Computing Summary Statistics

Dave Lorenz

January 20, 2016

Abstract

These examples demonstrate some of the functions and statistical methods for computing summary statistics that are available in the `smwrQW` package.

Contents

1 Introduction

The examples in this vignette use data 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; the class "lcens" is most appropriate for these data as they are only left-censored and have only the value and an indicator of censoring. The functions demonstrated in these examples will also accept data of class "qw." The R code following this paragraph gets the data and converts the column named "As" to class "lcens."

Helsel (2012) reviews methods and makes some recommendations regarding the choice of methods. Lorenz (2016) discusses the performance of the methods specific to the functions in the `smwrQW` package. Details of the computation of the methods are in Helsel (2012). Details pertaining to how censored data are handled in these functions are in Lorenz (2016).

```
> # Load the smwrQW package
> library(smwrQW)
> # And the data
> data(Oahu, package="NADA")
> # Convert the data
> Oahu <- transform(Oahu, As=as.lcens(As, censor.codes=AsCen))
```

2 Maximum Likelihood Estimation

Maximum likelihood estimation will estimate the most likely parameters of an assumed distribution, like the mean and standard deviation of a normal distribution, to observed data.

The example below illustrates the computation of the mean and standard deviation using the "log MLE" method for arsenic data in the Oahu dataset. The example uses the `censStats` function to perform the computation, which prints the output in a fairly readable form. The quantiles are also computed and printed. The `quantile` function for censored data and the `censQuantile` function (not shown in the example), impute values for the censored data, much like regression on order statistics to compute quantiles that are more representative of the sample data than simply computing the quantiles of the computed distribution. That behavior replicates other U.S. Geological Survey software based on Helsel and Cohn (1988).

```
> # The mean and standard deviation.
> with(Oahu, censStats(As, method="log MLE"))

     mean std. dev.
   0.9453    0.6559
Statistics for the log transforms:
     mean std. dev.
  -0.2528    0.6269

> # And the quantiles
> with(Oahu, quantile(As, method="log MLE"))

       0%       25%       50%       75%      100%
0.3036237 0.5247952 0.7000000 1.0934697 3.2000000
```

The `smwrQW` functions also include adjusted maximum likelihood estimates (Helsel and Cohn, 1988). The adjusted maximum likelihood estimate eliminates the first-order bias in maximum likelihood estimates for the parameters of the normal or log-normal distribution. The example below demonstrates it use on the Oahu data. Quantiles are not computed using "AMLE."

```
> # The mean and standard deviation.
> with(Oahu, censStats(As, method="log AMLE"))

     mean std. dev.
   0.9376    0.6435
Statistics for the log transforms:
     mean std. dev.
  -0.2528    0.6404
```

3 The Nonparametric Method

The nonparametric method employs techniques from survival analysis. That method is often referred to as the Kaplan-Meier method using flipped data. Flipping the data transformed left-censored data to right-censored data, which is common in the field of survival analysis and on which the Kaplan-Meier method is based. That method is called "flipped K-M" in the `smwrQW` functions. The example below illustrates the `censStats` and the `quantile` functions using the "flipped K-M" method.

```
> # The mean and standard deviation.
> with(Oahu, censStats(As, method="flipped K-M"))

     mean std. dev.
    0.949    0.8068

> # And the quantiles
> with(Oahu, quantile(As, method="flipped K-M"))

  0%  25%  50%  75% 100%
 0.5  0.5  0.7  0.9  3.2
```

The arsenic data in the Oahu dataset are unusual in that there are quantified values less than the minimum reporting level—the "flipped K-M" method distributes the left-censored values among those lowest values. For most water-quality data, there will be no quantified values less than the minimum reporting level. In that case, the `censStats` function will report the mean and standard deviation as interval censored values, the `quantile` function will will indicate the incomplete quantiles by appending an "*" after the percentile name; the `censQuantile` function can be used to compute interval censored values for the incomplete quantiles. The example below modifies the Oahu dataset by changing one of the uncensored 0.5 values to a less than value and demonstrates the `censStats`, `quantile`, and `censQuantile` functions. Note that this is a special modification for water-quality data, where the true value cannot be less than 0 and is not appropriate for truly left-censored data.

```
> # Modify the Oahu dataset, censoring the first 0.5 value
> OahuX <- Oahu
> OahuX$AsCen[which(OahuX$As == 0.5)[1]] <- TRUE
> # The mean and standard deviation.
> with(OahuX, censStats(As, method="flipped K-M"))

     mean std. dev.
    0.949    0.8068

> # And the quantiles
> with(OahuX, quantile(As, method="flipped K-M"))

  0%  25%  50%  75% 100%
 0.5  0.5  0.7  0.9  3.2

> with(OahuX, censQuantile(As, method="flipped K-M"))
```

```
0%: 0.5
25%: 0.5
50%: 0.7
75%: 0.9
100%: 3.2
```

4 Regression on Order Statistics

Regression on order statistics (ROS) computes the parameters of the normal distribution (mean and standard deviation) using ordinary least squares regression on the data displayed in a q-normal plot. See the "The Q-Q and Q-normal Plots" section in the "Plotting Censored Data" vignette for an example. The default value for the `alpha` value for the plotting position is 0.4 rather than the original 0. Lorenz and others (2011) reported that an `alpha` value of 0 results in a high bias for standard deviation when the percentage of censoring is small. The value of 0.4 was selected to be consistent with the other functions in `smwrQW`. The example below illustrates the `censStats` and the `quantile` functions using the "log ROS" method with the default value for `alpha`.

```
> # The mean and standard deviation.
> with(Oahu, censStats(As, method="log ROS"))

     mean std. dev.
   0.9854    0.7217
Statistics for the log transforms:
     mean std. dev.
  -0.2114    0.6057

> # And the quantiles
> with(Oahu, quantile(As, method="log ROS"))

       0%       25%       50%       75%      100%
0.3266900 0.5397979 0.7000000 1.1265899 3.2000000
```

References

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

[2] Helsel, D.R. and Cohn, T.A., 1988, Estimation of descriptive statistics for multiply censored water quality data: Water Resources Research v. 24, n. 12, p.1997–2004

[3] Helsel, D.R., and Hirsch, R.M., 2002, Statistical methods in water resources: U.S. Geological Survey Techniques of Water-Resources Investigations, book 4, chap. A3, 522 p.

[4] 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.

[5] Lorenz, D.L., Ahearn, E.A., Carter, J.M., Cohn, T.A., Danchuk, W.J., Frey, J.W., Helsel, D.R., Lee, K.E., Leeth, D.C., Martin, J.D., McGuire, V.L., Neitzert, K.M., Robertson, D.M., Slack, J.R., Starn, J., Vecchia, A.V., Wilkison, D.H., and Williamson, J.E., 2011, USGS library for S-PLUS for Windows—Release 4.0: U.S. Geological Survey Open-File Report 2011-1130.