

NWQMC webinar series

Nov. 15, 2016

Nondetects and Data Analysis

Dennis R. Helsel
PracticalStats.com

Practical Stats
© 2016

What I'll present today:

Introduction and Terminology

1. What's wrong with substitution?
2. Plotting data with nondetects
3. Estimating summary stats with nondetects
4. Hypothesis tests with nondetects
5. Regression/correlation with nondetects
6. Available software

Conclusion

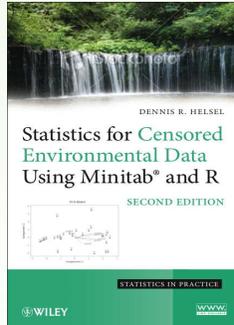
Practical Stats
© 2016

2

For more detail:

Statistics for Censored Environmental Data

by Dennis R. Helsel
Wiley (2012)



And the online course

Nondetects and Data Analysis

www.practicalstats.com/training/



Terminology: Nondetects

- are "real data" !
- "Less-thans", "qualified data"
- "Censored data" in statistics jargon
- left- or interval-censored values
- data known only to be below laboratory reporting (detection) limits

left-censored: <1

interval-censored: $[0 \text{ to } 1]$

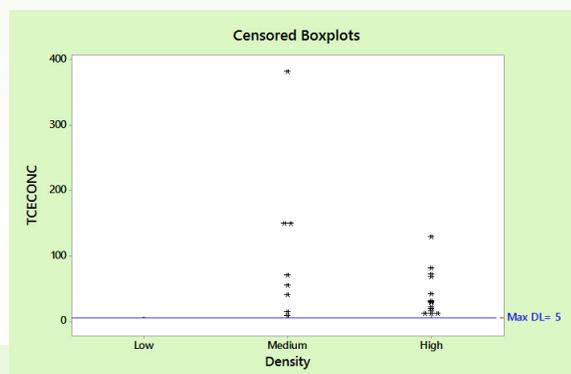
Reporting limit

- Laboratory reporting threshold
- General term
- Are several types of reporting limits, including
 - Detection limits
 - Quantitation limits
- Today I'll use "detection limit" and "reporting limit" interchangeably

1. What's wrong with substitution?

- Substitution of one-half or ($1/\sqrt{2}$) times the DL are most common
- Produces **invasive data** alien to the concentrations actually in samples
- Results in poor estimates and incorrect statistical tests
- Example: TCE concentrations in groundwater under 3 land-use groups

In Minitab:
%cbox c5 c4;
by c1.



ANOVA after substitution of 1/2 DL doesn't find a difference between the density groups

One-way ANOVA: Half DL versus Density

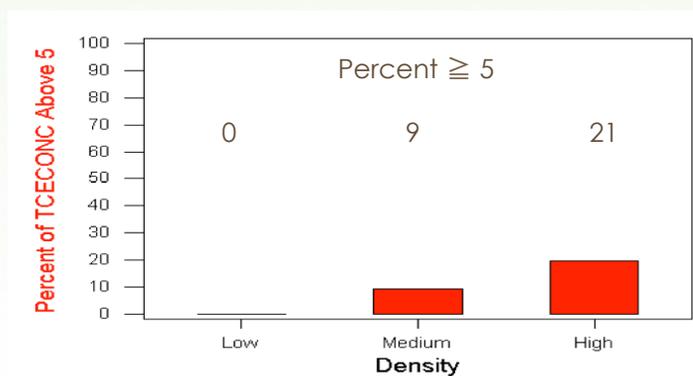
Null hypothesis All means are equal
 Alternative hypothesis At least one mean is different

Density	N	Mean	StDev	95% CI
Low	25	1.020	0.784	(-11.029, 13.069)
Medium	130	8.02	38.70	(2.74, 13.31)
High	92	7.76	19.61	(1.48, 14.04)

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Val
Density	2	1072	536.0	0.57	0.565
Error	244	228243	935.4		
Total	246	229315			

Binary: Contingency table test after re-censoring all values below 5 ug/L to <5



$$\chi^2 = 9.2$$

$$p = 0.001$$

(ANOVA p-value was 0.565)

% \geq 5 ug/L differs between groups, according to the contingency table. A simple nonparametric test.

Simple Nonparametric: Kruskal-Wallis test after re-censoring all values below 5 ug/L to <5

MTB > %censkw c5 c4 c1

Kruskal-Wallis Test on TCECONC.

Density-	N	Median	Ave Rank	Z
Low	25	-1.000	109.0	-1.11
Medium	130	-1.000	120.4	-0.84
High	92	-1.000	133.2	1.56
Overall	247		124.0	

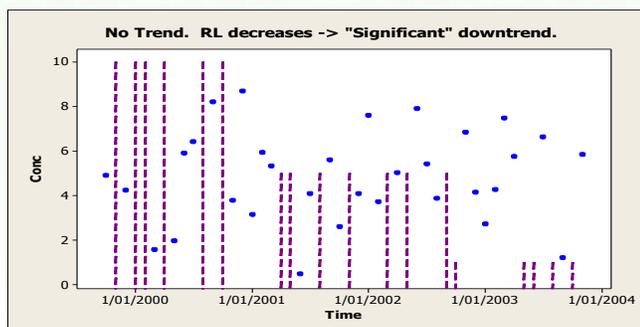
H = 9.17 DF = 2 **P = 0.010** (adjusted for ties)

Remember:
ANOVA p-value was 0.565

This should shock you! A strong signal (0.01) by using these methods, and a strong "no-signal" (0.56) by substituting 1/2RL. Never perform a parametric test like t-test or ANOVA after substitution!

Most common error when substituting for nondetects: trend analysis

- No trend in original data
- Dashed lines drawn up to the RL for nondetects
- The RL decreased over time
- After substitution, correlation with time (trend) becomes significant



Two other bad (and unfortunately, common) practices

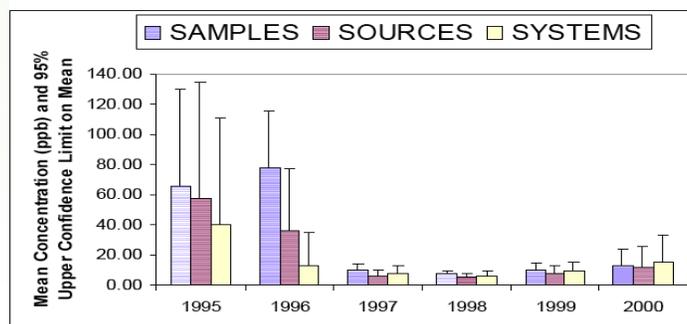
1. Deleting nondetects and just looking at detections
2. Comparing groups or trends using "% detections" when the RL changes

Summarizing only detections

The drop in the mean in 1997 does not necessarily show a decrease with time in the original data.

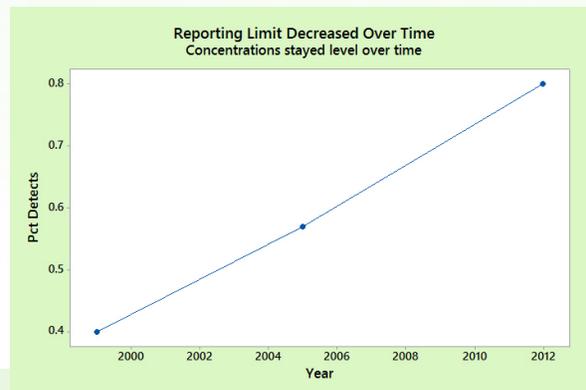
Perhaps the RL decreased in 96-97? Now smaller concentrations are "detects" and used to compute the mean.

Detected MTBE Concentration for Drinking Water Supplies in California in 1995-2000



Comparing % detections when the RL changes

- Comparing % detections only makes sense when the mix of RLs is identical across groups.
- In practice, this happens only when there is one reporting limit.
- Here the RL decreases with time. Concentrations stay the same. %detections go up.
- Instead, interpret the % >5 or another number. Use a consistent definition of “detect”.



Practical Stats
© 2016

Three Better Approaches

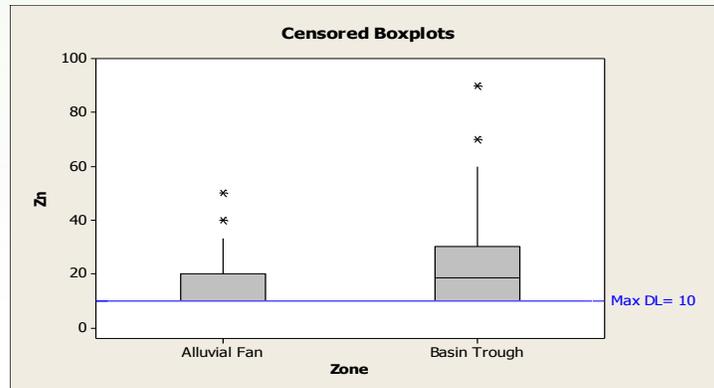
1. Binary methods
 - Simple. Data are either below or above a specified, single limit. Report % above; test difference in percentages with contingency tables; logistic regression.
2. Simple Nonparametric tests
 - Rank all data below highest RL as tied. Report percentiles; run simple nonparametric tests; Kendall's tau methods for correlation and regression.
3. Survival Analysis methods
 - More complicated. Can use data with multiple RLs without re-censoring to highest. Both parametric and nonparametric methods are available.

Practical Stats
© 2016

14

2. Plotting Data with Nondetects

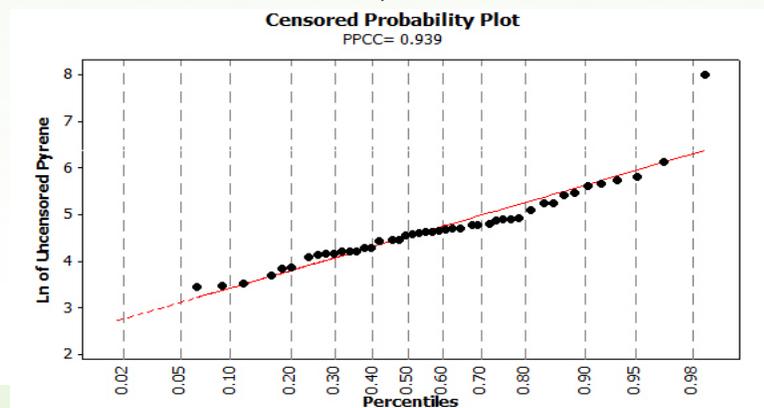
Censored Boxplots (boxplots at sunrise). All data below highest RL wiped off plot. Data above are same as if there were no RLs.



Censored Probability Plot

All detected obs plotted as points, even those between DLs. Their percentiles (on y-axis) adjusted for presence of nondetects, but nondetects are not plotted.

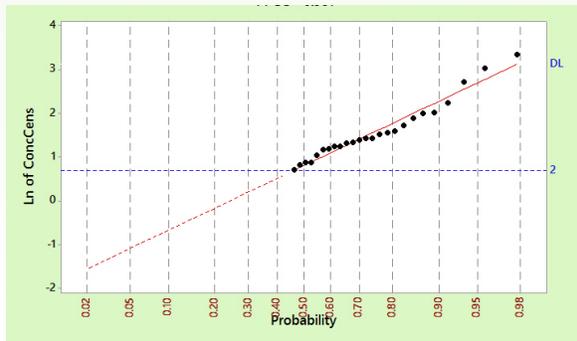
DO NOT just delete nondetects and create probability plot as usual!



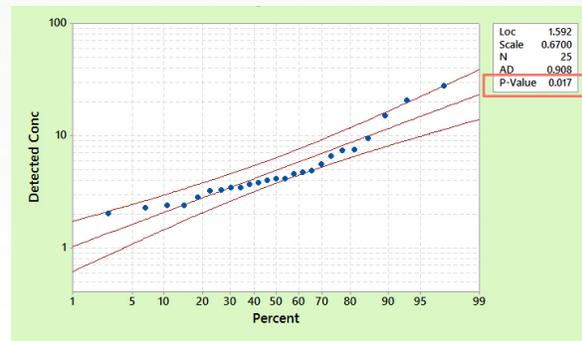
Don't delete nondetects and test whether data follow a distribution

Lognormal data. 44% NDs at one RL, at 2 ($\ln = 0.69$)

No longer lognormal, test for distribution will be wrong



Censored Probability Plot

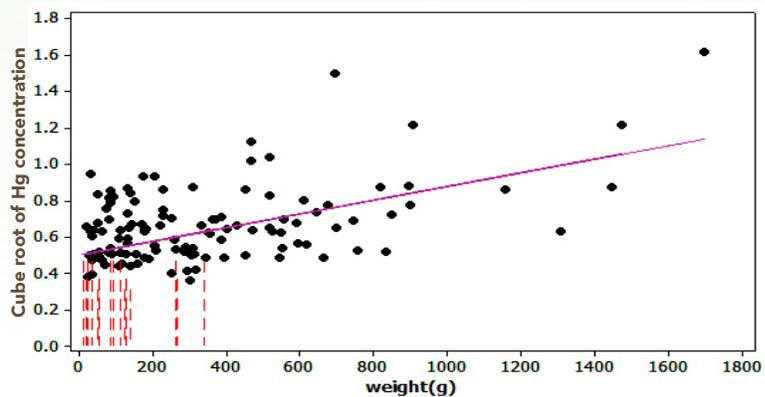


Standard probability plot after deleting nondetects

Scatterplots

Nondetects shown as intervals, not as points

Showing that nondetected Hg occurs only at low fish weight is important



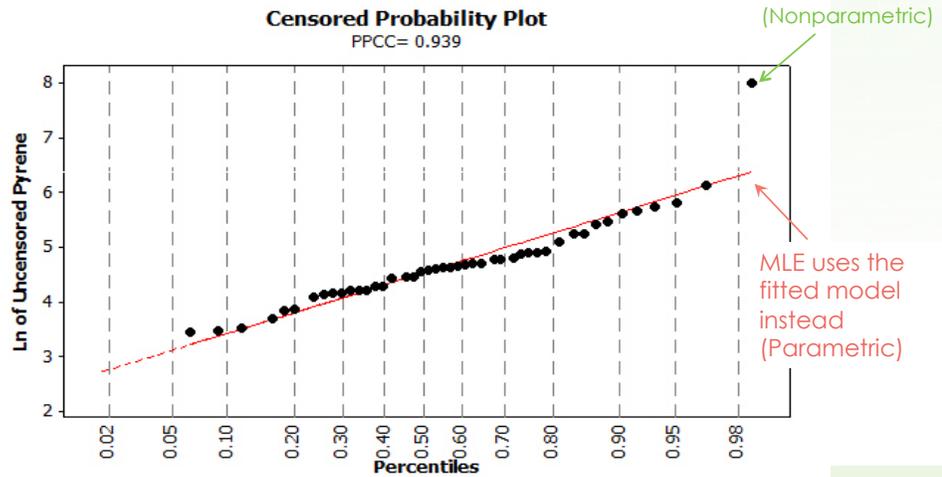
3. Descriptive statistics for data with nondetects

1. Binary methods
 - report % exceeding a single limit
 - re-censor data to below/above highest RL
 - Pros: simple
 - Cons: loses information
2. Simple nonparametric methods
 - report the median, IQR
 - must re-censor data below highest RL
 - Pros: simple
 - Cons: loses information, but maybe not much
3. Survival analysis methods [Our focus today](#)
 - provide numerical values
 - Pros: Can handle multiple RLs
 - Cons: Not familiar to environmental scientists

Three survival analysis methods to estimate descriptive statistics

- MLE (Maximum Likelihood Estimation)
 - theoretically best method if data follow a specified distribution. Parametric.
- Kaplan-Meier / Turnbull
 - estimate the percentiles (cdf) for detected data, accounting for the positions of nondetects.
 - Nonparametric (no distribution shape assumed).
- Robust ROS
 - regression on probability plot
 - parametric method for nondetects; nonparametric method for detects.

Three survival analysis methods to estimate descriptive statistics



Practical Stats
© 2016

Stats for the Pyrene data

Method	Mean	StDev	Pct25	Median	Pct75
MLE(ln)	133.9	142.7	50.9	91.6	164.9
K-M	164.2	393.9	63.0	98.0	133.0
ROS(ln)	163.2	393.1	63.2	90.5	132.8

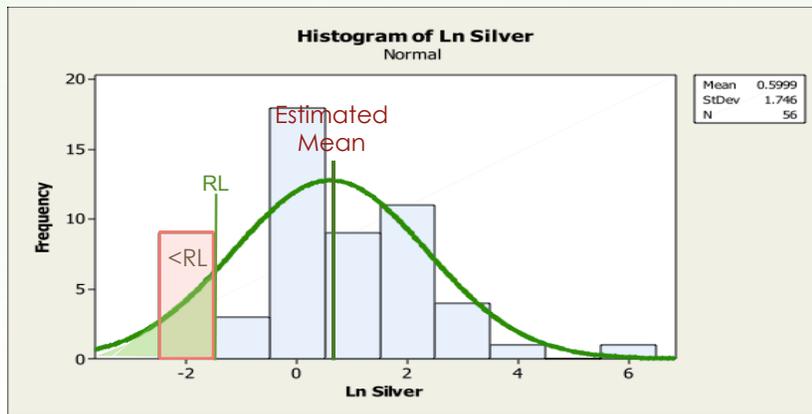
The 3 methods for censored data

- None of these 3 methods uses substitution
- Each of these 3 methods handles multiple DLs
- With outliers and non-normality you must decide whether to believe the data or the model

Practical Stats
© 2016

22

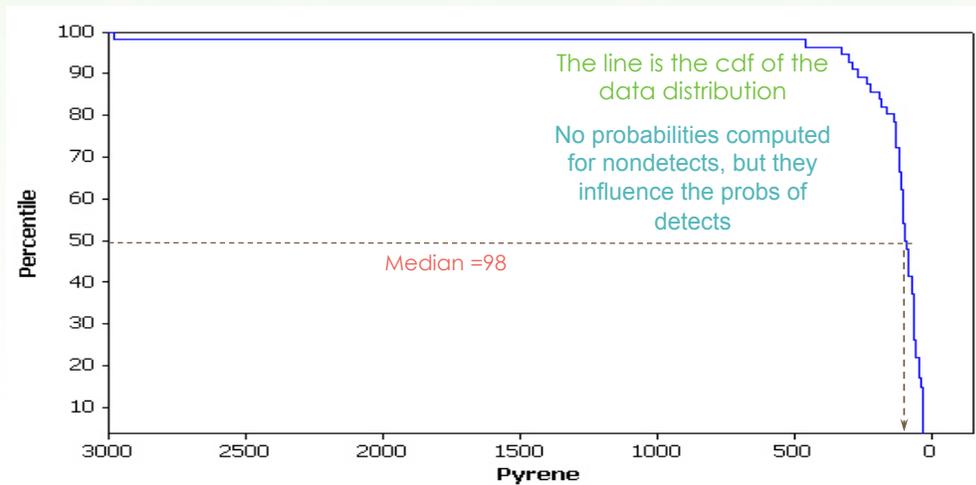
MLE: How MLE Works



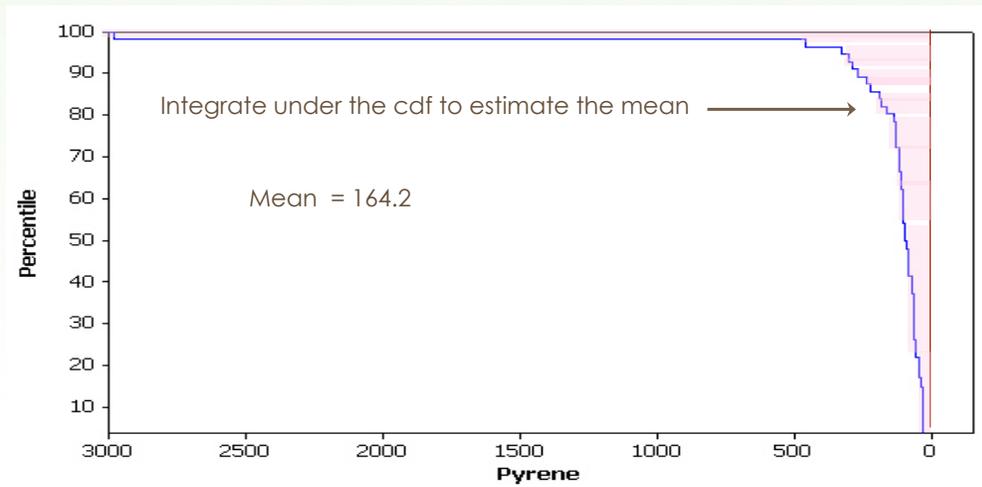
Curve is the solution for the chosen shape (here, normal dist) that best fits

1. The detected data (light blue bars) and
2. The percents of data below each RL (match the green area under the curve to the area of pink bar at -2)

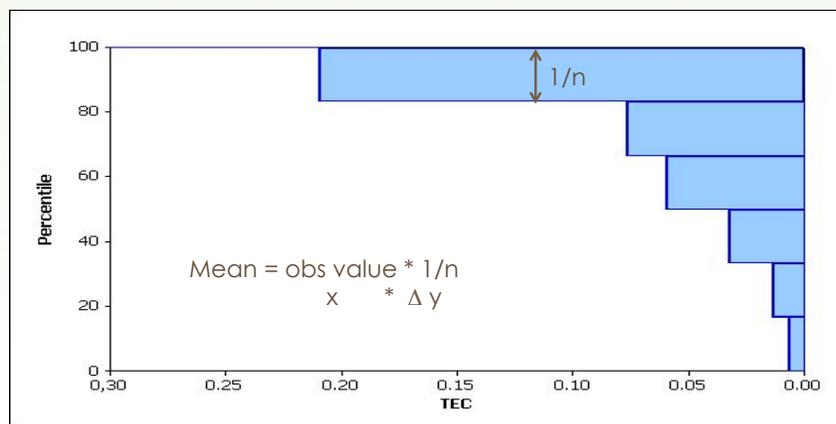
How Kaplan-Meier Works



K-M Estimate of the Mean

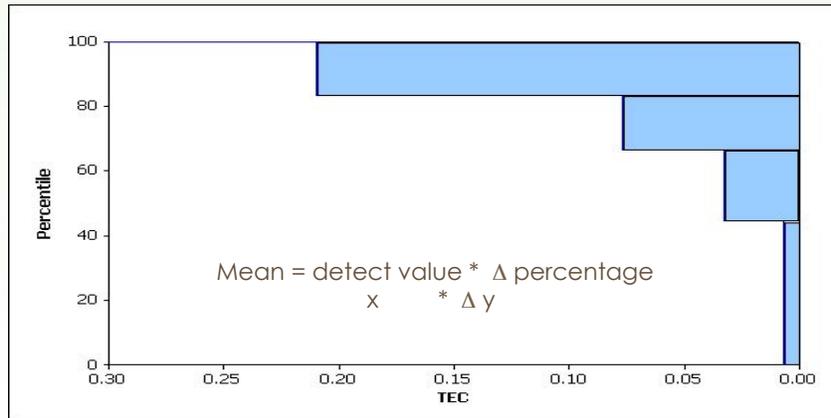


Without nondetects, each obs has a weight of $1/n$ (height of each bar) when computing the mean



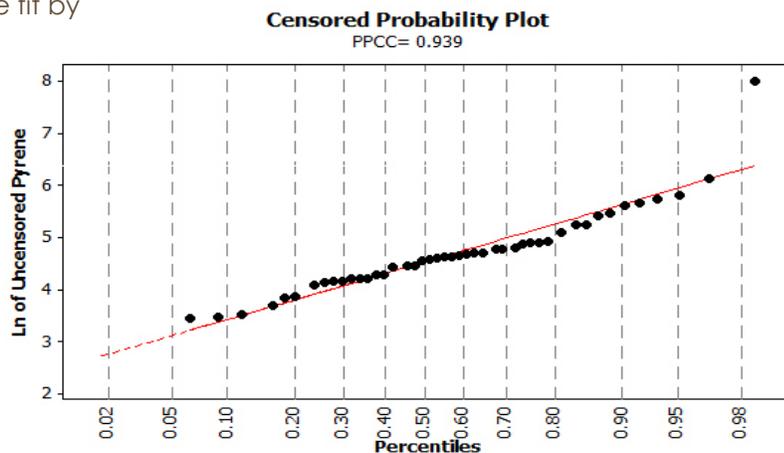
$$\text{KM mean w/o NDs} = \sum \text{data} / n$$

With nondetects, detects are unequally weighted based on # of points occurring above and below that point



Regression on Order Statistics (ROS)

- Detected values plotted at their percentiles (computed by including observed percent of nondetects)
- The distribution line fit by regression
- Regression model used to impute values for nondetects
- Detects plus imputed values used to compute statistics



Example: MLE estimates: NADA package for R

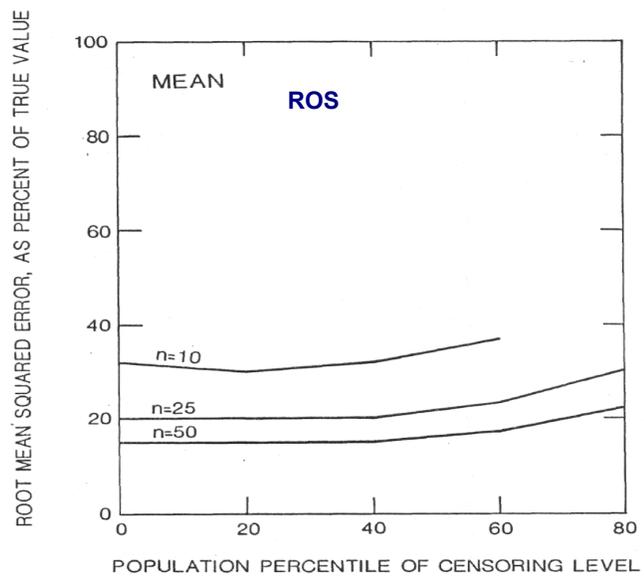
MLE:

```
> AsMle=cenmle(As,AsCen,dist="gaussian") [assume a normal distribution]
> AsMleLn = cenmle(As,AsCen) [default: assume a lognormal distribution]
```

Output (Typing the name prints the results that were computed):

```
> AsMleLn
      n      n.cen      median      mean      sd
24.0    13.00    0.7766007    0.9452585    0.6559261
```

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

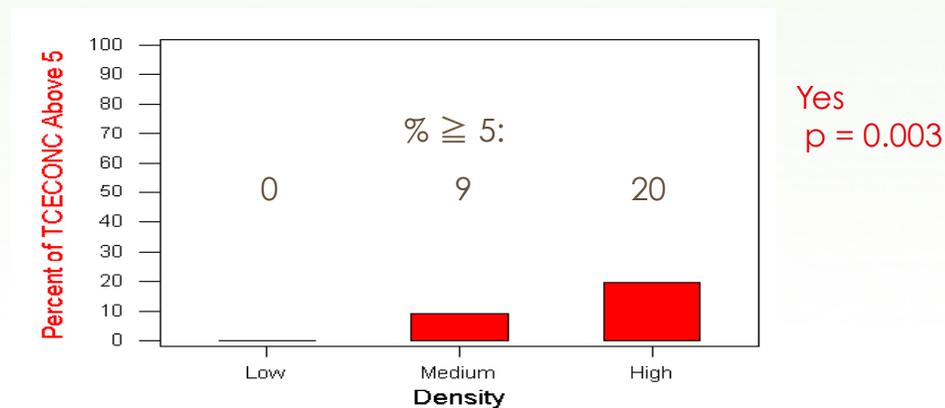


4. Hypothesis Tests

- Reminder: Substitution performs poorly. False “reject” or “do not reject” possible
- Methods without substitution:
 1. Binary methods
 2. Simple (nonparametric) methods -- re-censor at highest DL, run the test
 3. More complicated survival analysis methods provide a full solution for multiple DLs

Binary method: Contingency Table Test

Do % of TCE concentrations ≥ 5 ug/L differ?



Simple nonparametric methods

Data

<1	2	<2	3.1	4.0	<2	<5	<5
4.7	5.3	5.6	7.0	7.8	8.2	8.5	9.9



Ranks

5	5	5	5	5	5	5	5
5	10	11	12	13	14	15	16

9 smallest values all are <5. Ranks are tied at the mean of 1-9, = 5. Data ≥ 5 retain their individual ranks.

Better than substitution because it DOES NOT ADD any *invasive data*

Simple Nonparametric: Kruskal-Wallis test after re-censoring all values below 5 ug/L to <5

```
MTB > %censkw c5 c4 c1
```

```
Kruskal-Wallis Test on TCECONC.
```

Density-	N	Median	Ave Rank	Z
Low	25	-1.000	109.0	-1.11
Medium	130	-1.000	120.4	-0.84
High	92	-1.000	133.2	1.56
Overall	247		124.0	

H = 9.17 DF = 2 **P = 0.010** (adjusted for ties)

Remember:
ANOVA p-value was 0.565

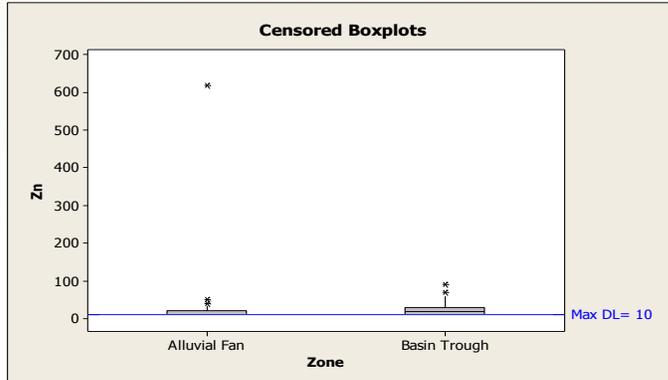
Drastically different from ANOVA. Never perform a parametric test like t-test or ANOVA after substitution!

Survival analysis methods to test data with multiple RLs

The Generalized Wilcoxon test: like a Kruskal-Wallis or Wilcoxon rank-sum test (nonparametric) but handles data with multiple RLs without re-censoring to the highest RL

Do zinc concentrations differ between 2 ground water groups?

Two RLs, at 3 and 10 ug/L.



Survival analysis methods to test data with multiple RLs

(data from Millard and Deverel (1988))

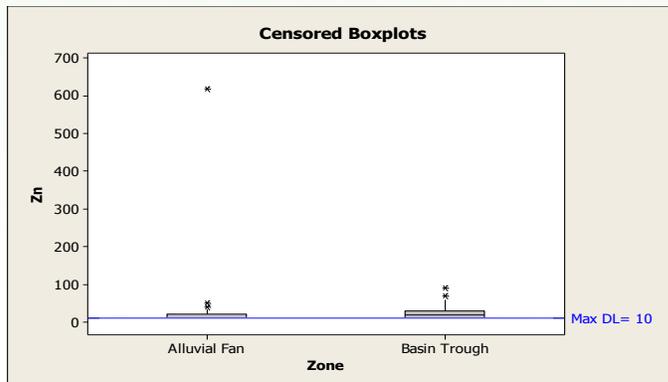
The Generalized Wilcoxon (GW) test: like a Kruskal-Wallis or Wilcoxon test (nonparametric) that handles data with multiple RLs without re-censoring to the highest RL

Do zinc concentrations differ between 2 ground water groups?

The t-test after subbing 1/2 RL doesn't find a difference.

MTB > %gw c3 c4 c5
 Wilcoxon p= 0.019

The GW test easily finds it.



5. Correlation and regression for data with nondetects

- A. Distributional (parametric) methods
 - Likelihood correlation coefficient
 - Censored regression -- Issue: are residuals normal?
- B. Nonparametric methods
 - Kendall's tau correlation coefficient
 - Akritas-Theil-Sen line

Evaluation of Substitution for regression models

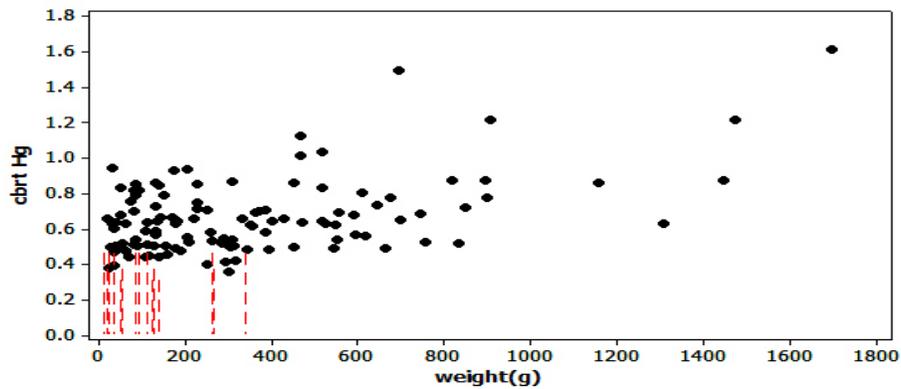
Thompson and Nelson (2003) found that for censored response (y) variables, substituting one-half the DL for nondetects produced

1. biased parameter estimates (slopes too close to zero) and
2. artificially small standard error estimates (x variables falsely significant)

The result of substitution in regression? Bad regression models, and inaccurate statements about which variables are correlated with and can be used to predict the Y variable.

Can methyl mercury concs in fish be predicted from the weight of the fish?

Cube root of Methyl Mercury versus Weight of Fish
 Nondetects shown as dashed red lines



Why cube root? Residuals in original units not normally distributed

Probability Plot for Std Resids of Censored Hg (ug/g wet)

Normal - 95% CI

Interval Censoring - ML Estimates

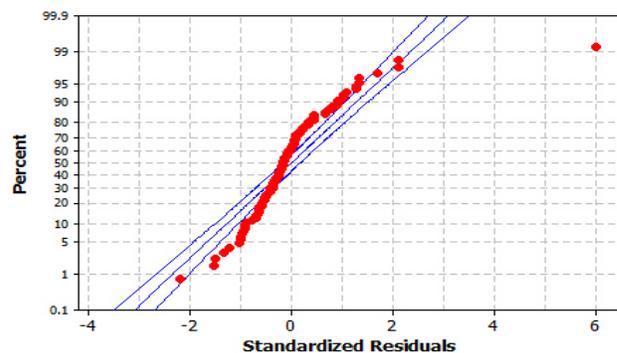
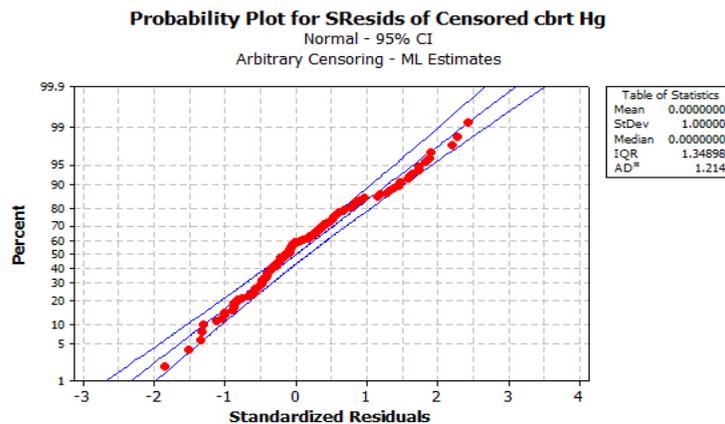


Table of Statistics	
Mean	0.000000
StDev	1.000000
Median	0.000000
IQR	1.34898
AD*	6.570

Residuals of cube roots much more like a normal distribution



Correlation coefficient by MLE

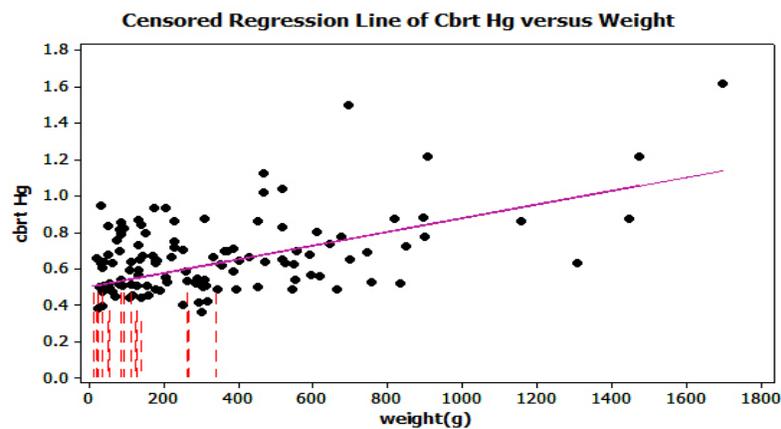
Parametric approach: MLE –based correlation:

the Likelihood r correlation coefficient should be used in the same context as Pearson's r – a linear correlation (not curved) with normal residuals.

It is based on the likelihood ratio test, which determines whether the regression equation explains a significant amount of variation.

Correlation and Regression by MLE

LR corr coeff = 0.52 slope = 0.00038, p < 0.001
 Nondetects included. No substitution.



lowest Hg (nondetects) occur only at low weights

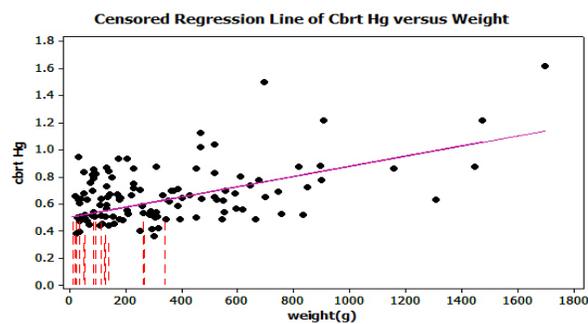
Nonparametric correlation coefficient: Kendall's tau

Compares each point with subsequent points in order of x
 How many Ys (here, Hg) increase, how many decrease?

Null hyp: half +, half -

$$\text{Tau is } \frac{\# + \text{ minus } \# -}{\text{total } \#}$$

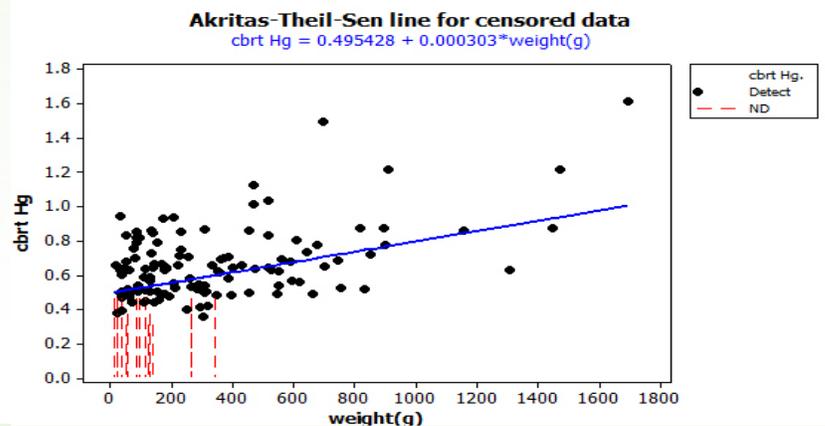
- <0.4 to 1.0 +
- 1.1 to 1.5 +
- <0.6 to 2.3 +
- <0.4 to <0.6 0
- 1.1 to <3 0



Nonparametric regression: Akritas-Theil-Sen line for data with nondetects

Cube root of Hg vs. weight

$\tau = 0.27$ (equiv to 0.47 for r) $p = 0.000$



Nonparametric regression: Akritas-Theil-Sen line

ATS is an optimization procedure

Searches for the slope that when subtracted from the data, the resulting residuals have $\tau=0$

Intercept is the median residual

No assumption of normality of residuals required.

However, this is a linear model. So data should be approx. linear, or else a transformation employed to produce linearity.

6. Software for data analysis with nondetects

- Minitab
 - Macros at <http://www.PracticalStats.com/nada>
- NADA for R on the CRAN site:
 - <http://www.r-project.org/>
- NCSS statistical software
 - <http://www.ncss.com>
- Other commercial software
 - Survival analysis routines for "greater thans". Must first "flip" the data
- ProUCL5 software

Routines for everything discussed here and more. Can input nondetects.

Has several routines. Can input nondetects.

Has routines for all but ATS line. Must first flip the data and run as "greater thans".

KM, ROS; Tarone-Ware (GW) test. No corr/regression for nondetects. Can input NDs.

Conclusions

- Survival analysis methods are available to compute descriptive statistics, perform hypothesis tests, and build regression models. Parametric and nonparametric methods available
- These methods work with data censored at multiple detection limits. No substitution necessary or allowed. Let nondetects be nondetects
- Much (but not all) of the information can be extracted using binary or simple nonparametric methods, after first re-censoring to the highest RL
- Any of the above will be far better than substituting (0, 1/2RL, etc) for nondetects and computing means or running parametric tests.

Further Resources

- Stats for Censored Environmental Data textbook
 - Contains much more detail
- Online materials from this webinar
 - <http://www.PracticalStats.com/training/>
- Free Newsletters
 - <http://www.PracticalStats.com/news/>
- Upcoming Training classes
 - <http://www.PracticalStats.com/training/>
- Journal articles
 - Send me an email request at [ask\[at\]practicalstats.com](mailto:ask[at]practicalstats.com)

Thanks for your attention

- I hope this enables you to use methods for censored data in your work

Practical Stats
---- make sense of your data

Questions?