Environmental Statistics

Chapter 2: Modelling variability and handling uncertainty

Session 2020/2021

University of Glasgow

University of Glasgow

What we will cover

- Distributions (revision!)
- Uncertainties
- **Dealing with censored observations**
- Outlier detection
- Missing data

Dealing with Censored Data

University of Glasgow

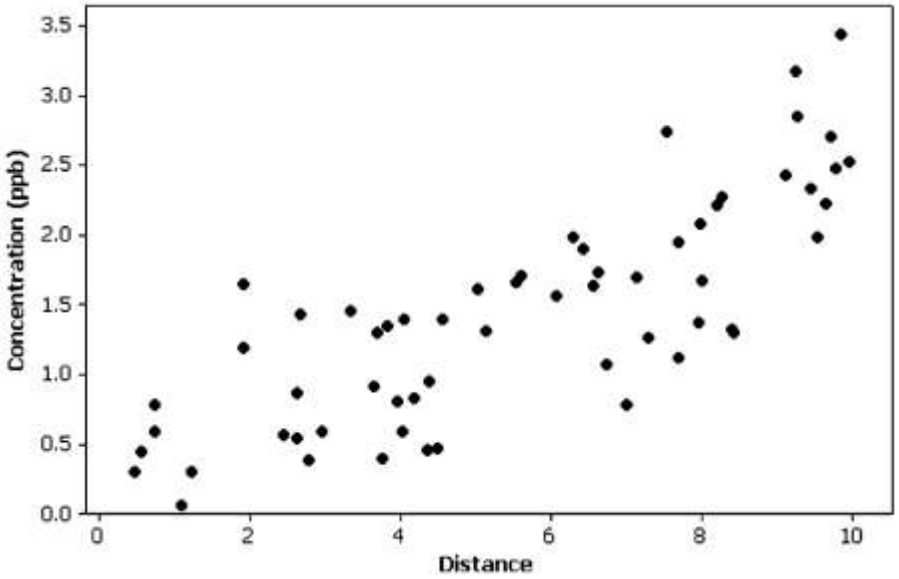
Dealing with censored data

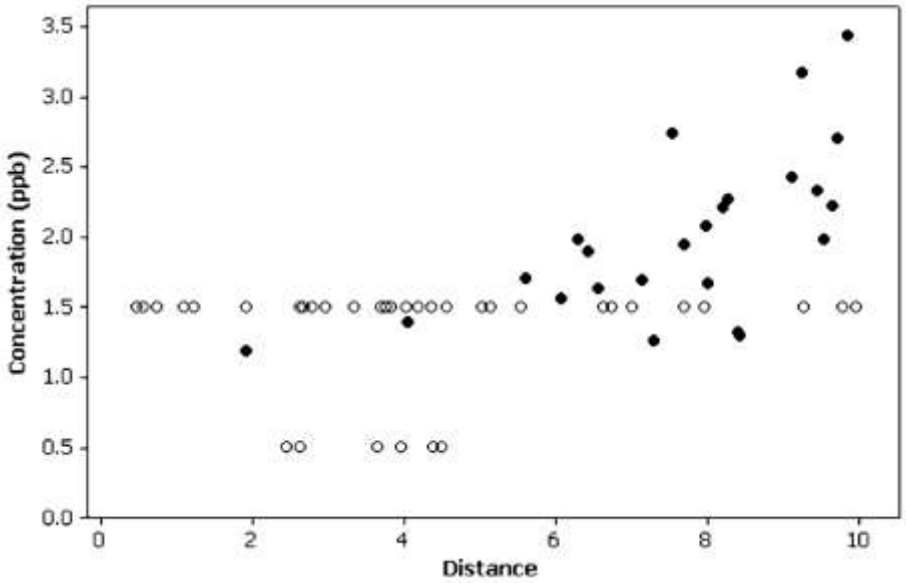
- What are censored data?
- Data which we don't know about – e.g. censored observations in survival analysis (often right censored)
- With environmental data: observations that fall **below analytical detection limits** –known as Limits of Detection of (LOD)
- observations are recorded as < … or > (most typically less than – left censored)
- So what effects do such values have in our analysis and how should we handle them?

University of Glasgow

LOD - example

- The plots below show the same set of data, however one set has observations marked as being at the limit of detection.





- Comment on these plots.

University of Glasgow

What is the limit of detection?

- **Definition:**
  Lowest concentration that can be distinguished with **reasonable confidence** from a *blank* - a hypothetical sample containing zero concentration of the analyte of interest.

- The definition, estimation and inference about it is a statistical matter related to the random error in measurements.

- What is meant by ***reasonable confidence***?

University of Glasgow

What is the limit of detection?

- limit of detection: often denoted c_L or q_L

- Related to the smallest measure of response x_l that can be detected where

$$x_L = x_B + ks_B$$

x_B is the mean of blanks
s_B is the standard deviation of the blanks,
k a numerical constant (often $k = 3$ is used)

University of Glasgow

Measurements of concentrations falling below the detection limit

- At low levels, many observations may fall below c_L
- They are not devoid of meaning, how are they reported?
  - not detected
  - less than c_L
- Do we remove them?
- No:
  - may invalidate any conclusions reached
  - analysis must take them into account.
- How are they used in any subsequent analysis?

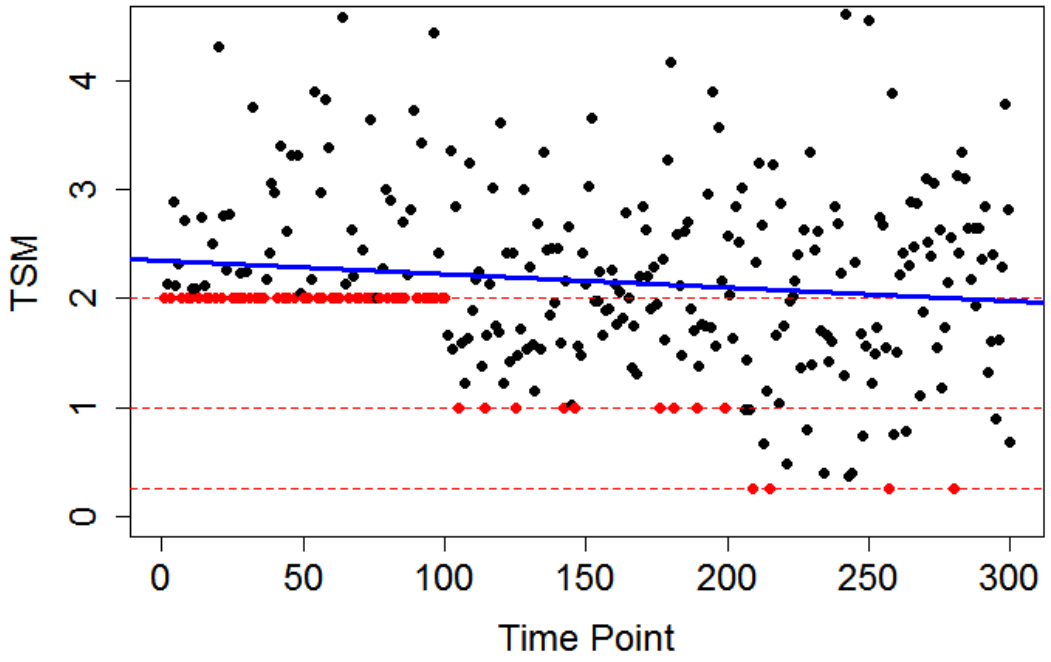
University of Glasgow

Measurements of concentrations falling below the detection limit

TSM (total suspended matter):

reduction in the limit of detection:

- falsely indicates a trend in the data
- appearance of a change in the seasonal pattern over time.



University of Glasgow

Approach 1: Simple Substitution

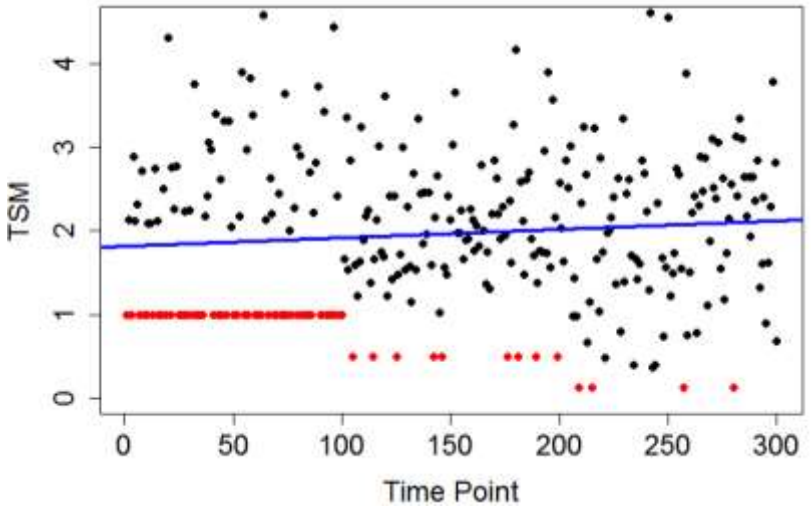
Common procedure is to take the LOD value and replace it by a fixed constant (e.g. 0.5 C_L)

Strengths;

Simple, Easy to implement

Acceptable if % of non detects is low (10-15%) - simulation has shown it works well

Better than ignoring them altogether (Eastoe, 2006)



Weaknesses;

Similar problems to before if percentage of non-detects is high (>10-15%) could result in a sample mean that is biased high

Helsel (1990) advises against this approach

University of Glasgow

Measurements of concentrations falling below the detection limit

- Statistical Approaches
  - **Maximum Likelihood Estimator**
  - **Kaplan Meier Estimator**
  - **Regression on Order Statistics**

- General idea behind each of the methods:
  - estimate summary statistics for the **distribution of the data** which take into account the censored observations.

- Using estimated distributions, values are simulated, subject to the constraint that they fall below the stated limit of detection values.

- Values generated are subsequently imputed in place of the censored observations.

University of Glasgow

Statistical Approach 1 : Maximum Likelihood

- **Parametric approach** – requires specification of a distribution which is a close fit to the data

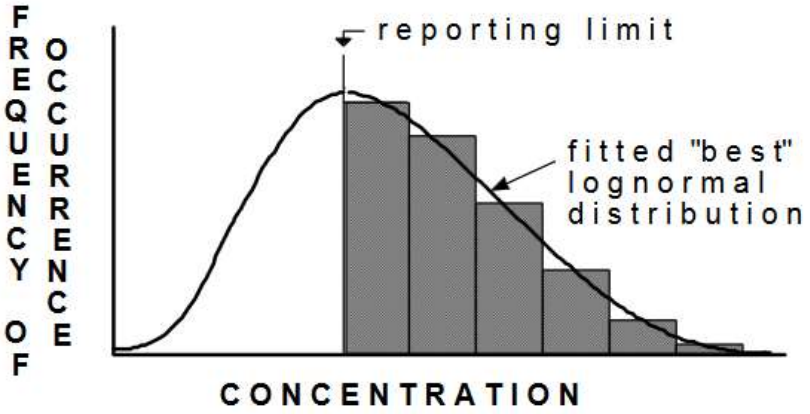
- Parameter estimates describe a distribution with the ML of producing a dataset with
  - the observed detected values and
  - the proportion of censored data which falls below each of the stated detection limits.

- Cohen's method is a simplified application of the MLE approach, underlying model is assumed to be **normal** - or transformed to normality

University of Glasgow

Statistical Approach 1 : Maximum Likelihood

Maximum Likelihood (MLE) -- fits 'best' lognormal distribution to the data, and then



determines summary statistics of the fitted distribution to represent the data.

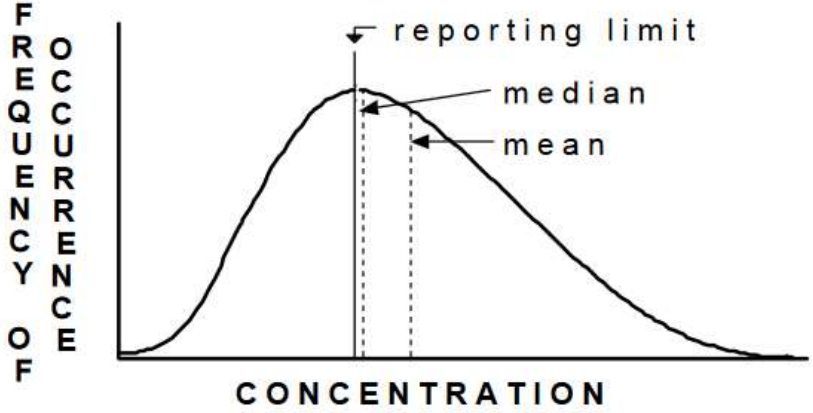


Figure 13.2. Distributional (MLE) method for computing summary statistics.

Helsel

University of Glasgow

Statistical Approach 1 : Maximum Likelihood

Strengths

- General MLE approach can handle multiple LOD – Cohen's can only handle a single LOD
- Can be a rigorous way to estimate summary statistics of data sets when the sample size is sufficiently large.
- If the underlying distribution is known, MLE will explicitly account for distribution type in calculating estimates.

University of Glasgow

Statistical Approach 1 : Maximum Likelihood

Weaknesses

- Poor specification of distribution can produce incorrect estimates
- MLE is most generally applicable to larger data sets (n > 50) with high detection frequencies.
- If data is log normal and log transformation applied, parameter estimates produced using this method are on log-transformed scale
- **Back-transforming** potentially produces **biased** estimators (non-linear relationship between the different scales)

University of Glasgow

Statistical Approach 2 : Kaplan Meier

Non-parametric approach

- Often used in survival analysis for estimating the summary statistics for data where there are right censored observations.
- It can also be applied to data where there are left censored observations by 'flipping' the data and subtracting them from a fixed constant
- Kaplan-Meier estimator estimates the survival function, maps the probability that obs will survive onto time.
- Translated into the context of left censored obs this is the probability that obs will fall below the limits of detection.

University of Glasgow

Example: Cadmium in Fish

Cadmium concentrations in fish for two regions of the Rocky Mountains.

Are the Cd concentrations the same or different in fish livers of the two regions?

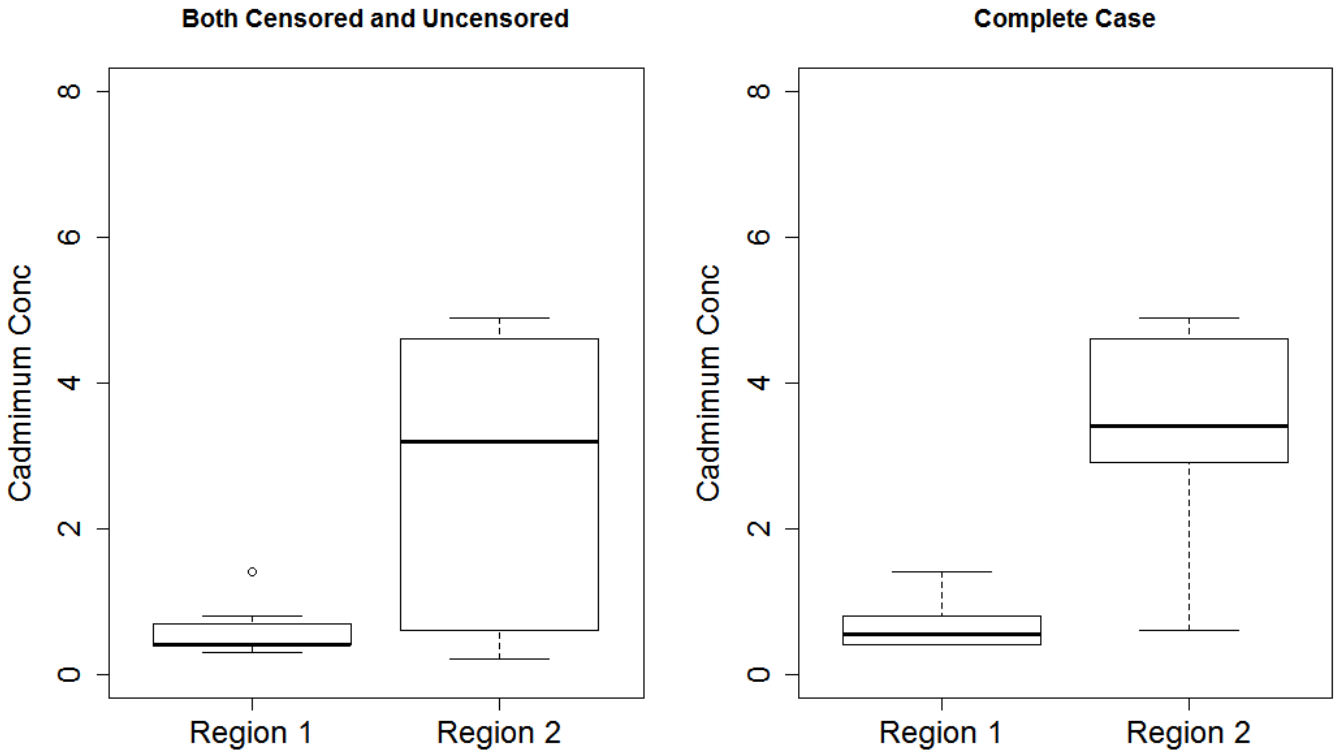
Four detection limits, at 0.2, 0.3, 0.4, and 0.6 ug/L

Cd	Region	CdCen
81.3	2	FALSE
3.5	2	FALSE
4.6	2	FALSE
0.6	2	FALSE
2.9	2	FALSE
3	2	FALSE
4.9	2	FALSE
0.6	2	FALSE
3.4	2	FALSE
0.4	1	FALSE
0.8	1	FALSE
0.3	1	TRUE
0.4	1	FALSE
0.4	1	FALSE
0.4	1	TRUE
1.4	1	FALSE
0.6	1	TRUE
0.7	1	FALSE
0.2	2	TRUE

Fig from USGS, *Example from Helsel*

University of Glasgow

Statistical Approach 2 : Kaplan Meier
Example: Cadmium



We're going to use the NADA package in R

Nondetects And Data Analysis for environmental data

University of Glasgow

Statistical Approach 2 : Kaplan Meier
Example: Cadmium

`cenfit` calculates ECDF for censored data using Kaplan-Meier method.

NADA package 'flips' the data

Group medians very different (both close to LoD)

Standard deviations are very different

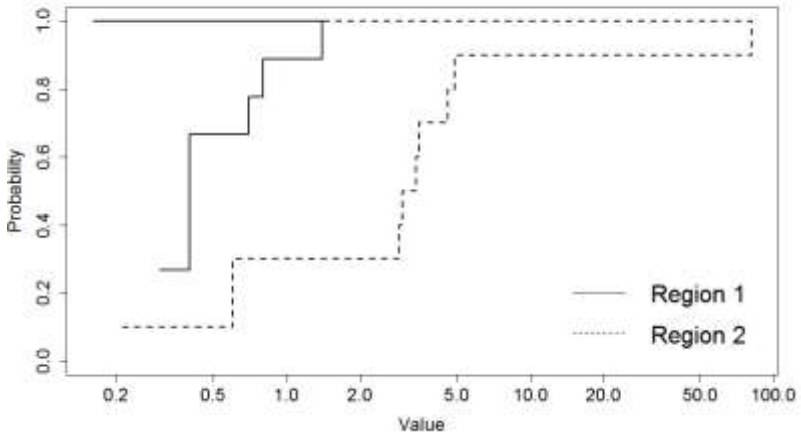
```
fishy <- cenfit(obs, censored, groups)
```

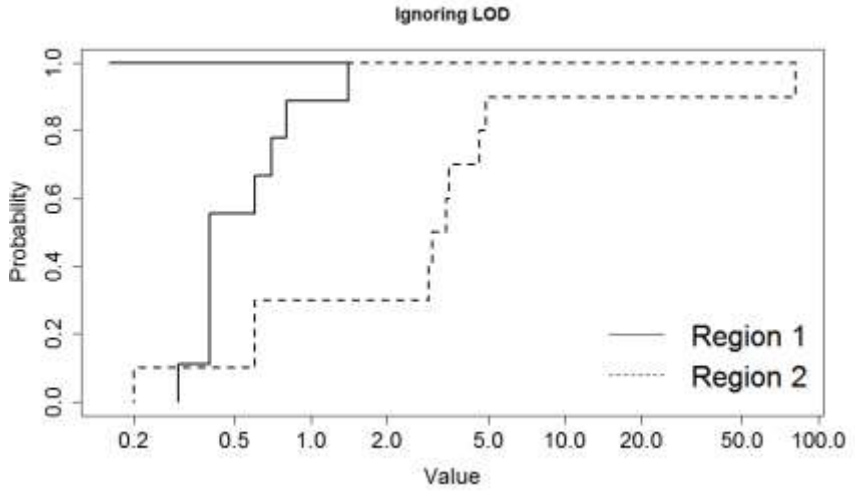
```
           n    n.cen   median        mean        sd
groups=1   9        3      0.4       0.589     0.352
groups=2  10        1      3.0      10.540    25.069
```

University of Glasgow

Statistical Approach 2 : Kaplan Meier
Example: Cadmium

Note that ecdf plots produced in NADA appear "backwards" from survival function plots for right-censored time to event data, as these are left-censored values.





University of Glasgow

Statistical Approach 2 : Kaplan Meier
Example: Cadmium

`cendiff` function tests for difference between the groups

H0: Median Cadmium levels are the same in Region 1 and Region 2
H1: Median Cadmium levels are the different in Region 1 and Region 2

```
> cendiff(obs, censored, groups)
                 N  Observed Expected (O-E)^2/E (O-E)^2/V
groups=Region1   9      2.84     6.13      1.76      7.02
groups=Region2  10      6.84     3.55      3.05      7.02

 Chisq= 7  on 1 degrees of freedom, p= 0.00808
```

University of Glasgow

Statistical Approach 2 : Kaplan Meier

Strengths

- Non parametric so no underlying distribution needs to be assumed – often useful for many environmental data
- Can accommodate multiple reporting limits
- Routinely used with data sets having a lower than 50% detection frequency.

Weaknesses

- Cannot rank censored data points with reporting limits above the highest detected concentration.
- If there is only a single LOD then this is the same as simple substitution

University of Glasgow

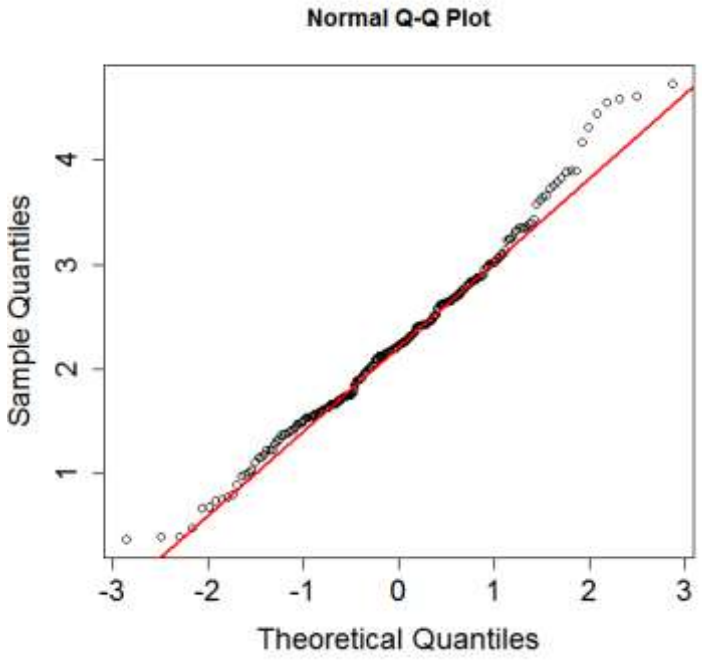
Statistical Approach 3 : Regression on Order Statistics

- Semi-parametric method for computing summary statistics of a distribution where there are left censored non-detect data.
- Within this method, left censored observations are modelled using a linear regression model of the observed un-censored values against their normal quantiles.
- 2 steps
- Use the `cenros` function within NADA package in R

University of Glasgow

Statistical Approach 3 : ROS STEP 1

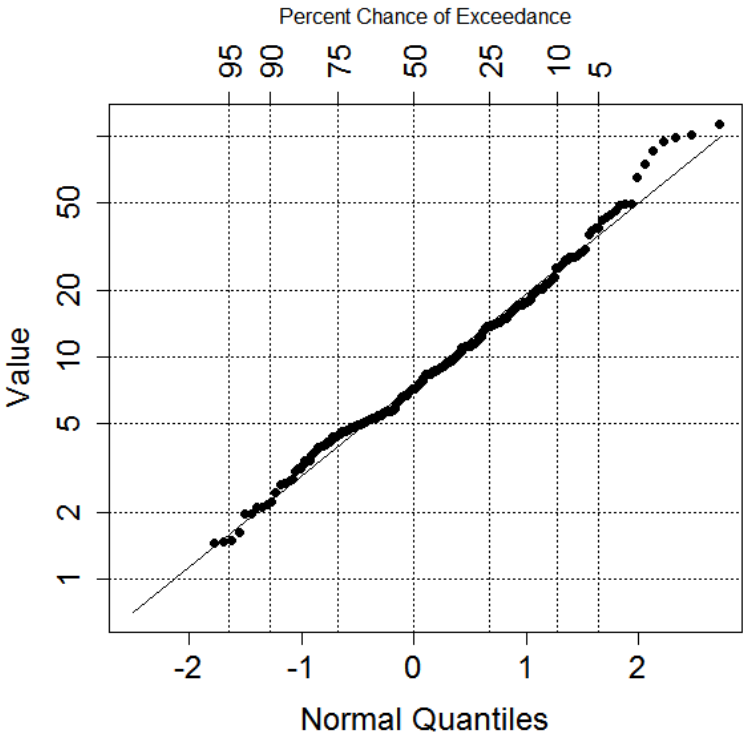
- A linear regression is formed using the plotting positions of the uncensored observations and their normal quantiles.



University of Glasgow

Statistical Approach 3 : ROS STEP 1 cont

- Fitted model is used to estimate the values of the censored observations as a function of their normal quantiles.
- The plot shows the uncensored observations and the probability plot - regression model
- NADA uses lognormal by default. Plot shows that a log-normal model fits these data well



University of Glasgow

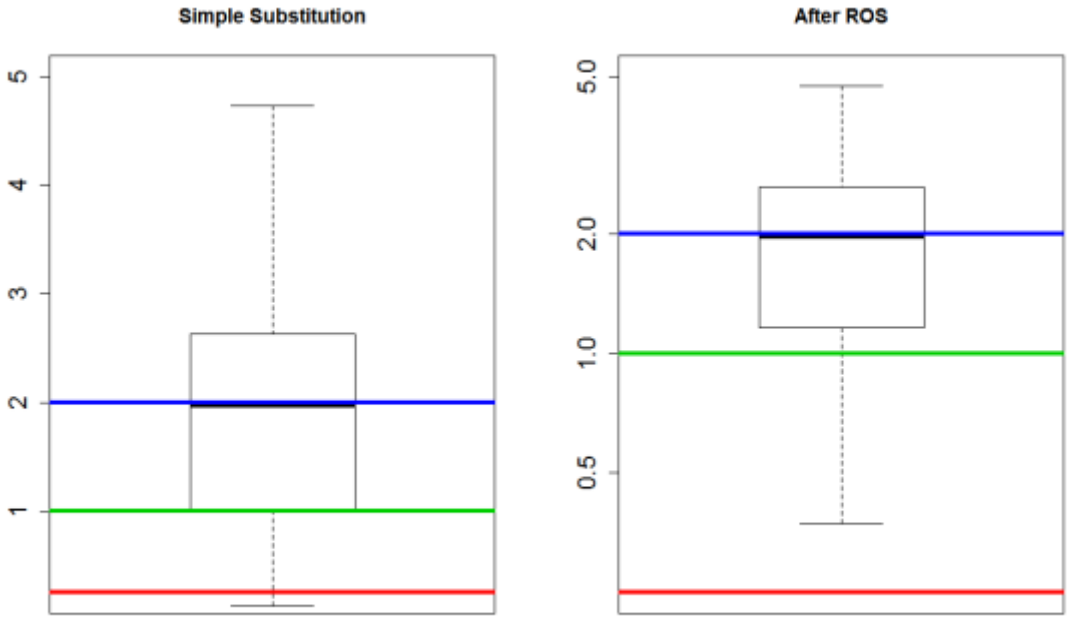
Statistical Approach 3 : ROS STEP 1 cont

```
> summary(tt)
Call:
lm(formula = obs.transformed ~ pp.nq)

Residuals:
     Min       1Q   Median       3Q      Max
-0.89861 -0.04406  0.06315  0.08993  0.14522

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.55401    0.01183   46.83   <2e-16 ***
pp.nq        0.52179    0.01410   37.02   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1638 on 233 degrees of freedom
Multiple R-squared:  0.8547,Adjusted R-squared:  0.854
F-statistic:  1370 on 1 and 233 DF,  p-value: < 2.2e-16
```

University of Glasgow

Statistical Approach 3 : ROS STEP 2

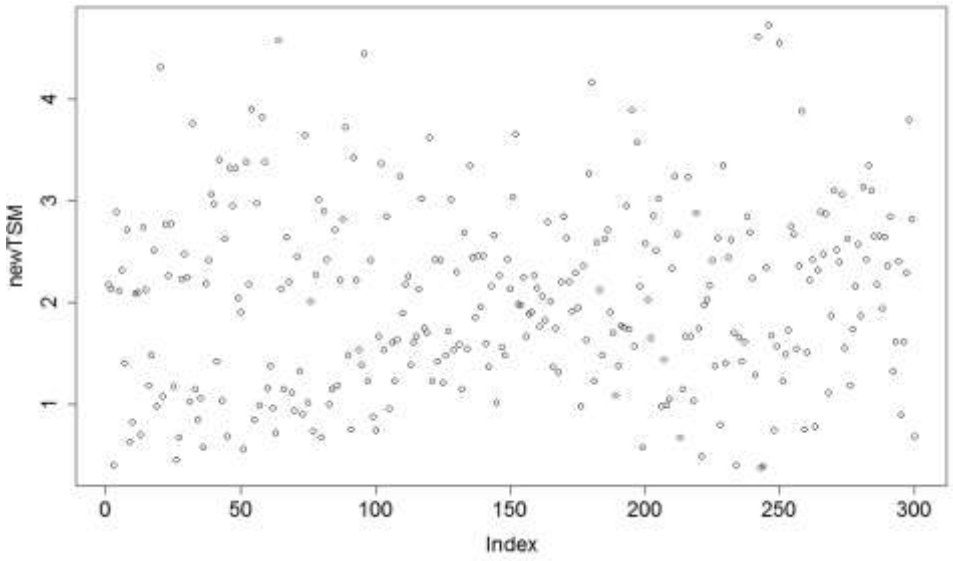
- Observed uncensored values are combined with modelled censored values to estimate summary statistics of the entire population



University of Glasgow

Statistical Approach 3 : ROS STEP 2

- Plot shows TSM with LOD values imputed using ROS method
- No significant trend over time when LOD values are imputed (unlike before)



University of Glasgow

Statistical Approach 3 : ROS
Example: Lead in Herons

Golden (2003) measured lead concentrations in the bodies of black-crowned night herons before and after exposure to doses of lead nitrate in water.

One group were exposed to high concentrations, the other low concentration – is there a difference in the levels found in the blood of the two grousps?

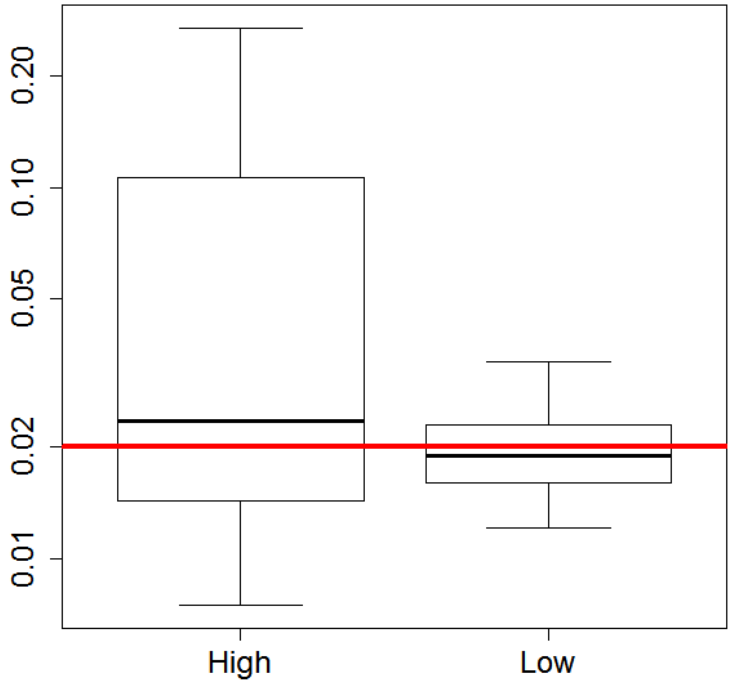
There is one detection limit, at 0.02 ug/g.

Example from Helsel

University of Glasgow

Statistical Approach 3 : ROS
Example: Lead in Herons

```
data(Golden)
attach(Golden)
cenboxplot(Blood, BloodCen, DosageGroup, range=0, lty=1)
```



University of Glasgow

Statistical Approach 3 : ROS

Strengths

- Widely applicable to many environmental data sets.
- Can accommodate multiple reporting limits as well as (unlike Kaplan-Meier) a single reporting limit.
- Can be used where there are up to 80% of observations listed as non-detects (recommended by Helsel, 2005).

Weaknesses

- Semi-parametric method, requires a known distributional model to describe the detected measurements.

University of Glasgow

What we have learned …

Limits of detection are often seen in environmental data

They should not be ignored – this can potentially have a large effect on the analysis!

There are several methods of dealing with LOD – each have strengths and weaknesses

The method chosen will be based on the percentage of LOD values and the assumed distribution of the data