Comparing Groups Example

Dave Lorenz

May 16, 2014

This example demonstrates some functions in the `USGSwsStats` package that facilitate comparing multiple groups. The example uses the subset of Knopman (1990) dataset C7 from Helsel and Hirsch (2002) to complement the analysis done in sections 7.1 and 7.4 in their book. All section references in this example are from Helsel and Hirsch (2002). The critical alpha value for all tests is 0.05.

```
> # Load the stats, USGSwsData, and USGSwsStats packages
> library(stats)
> library(USGSwsData)
> library(USGSwsStats)
> # Get the dataset
> data(AppalachianSpecCap)
```

1 Plot the Data

Helsel and Hirsch (2002) stress plotting the data to help undserstand the data and find an appropriate statistical technique. For these data, a box plot by group should indicate whether a parametric analysis can be used. Figure 1 replicates figure 7.14 from Helsel and Hirsch (2002) and indicates that the parametric analysis can be used, as well as the nonparametric analysis.

```
> # setSweave is a specialized function that sets up the graphics page for
> # Sweave scripts. For interactive use, it should be removed and the
> # default setting for set.up can be used.
> setSweave("CG01", 5, 5)
> with(AppalachianSpecCap, boxPlot(LogSpecCap, group=RockType,
+
> # Required call to close PDF output graphics
> graphics.off()
```

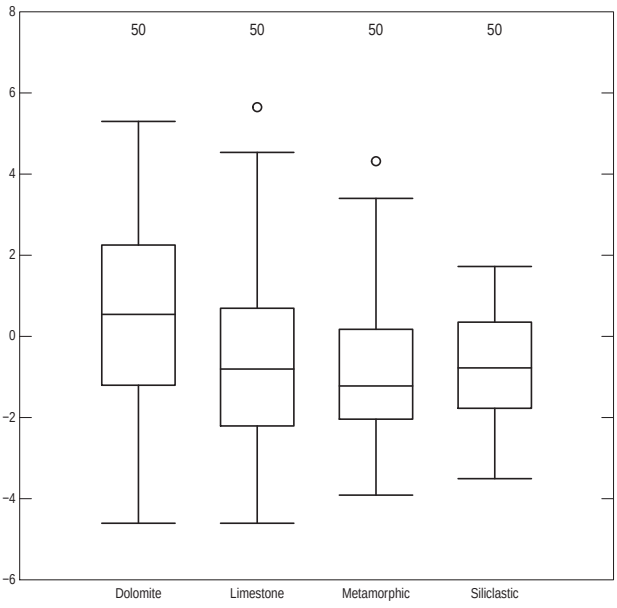

Figure 1. The box plot of log specific capacities by rock type.

2 Nonparametric Analysis

The `kruskal.test` function (stats package) performs the Kruskal-Wallis test decribed in section 7.1.1.

```
> # Perform the analysis using a formula
> kruskal.test(LogSpecCap ~ RockType, data=AppalachianSpecCap)

        Kruskal-Wallis rank sum test

data:  LogSpecCap by RockType
Kruskal-Wallis chi-squared = 11.544, df = 3, p-value = 0.00912
```

The attained p-value for the test is 0.00912, which is less than the critical alpha level set for the test, so one can proceed with a multiple comparison test to identify which rock types are different from one another. The `multicomp.test` in the USGSwsStats package can be used to identify similar groups using either parametric or nonparamteric methods. The nonparametric methods are closely related to the Krusak-Wallis test, in that the known statistics of ranks are used and not simply using a rank transform and parametric methods for computing the critical value separating groups. The code below demonstrates how to use `multicomp.test` to indentify similar groups using the nonparametric method. These nonparametric methods were developed after the original version of Helsel and Hirsh, from section 7.4.2.

```
> # Perform the MCT using the default Tukey method for determining the
> # critical value for separating groups.
> with(AppalachianSpecCap, multicomp.test(LogSpecCap, RockType,
+

         Nonparametric Multiple Comparison Test
Overall error rate: 0.05
Critical value: 2.569 by the tukey method

Response variable: Rank of LogSpecCap
Group variable: RockType

Table of paired comparisons, 95 percent confidence intervals
 excluding 0 are flagged by *.
                        estimate stderr   lower    upper    flag
Dolomite-Siliclastic     29.0500  11.5700  -0.6857  58.7900
Dolomite-Limestone       29.4400  11.5700  -0.2957  59.1800
Dolomite-Metamorphic     35.9500  11.5700   6.2140  65.6900 *
Siliclastic-Limestone     0.3900  11.5700 -29.3500  30.1300
Siliclastic-Metamorphic   6.9000  11.5700 -22.8400  36.6400
Limestone-Metamorphic     6.5100  11.5700 -23.2300  36.2500

Table of associations among groups
```

	Mean	Size	A	B
Dolomite	124.1	50	X	
Siliclastic	95.06	50	X	X
Limestone	94.67	50	X	X
Metamorphic	88.16	50		X

The critical value computed for the rank data using Tukey's method is 2.569. Groups are different if the difference in the mean rank (estimate) is is significantly different from 0. The lower and upper confidence interval are computed by subtracting and adding the stderr times the critrical value from the difference. The table of paired comparisons indicates that only Dolomite is different from Metamorhic. The table of associations reflects that by forming two association, one excluding Dolomite and the other excluding Metamorphic. Because Dolomite is more different from the others than Metamorphic, Group B is the more likely association.

3 Parametric Analysis

The `oneway.test` function (stats package) performs the Kruskal-Wallis test decribed in section 7.1.2. It is more striaghtforward than the `aov` function (also in the stats package) because it presents the results as a simple hypothesis test and it adjusts for unequal variances, much like `t.test`.

```
> # Perform the analysis using a formula
> oneway.test(LogSpecCap ~ RockType, data=AppalachianSpecCap)

        One-way analysis of means (not assuming equal variances)

data:  LogSpecCap and RockType
F = 3.2291, num df = 3.000, denom df = 106.146, p-value = 0.0254
```

The attained p-value for the test is 0.0254, which is less than the critical alpha level set for the test, so one can proceed with a multiple comparison test to identify which rock types are different from one another. The `multicomp.test` in the USGSwsStats package can be used to identify similar groups. The parametric method uses the pooled variance and does not assume unequal variances. The code below demonstrates how to use `multicomp.test` to indentify similar groups using the parametric method.

```
> # Perform the MCT using the default Tukey method for determining the
> # critical value for separating groups.
> with(AppalachianSpecCap, multicomp.test(LogSpecCap, RockType,
+

         Parametric Multiple Comparison Test
Overall error rate: 0.05
Critical value: 2.5912 by the tukey method

Response variable: LogSpecCap
Group variable: RockType

Table of paired comparisons, 95 percent confidence intervals
 excluding 0 are flagged by *.
                       estimate stderr   lower    upper    flag
Dolomite-Limestone       1.09600  0.41460  0.02217  2.17100 *
Dolomite-Siliclastic     1.16600  0.41460  0.09208  2.24100 *
Dolomite-Metamorphic     1.30200  0.41460  0.22750  2.37600 *
Limestone-Siliclastic    0.06991  0.41460 -1.00400  1.14400
Limestone-Metamorphic    0.20530  0.41460 -0.86890  1.28000
Siliclastic-Metamorphic  0.13540  0.41460 -0.93880  1.21000

Table of associations among groups
            Mean    Size A B
```

```
Dolomite    0.4081  50   X
Limestone   -0.6883 50     X
Siliclastic -0.7582 50     X
Metamorphic -0.8936 50     X
```

The critical value computed for the rank data using Tukey's method is 2.5912. Groups are different if the difference in the mean (estimate) is is significantly different from 0. The lower and upper confidence interval are computed by subtracting and adding the stderr times the critrical value from the difference. The table of paired comparisons indicates that only Dolomite is different from all others. The table of associations reflects that by forming two association, one including only Dolomite and the other excluding Dolomite.

References

[1] 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.

[2] Knopman, D. S., 1990, Factors related to the water-yielding potential of rocks in the Piedmont and Valley and Ridge provinces of Pennsylvania: U.S. Geological Survey Water-Resources Investigations Report 90-4174, 52 p.