Why Not Substitute ½ DL for Nondetects?

Dennis R. Helsel

PracticalStats.com

1

What's wrong with substitution?

a. Produces invasive data alien to the concentrations actually in samples
b. Adds a pattern to your data. Substitution is NOT neutral
c. Produces poor estimates and incorrect statistical tests
d. Far better methods are available. Right now. You don't need a PhD to do them.

2

For More Information

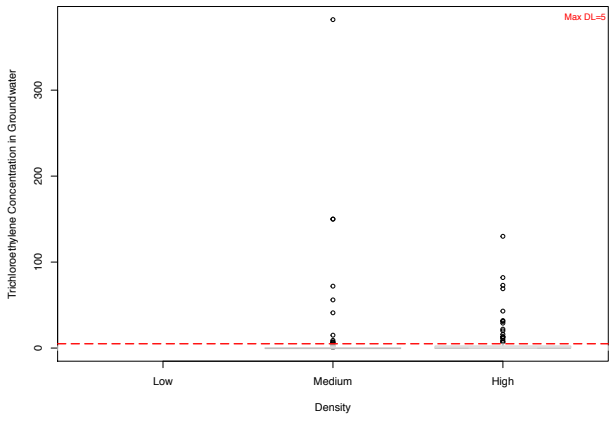
on this topic and other methods for data analysis with nondetects, see our online course

Nondetects And Data Analysis

at https://practicalstats.teachable.com

3

Example: TCE concentrations in GW



- Reporting limits at 1, 2, 3, 4, and 5 ug/L. ~ 80% censoring
- Do TCE distributions differ among the three land-use groups?

4

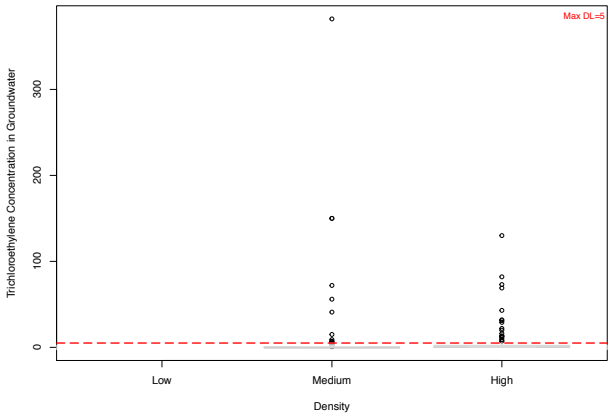
Substitution versus Kruskal-Wallis

ANOVA sub <5 = 0
p = 0.57

ANOVA sub <5 = 2.5
p = 0.56

ANOVA sub <5 = 5
p = 0.50

No differences found!



Kruskal-Wallis test
p = 0.01 .

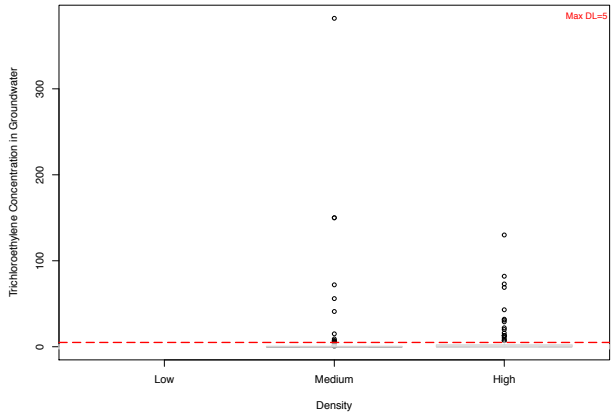
No substitution

Distribution of TCE differs between land-use groups

Simplest Tests

5

Better Tests For Censored Data




```
> cen1way (TCECONC, BDL_1, Density)
                n n.cen median     mean          sd
Factor=High    92    58     <5 7.778019 19.7826151
Factor=Low     25    23     <5 2.083333  0.5768489
Factor=Medium 130   113     <5 7.867264 39.7664396

Oneway Peto-Peto test of CensData: TCECONC by Factor:
Density

Chisq = 16.25   on 2 degrees of freedom  p = 0.000295
```

Distribution of TCE differs between land-use groups

```
Pairwise comparisons using Peto & Peto test
       High   Low
Low    0.0098 -
Medium 0.0024 0.2796
```

Shows which groups differ from others

Nonparametric Test using Multiple Detection Limits
from our Nondetects And Data Analysis course

6

There is still much confusion over the value of nondetects

Dr. Ronald Hites in Environ. Science & Technology (2019) in a Scientific Opinion piece stated:

"…. *Any time you have a % non-detected >20%, for whatever reason, it is unlikely that the data set can give useful results*."

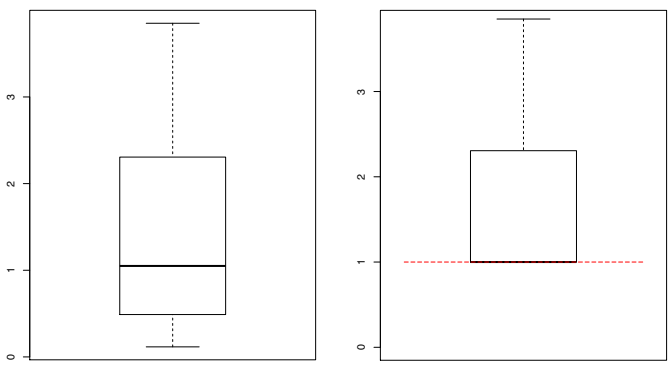
7

Substitution: Invasive data

Estimating the Standard Deviation

what happens to it when the same number is substituted for 60% of the observations?

Before substituting.
True std. dev.
s = 1.02



After substituting.
std. dev.
s = 0.84

t-tests, ANOVA, regression, intervals all depend on the std. dev.

8

Substitution when computing confidence intervals?

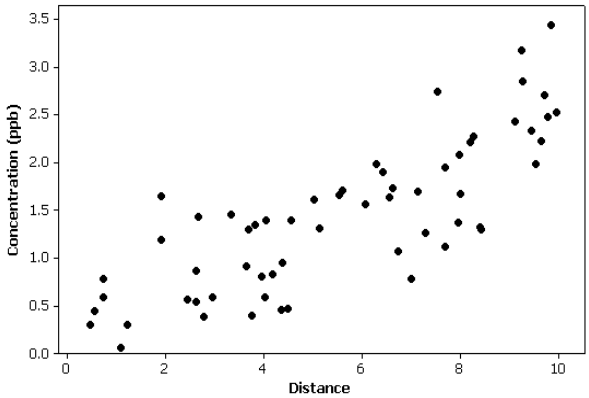
- Singh et al (2006), developers of the ProUCL software, determined that substituting ½ DL "does not provide adequate coverage [UCL95 is not high enough] …even for [% NDs] as low as 10%"
- Lower standard deviations produce lower confidence limits, too-short intervals.
- They summarize their results with "Do not use DL/2 (t) method to compute a UCL".
- Recommended methods were based on a Kaplan-Meier estimator, which is available in our course software.

9

Substitution: Invasive data

Correlation and Regression

Before censoring.
True correlation
r=0.81



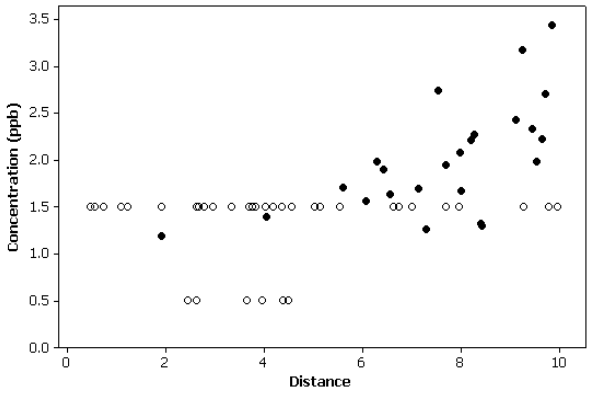
10

Substitution: Invasive data

Correlation and Regression

Two DLs @ 3 and 1.
<3s become 1.5,
<1s become 0.5.

After substitution. invasive data form flat (zero-slope) lines, lowering correlation to r=0.55



11

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

- biased parameter estimates (slopes) and
- artificially small standard error estimates (explanatory variables who shouldn't be in the regression appear to be significant)

There are better ways!

12

Regression with censored data

No Substitution for Nondetects!

Software comes with NADA course

```
> Pbreg <- cencorreg(Blood, BloodCen, Kidney)
 Likelihood R = 0.8236
 Rescaled Likelihood R = 0.8721
 McFaddens R = 0.714
> summary(Pbreg)
```

Correlation Coefficients

Regression Equation

Blood lead concentrations are related to lead in kidneys

```
             Value Std. Error      z       p
(Intercept) -4.4573     0.1733 -25.72 < 2e-16
Kidney       0.2436     0.0302   8.07 7.1e-16
Log(scale)  -0.6737     0.2036  -3.31 0.00094

Loglik(model)= -14.7   Loglik(intercept only)= -30
Chisq= 30.62 on 1 degrees of freedom, p= 3.1e-08
```

ln(blood Pb) = -4.457 + 0.244*kidney Pb or blood Pb = $e^{-4.457} \bullet kidneyPb^{0.244}$

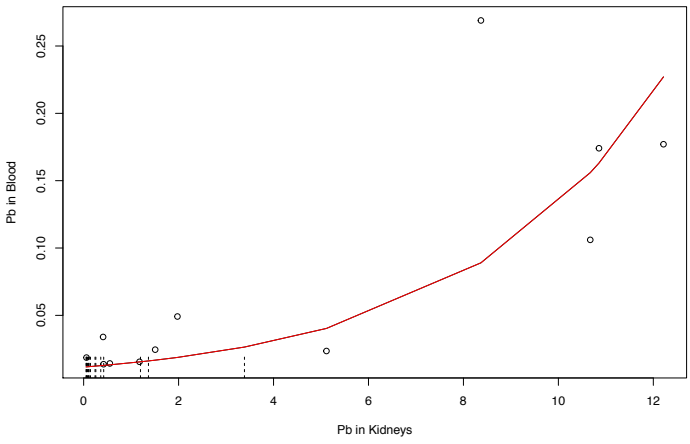
13

Plotting the regression line

Regression straight line in log units becomes a curve in original units

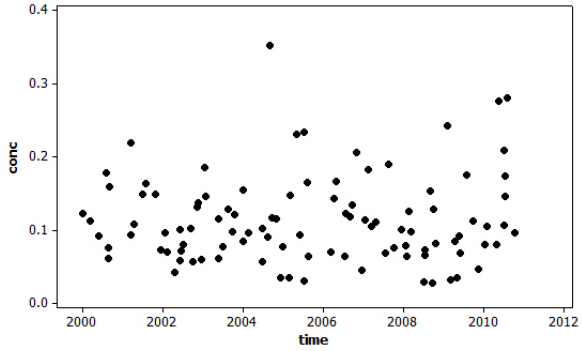
```
> cenxyplot(Kidney,
  KidneyCen, Blood, BloodCen,
  xlab = "Pb in Kidneys",
  ylab = "Pb in Blood")

> ik <- order(Kidney)

> lines(Kidney[ik],
  exp(predict(Pbreg)[ik]),
  col = "red")
```

Nondetects are plotted as dashed vertical lines. The info that they occur only at the low kidney Pb concentrations is used in computing the regression model.



14

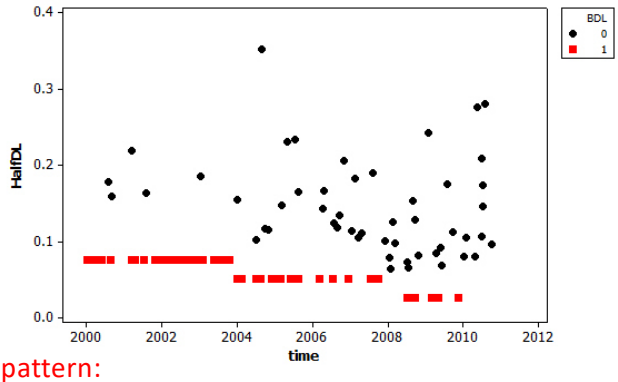
Classic faux pas with substitution: finding a trend that isn't there



- No change over time. No trend.
- Will replace smallest values with a decreasing pattern of detection limits, mimicking what often happens in labs.

15

Finding a trend that isn't there



Invasive pattern:

- DLs decrease over time.
- After substitution with 1/2DL, decreases over time (trend) may test as significant

16

Trend -- the censeaken function

The seasonal Kendall test for data with multiple DLs. No substitution. Available in our NADA course materials.

```
> censeaken (dectime, `Total Recoverable Chromium`, CrND, group = Season)
 DATA ANALYZED: Total Recoverable Chromium vs dectime by Season
----------
  Season  N    S    tau      pval intercept    slope
1    Dry 34 -176 -0.314 0.0091337    79.103 -0.03901
----------
  Season  N   S     tau    pval intercept    slope
1    Wet 29 -24 -0.0591 0.66604    24.355 -0.01169
----------
Seasonal Kendall test and Theil-Sen line
   N    S    Tau Pvalue_SK Nreps Intercept    Slope
1 63 -200 -0.207     0.014   999    74.232 -0.03655
----------
```

Significant downtrend in Dry season

No significant trend in Wet season

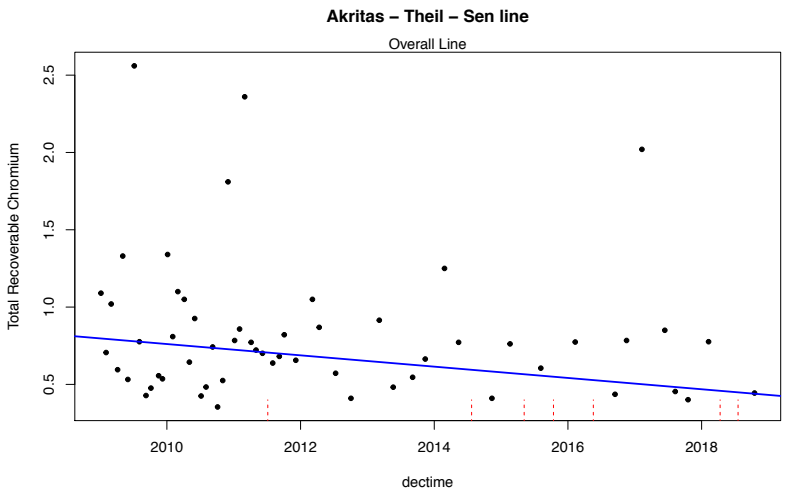
Significant trend overall. SK slope is -0.036 ug/L per year

17

Trend -- the censeaken function

Nondetects (dashed vertical lines) influence the trend line and test.

They occur more frequently at later times, adding to the evidence of a downtrend.

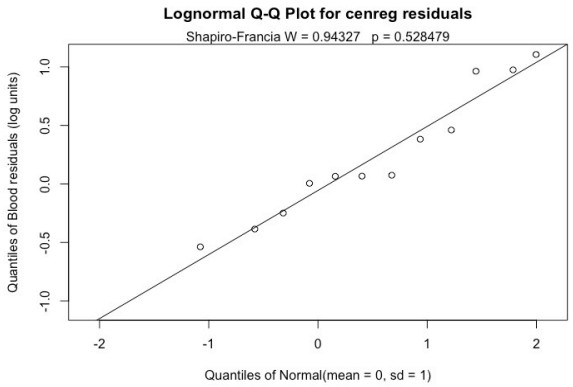


18

Check Assumptions with QQ Plot

Regression is a parametric method:

Check the assumption of a normal distribution with a Q-Q plot of residuals. Here using the default of log(Y) fits well. It often does for data with NDs (because they are close to zero)



 19

19

What else is available?

- Our Nondetects And Data Analysis course covers:

- parametric and nonparametric methods
- computing summary statistics
- confidence, prediction and tolerance intervals
- plotting data
- comparing to standards
- testing differences of two groups
- 3+ group tests with multiple comparisons
- regression and correlation
- trend analysis
- multivariate methods

All without substitution of values like 1/2DL for nondetects

 20

20

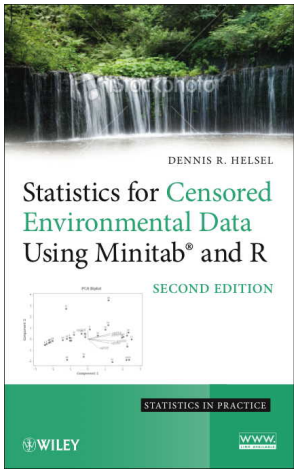
For more detail:

Statistics for Censored Environmental Data
(the second edition)

by Dennis R. Helsel

Wiley (2012)

www.PracticalStats.com/nada



 21

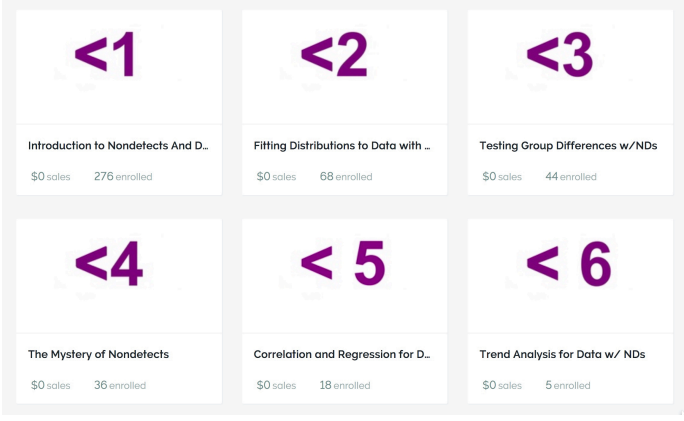
21

Free videos introducing methods for nondetects available now for streaming

on our Online Training Center
https://practicalstats.teachable.com/

Let colleagues know about them.

But they're only a fraction of the material in the NADA course itself.



 22

22

Thank you for viewing

Statistical methods for data with nondetects at 1 or more detection limits is the focus of our Nondetects And Data Analysis course

Questions about this material or our courses?

Get in touch!

Dennis Helsel ask@practicalstats.com

Courses & free recordings at:
http://practicalstats.teachable.com

23

23