PracticalStats LLC

NADA2: Everything You Can Do Today With Nondetects

Dennis R. Helsel
PracticalStats.com

1

To download the pdf of webinar slides

- on the PracticalStats Live page, the link is just below the video feed
- on our YouTube channel, the link is in the ABOUT tab

2

You may submit questions during and after the webinar

- on the PracticalStats Live page, the link is just below the video feed
- on our YouTube channel, the link is in the first entry in the DISCUSSION tab

- Questions will be answered and posted within two days, along with the webinar recording, at our online Training Site, https://practicalstats.teachable.com
- Your email address will not be included in the posted answer file.

3

Outline of this presentation

Introduction -- What's wrong with Substitution?

A. A survey of statistical methods for data with nondetects -- without substitution.
B. The NADA2 package for R -- when it will become available.

The statistical term for data known only to be below (or above) a threshold is "Censored Data"

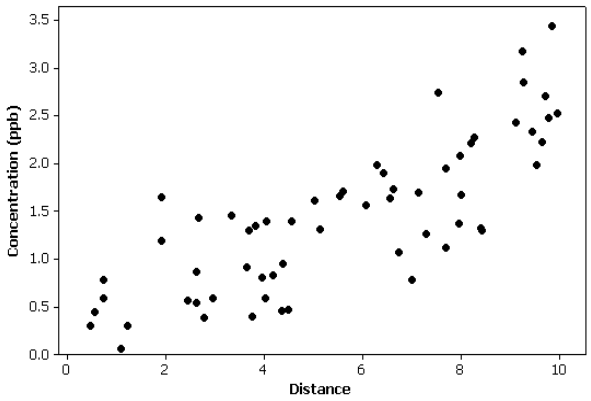
 4

4

What's Wrong With Substitution -- it adds unintentional patterns to data

Before censoring,

True correlation r=0.81



 5

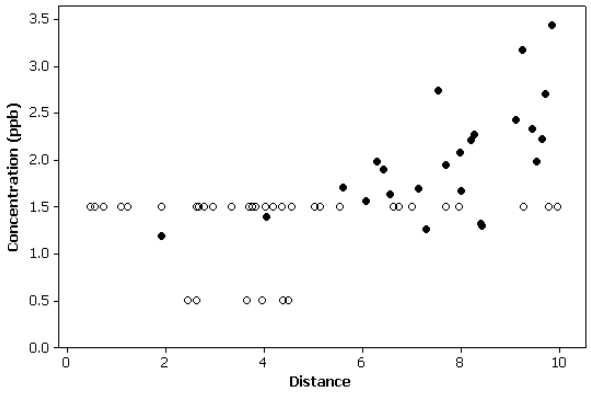
5

Standard deviation is lowered as the same value replaces numbers with variation

Don't just think about substitution for one measurement. The same number is substituted for multiple measurements, reducing variation.

Lower standard deviations mean:

- too-short confidence intervals
- incorrect p-values in tests
- incorrect decisions in regression
- incorrect surfaces in kriging
- and more.



 6

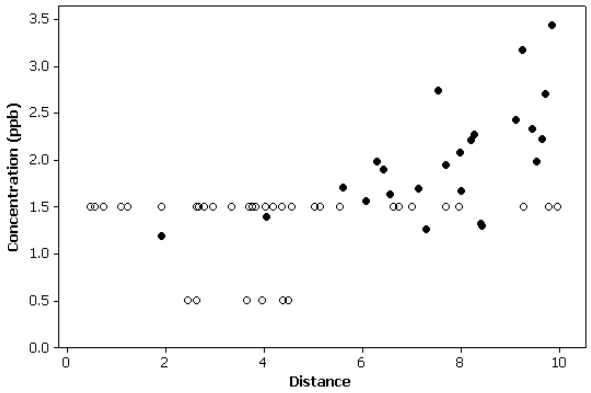
6

Lower correlation and regression slopes. It's the pattern: flat zero-slope lines

(Two DLs @ 3 and 1)
<3s become 1.5
<1s become 0.5

After substitution,
correlation is lowered to r=0.55

Slope is lowered as 0 slopes replace generally increasing slopes in the uncensored data



 7

7

A. Which statistical methods can incorporate nondetects now, without substitution?

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

1. Plot data and compare its shape to standard distributions
2. Estimate means, confidence intervals, UCLs
3. Run hypothesis tests between groups, compare data to standards
4. Compute correlation coefficients
5. Build regression models and evaluate if residuals match the assumed distribution
6. Compute nonparametric regression: the Theil-Sen line for censored data
7. Perform trend analyses (parametric and nonparametric)
8. Draw high-dimensional plots, cluster analysis, test hypotheses on multivariate data
9. Model with Kriging: spatial statistics

 8

8

More Information on Methods

for these topics is taught in our online course

Nondetects And Data Analysis

at https://practicalstats.teachable.com

 9

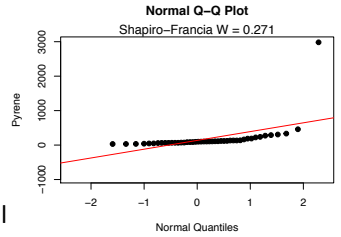
9

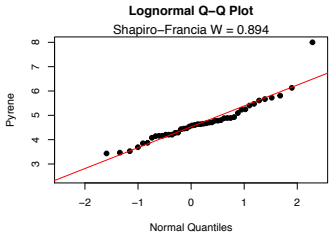
1. Compare data to a distribution's shape

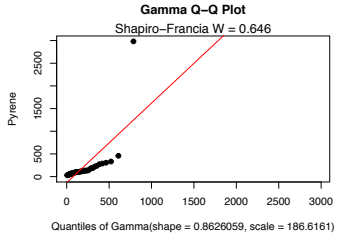
```
> cenCompareQQ (Pyrene, PyreneCen)
```

QQ Plots for
Pyrene concentrations in sediments

Built for censored data -- no substitution wanted. Points shown are detected data. Their positions ("Normal Quantiles", i.e. percentiles) take into account the presence of the NDs.







Best of the three distributions is the lognormal

 10

10

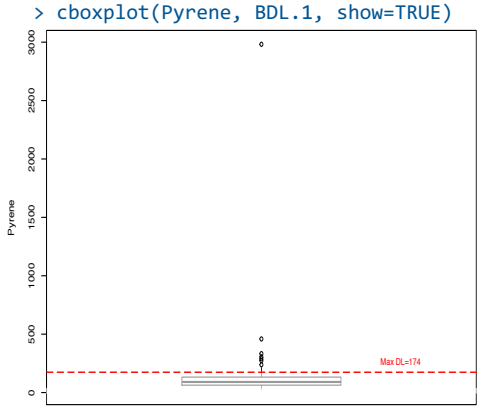
2. Compute mean, confidence limits

```
> enparCensored(Pyrene,PyreneCen, ci=TRUE, ci.method="bootstrap", n.bootstraps = 5000)

Censoring Level(s):  28  35  58  86 117 122 163 174
Estimated Parameter(s):          mean    = 164.09450
  (Kaplan-Meier)                 sd      = 389.41379
                                 se.mean =  49.75292

Sample Size:                     56
Percent Censored:                19.64286%
Confidence Interval Method:      Bootstrap
Confidence Level:                95%
Confidence Interval:             Pct.LCL = 100.10254
                                 Pct.UCL = 264.47772
```

```
> cboxplot(Pyrene, BDL.1, show=TRUE)
```



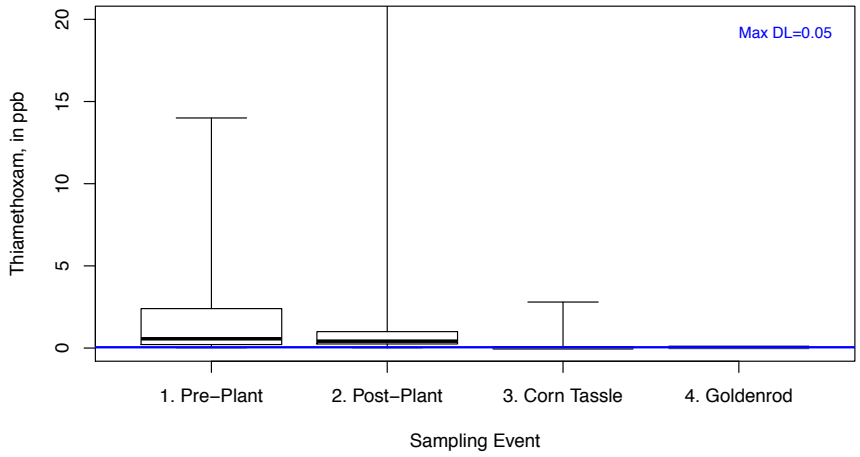
ROS estimates of boxplot below the Max DL

 11

11

3. Test Group Differences: Thiamethoxam on Pollen

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



 12

12

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

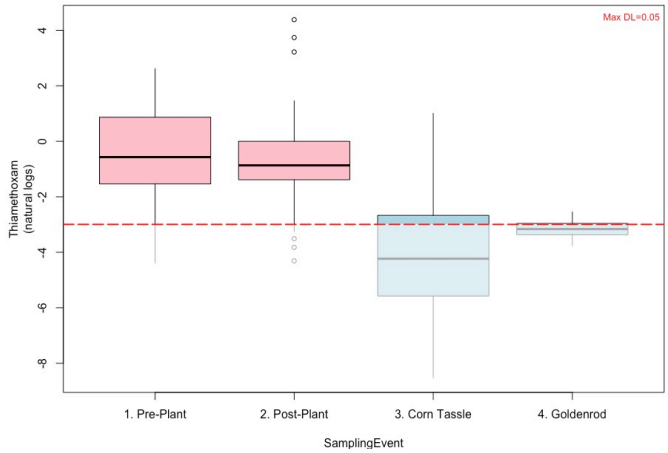
 13

13

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.



 14

14

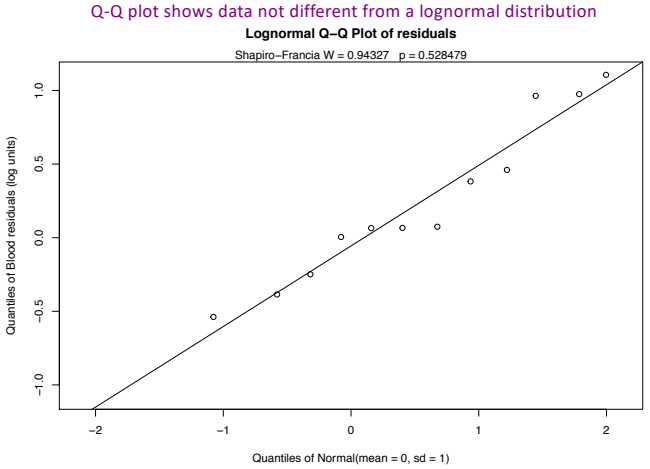
4. Correlation Coefficient: Likelihood R (analogous to Pearson's r)

Parametric method: coefficients fit by MLE
(Lead in blood and kidneys)

```
> cencorreg(Blood, BloodCen, Kidney)
 Likelihood R = 0.8236
 Rescaled Likelihood R = 0.8721
 McFaddens R = 0.714
 (Plot on next slide)

Chisq= 30.62 on 1 degrees of freedom,
 p= 3.14e-08     n= 27
```

Q-Q plot shows data not different from a lognormal distribution



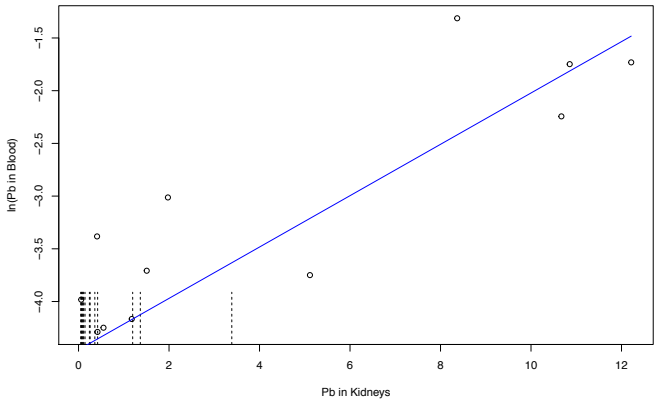
 15

15

Plot of the 2 variables

Plot of the logY regression in the units the regression wan run:
Notice the evidence from the nondetects -- they are much more frequent at low X than high X.
Nondetects shown as dashed lines.

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



 16

16

5. Regression with censored data

Regression by Maximum Likelihood Estimation: the cencorreg command.

```
> 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 = $e^{-4.457} \bullet kidneyPb^{0.244}$

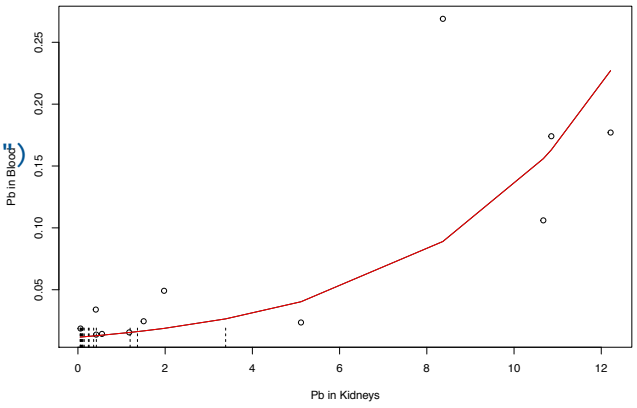
 17

17

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")
```



 18

18

5 (cont). Multiple Regression with Censored Data

To use the cencorreg function to do multiple regression you'll need to input the x variables as a single data frame. To create the data frame for all 4 x variables, and then for a 2 variable model:

```
> xvar4 <- data.frame(LandUse, PopDensity, PctIndLU, Depth)
> xvar3 <- data.frame(LandUse, PopDensity, Depth)
> xvar2 <- data.frame(PopDensity, Depth)
```

← Best model. Minimizes AIC

Which is the best model? I generally use the lowest AIC model:

```
> reg4 <- cencorreg(TCEConc, TCECen, xvar4)
 Likelihood R2 = 0.1075                    AIC = 395.8513
 Rescaled Likelihood R2 = 0.1326           BIC = 415.9318
 McFaddens R2 = 0.06833
```

 19

19

Four variable model: AIC = 395.8

```
> summary(reg4)
Call:
survreg(formula = "log(TCEConc)", data = "LandUse+PopDensity+PctIndLU+Depth",
    dist = "gaussian")
             Value Std. Error     z      p    Loglik(model)= -191.4
(Intercept) -5.38940    2.61512 -2.06  0.039    Loglik(intercept only)= -205.5
LandUse      0.32205    0.31035  1.04  0.299    Chisq=28.08 on 4 degrees of freedom, p=1.2e-05
PopDensity   0.21991    0.07829  2.81  0.005         Number of Newton-Raphson Iterations: 4
PctIndLU     0.03644    0.05274  0.69  0.490         n= 247
Depth       -0.00374    0.00238 -1.57  0.117
Log(scale)   1.02763    0.11058  9.29 <2e-16
Scale= 2.79
```

 20

20

Three variable model AIC = 394.3

- Is better than the 4 variable mode due to lower AIC. LandUse has a relatively high p-value. What about a 2-variable model?

```
> reg3 <- cencorreg(TCEConc, TCECen, xvar3)
 Likelihood R2 = 0.1057                    AIC = 394.3252
 Rescaled Likelihood R2 = 0.1305           BIC = 410.8924
 McFaddens R2 = 0.06718

> summary(reg3)
               Value Std. Error     z      p
(Intercept) -5.44065    2.62890 -2.07 0.0385
LandUse      0.33855    0.31107  1.09 0.2764
PopDensity   0.22621    0.07797  2.90 0.0037
Depth       -0.00367    0.00239 -1.54 0.1239
Log(scale)   1.02852    0.11059  9.30 <2e-16
Scale= 2.8

Loglik(model)= -191.7   Loglik(intercept only)= -205.5
        Chisq= 27.61 on 3 degrees of freedom, p= 4.4e-06
```


21

Two variable model AIC = 393.6

- This is better than the 3 variable model due to lower AIC
- Depth is now at p=0.06
- I generally keep variables with p < 0.10, as model selection stats like AIC and BIC generally underfit (choose too few explanatory variables)
- Just as in ordinary regression, R^2 increases with each added variable, so is no help in choosing a model.
- What about a 1-variable model, with just PopDensity?

```
> reg2 <- cencorreg(TCEConc, TCECen, xvar2)
 Likelihood R2 = 0.1012                    AIC = 393.5758
 Rescaled Likelihood R2 = 0.1249           BIC = 406.6296
 McFaddens R2 = 0.06413

> summary(reg2)
               Value Std. Error     z       p
(Intercept) -2.79067    0.81018 -3.44 0.00057
PopDensity   0.25959    0.07405  3.51 0.00046
Depth       -0.00434    0.00234 -1.85 0.06367
Log(scale)   1.03487    0.11068  9.35 < 2e-16

Scale= 2.81
Gaussian distribution
Loglik(model)= -192.3   Loglik(intercept only)= -205.5
        Chisq= 26.35 on 2 degrees of freedom, p= 1.9e-06
```


22

One variable model AIC = 395.7

- AIC is higher for the 1-varaible model. So AIC picks the 2-variable model.
- Summary: Choose the 2-var model.

```
> reg1 <- cencorreg(TCEConc, TCECen, PopDensity)
 Likelihood R = 0.2934                    AIC = 395.6935
 Rescaled Likelihood R = 0.3259           BIC = 405.2338
 McFaddens R = 0.2326

> summary(reg1)
survreg(formula = "log(TCEConc)", data = "PopDensity", dist =
"gaussian")
              Value Std. Error     z       p
(Intercept) -3.7343     0.7493 -4.98 6.2e-07
PopDensity   0.3087     0.0736  4.20 2.7e-05
Log(scale)   1.0418     0.1109  9.39 < 2e-16
Scale= 2.83

Loglik(model)= -194.3   Loglik(intercept only)= -205.5
        Chisq= 22.24 on 1 degrees of freedom, p= 2.4e-06
```

 23

23

6. Nonparametric Regression with Censored Data

Nonparametric method: the Akritas-Theil-Sen line.
The ATS command: ATS (Y, Ycen, X, Xcen)

```
> Pbk<- ATS(Blood, BloodCen, Kidney, KidneyCen)
  Akritas-Theil-Sen line for censored data
              ln(Blood) = -4.5128 + 0.295 * Kidney
              Kendall's tau = 0.4217   p-value = 0.00043
```

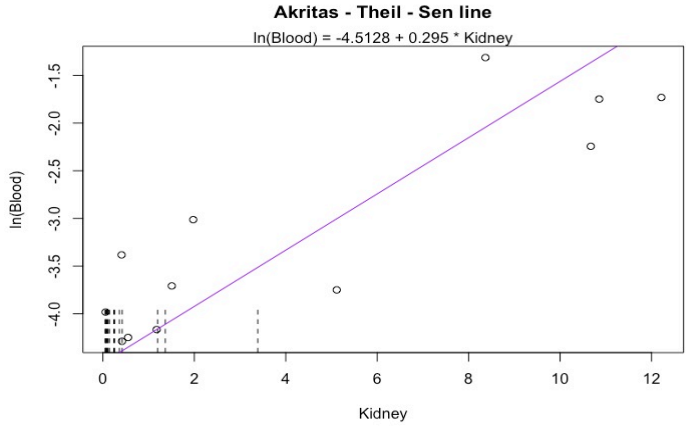
 24

24

The ATS line

```
> Pbk<- ATS(Blood, BloodCen, Kidney, KidneyCen)
```

Using log of the Y variable (the default) is the best straight line fit.

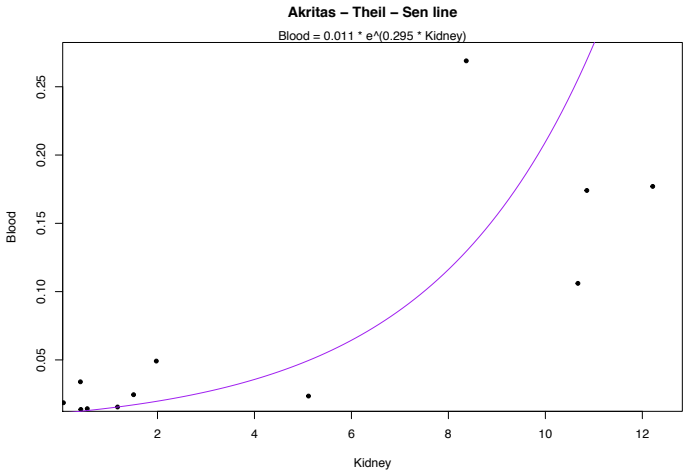


LLC 25

25

The ATS line is curved in the original units

```
> Pbk <- ATS(Blood, BloodCen, Kidney, KidneyCen, retrans = TRUE)
```



; LLC 26

26

7. Trend Analysis with a Covariate: "nonparametric multiple regression"

1. Compute a smooth of censored Y vs X, where X is a non-time covariate, using Generalized Additive Models (GAM)
2. Compute an ATS on the residuals -- Kendall's tau test of change in residuals over time. Slope is still in Y units per time.

   R function centrend (Y, Y.cen, X, time)

   time is often as decimal time (i.e. 2013.5 for halfway through the year)

 27

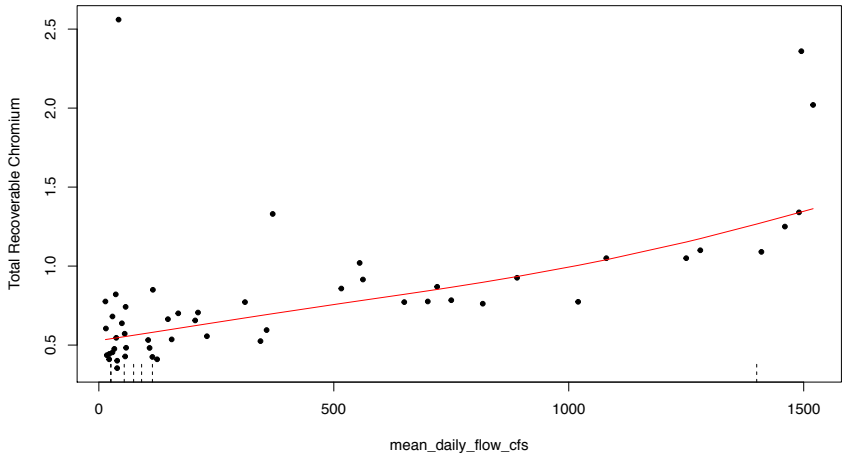
27

Note that the relationship is not linear

```
> resid.trend <- centrend(`Total Recoverable Chromium`, CrND, mean_daily_flow_cfs, dectime)
```

Censored GAM Smooth of Cr concentrations vs. mean daily flow.

dashed lines are the censored data

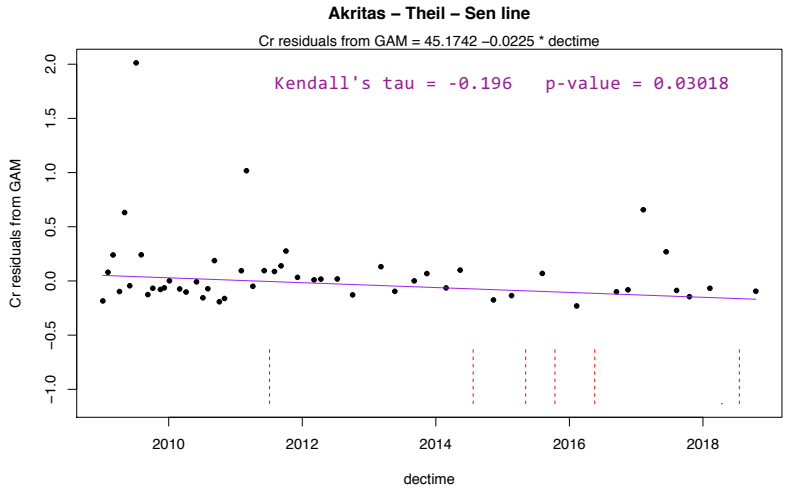


 28

28

ATS (Mann-Kendall) test on censored GAM residuals

Low censored data (dashed lines) contribute to the test results. Note that they appear mostly at later times, increasing evidence for a downtrend.



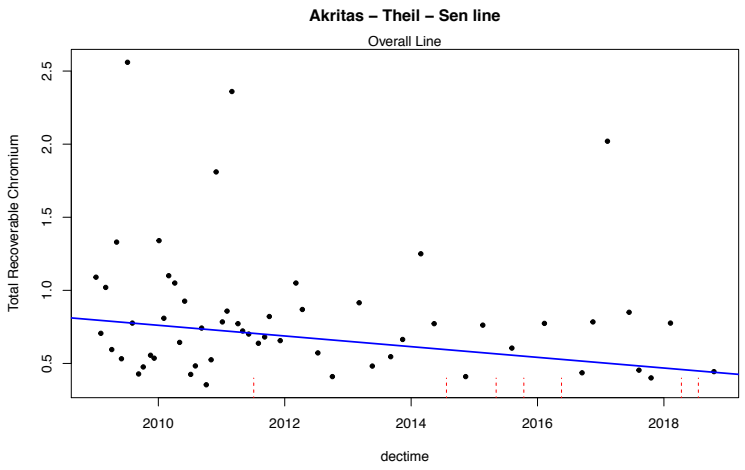
 29

29

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.



30

 30

30

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

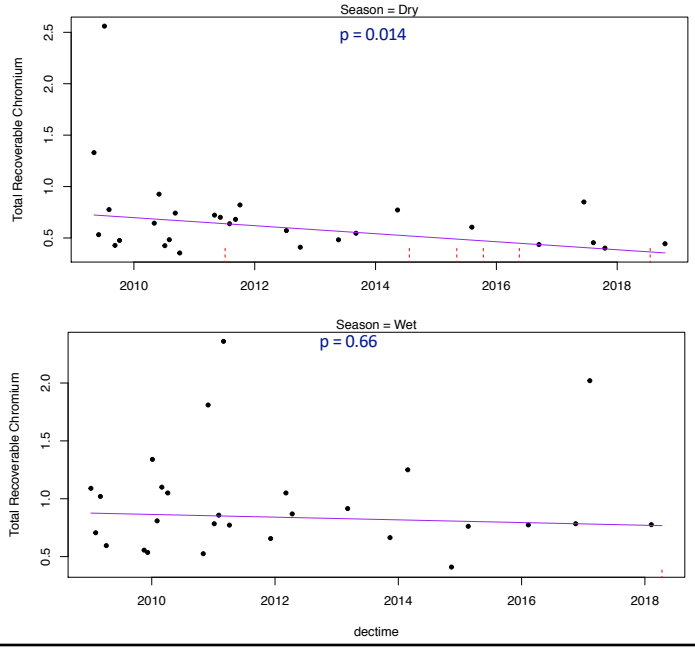
Significant trend overall. SK slope is -0.036 ug/L per year

 31

31

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

Optional graphs for each season



 32

32

8. Multivariate Methods

Multivariate methods provide a clue for how to deal with nondetects in a 'big data' context

 33

33

Three methods for incorporating censored data

a. BINARY. Convert data to detect/nondetect (at the single, or highest of multiple, DLs). Cluster, test for differences in proportions of 0/1, etc.

b. RANKS. Re-censor all values below a single (or the highest of multiple) detection limit as equal. Rank within each variable and perform MV analyses on the ranks.

c. USCORES for continuous data. Use u-scores to define relationships within multiply-censored variables. Perform MV analyses on the uscores (or their ranks).

 34

34

8a. BINARY. Simple matching coeff (similarity measure) between rows j and k

$$S_{jk} = {}^{(a+d)}/_{(a+b+c+d)}$$

where $a_{jk} = (1,1)$, $d_{jk} = (0,0)$ Matches

$b_{jk} = (0,1)$ $c_{jk} = (1,0)$ Mismatches

$1 = \text{value} < \text{RL}$, $0 = \text{value} \geq \text{RL}$

Dissimilarity (distance) measure $D_{jk} = 1 - S_{jk}$

 35

35

Example data: FishDDT.xls

```
> head(FishDDT)
```

concentrations group indicators (1 = nondetect)

	opDDD	ppDDD	opDDE	ppDDE	opDDT	ppDDT	Age	Date	oD_LT1	pD_LT1	oE_LT1	pE_LT1	oT_LT1	pT_LT1
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	1	1	1	14	1	1	Young	1996	1	1	1	0	1	1
2	1	42	8.4	130	1	31	Mature	1990	1	0	0	0	1	0
3	5.3	38	1	250	1	11	Mature	1994	0	0	1	0	1	0
4	1	12	1	57	1	1	Mature	2002	1	0	1	0	1	1
5	1	1	1	16	1	1	Young	2000	1	1	1	0	1	1
6	1	1	1	1	1	1	Young	2000	1	1	1	1	1	1

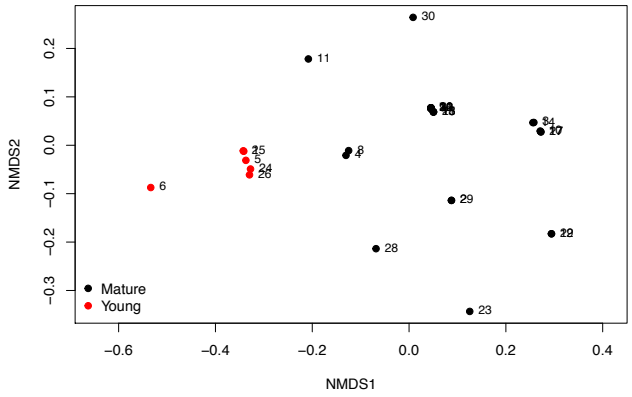
 36

36

NMDS on simple-matching distances

```
> DDTcen <- data.frame(oD_LT1, pD_LT1, oE_LT1, pE_LT1, oT_LT1, pT_LT1)
> binaryMDS (DDTcen, Age, title = "NMDS of DDT in fish")
```

- Create a data frame just with the indicators
- Distances between points on the 'map' are in the same rank order as distances ($1-S_{jk}$) between rows
- There is definitely a left-right differentiation between Mature and Young fish.
- The overlap of some sites is due to ties between pairs -- there's 6 variables, each with 2 possible values, 0 and 1.



 37

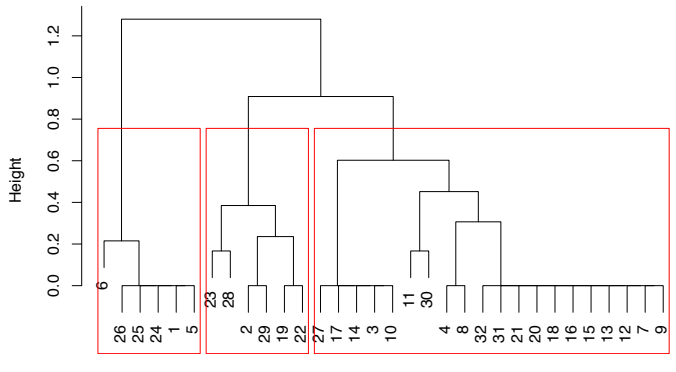
37

Clustering 0/1 data

```
> binaryClust(DDTcen, ncluster = 3)
```

All "Young" fish are together in the left-most cluster. The other two clusters are all "Mature" fish.

3-cluster solution



 38

38

8b. Distance Measure for Ordinal (Ranked) Methods

- Rank data within each variable (column). The `ordranks.R` script makes this quick.
- Input to ordranks is a dataframe with both the concentration and associated indicator columns. Format can be `C1 I1 C2 I2 C3 I3` (`paired - TRUE`, the default) or `C1 C2 C3 I1 I2 I3` (`paired = FALSE`).
- If there are multiple RLs, the highest must be used and data re-censored to it. All <RL are assigned a tied rank. The `ordranks.R` script makes this quick, it will re-censor each variable for you at its highest RL.
- Euclidean distance E on ranks used for MV analysis

$$E = \sqrt{\sum_i (y_{i1} - y_{i2})^2}$$

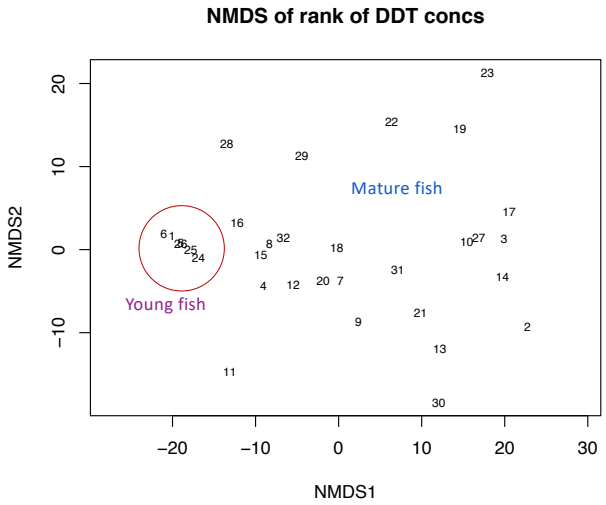
 39

39

NMDS on Euclean rank matrix

- `> ddt.eumds <- metaMDS(euclid.ddt)`
- `> p1=ordiplot(ddt.eumds, type="t", main="NMDS of rank of DDT concs")`

- Young and mature fish are distinguished using Euclidean distances on ranks.
- More differentiation between sites with ordinal data than for binary data.



40

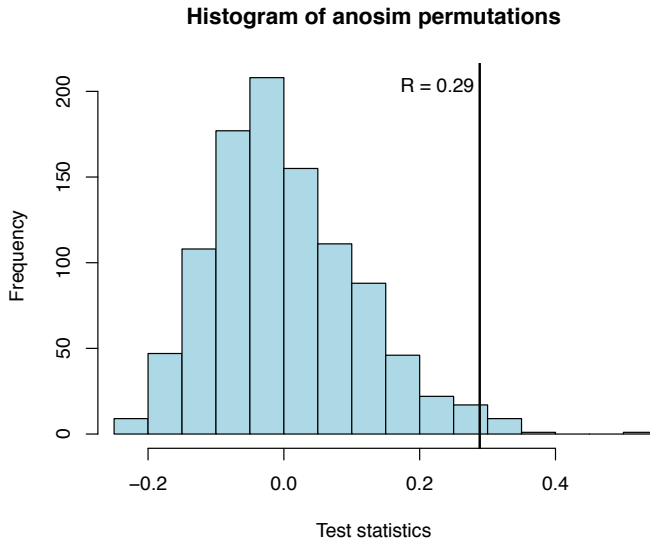
40

ANOSIM test: do the patterns of ddt metabolites differ in mature vs. young fish?

```
> rnk.ano <- anosim(euclid.ddt, FishDDT$Age)
> rnk.ano
anosim(x = euclid.ddt, grouping = FishDDT$Age)
Dissimilarity: euclidean

ANOSIM statistic R: 0.2884
      Significance: 0.013
Permutation: free
Number of permutations: 999
> anosimPlot(rnk.ano)
```

DDT metabolite patterns in the two groups differ significantly

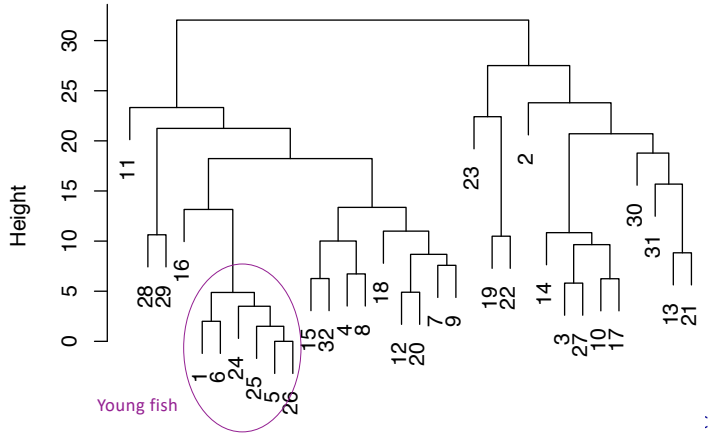


41

41

Cluster on ranks of concentrations

```
> rankclust <- hclust(euclid.ddt, method = "average")
> plot(rankclust)
```



42

42

8c. Uscores. For variables with multiple RLs

- U-scores u_i are a measure of order within a variable
- They are the sum of the sign of differences between the ith observation and all other observations in the data set
- To say it another way, the u-score is the number of observations the i^{th} value is above, minus the number of observations the i^{th} value is below. The larger the i^{th} observation's value, the higher the u-score. The median observation would have a u-score of 0.
- u-scores are the basis for the Mann-Whitney test, are the numerator in computation of Kendall's tau correlation, and are related to Kaplan-Meier and other nonparametric methods
- The distance measure will be the Euclidean distance on either the u-scores, or the ranks of the u-scores. The latter insures all values are positive, which may be required by some multivariate software.

$$u_i = \sum_{i \neq k} sign(x_i - x_k)$$

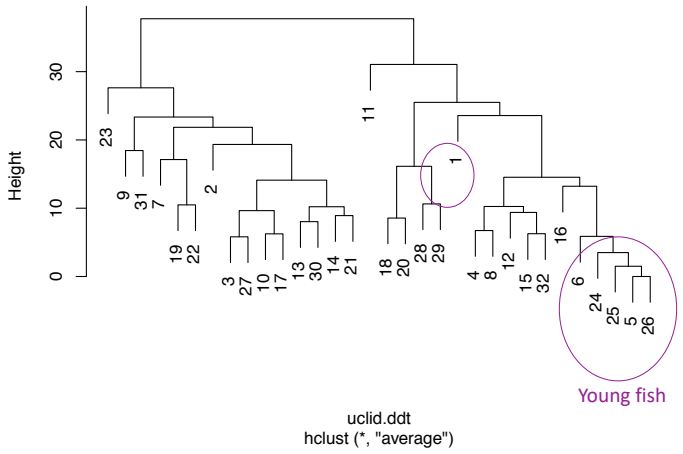
- I changed site 1's data so that the DL = 5 instead of 2 ug/g.

 43

43

Cluster analysis using uscores

```
> uclust <- hclust(uclid.ddt, method = "average")
> plot(uclust)
```

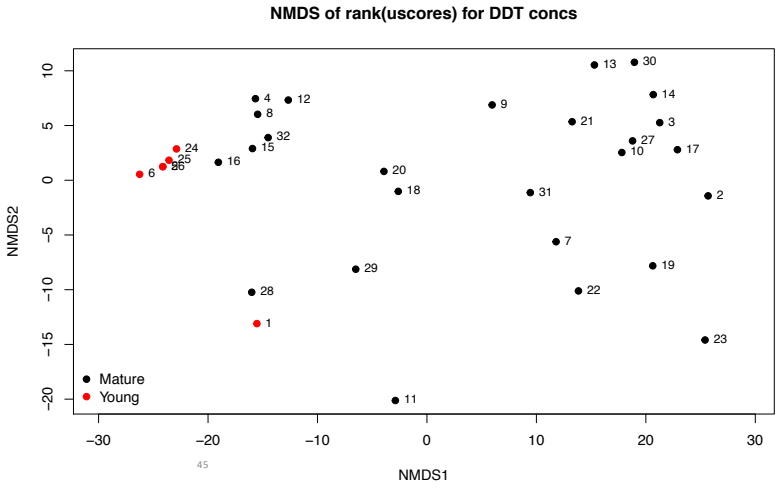


 44

44

2. NMDS on u-scores

```
•> uMDS(u_scores, group = FishDDTalt$age, main="NMDS of rank(uscores) for DDT concs")
```



Better separation between Young fish (red dots) than with binary NMDS.

Site 1 is offset because its concentrations are <5 instead of <2.

45

45

Trend test on u-scores

```
> ddt.utrend <- mantel(time.dist, uclid.ddt, method="kendall", permutations =9999)
> ddt.utrend
Mantel statistic based on Kendall's rank correlation tau
Call:
mantel(xdis = time.dist, ydis = uclid.ddt, method = "kendall",  permutations = 9999)
Mantel statistic r: 0.432
      Significance: 1e-04            # very significant. There is a trend in the pattern of DDT and
Upper quantiles of permutations (null model):        its degradation products over time.
   90%    95%  97.5%    99%
0.0415 0.0596 0.0773 0.1038    # test statistic of 0.432 is >> the 99th percentile of the permutations
Permutation: free
Number of permutations: 9999
```

 46

46

9. Spatial Statistics

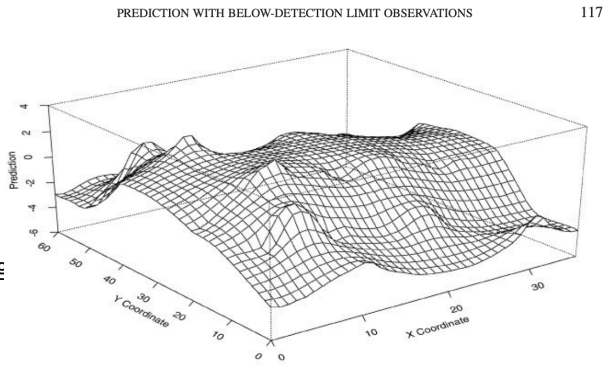
- Spatial applications model the cross-correlations between locations, incorporating that information into predictions
- Generalized Least Squares is the primary statistical procedure. In spatial applications it often goes by the name 'kriging'.
- For censored data the equations are solved by maximum likelihood ("expectation maximization" or EM) methods rather than least-squares
- Are also binary methods (above DL | below DL) for a single DL using indicator kriging, and
- Rank methods (percentiles) for Kaplan-Meier percentiles using probability kriging
- These are NOT part of the NADA2 package

 47

47

1. Kriging with Maximum Likelihood

- Expectation Maximization is most common approach. Define a function of best fit and then iteratively maximize the function. CensSpatial package in R.
- Incorporates explanatory variables (trend) and spatial correlation (variogram)
- Assumes normal distribution of residuals (leftovers from the model)
- Logarithms often taken first due to lack of realism for normality assumption
- Censored data modeled using data augmentation (DA) -- iterative cross correlation with nearby locations
- Resulting surface is a geometric mean due to the log transformation

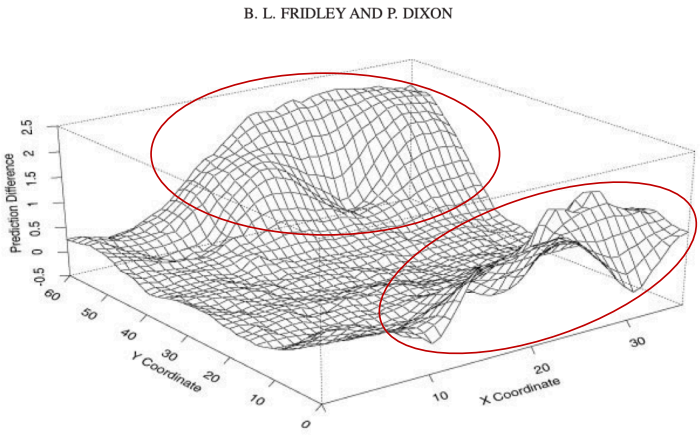


 48

48

Kriging with Maximum Likelihood

- DA found to have less bias than substitution of DL/2, etc. (Fridley and Dixon, 2007)
- Large differences between krige and DL/2 represent unrealistic DL/2 estimates "far away from the road".
- Substitution incorrectly resulted in larger cleanup regions and so higher estimated costs.
- Ordoñez et al (2018) found Data Augmentation to have lower AIC/BIC (errors) than DL/2 or other substitutions.



 49

49

B. The NADA2 Package for R

- The NADA2 package is available now with our *Nondetects And Data Analysis* online course at https://practicalstats.teachable.com
- NADA2 will be available through the CRAN site (R software site) in 1st Qtr of 2021
- The censSpatial package (kriging) and EnvStats package are available now on the CRAN site.

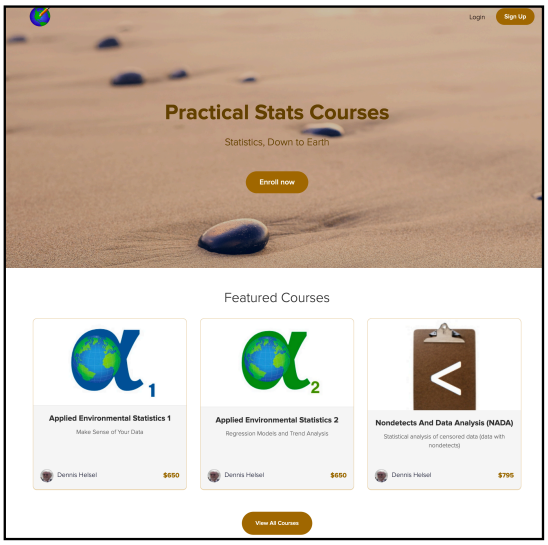
50

For more information: videos on now available for streaming at no charge on our Online Training Center

Practicalstats Online Training Center:
https://practicalstats.teachable.com/

Click on the "View All Courses" button at the bottom to list the free videos.
Let colleagues know about them.

Our full 'semester' course Nondetects And Data Analysis $
that covers the full range of statistical methods for nondetects, all without substitution of DL/2, is available there now.



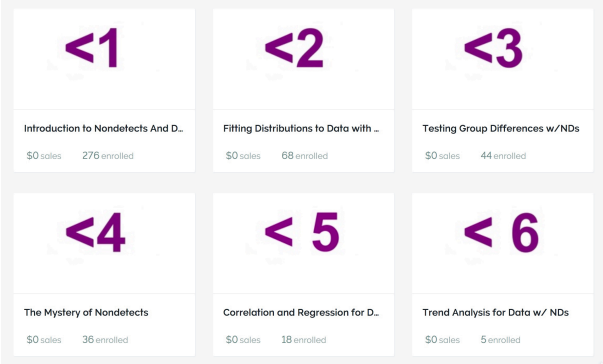
 51

51

Today's video will soon be available for streaming on our Online Training Center

… along with 8 other videos freely available there on the topic on statistics for data with nondetects.
https://practicalstats.teachable.com/

The recording of today's webinar will become <9 .

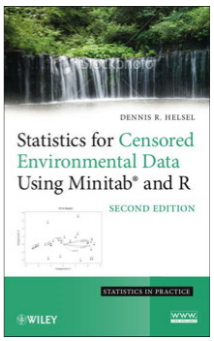


 52

52

Thank you for attending

- More on methods for nondetects can be found in my book Statistics For Censored Environmental Data Using Minitab and R by Dennis Helsel (2012)
- Our online course Nondetects And Data Analysis is available on our Training Site, https://practicalstats.teachable.com



Questions from this webinar?

Use the link on our Practical Stats Live page or the DISCUSSION tab in our YouTube Channel and I'll post all the questions with answers on our Training Site along with the recording of this webinar.

Thank you for attending today!

 53

53