Comparing Three or More Groups

Dave Lorenz

July 26, 2017

Abstract

These examples demonstrate some of the functions and statistical methods for comparing three or more groups that are available in the `smwrQW` package.

Contents

1 Introduction

The examples in this vignette use the TSE 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; 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 creates a column named "TCE" of class "lcens." With the exception of the binary method, only censored data techniques are included in this vignette. Techniques that apply to single reporting limits and can require recensoring and simple substitution are not included, as the censored techniques can be used directly by the functions in `smwrQW`.

```
> # Load the smwrQW package
> library(smwrQW)
> # And the data
> data(TCE, package="NADA")
> # Convert the data to column TCE
> TCE <- transform(TCE, TCE=as.lcens(TCEConc, censor.codes=TCECen))
```

2 Binary Method

The binary method simply recodes values as 0 or 1 depending on whether the value is less than or greater than or equal to a specified criterion. The recoded values can then be tabulated and tested for the equality of proportions.

The example below illustrates the recoding of values and used the `prop.test` to test for the equality of proportions between the "Low," "Medium," and "High" population density residential land use. The `code01` function returns a data frame of values with missing values removed from the input arguments. The data must be tabulated with the rows representing the groups, the first argument to `table` and the columns the counts of the 0/1 data. The printed output from the `prop.test` indicates that all of the Low density land use are 0, with decreasing percentages for Medium and High density land uses. The p-value, 0.009864, indicates that the null hypothesis of equal proportions should be rejected at the 0.05 significance level.

```
> # Create Density as a factor ordered Low-Medium-High
> TCE <- transform(TCE, Density=factor(Density, levels=c("Low", "Medium", "High")))
> # Append a column of 0/1 values to the data
> TCE <- cbind(TCE, with(TCE, code01(TCE01=TCE)))
> # Tabulate the 0/1 data and print it
> TCETbl <- with(TCE, table(Density, TCE01))
> print(TCETbl)

        TCE01
Density    0   1
  Low     25   0
  Medium 118  12
  High    74  18

> # And the test
> prop.test(TCETbl)

        3-sample test for equality of proportions without
        continuity correction

data:  TCETbl
X-squared = 9.2376, df = 2, p-value = 0.009864
alternative hypothesis: two.sided
sample estimates:
   prop 1    prop 2    prop 3
1.0000000 0.9076923 0.8043478
```

3 Maximum Likelihood Estimation Method

Comparing multiple groups using maximum likelihood estimation method extends censored regression as ANOVA extends ordinary least squares–the test is constructed by relating a censored response variable to a factor.

An important first step in any parametric statistical analysis is to plot the data. Figure 1 shows 2 graphs, the log-transformed q-normal plot for Medium and High density land use. The Low density graph is not shown because there are very few uncensored values. Only the log-transformed is shown because the log-transformed is the most likely to produce nearly normal distributions.

```
> setSweave("graph01", 3 ,6)
> # Set layout for 2 graphs
> AA.lo <- setLayout(num.cols=2, num.rows=1)
> # Create the graphs
> setGraph(1, AA.lo)
> with(subset(TCE, Density=="Medium"), qqPlot(TCE,
+   ytitle="TCE in Medium", yaxis.log=TRUE))
> setGraph(2, AA.lo)
> with(subset(TCE, Density=="High"), qqPlot(TCE,
+   ytitle="TCE in High", yaxis.log=TRUE))
> graphics.off()
```

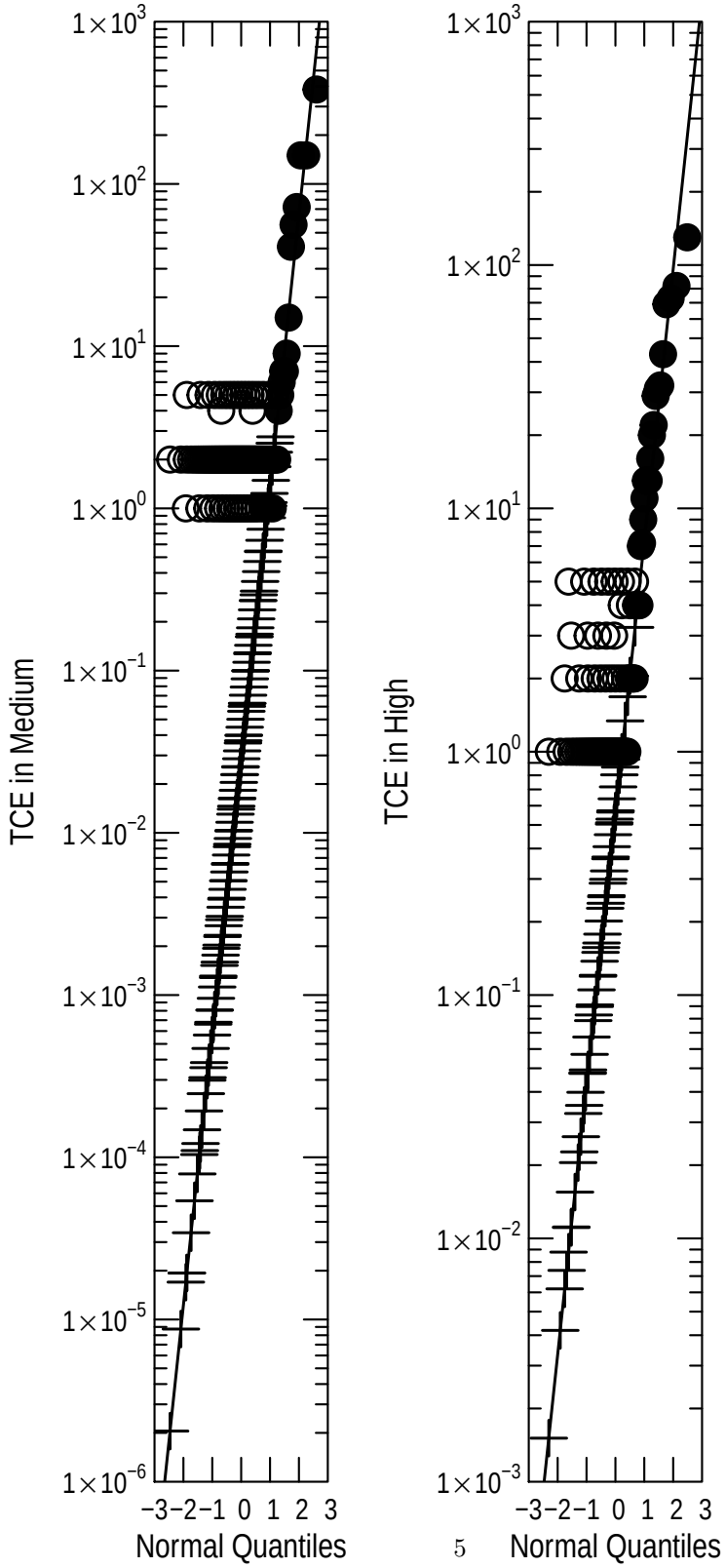
TCE in Medium
1×10^3
1×10^2
1×10^1
1×10^0
1×10^−1
1×10^−2
1×10^−3
1×10^−4
1×10^−5
1×10^−6
−3 −2 −1 0 1 2 3
Normal Quantiles
TCE in High
1×10^3
1×10^2
1×10^1
1×10^0
1×10^−1
1×10^−2
1×10^−3
−3 −2 −1 0 1 2 3
Normal Quantiles

Figure 1. Q-normal plots to check log-normal distribution assumption.

The `censReg` function is used for regression and comparing groups. It functions much like any modeling function, like `lm` in R—it constructs the model from a formula and data and has other options similar to `lm`. Its use for the censored equivalent of the two-sample t-test is shown below. Because the data are only left-censored, it uses adjusted maximum likelihood estimation (AMLE), which eliminates first-order bias from the maximum likelihood estimate. The p-value of the overall test result is 0.0005, suggesting the the null hypothesis of no difference among the three land use densities should be rejected.

```
> # The ANOVA analogue test:
> censReg(TCE ~ Density, data=TCE, dist="lognormal")

Call:
censReg(formula = TCE ~ Density, data = TCE, dist = "lognormal")

Coefficients:
              Estimate Std. Error z-score p-value
(Intercept)     -3.513      1.174 -2.9917  0.0000
DensityMedium    1.133      1.141  0.9927  0.1879
DensityHigh      2.790      1.161  2.4029  0.0022

Estimated residual standard error (Unbiased) = 2.867
Distribution: lognormal
Percent standard error: 6090
Positive percent error: 1658
Negative percent error: -94.31

Number of observations = 247, number censored = 194 (78.5 percent)

Loglik(model) = -197.8 Loglik(intercept only) = -205.5
  Chi-square = 15.41, degrees of freedom = 2, p-value = 0.0005

Computation method: AMLE
```

4 Nonparametric Methods

The nonparametric method in the `smwrQW` package for comparing three or more groups is a test that compares the flipped survival curves. The details are described by Helsel (2012).

The test that compares flipped survival curves can be used for two or more groups and is executed by the `censKSample.test`. There are two types of the test, "Peto" and "log-rank"; both are described by Helsel (2012). For these data, the both types return p-value substantially less than 0.05, suggesting that the null hypothesis of no difference be rejected.

```
> # The Peto type two-sample test
> with(TCE, censKSample.test(TCE, Density, type="Peto"))

        Left-censored k sample test

data:  TCE by Density
Peto & Peto chi-square = 16.255, df = 2, p-value =
0.0002953
alternative hypothesis: two.sided

> # The log-rank type two-sample test
> with(TCE, censKSample.test(TCE, Density, type="log-rank"))

        Left-censored k sample test

data:  TCE by Density
log-rank chi-square = 16.28, df = 2, p-value = 0.0002917
alternative hypothesis: two.sided
```

5 Multiple Comparison Tests

Multiple comparison tests for censored data are preformed by the `censMulticomp.test` in the `smwrQW` package. It runs repeated generalized Wilcoxon tests among all of the groups and uses `p.adjust` to modify the p-value to account for the multiple comparisons. The details are described by Helsel (2012).

The `censMulticomp.test` is demonstrated in the code below, using the default method for adjusting the p-values ("holm"). The table of paired comparisons in the print out indicates that Low density is not significantly different from Medium density and that High density is significantly different from both Low and Medium density land use areas.

```
> # The Peto type two-sample test
> with(TCE, censMulticomp.test(TCE, Density))


         Nonparametric left-censored multicomparison test
Overall error rate: 0.05
Attained P-values adjusted by the holm method

Response variable: TCE
Group variable: Density

Table of paired comparisons, attained p-values less than 0.05 are flagged by '*'
              Zscore  P.adjusted flag
Low - Medium  -1.1555  0.2479
Low - High    -2.8434  0.0089    *
Medium - High -3.3180  0.0027    *
```

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.