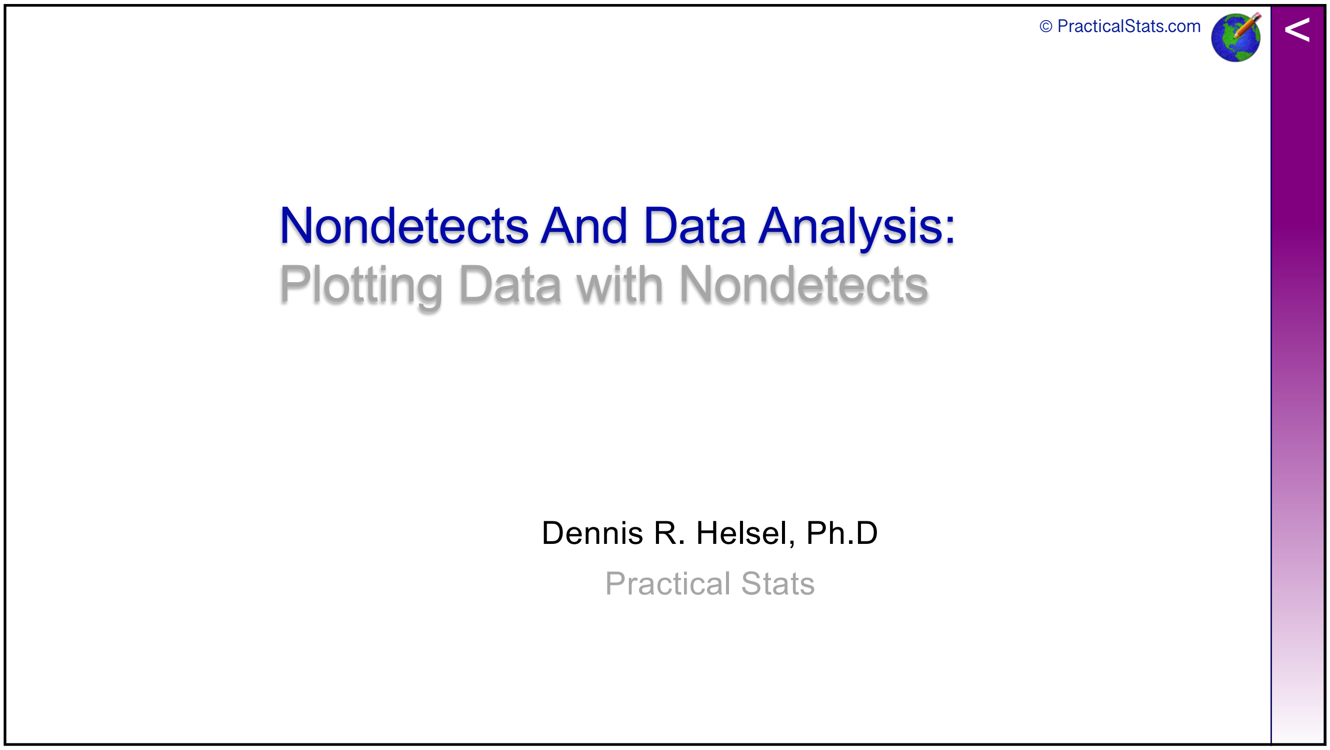
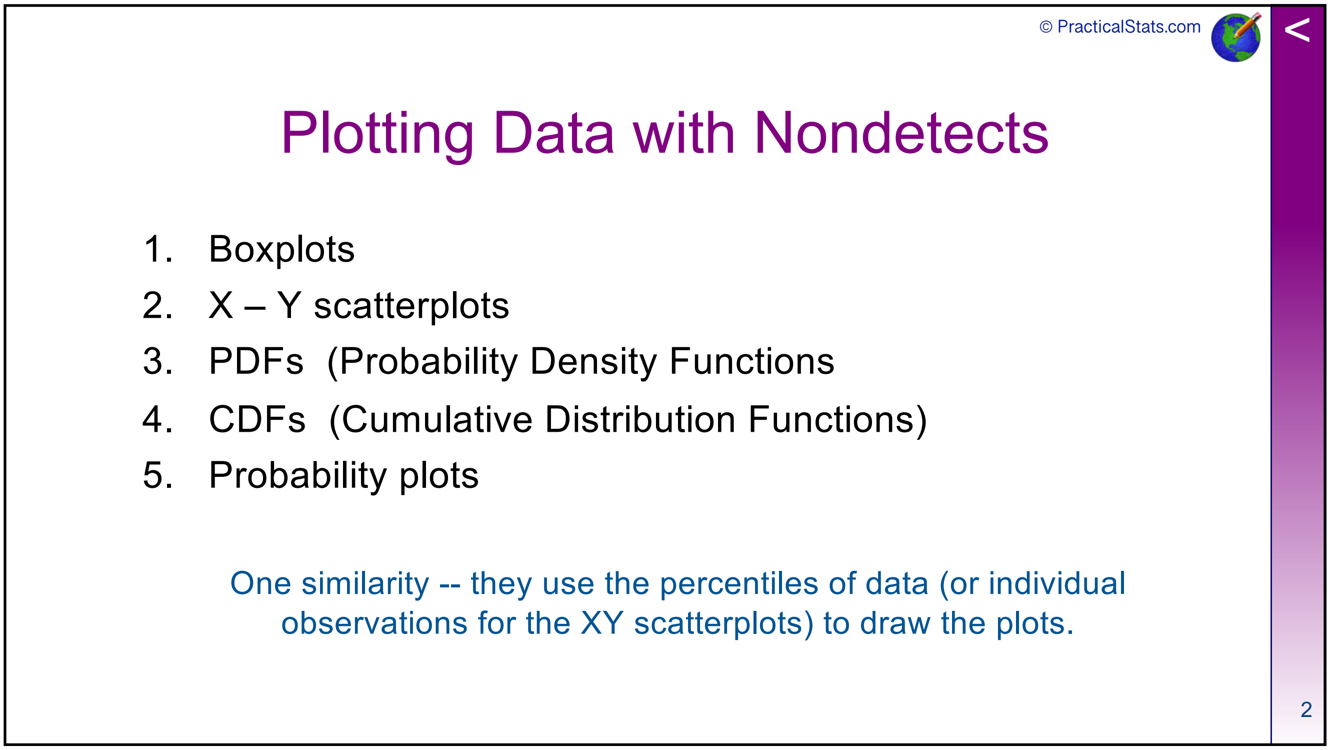


© PracticalStats.com

Nondetects And Data Analysis:
Plotting Data with Nondetects

Dennis R. Helsel, Ph.D
Practical Stats

1



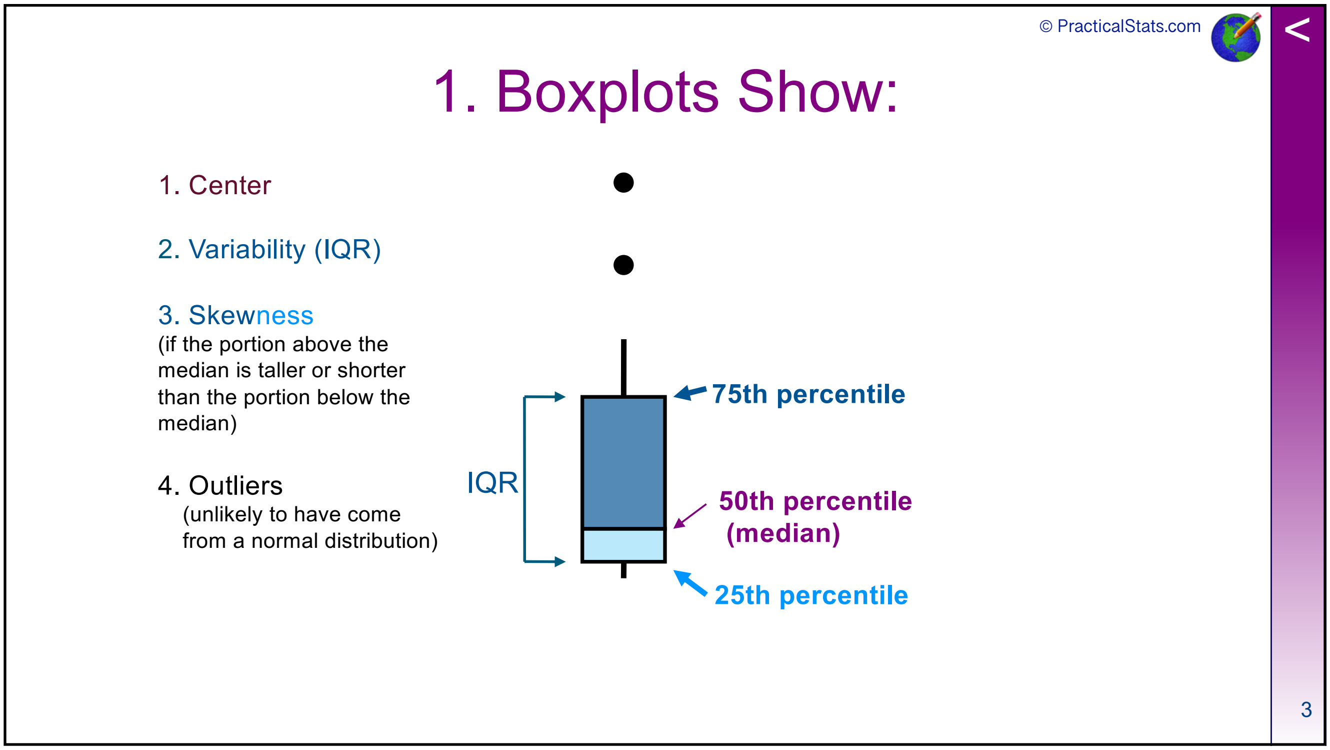
© PracticalStats.com

Plotting Data with Nondetects

1. Boxplots
2. X – Y scatterplots
3. PDFs (Probability Density Functions)
4. CDFs (Cumulative Distribution Functions)
5. Probability plots

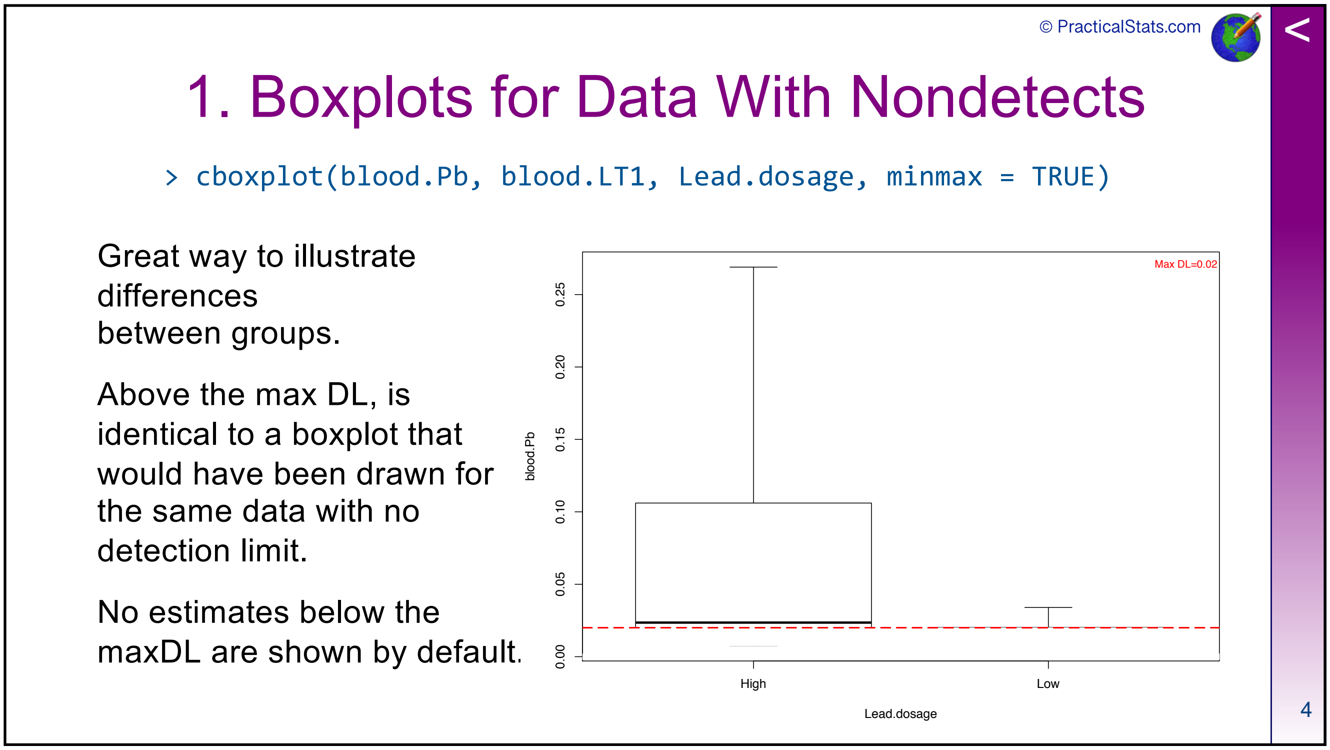
One similarity -- they use the percentiles of data (or individual observations for the XY scatterplots) to draw the plots.

2



1. Boxplots Show:

1. Center
2. Variability (IQR)
3. Skewness
(if the portion above the median is taller or shorter than the portion below the median)
4. Outliers
(unlikely to have come from a normal distribution)



1. Boxplots for Data With Nondetects

```
> cboxplot(blood.Pb, blood.LT1, Lead.dosage, minmax = TRUE)
```

Great way to illustrate differences between groups.

Above the max DL, is identical to a boxplot that would have been drawn for the same data with no detection limit.

No estimates below the maxDL are shown by default.

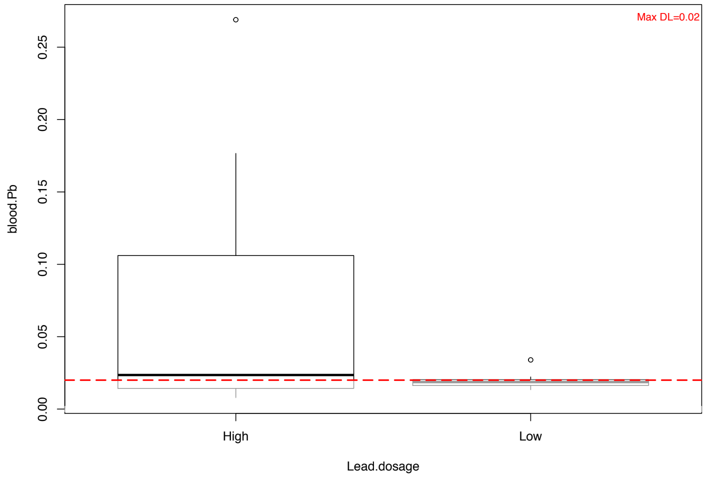
1. Boxplots for Data With Nondetects

```
> cboxplot(blood.Pb, blood.LT1, Lead.dosage, show =TRUE)
```

Portion below the max DL is estimated by ROS and shown by show = TRUE option.

Estimates are grayed out to indicate uncertainty.

Not using minmax = TRUE gives the default boxplot showing outliers.



2. X-Y Scatterplots With Nondetects

```
> cenxyplot(Pctl, PctlCen, AtraConc, AtraCen, log="y”, ylab="Atrazine
  Concentration", xlab = "Flow Percentile at Sampling")
```

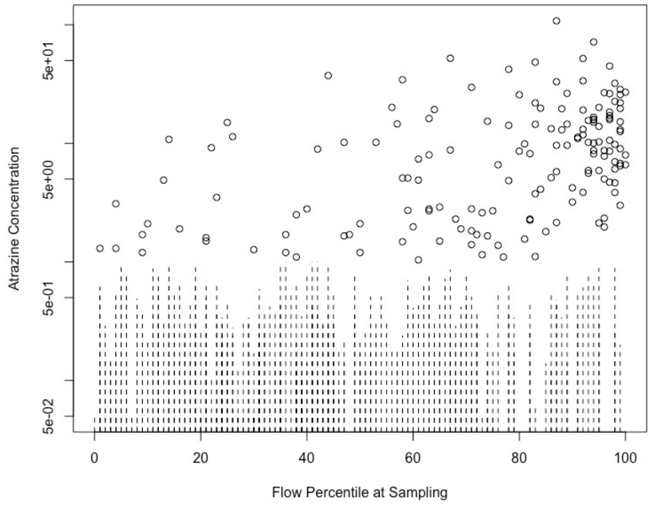
cenxyplot in the NADA package

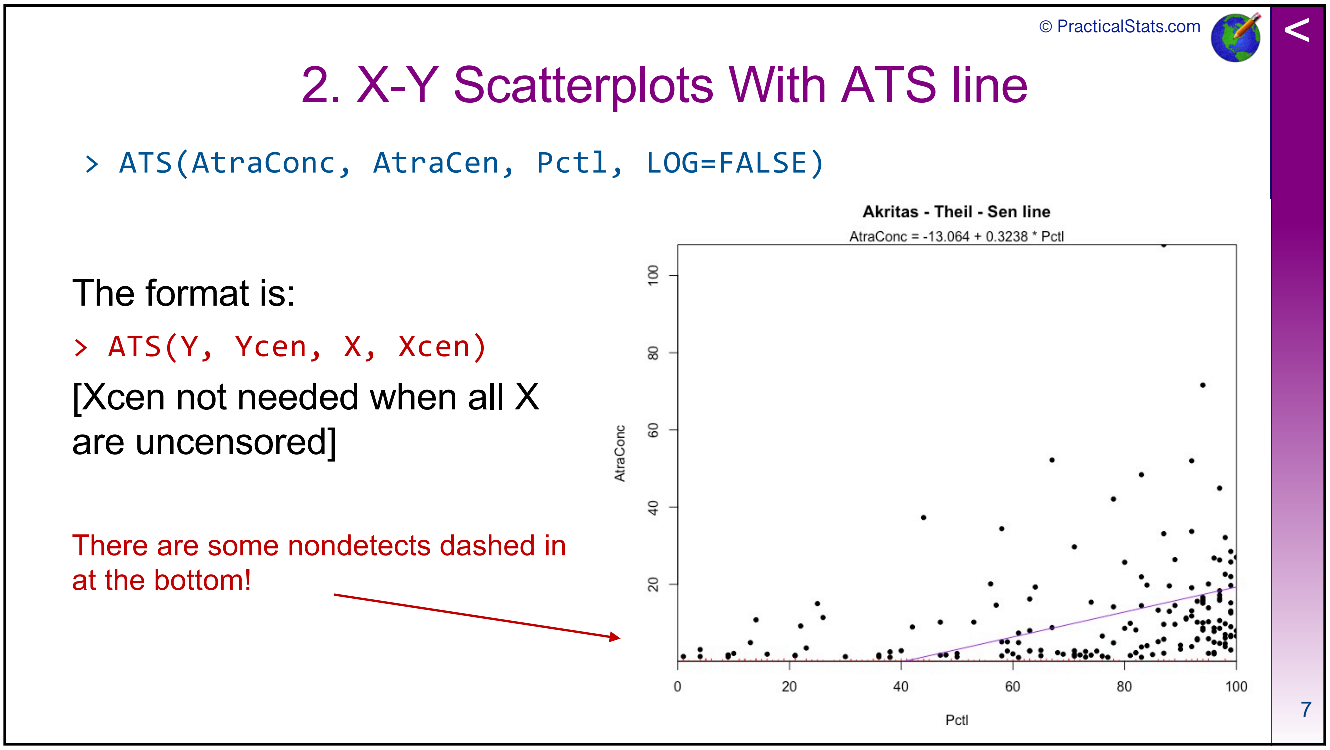
Format is cenxyplot (X, Xcen, Y, Ycen, …)

Detects plotted individually

Nondetects shown as an interval (dashed vertical lines).

Here the y axis is on a log scale





© PracticalStats.com

2. X-Y Scatterplots With ATS line

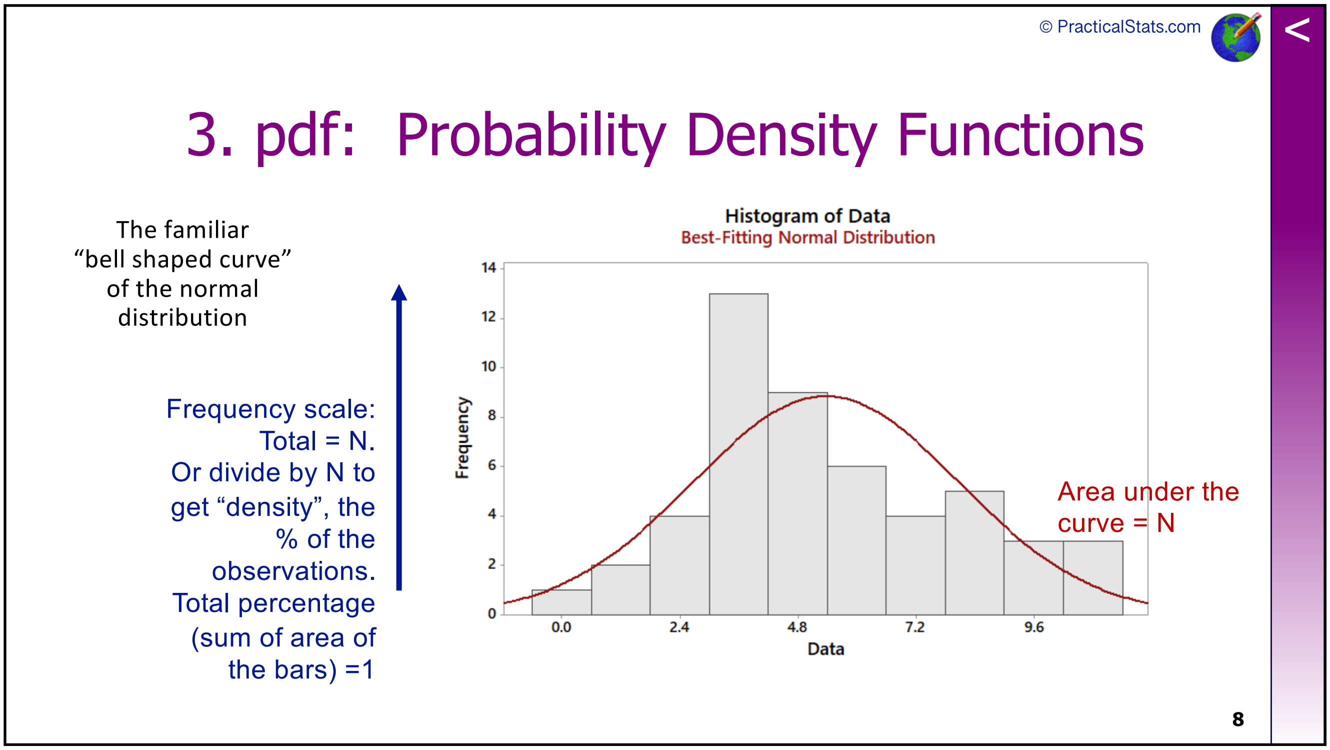
```
> ATS(AtraConc, AtraCen, Pctl, LOG=FALSE)
```

The format is:

```
> ATS(Y, Ycen, X, Xcen)
```

[Xcen not needed when all X are uncensored]

There are some nondetects dashed in at the bottom!



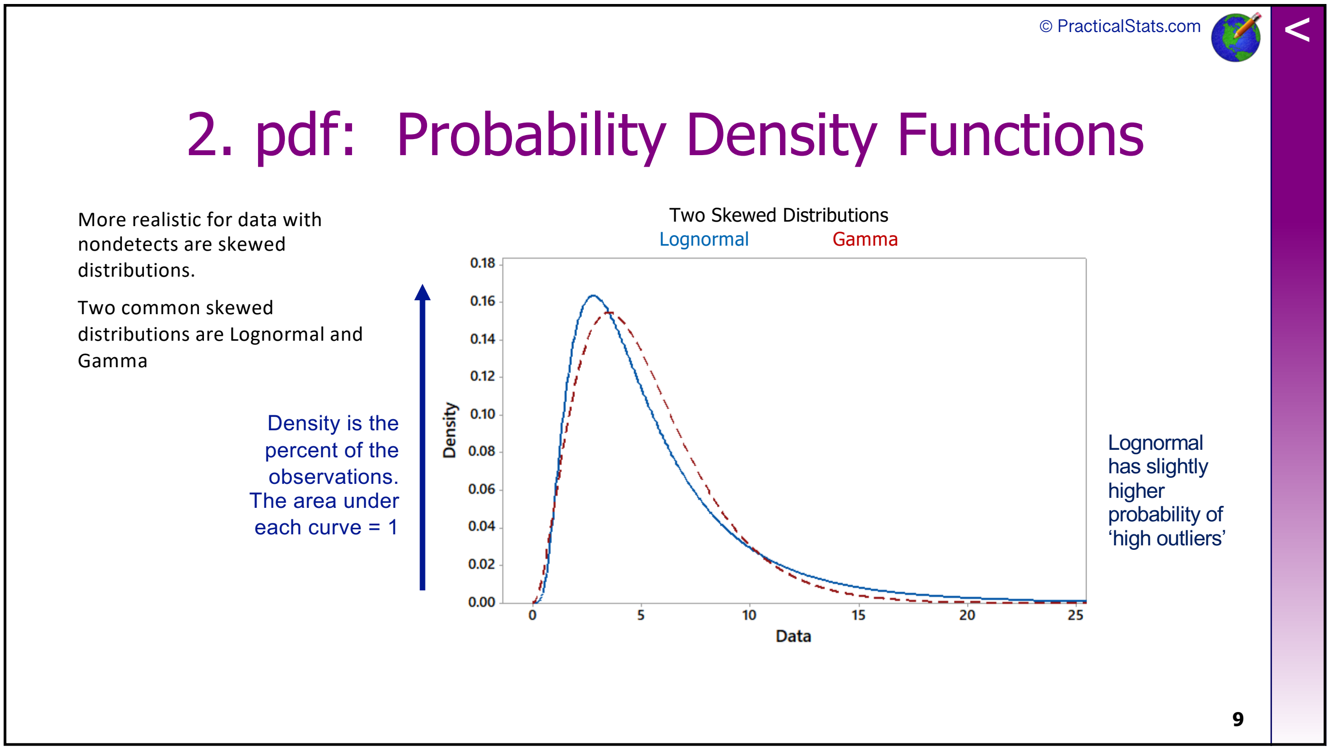
© PracticalStats.com

3. pdf: Probability Density Functions

The familiar
"bell shaped curve"
of the normal
distribution

Frequency scale: Total = N. Or divide by N to get "density", the % of the observations. Total percentage (sum of area of the bars) =1

Area under the curve = N



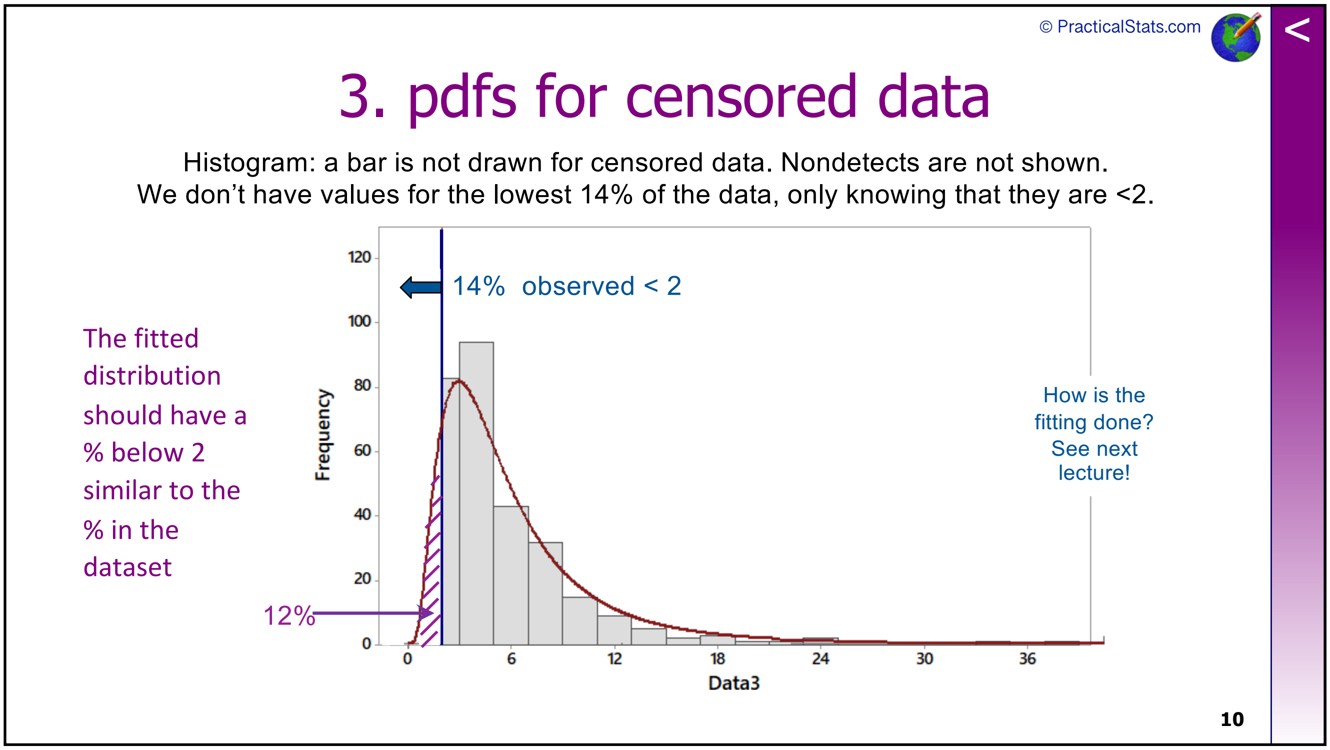
2. pdf: Probability Density Functions

More realistic for data with nondetects are skewed distributions.

Two common skewed distributions are Lognormal and Gamma

Density is the percent of the observations. The area under each curve = 1

Lognormal has slightly higher probability of 'high outliers'

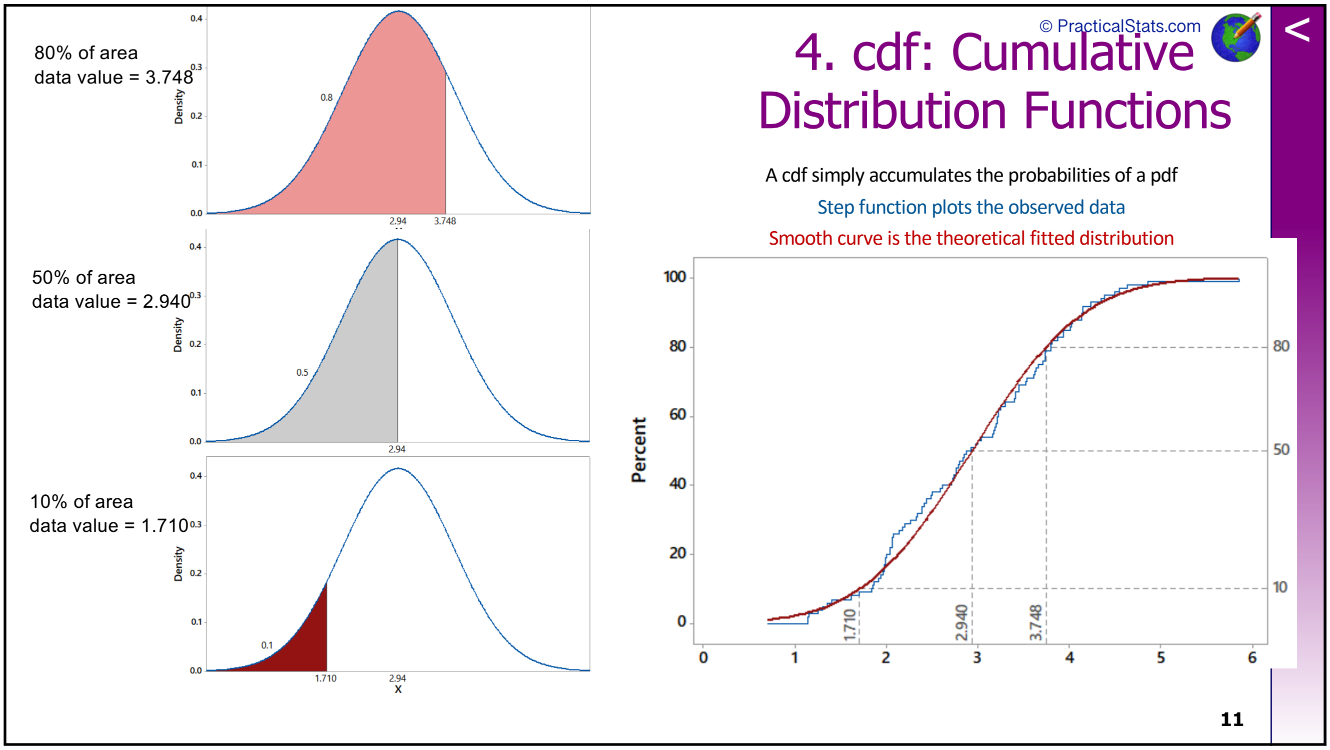


3. pdfs for censored data

Histogram: a bar is not drawn for censored data. Nondetects are not shown.
We don't have values for the lowest 14% of the data, only knowing that they are <2.

The fitted distribution should have a % below 2 similar to the % in the dataset

How is the fitting done? See next lecture!



© PracticalStats.com

4. cdf: Cumulative Distribution Functions

A cdf simply accumulates the probabilities of a pdf

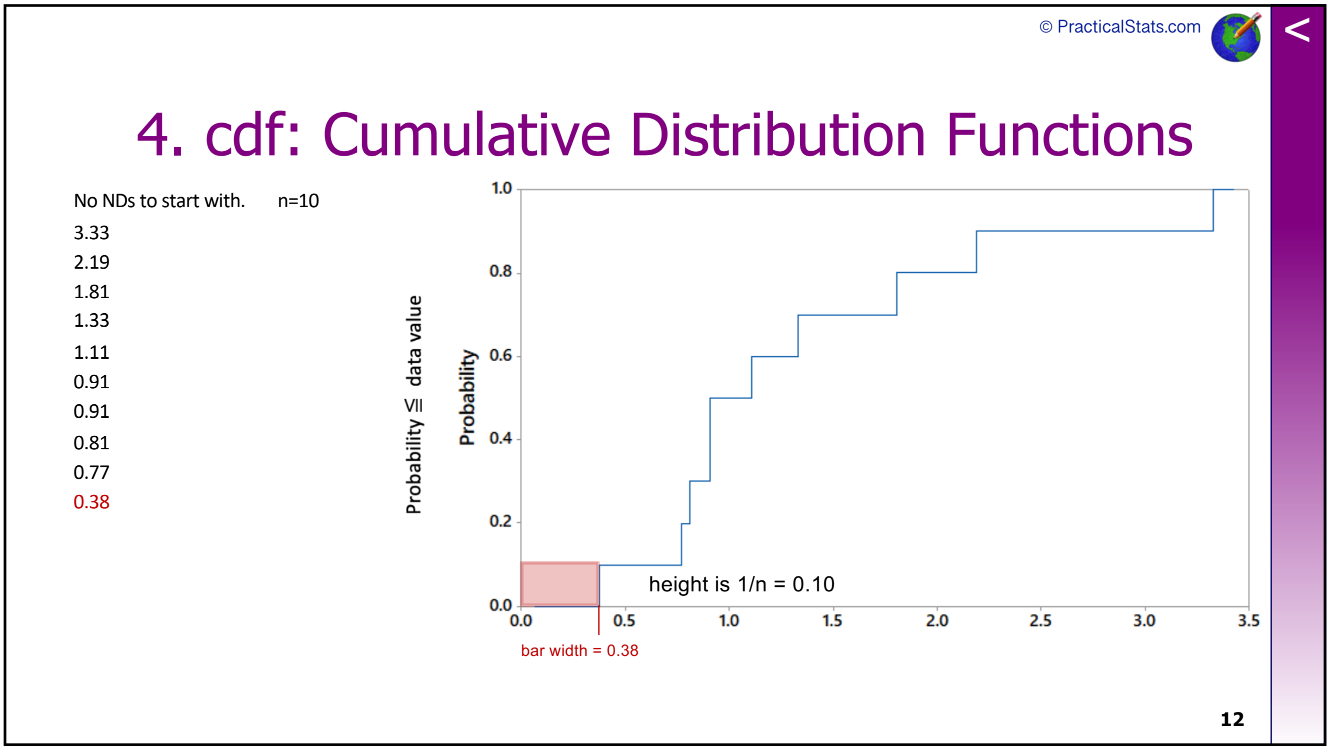
Step function plots the observed data

Smooth curve is the theoretical fitted distribution

80% of area
data value = 3.748

50% of area
data value = 2.940

10% of area
data value = 1.710



© PracticalStats.com

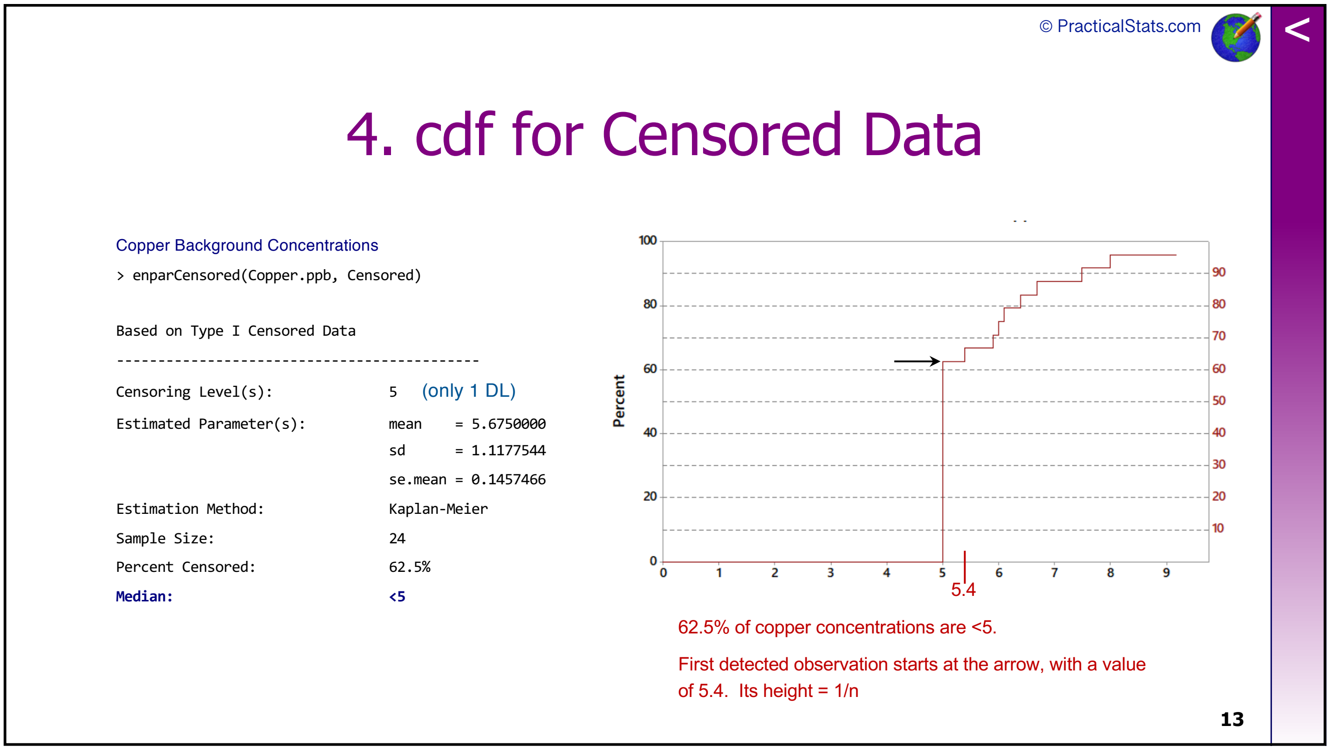
4. cdf: Cumulative Distribution Functions

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

Probability ≤ data value

height is 1/n = 0.10

bar width = 0.38



© PracticalStats.com

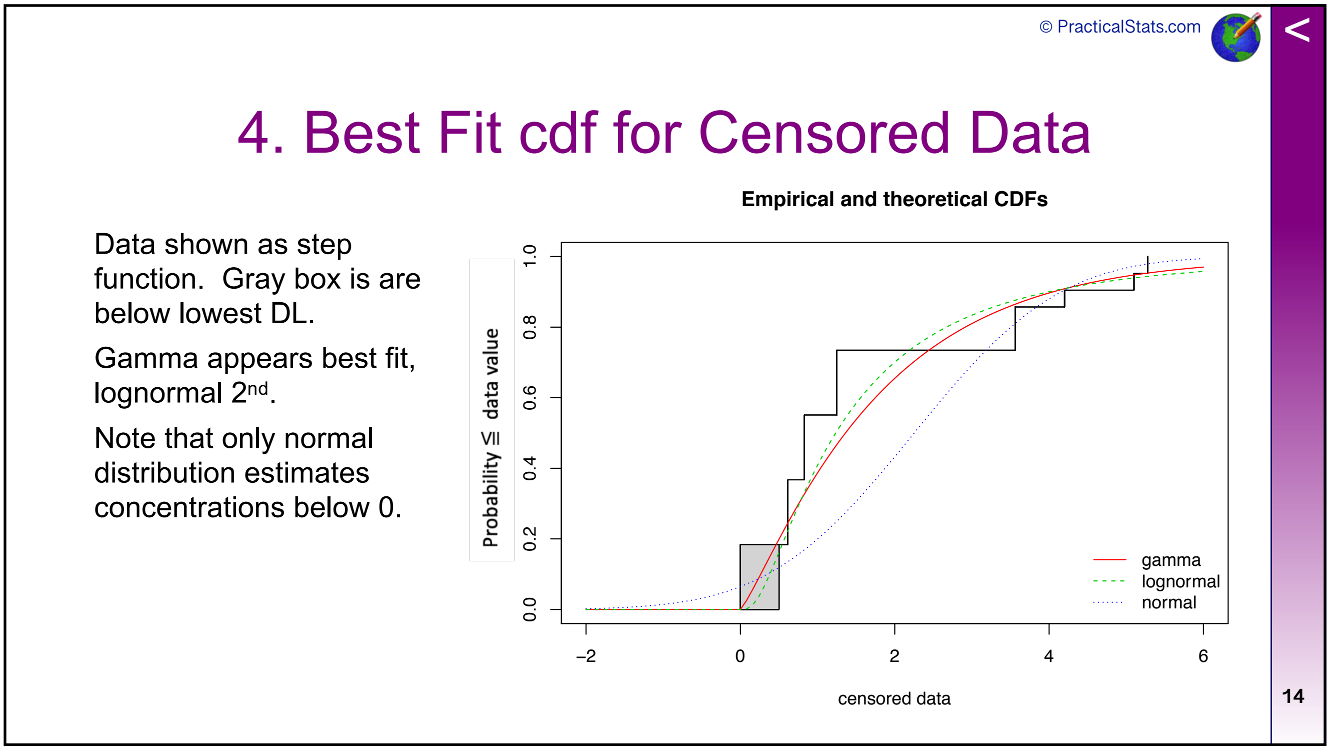
4. cdf for Censored Data

Copper Background Concentrations

```
> enparCensored(Copper.ppb, Censored)

Based on Type I Censored Data
------------------------------------------
Censoring Level(s):              5   (only 1 DL)
Estimated Parameter(s):          mean    = 5.6750000
                                 sd      = 1.1177544
                                 se.mean = 0.1457466
Estimation Method:               Kaplan-Meier
Sample Size:                     24
Percent Censored:                62.5%
Median:                          <5
```

62.5% of copper concentrations are <5.

First detected observation starts at the arrow, with a value of 5.4. Its height = 1/n



© PracticalStats.com

4. Best Fit cdf for Censored Data

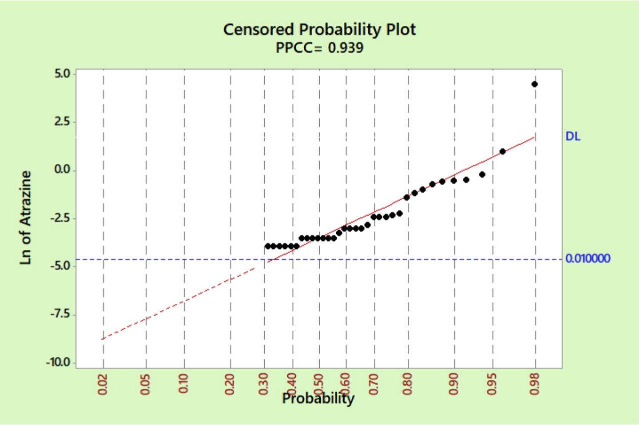
Data shown as step function. Gray box is are below lowest DL.

Gamma appears best fit, lognormal 2nd.

Note that only normal distribution estimates concentrations below 0.

5. Probability (or Q-Q) plots for data with NDs

- Plots the probability ≦ data value for detected observations on X axis
- 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.



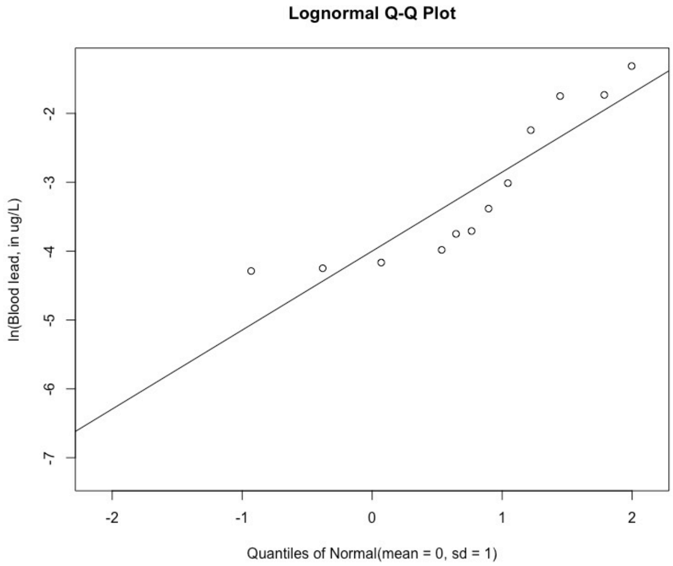
5. Probability (Q-Q) Plots for Data With Nondetects

Instead of the nonlinear Probability ≤ data value, software often plots a linear scale. One is Normal Quantiles.

Normal Quantiles are quantiles of a normal distribution with mean = 0 and a standard deviation of 1.

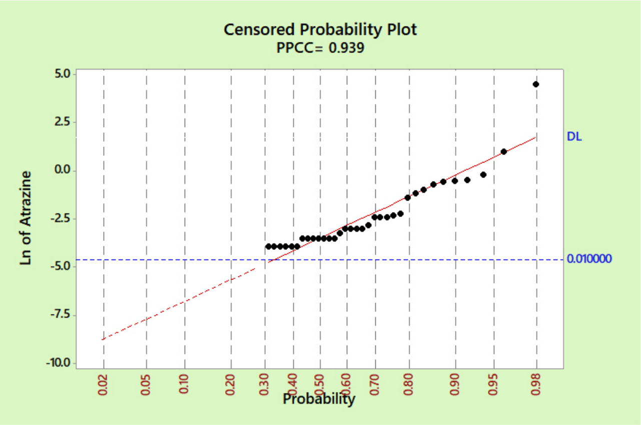
Nondetects influence quantiles of detects. Here the lowest detected value is just above -1, which is at the 16th percentile of the dataset. There are 16% of the data < lowest detection limit.

Mulitple DLs can be incorporated.



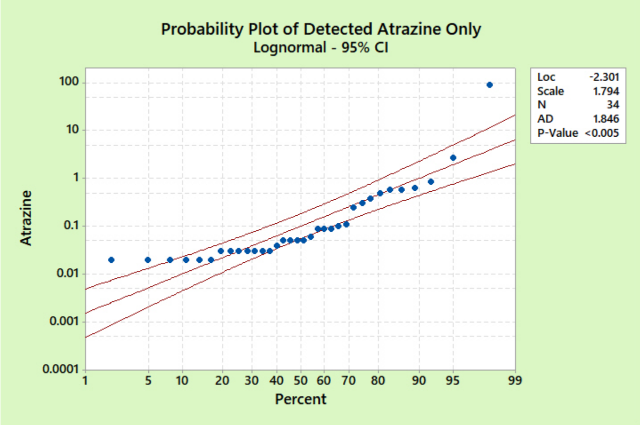
Q-Q plot to see fit of distribution to data with NDs

- 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.
- The qqPlotCensored command in the EnvStats package is one of those.



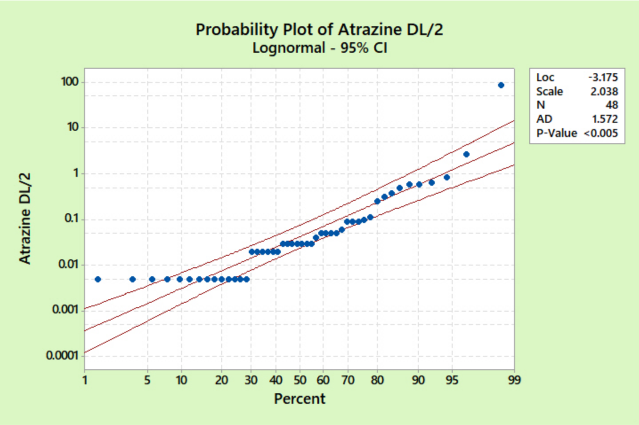
Q-Q plot when data with NDs are incorrectly deleted

- 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.



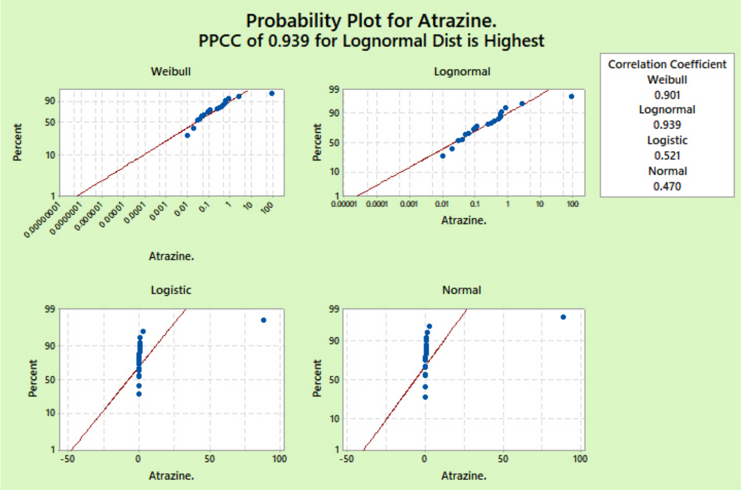
Q-Q plot with ½ DL substituted for NDs

- 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.



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.



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)

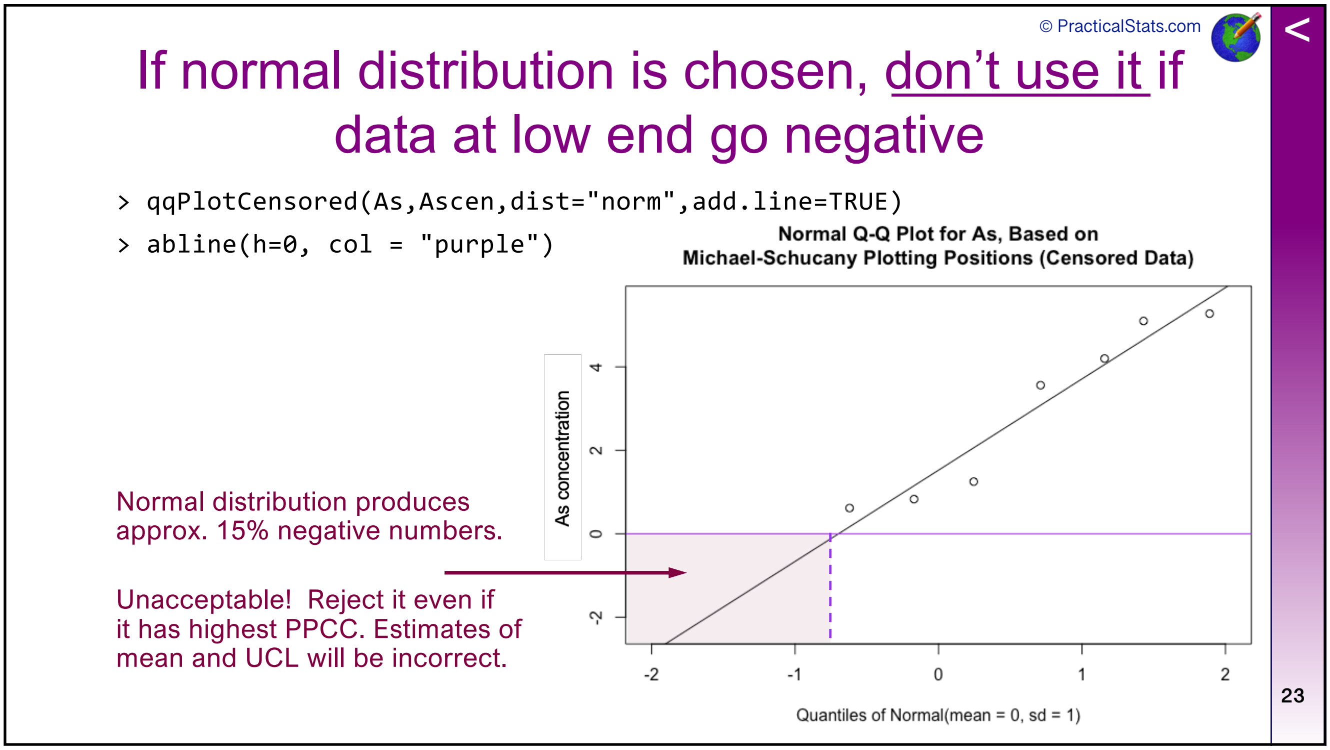
Compute PPCC or BIC for candidate distributions. Choose the best*

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

> 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
```

Best: 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.



If normal distribution is chosen, don’t use it if data at low end go negative

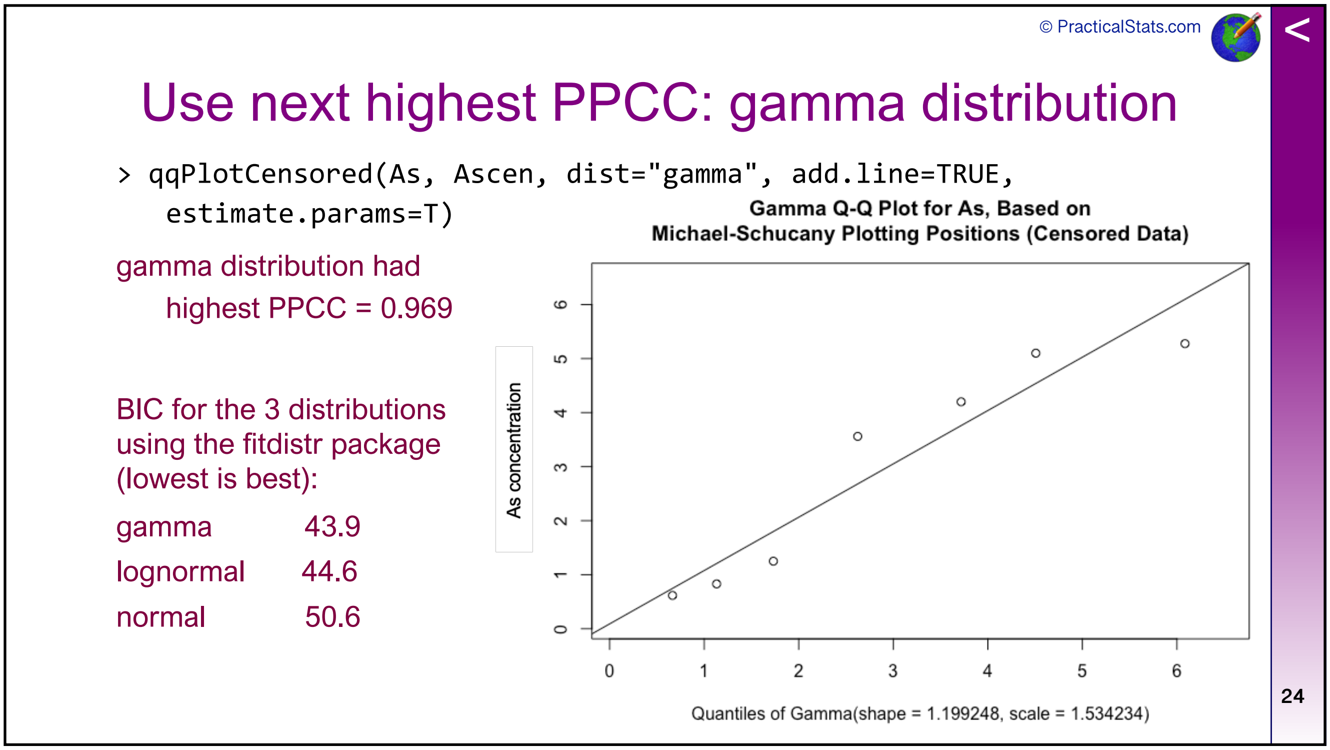
```
> qqPlotCensored(As,Ascen,dist="norm",add.line=TRUE)
> abline(h=0, col = "purple")
```

Normal Q-Q Plot for As, Based on Michael-Schucany Plotting Positions (Censored Data)

Normal distribution produces approx. 15% negative numbers.

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

23



Use next highest PPCC: gamma distribution

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

Gamma Q-Q Plot for As, Based on Michael-Schucany Plotting Positions (Censored Data)

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

24

© PracticalStats.com

<

Summary: Plotting Data with Nondetects

- Best methods are those using probabilities / quantiles
- This is because nondetect information is contained in probabilities of being <DL
- Boxplots, probability plots, PDFs and CDFs all portray this information
- Scatterplots can be used by portraying nondetects as dashed lines or interval bars rather than as points