NADA for R
A contributed package for censored environmental data

Dennis Helsel
Practical Stats

Lopaka (Rob) Lee
U.S. Geological Survey

Censored data

- Data known only to be above or below a threshold. The exact, single number is not known.
- In environmental studies, most frequent application is to “nondetects”, values known only to be below reporting (detection) limits.
- <10 = a value measured somewhere between 0 and 10

“Nondetects” occur in many fields

- Water quality
- Air quality
- Soil chemistry
- Geochemistry
- Astronomy
- Occupational health
- Risk analysis
- Biocontaminants

The Problem

- Substitution is the most commonly-used method for incorporating censored environmental data
- $\frac{1}{2}$ or $\frac{1}{\sqrt{2}}$ times RL are the most commonly-used substitutions
- Using ½, each <1 becomes 0.5, each <5 becomes 2.5, etc.

Survival analysis methods perform better than substitution

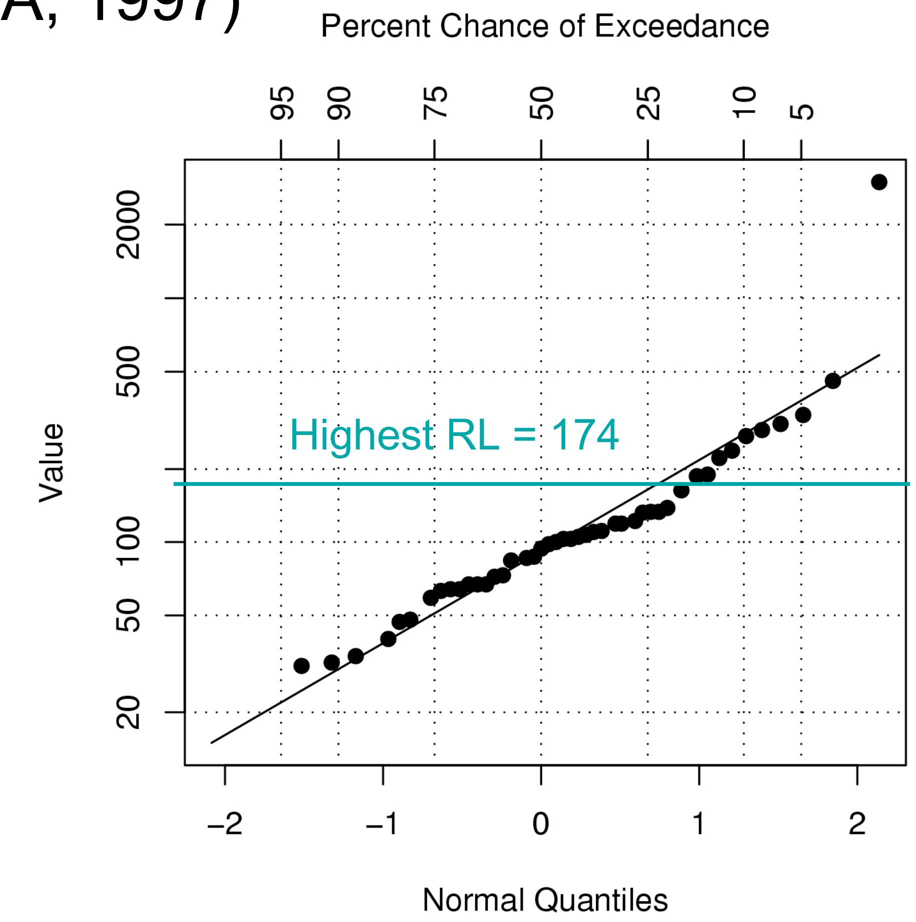
- Survival analysis methods explicitly incorporate censored data
- Substituted value is arbitrary
- No ‘invasive data’ added to the observations measured
- No reason to use substitution except that it is cheap and easy

NADA for R package

- Performs parametric and nonparametric methods for left-censored data
- Consistent function names and usage
- Almost all functions begin with the prefix “cen” -- for example, “cenfit”, and “cenmle”
- Generic functions such as “mean”, “quantile”, and “plot” can be used with output objects from any of the NADA for R functions

Example censored data set

- Pyrene concentrations in benthic sediments. 56 observations, 11 censored at 8 DLs. From She (Journal. AWRA, 1997)

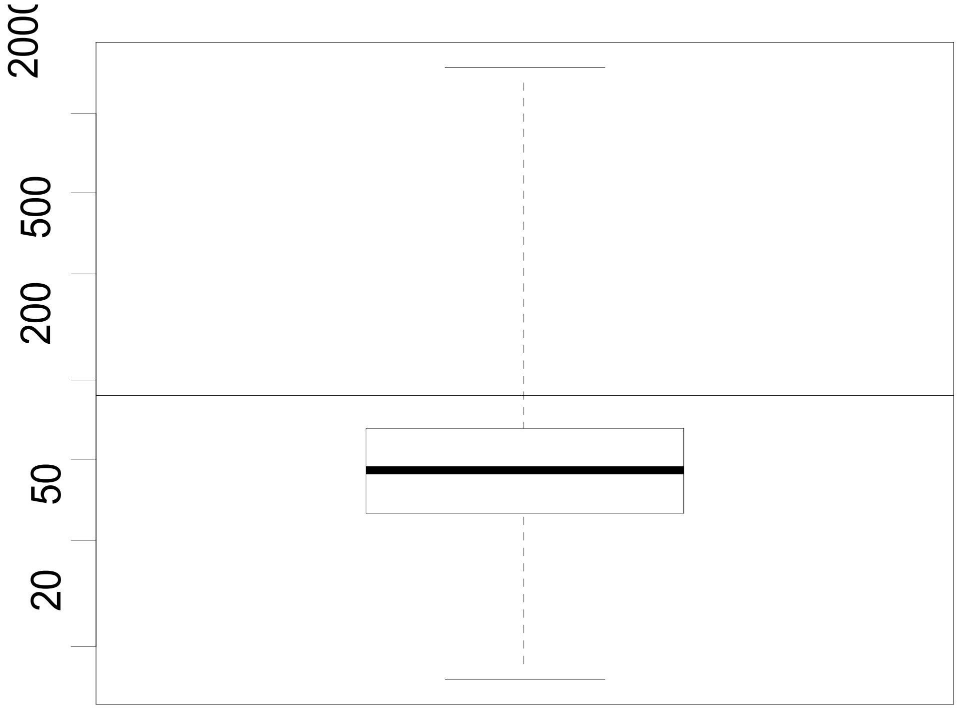


Entering and summarizing data

```
> ShePyrene
   Pyrene PyreneCen
1      28      TRUE
2      31     FALSE
3      32     FALSE
...

> censummary(ShePyrene)
all:
         n      n.cen    pct.cen        min        max
  56.00000   11.00000   19.64286   28.00000 2982.00000

limits:
  limit n uncen   pexceed        limit n uncen   pexceed
1    28 1     3 0.9629191       5   117 1     2 0.3325437
2    35 2     3 0.8516764       6   122 1     5 0.2920918
3    58 1    10 0.7775146       7   163 3     1 0.1964286
4    86 1    11 0.5550292       8   174 1    10 0.1785714
```

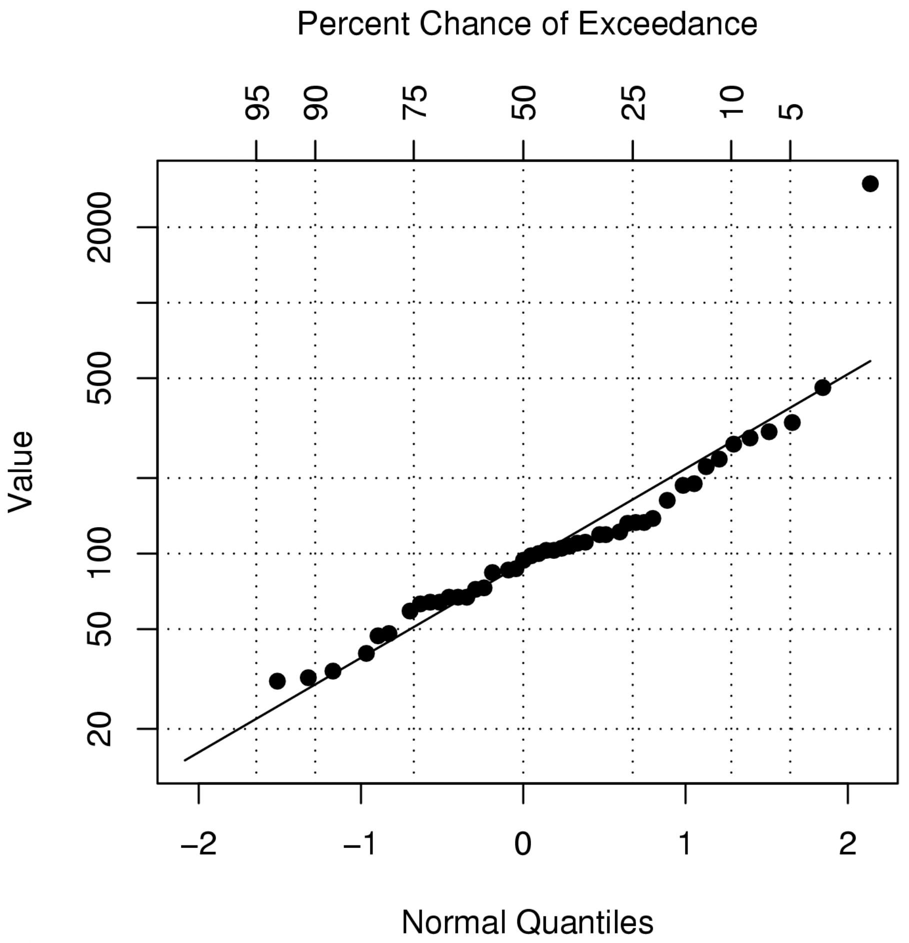
Plotting Censored Data

```
> cenboxplot(Pyrene, PyreneCen)
```



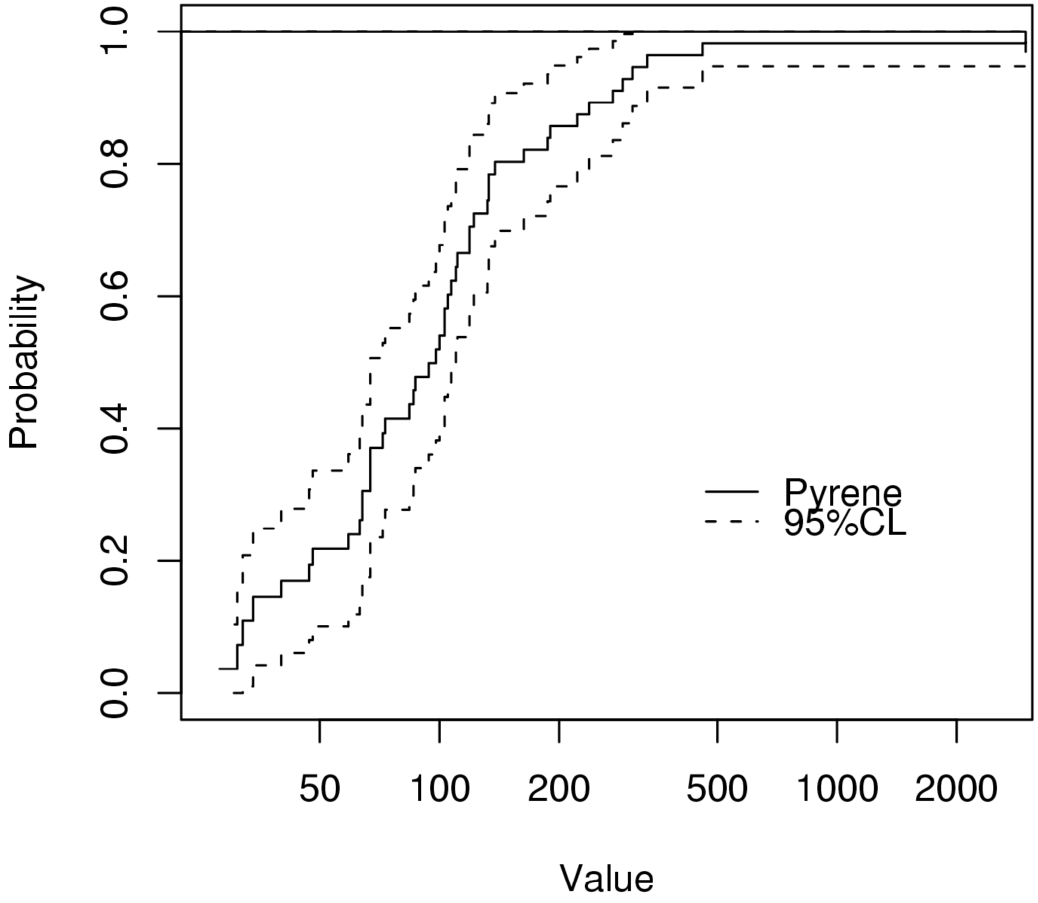
Plotting Censored Data

- Censored probability plot



Plotting Censored Data

- Survival curve (a cdf for left-censored data)



Three Valid Approaches for the Analysis of Censored Data

1. Parametric methods. Assume data follow a specific distribution.
   - Maximum likelihood estimation (MLE)
2. "Robust" methods
   - Regression on Order Statistics (ROS)
3. Nonparametric methods. Based on percentiles, ranks.
   - Kaplan-Meier
   - Wilcoxon score tests
   - Kendall's tau

Estimating Descriptive Statistics

MLE for Pyrene data - using cenmle function.

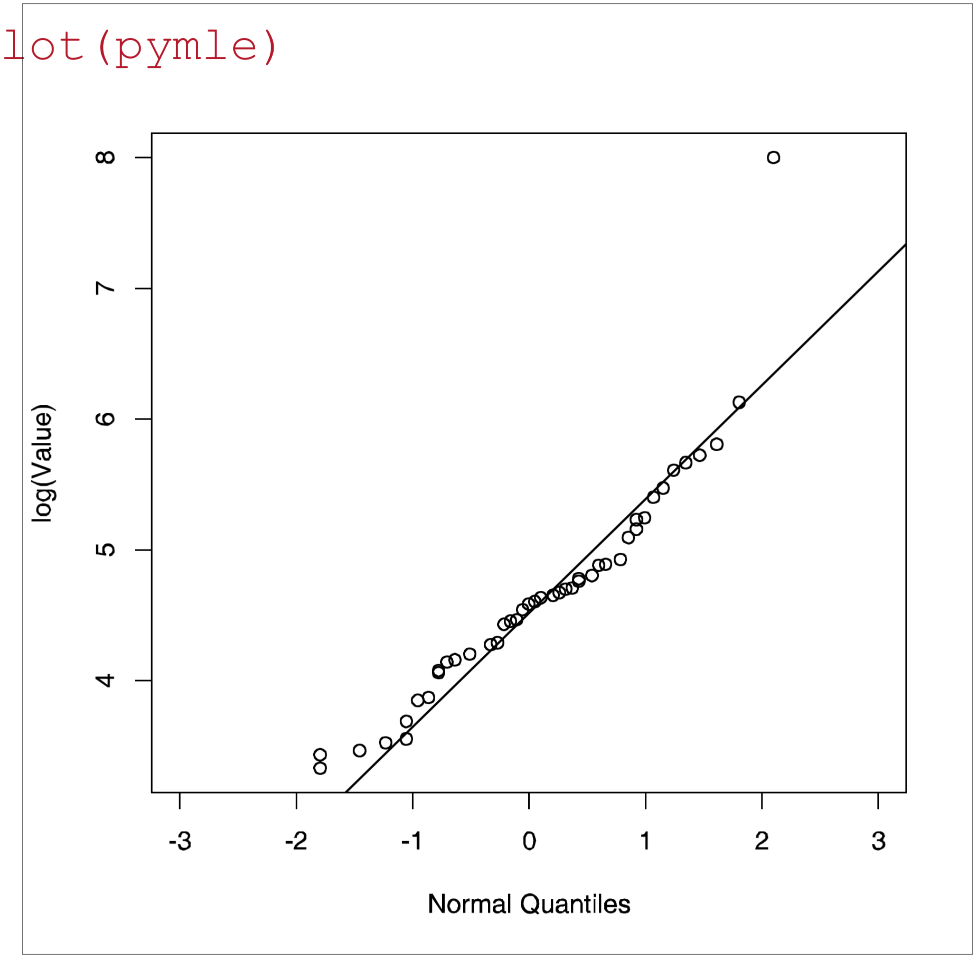
Lognormal distribution is assumed by default

```
> pymle = cenmle(Pyrene, PyreneCen)
> pymle
       n     n.cen    median      mean         sd
 56.0000   11.0000   90.5000 163.1531 393.1309

> summary(pymle)
             Value Std. Error     z          p
(Intercept)  4.518      0.122 37.08 6.22e-301
Log(scale)  -0.138      0.106 -1.30  1.94e-01

Scale= 0.871

Log Normal distribution
```

Parametric Method: MLE

Check residuals to see if they follow a lognormal distribution

```
> plot(pymle)
```



Estimating Descriptive Statistics

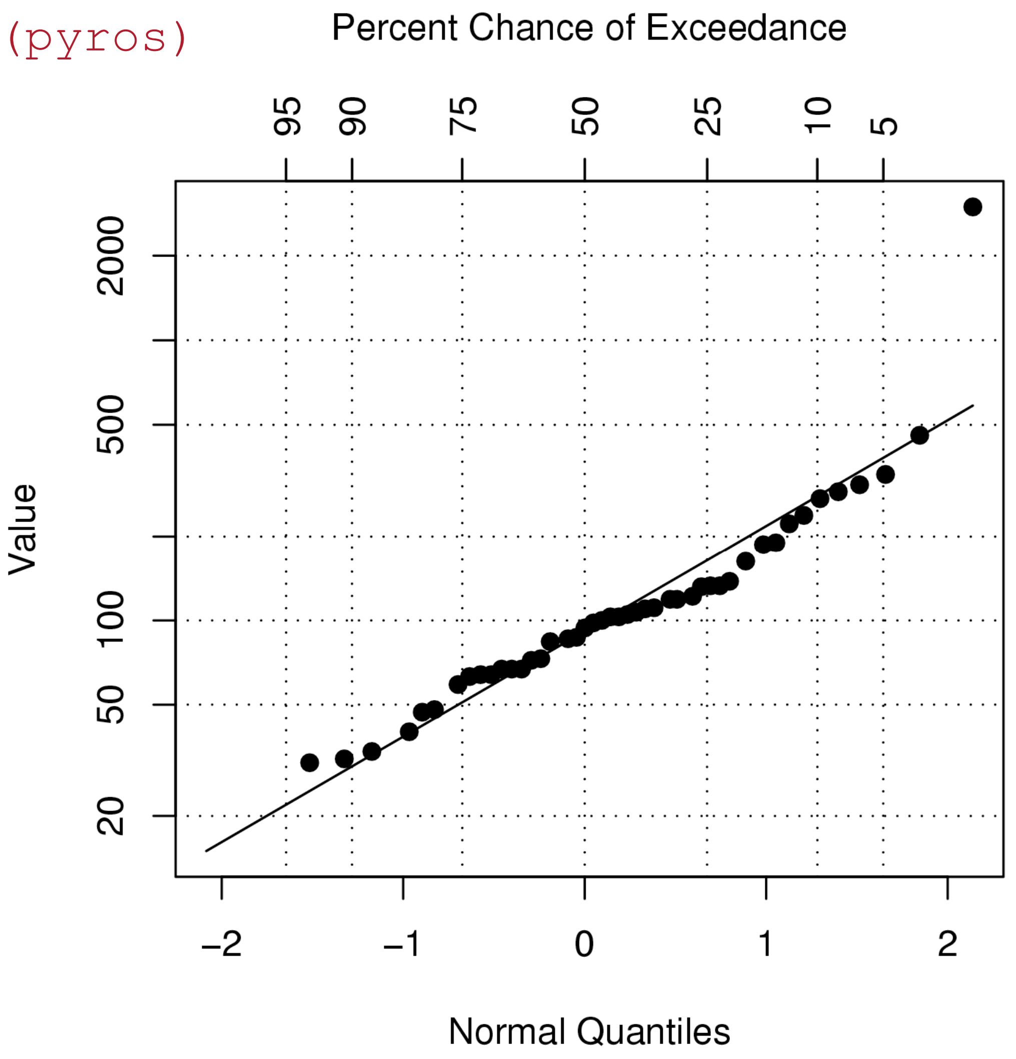
Robust Regression on Order Statistics (ROS)

```
> pyros = cenros(Pyrene, PyreneCen)
> pyros

       n    n.cen   median     mean       sd
 56.0000  11.0000  90.5000 163.1531 393.1309
```

ROS is not strongly sensitive to choice of distribution. Can check with probability plot.

Regression on Order Statistics

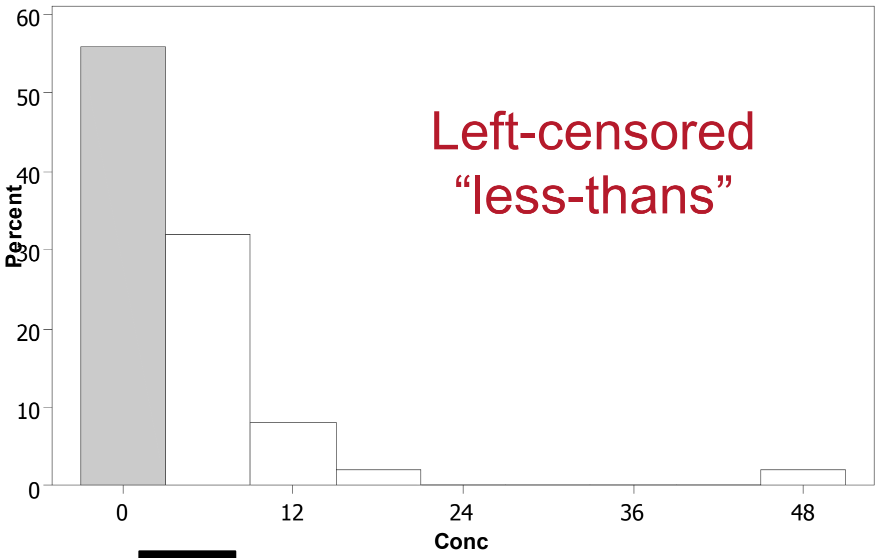
```
> plot(pyros)
```



Kaplan-Meier (nonparametric) method

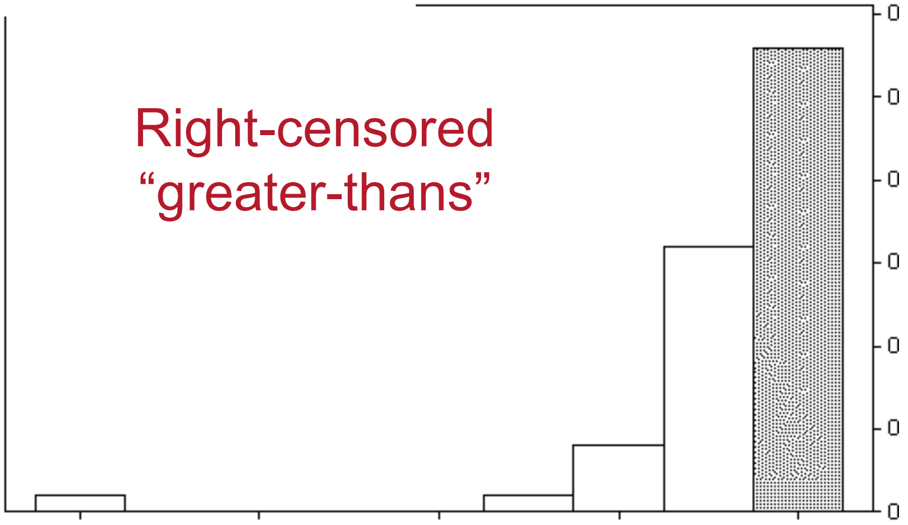
- Standard method in medical and industrial statistics
- Software currently built for right-censored data, so left-censored data must be flipped: flip = Constant - X.
- Estimates the survival function S, which becomes the CDF (percentiles) of the original X data.

Commercial stat software: must ‘flip’ the data manually



Left-censored
“less-thans”

Flip



Right-censored
“greater-thans”

Flipping done automatically in NADA for R

Estimating Descriptive Statistics

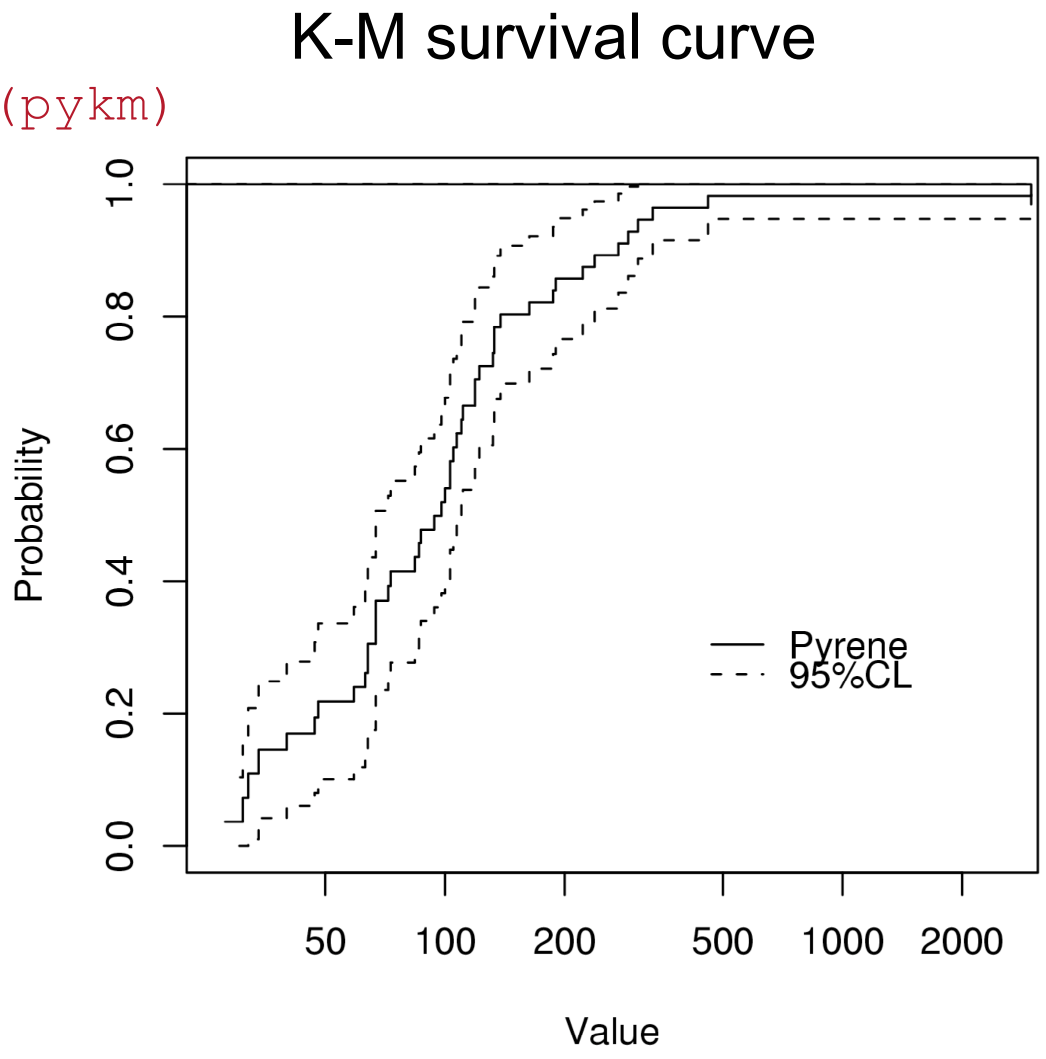
Kaplan-Meier using cenfit command

Cenfit is is analgous to the "survfit" function in the survival package

```
pykm = cenfit(Pyrene, PyreneCen)

> pykm
       n    n.cen   median     mean       sd
 56.0000  11.0000  98.0000 164.0945 389.5899
```

Estimating Descriptive Statistics

```
> Plot (pykm)
```



Estimating Descriptive Statistics

All 3 methods with censtats

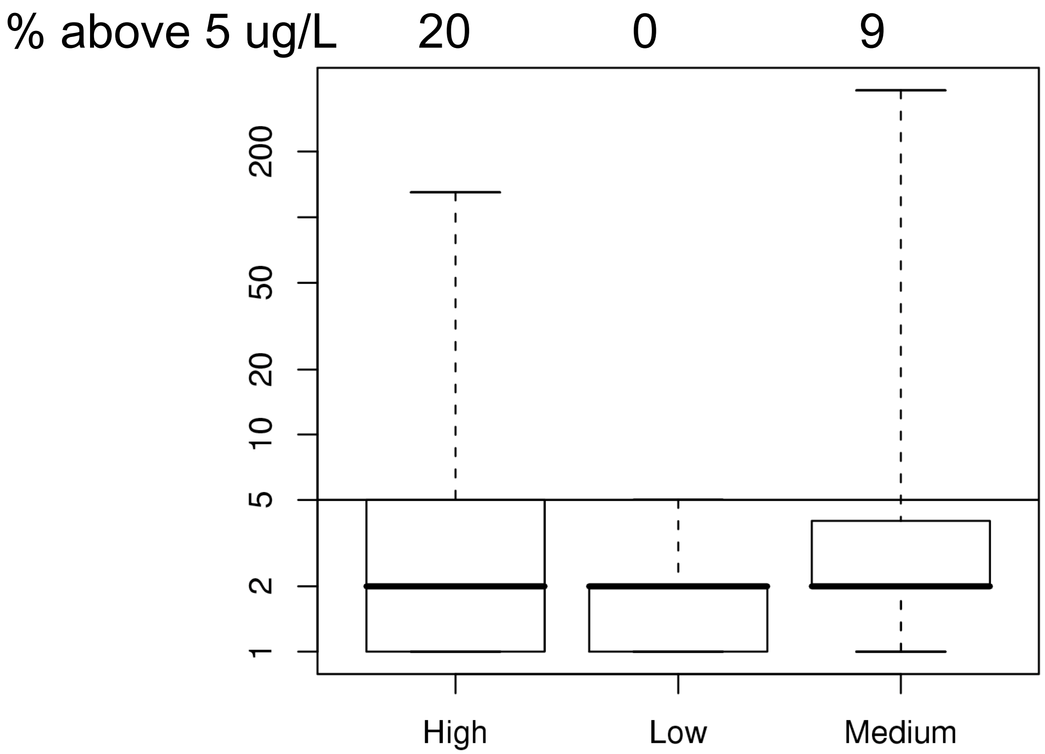
```
> Pystats =censtats(Pyrene, PyreneCen)
> pystats
       n    n.cen  pct.cen
56.00000 11.00000 19.64286

          median     mean        sd
K-M     98.00000 164.0945  389.5899
ROS     90.50000 163.1531  393.1309
MLE     91.64813 133.9142  142.6698
```

None of these 3 methods required substitution

ANOVA using censored regression

Are these 3 distributions the same, or different?

```
> cenboxplot(TCEConc, TCECen, Density)
```



ANOVA using censored regression

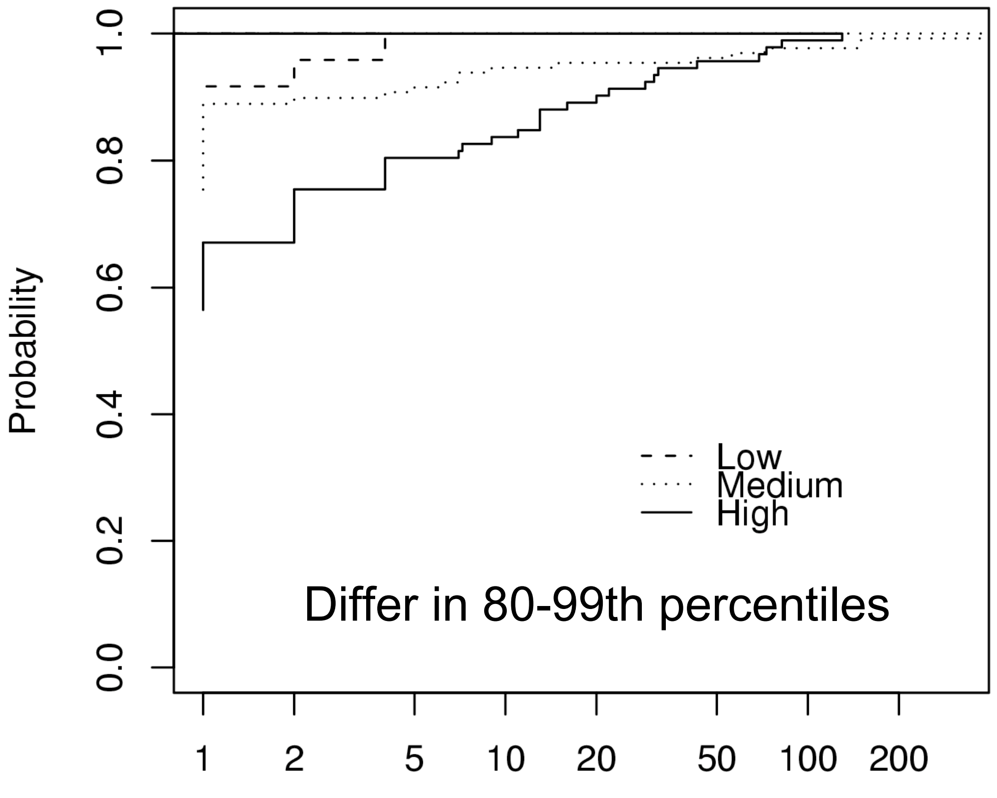
```
> tcemle = cenmle(TCEConc, TCECen, Density)
> summary(tcemle)
               Value Std. Error     z        p
(Intercept)   -0.722      0.416 -1.73 8.28e-02
DensityLow    -3.060      1.138 -2.69 7.17e-03
DensityMedium -1.656      0.553 -2.99 2.76e-03
Log(scale)     1.048      0.111  9.41 4.76e-21
Scale= 2.85

Log Normal distribution
Loglik(model)= -308.7
Loglik(intercept only)= -316.4
Loglik-r:  0.2459125
Chisq= 15.41 on 2 degrees of freedom, p= 0.00045
```

Wilcoxon tests with censored data

Nonparametric

```
> cendiff(TCEConc, TCECen, Density)
            N Observed Expected (O-E)^2/E (O-E)^2/V
Dens=High  92    30.45     18.2      8.26     15.65
Dens=Low   25     1.73      5.7      2.76      3.62
Dens=Med  130    15.47     23.8      2.89      6.76
 Chisq= 16.3  on 2 degrees of freedom, p= 0.000295
```

Wilcoxon tests for censored data

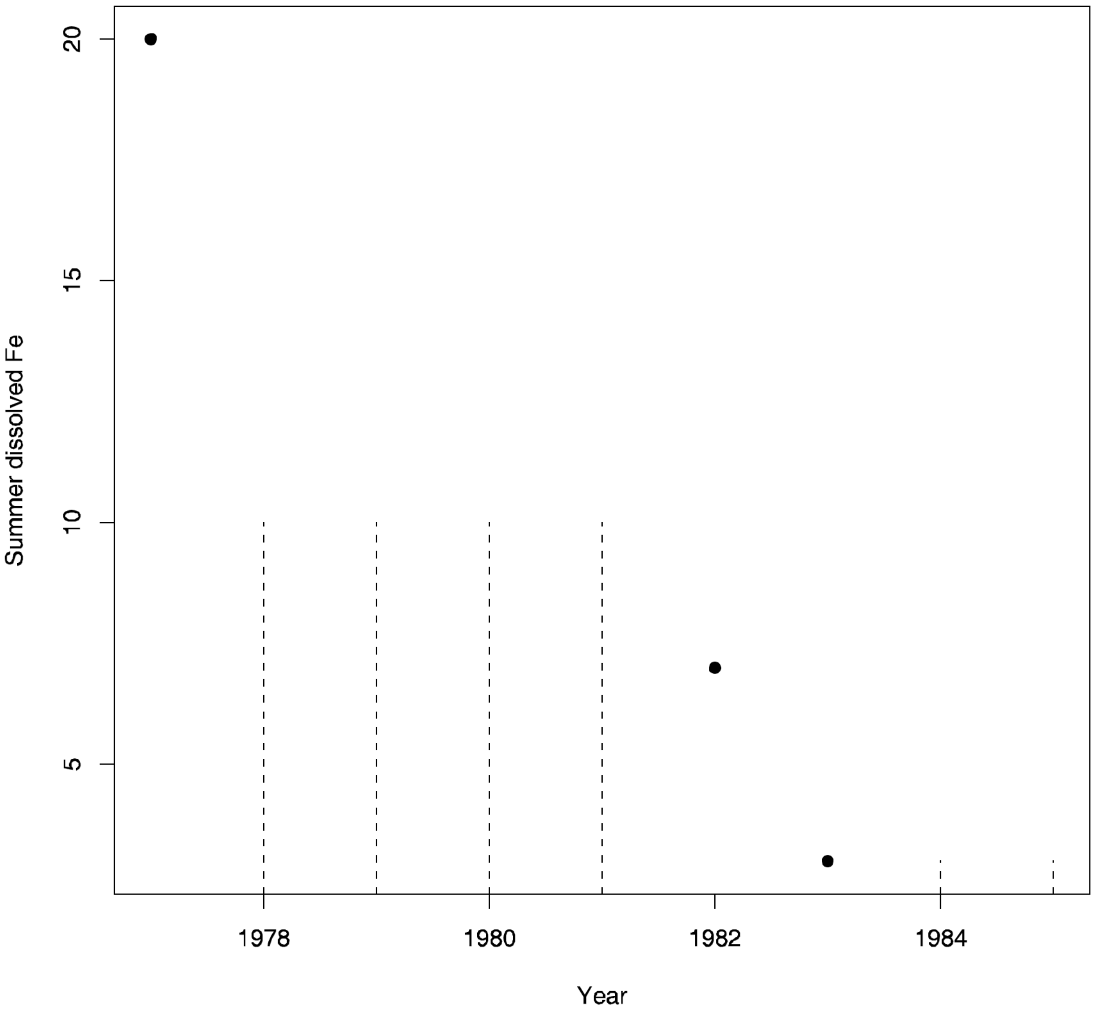


Differ in 80-99th percentiles

Score test looks for differences among survival curves (cdfs) for the three land-use groups.

Correlation and regression with censored data

```
> cenxyplot(Year, YearCen, Summer, SummerCen)
```



Is there a correlation between Dissolved Iron and Year?

What equation best describes the trend?

Parametric Censored Regression

```
> cenreg(Cen(Summer, SummerCen)~Year)

              Value Std. Error     z        p
(Intercept) 507.472   106.3237  4.77 1.82e-06
Year         -0.255     0.0537 -4.76 1.97e-06
Log(scale)   -1.118     0.4106 -2.72 6.48e-03

Scale= 0.327

Log Normal distribution
Loglik(model)= -9.3   Loglik(intercept only)= -12.8
Loglik-r:  0.7371631

Chisq= 7.06 on 1 degrees of freedom, p= 0.0079
```

cenreg is analogous to survreg in the survival package. Data are flipped within cenreg.

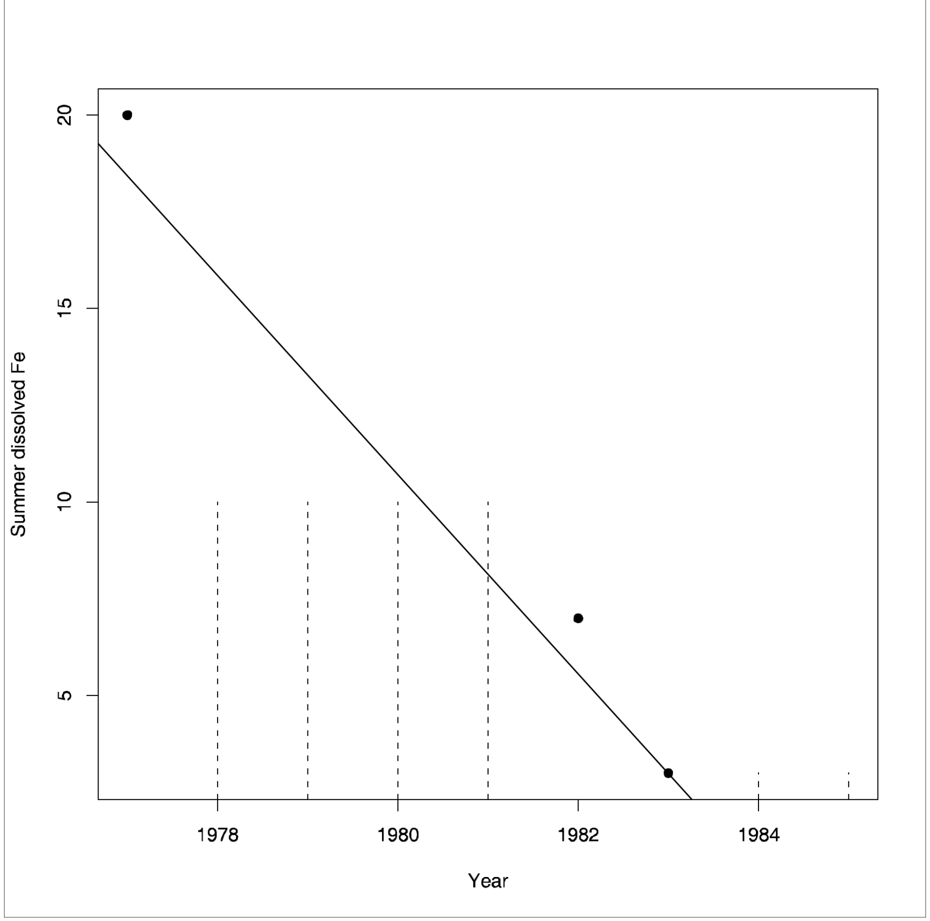
ATS nonparametric line for censored data

Nonparametric approach: ATS version of Thiel-Sen robust line (based on Kendall's tau)

```
> cenken(Summer, SummerCen, Year)

slope
[1] -2.572113

intercept
[1] 5103.5

tau
[1] -0.3611111

p
[1] 0.1315868
```

ATS nonparametric line for censored data

```
> cenxyplot(Year, YearCen, Summer, SummerCen)
> lines(cenken(Summer, SummerCen, Year))
```



More detail is available in the textbook:

Nondetects
And
Data
Analysis

Statistics for Censored Environmental Data

WILEY

Nondetects
And
Data
Analysis

Statistics for Censored Environmental Data

Dennis R. Helsel

Statistics in Practice

by Dennis R. Helsel

Wiley (2005)

www.PracticalStats.com/nada