

Incorporating $>$ and $<$ Values in Data Analysis

Dennis R. Helsel

PracticalStats.com



© 2020 PracticalStats.com

1

Objectives for this webinar

1. Show examples of what is available for incorporating $<$ and $>$ data into interval-censored methods to compute summary statistics, hypothesis tests and regression.
2. Discuss what interval-censoring is and how it is represented in software
3. Show you the major focus of my work for this year. Scripts you see here are 'in progress'



© 2020 PracticalStats.com

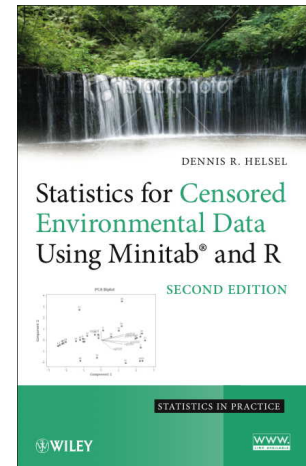
2

2

Storing Nondetects in Databases

Two methods used in statistics software

1. Indicator column
2. Interval Endpoints -- used for interval-censored data



© 2020 PracticalStats.com

3

3

Interval Endpoints Format

1st column is the lower limit, 2nd column is the upper limit

		<u>Start</u>	<u>End</u>
<1	----	0	1
<5	----	0	5
10	----	10	10
>30	----	30	NA (missing value for software)

- Detects have same value in both columns.
- Used by both parametric and nonparametric methods in R with the type = "interval2" option.



© 2020 PracticalStats.com

4

4

A Common Use for the Interval Endpoints Format

Can easily differentiate between a DL and a QL. Easy to explain to the general public.

For example, for a method detection limit $DL=1$ and a quantitation limit $QL=3$:

		<u>Start</u>	<u>End</u>
<1	---->	0	1
2J	---->	1	3



© 2020 PracticalStats.com

5

5

Methods for Analysis of Interval-Censored Data

1. Parametric methods by maximum likelihood estimation (MLE)
Must assume a distribution.
2. Nonparametric methods: Turnbull estimator, score tests
Distribution-free.

For more information go to our Newsletter Archive
<http://practicalstats.com/news/archive.html> to get:

Mar 2016	Seven perilous errors in environmental statistics
Nov 2015	Statistics for doubly-censored data
Sept 2015	Tests for Equal Variance
July 2015	Parametric vs Nonparametric vs Permutation Tests
May 2015	The power of permutation tests



© 2020 PracticalStats.com

6

6

Data: Ontario Pollen Monitoring Network

- Pesticide concentrations are measured in pollen at beehives located across the province.
- Neonicotinoids (neonics) 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 neonics.
- 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



© 2020 PracticalStats.com

7

7

Data: Ontario Pollen Monitoring Network

```
> 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

I changed values of
40 and 80 in the
2.Post-Plant group
to be >30.



© 2020 PracticalStats.com

8

8

Testing with Greater Thans and Less Thans

I changed values of 40 and 80 in the 2.Post-Plant group to be >30.

<0.05
<0.05
25
<0.05
<0.05
> 30
0.06

SamplingEvent	ThiaAbvBelow	Thia.lo	Thia.high
3. Corn Tassle	Below	NA	0.05
4. Goldenrod	Below	NA	0.05
2. Post-Plant	Above	25.00	25.00
1. Pre-Plant	Below	NA	0.05
4. Goldenrod	Below	NA	0.05
2. Post-Plant	Above	30.00	NA
1. Pre-Plant	Above	0.06	0.06
4. Goldenrod	Below	NA	0.05
2. Post-Plant	Above	30.00	NA
3. Corn Tassle	Below	NA	0.05



© 2020 PracticalStats.com

9

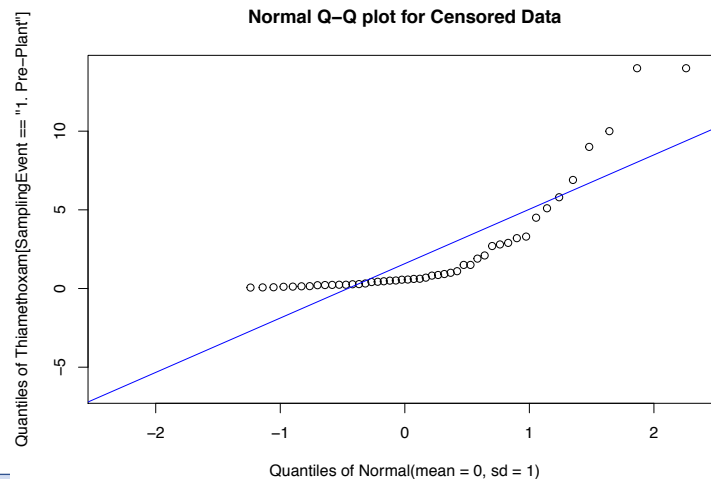
9

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.



© 2020 PracticalStats.com



10

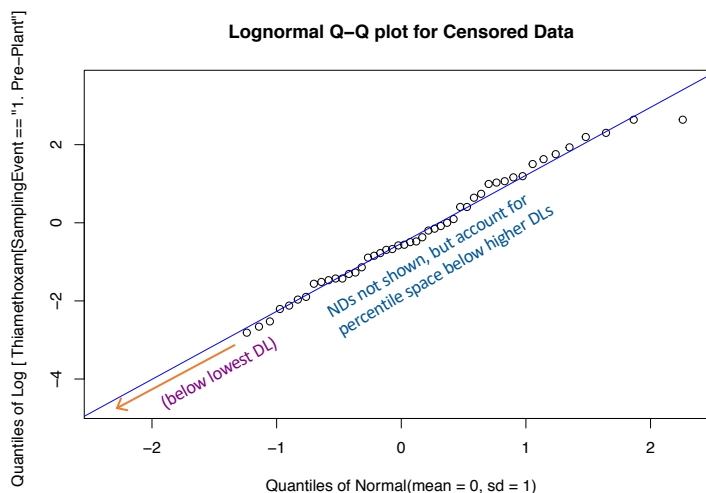
10

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



© 2020 PracticalStats.com



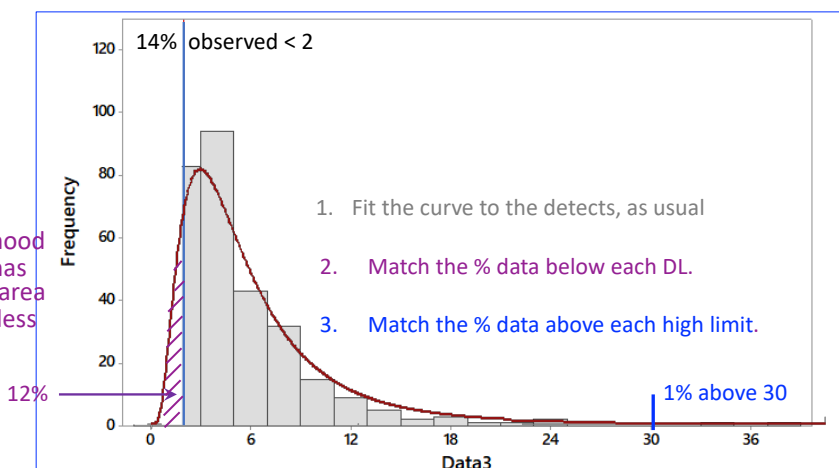
11

11

1. Parametric MLE: Fit distribution to censored data

Minimize the log-likelihood. For censored data it has two parts, one for detects and one for nondetects.
We don't have values for the lowest 14% of the data, only knowing that they are <2.

Maximum Likelihood (MLE) best fit has 12% of the total area under its curve less than 2.



© 2020 PracticalStats.com

12

12

1b. Testing with Greater Thans and Less Thans using MLE

```
> thia.log <- survreg(Surv(Thia.lo, Thia.high, type="interval2")~SamplingEvent, dist = "lognormal")
> summary(thia.log)
```

	Value	Std. Error	z	p
(Intercept)	-0.5495	0.2383	-2.31	0.021
SamplingEvent[T.2. Post-Plant]	-0.2234	0.3342	-0.67	0.504
SamplingEvent[T.3. Corn Tassle]	-3.3699	0.3817	-8.83	< 2e-16
SamplingEvent[T.4. Goldenrod]	-4.4538	0.4926	-9.04	< 2e-16
Log(scale)	0.5331	0.0696	7.66	1.9e-14

"ANOVA" on censored data
assuming a lognormal distribution
of residuals.

Scale= 1.7

Log Normal distribution

Loglik(model)= -187.2 Loglik(intercept only)= -260.2

Chisq= 146.11 on 3 degrees of freedom, p= 1.8e-31

Number of Newton-Raphson Iterations: 5

n= 284



© 2020 PracticalStats.com

13

13

13

1c. Regression for Data with Greater Thans and Less Thans

```
> surv.thia <- Surv(Thia.lo, Thia.high, type="interval2")
> x <- survreg(surv.thia~SamplingTime, dist = "lognormal")
> summary(x)
```

	Value	Std. Error	z	p
(Intercept)	1.6124	0.3463	4.66	0.0000032
SamplingTime	-1.6701	0.1528	-10.93	< 2e-16
Log(scale)	0.5949	0.0703	8.47	< 2e-16

$\ln(\text{Thia conc}) = 1.61 - 1.67 * \text{SamplingTime}$

Scale= 1.81

regression slope

Log Normal distribution

Loglik(model)= -198.9 Loglik(intercept only)= -260.2

Chisq= 122.78 on 1 degrees of freedom, p= 1.6e-28



© 2020 PracticalStats.com

14

14

1. MLE Summary

- Must assume a distribution
- No substituted values are used
- Nondetects affect the computations of mean, standard deviation, and percentiles through their observed percentage of values below and above each DL (the percentile probability of each DL)
- Gamma, Weibull and lognormal distributions are most commonly used, given the skewness of environmental data
- Greater thans are just a mathematical reflection of the process for incorporating less thans (nondetects). Requires interval censoring.
- Methods are in the survival package of R with type = "interval2".



© 2020 PracticalStats.com

15

15

2. Nonparametric Estimation of Descriptive Statistics for Interval-Censored Data

- Similar to computing sample percentiles of the data
- Computes Kaplan-Meier percentiles -- percentiles for detected data that are influenced by the positions of the nondetects & greater-thans
- The interval-censored version: Turnbull estimator determines percentiles of the cdf
- Survfit command in R computes the Turnbull percentiles for interval-censored input data
- Scripts that will soon debut in a 2nd NADA course will allow both left and right censoring. Example: compute the mean with `rmean_int2`



© 2020 PracticalStats.com

16

16

2a. cdfs and the Kaplan-Meier Method

No NDs to start with. $n=10$

3.33

2.19

1.81

1.33

1.11

0.91

0.91

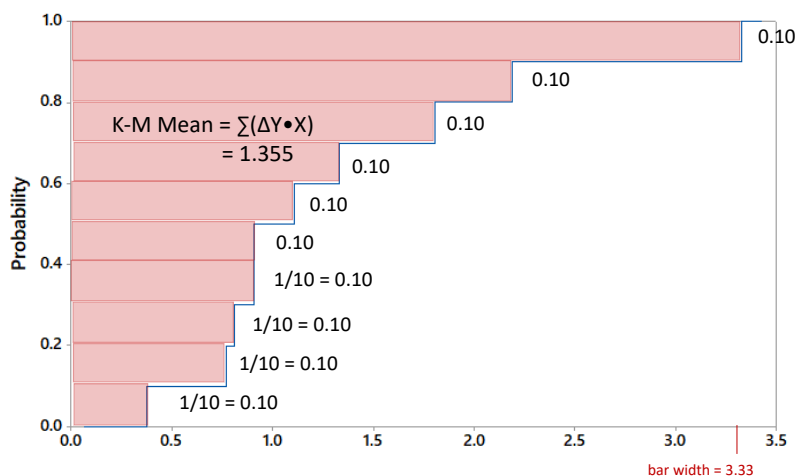
0.81

0.77

0.38

$$\begin{aligned}\text{Mean} &= \sum \text{data} / 10 \\ &= \sum (0.10 \cdot \text{data values}) \\ &= \sum (\Delta Y \cdot X)\end{aligned}$$

The mean equals the sum of the height * width for the bars, or the area in color to the left of the cdf, down to $x = 0$



© 2020 PracticalStats.com

17

17

2. cdfs and the Kaplan-Meier Method

- Kaplan-Meier takes each nondetect and reassigns its probability to the detects that occur below it.
- This assumes the observed shape of the data below the highest nondetect is the best indicator of the shape of the data in that region
- Observations below the lowest DL are treated as detected values, keeping their probabilities. The DL value is usually assigned to the < lowest DL data.
- The same process redistributes probabilities upward for > values.



© 2020 PracticalStats.com

18

18

2. cdfs and the Kaplan-Meier Method

Concentrations WITH NDs. n=10.

3.33

2.19

1.81

1.33

1.11

0.91 <1 no bar, redistributes this 0.10

0.91 + 0.025

0.81 + 0.025

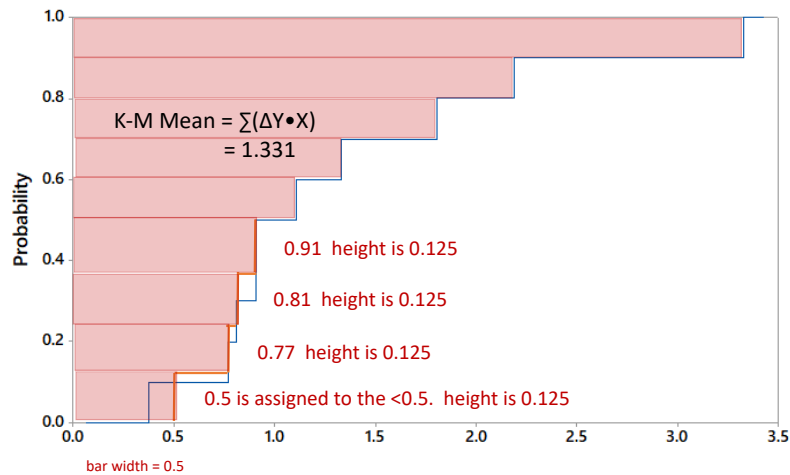
0.77 + 0.025

0.58 <0.5 + 0.025

The mean still = $\sum(\Delta Y \cdot X)$, the area to the left of the adjusted cdf down to X = 0.



© 2020 PracticalStats.com



19

2a. Nonparametric Estimation of Descriptive Statistics for Data with Greater Thans and Less Thans

Compute median (can also get other percentiles)

```
> survfit(Surv(Thia.lo, Thia.high, type="interval2")~SamplingEvent)
```

	records	n	events	median	0.95LCL	0.95UCL
SamplingEvent=1. Pre-Plant	52	53	47.90	0.57	0.32	1.10
SamplingEvent=2. Post-Plant	54	55	46.88	0.41	0.31	0.64
SamplingEvent=3. Corn Tassle	56	57	16.29	0.05	0.05	0.05
SamplingEvent=4. Goldenrod	42	43	6.14	0.05	0.05	0.05

Compute the mean

```
> rmean_int2(Thiamethoxam[SamplingEvent== "1. Pre-Plant"], ThiaCens[SamplingEvent== "1. Pre-Plant"])
Rmean = 2.002404
```



© 2020 PracticalStats.com

20

20

2b. Nonparametric Score Tests for Data with Greater Thans and Less Thans

- Nonparametric tests similar to the Wilcoxon rank-sum and Kruskal-Wallis tests
- Scores are ranks for the detected data, adjusted for censoring - similar to K-M or Turnbull percentiles
- R version is called the “Peto-Peto” test
- Similar tests called Peto-Prentice, HF1, Generalized Wilcoxon and Gehan
- Interval-censored nonparametric tests are in the **interval** package of R
- Easy to use version will be the `icen1way` command in our 2nd NADA course scripts



© 2020 PracticalStats.com

21

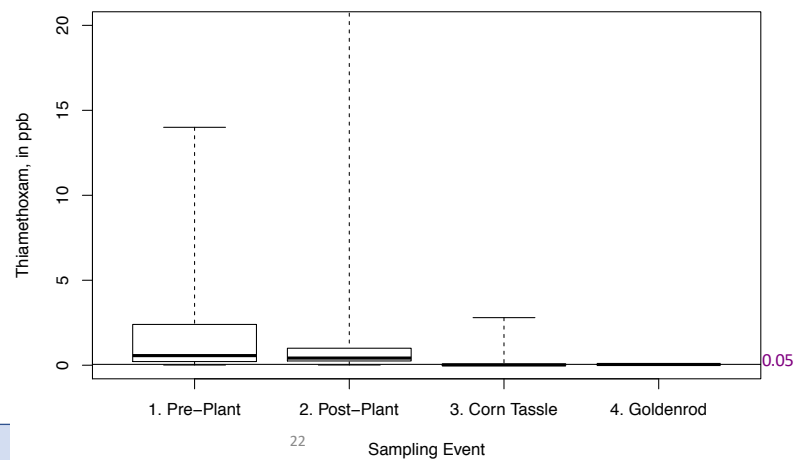
21

2b. Boxplots of Data for the Score Test

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

cenboxplot command is
from the **NADA** package

I changed values of
40 and 80 in the
2.Post-Plant group
to be >30.



© 2020 PracticalStats.com

22

22

22

2b. Nonparametric Score Tests for Data with Greater Thans and Less Thans (similar to Kruskal-Wallis)

```
> icen1way(Thia.lo, Thia.high, SamplingEvent)
Oneway Peto-Peto test of CensData: Thia.high by Factor: SamplingEvent
Chisq = 70.62 on 3 degrees of freedom p = 3.11e-15
```

Multiple Comparison Tests -- 'BH' False Discovery Rate

Group A	Group B	pval adj
2. Post-Plant	4. Goldenrod	0.00000001192
1. Pre-Plant	4. Goldenrod	0.00000016810
2. Post-Plant	3. Corn Tassle	0.00001178000
1. Pre-Plant	3. Corn Tassle	0.00001500000
3. Corn Tassle	4. Goldenrod	0.00243200000
1. Pre-Plant	2. Post-Plant	0.41630000000

For more info on this test, take our [Nondetects And Data Analysis](#) course or see our "Testing Groups for Data with Multiple DLs" webinar on practicalstats.teachable.com

Pre-A	Post-A	Corn Tassle B	Goldenrod C
-------	--------	---------------	-------------



© 2020 PracticalStats.com

23

23

2c. Nonparametric Correlation for Data with Greater Thans and Less Thans

```
> surv.thia <- Surv(Thia.lo, Thia.high, type="interval2")
> ictest(surv.thia ~ SamplingTime) ictest is a command in the interval package
I'm working on a "translation" of it
```

Asymptotic Logrank trend test(permutation form), Sun's scores
 data: surv.try by SamplingTime
 Z = 7.4931, p-value = 6.728e-14
 alternative hypothesis: survival distributions not equal

n	Score Statistic*	sort of a "correlation coeff." mainly shows direction	"earlier failures" = smaller Y values. So is a negative correlation (see 1 to 4 in boxplots)
[1,] 204	102.0895		

* positive so larger covariate values give earlier failures than expected
 X variable: SamplingTime



© 2020 PracticalStats.com

24

24

Summary: Incorporating Greater Thans and Less Thans in Data Analysis

- Interval-censored data methods use the information in the detected values, plus in the % of data below or above each limit, to compute statistics and hypothesis tests
- Input is in the “interval endpoints” format of Helsel (2012)
- MLE in the survival package allows “ANOVA” and regression style testing
- Nonparametric hypothesis tests (Peto-Peto) are computed for interval-censored data using methods in the interval package (there are other packages as well).
- R, SAS, and Stata all have routines like this. Other stat software is less likely to have them
- Talks are underway to incorporate Helsel’s interval-censored scripts into the NADA package of R. They are in, or will soon be in, the NADA online course at <https://practicalstats.teachable.com>



© 2020 PracticalStats.com

25

25

Next Month’s Video

Tuesday April 21st at anytime

- **Matched Pair Tests for Data With Nondetects**

Tuesday April 21, 2020

No need to register. Free to view in any time zone. Will be posted on April 21st to our Online Training Center

<https://practicalstats.teachable.com/>

- Sign up for our newsletter/announcement list to get the announcement and description of content emailed to you. <http://practicalstats.com/news/>
- Future webinars not directly related to our Training Courses will be located on our Videos page: <http://practicalstats.com/videos/>



© 2020 PracticalStats.com

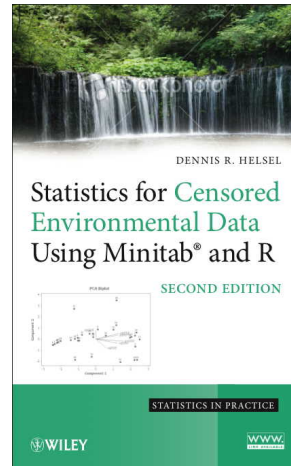
26

26

For more on stats for data with NDs:

Statistics for Censored Environmental Data (the second edition)

by Dennis R. Helsel
(2012)



© 2020 PracticalStats.com

27

27

Today's webinar will be available soon for streaming

at no charge on our Online Training Center
<https://practicalstats.teachable.com/>

Let colleagues who missed it know about it.

Current free videos on Stats with NDs at our Training Center:

<p><1</p> <p>Introduction to Nondefects And D...</p> <p>\$0 sales 276 enrolled</p>	<p><2</p> <p>Fitting Distributions to Data with ..</p> <p>\$0 sales 68 enrolled</p>	<p><3</p> <p>Testing Group Differences w/NDs</p> <p>\$0 sales 44 enrolled</p>
<p><4</p> <p>The Mystery of Nondefects</p> <p>\$0 sales 36 enrolled</p>	<p>< 5</p> <p>Correlation and Regression for D...</p> <p>\$0 sales 18 enrolled</p>	<p>< 6</p> <p>Trend Analysis for Data w/ NDs</p> <p>\$0 sales 5 enrolled</p>



© 2020 PracticalStats.com

28

28

Webinars on general statistics on our Video page

at no charge
<http://www.practicalstats.com/videos/>

In-Depth Videos (~ 1 hour)



40 Years of Water Quality Statistics
 1. pdf of Powerpoint slides
 2. Q&A from the webinar

VIDS: short (15 min or less) videos on practical topics

VIDS 1



Permutation Tests
 What's New in Statistics 1:
 Permutation Tests
 1. pdf of Powerpoint slides
 2. Q&A from the webinar

VIDS 2



R Software
 What's New in Statistics 2:
 R Software
 1. pdf of Powerpoint slides
 2. Q&A from the webinar

Current free videos on our video page:



How Many Observations Do I Need?
 How Many Observations Do I Need?
 1. pdf of Powerpoint slides
 2. Q&A from the webinar



© 2020 PracticalStats.com

29

29

Thank you for attending


- Some of the material is based on my book
Statistics For Censored Environmental Data by Dennis Helsel (2012).
- All opinions are my own and do not represent those of anyone else.

Questions?

Get in touch!

Dennis Helsel ask@practicalstats.com

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



© 2020 PracticalStats.com

30

30