NORTHWESTERN UNIVERSITY | SCHOOL OF PROFESSIONAL STUDIES

R Lesson 11 - Solutions

MSPA 401 – Introduction to Statistical Analysis

Lesson 11 Analysis of Variance

1) Use the **tableware.cvs** data to test the hypothesis that the mean RATE for the five levels of TYPE are equal. Test the hypothesis using a 0.05 significance level. Produce and plot means and confidence intervals for each level of TYPE. (You may use the example given in Section 16.1.1 of Lander as a reference and guide. Try the analysis two different ways. Use -1 in the model to suppress the intercept, and alternatively without using -1. Compare results. Load the ggplot2() and plyr() packages.)

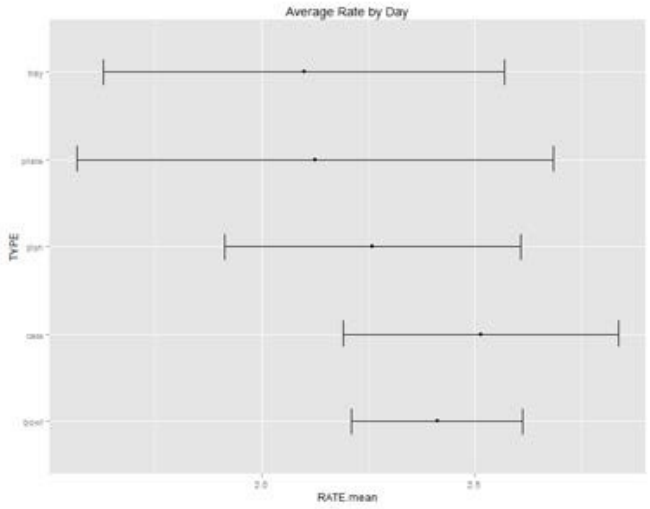
   The first output uses -1 in the model and compares the means to zero.

```
> summary(RATE_anova)
             Df Sum Sq Mean Sq F value Pr(>F)
TYPE          5  317.4   63.48   219.9 <2e-16 ***
Residuals    54   15.6    0.29
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$p < 0.001$. We reject the null hypothesis that the mean RATE of all five levels is equal to zero. There is at least one significant difference between oneRATE and zero.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
TYPEbowl       2.4113     0.1120   21.52  < 2e-16 ***
TYPEcass       2.5140     0.1699   14.80  < 2e-16 ***
TYPEdish       2.2600     0.2031   11.13 1.36e-15 ***
TYPEplate      2.1256     0.1791   11.87  < 2e-16 ***
TYPEtray       2.0990     0.1699   12.35  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5373 on 54 degrees of freedom
Multiple R-squared:  0.9532,    Adjusted R-squared:  0.9488
F-statistic: 219.9 on 5 and 54 DF,  p-value: < 2.2e-16
```



The second output does not use -1 and shows how the means compare among themselves. They are not significantly different. Observation of the plot shown above confirms this result.

```
> summary(RATE_anova)
            Df Sum Sq Mean Sq F value Pr(>F)
TYPE         4   1.42  0.3550    1.23  0.309
Residuals   54  15.59  0.2887
> summary(RATE_lm)

Call:
lm(formula = RATE ~ TYPE, data = tableware)

Residuals:
    Min      1Q  Median      3Q     Max
-1.2690 -0.3726  0.0800  0.4335  0.9287

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.4113     0.1120  21.522   <2e-16 ***
TYPEcass      0.1027     0.2035   0.505    0.616
TYPEdish     -0.1513     0.2319  -0.652    0.517
TYPEplate    -0.2857     0.2113  -1.353    0.182
TYPEtray     -0.3123     0.2035  -1.534    0.131
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5373 on 54 degrees of freedom
Multiple R-squared:  0.08348, Adjusted R-squared:  0.01559
F-statistic:  1.23 on 4 and 54 DF,  p-value: 0.3092
```

2) Use the **tableware.cvs** data to test the hypothesis that the mean PRICE for the five levels of TYPE are equal. Test the hypothesis using a 0.05 significance level. Print out 95% confidence intervals for each level of TYPE.

```
Analysis of Variance Table

Response: PRICE
          Df Sum Sq Mean Sq F value    Pr(>F)
TYPE       4  36174  9043.5  6.8532 0.0001548 ***
Residuals 54  71258  1319.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

And, the confidence intervals for each level of TYPE:

```
                 2.5 %    97.5 %
(Intercept)  52.205239 82.57737
TYPEcass     31.941838 87.11555
TYPEdish    -18.686590 44.18970
TYPEplate   -44.193091 13.07715
TYPEtray      3.971838 59.14555
```

Data Set: hot_dogs.csv (Original source: Consumer Reports, June 1986, pp. 366-367.)

1) Use the hot_dogs.csv data. Perform a one---way AOV by Type on Calories and also Sodium. Use Tukey's Honest Significant Difference Test if the F---test is significant. Generate boxplots.

```
> summary(calories.anova)
            Df Sum Sq Mean Sq F value   Pr(>F)
Type         2  17692    8846   16.07 3.86e-06 ***
Residuals   51  28067     550
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$p < 0.001$. There is a significant difference between the mean calories of at least two types. Tukey's Honest Significant Difference Test (below) indicates that there are significant differences between the mean calories of the poultry and beef and poultry and meat hot dogs.

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

Fit: aov(formula = Calories ~ Type, data = hotdogs)

$Type
                   diff       lwr       upr     p adj
Meat-Beef      1.855882 -16.82550  20.53726 0.9688129
Poultry-Beef -38.085294 -56.76667 -19.40391 0.0000277
Poultry-Meat -39.941176 -59.36515 -20.51720 0.0000239
> summary(sodium.anova)
            Df Sum Sq Mean Sq F value Pr(>F)
Type         2  31739   15869   1.778  0.179
Residuals   51 455249    8926
```

$p = 0.179$. We do not find significant difference between the mean sodium contents of the different hot dog types.