Quality Control Data Analysis

Dave Lorenz, Jeffrey Martin, and Laura Medalie

December 30, 2015

Abstract

These examples demonstrate some of the functions and statistical methods for the analysis of water-quality control data that are available in the `smwrQW` package. The examples illustrate many of the concepts discussed by Mueller and others (2015).

Contents

1 Introduction

Most of the examples will use datasets supplied in the **smwrQW** package, but the Statistical Concepts section will use small, artificial datasets, with known properties to illustrate the use of the functions. The artificial data are class "lcens" rather than class "qw," but the statistical methods apply to both classes. The user should read the Working with Water-Quality Data vignette to become familiar with the management of water-quality data in the `smwrQW` package before attempting the examples in this vignette.

```
> # Load the smwrQW package
> library(smwrQW)
> # Generate normal data
> set.seed(132)
> Xn <- rnorm(26, mean=5)
> # And left-censor it at 2 levels
> Xc <- as.lcens(Xn, rep(c(4.5, 5.0), 13))
> # And log-normal data
> Xln <- rlnorm(26, mean=0.1, sd=0.3)
> # And left censor it at 2 levels
> Xlc <- as.lcens(Xln, rep(c(0.8, 1.2), 13))
> # Multiply censor the data
> Xmc <- censor(Xln, 0.8, 1.75)
```

2 Statistical Concepts

It is not possible, physically or financially, to measure all occurrences of every characteristic of interest in environmental studies. For some characteristics, any direct measurement is impossible. Thus, statistical methods are necessary to make estimates of these characteristics. Such estimates can be less than satisfying, and even the subject of disbelief or derision.

Water-quality data are complicated because the data are reported as censored values if the measured result is less than the reporting level, defined by the laboratory for each method and analyte. Censored values can be problematic for statistical analyses, particularly if some assumed value, such as zero or one-half the reporting level, is substituted for the censored result (see Helsel, 2012, for a detailed discussion). The methods in the **smwrQW** and described in this vignette provide statistically unbiased, or asymptotically unbiased, results with very little user adjustments and do not rely on simple substitution.

2.1 Confidence Interval of the Mean

The uncertainty of an inferential statistic often is indicated by reporting a range of values, referred to as a "confidence interval." Confidence intervals are constructed to contain an unknown characteristic of the population, such as the mean, median, standard deviation, or a percentile, with a specified probability. The width of the confidence interval is the uncertainty due to estimation of a population characteristic based on sample data. Note that the estimation of a confidence interval is appropriate for random samples from a single population.

The equations for computing confidence intervals are not reproduced in this vignette. Mueller and others (2015) and Helsel (2012) present and describe the equations for the user who desires more information about the actual computation.

The `censMean.CI` function can be used to compute the confidence interval for the sample mean. There are six computational methods for the `censMean.CI` function, each designed for different distributions and censoring. The "AMLE" method can be used for uncensored data or left-censored data that can be assumed to be normally distributed. The "MLE" method can be used for multiply-censored data that can be assumed to be normally distributed. The "ROS" method can be used for left-, multiply-, or uncensored data that are approximately normally distributed. The "ROS" method uses robust estimation and bootstrapping to relax the assumption of normality. The "log AMLE" method can be used for left-, or uncensored data that can be assumed to be log-normally distributed. The "log MLE" method can be used for multiply-censored data that can be assumed to be log-normally distributed. The "log ROS" method can be used for left-, multiply-, or uncensored data that are approximately log-normally distributed. The "log ROS" method uses robust estimation and bootstrapping to relax the assumption of normality.

The confidence interval for the mean of normally distributed, uncensored data can also be computed using the `t.test` function in base R as demonstrated in the code immediately following this paragraph. The code also illustrates the use of the `censMean.CI` in the **smwrQW** package. The method "AMLE" is used to compute the confidence interval. The adjusted maximum likelihood method, "AMLE," is first-order unbiased and replicates the results from `t.test`.

```
> # The t.test for confidence interval of the mean
> t.test(Xn, conf.level=.9)

        One Sample t-test
```

```
data:  Xn
t = 27.638, df = 25, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:
 4.676905 5.293098
sample estimates:
mean of x
 4.985001
```

```
> # The adjusted maximum likelihood
> censMean.CI(Xn, method="AMLE", CI=.9)
```

```
     estimate      lcl      ucl  ci
mean 4.985001 4.676905 5.293098 0.9
```

```
> # Compare to the left-censored data
> censMean.CI(Xc, method="AMLE", CI=.9)
```

```
     estimate      lcl      ucl  ci
mean 4.886856 4.525555 5.248157 0.9
```

Most often, water-quality data can be assumed to be from a log-normal distribution. The `method` argument must be "log AMLE" for an unbiased estimate of uncensored or left0censored data, "log MLE" for an asymptotically unbiased estimate of any censored data, or "log ROS" for a robust estimate for any censored data. The confidence intervals of the mean of a log-normally distributed sample are not simply back-transformed confidence intervals of the mean of the logs, see Helsel (2012) for details. The code following this paragraph demonstrates the "log AMLE" method for both the uncensored and leftcensored log-normal data and uses "log MLE" and "log ROS" for the mutliply censored data.

```
> # The AMLE, uncensored
> censMean.CI(Xln, method="log AMLE", CI=.9)
```

```
     estimate      lcl      ucl  ci
mean  1.22684 1.121109 1.345951 0.9
```

```
> # The AMLE, left censored
> censMean.CI(Xlc, method="log AMLE", CI=.9)
```

```
     estimate      lcl      ucl  ci
mean 1.245633 1.146584 1.357304 0.9
```

```
> # MLE for multiply censored
> censMean.CI(Xmc, method="log MLE", CI=.9)
```

```
     estimate      lcl      ucl  ci
mean 1.225595 1.122493 1.338166 0.9
```

```
> # and ROS
> censMean.CI(Xmc, method="log ROS", CI=.9)

     estimate      lcl      ucl  ci
mean  1.22873 1.121265 1.352765 0.9
```

2.2 Confidence Interval for the Median and other Percentiles

The median and other percentiles are statistics of particular importance in QC analyses. Nonparametric confidence intervals on percentiles are calculated using the binomial probability function when applied to uncensored data (Mueller and others, 2015) and an approximation using the Kaplan-Meier method described by Helsel (2012) when the data are censored. In general, the confidence interval for the approximation is slightly smaller than that based on the binomial distribution, as demonstrated by the first part of the code following this paragraph. The `qtiles.CI` function in the **smwrStats** package computes confidence intervals for uncensored data and the `censQtiles.CI` function in the **smwrQW** package computes confidence intervals for left-censored data. The code following this paragraph demonstrates both functions.

```
> # Uncensored data, by default, 90% CI for median
> qtiles.CI(Xln)

    estimate      lcl      ucl        ci
50% 1.200256 1.054237 1.335974 0.9244813

> censQtiles.CI(Xln)

    estimate      lcl      ucl  ci
50% 1.200256 1.054609 1.289765 0.9
attr(,"minimum")
[1] 0.6258476
attr(,"maximum")
[1] 2.015399

> # Compare to censored data
> censQtiles.CI(Xlc)

    estimate      lcl      ucl  ci
50% 1.200256 1.054609 1.289765 0.9
attr(,"minimum")
[1] 0.8
attr(,"maximum")
[1] 2.015399

> # Compute the upper confidence interval for selected probabilities
> censQtiles.CI(Xlc, probs=c(.75, .9, .95), bound="upper")

    estimate lcl      ucl  ci
75% 1.497274   0 1.561259 0.9
```

```
90% 1.609559   0 1.681648 0.9
95% 1.681648   0 2.015399 0.9
attr(,"minimum")
[1] 0.8
attr(,"maximum")
[1] 2.015399
```

```
> # Compare to "filled in" values
> Xlc.f <- fillIn(Xlc, method="log ROS")
> qtiles.CI(Xlc.f, probs=c(.95), bound="upper")
```

```
    estimate  lcl ucl ci
95% 1.681648 -Inf  NA NA
```

The maximum reported for the upper confidence level by the approximation method is the largest value in the data, which is included as an attribute of the returned value. There can be times when that is acceptable. However, if the user must know that the actual upper limit could be greater than the largest value, then the user can estimate all values using the `fillIn` function (assuming either a normal or log-normal distribution) and use the `qtiles.CI` function to determine if the upper level matches the largest value or exceeds it (returned as `NA` as in the example).

2.3 Confidence Interval for a Proportion

Proportions can be computed for a dataset by dividing the observations into groups, such as those less than or greater than a specified value. In water-quality analyses, proportions often are used to indicate the frequency of analyte detection or exceeding a specified threshold within the total number of observations in a sample dataset. The code following this paragraph demonstrates a simple application of determining the confidence interval of a proportion–for the largest detection limit. The code uses the `binom.test` rather than `prop.test` because it computes the exact test statistic. Note that this example assumes no missing values; if necessary they should be removed before the analysis.

```
> # What is the maximum detection limit?
> Xlc.max <- max(censorLevels(Xlc))
> Xlc.max
```

```
[1] 1.2
```

```
> # What percentage?
> percentile(Xlc, 1.2)
```

```
Percent >= 1.2
           50
```

```
> # Compute the confidence interval
> binom.test(sum(Xlc >= Xlc.max),
+         length(Xlc), percentile(Xlc, 1.2, percent=FALSE))
```

```
        Exact binomial test

data:  sum(Xlc >= Xlc.max) and length(Xlc)
number of successes = 13, number of trials = 26, p-value = 1
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.2992722 0.7007278
sample estimates:
probability of success
                   0.5
```

3 Analysis of Blanks

Blanks are used to estimate the positive bias that can be caused by extraneous contamination introduced into environmental samples during collection, processing, shipment, and laboratory analysis. Evaluation of data from field blanks depends on the inference space represented by the blanks. In general, there are two possibilities: (1) a single blank is prepared to represent potential sources of contamination that affect a specific, small set of environmental samples, or (2) multiple blanks are prepared periodically over time and space to represent potential sources of contamination that might affect a much larger set of environmental samples.

This example presents a technique for evaluating potential contamination by comparing the distribution of atrazine in environmental samples to the distribution of the 90-percent upper confidence limit for percentiles of concentration in associated field blanks (Mueller and others (2015). Out of a total of 637 data points for atrazine concentrations in field blanks associated with surface-water sampling across the United States, 630 are censored at a consistent DL of 0.004 μg/L. The lowest concentration of the 7 quantified values is 0.0035 μg/L, which is less than the DL (atrazine is an information-rich analyte for samples analyzed at the NWQL). The last 10 rows of the example dataset of atrazine field blank results are shown below. The dataset is ordered numerically by RESULT with censored results ahead of uncensored results.

4 Analysis of Spikes

Spikes are used to estimate the positive or negative bias that can affect the measured results for environmental samples because of analyte degradation or problems with the analytical methods. This bias is estimated by determining the recovery of known concentrations of the analytes in the spiked sample. Calculation of recovery for matrix spikes requires a separate environmental sample to determine the background concentration of the analyte in the unspiked matrix. Recovery in the spiked matrix samples can be compared to some criteria or to typical recovery for the analytical method based on laboratory reagent spikes.

5 Analysis of Replicates

Replicates are used to measure variability, which is defined as the random error in independent measurements as the result of repeated application of the measurement process under identical conditions. Statistical evaluation of replicate variability is based on the standard deviation of measured values in the primary environmental sample and the replicate sample or samples. If only one set of a large number of replicates was collected, the standard deviation could be calculated directly; however, the general practice is to collect many sets of a small number of replicates under different conditions.

References

[1] Helsel, D.R. 2012, Statistics for Censored Environmental Data Using Minitab and R: New York, Wiley, 324 p.

[2] Lorenz, D.L., 2016, smwrQW–an R package for managing and analyzing water-quality data, version 1.0.0: U.S. Geological Survey Open File Report 2016-XXXX.

[3] Mueller, D.K., Schertz, T.L., Martin, J.D., and Sandstrom, M.W., 2015, Design, analysis, and interpretation of field quality-control data for water-sampling projects: U.S. Geological Survey Techniques and Methods book 4, chap. C4, 54 p.

Glossary

censored data Any data containing or potentially containing censored values.

left-censored data Any data containing left-censored values.

multiply-censored data Any data containing right- or interval-censored values or a mixture of any types of censored values.

uncensored data Any data not containing any censored values.