MATH 424: Homework Chapter 1: Analysis of Variance

Due on Thursday, December 14, 2017

Kafai 11:10am

Jonathan Dombrowski

Contents

Q **3**
a . . . 3

Q 20 **4**
a . . . 4
b . . . 5
c . . . 5
d . . . 5

Q 26 **5**
a . . . 5
b . . . 6
c . . . 7

Q 36 **8**
a . . . 8
b . . . 8
c . . . 10

Q 56 **10**
a . . . 11
b . . . 11
c . . . 11
d . . . 11
e . . . 11
f . . . 12
g . . . 12

Q 65 **12**
a . . . 12

Q

IS there sufficient evidence to indicate the differences among the mean Al/Be ratios for the five boreholes? Test using $\alpha = 0.10$.

a

Model to test : $E(y) = \beta_0 = \beta_1 x_1$
Creating dummy variables for the Borehole type, the model changes to :

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

$$\beta_0 \begin{cases} 0: \text{if not} \\ 1: \text{SD} \end{cases}$$

$$x_1 \begin{cases} 0: \text{if not} \\ 1: \text{SWRA} \end{cases}$$

$$x_2 \begin{cases} 0: \text{if not} \\ 1: \text{URMB-1} \end{cases}$$

$$x_3 \begin{cases} 0: \text{if not} \\ 1: \text{URMB-2} \end{cases}$$

$$x_4 \begin{cases} 0: \text{if not} \\ 1: \text{URMB-3} \end{cases}$$

Testing for overall model validity

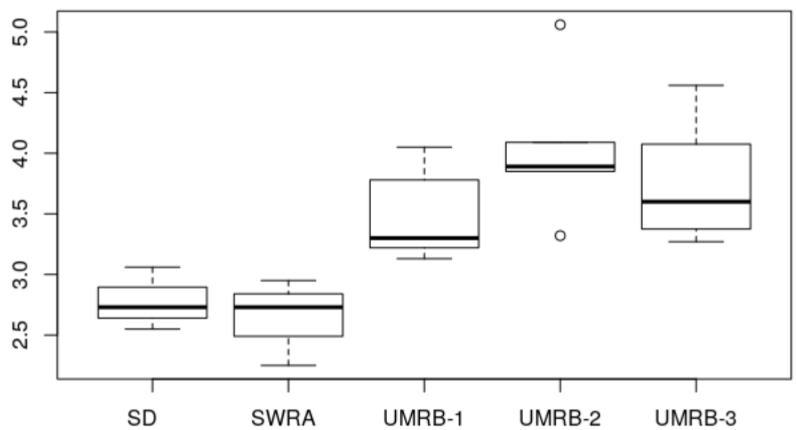
$$H_o = \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$
$$H_a = \beta_i \neq \beta_j, \forall i, j : i \neq j$$

```
Call:
glm(formula = RATIO ~ BOREHOLE, data = TILLRATIO)

Deviance Residuals:
    Min        1Q    Median        3Q       Max
-0.6967   -0.2821   -0.1267    0.2939    1.0433

Coefficients:
               Estimate Std. Error t value Pr(>t)
(Intercept)      2.7800     0.2590  10.734 5.52e-10 ***
BOREHOLESWRA    -0.1367     0.3663  -0.373 0.712778
BOREHOLEUMRB-1   0.7171     0.3095   2.317 0.030701 *
BOREHOLEUMRB-2   1.2367     0.3172   3.899 0.000827 ***
BOREHOLEUMRB-3   0.9814     0.3095   3.171 0.004607 **
---
Signif. codes:
0    ***    0.001    **    0.01    *    0.05    .    0.1          1

(Dispersion parameter for gaussian family taken to be 0.2012109)

    Null deviance: 10.0612  on 25  degrees of freedom
Residual deviance:  4.2254  on 21  degrees of freedom
AIC: 38.543
```

Calling analysis of variance on the model gives us the p-value we are looking for:

```
             Df  Sum Sq  Mean Sq  F value    Pr(>F)
BOREHOLE      4   5.836   1.4589    7.251  0.000784 ***
Residuals    21   4.225   0.2012
---
Signif. codes:
0    ***    0.001    **    0.01    *    0.05    .    0.1          1
```

From this we can complete the test and since $\alpha = 0.10$ is greater than the p-value of 0.000784, therefore we reject the null hypothesis and state that there is sufficient evidence to sugget that there are differences between the mean Al/Be ratios for the five boreholes. The visual below supports these claims.



Q 20

(a) Explain why this data should be analyzed using a randomized block design. As part of your answer, identify the blocks and the treatments.
(b) A partial ANOVA table for the experiment is shown below. Explain why there is enough information in the table to make conclusions.
(c) State the null hypothesis of interest to the researcher.
(d) Make the appropriate conclusion.

a

The Randomized Block Design helps remove biases in analyzing data with factors that can influence the main factors, but are not the main focus of the experiment. In this case, we expect that the scores at different times will be more alike than the the total comparison. The goal of the experiment was to compare the mean competence levels of the three periods. In order to abstract the bias of different persons' starting scores, we can block the experiment in such a way that the 222 participants are the different blocks and the treatments are the three time periods: 1wk before, 2d after and 2m after.

Q 20 continued on next page...

b

Being able to make a conclusion is dependent on accepting or rejecting a hypothesis based on a F-test/p-value. In order to find that information, we have to calculate various terms from the data. If we are already given the p-values, then the other information is unnecessary in making a reasonable conclusion. Therefore we are given enough information to make conclusions.

c

The researcher is interested in comparing the mean competence levels of the three periods. This relies on testing whether or not the means are all the same. The null hypothesis in the interest of the researchers is as follows:

$$H_o = \mu_1 = \mu_2 = \mu_3 = 0$$

with the accompanying alternative hypothesis being:

$$H_a = \mu_i \neq \mu_j \forall i \neq j$$

d

Based on the partial ANOVA table given to us in the text, we can reject the null hypothesis in the previous question and state that there is at least one pair of μ's that are different therefore the time period has an effect on determining the competence level. If we are trying to build a model to predict the competency of trainees, then taking the 222 different trainees into account is going to be a significant factor in reducing the variance in the model.

Q 26

Massage therapy for boxers (contd). Refer to Exercise 12.25. The models of parts a, b, and c were fit to data in the table using MINITAB. The MINITAB printouts are displayed here.
(a) Construct an ANOVA summary table.
(b) Is there evidence of differences in the punch- ing power means of the four interventions? Use = .05.
(c) Is there evidence of a difference among the punching power means of the boxers? That is, is there evidence that blocking by box- ers was effective in removing an unwanted source of variability? Use = .05.

a

From the General format of an ANOVA table for Randomize Block Design we can see the formulas for calculating the Anova Table.

```
Source              df         SS        MS                   F
Treatments          p - 1      SST   MST = SST/(p-1)          F = MST/MSE
Blocks              b - 1      SSB   MSB = SSB/(b-1)          F = MST/MSE
Error     n - p - b + 1        SSE   MSE=SSE/(n-p-b+1)
Total               n - 1      SS(Total)
```

Followed by the filled table with the values from the text

```
----------------------------------------------------------------------

               d.f.      SS        MS      F(ratio)    p-value
Treatments       3     15,754     5,251      4.158     0.133
Block            7    117,044    16,721     13.230     0.00092
```

```
Error           21     26,525     1,263
Total           31    159,323    13,280
```

b

We are checking the following test for the reduced model of just the Treatments consisting of the different times for the massages.

$$H_o = I_0 = I_1 = I_2 = I_3 = 0$$

$$H_a = I_i \neq I_j; \forall i \neq j$$

Using the F-test we derived in the Anova table:

$$F = \frac{MST}{MSE} = 4.158$$

The accompanying p-value for this F-score is : 0.133.
Since we are using $\alpha = 0.05$, then we can conclude that p is greater than α and therefore we fail to reject the null hypothesis. There is not substantial evidence to show differences in the punching power means of the four interventions.

c

We are checking the following test for the reduced model of just the Blocks consisting of the different boxers.

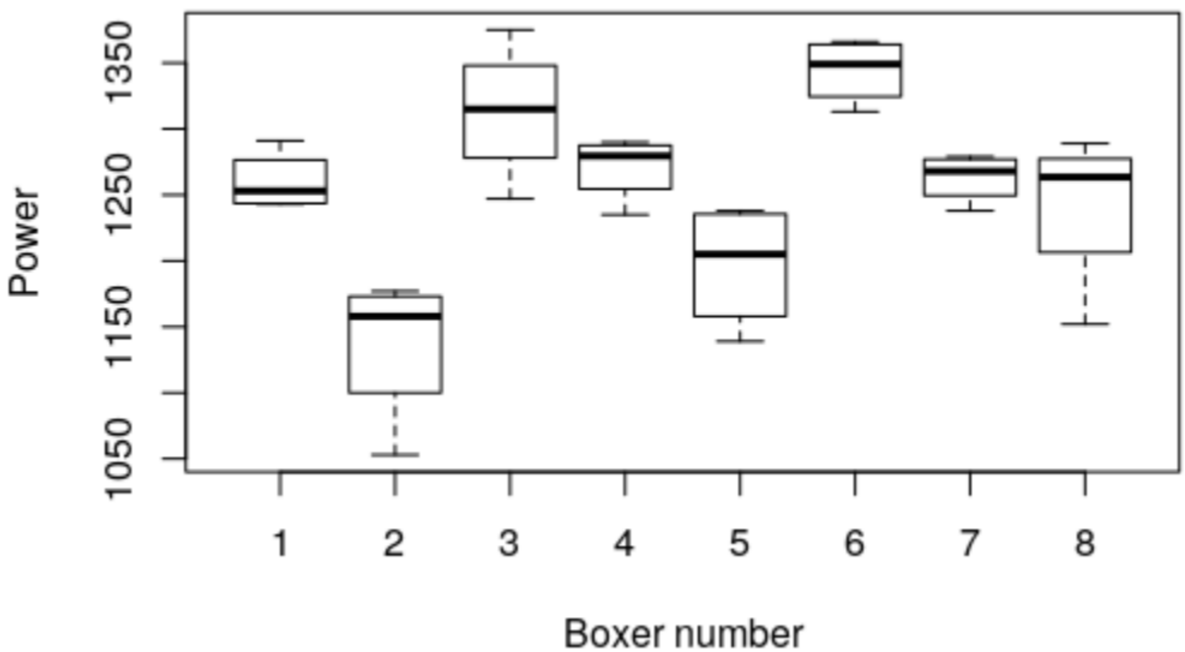
$$H_o = \beta_0 = \beta_1 = \beta_2 = \ldots = \beta_7 = 0$$

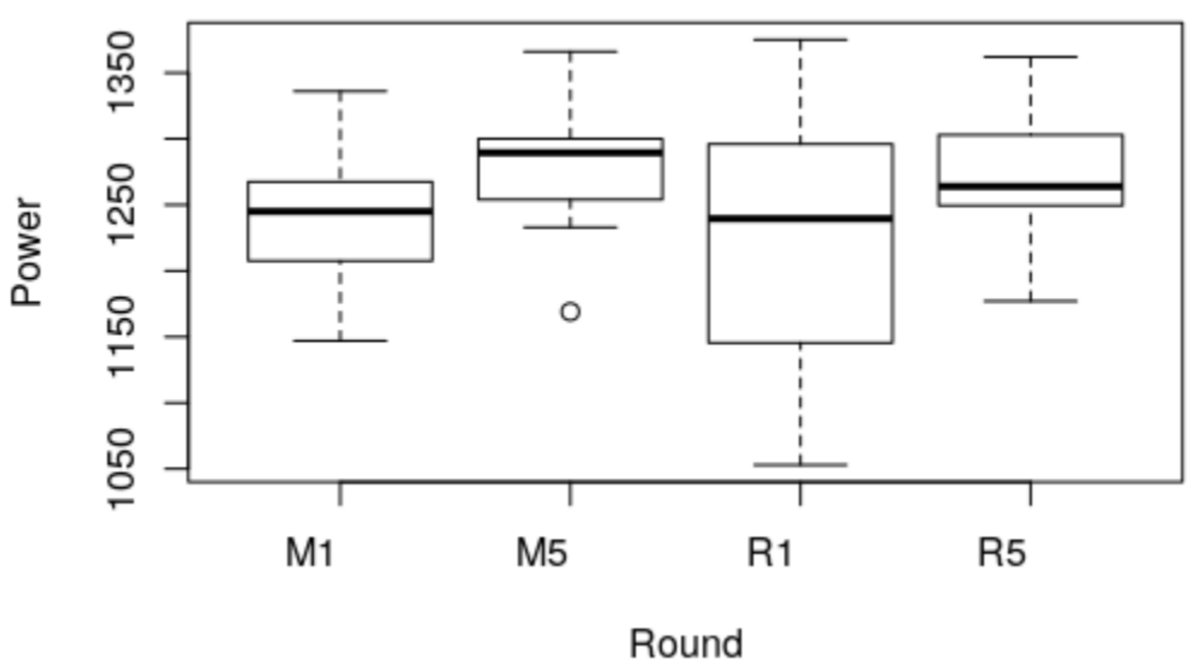
$$H_a = \beta_i \neq \beta_j ; \forall i \neq j$$

Using the F-test we derived in the Anova table:

$$F = \frac{MST}{MSE} = 13.230$$

The accompanying p-value for this F-score is : 0.00092.
Since we are using $\alpha = 0.05$, then we can conclude that p is less than α. Therefore we can conclude that the blocking by boxers is effective in removing unwanted variability. We can conclude that the mean punching power of a boxer is dependant on which boxer is the individual in question. This conclusion of dependency is further shown in the below tables





Q 36

Commercial eggs produced from different hous- ing systems. In the production of commercial eggs in Europe, four different types of housing systems for the chickens are used: cage, barn, free range, and organic. The characteristics of eggs produced from the four housing systems were investigated in Food Chemistry (Vol. 106, 2008). Twenty-four commercial grade A eggs were ran- domly selectedsix from each of the four types of housing systems. Of the six eggs selected from each housing system, three were Medium weight class eggs and three were Large weight class eggs. The data on whipping capacity (percent overrun) for the 24 sampled eggs are shown in the next table. The researchers want to investigate the effect of both housing system and weight class on the mean whipping capacity of the eggs. In par- ticular, they want to know whether the difference between the mean whipping capacity of medium and large eggs depends on the housing system.
(a) Identify the factors and treatments for this experiment.
(b) Use statistical software to conduct an ANOVA on the data. Report the results in an ANOVA table.
(c) Is there evidence of interaction between housing system and weight class? Test using = .05. What does this imply, practically?

a

There are two factors in this experiment: Housing and Weight Class. The treatments for this are the {Medium,Large} * {Cage, Barn, Free-Range, Organic} for a total of 8 treatments.

b

Fitting the model onto the data we get the following readout:

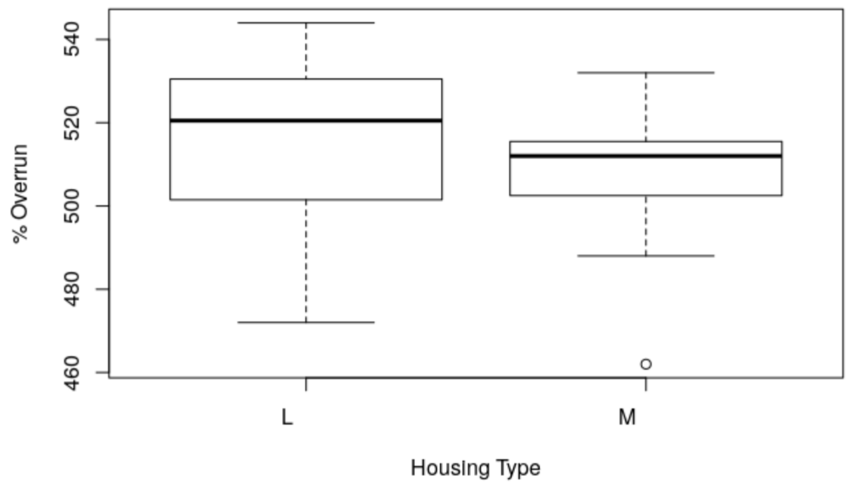
```
Call:
lm(formula = OVERRUN ~ HOUSING + WTCLASS + HOUSING:WTCLASS, data = EGGS2)

Residuals:
   Min     1Q Median     3Q    Max
-19.67  -4.25  -1.00   6.50  19.33

Coefficients:
                               Estimate Std. Error t value Pr(>t)
(Intercept)                     511.667      6.365  80.387  < 2e-16 ***
HOUSINGCAGE                     -29.000      9.002  -3.222  0.00533 **
HOUSINGFREE                      12.333      9.002   1.370  0.18956
HOUSINGORGANIC                   23.333      9.002   2.592  0.01965 *
WTCLASSM                          3.333      9.002   0.370  0.71601
HOUSINGCAGE    :WTCLASSM         -4.333     12.730  -0.340  0.73798
HOUSINGFREE    :WTCLASSM        -16.333     12.730  -1.283  0.21775
HOUSINGORGANIC :WTCLASSM        -15.000     12.730  -1.178  0.25590
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

Residual standard error: 11.02 on 16 degrees of freedom
Multiple R-squared:  0.7989,    Adjusted R-squared:  0.7109
F-statistic:  9.08 on 7 and 16 DF,  p-value: 0.0001454
```

With the accompanying ANOVA table:

```
                Df  Sum Sq Mean Sq F value  Pr(>F)
HOUSING          3   7249  2416.5  19.882 1.2e-05 ***
WTCLASS          1    187   187.0   1.539   0.233
HOUSING:WTCLASS  3    289    96.3   0.792   0.516
Residuals       16   1945   121.5
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1
```

The WTCLASS seems to have no effect in determining the percent Overrun. This is further confirmed by the following boxplot.

Q 36 [b] continued on next page. . .



f

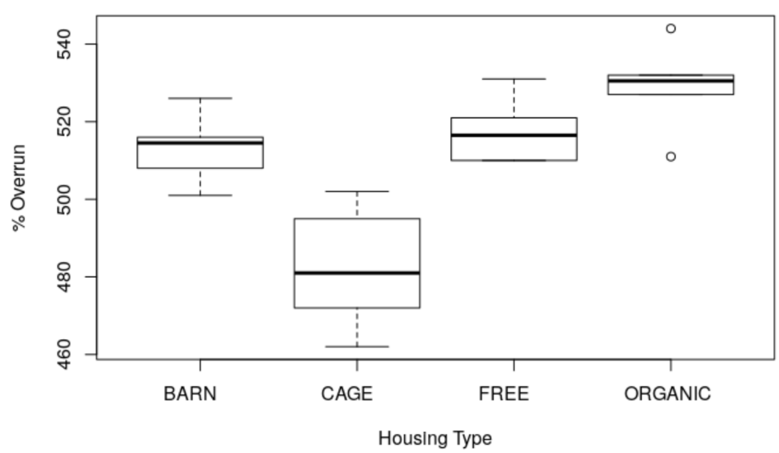
c

Testing whether or not the HOUSING and WTCLASS factors have any interaction:

$$H_o : \beta_3 = 0$$

$$H_a : \beta_3 \neq 0$$

Using the output from the ANOVA table above, we can see that the p-value for the interaction variables is 0.516. From this we can see that p is far greater than α and can fail to reject H_0. From this we can conclude that there is no interaction between the HOUSING and WTCLASS in predicting OVERRUN.



Q 56

Commercial eggs produced from different hous- ing systems. Refer to the Food Chemistry (Vol. 106, 2008) study of four different types of egg housing systems, Exercise 12.36 (p. 659). Recall that you discovered that the mean whipping capacity (percent overflow) differed for cage, barn, free range, and organic egg housing systems. A multiple comparisons of means was conducted using Tukeys method with an experimentwise error rate of .05. The results are displayed in the SPSS printout above.

(a) Locate the confidence interval for (l CAGE BARN) on the printout and interpret the result.
(b) Locate the confidence interval for (CAGE FREE) on the printout and interpret the result.
(c) Locate the confidence interval for (CAGE ORGANIC) on the printout and interpret the result.
(d) Locate the confidence interval for (BARN FREE) on the printout and interpret the result.
(e) Locate the confidence interval for (BARN ORGANIC) on the printout and interpret the result.
(f) Locate the confidence interval for (FREE ORGANIC) on the printout and interpret the result.
(g) Based on the results, parts af, provide a ranking of the housing system means. Include the experimentwise error rate as a statement of reliability.

a

The CI for $\mu_{cage} - \mu_{barn}$ is (-49.38, -12.95). The true difference between the cage raised and the barn raised percent overflow will lie within the interval of (-49.38, -12.95) for 95% of samples from the population. This difference is statistically significant with a significance level lower than the threshold: 0.001 vs. 0.05; therefore the pair of treatments has a decisive difference. Barn raised eggs have a significantly higher percentage overrun in regards to cage raised.

b

The CI for $\mu_{cage} - \mu_{free}$ is (-53.54,-17.12). The difference between cage raised and the free range percent overflow will lie within the interval of (-53.54,-17.12) for 95% of the samples from the population. This difference is statistically significant with a significance level lower than the threshold: .000 vs. 0.05; therefore cage-raised chickens have a decisively lower percent overrun than free-range.

c

The CI for $\mu_{cage} - \mu_{organic}$ is (-65.21,-28.79). The difference between the cage and organic percent overflow will lie withing the interval (-65.21,-28.79) 95% of the time. This difference is statistically significant with a significance level lower than the threshold: .000 vs. 0.05. Therefore the cage raised chickens have a decisively lower percent overrun than the organically raised.

d

The CI for $\mu_{barn} - \mu_{free}$ is (-22.38,14.04). The difference between the barn and free percentage overflow will lie in the interval (-22.38,-17.12) 95% of the samples from the population. The mean difference is NOT statisticaly significant with a significance level of 0.912 vs. 0.05. Therefore we cannot reject the hypothesis that the two means are different.

e

The CI for $\mu_{barn} - \mu_{organic}$ is (-34.04,2.38). The difference between the barn and the organic percentage overflow lie in the interval (-34.04,2.38) 95% of the samples from the population. The mean difference is NOT statisticaly significant with a significance level of 0.100 vs. 0.05. Therefore we cannot reject the hypothesis that the two means are different.

f

The CI for $\mu_{free} - \mu_{organic}$ is (-29.88,6.54). The difference between the free and the organic percentage overflow wil lie between the interval (-29.88,6.54) 95% of the samples from the population. The mean difference is NOT statisticaly significant with a significance level of 0.295 vs. 0.05. Therefore we cannot reject the hypothesis that the two means are different.

g

Ordering the Tukey readout from R gives us the following table, which orders the difference of th means from highest to lowest, and also from lowest p-value to highest. The p-values in the table are significant at $\alpha = 0.05$.

```
  Tukey multiple comparisons of means
    95 percent family-wise confidence level

Fit: aov(formula = modelA)

$HOUSING
                        diff        lwr       upr     p adj
ORGANIC -CAGE      47.000000  29.222532  64.77747 0.0000021
FREE    -CAGE      35.333333  17.555865  53.11080 0.0001048
CAGE    -BARN     -31.166667 -48.944135 -13.38920 0.0004576
------------------------------------------------------------
ORGANIC -BARN      15.833333  -1.944135  33.61080 0.0917490
ORGANIC -FREE      11.666667  -6.110801  29.44413 0.2861315
FREE    -BARN       4.166667 -13.610801  21.94413 0.9121909
```

Q 65

a

The R printout is at the bottom of this, but for proof of concept, verifying the method used in R seemed appropriate. The final R printout is below the proof of concept.
Using Tukeys' comparison method:

$$w_{ij} = q_\alpha(p, v)\sqrt{\frac{MSE}{2}} * \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

$q_\alpha(p, v) = q_\alpha(5, 21) = 3.578$

$$w = 3.578 * \sqrt{\frac{0.2012}{2}} * \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

$$= 1.117 * \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

In order to calculate the w_{ij} we have to have the specific treatments in mind. Testing this on two different pairs:

$$w_{URMB-2/SD} = 1.117 * \sqrt{\frac{1}{6} + \frac{1}{3}} = 0.7898$$

Q 65 [a] continued on next page. . .

$$w_{SWRA/SD} = 1.117 * \sqrt{\frac{1}{7} + \frac{1}{3}} = 0.7710$$

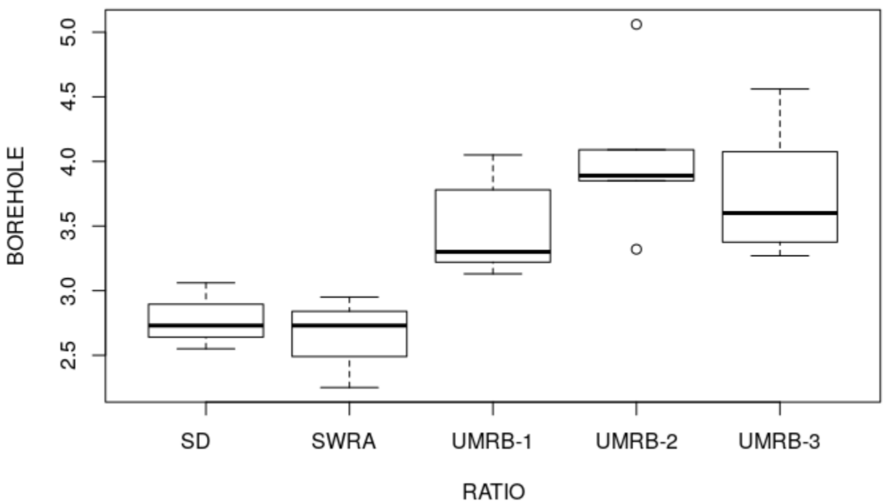
The distance in the means is greater than the w for URMB-2/SD (0.7898 vs. 1.23) and less than for the SWRA/SD pair (0.771 vs. -0.0133). From this we can conclude that the difference in means for URMB-2/SD is statistically significant and the means for SWRA/SD are not statistically significant. This reaffirms the p-values in the following R-output table ran from the built in Tukey multipe analysis.

```
    90 percent family-wise confidence level

Fit: aov(formula = model)

$BOREHOLE
                    diff         lwr       upr     p adj
SWRA  -SD     -0.1366667 -1.10100241 0.8276691 0.9955440
UMRB-1-SD      0.7171429 -0.09786960 1.5321553 0.1788060
UMRB-2-SD      1.2366667  0.40152741 2.0718059 0.0066199
UMRB-3-SD      0.9814286  0.16641611 1.7964410 0.0333957
UMRB-1-SWRA    0.8538095  0.03879707 1.6688220 0.0782702
UMRB-2-SWRA    1.3733333  0.53819408 2.2084726 0.0024546
UMRB-3-SWRA    1.1180952  0.30308278 1.9331077 0.0126734
UMRB-2-UMRB-1  0.5195238 -0.13756024 1.1766079 0.2643848
UMRB-3-UMRB-1  0.2642857 -0.36702022 0.8955916 0.8034378
UMRB-3-UMRB-2 -0.2552381 -0.91232215 0.4018460 0.8420582
```

Further examination of the p-values listed in the table show that the difference of the means of the highlighted pairs are statistically significant through the use of the Tukey Multiple Comparisons method at an experiment wise error rate of 0.1. These pairs will have decisive impact on the prediction of glacial drift age.
This includes:

- UMRB-2-SD
- UMRB-3-SD
- UMRB-1-SWRA
- UMRB-2-SWRA
- UMRB-3-SWRA

To graphically support this conclusion:


BOREHOLE
5.0
4.5
4.0
3.5
3.0
2.5
SD
SWRA
UMRB-1
UMRB-2
UMRB-3
RATIO