Two Things I've Learned After 43 Years of Applying Statistics to Environmental Science

Dennis R. Helsel

PracticalStats.com

1

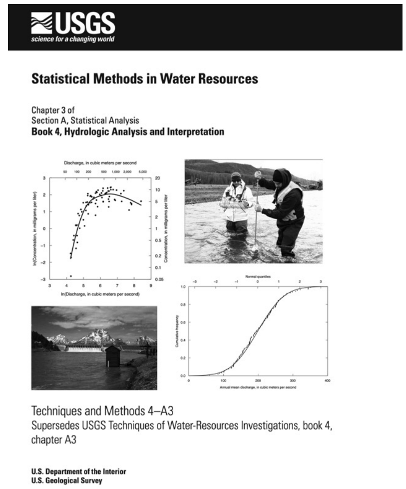
Two Things I've Learned in 43 Years

1. Use methods that fit the objectives of your data. That often means using distribution-free methods
2. Use methods for censored data developed in other disciplines for data with nondetects

 2

2

1. Use methods that fit the objectives of your data.



Much more detail is available in the 2020 edition of **Statistical Methods in Water Resources**, available for free at

https://doi.org/10.3133/tm4A3

See especially Chapter 4.

3

Simple Example: Which statistic (mean or median) meets my objective?

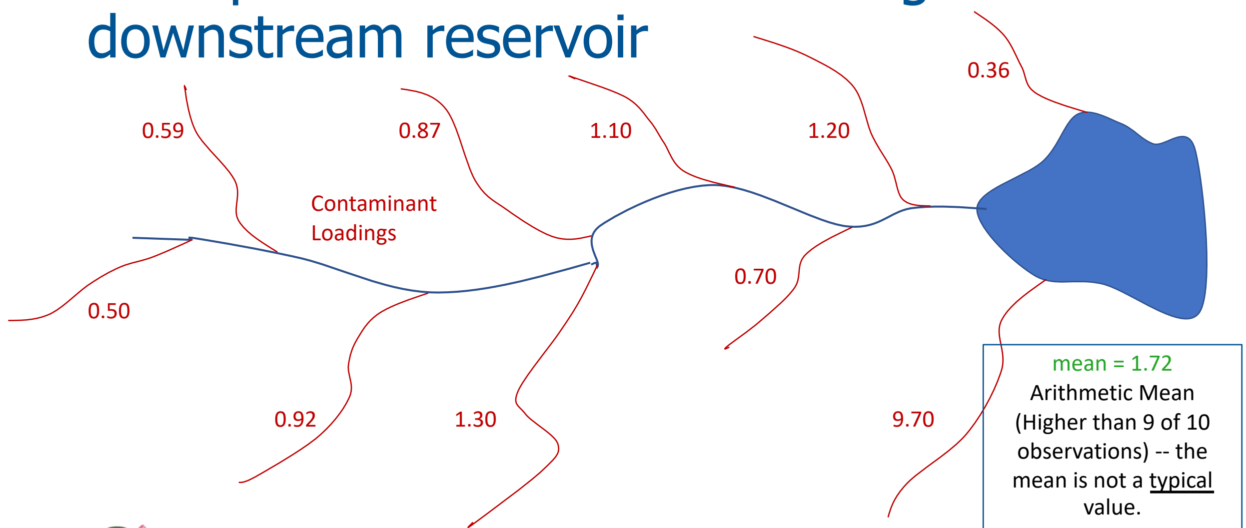
- the mean is a per-sample estimate of the total

$$\bar{x} = {}^{\Sigma x_i}/_{n}$$

- are you interested in a total? Does it make physical sense to sum observations? Yes for estimates of mass, volume, cumulative exposure, etc…

4

Example 1: Ten streams flowing into a downstream reservoir

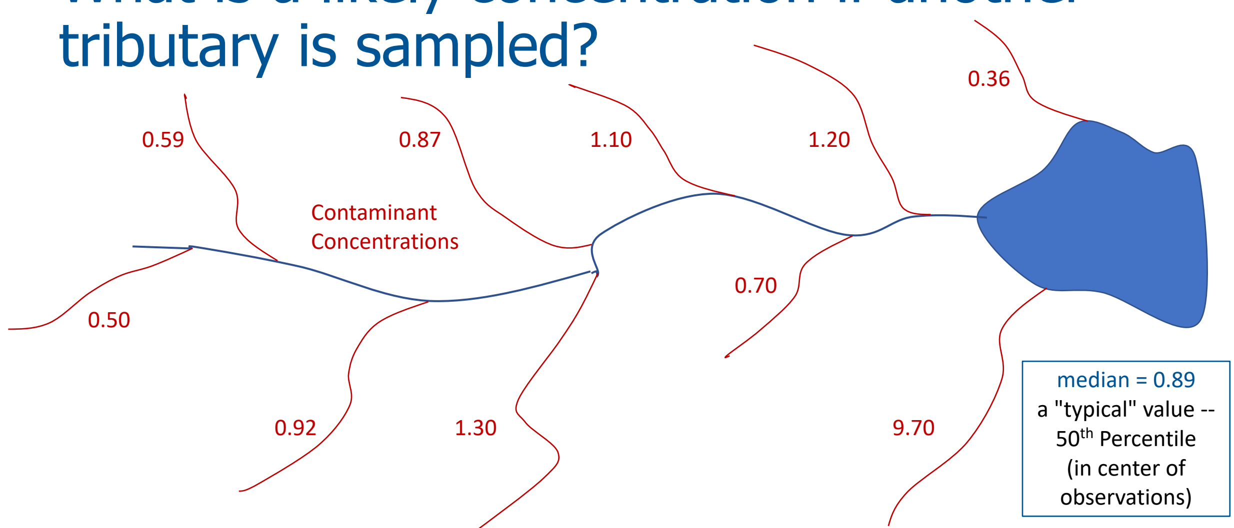


If being sensitive to a few high outliers is a good thing, you want the mean.

 5

5

What is a likely concentration if another tributary is sampled?

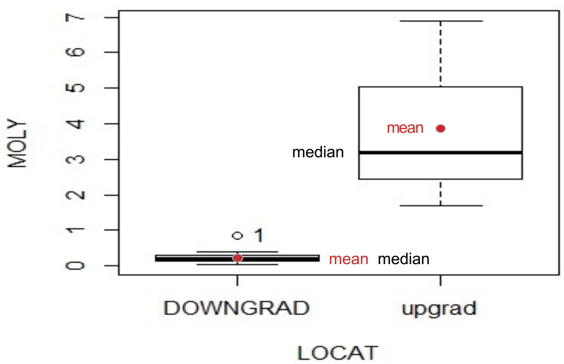


If being sensitive to a few high outliers is annoying, you want the median.

 6

6

Example 2: Do the two groups differ in molybdenum concentrations?



Note: All of the upgradient data are higher than all of the downgradient data.

mean in DOWNGRAD 0.25   mean in upgrad 3.93 ~(16x)
t-test: (parametric test)
means not significantly different. p=0.14

median DOWNGRAD 0.20   median upgrad 3.20
rank-sum test (distribution free test)
cdfs (percentiles) significantly different p=0.01

The answer depends on which question you are asking, which should control which test you use.

Reminder: p-values present strength of evidence in the data for there being no signal (no difference between groups, no trend, etc.). Smaller p-values indicate stronger evidence that a signal exists.

 7

7

Two different questions

- Parametric tests: test differences in means (t-test)
  answers questions about mass, volume, totals
  assume a normal distribution fits the data
  and groups have equal variance.
- Nonparametric tests: tests difference in cdfs (rank-sum test)
  answers questions of frequency
  "does one group have more frequent high values than the other?"
  distribution-free. No assumptions of data shape.

Which type of question do you normally ask, mass or frequency?

 8

8

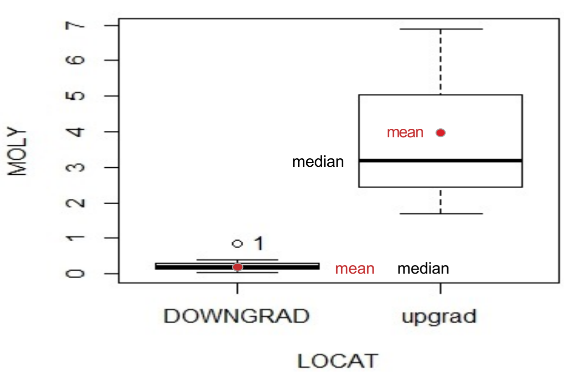
Three Classes of Hypothesis Tests

- Parametric tests are tests on means or other distributional parameters. They assume that data follow some specific distribution, often the normal, to get accurate p-values. Strong affect by outliers.
- Nonparametric tests are tests on ranks (percentiles). They compute all possible outcomes to get a p-value. No distribution assumed – the observed distribution of data is used. Resistant to outliers.
- Permutation tests are tests on any statistic. They determine the likelihood of getting the observed test statistic out of thousands of possible rearrangements of the data. Used as an alternative to parametric tests, to test means without assuming a distribution, and to minimize the effect of outliers -- <u>they solve the major problems with parametric tests</u>.

9

9

Example 2: Do the means of two groups differ in molybdenum concentrations?



Note: All of the upgradient data are higher than all of the downgradient data.

mean in DOWNGRAD 0.25 mean in upgrad 3.93 ~(16x)

t-test: (parametric test)
means not significantly different. p=0.14

permutation test (distribution free test)
means significantly different p=0.002

If you are interested in testing means, use a permutation test. It works well on non-normal skewed data as well as for data that can be fit well by a normal distribution.

10

10

Tests compute the p-value in different ways

The three classes of hypothesis tests have three different methods for obtaining a p-value from the data:

1. Parametric tests. Assume the data follow a distribution (normal). The distribution of the test statistic is then used to compute the p-value, but this is only valid if the data actually have the assumed distributional shape
2. Nonparametric tests. Compute all possible test statistics that could result from datasets of the size of the current dataset - distribution-free
3. Permutation tests. Compute a large number of (or all) possible test statistics that could result from re-arrangements of the current dataset - distribution-free

11

11

Power: the ability of tests to find a signal when it is present

- Parametric tests have low power whenever data have outliers, or are skewed, or groups have different variability.
- Field data in environmental sciences usually have all three of these characteristics. So ANOVA, t-tests, and t confidence intervals don't work well for the type of data we usually encounter.

Alternatives:

Permutation tests. They don't assume a normal distribution. Not bothered by outliers. <u>Can test for differences in means</u>.

Nonparametric tests. Do not assume a normal distribution. Not bothered by outliers or changing variance. Test for differences in percentiles (typical patterns). Example: trend tests. "Are high concentrations getting more frequent with time?"

12

12

A Decision Tree

- What is your objective?

If you are interested in

Mass/Totals	Frequency/Typical patterns
chronic effects	acute effects
↓	↓
Test means	Test percentiles
using permutation tests	using nonparametric (rank-based) tests

No need to test for whether data follow a normal distribution. Use distribution-free tests instead.

Info on Permutation Tests

- Can be used for any statistic of interest
- Can be used for more than 2 groups
- With smaller datasets permutation tests compute all of the possible test statistics. The p-value is the % of cases where the test statistic using the rearranged data is ≥ the observed test outcome.
- With larger datasets there are too many possible rearrangements. Thousands of rearranged data are tested and the p-value is the % of cases where these outcomes are ≥ the observed test statistic.

Exact Permutation Test for a Small Dataset

Using the perm2 command in the NADA2 package of R:

```
> perm2(MOLY,LOCAT)
Data analyzed = MOLY LOCAT
 Group names are  DOWNGRAD upgrad

PERMUTATION TEST OF DIFFERENCE IN 2 MEANS
Number of Possible Permutations =  560 is less than 1000
 ALTERNATIVE: MEAN of DOWNGRAD  NOT EQUAL TO MEAN of upgrad

Diff of means = -3.685949   pvalue = 0.0018   nrep = 560
```

p-value is two orders of magnitude lower than the t-test p-value of 0.14!
Same data -- Much more power

15

15

Many Reps Permutation Test for a Larger Dataset

Larger datasets produce too many outcomes to compute all. Instead compute several thousand random outcomes.

1. Compute the observed test statistic, δ_{obs}
2. Take a random sample from the data set, without replacement and without regard to which group the data came from. Assign the correct number randomly to each group
3. Compute the test statistic δ for the new sample
4. Do this 10,000 times: $\delta_1, \delta_2, \ldots, \delta_n$ to represent the null hypothesis
5. Compare δ_{obs} to these n test statistics How unusual is δ_{obs}? The answer is the permutation p-value

16

16

Summary: Use methods that fit the objectives of your data

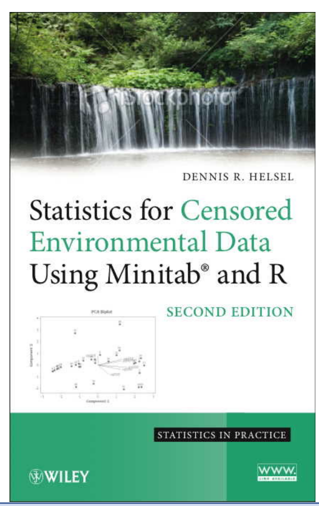
1. Decide whether you want to test questions of mass / volume, or of frequency. IMO most questions in environmental science are of frequencies.
2. If interested in mass/volume/means use a permutation, not a parametric, test.
3. If interested in frequency questions use a nonparametric test.
4. Don't rely on parametric tests like t-tests and ANOVA. They have a considerable loss in power (producing false negatives, missing trends or higher concentrations) with skewed, unequal variance data such as most environmental datasets. Instead use permutation tests.

17

2. Use methods for censored data from other disciplines for data with nondetects

Much more detail is available in the 2012 textbook **Statistics for Censored Environmental Data using Minitab and R**, by Helsel, published by Wiley.

https://practicalstats.com/info2use/books.html



18

1. What's wrong with substitution (of 1/2DL, etc.)?

a. Strongly affects the variation of data (commonly decreases it but may increase it).

b. Adds invasive patterns alien to the collected data. Substitution is NOT neutral

c. Produces poor estimates and incorrect statistical tests

d. Changes the shape of the data distribution

e. Far better methods are available. Right now. You don't need a PhD to do them.

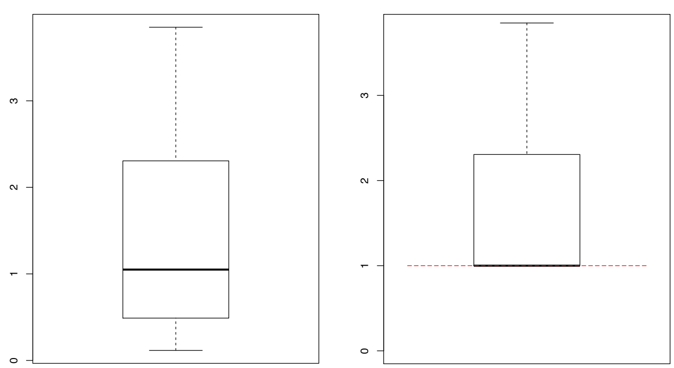
1a. Substitution → Changes Std Dev.

Example: Estimating the Standard Deviation

What happens (here with one DL) when the same number (1/2 DL) is substituted for 60% of the observations?

=== All <2 become = 1 ===>

Before substituting.
True std. dev.
s = 1.02



After substituting.
std. dev.
s = 0.84
-- an 18% decrease

originally had a variety of values below 1/2DL that are now replaced with a constant

How does this affect confidence intervals?

- Singh et al (2006), developers of ProUCL software, determined in a simulation study that substituting ½ DL "does not provide adequate coverage [UCL95 is not high enough] …even for [% non-detects] 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".
- In addition to confidence intervals, t-tests, ANOVA, regression and many other procedures all depend on the std. dev.

How long has this been known?

Gilliom and Helsel (1986) in Water Resources Research:

- Compared substitution to other methods for estimating means, medians, std dev, percentiles
- Found that the other methods were predominantly better than substitution, often greatly better
- ½ DL gave reasonable estimates for the mean with one DL, but not for other statistics, and not with multiple limits
- For example, the bias of subbing 1/2DL for estimating the median was about 4.5 times that for a better method (Regression on Order Statistics).

Methods relying on the standard deviation will be off the mark after substituting a fraction of the detection limits for non-detects

1b. Substitution Adds Invasive Patterns

Characteristics other than pollutant concentrations often affect detection limits. Substituting ½ DL adds the pattern of that characteristic to the concentrations -- a pattern that has nothing to do with the concentration itself.

Example 1: Arsenic (As) in leaves measured in ashes in the lab.

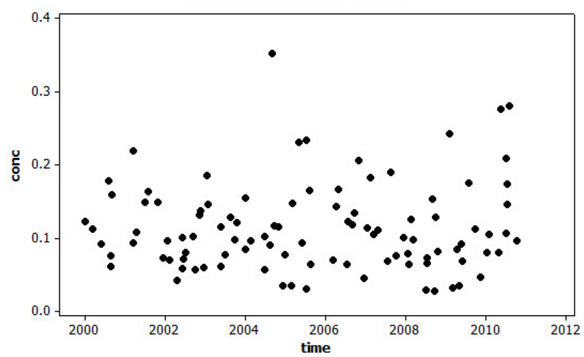
As in dry weight = As in ash weight*(%ash/100). The DL in the ash may be 0.5 for As but the % ash (% leaf material minus water) differs between samples, so the resulting dry weights have many different DLs. 1/2DL adds a pattern of the water weight to the As concentrations unrelated to As concentration in the leaf.

Example 2: Concentrations in a river over 20 years. DLs decrease over time. 1/2DL adds a decreasing pattern over time that was not in the river.

 23

23

Example 2: Finding a trend that isn't there (or may obscure a trend that is there)

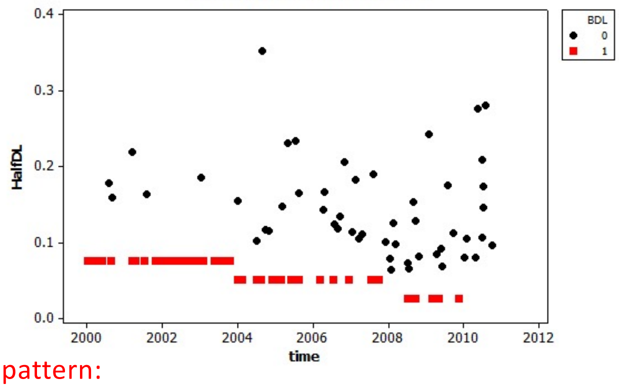


- True pattern -- no change over time. No trend in the river.
- Will replace smallest values with a decreasing pattern of detection limits, mimicking what often happens in labs.

 24

24

Put a trend into the data that isn't actually there



Invasive pattern:

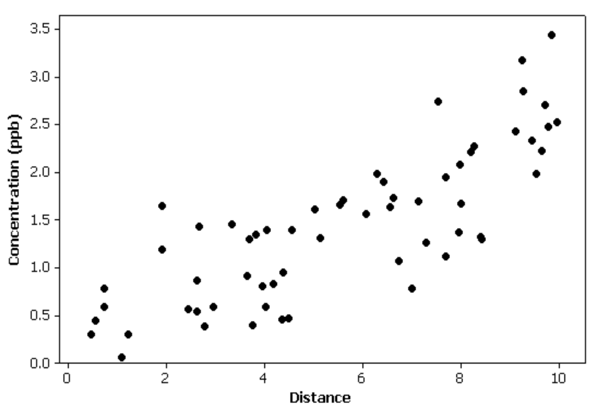
- DLs decrease over time. Data did not.
- After substitution with 1/2DL, a portion of data decrease over time (trend often tests as significant)

25

1b cont. Invasive patterns in regression

Correlation and Regression

Before censoring.
True correlation
r=0.81



26

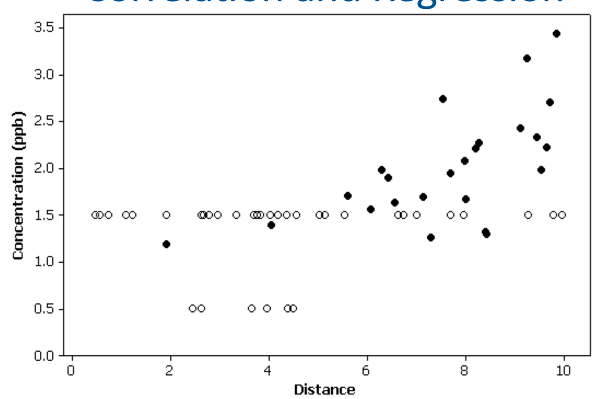
1b cont. Substitution adds invasive patterns not in the original data

Two DLs at 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 from the true 0.81.

It's like watering down good wine.

Correlation and Regression



 27

27

Evaluation of Substitution for regression models

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

- biased parameter estimates (slopes too close to 0) and
- artificially small standard errors (std deviation of residuals). Causes explanatory variables that shouldn't be in the regression to appear significant

There are better ways! The NADA2 package for R contains methods for "censored data" that do not substitute fractions of the DLs and are valid for both 1 and multiple DLs. See https://practicalstats.com/training

 28

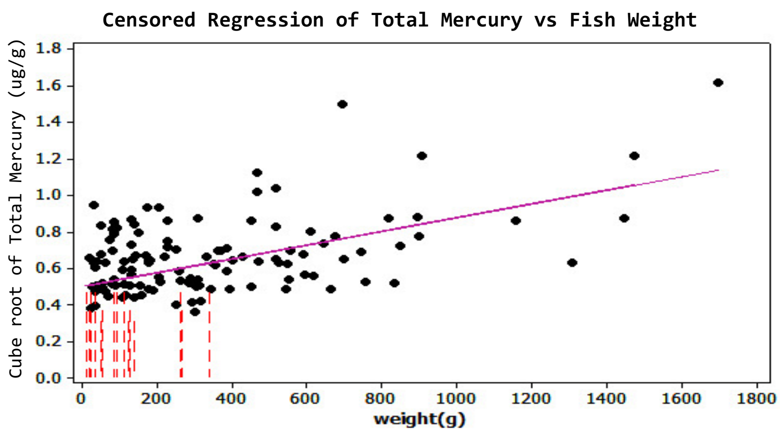
28

1e. What else can be done? Censored Regression

corr coeff = 0.52
slope = 0.00038,
p < 0.001 (weight is significantly correlated with Total Mercury)

non-detects included without substitution.

Lowest Hg (non-detects) occur only at low weights.
Information in the NDs adds to the regression results.



29

1c. Substitution produces incorrect statistical tests

- Data from the Ontario (CA) Pollen Monitoring Network
- Pesticide concentrations are measured in pollen at beehives located across the province.
- Neonicotinoids are neurotoxins that kill insects through attacking receptors in nerve synapses.
- Nearly 100% of corn seed and roughly 60% of soybean seed are treated with neonicotinoids.
- Thiamethoxam is a neonicotinoid pesticide; the concern is its affect on honeybees.
- Do thiamethoxam concentrations differ in pollen between 2 stages of plant growth (post-planting vs. corn tassel appearance)?

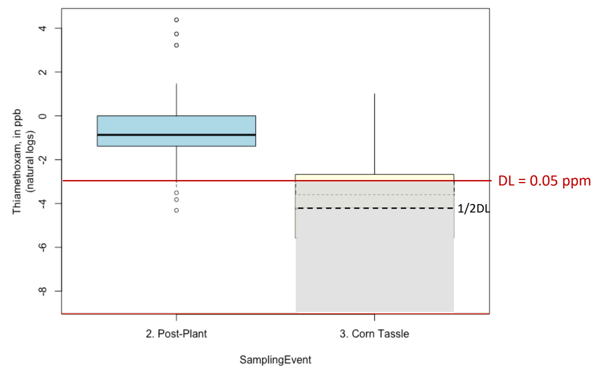
Source: Ontario Ministry of the Environment, Conservation and Parks

30

t-test after substituting 1/2DL finds no differences

Welch Two Sample t-test
data: Thiamethoxam by SamplingEvent
t = 1.9309, df = 53.092, p-value = 0.05884
Significant difference NOT FOUND

Reminder: p-values present strength of evidence against there being a signal (difference between groups, trend, etc.). Smaller p-values indicate stronger evidence that a signal exists.



Substitution followed by t-tests and ANOVA are unfortunately very commonly used to test for evidence of contamination and for determination of levels affecting organisms.

31

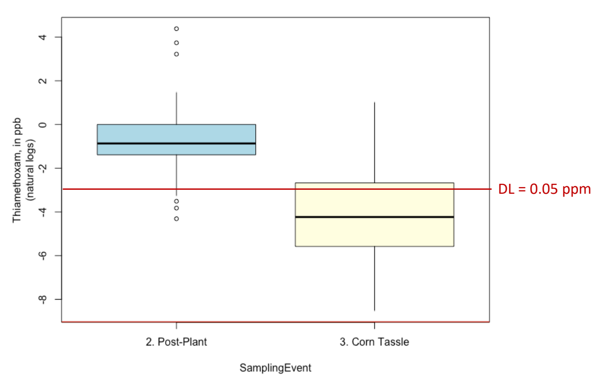
31

1e. Better Tests For Censored Data Exist in the NADA2 package for R

```
> cen1way(Thiamethoxam, ThiaCens, SamplingEvent)

Oneway Peto-Peto test of CensData: Thiamethoxam
                        by Factor: SamplingEvent

Chisq = 62.11   on 3 degrees of freedom
p = 2.08e-13
```

Peto-Peto nonparametric test finds strong evidence for difference between the two groups! No assumption of normal distribution is needed for this test.

The t-test after substitution found much less evidence for a difference (p=0.0588). The p-value from the Peto-Peto test (no substitution) is 11 orders of magnitude lower!
It extracts much more evidence from the same data.



- Testing is used in development of no-effect levels and adverse-effect levels. Not finding effects due to substituting for non-detects has likely caused frequent mis-specification of these levels.

32

32

B. What statistical methods can incorporate nondetects now?

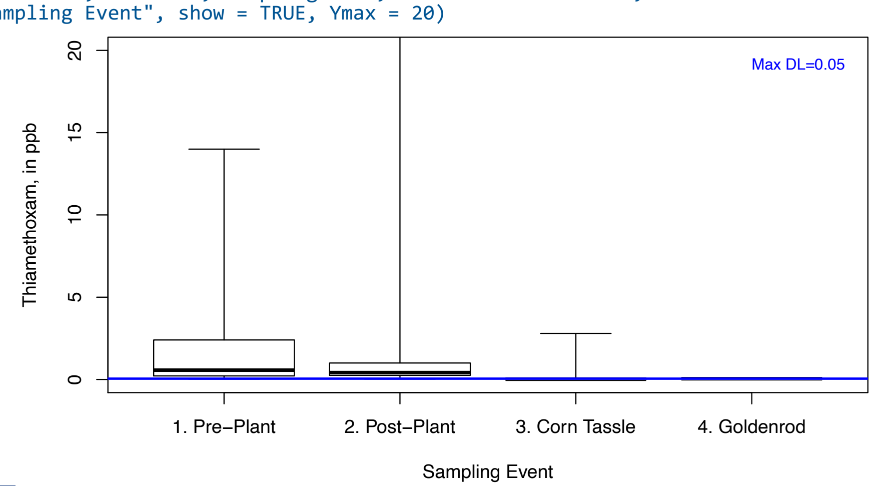
Without substituting numbers for nondetects or throwing variables away, you can:

- Plot data and compare its shape to standard distributions
- Estimate means, confidence intervals, UCLs
- Run hypothesis tests between groups, compare data to standards
- Build regression models and evaluate if residuals match the assumed distribution
- Perform trend analyses (parametric and nonparametric)
- Draw high-dimensional plots, cluster analysis, test hypotheses on multivariate data

33

3. Test Group Differences: Thiamethoxam on Pollen

```
> cboxplot(Thiamethoxam, ThiaCens, SamplingEvent, Ylab = "Thiamethoxam, in
ppb", Xlab = "Sampling Event", show = TRUE, Ymax = 20)
```



34

Peto-Peto test of Difference in Group Concentration Percentiles

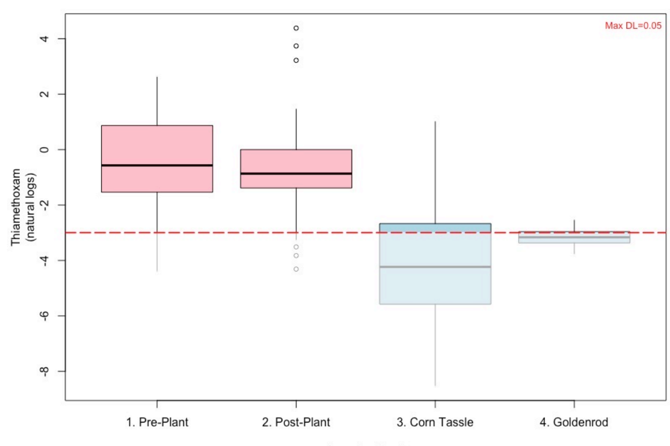
```
> cen1way (Thiamethoxam, ThiaCens, SamplingEvent)

     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
```

Pre-	Post-	Corn Tassle	Goldenrod
A	A	B	B

Logscale to clearly see differences

```
> cboxplot(Thiamethoxam, ThiaCens, SamplingEvent, show = TRUE, LOG=TRUE,
bxcol = c(”pink”, “pink”, “light blue”, “light blue”)
```

- Test results would be identical on a log scale (nonparametric tests).
- Significant difference between groups shown as different colors.
- ROS estimates below the highest DL shown as faded colors.



5. Regression with censored data

Lead in blood and kidneys of herons; Regression by Maximum Likelihood Estimation.

```
> Pbreg <- cencorreg(Blood, BloodCen, Kidney)
 Likelihood R = 0.8236
 Rescaled Likelihood R = 0.8721
 McFaddens R = 0.714
> summary(Pbreg)
Call:
survreg(formula = "log(Blood)", data = "Kidney", dist = "gaussian")
              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
Chisq= 30.62 on 1 degrees of freedom, p= 3.1e-08
```

slope

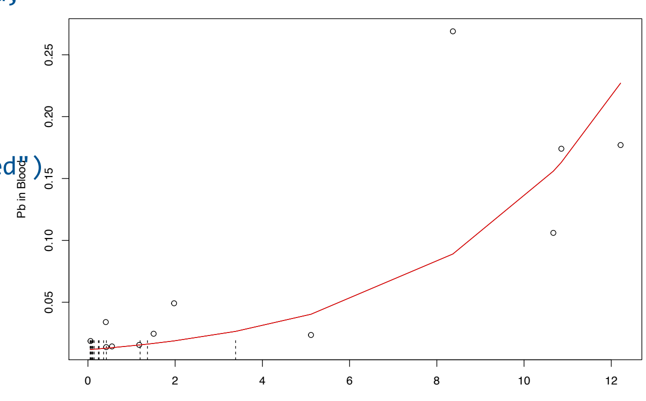
ln(blood Pb) = -4.457 + 0.244*kidney Pb

or blood Pb = $e^{-4.457} \bullet kidneyPb^{0.244}$

Plotting the regression line

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

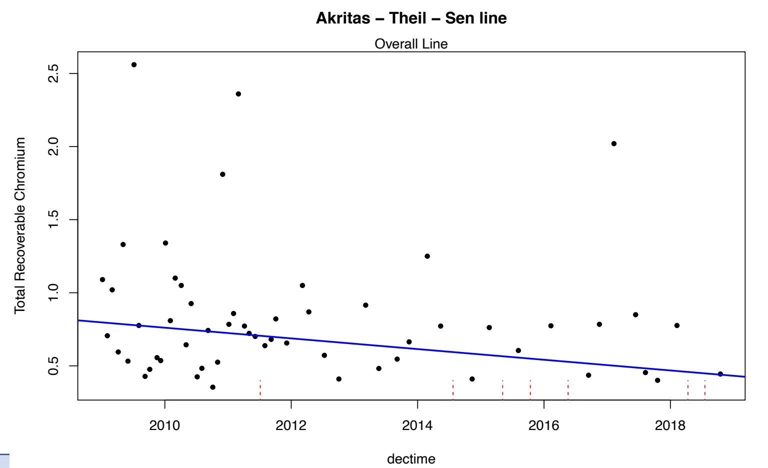
```
> cenxyplot(Kidney, KidneyCen, Blood,
   BloodCen, xlab = "Pb in Kidneys",
   ylab = "Pb in Blood")
> ik <- order(Kidney)
> lines(Kidney[ik],
  exp(predict(Pbreg)[ik]), col = "red")
```



9. Seasonal Kendall Test with nondetects

Nondetects influence the line and test.

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



 39

39

censeaken function

```
> 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

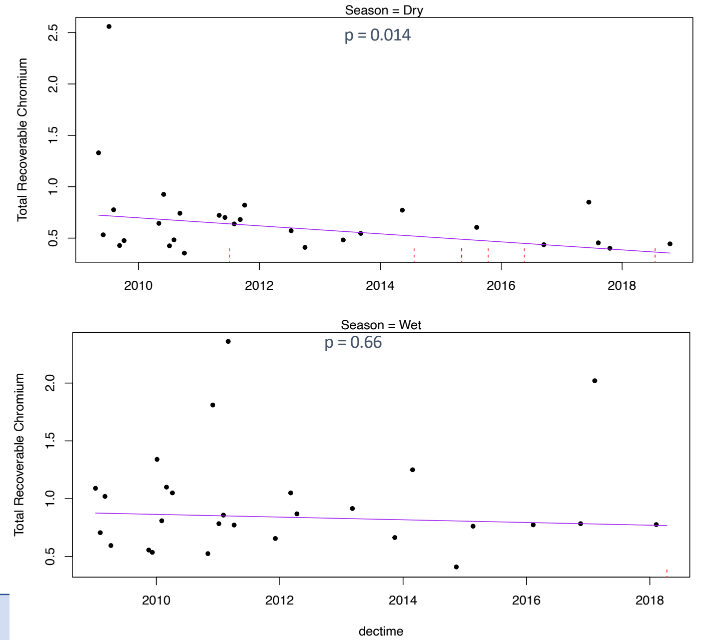
Significant trend overall. SK slope is -0.036 ug/L per year

 40

40

```
censeaken (dectime, `Total Recoverable Chromium`, CrND, group = Season, seaplots = TRUE)
```

Optional graphs for each season



41

41

For more information -- 9 free videos

https://practicalstats.com/videos/nadavids.html

1. The Cost of Complacency *
2. The Mystery of Nondetects -- How Censored Data Methods Work *
3. Testing Groups of Data with Multiple Detection Limits
4. Fitting Distributions to Data with Nondetects
5. Correlation and Regression for Data with Nondetects
6. Trend Analysis for Data with Nondetects
7. Incorporating > and < Values in Data Analysis
8. Matched Pair Tests with Nondetects
9. NADA2: Everything You Can Do Today with Nondetects *

* introductory

42

42

For even more information -- a free training course Nondetects And Data Analysis

Course Outline By Section

Section 1. Get Started with RStudio

Section 2. Detection and Reporting Limits

Section 3. Store Censored Data in Databases

Section 4. Plot Data with Nondetects

Section 5. Estimating Descriptive Statistics

Section 6. Intervals (Confidence, Prediction, Tolerance)

Section 7. Matched Pair Tests & Comparing Data to Standards

Section 8. Two-Group Tests with Nondetects

Section 9. Three+ Group Tests with Nondetects

Section 10. Correlation and Regression with Nondetects

Section 11. Trend Analysis with Nondetects

Section 12. Logistic Regression

Section 13. Multivariate Methods with Nondetects

https://practicalstats.com/training/

43

43

Summary: After 43 Years, These Two Principles Will Help Guide Data Analysis

1. Decide your objectives, which then determine which type of test to use, permutation tests for means or nonparametric tests for percentiles.
2. Use methods designed for censored nondetect data. Substituting a fraction of the DL and running a routine test may produce false positives OR false negatives.

Questions?

44

44

Thank you for attending

- The first half of this webinar is based on the book Statistical Methods in Water Resources, 2nd Edition by Helsel, Hirsch, Archfield, Ryberg and Gilroy. *USGS Techniques and Methods 4-A3* (2020). Especially see Chapter 4.
- The second half of this webinar is based on the book Statistics for Censored Environmental Data using Minitab and R by Helsel (2012).
- All opinions are my own and do not represent those of anyone else.

Questions?

Also see the free course & free webinars at: http://practicalstats.com

45

45