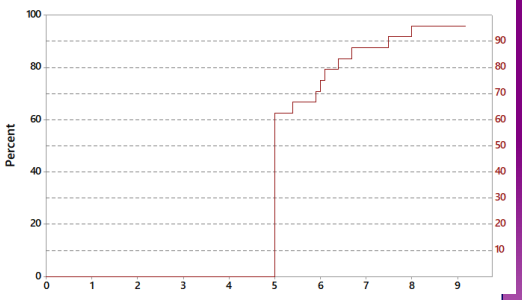
2. Kaplan-Meier

An example of when KM doesn't work well
(large % of data below the lowest DL)

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       = 1.1777544
                                 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 few data below the lowest DL (as often is for 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.

18

2b. 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 probability of ≤ each DL)
- K-M works best with a small % below the lowest detection limit (as is often true for multiple DLs)

19

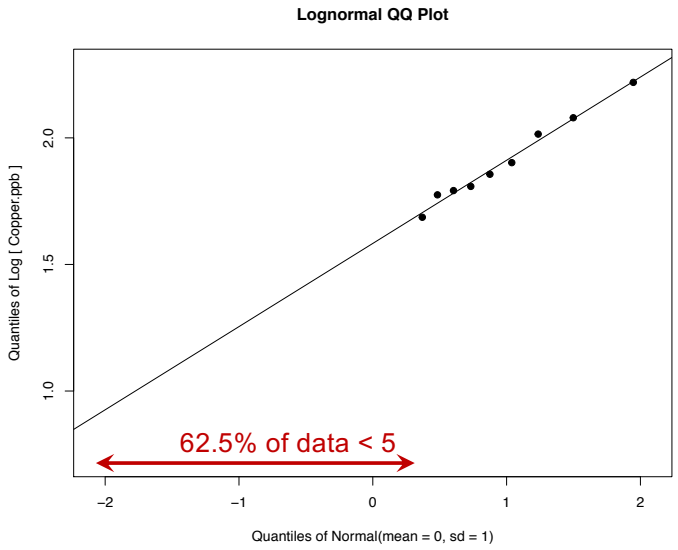
2c. ROS and the Censored Q-Q plot

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).



20

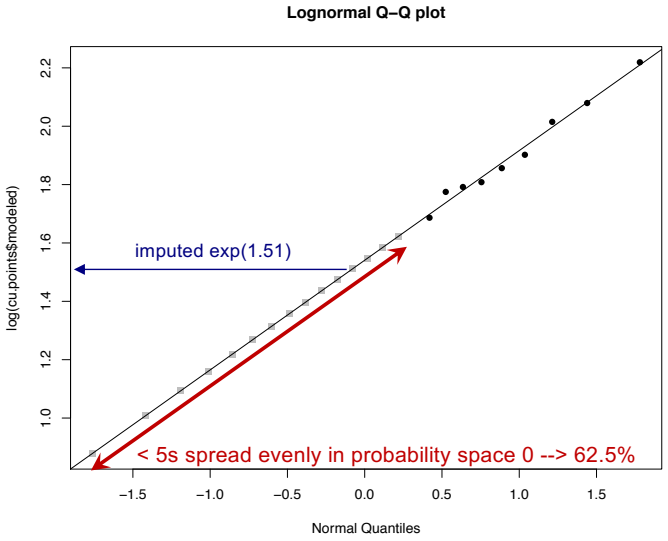
2c. ROS and the Censored Q-Q plot

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 <DL.

The probabilities for nondetect data are corporately known, though their individual concentrations are not known. Therefore unique values cannot be assigned to each observation.

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).



21

2c. 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 a <DL by the laboratory.

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

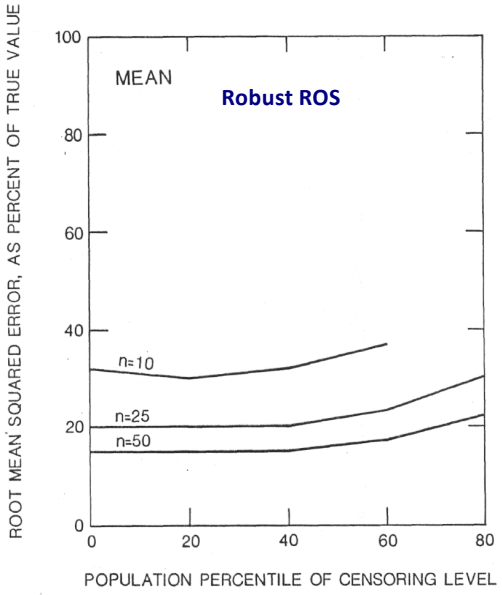
ROS summary statistics

mean (of imputed + detected data) = 4.95

stdev = 1.74

22

If a good estimation method is used, for up to 60% censoring the estimate for the mean has no more error than if there were no censoring



23

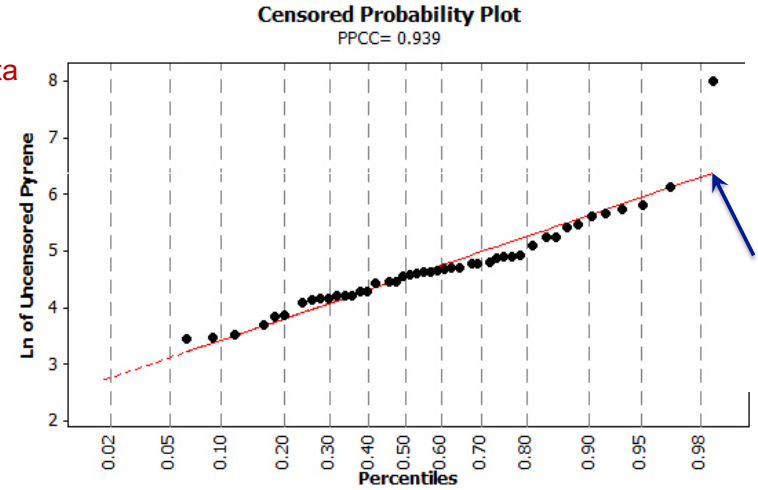
2c. ROS Summary

- A distribution must be assumed for modeling the nondetect portion of the distribution. Lognormal is the most common.
- No substituted values are used (imputed is defined as data that come from a model, not just a fraction of the DL)
- Nondetects affect the computations of mean, standard deviation, and percentiles through their observed percentage of values below each DL (the probability of ≤ each DL)
- ROS is the most generally applicable method of the three. It works well in most situations

24

ROS, K-M vs MLE

ShePyrene data



KM, ROS use actual value

MLE uses the model, the estimated percentile on the line

25

Comparing the three methods

	MLE	K-M	ROS
Gives the same mean as ∑x/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 (62% < 1 DL: biased high)	4.95

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

26

Helsel's Recommendations

For more than 50 detected obs that follow a standard distribution (gamma, lognormal):
use MLE.

For fewer than 50 detected observations:
With multiple DLs, use Kaplan-Meier or ROS.
With one DL, use ROS.

Otherwise, use ROS.

27

Methods for censored data

Method	Parametric	Nonparametric
Descriptive stats	MLE ROS	Kaplan-Meier
Intervals	Bootstrapping MLE	Bootstrapping K-M
Paired Data	CI on differences by MLE	PPW
2 Indep Groups	MLE Regression on 0/1 factor	Generalized Wilcoxon
3+ Indep Groups	MLE Regression on 0/1 factor	Generalized Wilcoxon
Correlation	Likelihood R by MLE	Kendall's tau
Regression	MLE Regression	ATS line

28