Long Problem Set 4

1. The data at this link (http://www.rsc.org/images/CO2_methods_tcm18-57755.txt) reports results for the determination of CO_2 by six different methods. The data itself uses Na_2CO_3 as a reference standard; presumably, a portion of the standard was treated to release the CO_2, which subsequently was determined and reported as % w/w CO_2 in the sample. Complete an analysis of variance on this data and identify sources of difference between the methods at $\alpha = 0.5$. The file carbonDioxide.RData contains vectors with the results for each method and a data frame that gives the concentration of CO_2 in the first column and the analytical method in the second column.

Answer. First, let's load the data and examine the contents of the data frame

```
CO2.df
```

```
##      CO2  method
## 1  41.41    grav
## 2  41.62    grav
## 3  41.48    grav
## 4  41.44    grav
## 5  41.50    grav
## 6  41.51    grav
## 7  41.43    grav
## 8  41.51    grav
## 9  41.59    grav
## 10 41.22 method1
## 11 41.41 method1
## 12 40.30 method1
## 13 40.34 method1
## 14 40.70 method1
## 15 41.16 method1
## 16 40.74 method1
## 17 40.23 method1
## 18 40.82 method1
## 19 40.93 method1
## 20 40.89 method1
## 21 40.91 method2
## 22 40.61 method2
## 23 40.02 method2
## 24 40.16 method2
## 25 40.54 method2
## 26 40.61 method2
## 27 40.61 method2
## 28 41.95 method3
## 29 41.31 method3
## 30 39.09 method3
## 31 40.59 method3
## 32 41.65 method3
## 33 42.90 method3
## 34 41.30 method3
## 35 40.30 Method4
## 36 40.46 Method4
## 37 39.50 Method4
## 38 39.53 Method4
## 39 39.95 Method4
```

```
## 40 40.80 method5
## 41 40.88 method5
## 42 41.24 method5
## 43 41.03 method5
## 44 41.53 method5
## 45 41.21 method5
## 46 40.96 method5
## 47 41.09 method5
```

As this is in the proper form for an analysis of variance, let's complete a one-way analysis of variance and examine the results

```
co2.aov = aov(CO2 ~ method, data = CO2.df)
summary(co2.aov)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## method       5  10.21  2.0414   7.145 6.92e-05 ***
## Residuals   41  11.71  0.2857
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

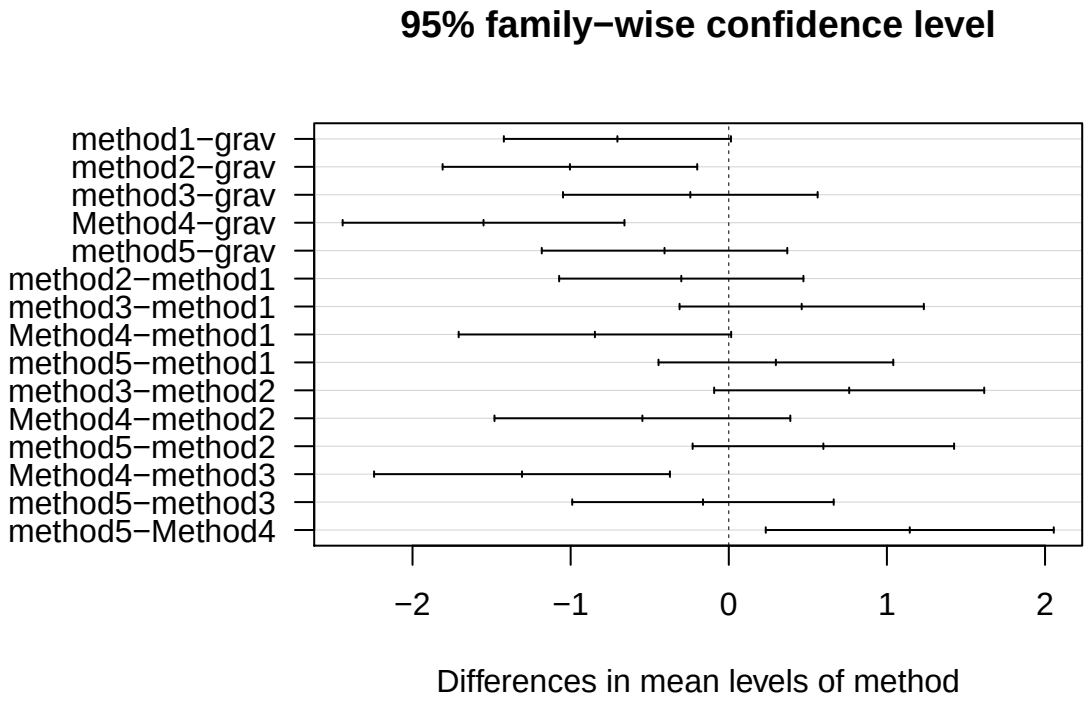
which, given the value for p, suggests that there are significant differences between the methods. To evaluate the sources of these significant differences, we use a TukeyHSD test

```
co2.hsd = TukeyHSD(co2.aov)
co2.hsd
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = CO2 ~ method, data = CO2.df)
##
## $method
##                        diff         lwr         upr     p adj
## method1-grav    -0.7043434 -1.42238620  0.01369933 0.0572738
## method2-grav    -1.0046032 -1.80968983 -0.19951652 0.0071747
## method3-grav    -0.2431746 -1.04826126  0.56191205 0.9436047
## Method4-grav    -1.5508889 -2.44195653 -0.65982125 0.0000817
## method5-grav    -0.4063889 -1.18265589  0.36987811 0.6256638
## method2-method1 -0.3002597 -1.07266247  0.47214299 0.8520417
## method3-method1  0.4611688 -0.31123390  1.23357156 0.4866358
## Method4-method1 -0.8465455 -1.70819678  0.01510587 0.0566511
## method5-method1  0.2979545 -0.44436069  1.04026978 0.8344809
## method3-method2  0.7614286 -0.09249478  1.61535192 0.1047391
## Method4-method2 -0.5462857 -1.48171188  0.38914045 0.5110025
## method5-method2  0.5982143 -0.22859344  1.42502201 0.2771764
## Method4-method3 -1.3077143 -2.24314045 -0.37228812 0.0019541
## method5-method3 -0.1632143 -0.99002201  0.66359344 0.9911619
## method5-Method4  1.1445000  0.23375962  2.05524038 0.0066598
```

Let's plot the results of the TukeyHSD test to make it easier to see where differences seem significant

```
old.par = par(mar = c(5, 8, 4, 2))
plot(co2.hsd, las = 1)
```




```
par(old.par)
```

Although we can see where there are differences between methods, there is no clear way to divide the methods into, say, one or two groups with distinctly different results; that is, none of the methods is an apparent "outlier" when compared to all other methods, although the gravimetric method and method 4 show the greatest number of significant differences (2 and 3 respectively; or 3 and 4, respectively, if we expand our value of α by a bit). If we rank the methods from the smallest-to-largest mean values,

```
mean(grav)
```

```
## [1] 41.49889
```

```
mean(method1)
```

```
## [1] 40.79455
```

```
mean(method2)
```

```
## [1] 40.49429
```

```
mean(method3)
```

```
## [1] 41.25571
```

```
mean(method4)
```

```
## [1] 39.948
```

```
mean(method5)
```

```
## [1] 41.0925
```

$$\text{method 4} < \text{method 2} < \text{method 1} < \text{method 5} < \text{method 3} < \text{gravimetric}$$

we are not surprised to find that the gravimetric method and method 4 show the greatest number of differences. This result is typical of data collected by different methods that, although subject to different biases, return

results that are not all that different from each other or at least not much larger than the random errors associated with each method.

2. One important use of an analysis of variance is the ability to use the results to classify samples (a topic to which we will return later in the semester). For example, the file Pottery.RData has five objects that give the concentrations of Al, Ca, Fe, Mg, and Na (as %w/w metal oxide) in pottery shards collected at four different sites, which are identified as the factors "A", "C", "I", and "L" in the object Site. If there are significant differences in the concentration of a metal in pottery shards collected at different sites, then these differences might serve as characteristic markers that can help classify pottery shards of unknown origin. Use a separate one-way analysis of variance for each metal, using $\alpha = 0.05$, and determine if there is a way you can determine the origin of a pottery shard based on the concentration of one or more of the metals. Your answer should explain clearly how you could identify whether a pottery shard is from site "A", "C", "I", or "L".

Answer. First, let's gather the data together into a data frame with the results for each metal in separate columns and the site in a column; note that we will not need to stack the data frame as we plan to consider the metals individually

```
df = data.frame(Al, Ca, Fe, Mg, Na, Site)
df
```

```
##       Al   Ca   Fe   Mg   Na Site
## 1   14.4 0.15 7.00 4.30 0.51    L
## 2   13.8 0.12 7.08 3.43 0.17    L
## 3   14.6 0.13 7.09 3.88 0.20    L
## 4   11.5 0.16 6.37 5.64 0.14    L
## 5   13.8 0.20 7.06 5.34 0.20    L
## 6   10.9 0.17 6.26 3.47 0.22    L
## 7   10.1 0.20 4.26 4.26 0.18    L
## 8   11.6 0.18 5.78 5.91 0.16    L
## 9   11.1 0.29 5.49 4.52 0.30    L
## 10  13.4 0.28 6.92 7.23 0.20    L
## 11  12.4 0.22 6.13 5.69 0.54    L
## 12  13.1 0.31 6.64 5.51 0.24    L
## 13  12.7 0.20 6.69 4.45 0.22    L
## 14  12.5 0.22 6.44 3.94 0.23    L
## 15  11.8 0.30 5.44 3.94 0.04    C
## 16  11.6 0.29 5.39 3.77 0.06    C
## 17  18.3 0.03 1.28 0.67 0.03    I
## 18  15.8 0.01 2.39 0.63 0.04    I
## 19  18.0 0.01 1.50 0.67 0.06    I
## 20  18.0 0.01 1.88 0.68 0.04    I
## 21  20.8 0.07 1.51 0.72 0.10    I
## 22  17.7 0.06 1.12 0.56 0.06    A
## 23  18.3 0.06 1.14 0.67 0.05    A
## 24  16.7 0.01 0.92 0.53 0.05    A
## 25  14.8 0.03 2.74 0.67 0.05    A
## 26  19.1 0.10 1.64 0.60 0.03    A
```

Next, let's complete a one-way analysis of variance on each metal using a linear model with the general form metal ~ Site.

```
al.aov = aov(Al ~ Site, data = df)
summary(al.aov)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Site         3 175.61   58.54   26.67 1.63e-07 ***
```

```
## Residuals   22  48.29    2.19
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
ca.aov = aov(Ca ~ Site, data = df)
summary(ca.aov)
```

```
##             Df  Sum Sq Mean Sq F value   Pr(>F)
## Site         3 0.20470 0.06823   29.16 7.55e-08 ***
## Residuals   22 0.05149 0.00234
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fe.aov = aov(Fe ~ Site, data = df)
summary(fe.aov)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Site         3 134.22   44.74   89.88 1.68e-12 ***
## Residuals   22  10.95    0.50
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mg.aov = aov(Mg ~ Site, data = df)
summary(mg.aov)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Site         3 103.35   34.45   49.12 6.45e-10 ***
## Residuals   22  15.43    0.70
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
na.aov = aov(Na ~ Site, data = df)
summary(na.aov)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Site         3 0.2582 0.08608   9.503 0.000321 ***
## Residuals   22 0.1993 0.00906
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For each metal, the p-value strongly suggests that the variance in metal concentrations between the sites is much greater than the variance in metal concentrations within each site. To evaluate possible significant differences between sites, we use the TukeyHSD test at $\alpha = 0.05$.

```
TukeyHSD(al.aov)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Al ~ Site, data = df)
##
## $Site
##           diff       lwr       upr     p adj
## C-A -5.6200000 -9.061978 -2.178022 0.0008775
## I-A  0.8600000 -1.741891  3.461891 0.7956408
## L-A -4.7557143 -6.899034 -2.612395 0.0000188
## I-C  6.4800000  3.038022  9.921978 0.0001671
## L-C  0.8642857 -2.245569  3.974140 0.8663153
```

```
## L-I -5.6157143 -7.759034 -3.472395 0.0000016
```

```
TukeyHSD(ca.aov)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Ca ~ Site, data = df)
##
## $Site
##            diff         lwr          upr     p adj
## C-A  0.24300000  0.13060908  0.355390922 0.0000270
## I-A -0.02600000 -0.11095955  0.058959551 0.8301234
## L-A  0.15014286  0.08015705  0.220128668 0.0000301
## I-C -0.26900000 -0.38139092 -0.156609078 0.0000063
## L-C -0.09285714 -0.19440323  0.008688944 0.0811915
## L-I  0.17614286  0.10615705  0.246128668 0.0000029
```

```
TukeyHSD(fe.aov)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Fe ~ Site, data = df)
##
## $Site
##           diff        lwr       upr     p adj
## C-A  3.9030000  2.2638764  5.542124 0.0000068
## I-A  0.2000000 -1.0390609  1.439061 0.9692779
## L-A  4.8601429  3.8394609  5.880825 0.0000000
## I-C -3.7030000 -5.3421236 -2.063876 0.0000146
## L-C  0.9571429 -0.5238182  2.438104 0.3023764
## L-I  4.6601429  3.6394609  5.680825 0.0000000
```

```
TukeyHSD(mg.aov)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Mg ~ Site, data = df)
##
## $Site
##           diff        lwr       upr     p adj
## C-A  3.2490000  1.3033471  5.194653 0.0006856
## I-A  0.0680000 -1.4027753  1.538775 0.9992199
## L-A  4.2204286  3.0088708  5.431986 0.0000000
## I-C -3.1810000 -5.1266529 -1.235347 0.0008651
## L-C  0.9714286 -0.7864842  2.729341 0.4349620
## L-I  4.1524286  2.9408708  5.363986 0.0000000
```

```
TukeyHSD(na.aov)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Na ~ Site, data = df)
##
```

```
## $Site
##          diff           lwr       upr     p adj
## C-A 0.0020000 -0.2191229048 0.2231229 0.9999941
## I-A 0.0060000 -0.1611532044 0.1731532 0.9996339
## L-A 0.2027143  0.0650210863 0.3404075 0.0025463
## I-C 0.0040000 -0.2171229048 0.2251229 0.9999530
## L-C 0.2007143  0.0009279937 0.4005006 0.0486676
## L-I 0.1967143  0.0590210863 0.3344075 0.0033933
```

There is a lot of output here, but the significant differences correspond to site comparisons where p is less than 0.05. This table helps us summarize the results where yes indicates a likely significant difference between the sites for a particular metal.

metal	C vs. A	I vs. A	L vs. A	I vs. C	L vs. C	L vs. I
aluminum	yes	no	yes	yes	no	yes
calcium	yes	no	yes	yes	no	yes
iron	yes	no	yes	yes	no	yes
magnesium	yes	no	yes	yes	no	yes
sodium	no	no	yes	no	yes	yes

From the patter of yes and no, we see that we can use an analysis for any one of Al, Ca, Fe, or Mg to distinguish between sites C and L relative to sites A and I, but not between sites C and L or sites A and I. More specifically, if we look at the signs of the differences between sites, we see that a relatively low concentration of Al or a relatively high concentration of Ca, Fe, or Mg suggests that a shard is from either site C or L, with the opposite outcome suggesting that a shard is from either site A or I.

Having determined that a pottery shard is from either site C or L, we can distinguish between the two possibilities by analyzing for Na, with a relatively high concentration suggesting the shard is from site L and a relatively low concentration suggesting it is from site C (although we do need to note two limitations to this conclusion: first, we have but two samples from site C, which is not a lot of samples on which to basis a firm conclusion; and, two, the two samples from site C have concentrations of Na that exceed that for two samples from site L, which means that our ability to separate sites C and L using Na is not perfect). Unfortunately, for pottery shards narrowed to sites A or I, there is no single comparison that will distinguish between the two sites.