Chapter 11

Censored Data

Dealing with "Below Detection Limit" Values

In the last chapter we discussed calibration and detection limits. Environmental data often include values reported as "below detection limit" along with the stated detection limit (e.g., Porter et al., 1988; USEPA, 2009; Helsel, 2012). A sample contains ***censored observations*** if the only information about some of the observations is that they are below or above a specified value. In the environmental literature, such values are often called "nondetects". Although this results in some loss of information, we can still do graphical and statistical analyses of data with nondetects. Statistical methods for dealing with censored data have a long history in the field of survival analysis and life testing (Kalbfleisch and Prentice, 1980; Miller, 1981b; Nelson, 1982; Lee and Wang, 2003; Hosmer et al., 2008; Therneau and Grambsch, 2010). In this chapter, we will discuss how to create graphs, estimate distribution parameters, perform goodness-of-fit tests, compare distributions, and fit linear regression models to data with censored values.

11.1 Types of Censored and Related Data

Environmental data with below detection-limit observations are an example of Type I left censored data. A data set may have a single or multiple detection limits. Type I, left, censored, and single are specific choices of four characteristics of data (Cohen, 1991, pp. 3-5):

1. observed, truncated, or censored
2. left, right, or interval censored
3. Type I, Type II, or randomly censored
4. single or multiple censoring values

11.1.1 Uncensored, Censored, or Truncated

Uncensored values are those we are used to dealing with. The value is reported and used as given. ***Censored values*** are those reported as less than some value (e.g., < 5 ppb), greater than some value (e.g., > 100 days), or as an interval (e.g., a value between 67 and 75 degrees). ***Truncated values*** are those that are not reported if the value exceeds some limit. Consider an analytical laboratory reporting the

concentration of atrazine in a ground water sample. If they report 0.02 ppb, that value is observed. If they report $<$ 0.05 ppb, the value is censored. If they only report the concentration if it exceeds 0.1 ppb, the values are truncated. The practical difference between censored and truncated data is that the number of censored values is known, but the number of truncated values is not.

Observed and censored environmental data are much more common than truncated data. The two most common situations producing truncated data are when exceedances, i.e. concentrations above a reporting limit, are the only reported values and when below detection-limit values are incorrectly read into a computer program. When below detection-limit values are stored as"<5", most software will read that as a character string, not a number. When used in a numerical analysis, that value will be converted to a missing value, erroneously truncating the data.

11.1.2 Left, Right, or Interval Censoring

A ***left censored*** value is one that is known only to be less than some value, e.g. $<$ 5 ppm. A ***right censored*** value is one that is known only to be more than some value. A value is ***interval censored*** if it is reported as being within a specified interval, e.g. 5 ppb $< X \leq$10 ppb. Any observation of a continuous random variable could be considered interval censored, because its value is reported to a few decimal places. A value of $X =$ 25 ppb might be interpreted as 24.5 ppb$\leq X <$ 25.5 ppb. This sort of fine-scale interval censoring is usually ignored and the values are treated as exactly observed. When the intervals are large and the range of the data is small, e.g., 10 or fewer intervals over the range of the data, it is better to consider values as interval censored (Dixon and Newman 1991). Truncated data are also described as left truncated, right truncated, or (very rarely) interval truncated.

11.1.3 Type I, Type II, or Random Censoring

A sample is ***Type I censored*** when the censoring levels are known in advance. The number of censored observations c (and hence the number of uncensored observations n) is a random outcome, even if the total sample size, N, is fixed. Environmental data are almost always type I censored.

A sample is ***Type II censored*** if the sample size N and number of censored observations c (and hence the number of uncensored observations n) are fixed in advance. The censoring level(s) are random outcomes. Type II censored samples most commonly arise in time-to-event studies that are planned to end after a specified number of failures, and Type II censored samples are sometimes called failure-censored samples (Nelson, 1982, p.248).

A sample is ***Randomly Censored*** when both the number of censored observations and the censoring levels are random outcomes. This type of censoring commonly arises in medical time-to-event studies. A subject who moves away from the study area before the event of interest occurs has a randomly censored value. The outcome for a subject can be modeled as a pair of random variables, (X, C), where X is the random time to the event and C is the random time until the subject moves away. X is an observed value if $X < C$ and right censored at C if $X > C$.

11.1.4 Single or Multiple Censoring

A sample is ***singly censored*** (e.g., singly left censored) if there is only one censoring level T. (Technically, left censored data are ***singly left censored*** only if all n uncensored observations are greater than or equal to T, and right-censored data are ***singly right censored*** only if all n uncensored observations are less than

or equal to T (Nelson, 1982, p.7); otherwise, the data are considered to be ***multiply censored***.)

A sample is ***multiply censored*** if there are several censoring levels $T_1, T_2, ..., T_p$, where $T_1 < T_2 < \cdots < T_p$. Multiple censoring commonly occurs with environmental data because detection limits can change over time (e.g., because of analytical improvements), or detection limits can depend on the type of sample or the background matrix. The distinction between single and multiple censoring is mostly of historical interest. Some older statistical methods are specifically for singly censored samples. Most currently recommended methods can be used with either singly or multiply censored samples, but the implementation is often easier with one censoring level.

11.2 Examples of Data Sets with Censored values

We have already seen some data sets with censored observations. We use these and others to illustrate the various types of censored data.

11.2.1 Type I Left Singly Censored Data

The benzene data presented in Table **??** illustrate type I left singly censored data with a single censoring level of 2 ppb. There are $N = 36$ observations, with c= 33 censored observations and n= 3 uncensored observations. The trichloroethylene data presented in Table **??** have a single censoring level of 5 ppb, with N= 24, c = 10, and n= 14. The Skagit data set (stored as Skagit.NH3_N.df in EnvStats) contains 395 monthly measurements of ammonia nitrogen (NH_3-N) concentrations (mg/l) in the Skagit River (Marblemount, Washington station) made from January 1978 through December 2010. Table 11.1 shows the 60 observations made between January 1978 and December 1982, ordered from smallest to largest, for which all censored observations were censored at 0.01 mg/l. Notice that there are samples with observed values of 0.01 mg/l, which will be treated differently from the 16 samples reported as < 0.01 mg/l.

Table 11.1: NH_3-N concentrations, as mg/l, measured in the Skagit River, Washington State, between January 1978 and December 1982.

NH_3-N concentration (mg/l)								
<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.01	0.01
0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
0.02	0.02	0.02	0.02	0.02	0.02	0.03	0.03	0.03
0.03	0.03	0.03	0.03	0.04	0.04	0.04	0.04	0.04
0.04	0.05	0.05	0.05	0.06	0.47			

11.2.2 Type I Left Multiply Censored Data

Table 11.2 displays copper concentrations (μg/L) in shallow groundwater samples from two different geological zones in the San Joaquin Valley, California (Millard and Deverel, 1988). The alluvial fan data include four different detection limits and the basin trough data include five different detection limits. The Olympic data set (stored as Olympic.NH4.df in EnvStats) contains approximately weekly or bi-weekly observations of ammonium (NH_4) concentrations (mg/L) in wet deposition measured between January 2009 and December 2011 at the Hoh Ranger Station in Olympic National Park, Washington (part of the National Atmospheric

Deposition Program/National Trends Network (NADP/NTN)). Table 11.3 displays the data from the first eight and last 2 months. There are 56 observed values and 46 censored values, with four different detection limits, although only two detection limits occur in the data shown in the table.

Table 11.2: Copper concentrations in shallow groundwater in two geological zones (Millard and Deverel, 1988)

Zone	**Copper (μg/L)**										
Alluvial Fan	<1	<1	<1	<1	1	1	1	1	1	2	2
	2	2	2	2	2	2	2	2	2	2	2
	2	2	2	2	2	2	2	2	3	3	3
	3	3	3	4	4	4	<5	<5	<5	<5	<5
	<5	<5	<5	5	5	5	7	7	7	8	9
	<10	<10	<10	10	11	12	16	<20	<20	20	
Basin Trough	<1	<1	1	1	1	1	1	1	1	<2	<2
	2	2	2	2	3	3	3	3	3	3	3
	3	4	4	4	4	4	<5	<5	<5	<5	<5
	5	6	6	8	9	9	<10	<10	<10	<10	12
	14	<15	15	17	23						

Table 11.3: NH_4 ion concentration (mg/l) in the National Atmospheric Deposition Program/National Trends Network (NADP/NTN) wet deposition collector at Hoh Ranger Station, Olympic National Park, Washington. Data only shown for the first eight and last two months.

	Week 1	Week 2	Week 3	Week 4	Week 5
Month 1		<0.006	<0.006		
Month 2	0.006		0.016	<0.006	
Month 3	0.015	0.023	0.034	0.022	
Month 4	0.007	0.021	0.012	<0.006	
Month 5	0.021	0.015	0.088	0.058	
Month 6	<0.006	<0.006			
Month 7	<0.006	<0.006	0.074		
Month 8	0.011	0.121	<0.006		
⋮	⋮	⋮	⋮	⋮	⋮
Month 35	0.036	<0.008	0.012	0.03	0.022
Month 36	0.008				

11.3 Graphical Assessment of Censored Data

In Chapter **??** we discussed graphics for a single variable, including histograms, empirical cdf plots, and probability (Q-Q) plots. All these graphics can be produced when you have censored data, but you have to account for the censoring. This is especially important if you have multiply censored data. In this section we show how to create quantile plots and probability plots, as well as determine "optimal" Box-Cox transformations. We also talk briefly about plotting histograms.

11.3.1 Empirical CDF Plots for Censored Data

An empirical cumulative distribution function plot or empirical cdf plot plots the ordered data on the x-axis against the estimated cumulative probabilities on the y-axis. The formula to compute the estimated cdf with

completely observed data was given in Equation (**??**).

When you have Type I left-censored data with one censoring level, T, that is less than or equal to any uncensored value, the empirical cdf is easy to calculate because all the observations can be ordered. All the censored values are considered equal, but they are all less than the smallest observed value. If there is an uncensored value at a censoring limit, e.g., an uncensored value of 5 and censored value of <5, the uncensored value is considered larger. The general empirical cdf equation (**??**) can be applied to the uncensored values without any modification. The estimated cdf, $\hat{F}(x)$ is undefined for $x < T$, equal to $\#censoredvalues/n$ at T, and calculated using Equation **??**, for $x > T$.

When the data set includes observations censored at some value larger than the smallest observed value, Equation **??** must be modified. For right-censored data, the standard estimator of the survival function, $\hat{S}(t) = 1 - \hat{F}(t)$ is the Kaplan-Meier (KM) estimator (Kaplan and Meier, 1958; Hosmer, et al. 2008, pp. 17-26). This can be adapted to left-censored data.

The KM estimator uses the concept of "# known at risk". For left censored data, the "# known at risk" for any given Y is the number of observations known to have values $\leq Y$. Consider eight sorted observations: 3, <5, <5, 5, 6, 6, <10, 12. The "# known at risk" at Y=12 is 8 because all 8 values are known to be less than or equal to 12. The "# known at risk" at Y=7 is 6 because there four observed values and two censored values known to be less than 7. The value of <10 may be less than 7 but it cannot be counted as known to be less than 7.

We need notation that describes the uncensored values and the censoring pattern. Define J as the number of unique uncensored values in the data set. The N=8 observations above have $J = 4$ unique uncensored values. Define Z_j, $j = 1, \ldots J$ as those unique values. Their indices, $j = 1, \ldots, J$, are assigned in *reverse* order, i.e. from largest to smallest. For the observations above, $Z_1 = 12$, $Z_2 = 6$, $Z_3 = 5$, and $Z_4 = 3$. Define R_j as the number of observations known at risk at Z_j. Define n_j as the number of uncensored values at Z_j. (If there are no tied uncensored values, $n_j = 1$ for all j.) For the observations above, the values of R_j and n_j are given in Table 11.4.

j	Z_j	R_j	n_j	$\frac{R_j - n_j}{R_j}$	$\hat{F}(Z_j)$
1	12	8	1	7/8	1.0
2	6	6	2	4/6	1.0*(7/8) = 0.875
3	5	4	1	3/4	0.875*(4/6) = 0.5833
4	3	1	1	0/1	0.583*(3/4) = 0.4375

Table 11.4: Example of Kaplan-Meier notation and calculations.

To compute the KM estimator of $F(x)$, the observations are sorted in decreasing order and R_j and n_j are computed for each unique uncensored value. $\hat{F}(x)$ is set to 1 for $X \geq Z_1$. Here, $\hat{F}(x) = 1$ for $X \geq 12$. $\hat{F}(x)$ for X between Z_1 and Z_2 is the product of $\hat{F}(Z_1)$ and $(R_1 - n_1)/R_1$. $\hat{F}(x)$ for X between Z_2 and Z_3 is the product of $\hat{F}(Z_2)$ and $(R_2 - n_2)/R_2$. $\hat{F}(x)$ has jumps at each Z_j. The estimate for any Z_j is given by:

$$\hat{F}(Z_j) = \begin{cases} 1 & j = 1 \\ \prod_{i=1}^{j} \frac{R_i - n_i}{R_i} & 1 < j \leq J \end{cases} \tag{11.1}$$

When all observations are uncensored, the KM estimator reduces to the empirical probabilities estimator shown in equation (**??**) for $\hat{F}(x)$. That is because the numerator of one ratio in the product cancels the denominator of another. This is illustrated in table 11.5, using a data set like the one used to construct table 11.4 except that all the previously censored observations are now uncensored. $\hat{F}(5)$ is the product of 7/8, 6/7, and 4/6. The estimates of $\hat{F}(3)$, $\hat{F}(5)$, and $\hat{F}(6)$ for the data with censored values are larger than those for the uncensored data. This is because these values of X are below one or more censored values. The probability mass associated with the censored value at $X = 10$ has been "redistributed" to the left onto each of the smaller uncensored values (3, 5, and 6). The probability mass associated with the two censored values

at $X = 5$ is "redistributed" to the left onto X=3. This is the left-censored equivalent of the "redistribute to the right" interpretation of the KM estimator for right-censored data (Efron 1967)

The nature of this "redistribution" can be seen by comparing terms in Equation (11.1) shown in tables 11.4 and 11.5. If you combine the step from $X = 10$ to $X = 6$ with the step from $X = 6$ to $X = 5$, $\hat{F}(5)$ from the "all uncensored" data is the product of 7/8 and 4/7. For the data set with censored values, $\hat{F}(5)$ is the product of 7/8 and 4/6. The second term in the product for the censored estimate is larger ($4/6 = 0.6667$) than that in the product for the uncensored estimate ($4/7 = 0.571$). The increase reflects the redistribution of probability mass from $X < 10$ onto the observed value at $X = 5$.

j	Z_j	R_j	n_j	$\frac{R_j - n_j}{R_j}$	$\hat{F}(Z_j)$
1	12	8	1	(7/8)	1.0
2	10	7	1	(6/7)	(7/8) = 0.875
3	6	6	2	4/6	(6/7)*(7/8) = 6/8 = 0.75
4	5	4	3	1/4	(6/8)*(4/6) = 4/8 = 0.5
5	3	1	1	0/1	(4/8)*(1/4) = 1/8 = 0.125

Table 11.5: Example of Kaplan-Meier notation and calculations for 8 uncensored values.

Example 1. Empirical CDF Plots of the NH_4 wet deposition data Data

The Olympic ammonium data (part of which is shown in Table 11.3) are skewed to the right. All but two values are less than 0.09 mg/L while the two largest values are 0.12 mg/L and 0.25 mg/L. Figure 11.1 displays the Kaplan-Meier estimate of the empirical cdf plot for the log-transformed observations. The plot was produced using:

```
with(Olympic.NH4.df, ecdfPlotCensored(log(NH4.mg.per.L), Censored,
  prob.method = "kaplan-meier", type = "s", include.cen = TRUE,
  xlab = expression(paste("log [ ", NH[4], " (mg/L) ]")),
  main = ""))
```

The upside-down triangles indicate the censoring levels of censored observations. They are included on the plot when `include.cen=TRUE`. The `type="s"` argument plots the data as a step function.

Figure 11.2 overlays the CDF of a fitted normal distribution on top of the empirical CDF of the log-transformed NH_4 data. This plot was produced using:

```
with(Olympic.NH4.df, cdfCompareCensored(log(NH4.mg.per.L), Censored,
  prob.method = "kaplan-meier", type = "s",
  xlab = expression(paste("log [ ", NH[4], " (mg/L) ]")), main = ""))
```

This plot suggests that a lognormal distribution is an adequate fit to these data. If you wanted to compare the empirical cdf based on the original, untransformed observations to a fitted lognormal distribution, the EnvStatscode is:

```
with(Olympic.NH4.df, cdfCompareCensored(NH4.mg.per.L, Censored,
  dist = "lnorm", prob.method = "kaplan-meier", type = "s",
  xlab = expression(paste(NH[4], " (mg/L)")), main = ""))
```

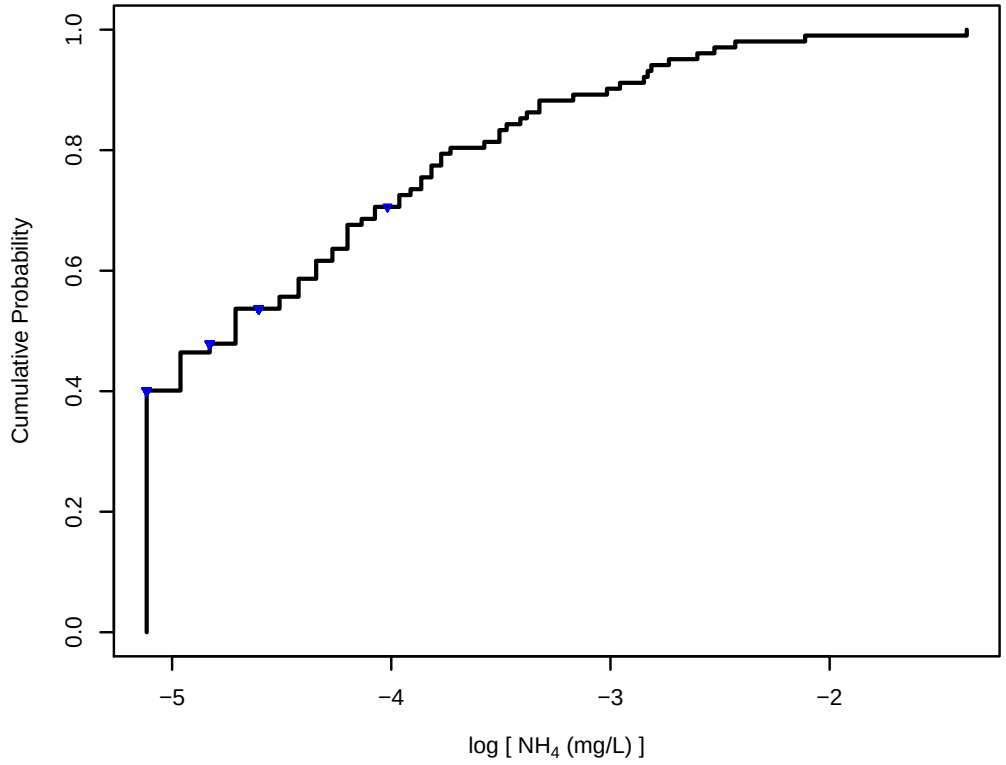


Figure 11.1: Empirical cdf plot of the log-transformed NH_4 data of Table 11.3

Example 2. Comparing the Empirical CDF of Copper between Two Geological Zones

Figure 11.3 compares the empirical cdf of copper concentrations from the alluvial fan zone with those from the basin trough zone using the data shown in Table 11.2. The code to produce this plot is:

```
with(Millard.Deverel.88.df, cdfCompareCensored(
  x = Cu[Zone == "Alluvial.Fan"],
  censored = Cu.censored[Zone == "Alluvial.Fan"],
  y = Cu[Zone == "Basin.Trough"],
  y.censored = Cu.censored[Zone == "Basin.Trough"],
  prob.method = "kaplan-meier", type = "s",
  xlab = expression(paste("Order Statistics for Cu in ",
  "Alluvial Fan and Basin Trough (", mu, "g/L)")), main=""))
```

This plot shows that the two distributions are fairly similar in shape and location.

11.3.2 Q-Q Plots for Censored Data

As explained in Section **??**, a probability or quantile-quantile (Q-Q) plot plots the ordered data (the empirical quantiles) on the y-axis vs. the corresponding expected quantiles from the assumed theoretical probability distribution on the x-axis, where the expected quantiles are computed from plotting positions. The same principle applies to censored data. The only changes are that plotting positions are only defined for observed

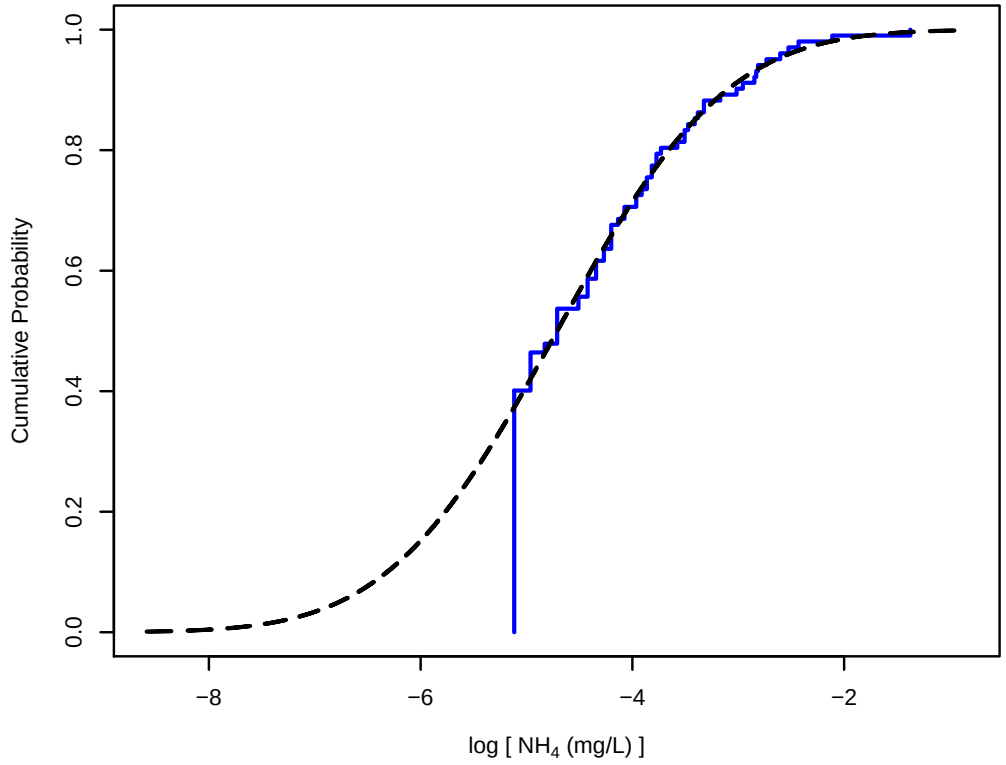


Figure 11.2: Empirical cdf plot of the log-transformed NH_4 data with a fitted normal distribution.

values and they need to be calculated in a way that accounts for the censoring.

Plotting positions are very easy to calculate when there is only one censoring level, T, that is less than or equal to any uncensored value, because all observations can be ordered. In this simple, but common, situation, all the censored values are considered equal, and they are all less than the smallest observed value. If there is an uncensored value at a censoring limit, e.g. an uncensored value of 5 and censored value of <5, the uncensored value is considered larger. For the observed values, i.e. $Y_{(i)} \geq T$, the plotting position is calculated using the usual formula:

$$\hat{p}_i = \frac{i - a}{N - 2a + 1}. \tag{11.2}$$

The plotting positions, $\hat{p}(i)$, are undefined for the censored values. Recommended choices for the constant a depend on the assumed distribution for the population (Table **??**).

When the uncensored and censored values are intermingled, that is the data set has one or more uncensored observations with values less than a censoring level, then the computation of the plotting positions has to account for those censored values. Both Michael and Schucany (1986) and Hirsch and Stedinger (1987) have developed methods for this situation.

Michael and Schucany's (1986) method generalizes the Kaplan-Meier estimate of the cdf to include the constant a. To calculate plotting positions, the observations are sorted by their values (uncensored or censoring limit). If a censoring limit is the same as an uncensored value, the uncensored value is considered larger. The set Ω is the set of indices corresponding to uncensored values. The Michael and Schucany (1986) plotting position formula is

$$\hat{p}_i = \frac{N - a + 1}{N - 2a + 1} \prod_{j \subset \Omega, j \geq i} \frac{j - a}{j - a + 1} \tag{11.3}$$

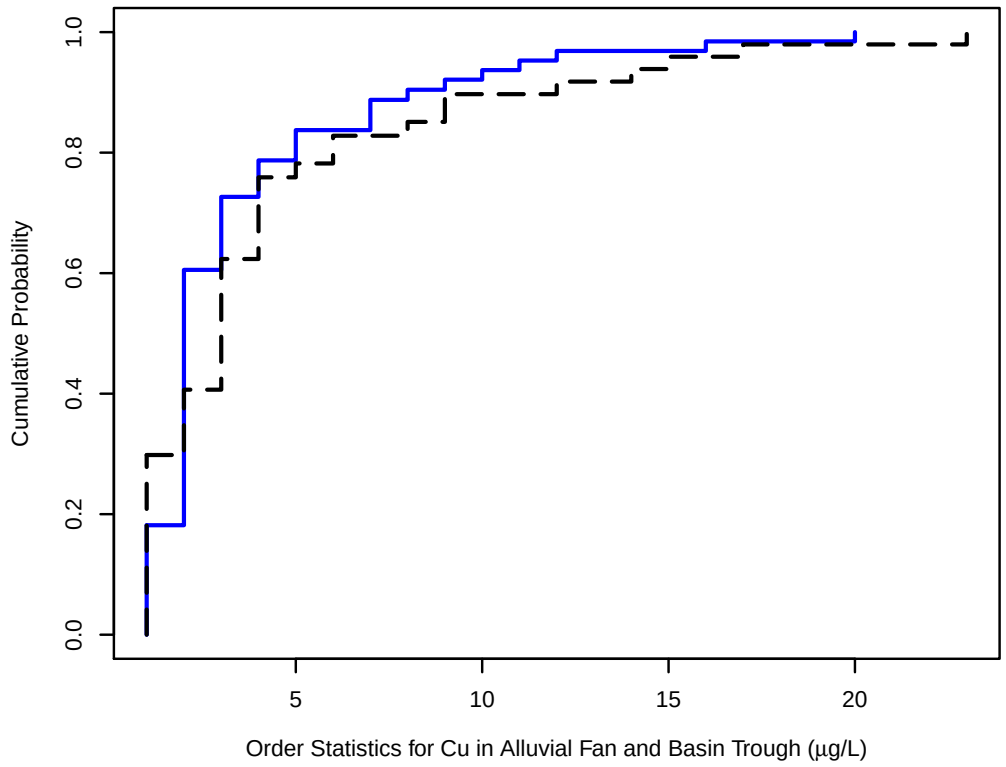


Figure 11.3: Empirical cdf's of copper concentrations in the alluvial fan (solid line) and basin trough (dashed line) zones

To illustrate the use of equation (11.3), consider a small data set with five values of <4, <4, 5, <14, 15. The set Ω is (3,5). The plotting position for the uncensored value of 5 ($i = 3$) using $a = 3/8$ is calculated as

$$\left(\frac{5-3/8+1}{5-6/8+1}\right)\left(\frac{3-3/8}{3-3/8+1}\right)\left(\frac{5-3/8}{5-3/8+1}\right) \approx 0.64$$

Table 11.3.2 compares plotting positions from a hypothetical version of this data in which all values were reported as measured to the plotting positions when the data set is multiply censored. Note that the plotting position for the value of 15 is the same for the censored and uncensored data sets. This is true in general for values above the largest censored value or when there are no censored values. The plotting positions for observed values below one or more censored values, e.g. for 5, are adjusted to account for the censoring.

Table 11.6: Michael-Schucany (M-S) and Hirsch-Stedinger (H-S) plotting positions for a data set with multiply censored observations, compared to Blom plotting positions for uncensored values.

Uncensored	Blom	Censored	Plotting Posision	
Data Set	Plotting Position	Data Set	M-S	H-S
1	0.12	<4		
2	0.31	<4		
5	0.5	5	0.64	0.67
10	0.69	<14		
15	0.88	15	0.88	0.90

Hirsch and Stedinger (1987) derive a different estimator of plotting positions for uncensored observations. Their starting point is the observation that a modified survival function, $S^*(x)$, defined as $P[X \geq x]$, can be estimated at each censoring limit even with multiple censoring values. Define T_j as the j'th sorted censoring limit, i.e. $T_1 < T_2 < \ldots < T_K$, c_j as the number of values with the j'th censoring limit, and K as the number of distinct censoring limits. Then,

$$\begin{aligned} \hat{S}^*(T_j) &= \hat{P}[X \geq T_j] \\ &= \hat{P}[X \geq T_{j+1}] + \hat{P}[T_j \leq X < T_{j+1}] \\ &= \hat{S}^*(T_{j+1}) + \hat{P}[T_j \leq X < T_{j+1} \mid X < T_{j+1}]\, \hat{P}[X < T_{j+1}] \\ &= \hat{S}^*(T_{j+1}) + \hat{P}[T_j \leq X < T_{j+1} \mid X < T_{j+1}]\, (1 - \hat{S}^*(T_{j+1})) \end{aligned} \tag{11.4}$$

The conditional probability in Equation (11.4) can estimated by

$$\hat{P}[T_j \leq X < T_{j+1} \mid X < T_{j+1}] = \frac{A_j}{A_j + B_j},$$

where A_j is the number of uncensored observations in the interval $[T_j, T_{j+1})$ and B_j is the total number of observations $< T_j$. Substituting this estimate into (11.4) gives

$$\hat{S}^*(T_j) = \hat{S}^*(T_{j+1}) + \left(\frac{A_j}{A_j + B_j}\right)\left[1 - \hat{S}^*(T_j)\right]. \tag{11.5}$$

This can be solved iteratively for $j = K, K-1, \ldots 0$. Note that $\hat{S}^*(T_{K+1})$ is defined to be 0 and $\hat{S}^*(T_0)$ is defined to be 1. The plotting positions for uncensored observations are calculated by linear interpolation of $\hat{S}^*$ between the bracketing censoring limits. The A_j uncensored observations in the interval $[T_j, T_{j+1})$ are ranked $r = 1, 2, \ldots A_j$ within the interval. These observations are also indexed by i, their rank within the full data set. The plotting position for the i'th uncensored value is

$$\hat{p}_i = \left[1 - \hat{S}^*(T_j)\right] + \left[\hat{S}^*(T_j) - \hat{S}^*(T_{j+1})\right] \frac{r - a}{A_j - 2a + 1}, \tag{11.6}$$

where the constant a is chosen appropriate for the presumed distribution, i.e. $a = 3/8$ for a normal or lognormal distribution. However, Helsel and Cohn (1988, p. 2001) found very little effect of changing the value of a on the Hirsch-Stedinger plotting positions (11.6).

Table 11.3.2 gives the Hirsh-Stedinger plotting positions for the five observations in Table 11.3.2 using $a = 3/8$. The Hirsch-Stedinger plotting positions are very close to the Michael-Schucany even though the example data set is small with a lot of censoring. In general, there is little effect of the choice of plotting position formula.

Either the Michael-Schucany (11.3) or Hirsh-Stedinger (11.6) formulae can be used with the choice of a that is appropriate for the presumed distribution (Table **??**). For example, a would be set at $a = 3/8$ if the presumed distribution is normal or lognormal.

Example 3. Comparing the Multiply Censored NH_4 Data to a Lognormal Distribution

Figure 11.4 shows the normal Q-Q plot for the log-transformed NH4 deposition concentration data shown in Table 11.3. As in the case of the empirical cdf plot shown in Figure 11.2, this plot indicates that the lognormal distribution provides an adequate fit to these data. This plot was produced using:

```
with(Olympic.NH4.df, qqPlotCensored(NH4.mg.per.L, Censored,
  distribution = "lnorm", add.line = TRUE, main = ""))
```

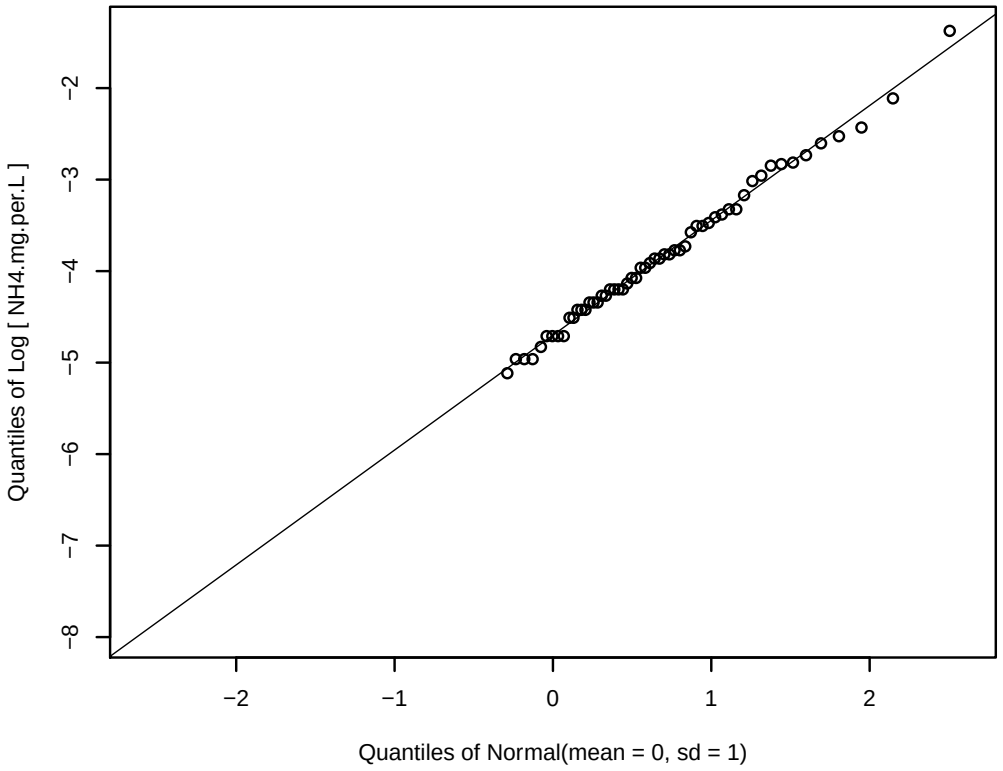


Figure 11.4: Normal Q-Q plot for the log-transformed multiply censored NH4 data of Table 11.3

11.3.3 Box-Cox Transformations for Censored Data

In Chapter **??** we discussed using Box-Cox transformations as a way to satisfy the normality assumption for standard statistical tests, and also sometimes to satisfy the linear and/or the equal variance assumptions for a standard linear regression model (see Equations (**??**) and (**??**)). We also discussed three possible criteria to use to decide on the power of the transformation: the probability plot correlation coefficient (PPCC), the Shapiro-Wilk goodness-of-fit test statistic, and the log-likelihood function. All three can be extended to the case of singly and multiply censored data (e.g., Shumway et al., 1989). For example, the PPCC is the correlation between the observed values and the expected order statistics for the assumed distribution. This is adapted to censored data by computing the Hirsh-Stedinger or Michael-Schucany plotting positions using Equations (11.6) or (11.3), using the plotting positions to compute the expected order scores for the observed values, then calculating the correlation between the expected order scores and observed values. The Shapiro-Wilk approach suggested by Shumway (1989) maximizes a variation of the Shapiro-Wilk statistic using plotting positions that account for the censoring. The log-likelihood approach maximizes the log-likelihood for the transformed data. The log-likelihood function includes a term for the Jacobian of the transformation.

Example 4. Determining the "Optimal" Transformation for the Skagit river NH_3 Data

The `boxcoxCensored()` function will estimate the Box-Cox transformation from data with single or multiple censoring limits. Figure 11.5 displays a plot of the probability plot correlation coefficient vs. various values of the transform power l for the singly censored NH_3 data shown in Table 11.1. For these data, the log likelihood reaches its maximum at about $l = 0$. This plot was produced using:

```
index <- with(Skagit.NH3_N.df, Date >= "1978-01-01" &
  Date <= "1982-12-31")
nh3.BClist <- with(Skagit.NH3_N.df, boxcoxCensored(
  NH3_N.mg.per.L[index], Censored[index],
  objective = "Log-Likelihood", lambda = seq(-1, 1, by = 0.05)))
plot(nh3.BClist, main = "")
```

The optimum value can be found by:

```
nh3opt <- with(Skagit.NH3_N.df, boxcoxCensored(
  NH3_N.mg.per.L[index], Censored[index],
  objective = "Log-Likelihood", optimize = TRUE))
```

Based on the maximum log likelihood, the best choice of l is -0.29, i.e. a log transformation. Because of the large data set, this estimate is reasonably precise. A 95% confidence interval using profile likelihood is approximately (-???, ???). Another estimator of the transformation using maximum PPCC gives a similar result: -0.26.

11.4 Estimating Distribution Parameters

In Chapter **??** we discussed the maximum likelihood (MLE) estimators of distribution parameters. These can easily be extended to account for censored observations (e.g., Cohen, 1991; Schneider, 1986). We give the details in section 11.4.3. But, MLE's assume that the data are a sample from a specific probability distribution and may not be robust to misspecification of that distribution. Recent research in environmental

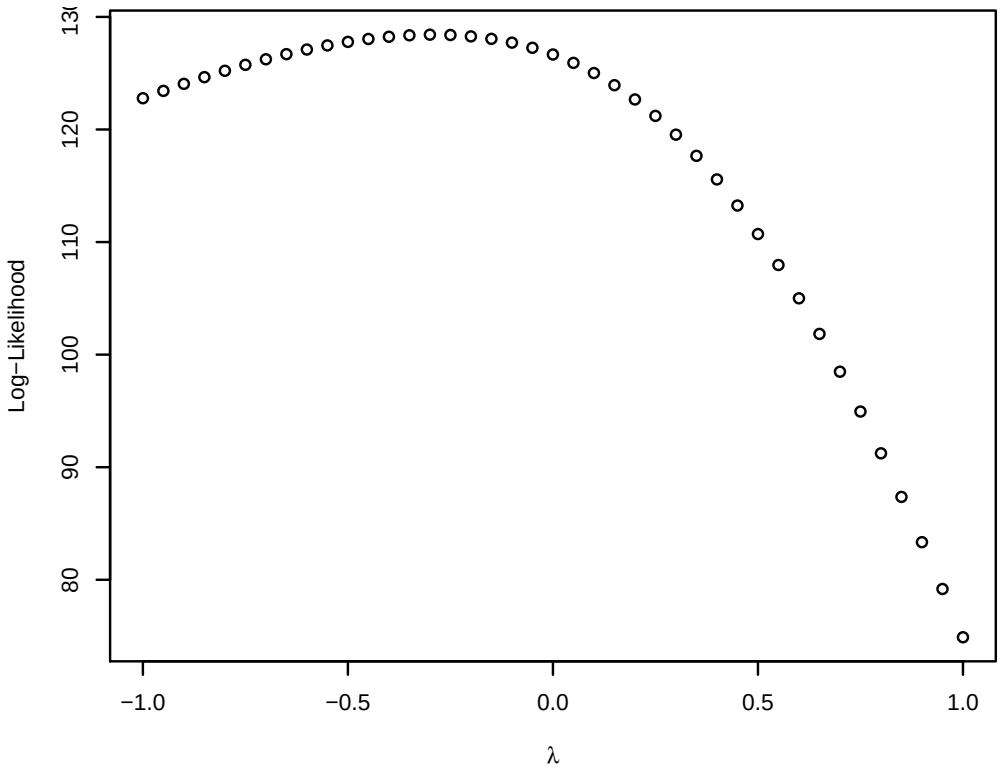


Figure 11.5: Probability plot correlation coefficient vs. Box-Cox transform power (l) for the singly censored NH_3 data of Table 11.1

statistics has focused on finding estimators that are more robust to mis-specification. Two such estimators the Robust Order Statistics estimator, discussed in section 11.4.3 and estimators based on Kaplan-Meier estimate of $\hat{F}(x)$, discussed in section 11.4.3.

In this section, we will discuss MLE's and the two robust estimators for both the normal and lognormal distributions and briefly mention other methods. We discuss both estimation and construction of prediction and confidence intervals. The discussion of constructing confidence intervals for the mean focuses on data with multiple censoring limits, but we also discuss the singly censored case.

11.4.1 Notation

Before discussing various methods, we need to introduce some notation. This notation is consistent with the notation used earlier in the chapter. It is collected here for convenience. We will start with the left multiply censored case, and then discuss the singly censored case.

Notation for Left Censored Data

Let x_1, x_2, $\cdots$, x_N denote N observations from some distribution. Of these N observations, assume n ($0 \geq n \geq N$) of them are recorded (uncensored) values and c of them are left censored at fixed ordered censoring levels, T_j, where $T_1 < T_2 < \cdots < T_k$ where $k \geq 2$ (for the multiply censored case). Let c_j denote the number of observations censored at T_j ($j = 1, 2, \cdots, k$). Hence, $c = \sum_{j=1}^{k} c_j$, the total number of censored observations. Let $x_{(1)}$, $x_{(2)}$, $\cdots$, $x_{(N)}$ denote the ordered "values", where now "value" is either

the uncensored value or the censoring level for a censored observation. If a censored observation has the same value as an uncensored one, the censored observation is placed first. Finally, let Ω denote the set of n subscripts in the ordered sample that correspond to uncensored observations.

When there is a single censoring limit that is less than or equal to any observed data value, e.g., observations of 10, <5, 15, 7, and 15 ppm, the notation above can be simplified. Define c as the number of observations left-censored at the fixed censoring level T. Using the notation for "ordered" observations, $x_{(1)} = x_{(2)} = \cdots = x_{(c)} = T$. For the five observations above, $n = 4$, $c = 1$, $x_{(1)} = 5$, $x_{(2)} = 5$, $x_{(3)} = 7$, $x_{(4)} = 10$, $x_{(5)} = 15$, and $\Omega = \{2, 3, 4, 5\}$.

11.4.2 Substituting a Constant for Censored Observations

Before we discuss the recommended methods for estimating distribution parameters from data with censored observations, it is important to mention a method that **should not be used**. That is substituting a constant for a censored value. Often the constant is one-half of the detection limit, but 0 and the detection limit have also been suggested. After substitution, the mean and standard deviation are estimated by the usual formulae (El-Shaarawi and Esterby, 1992; Helsel, 2006).

Substitution is simple abd frequently used. But, it often has terrible statistical properties. El-Shaarawi and Esterby (1992) show that substitution estimators are biased and inconsistent (i.e., the bias remains even as the sample size increases). To quote Helsel (2012, p xix),

> In general, do not use substitution. Journals should consider it a flawed method compared to the others that are available, and reject papers that use it. ... Substitution is fabrication. It may be simple and cheap, but its results can be noxious.

11.4.3 Normal Distribution

The recommended stimators for Type I left-censored data assuming a normal distribution can be grouped into three categories: likelihood, order statistics, and imputation (Helsel 2012). The likelihood estimators are based on maximizing the likelihood function for a specific distribution. Order statistics estimators are based on a linear regression fitted to data in a probability plot. Imputation estimators replace the censored values with new "uncensored" observations derived from the original data. There are many variations within each group.

Maximum Likelihood Estimators

For left censored data, the log likelihood for independent observations from a distribution with parameter vector $\boldsymbol{\theta}$ is given by:

$$\log L\left(\boldsymbol{\theta} \mid \{\boldsymbol{x}\}\right) = \log\left(\begin{array}{c} N \\ c_1\, c_2 \cdots c_k\, n \end{array}\right) + \sum_{j=1}^{k} c_j \log F\left(T_j\right) + \sum_{i \in \Omega} \log f\left(x_{(i)}\right), \tag{11.7}$$

where

$$\left(\begin{array}{c} N \\ c_1\, c_2 \cdots c_k\, n \end{array}\right) = \frac{N!}{c_1!\; c_2! \cdots c_k!\; n!} \tag{11.8}$$

denotes the multinomial coefficient, and $f()$ and $F()$ denote the population pdf and cdf, as in Equations (**??** and **??**). The log likelihood is the sum of three parts: a term based on the number of censored and

uncensored values, a part based on the probability of a censored observation being less than T_j, and a part based on the log likelihood for the uncensored observations. Because the first term does not depend on the parameters, it is often omitted from the log likelihood.

In the case of a normal distribution, the parameter vector $\boldsymbol{\theta}$ is (μ, σ) or perhaps $\left(\mu, \sigma^2\right)$, the pdf is

$$f(t) = \phi\left(\frac{t-\mu}{\sigma}\right), \tag{11.9}$$

and the cdf is

$$F(t) = \Phi\left(\frac{t-\mu}{\sigma}\right). \tag{11.10}$$

Cohen (1963, 1991) shows that the MLEs of μ and σ are the solutions to two simultaneous equations. That solution can be found by numerical maximization of the log likelihood function, Equation 11.7. When there is a single censoring limit, the MLEs can also be found with the aid of tables in Cohen (1991).

If all observations are censored, the MLE is undefined when there is a single censoring limit and defined but extremely poorly estimated when there are multiple censoring limits. The best approach for such data is to report the median, calculated as the median of the censoring levels (Helsel 2012, p. 143-4).

Variations on the theme

The maximum likelihood estimates of μ and σ are biased when the data include censored values. The bias tends to 0 as the sample size increases; Schmee et al. (1985) found that the bias is negligible if N is at least 100, even with 90% censoring. For less intense censoring, fewer observations are needed for negligible bias. Approximate bias corrections have been developed (Saw 1961b, Schneider 1986, pp. 107–110) but are rarely used. In general, correcting for bias increases the variance of the estimate and you can no longer construct confidence intervals using the profile likelihood method.

The MLE does not directly include the known values of the censoring limits. If there is only one censoring limit, the proportion of censored observations estimates the population probability that an random value is less than the censoring limit. If these two are set equal, the equations defining the MLE's have closed form solutions, which are called the Restricted Maximum Likelihood Estimates (Perrson and Rootzen 1977). It is not clear how to extend this to multiple censoring limits. Also, the proportion of censored observations is a random variable with a large variance when N is small. It seems inappropriate to condition on that proportion.

The MLE can be very sensitive to outliers. One option to reduce the influence of outliers is to downweight the influence of extreme values (Singh and Nocerino 2002). Such an approach would be appropriate if the outliers are a result of recording errors, so those observations should be downweighted or ignored when estimating the mean. When apparent outliers represent large values that are part of the population of interest, they should be given full weight.

Robust Order Statistics Estimator

Maximum likelihood is a fully parametric method; all inference is based on the assumed distribution. Robust Order Statistics (ROS) is a semi-parametric method in the sense that part of the inference is based on an assumed distribution and part of the inference is based on the observed values, without any distributional assumption. Hence, ROS estimators are more robust to misspecification of the distribution than are MLE's.

The key insight behind the ROS method is that estimating the mean and standard deviation is simple if we could impute the value of each censored observation. Those imputed values are treated just like uncensored

values. As an aside, it seems very silly to us as statisticians to be making up values that an analytic chemist has measured but not reported.

The ROS method uses Quantile-Quantile Regression to impute values for each censored observation. In the case of no censoring, it is well known (e.g., Nelson, 1982, p. 113; Cleveland, 1993, p. 31) the mean and standard deviation of a normal distribution can be estimated by fitting a least-squares regression line to the values in the standard normal Q-Q plot (Section **??**). The intercept of the regression line estimates the mean; the slope estimates the standard deviation. Given the ordered values, $x_{(i)}$, and their plotting positions, p_i, the regression model is:

$$x_{(i)} = \mu + \sigma\, \Phi^{-1}(p_i) + \varepsilon_i, \tag{11.11}$$

where Φ is the cdf of the standard normal distribution. The Blom plotting positions, Equation (**??**) are the most frequently used when the data are assumed to come from a normal distribution.

When the data include censored values, plotting positions are defined for both censored and uncensored observations based on the total sample size, N. When the data has a single censoring limit that is smaller than any uncensored value, the c censored values are assigned indices $i = 1, 2, , \cdots, c$ and the n uncensored values are assigned indices $i = c+1, c+2, N$. The values of p_i for each observation are the plotting position corresponding to that index, e.g. the Blom plotting positions given by Equation (**??**). When the data have multiple censoring limits, or some uncensored values are smaller than the censoring limit, plotting positions are computed using either the Michael-Schucany or Hirsch-Stedinger plotting positions described in section 11.3.2.

Equation (11.11) is then fit using the uncensored values and their plotting positions. The imputed value for each censored observation is its predicted value from the regression, computed as:

$$\hat{x}_{(i)} = \hat{\mu}_{qqreg} + \hat{\sigma}_{qqreg}\, \Phi^{-1}\,(p_i) \tag{11.12}$$

for $i = 1, 2, ..., c$. The mean is estimated by the sample average, equation (**??**), of the uncensored values combined with the imputed values. The standard deviation is estimated by the sample standard deviation, equation (**??**).

Variations on the theme The mean and standard deviation can be estimated by $\hat{\mu}$ and $\hat{\sigma}$ for the quantile regression, Equation (11.11). These are sometimes called the Quantile Regression estimators. However, they lack the robustness to misspecification of the distribution because the estimators are fully parametric.

Imputation can also be done using the MLE's of μ and σ in Equation (11.12). This is the Modified Maximum Likelihood Method proposed by El-Shaarawi (1989).

When there is a single censoring limit below all uncensored observations, the proportion of censored values also provides some information about the distribution parameters. El-Shaarawi (1989) proposed using this information by including the censoring level, T, and an associated plotting position, p_c with the uncensored observations and their plotting positions. This idea could be extended to multiple censoring limits if information is available about the fraction of points below each detection limit. For example, if the multiple censoring limits arise because of improvements in analytical methods, it may be known that 24 samples were measured with a censoring limit of 5ppm, of which 12 were censored, 36 samples were measured with a censoring limit of 2 ppm, of which 3 were censored, and 30 samples were measured with a censoring limit of 1 ppm, of which 1 was censored. The censoring limits are represented by the points ($\Phi^{-1}(0.5)$, 5), ($\Phi^{-1}(0.0825)$, 2), and ($\Phi^{-1}(0.0333)$, 1).

Kaplan-Meier-based estimator

Estimates of the mean and standard deviation can be computed from an estimated cdf, $\hat{F}(x)$. For censored data, the Kaplan-Meier estimator, described in Section 11.3.1, provides that estimated cdf. The Kaplan-Meier (KM) estimators of the mean and standard deviation are non-parametric and make no assumption about the distribution of the population.

Consider a data set with all uncensored values. For simplicity, assume all values are unique, so each occurs only once. The estimated cdf, $\hat{F}(x)$, is a step function with an increase of $1/N$ at each observed value. When the data are sorted in increasing order, the usual estimators of the mean and variance can be written as functions of the estimated cdf:

$$\begin{aligned} \hat{\mu} = \overline{Y} &= \left(\sum_{i=1}^{N} y_i\right)/N \\ &= \sum_{i=1}^{N} y_i \left(\hat{F}(y_i) - \hat{F}(y_{i-1})\right) \qquad (11.13) \\ \hat{\sigma}^2 = s^2 &= \frac{N}{N-1}\left(\sum_{i=1}^{N}(y_i - \overline{Y})^2\right)/N \\ &= \frac{N}{N-1}\sum_{i=1}^{N}(y_i - \overline{Y})^2 \left(\hat{F}(y_i) - \hat{F}(y_{i-1})\right), \qquad (11.14) \end{aligned}$$

where $\hat{F}(y_0)$ is defined as 0.

When the data include censored values, the same estimators of the mean (Equation 11.13) and standard deviation (Equation 11.14) can be used. The only difference is that Kaplan-Meier estimate of the cdf for data with left-censored values, $\hat{F}(x)$ is used in these equations (Gillespie et al., 2010). Table 11.7 provides the calculations for the data in Table 11.4. The index i indexes the unique uncensored values in increasing order. Table 11.7 illustrates the computations for the 8 observations 3, <5, <5, 5, 6, 6, <10, 12.

x_i	$\hat{F}(x_i)$	$\hat{D}_i = \hat{F}(x_i)$ - $\hat{F}(x_{i-1})$	$A_i = x_i * D_i$	$(x_i - \hat{\mu})^2 * D_i$
3	0.4375	0.4375	1.312	2.298
5	0.5833	0.1458	0.729	0.012
6	0.875	0.2917	1.750	0.146
12	1.0	0.125	1.500	5.625
sum			5.292	9.236

Table 11.7: Calculations for the KM estimates of mean and variance for the data in Table 11.4.

Another KM estimator can be obtained using software for the more typtical right-censored survival data by first "flipping" the data by subtracting all values from a large constant, M (Helsel 2012). Because most implementations of the right-censored KM estimator are designed for values that correspond to lifetimes, M should be larger than the largest data value so all "flipped" observations are positive. The subtraction turns left-censored values into right censored values. The appeal of this is that the KM estimator for data with right-censored values is available in many more software packages than is the estimator for left-censored values. The problem with "flipping" is that "flipped" estimates are biased in small samples. The problem is that the estimated cdf is defined as $\hat{F}(x) = P[X \leq x]$. Because of the flipping followed by estimation, followed by a flip back, the flipped estimates correspond to $P[X < x]$. The probability of a value $= x$ in the sample is put in the wrong place (Gillespie et al., 2010). If N is large, this probability is small, but it is a problem in small samples. The enparCensored() function in EnvStatscorrectly calculates $\hat{F}(x) = P[x \leq X]$.

The data set used in table 11.4 is unusual for environmental data; the smallest value is an uncensored value.

Usually, the smallest value (or values) are censored. In this case, $\hat{F}(x)$ is undefined for values less than the smallest censored value. The redistribute to the right interpretation provides the reason. The probability mass associated with a censored observation is redistributed onto the smaller uncensored values. When the smallest value is censored, there are no smaller uncensored values. The usual solution is to consider the smallest censored value as an uncensored value and assign the appropriate probability to that value of X. This causes no troubles when the objective is to estimate quantiles of the distribution, except perhaps for very small quantiles. However, it does cause problems when the goal is to use the KM method to estimate sample moments. If there is a single censoring limit smaller than all uncensored values, this solution is equivalent to substituting the censoring limit for the censored values and the moment estimators are biased. Hence, the KM approach is most useful for data with multiple censoring limits.

11.4.4 Lognormal Distribution

Section **??** describes two ways to parameterize the lognormal distribution: one based on the parameters associated with the log-transformed random variable and one based on the parameters associated with the original random variable (see Equations (**??**) to (**??**)). If you are interested in characterizing the log-transformed random variable, you can simply take the logarithms of the observations and the censoring levels, treat them like they come from a normal distribution, and use the methods discussed in the previous section. Often, however, we want an estimate and confidence interval for the mean on the original scale. In this section we will discuss various methods for estimating the mean and coefficient of variation of a lognormal distribution when some of the observations are censored. Most of these method combine a method to deal with censoring (Section 11.4.3) and a method to deal with the lognormal distribution (Section **??**).

For convenience, we repeat the notation for quantities associated with lognormal distributions described in Section **??**. The random variable Y has a lognormal distribution when $Y = \log(X) \sim N(\mu, \sigma)$.

symbol	equation	description
μ		mean of log transformed values
σ		s.d. of log transformed values
θ	$= \exp\left(\mu + \sigma^2/2\right)$	mean of X, the untransformed values
τ	$= \sqrt{\exp(\sigma^2) - 1}$	c.v. of X, the untransformed values
σ_X	$= \theta \times \tau$	s.d. of X, the untransformed values

Table 11.8: Notation for lognormal quantities.

Maximum Likelihood Estimators

The mean, θ, coefficient of variation, τ, and standard deviation σ_X, of a lognormal distribution are functions of μ and σ (see Table 11.8). By the invariance property of MLE's, the MLE's of θ, τ, and σ_X are those functions of the MLE's of μ and σ. If some observations are censored, the MLE's of μ and σ are calculated by maximizing the log-likelihood function for censored data, Equation (11.7). The MLE's of θ, τ, and σ_X are obtained by substituting $\hat{\mu}$ and $\hat{\sigma}$ into the equations for θ, τ, and σ_X.

Variations on the theme

The MLE's of μ and σ can be substituted into other estimators of θ. For example, the quasi MVUE is obtained using the equation for the minimum variance unbiased estimator (MVUE) for θ, Equation **??**, and the quasi bias corrected MLE is obtained using the equation for the bias-corrected MLE of θ, Equation **??**. These estimators are biased when used for data with censored values. The MVUE and bias-corrected

MLE are based on the statatistical properties of $\hat{\mu}$ and $\hat{\sigma}$ for uncensored data. When some observations are censored, $\hat{\mu}$ and $\hat{\sigma}$ have different properties.

Robust Order Statistics

The Robust Order Statistics method (section 11.4.3) is a very effective method to estimate parameters for the original scale. The ROS method is used to impute values for the censored observations. Again, we note the silliness (to statisticians) of making up values for quantities that analytical chemists have measured but not reported. The imputed values (on the log scale) are back transformed by exponentiation. The sample mean and standard deviation are calculated for the full data set (observed and imputed values) by the usual formulae (Hashimoto and Trussell, 1983; Gilliom and Helsel, 1986; and El-Shaarawi, 1989).

This method has the very big advantage of being robust to minor departures from a lognormal distribution. Many environmental observations have distributions that are skewed, like a lognormal distribution, but the log transformed values are not exactly normally distributed. In our experience, the distribution of log transformed values is often slightly negatively skewed, i.e. has a longer tail to the left of the mean. The equations for θ and τ depend on the assumption that the log transformed values are normally distributed. When the transformed distribution is slightly negatively skewed, the MLE of θ tends to overestimate the mean. The lognormal ROS estimator uses a distributional assumption only to impute the censored values. The uncensored values are used as reported. Because the censored values tend to be the smaller values, and they are exponentiated before the mean is calculated, errors in the log-scale imputed values (e.g. imputing -5 instead of -4) have little effect on the estimated mean.

Variations on the theme

Kroll and Stedinger (1986) proposed a combination of ROS and likelihood ideas that they called "robust MLE" for log normal data. That is to use maximum likelihood to estimate the mean and standard deviation from log transformed values. Impute log-scale values for each censored observation using the MLE's of the mean and standard deviation and exponentiate those imputed values to get imputations on the data scale. Finally, calculate the mean and standard deviation using the observed and the imputed values.

Another variation is to use a transformation other than the log, e.g., a Box-Cox transformation (Section 11.3.3). Shumway et al. (2002) found that considering three transformations, none, square-root, or log, choosing the most appropriate one using log-likelihood, then using the ROS method improved the performance of the ROS method.

Kaplan-Meier Estimator

Because the KM estimator (Section 11.3.1) is non-parametric, it can be used without modification with lognormal data. Again, the KM method should be avoided in the common case of a single censoring level lower than any observed value, because then it is equivalent to substitution.

Example 5. Estimating the Mean and Standard Deviation of NH_4 Deposition Concentrations

We saw in Example 1 that the NH4 data of Table 11.3 appear to be adequately described by a lognormal distribution. The `elnormCensored()` function in EnvStatscomputes ML and ROS estimates. The KM estimates were computed using the `enparCensored()` function. The data set includes values censored at 0.006, which is below any observed value. The `enparCensored()` function substitutes the half of the detection limit for those censored values. This can be changed using the *** = argument to enparCensored(). When *** = "dl", the detection limit is used; when ***="zero", 0 is used. The EnvStatscode is:

```
attach(Olympic.NH4.df)
NH4.mle <- elnormCensored(NH4.mg.per.L, Censored, method='mle')
NH4.ros <- elnormCensored(NH4.mg.per.L, Censored, method='impute.w.qq.reg')
NH4.km <- enparCensored(NH4.mg.per.L, Censored)
detach()
```

Table 11.9 displays the estimated mean and standard deviation using these three estimators. All three estimators provide similar estimates for the untransformed data, but the KM estimates of the standard deviation are smaller.

The effect of substituting a constant for the smallest censored values is quite marked. We evaluated how sensitive the estimates are to the choice of constant by computing the KM estimates with one-half of the detection limit or zero as the constant (Table 11.9). Estimates on the original scale change a little bit with the choice of constant but estimates on the log scale are quite different. Using one-half the detection limit seems to be the best choice of constant, at least for the NH_4 deposition data.

	Estimated			
Estimation	Log NH4		NH4	
Method	Mean	s.d.	Mean	s.d.
MLE	-4.71	1.25	0.0197	0.0381
ROS	-4.70	1.23	0.0194	0.0376
KM/dl	-4.39	0.86	0.0202	0.0308
KM/h	-4.65	1.11	0.0191	0.0313
KM/z			0.0179	0.0319

Table 11.9: Estimated mean and s.d. for the NH_4 deposition data in Table 11.3 using maximum likelihood (MLE), robust order statistics (ROS) and Kaplan-Meier (KM/h, KM/dl, KM/z) methods. The KM estimates require specifying the value for the censored observations below the smallest observed value. The estimates labeled KM/h use one half of the detection limit. Those labeled KM/dl use the detection limit and those labeled KM/z use zero.

11.5 Confidence Intervals for the Mean

We have emphasized reporting a confidence interval for the mean because that interval describes the location of and the uncertainty in the estimate. The usual formula, Equation **??** in Chapter 6, is not appropriate when some values are censored. Appropriate methods include normal approximations, profile likelihood, and of bootstrapping.

11.5.1 Normal Approximation

The general equation for a $100(1-\alpha)$ confidence interval for a parameter θ using a normal approximation is

$$\left[\hat{\theta} - z_{1-\alpha/2}\hat{\sigma}_{\hat{\theta}},\ \hat{\theta} + z_{1-\alpha/2}\hat{\sigma}_{\hat{\theta}}\right] \tag{11.15}$$

We can use either of these formulas to construct an approximate confidence interval for the mean by simply plugging in the estimated mean and standard error of the mean.

The standard error of the MLE of μ is calculated from the negative inverse of either the Observed or the Fisher Information matrix (Efron and Hinkley, 1978). Expressions for the elements of the Fisher information

matrix are given in Peng (2010). Alternatively, the Hessian matrix, evaluated at the MLE's, is calculated numerically by many optimization routines. The negative inverse of the Hessian matrix is the estimated variance-covariance matrix of the parameter estimates. The variance of the Kaplan-Meier estimator of μ (Helsel 2012, based on Lee and Wang 2003)) is given by:

$$\hat{\sigma}_m^2 = \sum_{i=1}^{m} \frac{A_i^2}{(n-i)(n-i+1)}, \tag{11.16}$$

where m is the number of uncensored values, and A_i is the cumulative area under the estimated cdf, $\hat{F}(x)$. Sometimes, this formula is "adjusted" for estimating the mean, by multiplying by $\frac{n}{n-1}$, where n is the total number of observations.

$$\hat{\sigma}_m^2 = \left(\frac{n}{n-1}\right) \sum_{i=1}^{m} \frac{A_i^2}{(n-i)(n-i+1)},$$

This is analogous to using $N-1$ instead of N as the denominator in a variance calculation. However, the theory underlying the use of $N-1$ when the data have no censored values does not carry over to the censored case.

It is tempting to approximate the standard error of the mean by

$$\hat{\sigma}_{\hat{\mu}} = \frac{\hat{\sigma}}{\sqrt{m}} \tag{11.17}$$

or

$$\hat{\sigma}_{\hat{\mu}} = \frac{\hat{\sigma}}{\sqrt{n}} \tag{11.18}$$

where m is the number of uncensored observations and n is the total number of observations. These are ad-hoc methods of estimating the standard error. By considering the contribution of an observation to the Observed Information, you can show that Equation (11.17) overestimates the se of the MLE of m while Equation (11.18) underestimates the se.

Consider observations from a normal distribution with known variance. The log likelihood, Equation (11.7) is the sum of contributions from the observed values and contributions from the censored values. Hence, the Observed Information, given by the negative of the second derivative of the log likelihood, is a sum of contributions from observed values and contributions from the censored values. The information about the mean from each observed value is $1/\sigma^2$. If there were no censored values, the total information is n/σ^2, so the variance of the mean is the reciprocal, σ^2/n, the usual formula.

The information from a censored value depends on the mean and the detection limit, but, it is always less than that from an observed value (Peng 2010). Figure 11.6 plots the information for observations from a normal distribution with mean of 5 and variance 1. An observed value contributes 1 to the Observed Information. A value of <5, i.e. $< \mu$, contributes approximately 0.637 to the Observed Information. A value of < 3 contributes approximately 0.88 to the Observed Information, almost as much as an observed value, while a value of < 7 contributes approximately 0.11, which is very little. Equation (11.17) assumes that censored values contribute no information, which underestimates the total Observed Information and overestimates the standard error of the mean. Equation (11.18) assumes that censored values contribute as much information as an observed value, which overestimates the total Observed Information and underestimates the standard error of the mean.

The magnitude of the under- or over-estimation is illustrated in Table 11.10. The population is assumed to be normal with mean 5 and variance 1. The data are 50 observed values and 50 values censored at a single detection limit. As the detection limit increases, the Observed Information in the sample of 100 observations decreases and the standard error of the mean increases. The standard error of the mean in the more common situation of jointly estimating the mean and variance is only slightly larger than when assuming the variance is known.

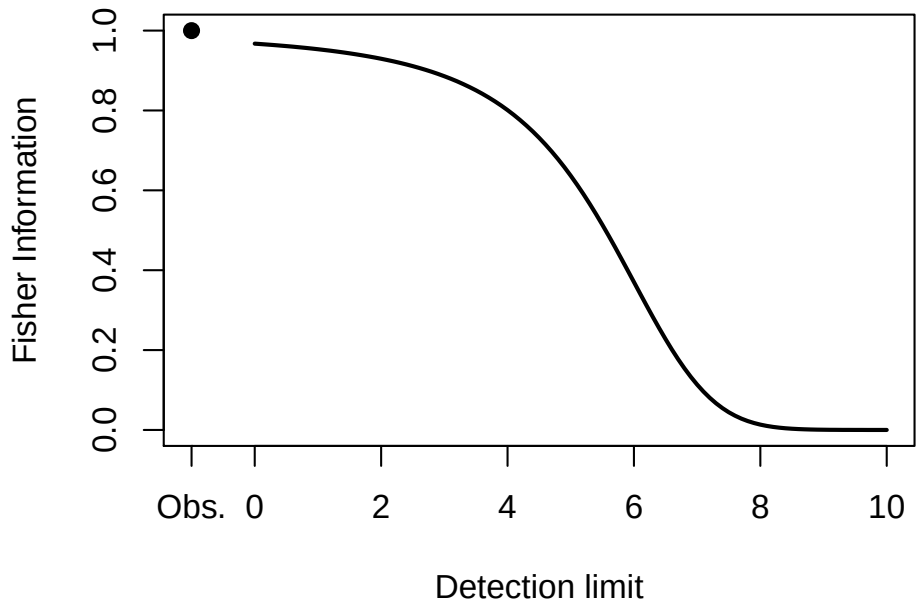


Figure 11.6: Information contributed by a single observed value (Obs.) or a single value censored at a detection limit ranging from 0 to 10. The population is assumed to be normal with mean 5 and variance 1.

It is also tempting to use a T distribution instead of a normal distribution. This is not supported by statistical theory. The derivation of the T distribution is based on independent estimates of the mean and variance. When some observations are censored, the estimated mean and estimated variance are correlated. The magnitude of the correlation depends on the method used to estimate the parameters, the sample size, and the number of censored observations. At best, using a t-distribution to calculate a confidence interval is an ad-hoc method. Even then, you need to choose the degrees of freedom. It is not clear whether this should be the number of observations - 1, or the number of uncensored observations - 1, or something else.

11.5.2 Profile likelihood

Profile likelihood is a general method of constructing a confidence interval when a likelihood function is available. The confidence interval is derived by inverting a likelihood ratio test (Cox and Hinkley 19744). Consider the likelihood function for a one-parameter distribution, e.g. a normal distribution with specified variance but unknown mean, μ. The likelihood ratio test of the simple null hypothesis that $\mu = \mu_0$ is based on the test statistic

$$T = -2 \log \left(\frac{L(\mu_0)}{L(\hat{\mu})} \right),$$

where $\hat{\mu}$ is the maximum likelihood estimator (MLE) of μ. In large samples, T has a χ^2 distribution with 1 degree of freedom (Cox and Hinkley 1974). In small samples, the distribution of T is usually very close to the large sample distribution. Hence, the hypothesis that $\mu = \mu_0$ is rejected at level α if $T > \chi^2_{1,\,1-\alpha}$, where $\chi^2_{1,\,1-\alpha}$ is the $1 - \alpha$ quantile of the Chi-squared distribution with 1 df. The profile $1 - \alpha$ confidence interval for μ is the set of μ_0 for which the hypothesis $\mu = \mu_0$ is not rejected at level α. Equivalently, it is

Table 11.10: Observed Information in a sample of 100 observations censored at values from 5.0 to 8.0. These observations are from a normal population with mean 5 and variance 1. For reference, the observed information in a sample of 100 non-censored observations is 100.

Detection Limit	Observed Information	standard error of the mean when: σ^2 known	σ^2 estimated
5.0	81.83	0.110	0.119
5.5	75.69	0.114	0.115
6.0	68.51	0.120	0.120
6.5	61.37	0.127	0.128
7.0	55.67	0.134	0.134
7.5	52.22	0.138	0.138
8.0	50.66	0.140	0.140

the set of μ_0 for which $T \leq \chi^2_{1,\,1-\alpha}$. When the MLE is estimated numerically, the profile confidence interval is calculated by numerically finding the values of μ_0 for which $T = \chi^2_{1,\,1-\alpha}$. Those two values are the lower and upper bounds of the profile $100(1-\alpha)\%$ confidence interval.

11.5.3 Generalized Pivotal Quantities

A different approach to construct a confidence bound is based on a generalized pivotal quantity (Krishnamorthy and Xu, 2011). A pivotal quantity is a function of parameters and their estimates that does not depend on the unknown parameters. For many problems, the pivotal quantity has a known distribution, e.g., T with a specified degrees of freedom. This distribution is then used to estimate confidence intervals and test hypotheses. The generalized pivotal quantity (GPQ) method uses parametric simulation to estimate the distribution of the pivotal quantity. For the population mean of a normal distributions, the pivotal quantity is $(\mu - \hat{\mu})/\hat{\sigma}$, For type II censored data, this pivotal quantity is free of the population parameters and the method is exact assuming the population distribution is normal. For type I censored data, the pivotal quantity is only approximately free of the population parameters, so the confidence bound is only approximate (*** need ref.).

The GPQ method requires simulating data containing a mix of censored and uncensored values from a specified distribution. With Type I singly censored data, there are two ways to handle the censored observations. One is to allow the number of censored observations to be a random variable, i.e., simulate from an appropriate distribution and record any observation less than the detection limit as censored at that detection limit. This is simple to implement when there is a single detection limit, because all observations share that detection limit. The other approach is to condition on the number of censored values and simulate new observed values from a truncated normal distribution, truncated at the detection limit.

Simulating Type I multiply censored data requires more thought and care. The analog of the first approach for singly censored data requires knowing the number of observations "subject" to each detection limit. For example, imagine that 24 observations were measured by a process that has a DL of 10 and the next 36 were measured by a process that has a DL of 5, and the last 12 were measured by a process with a DL of 1 (a pattern that would arise if equipment were replaced over time by more precise equipment). Data could be randomly generated, then observations censored at the appropriate detection limit for each observation.

However, the number of observations "subject" to each DL is usually not known, especially in cases where the DL varies because of matrix effects or interference from other compounds. In this situation, it is not clear how best to simulate the appropriate mix of censored and uncensored values.

11.5.4 Bootstrapping and Simulation

As explained in section **??**, you can use the bootstrap to construct confidence intervals for any population parameter. In the context of Type I left singly censored data, this method of constructing confidence intervals was studied by Shumway et al. (1989) and Peng (2010). Of the many ways to construct a confidence interval from the bootstrap distribution, the bootstrap-t method seems most promising for inference about the mean from data with censored values (Dixon, unpublished results).

One could also use simulation to determine the critical values for a specified distribution, sample size, and pattern of censoring. For normal distributions with Type II censoring, Schmee et al. (1985) provide tables for exact confidence intervals for sample sizes up to n= 100. These tables should work well for Type I censored data as well (Schmee et al., 1985), but they would apply only for data with a single detection limit below all observed values.

11.6 What analysis method should I choose?

There have been many studies of the performance of various estimators of the mean and standard deviation of data with below-detection limit observations. Helsel (2012, section 6.7) summarizes 15 studies and mentions four more. Our interpretation of those studies leads to recommendations similar to Helsel's:

1. Do not use Kaplan-Meier for data with a single detection limit smaller than the smallest observed value. In this situation, Kaplan-Meier is substitution in disguise.
2. For small-moderate amounts of censoring (e.g. $< 50\%$), use Robust Order Statistics or Kaplan-Meier, if multiple censoring limits.
3. For moderate-large amounts of censoring (e.g. 50% - 80%) and small sample sizes (e.g. < 50), use ROS or robust ML.
4. For very large amounts of censoring (e.g. $> 80\%$), don't try to estimate mean or standard deviation unless you are extremely sure of the appropriate distribution. Then use ML.

There have been few studies of the performance of confidence interval estimators. The performance of a confidence interval estimator depends on the properties of the estimators for *both* the mean and standard deviation, so confidence interval performance can not be assessed from results for the mean and standard deviation individually. Schmee et al. (1985) studied Type II censoring for a normal distribution and found that confidence intervals based on MLEs were too narrow (so the associated hypothesis tests have inflated Type I errors) for small samples with moderate to high censoring. Shumway et al. (1989) studied Type I left singly censored data that were transformed to normality. They considered three methods of constructing confidence intervals from MLE's of the mean and standard deviation of transformed data: the delta method, the bootstrap, and the bias-corrected bootstrap. Based on sample sizes of 10 and 50 with a censoring level at the 10th or 20th percentile, they found that the delta method performed quite well and was superior to the bootstrap method.

11.6.1 A Monte Carlo Simulation Study of Confidence Interval Coverage for Singly Censored Normal Data

Table 11.11 displays the results of a Monte Carlo simulation study of the coverage of two-sided approximate 95% confidence intervals for the mean based on using various estimators of the mean and using the normal approximation method of constructing the confidence interval. For each Monte Carlo trial, a set of observations from a N(3, 1) distribution were generated, and all observations less than the censoring level were

Table 11.11: Results of Monte Carlo simulation study based on Type I left singly censored data from a N(3, 1) parent distribution. Confidence intervals were computed using the normal approximation (z) and either profile likelihood (labeled profile) or generalized pivotal quantities (labeled gpq). For substitution, intervals were computed for two choices of standard error: one using the number of above dl values and the other using the total sample size. Relative bias more than 2% (or less than -2%), relative Root Mean Squared Error values more than 15%, and Coverage values less than 93% or more than 97% are highlighted.

Estimation Method	Sample Size	Percent % Censored	Bias (%)	rmse (%)	Coverage	
					z (%)	profile (%)
MLE	10	20	-0.7	11.6	93.7	94.9
		50	1.1	12.5	95.0	95.9
	20	20	-0.2	7.6	95.6	96.7
		50	-1.0	10.2	**97.3**	95.6
					z (%)	gpq (%)
Impute with Q-Q Regr. (ROS)	10	20	0.8	11.9	**90.5**	**91.7**
		50	**3.5**	14.0	**87.0**	95.5
	20	20	0.6	7.7	**93.9**	94.5
		50	0.4	10.9	**92.6**	94.6
					z (%)	
Impute with MLE (robust MLE)	10	20	0.4	11.5	**92.2**	
		50	**2.3**	12.3	**92.8**	
	20	20	0.4	7.6	94.3	
		50	-0.5	11.9	96.1	
					#>dl	n
Substitute dl/2	10	20	**-7.2**	**15.9**	94.4	**90.1**
		50	**-21.9**	**26.5**	98.2	**87.0**
	20	20	**-6.9**	11.8	96.0	93.2
		50	**-24.7**	**27.5**	**81.7**	**71.7**

censored. If there were less than three uncensored observations, this set of observations was discarded. If none of the observations was censored, the usual method of constructing a confidence interval for the mean was used (Equation (**??**). For each combination of estimation method, sample size, and censoring level, 1,000 trials were performed.

The results in Table 11.11 show that the estimators, other than substitution, behave fairly similarly and fairly well in terms of root mean squared error and confidence interval coverage. Profile likelihood confidence intervals perform slightly better than Normal approximation intervals, especially when the sample size is small and many observations are censored. For estimates from the ROS model, generalized pivotal quantity confidence intervals perform substantially better than those constructed using a normal approximation. The poor performance of Substitution with one-half of the detection limit is obvious.

Example 6. Estimating the Mean and Standard Deviation of Background Well Arsenic Concentrations

Table 11.12 displays various estimates and confidence intervals for the mean of the background well arsenic data shown in Table **??**. For these data, there are $n = 18$ observations, with $c = 9$ of them censored at 5 ppb. The estimates of the mean range from 5.1 to 6.1 ppb, and the estimates of the standard deviation range from 3.3 to 4.0 ppb. The Kaplan-Meier estimate of the mean is the same as the substitution estimate because these data have a single detection limit smaller than any observed values. The KM estimator is not recommended in this situation. The Kaplan-Meier estimate of the standard deviation is slightly smaller than the substitution estimate only because the Kaplan-Meier estimator divides by N while the substitution

estimator divides by $N-1$. (see Section 11.6).

Table 11.12: Estimates of the mean and standard deviation, and confidence intervals for the mean for the background well arsenic data of Table **??**

Estimation Method	**Estimated Mean, Sd**	**CI Method**	**CI**
MLE	5.3, 4.0	Normal Approx. (z)	[2.3, 8.3]
MLE	5.3, 4.0	Profile	[2.1, 7.3]
ROS	6.0, 3.3	Normal Approx. (z)	[3.5, 8.6]
Kaplan-Meier	6.1, 3.7	”	[4.6, 7.6]
Substitute dl/2	6.1, 3.8	”	[2.9, 8.2]

The EnvStatscode is:

11.7 Confidence Intervals for the Mean of Censored Lognormal Data

Constructing a confidence interval for the mean of censored lognormal data combines all the difficulties of estimating confidence intervals for the arithmetic mean of lognormal observations with all the difficulties of estimating confidence intervals for the mean of data with censored values. Three methods that appear to work well are a normal approximation based on Cox's method (El-Shaarawi, 1989), the bootstrap (Efron and Tibshirani, 1993), and profile likelihood (Cox and Hinkley, 1974).

11.7.1 Normal Approximation Based on Cox's Method

This method was proposed by El-Shaarawi (1989) and is an extension of the method derived by Cox and presented in Land (1972) for the case of complete data. The natural logarithm of the arthmetic mean of a logNormal distribution is $\log \hat{\theta} = m + s^2/2$, where m is the log-scale mean and s is the log-scale standard deviation. The standard error of $m + s^2/2$ is approximately:

$$\hat{\sigma}_{\hat{\theta}} = \sqrt{\hat{V}_{11} + 2\hat{\sigma}\hat{V}_{12} + \hat{\sigma}^2\hat{V}_{22}} \tag{11.19}$$

where V denotes the variance-covariance matrix of the MLE's of m and s. The endpoints of a $100(1-\alpha)\%$ confidence interval are then calculated as:

$$\left[\exp(\hat{\theta} + z_{1-\alpha/2}\hat{\sigma}_{\hat{\theta}},\ \exp(\hat{\theta} - z_{1-\alpha/2}\hat{\sigma}_{\hat{\theta}})\right]$$

This method can also be used with the Robust Order Statistics estimator. The sample mean and standard deviation are calculated from the combined collection of log-transformed observations and predictions. The covariance between the estimated mean and estimated standard deviation depends on data and is not easy to calculate. A simple approximate method is to ignore the covariances between predicted values and between predicted values and observations are ignored and consider all values to be independent. The standard error of $\hat{\theta}$ is then:

$$\hat{\sigma}_{\hat{\theta}} = \sqrt{\frac{\hat{\sigma}^2}{n} + \frac{\hat{\sigma}^4}{2(n-1)}} \tag{11.20}$$

because, for independent observations from a normal distribution, the estimated variance of the mean is $\hat{\sigma}^2/n$ and the estimated variance of the variance is $2\hat{\sigma}^4/(n-1)$.

We know of no studies of the robustness of the normal approximation to deviations from strict log-normality. The variance of the variance is very sensitive to the kurtosis (standardized 4th central moment). More generally, the variance of the variance is $\hat{\sigma}^4 \left(\frac{2}{n-1} + \frac{\kappa}{n}\right)$, where κ is the excess kurtosis. This suggests that the normal approximation will be sensitive to moderate deviations from a strictly log-normal distribution.

11.7.2 Profile likelihood

Profile likelihood can be used to construct a $100(1-\alpha)\%$ interval for θ by extending the principle used in subsection 11.5.2 to a function of two parameters. We want to find the set of values of t for which the hypothesis $\theta = t$ is not rejected by a Likelihood Ratio Test. The test statistic is

$$-2\left(\max_{\mu,\,\sigma|\theta=t} \log L(\mu,\sigma) - \max_{\mu,\,\sigma} \log L(\mu,\sigma)\right) \tag{11.21}$$

The first term reflects the constraint that the population mean $\theta = \exp(\mu + \sigma^2/2)$ is equal to t. That first term is equivalent to

$$\max_{\sigma^2} \log L(\log t - \sigma^2/2,\, \sigma^2) \tag{11.22}$$

or

$$\max_{\mu} \log L\left(\mu,\, 2(\log t - \mu)\right). \tag{11.23}$$

Expression (11.22) is more easily maximized numerically because expression (11.23) can result in a negative variance.

Each boundary of a two sided confidence interval is a value of t at which the hypothesis $\theta = t$ is rejected at exactly $\alpha/2$. This can quickly and easily be found numerically by a golden ratio search algorithm or bisection search algorithm (Monahan, 2011, p. 191). Define a test function

$$T(x) = \max_{\mu,\,\sigma} \log L(\mu,\sigma) - \max_{\mu,\,\sigma|\exp(\mu+\sigma^2/2)=x} \log L(\mu,\sigma) - 3.84/2. \tag{11.24}$$

and numerically find x for which $T(x) = 0$. The two values of x for which $T(x) = 0$ are the lower or upper bound of the 95If an interval other than a 95% two-sided interval were desired, 3.84 would be replaced by the appropriate quantile of a χ^2_1 distribution. An efficient algorithm to find the roots of $T(x)$ is to guess at a value, l, that is below the lower confidence interval bound. This can be checked by calculating the value of $T(l)$ for that choice of l and verifying that $T(l) > 0$. The upper end of the search interval is $u = \hat{\theta}$, for which the value of $T(u) = -1.92$. The values l and u now satisfy the requirement that $T(l)$ and $T(u)$ have opposite signs. Since $T(x)$ is a continuous function of x, there must be an x for which $l \leq x \leq u$ and $T(x) = 0$. Both the golden ratio and bisection algorithms repeatedly subdivide the search interval (l, u) until l and u converge on the same value, which is the desired lower bound of the confidence interval. To find the upper bound, a similar process is used with $l = \hat{t}$ and u chosen as some value larger than the upper bound of the confidence interval. This algorithm is implemented in the EnvStatsfunctions `elnormCensored()` to estimate confidence intervals for log scale values and `elnormAltCensored()` to estimate confidence intervals for the arithmetic mean.

11.7.3 Bootstrap

Various types of bootstrap methods can be used for data arising from a lognormal distribution. Peng (2010) considered the bias-corrected and accelerated (BCa) bootstrap and used the boostrap sampling distribution to assess the appropriateness of asymptotic methods based on a normal distribution for the estimated mean. Our experience is that the bootstrap-t intervals seems to have the best coverage, but this requires computing the standard error of the estimated mean for each bootstrap sample. If the distribution is extremely skewed,

e.g. a lognormal distribution with a large variance, the coverage of all bootstrap procedures drops well below the nominal coverage because the sample of observations is unlikely to include the very rare, very large values that have a big influence on the arithmetic mean (Table 5.13 in Millard and Neerchal, 2001).

The number of censored observations will vary from one bootstrap sample to the next, and it is possible to generate a bootstrap sample with too few uncensored values to estimate the mean. How best to proceed if this happens has not been well studied. One approach is to ignore that bootstrap sample, because the original estimate would not have been calculated, were there too few uncensored values. However, this has the potential of overestimating the lower bound of the confidence interval because samples with few uncensored values are more likely to arise from a population with a small mean. The alternative is to condition on the observed number of censored values, but this turns the censoring into type II censoring (fixed number of censored values). This remains a topic for further study.

11.7.4 Other methods to construct confidence intervals

If the sample size is sufficiently large for the sampling distribution of the estimated mean to be close to a normal distribution, Equation (**??**) or (**??**) can be used to construct an appropriate confidence interval. Peng (2010) proposed using the bootstrap sampling distribution to assess whether the sample size is sufficiently large.

If the likelihood function is parameterized in terms of the arithmetic mean, the standard error of the estimated mean is obtained from the negative inverse of the Hessian matrix. However, to obtain an appropriate confidence interval, the sample size needs to be sufficiently large that the distribution of the estimated arithmetic mean is sufficiently close to a normal distribution. If the sample size is large, the difference between quantiles of the normal distribution and quantiles of T distributions is small, so the choice of degrees of freedom is less important. This normal approximation is not recommended when the sample size is small, e.g. n=10 or 20, because the empirical coverage of confidence intervals constructed using Equation (**??**) or (**??**) is much smaller than nominal (Millard and Neerchal, 2001, table 10.16).

Shumway et al. (1989) and Peng (2010) developed a delta-method approximation to the variance of the arithmetic mean, which can be combined with equations (**??**) or (**??**) to construct a confidence interval. Again, when the sample size is small, e.g. n=10 or 20, the empirical coverage of delta-method confidence intervals is much smaller than nominal (Millard and Neerchal, 2001, table 10.16) and this method is not recommended.

11.7.5 Monte Carlo Simulation Study of Confidence Interval Coverage for Censored Lognormal Data

Table 11.13 displays the results of a Monte Carlo simulation study of the coverage of one-sided upper approximate 95% confidence intervals for the mean using various estimators of the mean and various methods of constructing the confidence interval. In this simulation, there was a single detection limit that was set at the 0.2 or the 0.5 quantile of a lognormal distribution.

For each Monte Carlo trial, a set of observations from a lognormal distribution with a log-scale mean of m = 3 and log-scale standard deviation s = 1 were generated, and all observations less than the censoring limit were converted to < limit. Any data set with 3 or fewer uncensored values was discarded. For each combination of estimation method, sample size, and censoring level, 1,000 trials were performed. The confidence intervals were constructed using the t-statistic with $n - 1$ degrees of freedom (Equation (**??**)). Based on the results in Table 11.13, it looks like using the MLE and either Cox's method or the profile method to construct the confidence interval will provide appropriate coverage when the distribution is correctly specified.

Table 11.13: Results of Monte Carlo simulation study based on Type I left singly censored data from a lognormal parent distribution with log-scale mean = 3 and log-scale standard deviation = 1. Sample sizes (N) are either 10 or 20. Average percent censored (%) are either 20% or 50%. For the MLE, coverage is computed for three nominal 95% confidence interval methods: delta method, cox method, and profile likelihood. For the ROS estimate, coverage is computed for the normal approximation and gpq intervals. For substitution (subs), coverage is computed for two standard errors, one using the number of observations > the dl and one using the total number. Estimated coverage is based on 1,000 simulations. Coverage less than 90% is highlighted.

Estimation			Bias	rmse	Coverage (%)		
Method	N	%	(%)	(%)			
MLE					delta	cox	profile
	10	20	3.9	46.4	**85.8**	90.3	94.4
		50	9.6	54.5	92.7	95.5	94.1
	20	20	1.9	30.4	90.1	94.0	94.4
		50	3.1	34.8	94.4	97.0	93.9
ROS					z (%)		
	10	20	-0.3	40.5	**82.4**		
		50	6.4	43.5	91.3		
	20	20	0.1	29.6	**86.4**		
		50	1.4	30.2	93.4		
subs					#>dl	n	
	10	20	-1.7	40.9	**82.2**	**43.8**	
		50	1.9	44.0	90.6	**38.9**	
	20	20	-0.8	29.8	**86.7**	**28.5**	
		50	-0.8	30.9	92.2	**30.3**	

You should note that this evaluation always used an underlying lognormal distribution, so it does not investigate robustness to departures from this assumption. Helsel and Cohn (1988) performed a Monte Carlo simulation study in which they used lognormal, contaminated lognormal, gamma, and delta (zero-modified lognormal) parent distributions, created multiply censored data, and looked at bias and RMSE. They recommend using the Impute with QQ regression estimator if you are not sure that the underlying population is lognormal.

Example 7. Estimating the Mean NH4 Concentration

In Example 5 we estimated the mean and sd of the log-transformed NH_4 concentration. Table 11.14 displays estimates of and confidence intervals for the mean of untransformed NH_4 concentrations. We see that in this case different methods yield slightly different estimates of the mean and coefficient of variation. This is probably due to the large amount of censoring (61%) and because looking at Table 11.3 we see that all of the observations are less than 0.09 except for two values of 0.12 and 0.25. The various methods of estimation will behave differently for data with unusually large values.

The EnvStatscode is:

11.7.6 Other Distributions

Methods for estimating distribution parameters for other distributions in the presence of censored data are discussed in Cohen (1991). Two that are implemented in EnvStatsare the Gamma distribution and the Poisson distribution. The `epoisCensored()` function estimates the rate parameter of the Poisson distribution in the presence of censored data. The `egammaCensored()` and `egammaAltCensored()` functions calculate the

Table 11.14: Estimates of the mean and coefficient of variation, and 95% confidence intervals for the mean NH_4 deposition.

Estimation Method	Estimated Mean, CV	95% CI calculated by Profile	Cox	normal
MLE	0.020, 1.95	(0.015, 0.026))	(0.014, 0.028)	
Impute with Q-Q Regression	0.019, 1.64			(0.011, 0.027)

maximum likelihood estimates of the parameters of a Gamma distribution. As with the `elnormCensored()` and `elnormAltCensored()` functions, `egammaCensored()` function estimates parameters on the log scale while `egammaAltCensored()` function estimates the arithmetic mean and cv.

The Gamma distribution and log normal distributions have very similar shapes, but the Gamma has a smaller probability of very large values (e.g., larger than 9 standard deviations above the mean). One consequence of this is that the upper confidence limit for the mean of a Gamma distribution is often smaller than the upper confidence limit for the mean of a lognormal distribution (Singh, Signh, and Iaci, 2002).

Example 8. Estimate and confidence limits for the mean NH_4 deposition under an assumed gamma distribution

We might prefer to assume that the population of NH_4 deposition values follows a gamma distribution, instead of a lognormal distribution. In practical terms, this change means that the population has fewer extremely large values (in the gamma model compared to the log normal model). Under this assumption, the estimated mean is 0.019, the estimated c.v. is 1.38, and the 95% profile ci for the mean is (0.014, 0.025). The ci using the normal approximation is (0.014, 0.024). These values were computed using the EnvStatscode:

11.8 Estimating Distribution Quantiles

In this section we will discuss how to estimate and construct confidence intervals for distribution quantiles (percentiles) for data with censored values.

11.8.1 Parametric Methods for Estimating Quantiles

In Chapter **??** we discussed methods for estimating and constructing confidence intervals for population quantiles or percentiles. For a normal distribution, the estimated quantiles are functions of the estimated mean and standard deviation (Equations (**??**) to (**??**). For a lognormal distribution, they are functions of the estimated mean and standard deviation based on the log-transformed observations (Equations (**??**) to (**??**)). In the presence of censored observations, population percentiles are estimated using the same formulae used for uncensored data, but the mean and standard error are estimated using censored data formulae. For example, the maximum likelihood estimate of the p'th percentile of a log-normal distribution is

$$\exp\left(\hat{\mu} + z_p\hat{\sigma}\right), \tag{11.25}$$

where $\hat{\mu}$ and $\hat{\sigma}$ are the ML estimates of μ and σ on the log scale, and z_p is the p'th percentile of a standard normal distribution. The same plug-in approach could be used with Kaplan-Meier (KM) or Regression on Order Statistics (ROS) estimates of μ and σ, but the estimated percentiles are no longer ML estimates. The

statistical properties of the estimated percentiles computed using KM or ROS estimates of the parameters are unknown, but they should be similar to ML estimates.

Very little statistical research has been done on constructing confidence bounds for percentiles when some observations are censored (Choudhary 2007). Confidence bounds can be approximated using the same equations used for uncensored data sets. For example Equation (**??**) for the $100(1-\alpha)\%$ upper confidence limit for the p'th percentile of a log-normal distribution becomes

$$\left[-\infty, \exp\left(\hat{\mu} + t_{n-1, z_p\sqrt{n}, 1-\alpha}\hat{\sigma}_\mu\right)\right] \tag{11.26}$$

when some observations are censored. These equations are approximate because the tolerance interval theory used to construct the confidence bound in equation (11.26) does not apply when observations are censored because the estimated mean and standard deviation are not independent.

Choudhary, P.K., 2007. A tolerance interval approach for assessment of agreement with left censored data. J. Biopharmaceutical Statistics 17:583-594.

A more appropriate method to construct a confidence interval for a quantile uses generalized pivotal quantities (Krishnamorthy and Xu, 2011). This applies the gpq method, described earlier to construct a confidence interval for the mean, to quantiles. For population quantiles from normal or lognormal distributions, the pivotal quantity is $(\Phi^{-1}(p) - \hat{\mu})/\hat{\sigma}$, where $\Phi^{-1}(p)$ is the 100p% percentile of a standard normal distribution, $\hat{\mu}$ is the estimated mean and $\hat{\sigma}$ is the estimated standard deviation. For type II censored data, this pivotal quantity is free of the population parameters and the method is exact assuming the population distribution is normal. For type I censored data, the pivotal quantity is only approximately free of the population parameters, so the confidence bound is only approximate (*** need reference).

The generalized pivotal quantity algorithm to calculate a 100(1-α)% upper bound for the 100p%'ile of a normal distribution is:

1. estimate $\hat{\mu}$ and $\hat{\sigma}$ for the real data set. Call these $\hat{\mu}_0$ and $\hat{\sigma}_0$

2. simulate a data set with the appropriate mix of censored and uncensored observations a normal distribution with mean = $\hat{\mu}_0$ and sd = $\hat{\sigma}_0$

3. estimate $\hat{\mu}$ and $\hat{\sigma}$ for the simulated data set

4. Compute the pivotal quantity for a normal distribution with mean $\hat{\mu}_0$ and sd $\hat{\sigma}_0$ as

$$Q_p^* = \frac{\hat{\mu}_0 + \hat{\sigma}_0\, \Phi^{-1}(p) - \hat{\mu}}{\hat{\sigma}}$$

5. repeat steps 2-4 a large number of times, e.g., 1000 or 10000 times.

6. estimate $Q^*_{p,(1-\alpha)}$, the empirical 1-α quantile of the distribution of Q_p^*

7. the required bound is then

$$\hat{\mu}_0 + Q^*_{p,(1-\alpha)}\, \hat{\sigma}_0 \tag{11.27}$$

If the population was assumed to follow a lognormal distribution, this approach would be used after log transforming the observations. The calculated bound from Equation (11.27) would then be exponentiated to get a bound on the raw data scale.

11.8.2 Monte Carlo Simulation of the generalized pivotal quantity and traditional methods to estimate tolerance intervals from data with censored values

The performance of both the generalized pivotal quantity method and the traditional method using a non-central T distribution were studied in a small simulation. The population follows a normal distribution with a mean of 3 and sd of 1. A sample of 10, 20, or 50 observations was randomly generated and censored at 2.16 to give an expected 20% censoring level or 3 to give an expected 50% censoring. The 90% upper bound for the 95%'ile was computed and compared to the population 95%'ile of 4.645. This was repeated 1000 times.

The results are shown in Table 11.15. The bounds computed using the traditional non-central T distribution, Equation (11.26), are generally too small and the coverage of the confidence bound is below the nominal 90%. The quantity $s/\sqrt{n}$ in Equation (**??**) for uncensored data was replaced by $\hat{\sigma}_\mu$ in Equation (11.26). Because $\hat{\sigma}_\mu \geq s/\sqrt{n}$ when some observations are censored, using Equation (**??**) for data with censored observations will have even worse coverage. The bounds computed by the generalized pivotal quantity are too large and the coverage is conservative. The degree to which the bounds are too large increases with increasing sample size and increasing proportion of censored observations.

Table 11.15: Performance of the generalized pivotal quantity (GPQ) and traditional, non-central T (NCT) estimators of the 90% upper confidence bound for the 95%'ile of a normal distribution.

				% of times	
#	%	mean bound		above pop. quantile	
obs.	censored	GPQ	NCT	GPQ	NCT
10	20	5.82	5.45	89.4	86.2
10	50	6.79	5.40	96.7	85.0
20	20	5.33	5.15	92.3	88.4
20	50	6.19	5.15	98.5	85.9
50	20	5.07	4.94	93.3	87.5
50	50	5.68	4.93	99.8	84.1
50	80	6.39	4.93	100	82.3

Example 9. Estimating and Constructing an Upper Confidence Limit for the 95th Percentile of the NH_4 deposition data

In examples **??** and 5, we found that the NH_4 deposition data was approximately modeled by a lognormal distribution with log-scale mean = -4.28 and sd = 1.25. The maximum likelihood estimate of the 95%'ile is -2.65 on the log scale and 0.070 on the concentration scale. An upper 90% confidence bound on the 95%'ile is 0.092 using the traditional method and 0.074 using the generalized pivotal quantity method. These values were computed in R by:

```
attach(Olympic.NH4.df)
tolIntLnormCensored(NH4.mg.per.L,Censored, ti.type = 'u', coverage = 0.95,
  conf.level = 0.90)
tolIntLnormCensored(NH4.mg.per.L,Censored, ti.type = 'u', ti.method='gpq',
  coverage = 0.90, conf.level = 0.95)
detach()
```

11.8.3 Nonparametric Methods for Estimating Quantiles

We saw in Section **??** that nonparametric estimates and confidence intervals for quantiles of the population are simply functions of the ordered observations (Equations (***) to (***)). If the data have a single detection limit below all the observed values, you can still estimate quantiles and create one-sided upper confidence intervals as long as there are enough uncensored observations that can be ordered in a logical way (see Example ***). The theory of nonparametric confidence intervals applies without any difficulty to this sort of censored data, because it requires only that observations can be ordered and the uncensored observations each add 1/n to the empirical CDF.

When the data are multiply censored, a quantile can easily be estimated by inverting the Kaplan-Meier estimate of $F(t)$. We know of no work on confidence intervals for quantiles of data with multiple censoring limits.

11.9 Prediction and Tolerance Intervals

In this section we will discuss how to construct prediction and tolerance intervals based on data with censored values.

11.9.1 Parametric Methods for Prediction and Tolerance Intervals

In Chapter **??** we discussed methods for constructing prediction and tolerance intervals. For a normal distribution, the prediction and tolerance intervals are functions of the estimated means and standard deviations (Equations (***) to (***)). For a lognormal distribution, they are functions of the estimated mean and standard deviation based on the log-transformed observations.

In the presence of censored observations, we can construct prediction intervals using the same formulae as for complete data, with mean and standard deviation estimated using censored data methods. It is not clear how well these methods behave in the presence of censored data. One concern is the correlation between the estimated mean and standard deviation, which means that a T distribution is at best an approximation. This is an area that requires further research.

The one-sided tolerance bound is the upper confidence bound of a proportion. Section 11.8.1 describes methods that are appropriate for data with censored values. A small simulation study (Table 11.15) shows that the traditional method based on a non-central T distribution is liberal and the the generalized pivotal quantity method is generally conservative for type I censored data.

Example 10. Prediction and tolerance intervals for NH_4 deposition

In examples **??** and 5, we found that NH_4 deposition data was approximately modeled by a lognormal distribution with log-scale mean = -4.28 and sd = 1.25. A two-sided 95% prediction interval for a single new observation, computed using Equation (**??**), is (-7.2, -2.2) on the log scale and (0.0007, 0.11) on the concentration scale. These values were computed in R by:

```
attach(Olympic.NH4.df)
NH4.mle <- elnormCensored(NH4.mg.per.L, Censored)
predIntLnorm(NH4.mle, k=1, pi.type='two-sided', conf.level=0.95, method='exact')
detach()
```

The argument `k=1` specifies the number of new observations.

11.9.2 Nonparametric Prediction and Tolerance Intervals

We saw in Chapter **??** that nonparametric prediction and tolerance intervals are simply functions of the ordered observations (Equations (***) to (***)). Thus, for left censored data, you can still estimate quantiles and create one-sided upper confidence intervals as long as there are enough uncensored observations that can be ordered in a logical way (see Example *** and Example ***).

When the data have multiple detection limits, so some observed points add more than 1/n to the empirical CDF, it is still easy to estimate a quantile by inverting the Kaplan-Meier estimate of the cumulative distribution function. It is not clear whether the theory of non-parametric confidence intervals for quantiles can be used without modification, can be used to get approximate intervals, or can not be used.

11.10 Hypothesis Tests

In this section, we discuss how to perform hypothesis tests when the data include censored values. In general, the concepts for each test are similar to the concepts for tests on uncensored data, but the details are different.

11.10.1 Goodness-of-Fit Tests

In Chapter **??** we showed that the Shapiro-Francia goodness-of-fit statistic can be written as the square of the sample correlation coefficient between the ordered observations and the expected values of the normal order statistics (Equation (***)). In practice we use an approximation to the true expected values of the normal order statistics based on Blom plotting positions, so that the Shapiro-Francia goodness-of-fit statistic is the square of the probability plot correlation coefficient (Equations (***) to (***)), and is therefore equivalent to the goodness-of-fit test based on the probability plot correlation coefficient (the PPCC test). Similarly, the Shapiro-Wilk goodness-of-fit statistic can be written as the square of the sample correlation coefficient between the ordered observations and the expected values of the normal order statistics weighted by their variance-covariance matrix (Equation (***)).

Using these formulations, Royston (1993) extended both the Shapiro-Francia (PPCC) and Shapiro-Wilk goodness-of-fit tests to the case of singly censored data by computing these statistics based on the uncensored observations, similar to the way we explained how to construct a Q-Q plot for singly censored data earlier in this chapter in Section 11.3.2. Royston (1993) provides a method of computing p-values for these statistics based on tables given in Verrill and Johnson (1988). Although Verrill and Johnson (1988) produced their tables based on Type II censoring, Royston's (1993) approximation to the p-value of these tests should be fairly accurate for Type I censored data as well.

The PPCC test is also easily extendible to the case of multiply censored data, but it is not known how well Royston's (1993) method of computing p-values works in this case.

We noted in Chapter **??** that goodness-of-fit tests are of limited value for small sample sizes because there is usually not enough information to distinguish between different kinds of distributions. This also holds true even for moderate sample sizes if you have censored data, because censored observations provide less information than uncensored values.

11.10.2 Hypothesis Tests on Distribution Parameters and Quantiles

In Chapter **??** we noted the one-to-one relationship between confidence intervals for a population parameter and hypothesis tests for the parameter (see Table ***). You can therefore perform hypothesis tests on distribution parameters and quantiles based on censored data by constructing confidence intervals for the parameter or quantile using the methods discussed in this chapter.

11.10.3 Nonparametric Tests to Compare Two Groups

In Chapter **??** we discussed various hypothesis tests to compare locations (central tendency) between two groups, including the Wilcoxon rank sum test, linear rank tests, and the quantile test. In the presence of censored observations, you can still use the Wilcoxon rank sum test or quantile test as long as there are enough uncensored observations that can be ordered in a logical way. For example, if both samples are singly censored with the same censoring level and all uncensored observations are greater than the censoring level, then all censored observations receive the lowest ranks and are considered tied observations. You can also use the Wilcoxon rank sum test even with multiply censored data as we will now discuss.

The Wilcoxon rank sum test is a linear rank test, see Equation (**??**), with a specific form of score function. Extensions of a linear rank test can be used for data with multiple censoring values (Prentice 1978). Most of this literature considers right censored survival data, but the same ideas can be used with left censored environmental data. Different linear rank tests use different score functions, and some tests may be better than others at detecting a small shift in location, depending upon the true underlying distribution.

Millard and Deverel (1988) studied the behavior of several linear rank tests for censored data. The normal scores test using the Prentice and Marek computation of the normal scores was the best at maintaining the nominal type-I error rate in both equal and unequal sample sizes (Millard and Deverel, 1988). When the sample sizes are equal, the Peto-Peto test (called Peto-Prentice in Millard and Deverel (1988)) with the asymptotic variance computation also maintained the nominal type I error rate and had a slightly higher power than the normal scores test (Millard and Deverel, 1988).

Example 11. Comparing Copper Concentrations between Two Geological Zones

In Example 2 we compared the empirical cumulative distribution functions of copper concentrations in the alluvial fan and basin trough zones (see Table **??** and Figure 11.3). The plot indicates the distributions of concentrations are fairly similar. The `test='normal.scores.2'` option requests the Prentice and Marek computation of the normal scores. The `test='peto-peto', variance='asymptotic'` options requests the Peto-Prentice test with asymptotic variance computation. The two-sample linear rank test based on normal scores and a hypergeometric variance yields a p-value of 0.2, indicating no evidence of a difference. This was computed in EnvStatsby:

```
attach(Millard.Deverel.88.df)
x <- Cu[Zone=='Alluvial.Fan']
x.cens <- Cu.censored[Zone=='Alluvial.Fan']
y <- Cu[Zone=='Basin.Trough']
y.cens <- Cu.censored[Zone=='Basin.Trough']
Cu.test <-  twoSampleLinearRankTestCensored(x, x.cens,
  y, y.cens, test='normal.scores.2')
Cu.test$p.value
```

```
[1] 0.2008913
```

```
Cu.test2 <-  twoSampleLinearRankTestCensored(x, x.cens,
  y, y.cens, test='peto.peto', variance='asymptotic')
Cu.test2$p.value
```

```
[1] 0.3557858
```

```
detach()
```

11.10.4 Parametric Regression and ANOVA

Methods for fitting parametric regression or ANOVA models in the presence of singly and multiply censored response observations have been developed in the fields of econometrics and life testing (e.g., Amemiya, 1985, Chapter 10; Meeker and Escobar, 1998). In econometrics, these are sometimes called Tobit models (Amemiya, 1985, p. 360). Maximum likelihood is the most common approach to estimate parameters, so hypothesis testing is by a likelihood ratio test (see section **??** for additional details).

The `survreg()` function in the survival library in R implements maximum likelihood estimation for right-, interval- and left-censored data. The potentially censored response information is captured by creating a "Surv" object that contains information about the value, whether each observation was censored (alive in survreg's terminology) and the type of censoring. Once created, just about any linear model (regression, a nonparametric regression using splines, ANOVA, ANCOVA) can be fit to that response object. We show two examples.

Example 12. Parametric test to compare Copper concentrations in two alluvial zones

In Example 11, we used a non-parametric test to compare Copper concentrations in two alluvial zones. Here, we use the survreg() function to conduct a parametric likelihood ratio test. The estimated difference in log transformed Copper concentrations is 0.12, with a standard error of 0.18. The p-value for the test of no difference is 0.51; the conclusion from the parametric test is the same as that from the non-parametric test. These results were obtained with the following R code:

```
temp <- Millard.Deverel.88.df
temp$Cu.surv <- Surv(temp$Cu, event=!temp$Cu.censored, type='left')
Cu.test <- survreg(Cu.surv~factor(Zone),
  data=temp, dist='lognormal')
summary(Cu.test)
```

```
Call:
survreg(formula = Cu.surv ~ factor(Zone), data = temp, dist = "lognormal")
                         Value Std. Error      z        p
(Intercept)              0.933     0.1162  8.036 9.28e-16
factor(Zone)Basin.Trough 0.116     0.1765  0.658 5.10e-01
Log(scale)              -0.151     0.0767 -1.966 4.93e-02

Scale= 0.86

Log Normal distribution
Loglik(model)= -217.6   Loglik(intercept only)= -217.8
```

```
        Chisq= 0.43 on 1 degrees of freedom, p= 0.51
Number of Newton-Raphson Iterations: 3
n=114 (4 observations deleted due to missingness)
```

```
anova(Cu.test)
```

```
             Df  Deviance Resid. Df    -2*LL  Pr(>Chi)
NULL         NA        NA       112 435.5367        NA
factor(Zone)  1 0.4315204       111 435.1052 0.5112439
```

Notice that the variable event for `Surv()` is defined as True = dead (observed value) and False = alive (censored). Hence, event is defined as the logical not of the value of the censoring indicator (`!Cu.censored`). It is also necessary to specify that the data left censored by adding `type='left'` in the call to `Surv()`. The `factor(Zone)` in the survreg model creates an indicator variable differentiating between the two alluvial zones. If you wanted to use a normal distribution instead of the lognormal, you would replace `dist='lognormal'` with `dist='gaussian'` in the call to `survreg()`.

Example 13. Testing for trend and predicting log NH4 deposition

The NH_4 deposition data are collected approximately weekly. One very relevant question is whether there is any temporal trend in deposition. One simple way to evaluate this is to fit a linear regression with week as the X variable. That can be done by the following R code:

```
attach(Olympic.NH4.df)
NH4.surv <- Surv(NH4.mg.per.L, event=!Censored, type='left')
NH4.trend <- survreg(NH4.surv~Week, data=Olympic.NH4.df,
  dist='lognormal')
summary(NH4.trend)
```

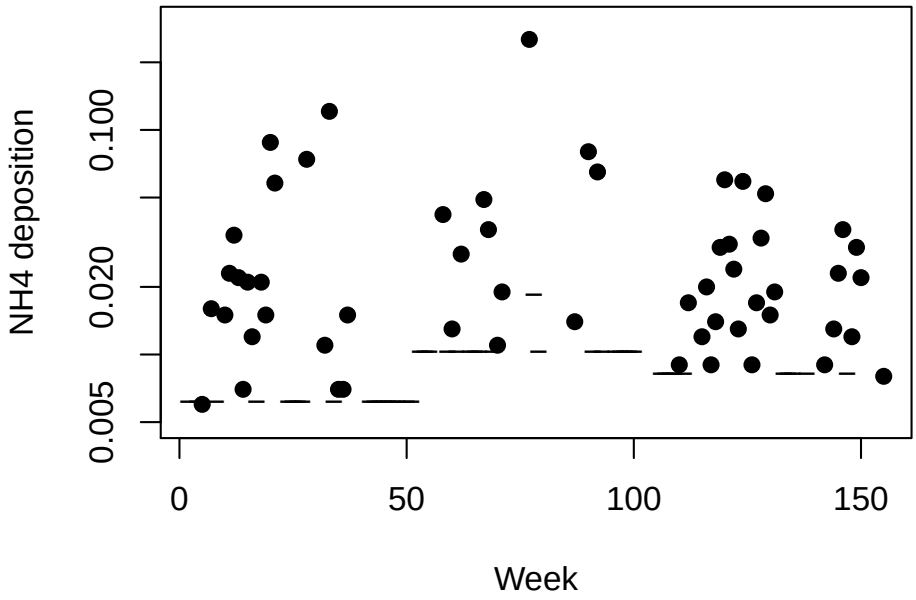
```
Call:
survreg(formula = NH4.surv ~ Week, data = Olympic.NH4.df, dist = "lognormal")
               Value Std. Error      z        p
(Intercept) -4.99062    0.26791 -18.63 1.91e-77
Week         0.00365    0.00288   1.27 2.04e-01
Log(scale)   0.21711    0.10369   2.09 3.63e-02

Scale= 1.24

Log Normal distribution
Loglik(model)= 89   Loglik(intercept only)= 88.2
        Chisq= 1.61 on 1 degrees of freedom, p= 0.2
Number of Newton-Raphson Iterations: 3
n= 102
```

```
detach()
```

The p-value for week is ca. 0.2, indicating little evidence of a linear trend over time.

However, a plot of the data, where two different symbols are used for censored (-) and observed (dot) values suggests there is a seasonal trend, with low amounts of deposition in winter (i.e. around week 0, week 52, week 104).

```
attach(Olympic.NH4.df)
plot(Week,NH4.mg.per.L,type='n', xlab='Week',
  ylab='NH4 deposition', log='y', ylim=c(0.005,0.30))
points(Week[Censored], NH4.mg.per.L[Censored],
  pch='-')
points(Week[!Censored], NH4.mg.per.L[!Censored],
  pch=19)
detach()
```



One way to model seasonality and trend is to describe the seasonal patterns by a trignometric regression, i.e. by a model term like sin(2*π*week/52), so one year is a full cycle of the trignometric term. That model suggests weak (at best) evidence of a trend. The R code, using the NH4.surv object previously created, is:

```
NH4.trend <- survreg(NH4.surv~I(sin(2*pi*Week/52))+
   I(cos(2*pi*Week/52))+ Week, data=Olympic.NH4.df, dist='lognormal')
summary(NH4.trend)
```

```
Call:
survreg(formula = NH4.surv ~ I(sin(2 * pi * Week/52)) + I(cos(2 *
    pi * Week/52)) + Week, data = Olympic.NH4.df, dist = "lognormal")
                            Value Std. Error      z        p
(Intercept)              -5.13595     0.2609 -19.68 3.06e-86
```

```
I(sin(2 * pi * Week/52))  0.38307     0.1733   2.21 2.71e-02
I(cos(2 * pi * Week/52)) -0.72432     0.1919  -3.77 1.60e-04
Week                      0.00495     0.0027   1.83 6.67e-02
Log(scale)                0.10916     0.1021   1.07 2.85e-01

Scale= 1.12

Log Normal distribution
Loglik(model)= 99   Loglik(intercept only)= 88.2
        Chisq= 21.68 on 3 degrees of freedom, p= 7.6e-05
Number of Newton-Raphson Iterations: 4
n= 102
```

since week 0 is Jan 1. The evidence for a linear trend, after adjusting for seasonal effects modeled as a sine curve and cosine curve, is weak. The estimated trend is 0.0049 per week (0.26 per year) with a standard error of 0.0027 (0.14 per year). The p-value for trend is 0.067 (weak, but suggestive, evidence of a trend).

If the goal were simply to describe patterns over time and predict NH_4 deposition, a more non-parametric approach might be appropriate. One very useful non-parametric regression approach is to regress Y on a polynomial-basis spline or a natural-cubic spline representation of week. The survreg() function can fit both using bs(week) for the polynomial basis and ns(week) for natural cubic split basis. The R code to fit these models, predict NH_4 deposition, and plot the data and predicted values is:

```
NH4.trend1 <- survreg(NH4.surv~bs(Week, knots=seq(26,150,26)),
  data=Olympic.NH4.df, dist='lognormal')
NH4.trend2 <- survreg(NH4.surv~ns(Week, knots=seq(26,150,26)),
  data=Olympic.NH4.df, dist='lognormal')
```

Both splines were fit with knots every 26 weeks to capture the annual rise and fall of NH_4 deposition. The fits of both the polynomial-basis and natural-cubic splines are shown in Figure 11.7.

11.10.5 Diagnostics for regression and ANOVA models

Our focus will be on visual assessment of models, as it was for assessing regression and ANOVA models for data without any censored observations. Previous model assessments were based on residuals, the difference between the observed value and the predicted value from some model. The difficulty applying this to censored values, e.g. < 5, are immediately obvious: the predicted value is unambiguous, but what "observed" value do you use to compute the residual? Simply ignoring the censored values is also not appropriate, because they will usually be the smaller values in the data set. Ignoring them biases the distribution of residuals.

One solution is to base model assessments on the deviance residual. Comparisons among models fit by maximum likelihood use $-2(lnL_{null} - lnL_{full})$ as the test statistic (see section **??**). If you define the deviance of a model as $D = -2lnL$, then there is a very clear correspondence between model comparison in a pair of linear models and model comparison in a pair of models fit by maximum likelihood. The change in Error sums-of-squares, $SSE_{null} - SSE_{full}$ in a linear model, is replaced by $D_{null} - D_{full}$ in a model fit by maximum likelihood. Notice that $SSE_{model} = \sum(Y_i - \hat{Y}_i)^2 = \sum r_i^2$, where r_i is the usual residual for the i'th observation. To continue the analogy, what can we square and sum to arrive at the deviance? The answer is the deviance residual:

$$D_i = sign(Y_i - \hat{Y}_i)\sqrt{\mid -2 * lnL_i \mid}$$

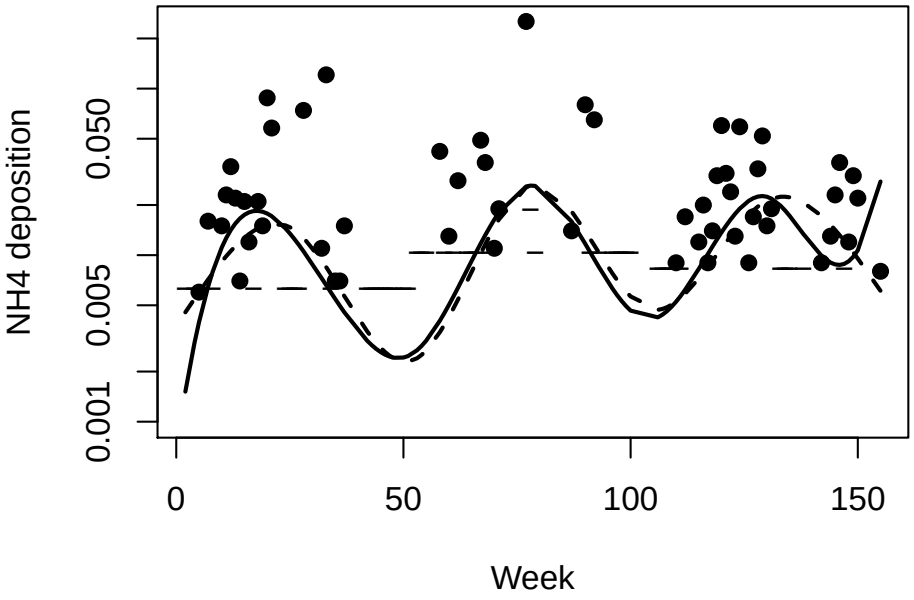


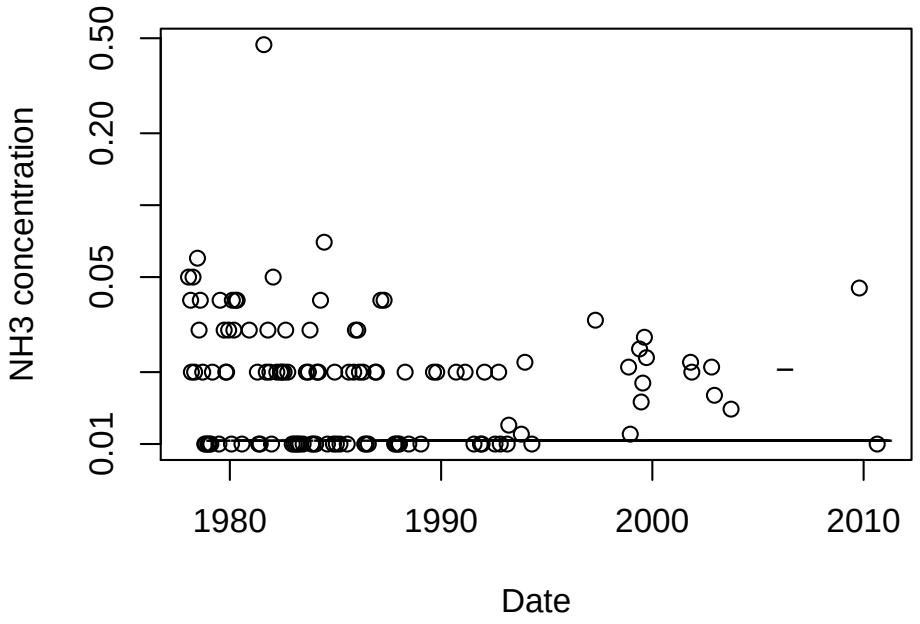
Figure 11.7: Plot of NH_4 deposition over time with predicted NH_4 deposition shown by lines. Solid line is the polynomial basis spline with knots every 26 weeks. The dashed line is the natural cubic spline with knots every 26 weeks. Observed values of NH_4 are indicated by solid dots; censored values are shown as the - symbol and plotted at their censoring limit.

where the $sign(Y_i - \hat{Y}_i)$ function is 1 if $Y_i > \hat{Y}_i$, 0 if $Y_i = \hat{Y}_i$, and -1 if $Y_i < \hat{Y}_i$. For censored values, Y_i is usually taken as the detection limit, but this is almost always not an issue because usually $\hat{Y}_i$ exceeds the detection limit, so the sign is unambiguous.

Example 14. Testing for trend in the Skagit River NH3 concentrations

Has the concentration of NH_3 in the Skagit River declined over time? Is this decline appropriately described by a linear trend over time, or is some more complicated model needed? The raw data give a pretty clear answer to the first question, but we will fit a model with a constant mean and look at the deviance residuals. Then we will then evaluate the residuals from a linear trend model.

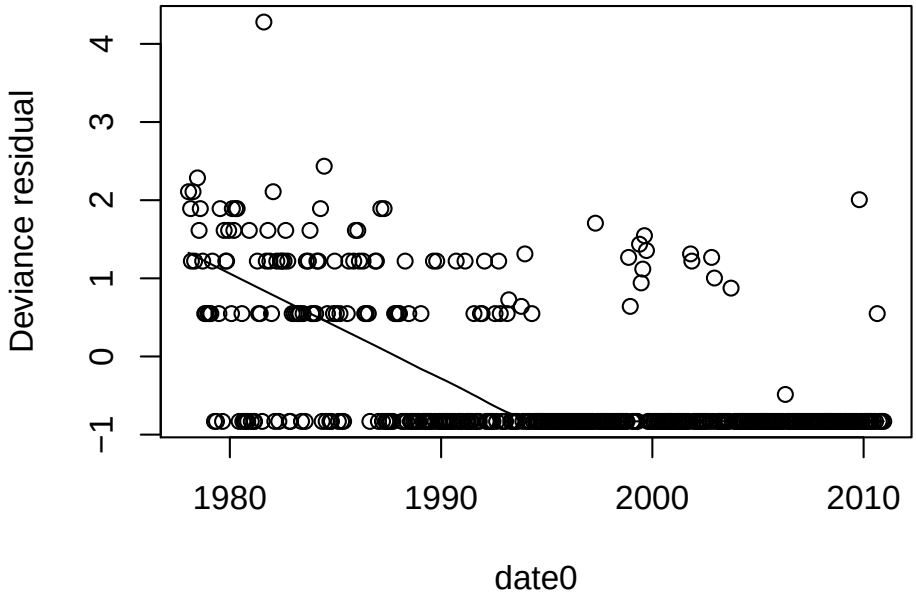
```
# plot the data
attach(Skagit.NH3_N.df)
date0 <- as.Date(Date,format='%m/%d/%Y')
plot(date0,NH3_N.mg.per.L,type='n', xlab='Date', ylab='NH3 concentration', log='y')
points(date0[!Censored], NH3_N.mg.per.L[!Censored])
points(date0[Censored], NH3_N.mg.per.L[Censored], pch='-')
detach()
```



The data suggests the mean and median concentrations of NH_3 have declined over time, and below detection-limit observations are more frequent in later years.

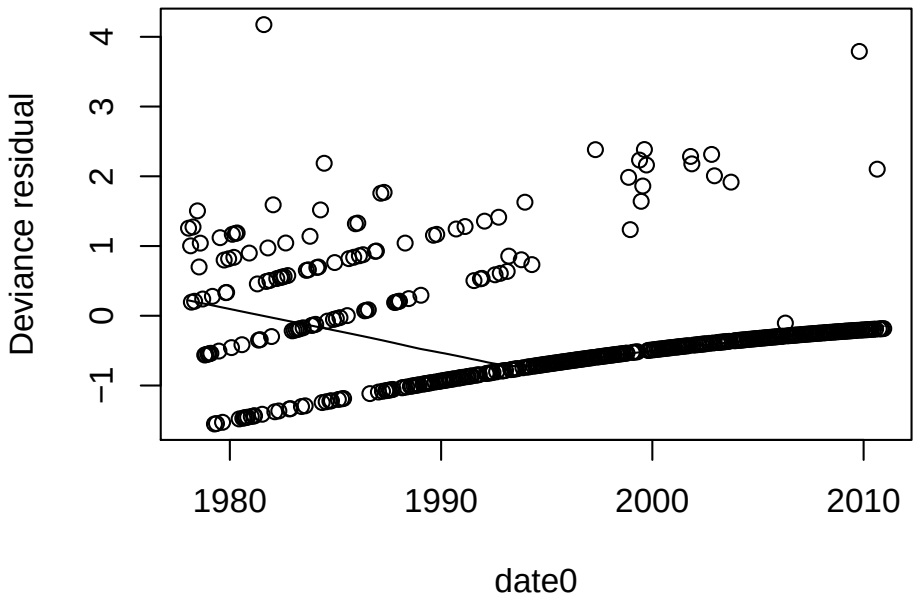
```
# plot the residuals from constant mean model
#    over time and add smooth
temp <- Skagit.NH3_N.df
```

```
temp <- temp[!is.na(temp$NH3_N.mg.per.L), ]
attach(temp)
date0 <- as.Date(Date,format='%m/%d/%Y')
Skagit.surv <- Surv(NH3_N.mg.per.L, !Censored, type='left')
Skagit.lm1 <- survreg(Skagit.surv~1, dist='lognormal')
plot(date0,residuals(Skagit.lm1, type='deviance'), ylab='Deviance residual')
lines(lowess(date0,residuals(Skagit.lm1,type='deviance')))
detach()
```



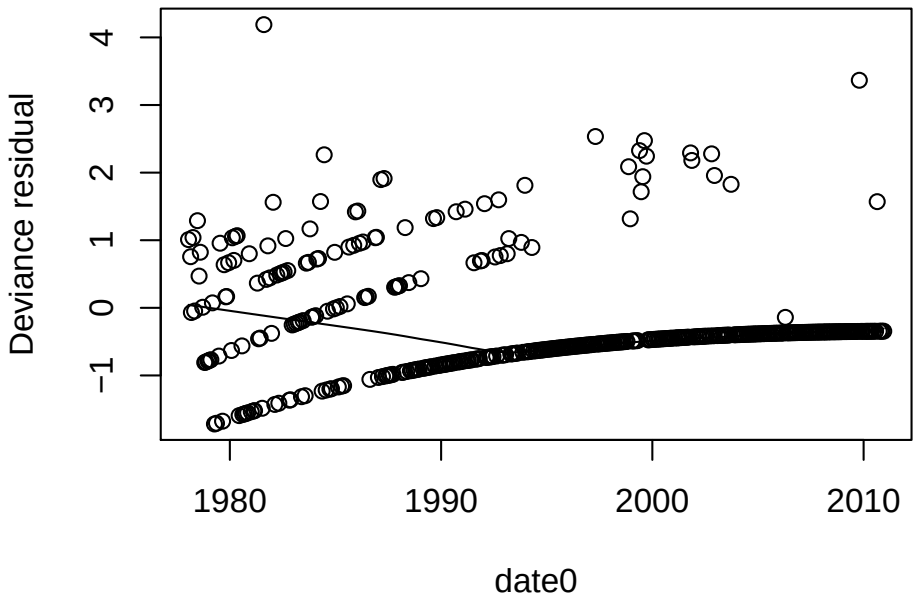
The residuals strongly indicate lack of fit to the constant mean model.

```
# plot residuals from quadratic model, with smooth
attach(Skagit.NH3_N.df)
Skagit.lm2 <- survreg(Skagit.surv~date0, dist='lognormal')
plot(date0,residuals(Skagit.lm2, type='deviance'), ylab='Deviance residual')
lines(lowess(date0,residuals(Skagit.lm2,type='deviance')))
detach()
```



There is a small amount of curvature in the residuals. Let's fit a model with a quadratic term, after centering and rescaling the date variable to avoid computing issues.

```
# plot residuals from quadratic model, with smooth
attach(Skagit.NH3_N.df)
date1 <- as.integer(date0)/1000
date1 <- (date1 - mean(date1))/sd(date1)
date2 <- date1^2
Skagit.lm3 <- survreg(Skagit.surv~date1+date2, dist='lognormal')
plot(date0,residuals(Skagit.lm3, type='deviance'), ylab='Deviance residual')
lines(lowess(date0,residuals(Skagit.lm3,type='deviance')))
detach()
```



This plot looks reasonably flat and fitting a cubic is not a significant improvement.

Akritas and Sen

11.11 Summary

1. Censored data can be classified as truncated vs. censored, left vs. right vs. double, singly vs. multiply, and Type I vs. Type II. Most environmental data sets with nondetect values are Type I left singly or multiply censored.
2. You can use quantile (empirical cdf) plots and Q-Q plots to compare censored data with a theoretical distribution or compare two samples.
3. Censored data have been studied extensively in the field of life testing and survival analysis. Several methods exist for estimating distribution parameters, creating confidence intervals, and performing hypothesis tests based on censored data.

New References

Efron, B. and Hinkley, D.V. 1978. Assessing the accuracy of maximum likelihood estimator: Observed versus Fisher Information. Biometrica 65:457-482

Kaplan and Meier

Monahan, J.F. 2001. *Numerical Methods of Statistics*. Cambridge University Press, New York.

Peng, C. 2010. Interval estimation of population parameters based on environmental data with detection limits. Environmetrics 21:645-658

Shumway, R.H., R.S. Azari, and M. Kayhanian, 2002, Statistical approaches to estimating mean water quality concentrations with detection limits. *Environmental Science and Technology* **36**:3345-3353.

Singh, A., Singh, A., and Iaci, R. 2002. Estimation of the exposure point concentration term using a Gamma distribution. US. EPA document EPA/600/R-02/084.