Confidence Intervals

Dave Lorenz

January 20, 2016

Abstract

These examples demonstrate some of the functions and statistical methods for computing confidence intervals for the mean and quantiles that are available in the `smwrQW` package.

Contents

1 Introduction

The examples in this vignette use the Golden dataset from the NADA package. The examples in this vignette use the function as.lcens to convert those data to a form used by the functions demonstrated. That conversion is not necessary for data of class "qw" or numeric data. Only the liver data in the Golden dataset are used and the data are subsetted to represent samples from a single population—the low dose.

```
> # Load the smwrQW package
> library(smwrQW)
> # And the data
> data(Golden, package="NADA")
> # Convert and subset the data
> Golden <- with(subset(Golden, DosageGroup ==  "Low"),
+   data.frame(Liver=as.lcens(Liver, censor.codes=LiverCen)))
```

2 Confidence Intervals for the Mean

Confidence intervals for the mean of an uncensored, normally distributed sample can easily be calculated using the one-sample `t.test` function. The confidence interval is set using the `conf.level` argument and the `alternative` argument controls whether two-sided, upper- or lower-confidence limits are reported. The `censMean.CI` function in the `smwrQW` package can be used to computed confidence intervals for the mean of censored, normally or lognormally distributed sample. Helsel (2012) describes the methods for maximum likelihood and the robust regression on order statistics. The methods for maximum likelihood are extended to adjusted maximum likelihood for the functions in `smwrQW`.

The example below illustrates each of the three methods for computing the confidence interval for the mean, "log AMLE," "log MLE," and "log ROS." The default method is "log AMLE" because maximum likelihood methods produce estimates of the standard deviation that are biased low, producing confidence interval that are too small. The methods "log ROS" and "ROS" use bootstrapping to compute confidence intervals and setting the seed for the random number generator will produce reproducible results. The results are reported as the estimate of the mean, the lower and upper confidence limits and the confidence interval. The assumption of lognormality is important because of issues related to the back-transformation and is simply assessed using the `censPPP.test` on the log-transformed data; the data could also be plotted using the `qqPlot` function, setting the `yaxis.log` argument to `TRUE`.

```
> # first assess the log-normality
> censPPCC.test(log(Golden$Liver))

        PPCC Normality Test for multiple detection limit

data:  log(Golden$Liver)
r(29 percent censored) = 0.96558, p-value = 0.4685
alternative hypothesis: Data are not from a normal distribution
null hypothesis: Data are from a normal distribution

> # compute the 90 percent (default) two-sided confidence intervals
> # The default log AMLE method
> censMean.CI(Golden$Liver)

      estimate        lcl       ucl  ci
mean 0.1217712 0.06753856 0.2538732 0.9

> # The log MLE method
> censMean.CI(Golden$Liver, method="log MLE")

      estimate        lcl       ucl  ci
mean 0.1254868 0.06689913 0.2353835 0.9

> # The log ROS method
> set.seed(123)
> censMean.CI(Golden$Liver, method="log ROS")

      estimate        lcl       ucl  ci
mean 0.1213158 0.07326139 0.1786908 0.9
```

3 Confidence Intervals for Quantiles

Helsel (2012) describes computing confidence intervals of the median and other quantiles for complete data using the properties of the binomial distribution. That method is coded as `qtiles.CI` in the `smwrStats` package (Lorenz 2016). He also describes several methods for computing those confidence intervals for left-censored data using nonparametric methods from survival analysis. The B-C method is coded as `censQtiles.CI` in the `smwrQW` package.

The example below illustrates computing the 90 percent two-sided and 95 percent upper confidence interval for median and 75th percentile for the Golden data used in the previous example. Those selected confidence interval should produce the same value for the upper confidence limit. For the two-sided or lower confidence level, `NA` is returned if the computed value is less than the minimum value of the data. In that case, the lower limit should be interpret as interval censored between 0 and the minimum, reported in the "minimum" attribute of the returned matrix. For the two-sided or upper confidence level, the maximum value in the dataset is returned even if the computed upper limit exceeds that value. The user must determine whether to accept the maximum or report the value as right censored, greater than that maximum. An example is shown in the R code below.

```
> # The two-sided confidence intervals
> censQtiles.CI(Golden$Liver, probs=c(.25, .5, .75))

      estimate          lcl        ucl  ci
25% 0.03930085           NA 0.05955189 0.9
50% 0.07923988   0.03930085 0.13614973 0.9
75% 0.17483254   0.09892787 0.30164120 0.9
attr(,"minimum")
[1] 0.03724138
attr(,"maximum")
[1] 0.3279693


> # and the upper confidence interval
> censQtiles.CI(Golden$Liver, probs=c(.25, .5, .75), bound="upper", CI=0.95)

      estimate lcl        ucl   ci
25% 0.03930085   0 0.05955189 0.95
50% 0.07923988   0 0.13614973 0.95
75% 0.17483254   0 0.30164120 0.95
attr(,"minimum")
[1] 0.03724138
attr(,"maximum")
[1] 0.3279693


> # Example where the maximum value is reported for the upper limit
> censQtiles.CI(Golden$Liver, probs=c(.9), bound="upper", CI=0.95)

     estimate lcl       ucl   ci
90% 0.3016412   0 0.3279693 0.95
attr(,"minimum")
[1] 0.03724138
attr(,"maximum")
[1] 0.3279693
```

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, smwrStats–An R Package for Analyzing Hydrologic Data, Version 0.7.4: U.S. Geological Survey Open File Report 2016-XXXX.