STAT9006: Multi-Variable Data Analysis with *R*
Part III

Outline

Outline

1. Between- and Within-Subjects Design
   - Formatting and exploring the data
   - Assumptions
   - Repeated Measures ANOVA
   - Plot of the main effects and simple main effects
   - Pairwise comparisons
   - Effect size
2. Exercise

Outline

Between- and Within-Subjects Design

Assumptions:

1. Follows a normal distribution;
2. Homogeneity of variance;
3. No difference in the variances of the differences.

The assumption is the sphericity assumption. If the sphericity assumption is:

- Satisfied, then the usual F-test is the most powerful test.
- Violated, then several choices are available - i.e., Greenhouse-Geisser, Huynh-Feldt etc.

Example

Repeated.xlsx contains part of the data for a study of oral condition of patients conducted at the Mid-Michigan Medical Center. The oral conditions of the patients were measured and recorded at the **initial stage**, at the end of the **second week**, at the end of the **fourth week**, and at the end of the **sixth week**. The variables age, initial weight and initial stage of the patients were recorded. Patients were divided into two groups at random: One group received a placebo and the other group received aloe juice treatment. Use the data to test the following hypotheses:

1. H_0: **No difference** exists in the oral condition over time;
2. H_0: **No difference** exists in the oral condition with respect to treatment;
3. H_0: **No difference** exists in the oral condition over time with respect to treatment.

Format the data

1. Check the data properties, missing values, etc.

```
### append a subject ID ... from workshop 03
(n<-dim(Repeated)[1]) # sample size
Patient<-seq(1:n)
library(dplyr)
Repeated<-mutate(Repeated,Patient)

#### ordering variables (if desired)
(cn<-dim(Repeated)[2])
Repeated<-Repeated[,c(cn,1:cn-1)]

# select data of interest
Rep<-select(Repeated,Patient,Treatment=`Treatment group`,
            Baseline=`Oral condition at the initial stage`,
            "Week 02"=`Oral condition at the end of week 02`,
            "Week 04"=`Oral condition at the end of week 04`,
            "Week 06"=`Oral condition at the end of week 06`)

# Convert to long format
library(tidyr)
(Long<-gather(Rep,Time,Oral,3:6))
View(Long)
```

Explore the data

```
# FIRST explore the data
# Step 01: check propoerties
is.factor(Long$Time)
Long$Time<-factor(Long$Time,levels=c("Baseline",
                                     "Week 02","Week 04","Week 06")) #specify order of levels

Long$Treatment<-factor(Long$Treatment,levels=c("Placebo","Aloe Juice"))
is.numeric(Long$Oral)

# Step 02: numerical descriptive statistics
# next line won't work because of missing data
(Stats<-Long %>% group_by(Treatment,Time) %>% summarise("Sample size"=n(),Mean = mean(Oral),
                                                        "Standard deviation"=sd(Oral),
                                                        Median = median(Oral),
                                                        "1st quartile"=quantile(Oral, 0.25),
                                                        "3rd quartile"=quantile(Oral, 0.75), Min=min(Oral),
                                                        Max=max(Oral)))

## locating missing value patients...from Workshop 03
Long[!complete.cases(Long),]
```

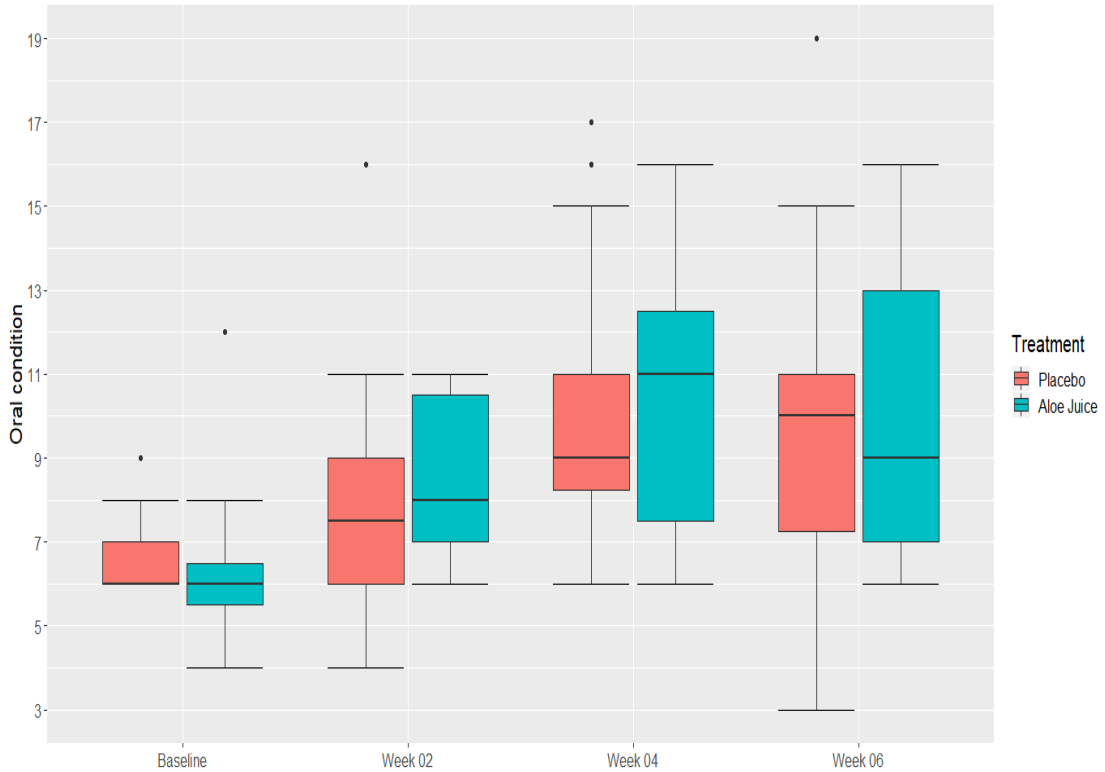
Handling missing data

```
## Two options
# 1. Per protocol (PP) analysis: remove patients from study and
#    complete analysis with  23 patients
Long23<-na.omit(Long) # removes patients with missing data
# 2. Intention to Treat (ITT) analysis: impute data values for patients with
#    missing values - e.g., LOCF
Long$Oral[Long$Patient==22 & Long$Time == "Week 06"]<-
  Long$Oral[Long$Patient==22 & Long$Time == "Week 04"]
Long$Oral[Long$Patient==24 & Long$Time == "Week 06"]<-
  Long$Oral[Long$Patient==24 & Long$Time == "Week 04"]

# now rerun the numerical descriptive statistics (ITT analysis)
Stats<-Long %>% group_by(Treatment,Time) %>% summarise("Sample size"=n(),Mean = mean(Oral),
                                                       "Standard deviation"=sd(Oral),
                                                       Median = median(Oral),
                                                       "1st quartile"=quantile(Oral, 0.25),
                                                       "3rd quartile"=quantile(Oral, 0.75), Min=min(Oral),
                                                       Max=max(Oral))
t(Stats)
```

Graphical descriptive statistics

```
# Step 03: graphical descriptive statistics (ITT analysis)
(g<-ggplot(Long,aes(x=Time, y=Oral,fill=Treatment))+stat_boxplot(geom = "errorbar")+
    geom_boxplot()+labs(x = "", y = "Oral condition"))
(g1<-g+coord_cartesian(ylim=c(3, 19)) + scale_y_continuous(breaks=seq(3,19,2))+
    theme(text = element_text(size=15)))
```



Assumptions (Normally distributed)

2. Check that the data does not violate the assumption of normality:

```
> # SECOND check that the assumptions are not violated (ITT analysis)
> # Step 01: tests of normality
> library(psych)
> Norm<-Long %>% group_by(Treatment, Time) %>% summarise("Sample size"=n(),Mean = mean(Oral),
+                                                        Median = median(Oral), Skewness=skew(Oral),
+                                                        "Normally distributed"=ifelse(
+                                                          shapiro.test(Oral)$p.value>0.05,"Yes","No"))
> t(Norm)
                    [,1]          [,2]          [,3]          [,4]
Treatment           "Placebo"     "Placebo"     "Placebo"     "Placebo"
Time                "Baseline"    "Week 02"     "Week 04"     "Week 06"
Sample size         "14"          "14"          "14"          "14"
Mean                " 6.571429"   " 8.142857"   "10.142857"   " 9.928571"
Median              " 6.0"        " 7.5"        " 9.0"        "10.0"
Skewness            "1.3688125"   "1.0997292"   "0.6321239"   "0.5214610"
Normally distributed "No"         "Yes"         "Yes"         "Yes"
                    [,5]          [,6]          [,7]          [,8]
Treatment           "Aloe Juice"  "Aloe Juice"  "Aloe Juice"  "Aloe Juice"
Time                "Baseline"    "Week 02"     "Week 04"     "Week 06"
Sample size         "11"          "11"          "11"          "11"
Mean                " 6.454545"   " 8.454545"   "10.636364"   "10.090909"
Median              " 6.0"        " 8.0"        "11.0"        " 9.0"
Skewness            "1.4701774"   "0.3343187"   "0.2390035"   "0.5337613"
Normally distributed "No"         "No"          "Yes"         "No"
```

Assumptions (Homogeneity of variances)

Check that the data does not violate the assumption of homogeneity of variances:

```
> # Step 02: Homogeneity of variances
> library(biotools) # #needs data to be (matrix,factor)
> ## return Long to Wide
> Wide<-spread(Long,Time,Oral)
> boxM(Wide[,c(3:6)],Wide$Treatment) #needs data to be (matrix,factor) - no missing values allowed

        Box's M-test for Homogeneity of Covariance Matrices

data:  Wide[, c(3:6)]
Chi-Sq (approx.) = 18.234, df = 10, p-value = 0.05114
```

Sphericity and Repeated Measures ANOVA

3. Depending on whether the assumptions are violated, apply the appropriate test:

```
> # THIRD Create a linear model and perform an ANOVA (ITT analysis)
> # Option 01: If conditions placed on normality and homogeneity of variances are violated
> # then it might be necessary to tranform the data
>
> # Option 02: If conditions are not violated
> library(ez)
> Long$Patient<-as.factor(Long$Patient)
> (res1<-ezANOVA(Long,dv=Oral,wid=Patient,between=Treatment,within=Time)) # won't work if missing data
Warning: Data is unbalanced (unequal N per group). Make sure you specified a well-considered value for
 the type argument to ezANOVA().
$ANOVA
          Effect DFn DFd           F             p p<.05         ges
2      Treatment   1  23  0.05581148 8.153341e-01       0.001317497
3           Time   3  69 13.93535622 3.365113e-07     * 0.216602017
4 Treatment:Time   3  69  0.07344145 9.740364e-01       0.001455026

$`Mauchly's Test for Sphericity`
          Effect         W           p p<.05
3           Time 0.4872109 0.008085645     *
4 Treatment:Time 0.4872109 0.008085645     *

$`Sphericity Corrections`
          Effect       GGe        p[GG] p[GG]<.05       HFe        p[HF] p[HF]<.05
3           Time 0.6749634 1.678638e-05         * 0.7405185 7.601877e-06         *
4 Treatment:Time 0.6749634 9.310951e-01           0.7405185 9.436746e-01
```

Plot of the main effects and simple main effects

4. Plot the means:

```
# FOURTH Effects plot (ITT analysis)
## Main effects
# Time
library(ggpubr)
ggline(Long, y ="Oral", x = "Time",add =c("mean_ci"),size=1)+theme_gray()+
  theme(text = element_text(size=15))+
  labs(y="95% CI of Oral Condition", x="")+
  coord_cartesian(ylim=c(6, 12)) + scale_y_continuous(breaks=seq(6,12,0.5))

# Treatment
ggline(Long, y ="Oral", x = "Treatment",add =c("mean_ci"),size=1)+theme_gray()+
  theme(text = element_text(size=15))+
  labs(y="95% CI of Oral Condition", x="")+
  coord_cartesian(ylim=c(7.5, 10)) + scale_y_continuous(breaks=seq(7.5,10,0.25))

## Simple Main effects (interaction effects)
ggline(Long, y ="Oral", x = "Time",color="Treatment",add =c("mean_ci"),size=1)+theme_gray()+
  theme(text = element_text(size=15))+
  labs(y="95% CI of Oral Condition", x="")+
  coord_cartesian(ylim=c(5, 14)) + scale_y_continuous(breaks=seq(5,14,1))
```

Pairwise comparisons (Main effects)

5. Difference in oral condition between time-points and gender, respectively.

```
> library(emmeans)
> ## Main effects
> # Time
> emmeans(res2, ~Time)%>%pairs(adjust="Tukey")
NOTE: Results may be misleading due to involvement in interactions
 contrast           estimate    SE df t.ratio p.value
 Baseline - Week.02    -1.79 0.673 69 -2.652  0.0477
 Baseline - Week.04    -3.88 0.673 69 -5.757  <.0001
 Baseline - Week.06    -3.50 0.673 69 -5.193  <.0001
 Week.02 - Week.04     -2.09 0.673 69 -3.105  0.0143
 Week.02 - Week.06     -1.71 0.673 69 -2.541  0.0624
 Week.04 - Week.06      0.38 0.673 69  0.564  0.9423

Results are averaged over the levels of: Treatment
P value adjustment: tukey method for comparing a family of 4 estimates
> # Treatment
> emmeans(res2, ~Treatment)%>%pairs(adjust="Tukey")
NOTE: Results may be misleading due to involvement in interactions
 contrast             estimate  SE df t.ratio p.value
 Placebo - Aloe Juice   -0.213 0.9 23 -0.236  0.8153

Results are averaged over the levels of: Time
```

Pairwise comparisons (Simple main effects)

Difference in oral condition between time-points given a fixed treatment.

```
> ## Simple Main effect (interaction effects)
> emmeans(res2, ~Time|Treatment)%>%pairs(adjust="Tukey")
Treatment = Placebo:
 contrast           estimate    SE df t.ratio p.value
 Baseline - Week.02   -1.571 0.893 69 -1.759  0.3018
 Baseline - Week.04   -3.571 0.893 69 -3.998  0.0009
 Baseline - Week.06   -3.357 0.893 69 -3.758  0.0020
 Week.02 - Week.04    -2.000 0.893 69 -2.239  0.1231
 Week.02 - Week.06    -1.786 0.893 69 -1.999  0.1984
 Week.04 - Week.06     0.214 0.893 69  0.240  0.9951

Treatment = Aloe Juice:
 contrast           estimate    SE df t.ratio p.value
 Baseline - Week.02   -2.000 1.008 69 -1.984  0.2038
 Baseline - Week.04   -4.182 1.008 69 -4.149  0.0005
 Baseline - Week.06   -3.636 1.008 69 -3.608  0.0032
 Week.02 - Week.04    -2.182 1.008 69 -2.165  0.1434
 Week.02 - Week.06    -1.636 1.008 69 -1.624  0.3723
 Week.04 - Week.06     0.545 1.008 69  0.541  0.9486

P value adjustment: tukey method for comparing a family of 4 estimates
```

Pairwise comparisons (Simple main effects)

Difference in oral condition between treatments given a fixed time-point.

```
> emmeans(res2, ~Treatment|Time)%>%pairs(adjust="Tukey")
Time = Baseline:
 contrast              estimate   SE df t.ratio p.value
 Placebo - Aloe Juice     0.117 1.22 63  0.096  0.9240

Time = Week.02:
 contrast              estimate   SE df t.ratio p.value
 Placebo - Aloe Juice    -0.312 1.22 63 -0.255  0.7993

Time = Week.04:
 contrast              estimate   SE df t.ratio p.value
 Placebo - Aloe Juice    -0.494 1.22 63 -0.404  0.6874

Time = Week.06:
 contrast              estimate   SE df t.ratio p.value
 Placebo - Aloe Juice    -0.162 1.22 63 -0.133  0.8946
```

Effect size

6. Determine the strength of the result (generalised eta-squared):

Size of effect	Generalized eta squared
Small	$0.02 \leq \eta_G^2 < 0.13$
Medium	$0.13 \leq \eta_G^2 < 0.26$
Large	$\eta_G^2 \geq 0.26$

```
> (res1<-ezANOVA(Long,dv=Oral,wid=Patient,between=Treatment,within=Time)) # won't work if missing data
Warning: Data is unbalanced (unequal N per group). Make sure you specified a well-considered value for
 the type argument to ezANOVA().
$ANOVA
          Effect DFn DFd           F             p p<.05         ges
2      Treatment   1  23  0.05581148 8.153341e-01       0.001317497
3           Time   3  69 13.93535622 3.365113e-07     * 0.216602017
4 Treatment:Time   3  69  0.07344145 9.740364e-01       0.001455026

$`Mauchly's Test for Sphericity`
          Effect         W           p p<.05
3           Time 0.4872109 0.008085645     *
4 Treatment:Time 0.4872109 0.008085645     *

$`Sphericity Corrections`
          Effect       GGe        p[GG] p[GG]<.05       HFe        p[HF] p[HF]<.05
3           Time 0.6749634 1.678638e-05         * 0.7405185 7.601877e-06         *
4 Treatment:Time 0.6749634 9.310951e-01           0.7405185 9.436746e-01
```

Outline

Exercise

An experiment was carried out to test the hypothesis that there is no difference in the maximum heart rate reached by athletes:

1. Using two types of machines (rowing machine and treadmill);
2. With respect to gender (male and female);
3. Interaction between machine type and gender.

Open the dataset, *Heart Rate*.xlsx. Report and interpret your findings using 0.05 level of significance.

This exercise should be answered using the 6 steps outlined in the above slides.