# Problem set 8

Professor McNamara

```r
library(tidyverse)
```

```
-- Attaching packages --------------------------------------- tidyverse 1.3.2 --
v ggplot2 3.3.6      v purrr   0.3.5
v tibble  3.1.8      v dplyr   1.0.10
v tidyr   1.2.1      v stringr 1.4.1
v readr   2.1.3      v forcats 0.5.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

```r
library(Stat2Data)
library(skimr)
library(agricolae)
```

Exercises to hand in: 5.14, 5.34, 5.38, 5.46, 8.14

## 5.14 Car ages

(No data)

### a. "You can't use ANOVA on the data, because there are four groups to compare."

I would say that Anova test can be used in this situation, because the 4 groups are independent from each other. All the explanatory variables are categorical.

### b. “You can’t use ANOVA on the data, because the response variable is not quantitative.”

No, anova test can be used because the response variable is quantitative. In this case the response variable is the age of the car. The explanatory variables for this study would be the 4 groups, Faculty, Students, staffs and administrators.

### c. “You can’t use ANOVA on the data, because the sample sizes for the four groups will probably be different.”

No, I would would disagree, because the total sample size of the study is 200. These includes all the members that are from different groups.

### d. “You can do the calculations for ANOVA on the data, even though the sample sizes for the four groups will probably be different, but you can’t generalize the results to the populations of all people with a parking permit at your college/university.”

No, this statement is wrong, because the sample size for this case study is 200 in total. Secondly, we can use the results obtained to generalize to a larger population, because this a observational study instead of a experimental study. More that this study is randomized, which states that the study is not biased to a certain group.

## 5.34 Aphid honeydew

(No data)

### a. Fill in the three missing values

(Fine to just tell me what they are)

DF model: 51-46 = 5

SS model: 64.77-39.87 = 24.9

F-value: 4.9807/0.8667 = 5.746741

### b. How many combos?

Number of combos = Degrees of freedom + 1 = 6

There are 6 different aphid/plant combinations for this analysis.

### c. Summarize the conclusion

Since the p value is 0.00 i.e. less than 0.05, we can reject the null hypothesis. We have enough evidence to suggest that at least one mean is different than the grand mean i.e. the mean of honeydew produced by at least one combination of aphid and host plant is different from the grand mean of amount of honeydew produced in the experiment by aphids for different combinations of type of aphid and type of host plant.

### 5.38 Meniscus: stiffness

```
data("Meniscus")
Meniscus <- Meniscus %>%
  mutate(Method = as_factor(Method)) %>%
  mutate(Method = fct_recode(Method,
                             "Vertical Suture" = "1",
                             "Meniscus Arrow" = "2",
                             "FastT-Fix" = "3"))
```

#### a. Hypotheses

Null hypothesis: The group means are all equal to each other and equal to the grand mean. This means that there is not a linear relationship between the method of repairing a meniscus and stiffness in the knee.

Alternative hypothesis: At least one of the group mean is different from another group or the grand mean. There is a linear relationship between the method of repairing a meniscus and stiffness in the knee.

#### b. Conditions

```
Meniscus %>%
  group_by(Method) %>%
  skim_without_charts(Stiffness)
```

Table 1: Data summary

| | |
|---|---|
| Name | Piped data |
| Number of rows | 18 |
| Number of columns | 4 |
| Column type frequency: | |
| numeric | 1 |
| Group variables | Method |

**Variable type: numeric**

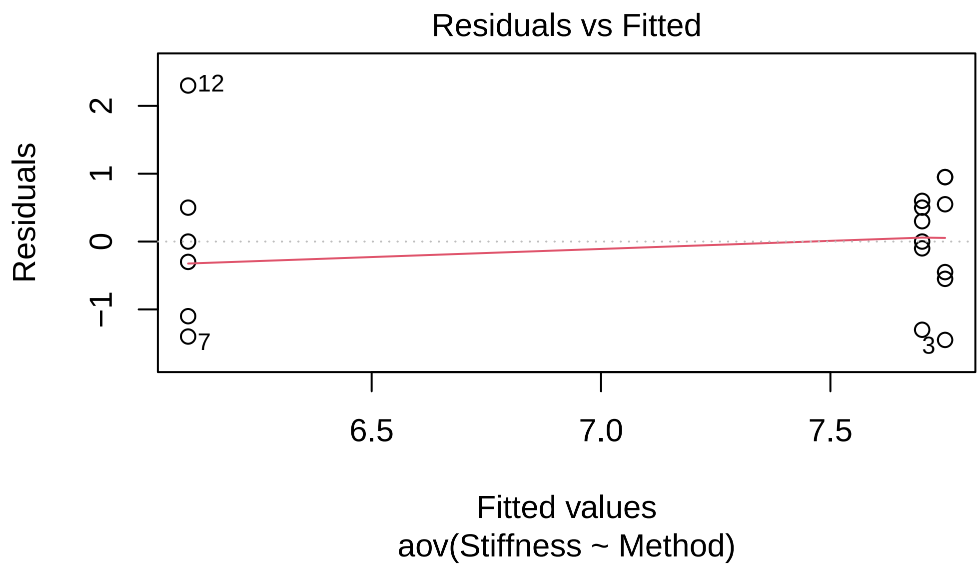
| skim_variable | Method | n_missing | complete_rate | mean | sd | p0 | p25 | p50 | p75 | p100 |
|---|---|---|---|---|---|---|---|---|---|---|
| Stiffness | Vertical Suture | 0 | 1 | 7.75 | 0.97 | 6.3 | 7.23 | 7.80 | 8.60 | 8.7 |
| Stiffness | Meniscus Arrow | 0 | 1 | 6.10 | 1.33 | 4.7 | 5.20 | 5.95 | 6.47 | 8.4 |
| Stiffness | FastT-Fix | 0 | 1 | 7.70 | 0.69 | 6.4 | 7.62 | 7.85 | 8.15 | 8.3 |

```r
a1 <- aov(Stiffness ~ Method, data = Meniscus)
model.tables(a1)
```

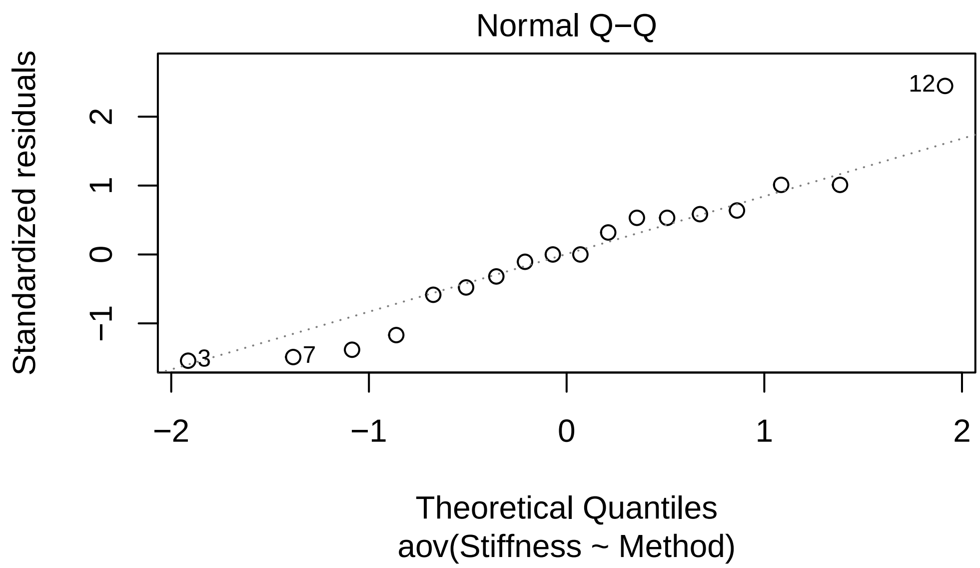
```
Tables of effects

 Method
Method
Vertical Suture  Meniscus Arrow       FastT-Fix
         0.5667         -1.0833          0.5167
```

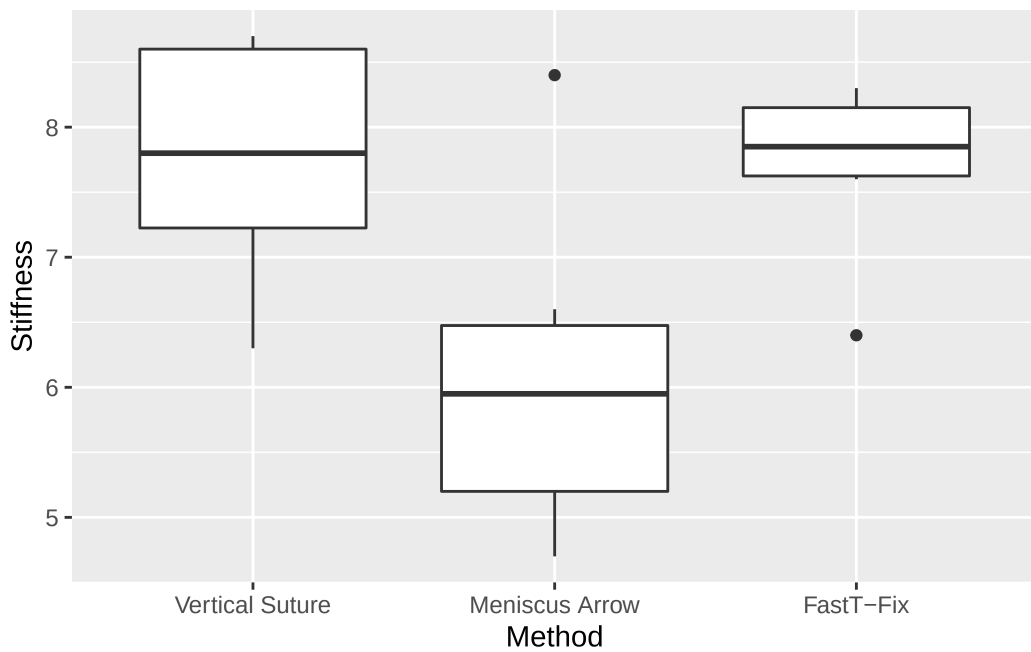
```r
plot(a1, which = 1)
```



```r
plot(a1, which = 2)
```




```r
ggplot(Meniscus) + geom_boxplot(aes(x = Method, y = Stiffness))
```




Independence: Since the methods are independent from each other, we can say that the condition for independence is met.

Normality: The QQ plot looks S-shaped, therefore, the condition for normality is not met.

Equality of variance: $sd_{max}/sd_{min} = 1.33/0.693 = 1.9 <= 2$ so the condition for equality of variance is met.

### c. Conduct an ANOVA

```
anova(a1)
```

```
Analysis of Variance Table

Response: Stiffness
          Df Sum Sq Mean Sq F value  Pr(>F)
Method     2 10.570   5.285  4.9811 0.02193 *
Residuals 15 15.915   1.061
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p value is statistically significant ($<0.05$) so we can reject the null hypothesis. So we have enough evidence to suggest that the mean value of stiffness

differs based on the method of meniscus repair.

## 5.46 Meniscus stiffness: Fisher's LSD

```r
print(LSD.test(a1, "Method"))
```

```
$statistics
  MSerror Df     Mean       CV t.value     LSD
    1.061 15 7.183333 14.33942 2.13145 1.26757

$parameters
        test p.ajusted name.t ntr alpha
  Fisher-LSD      none Method   3  0.05

$means
                Stiffness       std r      LCL      UCL Min Max   Q25  Q50
FastT-Fix            7.70 0.6928203 6 6.803692 8.596308 6.4 8.3 7.625 7.85
Meniscus Arrow       6.10 1.3266499 6 5.203692 6.996308 4.7 8.4 5.200 5.95
Vertical Suture      7.75 0.9710819 6 6.853692 8.646308 6.3 8.7 7.225 7.80
                  Q75
FastT-Fix       8.150
Meniscus Arrow  6.475
Vertical Suture 8.600

$comparison
NULL

$groups
                Stiffness groups
Vertical Suture      7.75      a
FastT-Fix            7.70      a
Meniscus Arrow       6.10      b

attr(,"class")
[1] "group"
```

We can see that the group for Vertical Suture and FastT-Fix is the same but the one for Meniscus Arrow is different. This means that the difference in effectiveness between Vertical Suture and FastT-Fix is not significant but the difference in effectiveness between FastT-Fix and Meniscus Arrow is significant. The stiffness value for Vertical Suture is the highest (best) followed by FastT-Fix and then Meniscus Arrow. Therefore, the use of FastT-Fix would be similar to Vertical Suture but significantly better than Meniscus Arrow.

### 8.14 Sea slugs

```r
data("SeaSlugs")
SeaSlugs <- SeaSlugs %>%
  mutate(Time = as_factor(Time))
```

#### a. Fisher's LSD

```r
a3 <- aov(Percent ~ Time, data = SeaSlugs)
summary(a3)
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)    
Time         5 0.6309 0.12618   5.965 0.000607 ***
Residuals   30 0.6346 0.02115                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
model.tables(a3)
```

```
Tables of effects

 Time 
Time
       0        5       10       15       20       25 
 0.26400  0.07383 -0.09400 -0.08833 -0.05250 -0.10300 
```

```r
print(LSD.test(a3, "Time"))
```

```
$statistics
     MSerror Df      Mean       CV  t.value       LSD
  0.02115452 30 0.2716667 53.53838 2.042272 0.1714963

$parameters
        test p.ajusted name.t ntr alpha
  Fisher-LSD      none   Time   6  0.05

$means
```

```
     Percent       std r        LCL       UCL   Min   Max     Q25    Q50
0  0.5356667 0.1687859 6 0.41440050 0.6569328 0.357 0.857 0.47525 0.5000
10 0.1776667 0.1238881 6 0.05640050 0.2989328 0.067 0.333 0.08350 0.1330
15 0.1833333 0.1470397 6 0.06206716 0.3045995 0.000 0.333 0.05350 0.2405
20 0.2191667 0.1383914 6 0.09790050 0.3404328 0.067 0.437 0.10775 0.2335
25 0.1686667 0.1484650 6 0.04740050 0.2899328 0.000 0.412 0.08350 0.1330
5  0.3455000 0.1423921 6 0.22423383 0.4667662 0.125 0.467 0.26050 0.4000
       Q75
0  0.52475
10 0.28300
15 0.28125
20 0.26700
25 0.23350
5  0.45025

$comparison
NULL

$groups
     Percent groups
0  0.5356667      a
5  0.3455000      b
20 0.2191667     bc
15 0.1833333     bc
10 0.1776667     bc
25 0.1686667      c

attr(,"class")
[1] "group"
```

Looking at the letters for the groups, we see that the alphabet for time = 0 i.e. 0 mins after the tide comes in is different from all other times. This means that the proportion of the sea slug larvae that metamorphose is significantly different exactly when tide comes in and after it. Using the same concept, there is no statistically significant difference between time intervals of 5, 10, 15, and 20 mins and 10, 15, 20, and 25 mins but there is statistically significant difference between 5 mins and 25 mins of when tide comes in.

Therefore, 0-5, 0-10, 0-15, 0-20, 0-25, and 5-25 are statistically significant intervals. This means that there are differences that exist between the percent of larvae that metamorphosed in the different water conditions.

### b. Tukey's HSD

```r
data("Alfalfa")
TukeyHSD(a3)
```

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

Fit: aov(formula = Percent ~ Time, data = SeaSlugs)

$Time
              diff        lwr         upr     p adj
5-0   -0.190166667 -0.4455792  0.06524590 0.2397208
10-0  -0.358000000 -0.6134126 -0.10258743 0.0023231
15-0  -0.352333333 -0.6077459 -0.09692077 0.0027831
20-0  -0.316500000 -0.5719126 -0.06108743 0.0085222
25-0  -0.367000000 -0.6224126 -0.11158743 0.0017407
10-5  -0.167833333 -0.4232459  0.08757923 0.3666256
15-5  -0.162166667 -0.4175792  0.09324590 0.4038772
20-5  -0.126333333 -0.3817459  0.12907923 0.6641386
25-5  -0.176833333 -0.4322459  0.07857923 0.3114499
15-10  0.005666667 -0.2497459  0.26107923 0.9999998
20-10  0.041500000 -0.2139126  0.29691257 0.9960188
25-10 -0.009000000 -0.2644126  0.24641257 0.9999978
20-15  0.035833333 -0.2195792  0.29124590 0.9980127
25-15 -0.014666667 -0.2700792  0.24074590 0.9999748
25-20 -0.050500000 -0.3059126  0.20491257 0.9901287
```

Looking at the p value, we see that the interval 10-0, 15-0, 20-0, and 25-0 are the most statistically significant. This means that there are differences that exist between the percent of larvae that metamorphosed in the different water conditions.

### c. Compare

Tukey HSD covers fewer intervals than Fisher LSD with common intervals being 0-10, 0-15, 0-20, 0-25 and Fisher LSD having 0-5 and 5-25 extra. So even though they share common intervals, the overall answer is different. I would prefer to use Tukey HSD in this case since it is more precise and narrow in terms of the intervals provided.