Analysis of Variance

Brenda Maldonado

9/26/2021

1) The heartbpchol.csv data set contains continuous cholesterol (Cholesterol) and blood pressure status (BP_Status) (category: High/ Normal/ Optimal) for alive patients.

```
chol <- read.csv('heartbpchol.csv')
head(chol,2)
```

```
##   Cholesterol BP_Status
## 1         221   Optimal
## 2         188      High
```

For the heartbpchol.xlsx data set, consider a one-way ANOVA model to identify differences between group cholesterol means. The normality assumption is reasonable, so you can proceed without testing normality.

1a) Perform a one-way ANOVA for Cholesterol with BP_Status as the categorical predictor. Comment on statistical significance of BP_Status, the amount of variation described by the model, and whether or not the equal variance assumption can be trusted.

```
aov(Cholesterol~BP_Status, data=chol) %>% summary()
```

```
##              Df  Sum Sq Mean Sq F value  Pr(>F)
## BP_Status     2   25211   12605   6.671 0.00137 **
## Residuals   538 1016631    1890
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Since the F value is 6.671 the null hypothesis is rejected. This means that the groups being tested are significantly different and the equal variance assumption cannot be trusted.

1b) Comment on any significantly different cholesterol means as determined by the post-hoc test comparing all pairwise differences. Specifically explain what that tells us about differences in cholesterol levels across blood pressure status groups, like which group has the highest or lowest mean values of Cholesterol.

```r
pairwise.t.test(chol$Cholesterol, chol$BP_Status, p.adjust.method = "bonf")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  chol$Cholesterol and chol$BP_Status
##
##         High   Normal
## Normal  0.0121 -
## Optimal 0.0064 0.7093
##
## P value adjustment method: bonferroni
```

2) For this problem use the bupa.csv data set. Check UCI Machine Learning Repository for more information (http://archive.ics.uci.edu/ml/datasets/Liver+Disorders). The mean corpuscular volume and alkaline phosphatase are blood tests thought to be sensitive to liver disorder related to excessive alcohol consumption. We assume that normality and independence assumptions are valid.

```r
bupa <- read.csv('bupa.csv')
bupa %>% head()
```

```
##   mcv alkphos drinkgroup
## 1  85      92          1
## 2  85      64          1
## 3  86      54          1
## 4  91      78          1
## 5  87      70          1
## 6  98      55          1
```

2a) Perform a one-way ANOVA for mcv as a function of drinkgroup. Comment on significance of the drinkgroup, the amount of variation described by the model, and whether or not the equal variance assumption can be trusted

```r
aov(mcv~drinkgroup,bupa) %>% summary()
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)    
## drinkgroup    1    596   596.4   32.94 2.09e-08 ***
## Residuals   343   6210    18.1                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Since the P value is 2.09e-08 the null hypothesis is rejected. This means that the groups being tested are significantly different and the equal variance assumption cannot be trusted.

2b) Perform a one-way ANOVA for alkphos as a function of drinkgroup. Comment on statistical significance of the drinkgroup, the amount of variation described by the model, and whether or not the equal variance assumption can be trusted.

```r
aov(alkphos~drinkgroup,data=bupa) %>% summary()
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## drinkgroup    1    320   320.0   0.951   0.33
## Residuals   343 115483   336.7
```

- Since the F value is 0.951 the null hypothesis can be accepted. This means that the groups being tested are significantly similar and the equal variance assumption can be trusted.

```r
psy <- read.csv('psych.csv')
head(psy,2)
```

3) The psychology department at a hypothetical university has been accused of underpaying female faculty members. The data represent salary (in thousands of dollars) for all 22 professors in the department. This problem is from Maxwell and Delaney (2004).

```
##   sex   rank salary
## 1   F Assist     33
## 2   F Assist     36
```

```r
aov(salary~sex+rank,data=psy) %>% summary()
```

3a) Fit a two-way ANOVA model including sex (F, M) and rank (Assistant, Associate) the interaction term. What do the Type 1 and Type 3 sums of squares tell us about significance of effects? Is the interaction between sex and rank significant? Also comment on the variation explained by the model.

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## sex          1  155.2  155.15   17.88 0.000454 ***
## rank         1  169.8  169.82   19.57 0.000291 ***
## Residuals   19  164.8    8.68                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the ANOVA results, you can conclude the following, based on the p-values and a significance level of 0.05:

- the p-value of sex is 0.000454 (significant), which indicates that the sex is associated with significant different in salary.
- the p-value of rank is 0.000291 (significant), which indicates that the rank is associated with significant different in salary.