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
• **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 \( \frac{1}{2} \), 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

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

> censummary(ShePyrene)
all:

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>n.cen</th>
<th>pct.cen</th>
<th>min</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>all</td>
<td>56.00</td>
<td>11.00</td>
<td>19.64</td>
<td>28.00</td>
<td>2982.00</td>
</tr>
</tbody>
</table>

limits:

<table>
<thead>
<tr>
<th>limit</th>
<th>n</th>
<th>uncen</th>
<th>p.exceed</th>
<th>limit</th>
<th>n</th>
<th>uncen</th>
<th>p.exceed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>28</td>
<td>1</td>
<td>0.9629</td>
<td>5</td>
<td>117</td>
<td>1</td>
<td>0.3325</td>
</tr>
<tr>
<td>2</td>
<td>35</td>
<td>2</td>
<td>0.8516</td>
<td>6</td>
<td>122</td>
<td>1</td>
<td>0.2920</td>
</tr>
<tr>
<td>3</td>
<td>58</td>
<td>1</td>
<td>0.7775</td>
<td>7</td>
<td>163</td>
<td>3</td>
<td>0.1964</td>
</tr>
<tr>
<td>4</td>
<td>86</td>
<td>1</td>
<td>0.5550</td>
<td>8</td>
<td>174</td>
<td>1</td>
<td>0.1786</td>
</tr>
</tbody>
</table>
```
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)

   - Kaplan-Meier
   - Wilcoxon score tests
   - Kendall’s tau
MLE for Pyrene data - using cenmle function.

Lognormal distribution is assumed by default

```r
> 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)

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

<table>
<thead>
<tr>
<th>n</th>
<th>n.cen</th>
<th>median</th>
<th>mean</th>
<th>sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>56.0000</td>
<td>11.0000</td>
<td>90.5000</td>
<td>163.1531</td>
<td>393.1309</td>
</tr>
</tbody>
</table>
```

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”

Flipping done automatically in NADA for R

Right-censored “greater-thans”
Estimating Descriptive Statistics

Kaplan-Meier using cenfit command

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

```r
pykm = cenfit(Pyrene, PyreneCen)

> pykm
   n  n.cen  median    mean    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

<table>
<thead>
<tr>
<th>n</th>
<th>n.cen</th>
<th>pct.cen</th>
</tr>
</thead>
<tbody>
<tr>
<td>56.00000</td>
<td>11.00000</td>
<td>19.64286</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>median</th>
<th>mean</th>
<th>sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-M</td>
<td>98.00000</td>
<td>164.0945</td>
</tr>
<tr>
<td>ROS</td>
<td>90.50000</td>
<td>163.1531</td>
</tr>
<tr>
<td>MLE</td>
<td>91.64813</td>
<td>133.9142</td>
</tr>
</tbody>
</table>

None of these 3 methods required substitution
ANOVA using censored regression

Are these 3 distributions the same, or different?

> cenboxplot(TCEConc, TCECen, Density)

% above 5 ug/L

20 0 9

High Low Medium
ANOVA using censored regression

> tcemle = cenmle(TCEConc, TCECen, Density)
> summary(tcemle)

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>Std. Error</th>
<th>z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-0.722</td>
<td>0.416</td>
<td>-1.73</td>
<td>8.28e-02</td>
</tr>
<tr>
<td>DensityLow</td>
<td>-3.060</td>
<td>1.138</td>
<td>-2.69</td>
<td>7.17e-03</td>
</tr>
<tr>
<td>DensityMedium</td>
<td>-1.656</td>
<td>0.553</td>
<td>-2.99</td>
<td>2.76e-03</td>
</tr>
<tr>
<td>Log(scale)</td>
<td>1.048</td>
<td>0.111</td>
<td>9.41</td>
<td>4.76e-21</td>
</tr>
</tbody>
</table>

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)

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Observed</th>
<th>Expected</th>
<th>(O-E)^2/E</th>
<th>(O-E)^2/V</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dens=High</td>
<td>92</td>
<td>30.45</td>
<td>18.2</td>
<td>8.26</td>
<td>15.65</td>
</tr>
<tr>
<td>Dens=Low</td>
<td>25</td>
<td>1.73</td>
<td>5.7</td>
<td>2.76</td>
<td>3.62</td>
</tr>
<tr>
<td>Dens=Med</td>
<td>130</td>
<td>15.47</td>
<td>23.8</td>
<td>2.89</td>
<td>6.76</td>
</tr>
</tbody>
</table>

Chisq= 16.3 on 2 degrees of freedom, p= 0.000295
Wilcoxon tests for censored data

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

Differ in 80-99th percentiles
Correlation and regression with censored data

Is there a correlation between Dissolved Iron and Year?

What equation best describes the trend?

cenxyplot(Year, YearCen, Summer, SummerCen)
Parametric Censored Regression

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

<table>
<thead>
<tr>
<th>Value</th>
<th>Std. Error</th>
<th>z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>507.472</td>
<td>106.3237</td>
<td>4.77</td>
</tr>
<tr>
<td>Year</td>
<td>-0.255</td>
<td>0.0537</td>
<td>-4.76</td>
</tr>
<tr>
<td>Log(scale)</td>
<td>-1.118</td>
<td>0.4106</td>
<td>-2.72</td>
</tr>
</tbody>
</table>

Scale= 0.327  

\(\text{cenreg is analogous to survreg in the survival package. Data are flipped within cenreg.}\)

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
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:

**Non-detects And Data Analysis**
Statistics for Censored Environmental Data

by Dennis R. Helsel
Wiley (2005)

www.PracticalStats.com/nada