Testing Groups of Data with Multiple Detection Limits

Dennis R. Helsel

A sampling of the online course

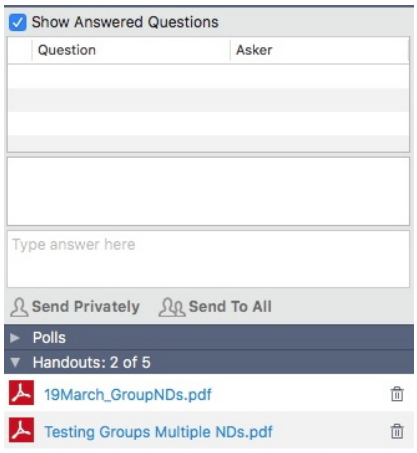
Nondetects And Data Analysis

Important Details -- your User Panel

The Questions pane: where you type questions and send to me. I'll answer as many as I can at the end of the webinar.

Handouts pane: The first is our most recent newsletter. The second is the file of PowerPoint slides for this webinar, two per page. Click on them to download before you leave the webinar.

If you are listening to the recording, you will find the 1st handout in our newsletter archive: http://practicalstats.com/news/archive.html
The 2nd is on our Downloads page: http://practicalstats.com/info2use/downloads.html



2

Testing Groups of Data with Multiple Detection Limits

Dennis R. Helsel

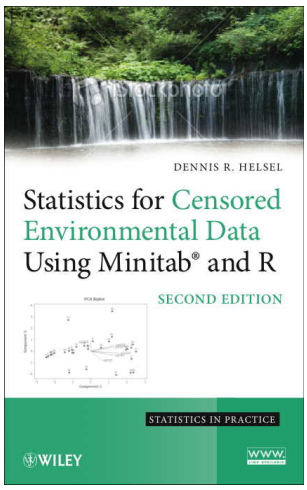
A sampling of the online course

Nondetects And Data Analysis

For more on stats for data with NDs:

Statistics for Censored Environmental Data
(the second edition)

by Dennis R. Helsel
(2012)



4

Methods for Testing Differences Between Groups

1. Parametric. Tests differences in group means - "Does one group have a higher mean than another group?" You must designate the assumed distribution that best matches the shape of your data.
2. Nonparametric. Tests differences in percentiles - "Is one group shifted higher than another?" No shape is assumed or necessary.

5

Data: Ontario 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 4 stages of plant growth (pre-plant, post-plant, corn tassel, goldenrod)?

Source: Ontario Ministry of the Environment, Conservation and Parks

6

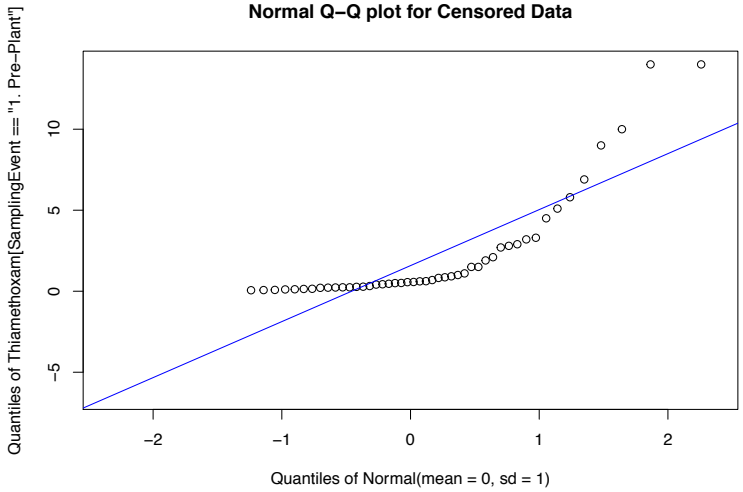
Format for reading in data with nondetects

```
> head(Pollen_Thiamethoxam)

row   Thiamethoxam  ThiaCens   SamplingEvent
              <dbl>     <dbl>   <chr>
 1  <0.05     0.05         1   3. Corn Tassle
 2  <0.05     0.05         1   3. Corn Tassle
 3            0.86         0   1. Pre-Plant
 4  <0.05     0.05         1   3. Corn Tassle
 5  <0.05     0.05         1   4. Goldenrod
 6            1            0   1. Pre-Plant

. . . . . 204 rows
```

Indicator column format
1. column of concs or DLs (Thiamethoxam)
2. column of 1/0 indicators, where 1 designates a DL and 0 designates a detected conc in column 1. (ThiaCens)

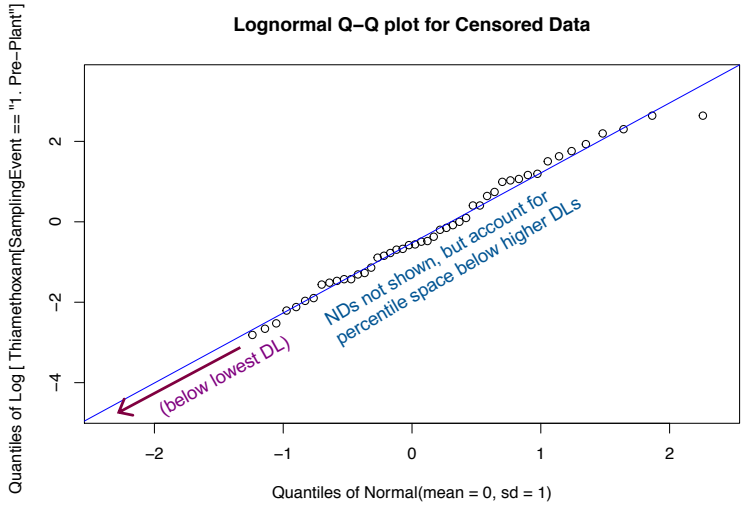
Q-Q plot to see fit of distribution to data with NDs

- Compares quantiles of detected observations to quantiles of the fitted distribution.
- Nondetects not plotted, but space for them left so that quantiles of detected observations are correct.
- Straight line represents the fitted distribution.
- These data are curved -- not a good fit to a normal distribution.



Thiamethoxam Data Fit Well by Lognormal Distribution

- Straight line pattern of data -- a good fit to a lognormal distribution.
- Will use the lognormal distribution for parametric methods

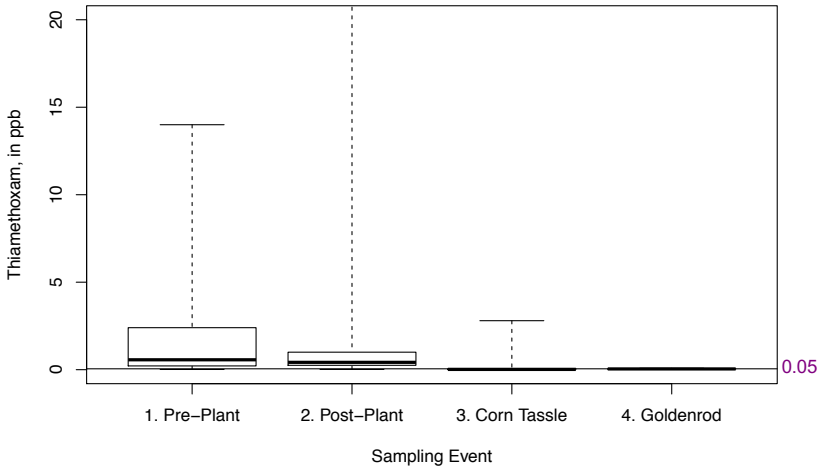


9

Concentrations are skewed

```
> cenboxplot (Thiamethoxam, as.logical(ThiaCens),
    as.factor(SamplingEvent), ylab = "Thiamethoxam, in ppb", xlab
    = "Sampling Event", log=FALSE, ylim =c(0,20))
```

command from the NADA package

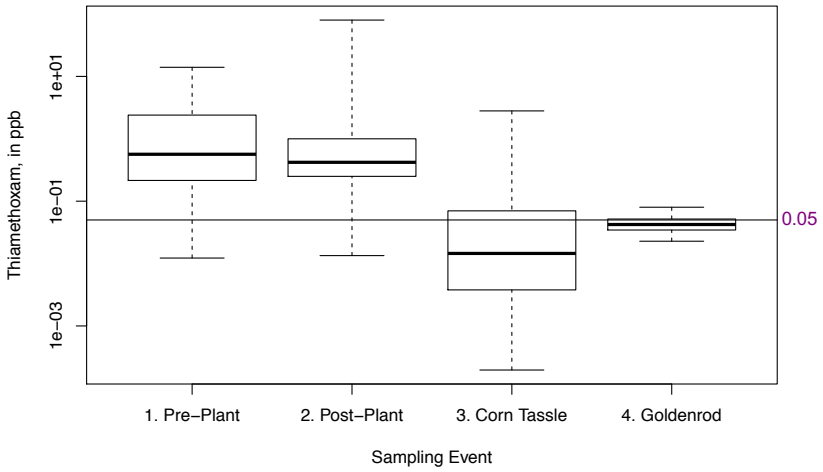


10

Logs of Concentrations very symmetric

```
> cenboxplot (Thiamethoxam, as.logical(ThiaCens),
   as.factor(SamplingEvent), ylab = "Thiamethoxam, in
   ppb", xlab = "Sampling Event")
```

command from the NADA package



11

Parametric Method -- Maximum Likelihood Estimation (MLE)

- Starts with the observed data (including nondetects), and your choice of which distribution should be used
- Given the observed data, what values for parameters (mean, standard deviation) for that distribution are most likely to have given rise to these data?
- Optimization performed to maximize the fit between observed data and the parameters
- Error in the fit is computed as a 'log likelihood', which is minimized as the parameters are changed in value until the best set of parameters are chosen
- No values are substituted for nondetects! Instead…….
- For censored data, the fit is to both observed detected observations and to the observed proportions of data (detects and nondetects) below each detection limit.

Parametric Method -- Maximum Likelihood Estimation (MLE)

- No values are substituted for nondetects. Information provided by nondetects is the observed percent below each detection limit.

For example, if there were two DLs, 0.05 and 0.10,

% <0.05 = (the number of observed <0.05) / n

% <0.10 = sum [# <0.05 + # <0.10 + # detects 0.05-0.099] / n

This is how the information in nondetects is used, without substituting values that you don't know for them.

How the MLE test is computed

- The log-likelihood of a "no model" fit (all group means = overall mean of the dataset) is compared to the log-likelihood for the "model" fit using the observed group means.
- For k = 4 groups there are k-1=3 degrees of freedom used by the model, just as in ANOVA.
- A Chisquare statistic is computed as -2*(difference in the two log likelihoods).
- The Chisquare statistic (the signal strength for the difference in means) is compared to a chi-square distribution with k-1 degrees of freedom to determine the p-value
- If the Chisquare statistic is small (not much evidence for a difference in group means), the p-value is large (>0.05) and the group means are not significantly different from one another.
- If the Chisquare statistic is large, the p-value is small (<0.05) and at least one group mean differs from the others.
- This is an overall test -- no statement of which group means differ from the others is made.

14

MLE "ANOVA" on logs of Concentration

available when taking the NADA online course

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


MLE test of mean natural logs of CensData: Thiamethoxam by
Factor: SamplingEvent

Assuming lognormal distribution of CensData

Chisq = 146.3  on 3 degrees of freedom

 p = 1.64e-31
```

-- the mean logs (geometric means) differ

Pre-A	Post-A	Corn Tassle B	Goldenrod B

```
      Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts

Fit: survreg(formula = logCensData ~ Factor, dist =
     "gaussian")

Linear Hypotheses:
```

	Estimate	Std. Error	z value	Pr(>|z|)
2. Post-Plant - 1. Pre-Plant == 0	-0.2197	0.3348	-0.656	0.912
3. Corn Tassle - 1. Pre-Plant == 0	-3.3733	0.3823	-8.823	<0.001 ***
4. Goldenrod - 1. Pre-Plant == 0	-4.4589	0.4932	-9.040	<0.001 ***
3. Corn Tassle - 2. Post-Plant == 0	-3.1536	0.3795	-8.310	<0.001 ***
4. Goldenrod - 2. Post-Plant == 0	-4.2392	0.4909	-8.636	<0.001 ***
4. Goldenrod - 3. Corn Tassle == 0	-1.0857	0.5004	-2.169	0.128

15

MLE Multiple Comparisons

- For each each comparison of two groups' means, an estimate of difference (mean1 - mean2) is computed along with its standard error.
- If zero is outside the confidence interval on that estimate, the two groups differ and p-values are below 0.05.
- p-values are adjusted for the multiple k(k-1)/2 comparisons of k group means that are made using Tukey's method
- Here 4 of the 6 differences in means are significant.

Pre-A	Post-A	Corn Tassle B	Goldenrod B

```
Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts

Fit: survreg(formula = logCensData ~ Factor, dist =
     "gaussian")

Linear Hypotheses:
```

	Estimate	Std. Error	z value	Pr(>|z|)
2. Post-Plant - 1. Pre-Plant == 0	-0.2197	0.3348	-0.656	0.912
3. Corn Tassle - 1. Pre-Plant == 0	-3.3733	0.3823	-8.823	<0.001 ***
4. Goldenrod - 1. Pre-Plant == 0	-4.4589	0.4932	-9.040	<0.001 ***
3. Corn Tassle - 2. Post-Plant == 0	-3.1536	0.3795	-8.310	<0.001 ***
4. Goldenrod - 2. Post-Plant == 0	-4.2392	0.4909	-8.636	<0.001 ***
4. Goldenrod - 3. Corn Tassle == 0	-1.0857	0.5004	-2.169	0.128

16

The two scripts used here come with our NADA online training course

If you want to write your own, here are the commands used:

cenanova	cen1way
based on survreg command in the survival package of R	based on survdiff command in the survival package
multiple comparisons based on the glht command in the multcomp package of R	multiple comparisons based on the pairwise_survdiff command in the survminer package

17

Nonparametric Peto-Peto test (cen1way)

- Nonparametric MLE test. Is a type of "linear rank test"
- Scores (ordered values like percentiles or ranks) are computed for the uncensored data and separately, the censored nondetect data. No distribution assumed.
- From the scores are computed a log-likelihood, a joint (detects and nondetects) measure of error
- The log-likelihood of a model with group differences is compared to a null situation of no group differences
- If the group difference model has significantly lower error than the null model, the p-value is small, and the group assignment is explaining part of the variation of the data
- No values are substituted for nondetects!
- There are several similar tests which differ in the details of their score function. These include the Peto-Peto test, the Tarone-Ware test and the logrank test. For data which look something like a lognormal distribution the first two have more power than the logrank test

18

Computing the Peto-Peto test statistic

- A Chisquare statistic is computed as -2*(difference in the two log likelihoods), -2*($loglikelihood_{null}$ - $loglikelihood_{model}$).
- The computed Chisquare statistic is compared to a chi-square distribution with k-1 degrees of freedom to determine the p-value
- If the Chisquare statistic (the signal of group differences) is small, the p-value is large (>0.05) and the group means are not significantly different from one another.
- If the Chisquare statistic is large, the p-value is small (<0.05) and at least one group mean differs from the others.
- This is an overall test -- no statement of which group means differ from the others is made. That is done with multiple comparisons.

19

Peto-Peto test of Difference in Group Concentration Percentiles

available when taking the NADA online course

```
> 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-A	Post-A	Corn Tassle B	Goldenrod B

20

Graph the Data: Sample CDFs Incorporating Nondetects

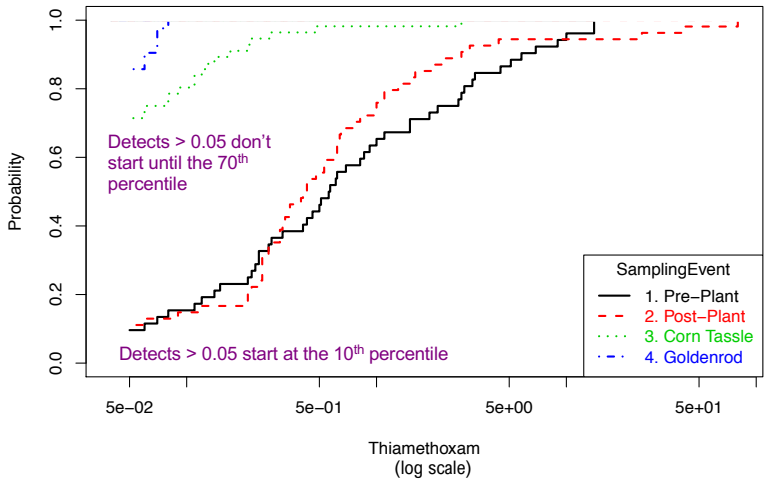
```
> plotcdf (Thiamethoxam, ThiaCens, SamplingEvent, logscale=TRUE)
```

Data shown as step function.

CDF is a plot of quantiles. Probability = 0.5 is a median, etc.

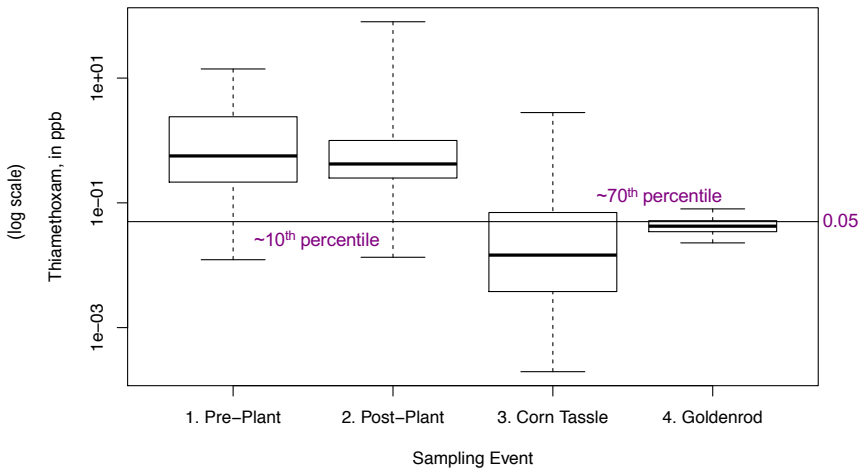
Higher quantiles for a given probability → cdf plots to the right.

Pre- and Post- Plant are similar. All other comparisons appear different.



Graph the Data: or Use Censored Boxplots

```
> cenboxplot (Thiamethoxam, ThiaCens, SamplingEvent, ylab
   = "Thiamethoxam, in ppb", xlab = "Sampling Event")
```



Conclusions

1. Hypothesis test methods are available for censored data WITHOUT substituting fabricated values for nondetects
2. These 'survival analysis' methods work well for both one and multiple DLs
3. Use parametric methods if you want to test for differences in means, and you feel confident that a particular distribution fits the data well (transformations change the meaning of a mean)
4. Use nonparametric methods if you want to test for whether some groups have generally higher values than others (different percentiles). No distribution needs to be assumed.
5. Multiple comparison tests are available for both types of methods.

23

Much More In Our Online Training Courses

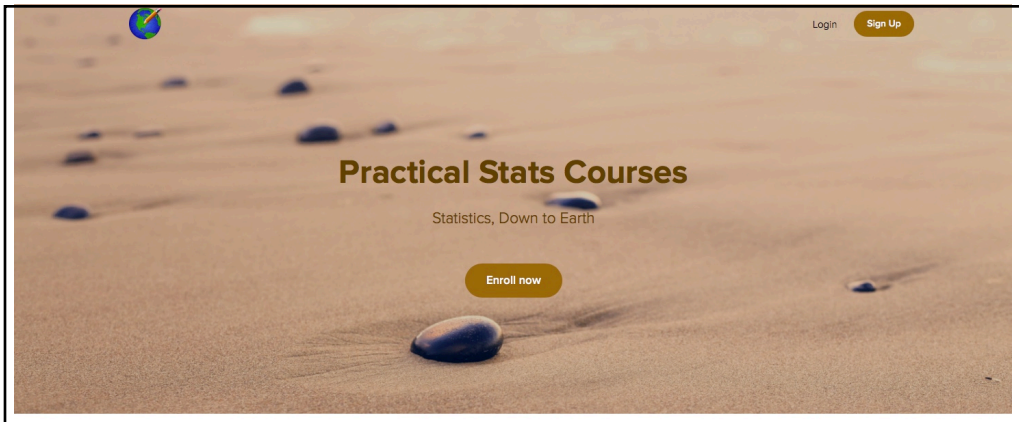
Coming Soon – Nondetects And Data Analysis (NADA) Online Course

Estimation, Hypothesis Tests, Regression, all without substitution, for data with nondetects

Our Training Site: see

https://practicalstats.teachable.com

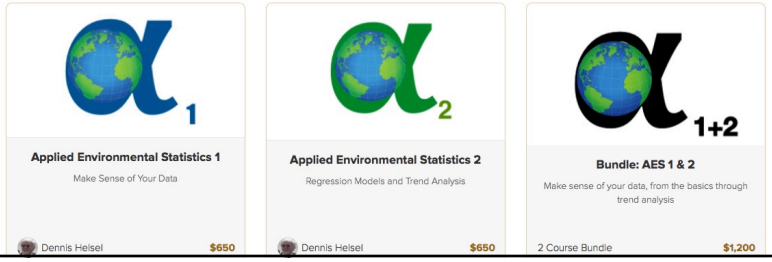
24



© PracticalStats.com

Our Training Site

Featured Courses



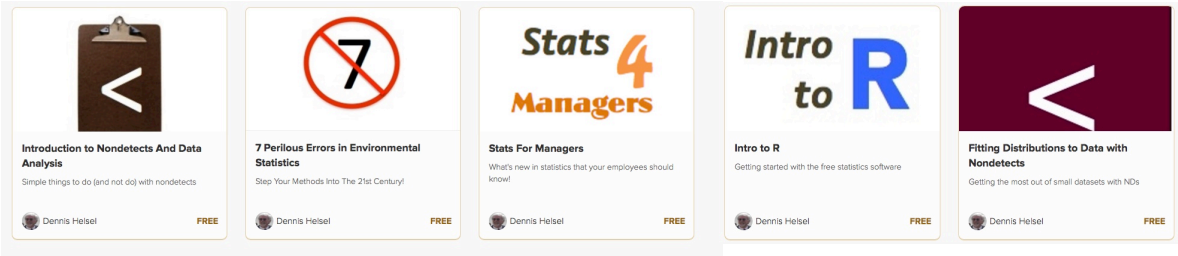
Applied Environmental Statistics

A Complete Survey of Environmental Stats

25

© PracticalStats.com

Free Webinars on Our Training Site



https://practicalstats.teachable.com

A recording of this webinar will be there tomorrow

26

Our Next Webinar

Forty Years of Water Quality Statistics: What's Changed, What Hasn't?

on April 23, 2019 11:00 AM MDT

You'll receive an email tomorrow from ask@practicalstats.com with the link to register:

https://attendee.gotowebinar.com/register/926744663037762061

The incentive for my forty years of mentoring/cajoling others to use good statistical methods is that advances in statistics have taken decades to reach the environmental science community. We don't read the necessary literature. Our training is often outdated. What we don't know does hurt us – it constrains our ability to see and quantify the important effects that are the objectives of our studies. Surveying the use of statistics in water quality over the last forty years (and I may have read one of your publications along the way), I discuss what advances have been made, what common errors are still being made, and where the field of applied statistics is heading into the future.

[A short version of this talk will be given next week at the National Water Quality Monitoring Conference]

27

Questions?

Thank you for attending!

Dennis Helsel

PracticalStats.com

28