Fitting Distributions to Data with Nondetects

Dennis R. Helsel

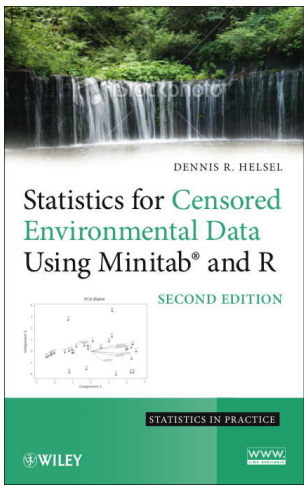
A sampling of the online course

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)



2

Why fit a distribution to data?

1. To compute summary statistics with small (8 to 20) observations
2. To compare the mean or percentiles of small datasets to a standard

- With few data it is difficult to do a good job of estimating 'extremes' such as the 90th or 95th percentile, or a UCL95.
- Instead, assume a model (distribution) of the shape of the data to project to the highest or lowest values.
- Result is only as good as the fit of the model to the actual shape of your data.
- Assuming a normal distribution is often a very poor model to use with environmental data. This is what you are doing if you use a t-statistic to compute the UCL95 or percentiles.
- If you had more data (20+ observations, more for more extreme parameters like UPLs and UTLs), bootstrapping can estimate statistics without assuming a distribution.

3

Is the Current Maximum a Good Estimate?

- Often the maximum of a small number of observations is used to estimate a UCL95 or high percentile.
- For small data sets there is a very high probability that new observations will exceed the current maximum, and a reasonable probability that with more data the UCL95 will exceed the current maximum.
- In other words, the maximum of <8 observations is frequently too low when estimating the UCL95 or a high percentile.
- For example, with 4 observations the maximum will likely be a number at or below the 60th percentile. It more likely estimates the median than is a good estimate of the UCL95 or 95th percentile.
- With 8-20 observations, use a model of the shape of the data distribution to estimate high values.

4

What distributional model should I use?

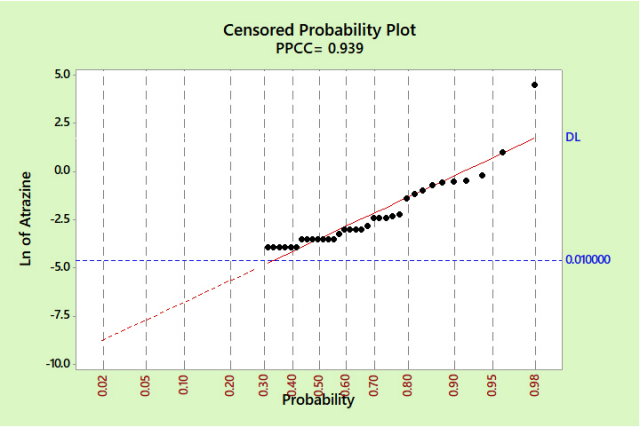
The one that best fits your data.

- If you have very few data, a larger dataset expected to be similar in characteristics could be used to determine the distribution – same parameter at a nearby longer-record location, etc.
- Most environmental data (air, water, soils, trace chemicals in biota) are skewed. This is caused by the lower bound of 0 in trace chemicals or other variables.
- Common skewed distributions include the lognormal, gamma and Weibull distributions.
- Use Q-Q plots and goodness of fit statistics to choose which distribution fits best.

5

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

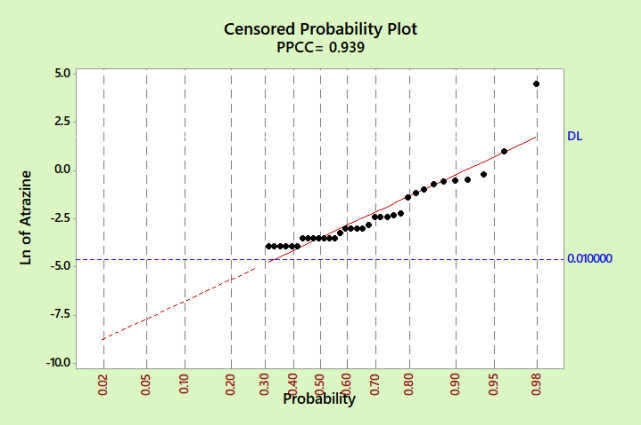
- Plots the probability ≦ data value for detected observations.
- Nondetects not plotted, but appropriate space for them left so that percentiles for detected observations are correct.
- Straight line represents a distribution such as normal, lognormal or gamma.
- PPCC measures fit. Max PPCC = 1. Choose the distribution with highest PPCC.



6

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

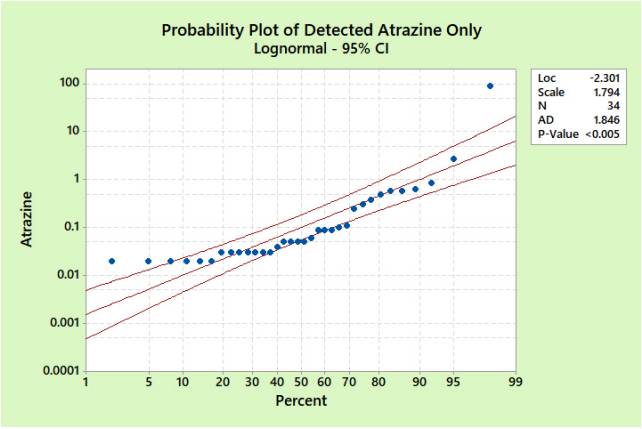
- This is plotted correctly. 30% nondetects not shown as points, but space is reserved for them at the lower end.
- Will find the routine to do this in the "survival analysis" or "censored data" sections of statistical software.



7

Q-Q plot when data with NDs are incorrectly deleted

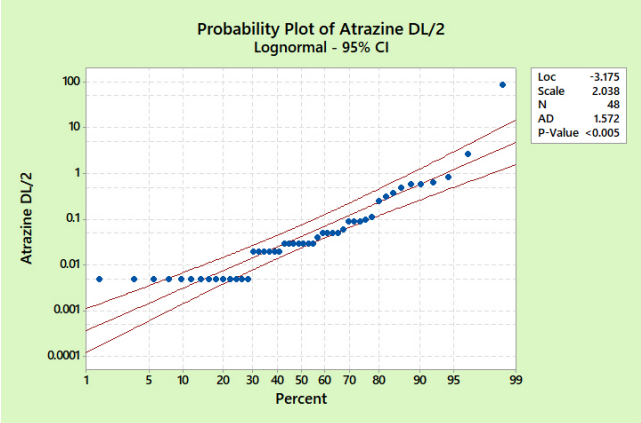
- NOT plotted correctly. Used standard Q-Q plots not designed for data with nondetects.
- Nondetects deleted, so all percentiles are too low (pushed to the left).
- Misfits the distribution compared to the true shape of data.



8

Q-Q plot with ½ DL substituted for NDs

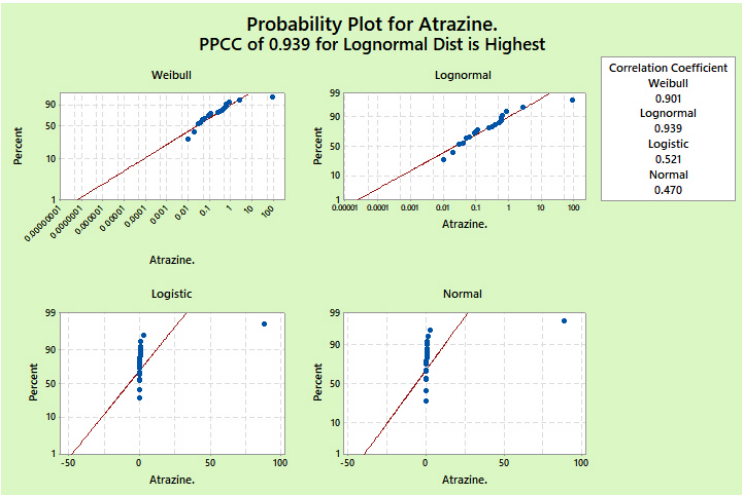
- Substituted values form straight line(s) at the low end.
- Distorts the distribution at low end compared to true shape of data.
- Leads to choosing the wrong distribution; bad estimates for percentiles at the low end.



9

Q-Q plots of possible distributions fit to data with NDs

- Distribution with data closest to a straight line, or with highest PPCC (correlation coefficient), is the best fit to the data.
- Here the lognormal distribution is the best fit compared to three other distributions.



10

Fitting Distributions with R

Arsenic concentrations in groundwater

```
> censummary(As, Ascen)
       n    n.cen  pct.cen      min      max
21.00000 14.00000 66.66667  0.50000  5.27628

limits:
  limit  n uncen   pexceed
1   0.5  1     3 0.8163265
2   2.0  1     0 0.2653061
3   3.0  1     1 0.2653061
4   4.0 11     3 0.1428571
```

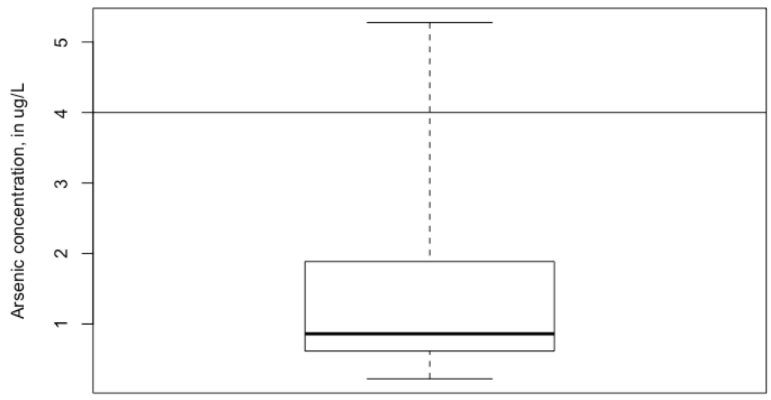
21 obs. Small enough to decide to use a distributional method.

(I am using the EnvStats, NADA and fitdistrplus packages)

11

Concentrations are skewed

```
> cenboxplot(As,Ascen, log=FALSE, ylab = "Arsenic
   concentration, in ug/L")
```



12

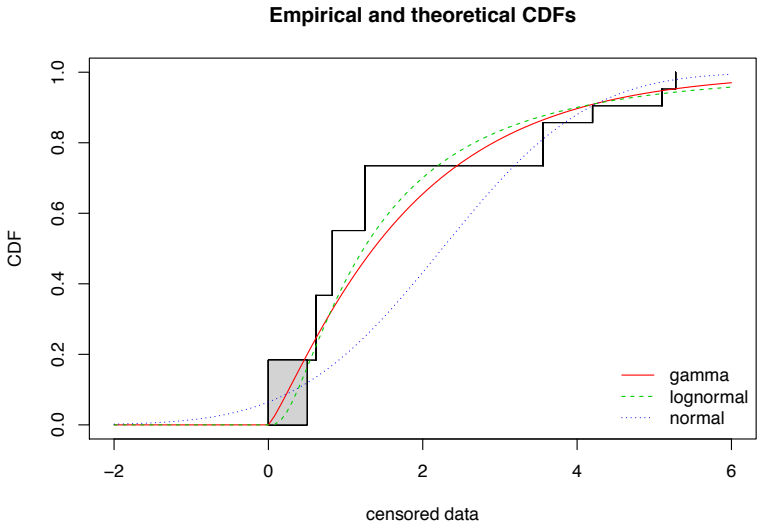
Fit of Three Distributions using MLE

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

Data shown as step function.

Gamma appears best fit, lognormal 2nd.

Note that only normal distribution estimates concentrations below 0.



13

How MLE Works

- Starts with the observed data, and your statement 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 match between observed data and the parameters
- For censored data, computed percentiles incorporate both observed detected observations and the observed proportions of data (detects and nondetects) below each detection limit.

How MLE Works

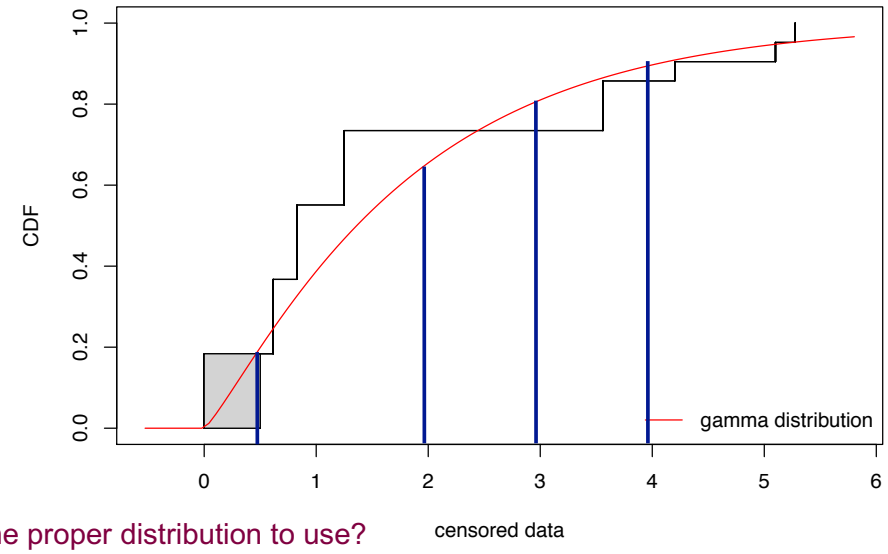
Once the distribution's shape is specified, MLE estimates the best-fitting parameters.

The parameters fit

1. the detected data (step function) and

2. proportions of data, both detects and NDs, below each DL (blue lines).

No substitution of values for nondetects is used.



So how do you choose the proper distribution to use?

Step 1. Compute PPCC or BIC for candidate distributions. Choose the best*

```
> gofTestCensored(As, Ascen, dist = "gamma", test = "ppcc")
Hypothesized Distribution:     Gamma
Estimated Parameter(s):        shape = 1.19924   scale = 1.534234
Test Statistic:                r = 0.969

> gofTestCensored(As, Ascen, dist = "lnorm", test = "ppcc")
Hypothesized Distribution:     Lognormal
Estimated Parameter(s):        meanlog = 0.2107019   sdlog   = 0.9159493
Test Statistic:                r = 0.966

> gofTestCensored(As, Ascen, dist = "norm", test = "ppcc")
Hypothesized Distribution:     Normal
Estimated Parameter(s):        mean = 1.646692    sd   = 1.966435
Test Statistic:                r = 0.968
```

Maximize the PPCC, minimize the BIC to obtain the best fitting distribution.

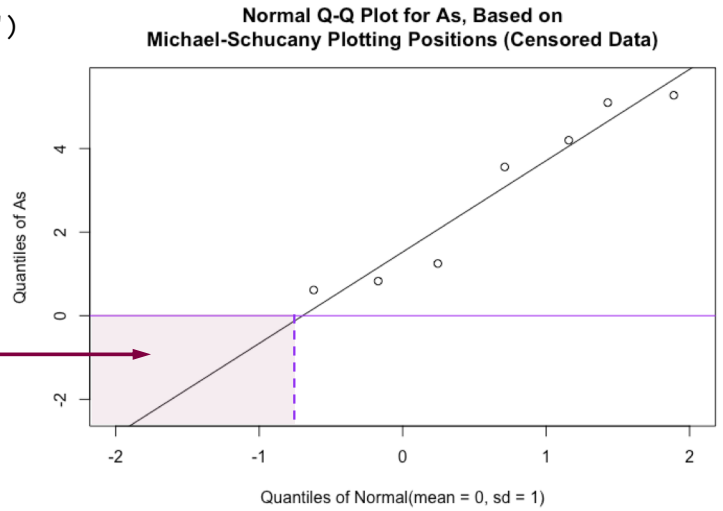
Highest PPCC of 0.969 is the gamma distribution. Almost the same at 0.968 is the normal distribution – could choose either? No! (Remember how it was off on the CDF plot?)

To illustrate an important issue with the normal distribution that you should always check, let's choose it as the one to use.

16

Step 1*. If normal distribution is chosen, don't use it if data at low end go negative

```
> qqPlotCensored(As,Ascen,dist="norm",add.line=TRUE)
> abline(h=0, col = "purple")
```



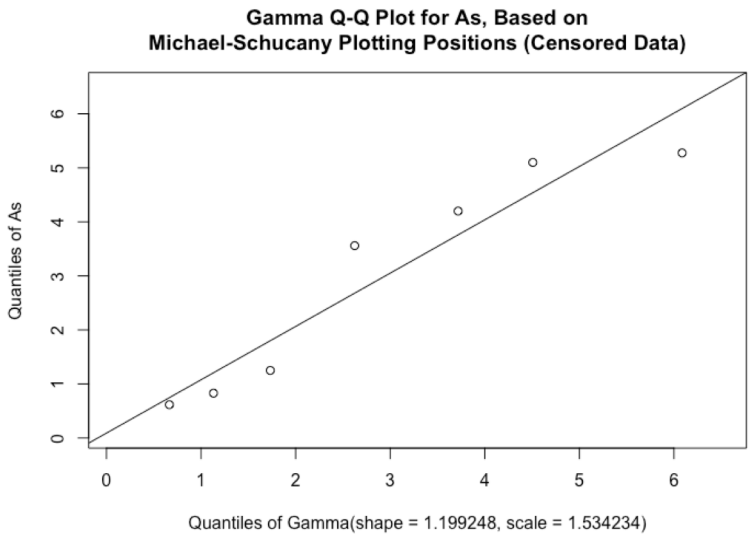
Normal distribution produces approx. 15% negative numbers.

Unacceptable! Reject it even if it has highest PPCC. Estimates of mean and UCL will be incorrect.

17

Step 1 Final. Use gamma distribution

```
> qqPlotCensored(As, Ascen, dist="gamma", add.line=TRUE,
    estimate.params=T)
```



gamma distribution had highest PPCC = 0.969

BIC for the 3 distributions using the fitdistr package (lowest is best):

gamma	43.9
lognormal	44.6
normal	50.6

18

Step 2. Compute mean, UCL95

```
> egammaAltCensored(As, Ascen, ci=TRUE, ci.type="upper", ci.method="normal.approx")

Results of Distribution Parameter Estimation Based on Type I Censored Data
--------------------------------------------
Assumed Distribution:            Gamma
Estimated Parameter(s):          mean = 1.8399269
                                 cv   = 0.9131572
Estimation Method:               MLE
Data:                            As
Censoring Variable:              Ascen
Sample Size:                     21
Percent Censored:                66.66667%
Confidence Interval for:         mean
Confidence Interval Method:      Normal Approximation
Confidence Interval Type:        upper 95%
Confidence Interval:             LCL =     -Inf
                                 UCL = 2.575537
```


Step 3. Compute upper percentiles

```
> As.gamma <- egammaCensored(As, Ascen)
> eqgamma(As.gamma, p= c(0.9, 0.95) )

Results of Distribution Parameter Estimation Based on Type I Censored Data
--------------------------------------------
Assumed Distribution:            Gamma
Censoring Side:                  left
Censoring Level(s):              0.5 2.0 3.0 4.0
Estimated Parameter(s):          shape = 1.199248
                                 scale = 1.534234
Estimation Method:               MLE
Estimated Quantile(s):           90'th %ile = 4.050705
                                 95'th %ile = 5.172334
Quantile Estimation Method:      Quantile(s) Based on
                                 MLE Estimators
Data:                            As
Censoring Variable:              Ascen
Sample Size:                     21
Percent Censored:                66.66667%
```

Nonparametric Alternative: Kaplan-Meier plus Bootstrap

```
> enparCensored(As, Ascen, ci=TRUE, ci.method="bootstrap", ci.type="upper", n.bootstraps=10000)
Results of Distribution Parameter Estimation Based on Type I Censored Data
------------------------------------------
Estimated Parameter(s):          mean     = 1.7169702
                                 sd       = 1.5928374
                                 se.mean = 0.1159666
Estimation Method:               Kaplan-Meier
Confidence Interval for:         mean
Confidence Interval Method:      Bootstrap
Number of Bootstraps:            10000
Number of Bootstrap Samples With No Censored Values:  0
Number of Times Bootstrap Repeated Because Too Few Uncensored Observations:  20
Confidence Interval Type:        upper 95%
Confidence Interval:             Pct.LCL = 0.000000
                                 Pct.UCL = 2.547658
                                 BCa.LCL = 0.000000
                                 BCa.UCL = 2.511977
```

Since there were 21 obs.
I thought I'd show you a bootstrap result.

Very similar to gamma dist. results.

21

Similar commands for other distributions

Other distributions: lognormal, Weibull, . . . more.

Other computation methods: robust ROS, robust M estimation, bias-corrected MLE, . . . more.

For more info, see:

Millard, S.P., 2013. EnvStats: An R Package for Environmental Statistics (2nd Edition). Springer, New York.

or the EnvStats free pdf package guide on the CRAN site.

22

Conclusions

1. Do not simply assume a normal distribution for small datasets and compute t-interval UCLs or percentiles
2. Use survival analysis versions of Q-Q plots to correctly incorporate nondetects, including multiple DLs
3. Use the most likely distribution (usually a skewed distribution) to compute statistics and compare them to standards
4. Use the PPCC or BIC (or AD) to find the best fitting distribution. Max PPCC or min BIC, AD correspond to best fit.

23

Much More In Our Online Training Courses

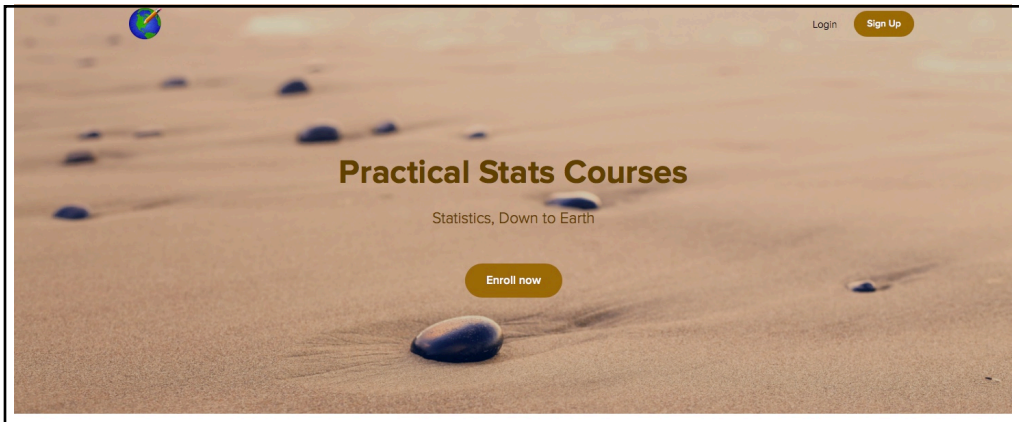
Coming Soon – Nondetects And Data Analysis

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

Our Training Site: see

https://practicalstats.teachable.com

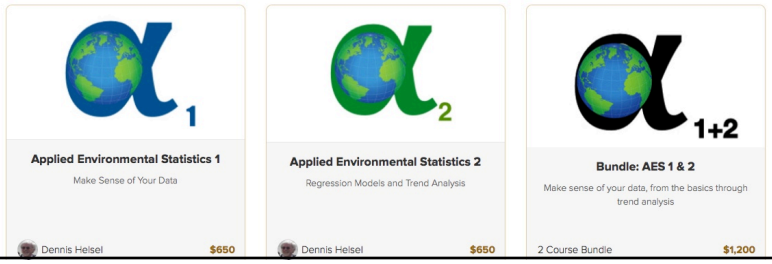
24



© PracticalStats.com

Our Training Site

Featured Courses



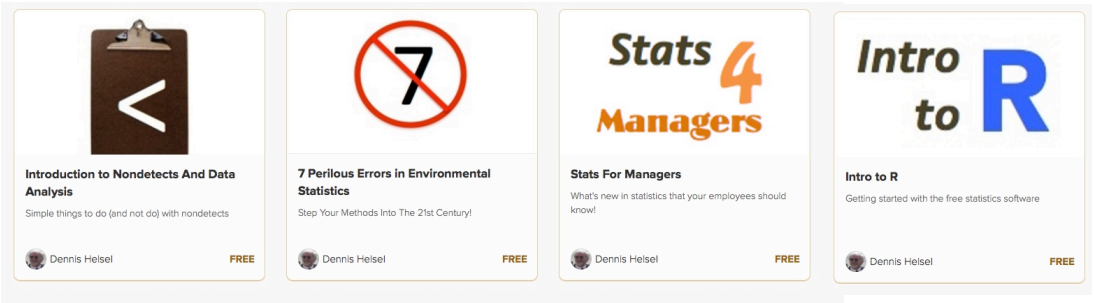
Applied Environmental Statistics

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 soon

26

Our Next Webinar

Testing Groups of Data With Multiple DLs

on Mar 19, 2019 11:00 AM MDT

You'll receive an email from ask@practicalstats.com in 2 days with the link to register:

https://attendee.gotowebinar.com/register/1074923646089662989

Testing differences between groups of data is familiar to scientists -- ANOVA for means and Kruskal-Wallis for cdfs (percentiles). If differences occur, multiple comparison methods determine which groups differ from others. Tests with the same objectives exist that are designed for censored data, data subject to (one or multiple) detection limits. In this webinar you will learn how these tests work and how you can compute them.

27

Questions?

Thank you for attending!

Dennis Helsel
PracticalStats.com

28