Comparing Two Groups

Dave Lorenz

January 20, 2016

Abstract

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

Contents

1 Introduction

The examples in this vignette use the CuZn 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 converts the column named "Zn" to 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(CuZn, package="NADA")
> # Convert the Zn data, Cu is not used
> CuZn <- transform(CuZn, Zn=as.lcens(Zn, censor.codes=ZnCen))
```

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 "AlluvialFan" and "BasinTrough" geologic zones. The `code01` function returns a data frame of values with missing values removed from the input arguments. The first step is to subset the CuZn dataset to remove the Cu data and the missing values. The [[1]] extraction from the returned value from the `code01` function simply converts the data to a vector of 0/1 values rather than a data frame. 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 about 29.85 percent of the AlluvailFan zone are 0, less than the largest reporting level and 24 percent of the BasinTrough zone are 0. The p-value, 0.6222, indicates that the null hypothesis of equal proportions cannot be rejected at the 0.05 significance level.

```
> # Subset the data to remove the Cu (columns 1 and 2) and the missing values
> ZnData <- na.omit(CuZn[, 3:5])
> # Add a column of 0/1 values
> ZnData <- transform(ZnData, Zn01=code01(Zn)[[1]])
> # Tabulate the 0/1 data and print it
> ZnTbl <- with(ZnData, table(Zone, Zn01))
> print(ZnTbl)

             Zn01
Zone           0  1
  AlluvialFan 20 47
  BasinTrough 12 38

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

        2-sample test for equality of proportions with continuity correction

data:  ZnTbl
X-squared = 0.24276, df = 1, p-value = 0.6222
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.1202614  0.2372763
sample estimates:
   prop 1    prop 2
0.2985075 0.2400000
```

3 Maximum Likelihood Estimation Method

Comparing two groups using maximum likelihood estimation method extends censored regression as the t-test extends ordinary least squares–the test is constructed by relating a censored response variable to a 2-level factor.

An important first step in any parametric statistical analysis is to plot the data. Figure 1 shows 4 graphs, two for each zone, one log-transformed and one untransformed. It is clear that there is no reason to pursue the untransformed approach and the log-transformed approach looks reasonable, but the AlluvialFan Zinc does have one very large outlying value. The coefficient for ZoneBasinTrough is 0.2575, measured in natural log units and represents the difference between the AlluvailFan and the BasinTrough Concentrations of zinc. The p-value, 0.1105, indicates that the null hypothesis of equal means cannot be rejected at the 0.05 significance level.

```
> setSweave("graph01", 6 ,6)
> # Set layout for 2 graphs
> AA.lo <- setLayout(num.cols=2, num.rows=2)
> # Create the graphs
> setGraph(1, AA.lo)
> with(subset(CuZn, Zone=="AlluvialFan"), qqPlot(Zn,
+   ytitle="Alluvial Zinc", yaxis.log=FALSE))
> setGraph(2, AA.lo)
> with(subset(CuZn, Zone=="BasinTrough"), qqPlot(Zn,
+   ytitle="BasinTrough Zinc", yaxis.log=FALSE))
> setGraph(3, AA.lo)
> with(subset(CuZn, Zone=="AlluvialFan"), qqPlot(Zn,
+   ytitle="Alluvial Zinc", yaxis.log=TRUE))
> setGraph(4, AA.lo)
> with(subset(CuZn, Zone=="BasinTrough"), qqPlot(Zn,
+   ytitle="BasinTrough Zinc", yaxis.log=TRUE))
> graphics.off()
```

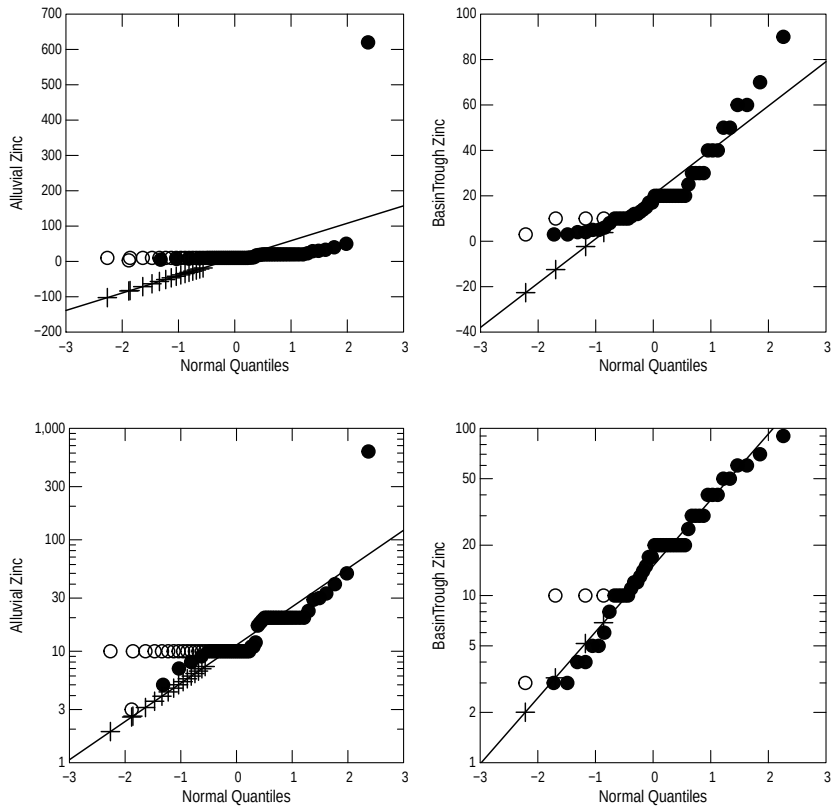

Figure 1. Q-normal plots to check log-normal distribution assumption.

The `censReg` function is used for regression and comparing two 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 two-sample test:
> censReg(Zn ~ Zone, data=CuZn, dist="lognormal")


Call:
censReg(formula = Zn ~ Zone, data = CuZn, dist = "lognormal")

Coefficients:
                Estimate Std. Error z-score p-value
(Intercept)       2.4669     0.1137  21.698  0.0000
ZoneBasinTrough   0.2575     0.1661   1.551  0.1105

Estimated residual standard error (Unbiased) = 0.849
```

```
Distribution: lognormal
Percent standard error: 102.8
Positive percent error: 133.7
Negative percent error: -57.22

Number of observations = 117, number censored = 20 (17.1 percent)

Loglik(model) = -138 Loglik(intercept only) = -139.2
  Chi-square = 2.547, degrees of freedom = 1, p-value = 0.1105

Computation method: AMLE
```

4 Nonparametric Methods

Two nonparametric methods are available in the `smwrQW` package for comparing two groups—the generalized Wilcoxon test and a test that compares the flipped survival curves. The details for both are described by Helsel (2012).

The generalized Wilcoxon test is illustrated in the example below using the CuZn data. There are two options for the method used to compute the test, "gehan" and "peto." The "gehan" method computes the Gehan test statistic and "peto" computes the Peto-Prentice or Peto-Peto test statistic; both are described by Helsel (2012). The default method selects "gehan" if the data are right- or multiply-censored and "peto" otherwise. Both methods return p-values less than the 0.05 alpha level and reject the null hypothesis of no difference. Not shown in this example, but data with a single reporting level could use the Wilcoxon rank-sum test using simple substitution for the censored values.

```
> # The Gehan two-sample test:
> with(CuZn, genWilcox.test(Zn, Zone, method="gehan"))


        Gehan generalized Wilcoxon test

data:  AlluvialFan and BasinTrough
Gehan Z = -2.3037, n = 67, m = 50, p-value = 0.02124
alternative hypothesis: true difference is not equal to 0
sample estimates:
                   n events median
group=AlluvialFan 67     51   10.0
group=BasinTrough 50     46   18.5

> # The Peto-Prentice two-sample test:
> with(CuZn, genWilcox.test(Zn, Zone, method="peto"))


        Peto-Prentice generalized Wilcoxon test

data:  AlluvialFan and BasinTrough
Peto-Prentice Z = -2.2269, n = 67, m = 50, p-value = 0.02596
alternative hypothesis: true difference is not equal to 0
sample estimates:
                   n events median
group=AlluvialFan 67     51   10.0
group=BasinTrough 50     46   18.5
```

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 "Peto" type returns a p-value less than 0.05 and the "log-rank" type returns a p-value greater than 0.05, thus giving conflicting conclusions.

```
> # The Peto type two-sample test
> with(CuZn, censKSample.test(Zn, Zone, type="Peto"))
```

```
        Left-censored k sample test

data:  Zn by Zone
Peto & Peto chi-square = 5.1835, df = 1, p-value = 0.0228
alternative hypothesis: two.sided


> # The log-rank type two-sample test
> with(CuZn, censKSample.test(Zn, Zone, type="log-rank"))


        Left-censored k sample test

data:  Zn by Zone
log-rank chi-square = 2.8426, df = 1, p-value = 0.0918
alternative hypothesis: two.sided
```

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.