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

Outline

Outline

1. Within-Subjects Designs: Repeated Measures
   - Formatting and exploring the data
   - Assumptions
   - Repeated Measures ANOVA
   - Main effects plot
   - Pairwise comparisons
   - Effect size

2. Exercise

Outline

Within-Subjects Designs: Repeated Measures

Assumptions:

1. Follows a normal distribution
2. Homogeneity of variance (unless the sample sizes are equal)
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.

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,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,2:5))
```

Explore the data

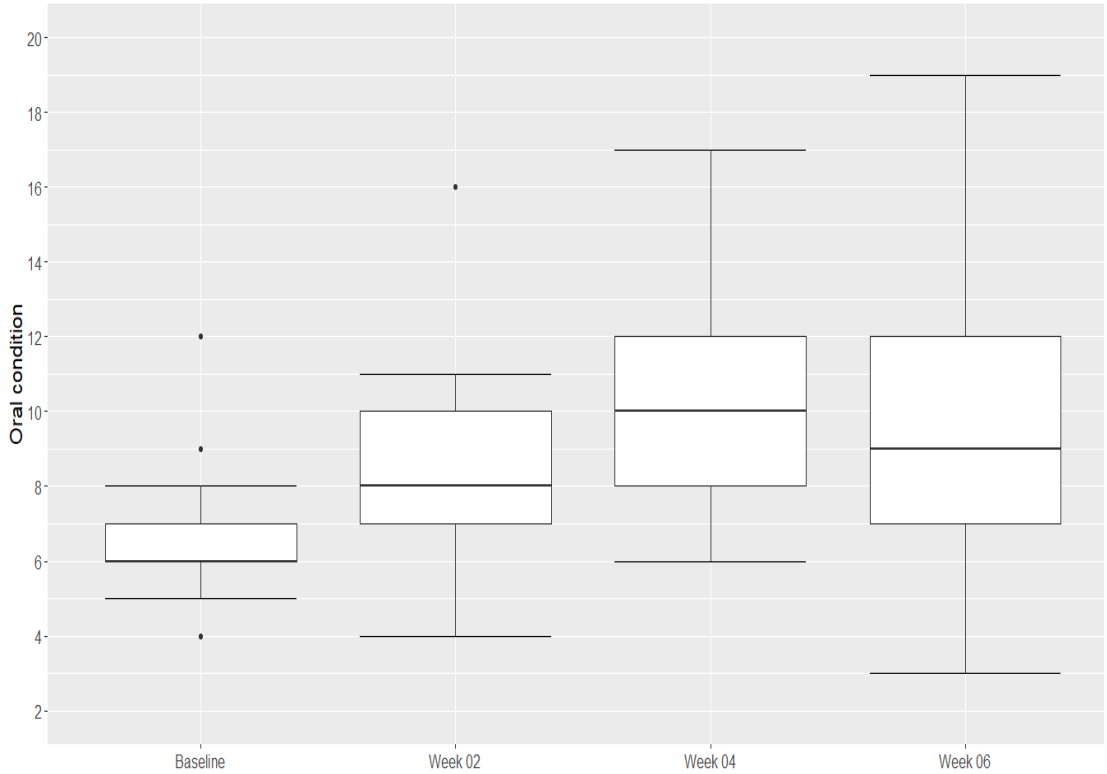
```
# FIRST explore the data
# Step 01: check propoerties
is.factor(Long$Time)
Long$Time<-factor(Long$Time,levels=c("Baseline",
                                     "Week 02","Week 04","Week 06")) #specify order of levels
is.numeric(Long$Oral)

# Step 02: numerical descriptive statistics
# next line won't work because of missing data
(Stats<-Long %>% group_by(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(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))+stat_boxplot(geom = "errorbar")+
    geom_boxplot()+labs(x = "", y = "Oral condition"))
(g1<-g+coord_cartesian(ylim=c(2, 20)) + scale_y_continuous(breaks=seq(2,20,2))+
    theme(text = element_text(size=15)))
```



Assumptions

2. Check that the data does not violate the assumptions of normality and homogeneity of variances:

```
> # Step 03: graphical descriptive statistics (ITT analysis)
> (g<-ggplot(Long,aes(x=Time, y=Oral))+stat_boxplot(geom = "errorbar")+
+     geom_boxplot()+labs(x = "", y = "Oral condition"))
> (g1<-g+coord_cartesian(ylim=c(2, 20)) + scale_y_continuous(breaks=seq(2,20,2))+
+     theme(text = element_text(size=15)))
> # SECOND check that the assumptions are not violated (ITT analysis)
> # Step 01: tests of normality
> library(psych)
> Norm<-Long %>% group_by(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]
Time                "Baseline"   "Week 02"    "Week 04"    "Week 06"
Sample size         "25"         "25"         "25"         "25"
Mean                " 6.52"      " 8.28"      "10.36"      "10.00"
Median              " 6"         " 8"         "10"         " 9"
Skewness            "1.8036582"  "1.0139853"  "0.4871346"  "0.5568237"
Normally distributed "No"        "No"         "No"         "Yes"
>
> # Step 02: Homogeneity of variances
> # There is no between-subjects factor => no need to check for equal variances
```

Sphericity and Repeated Measures ANOVA

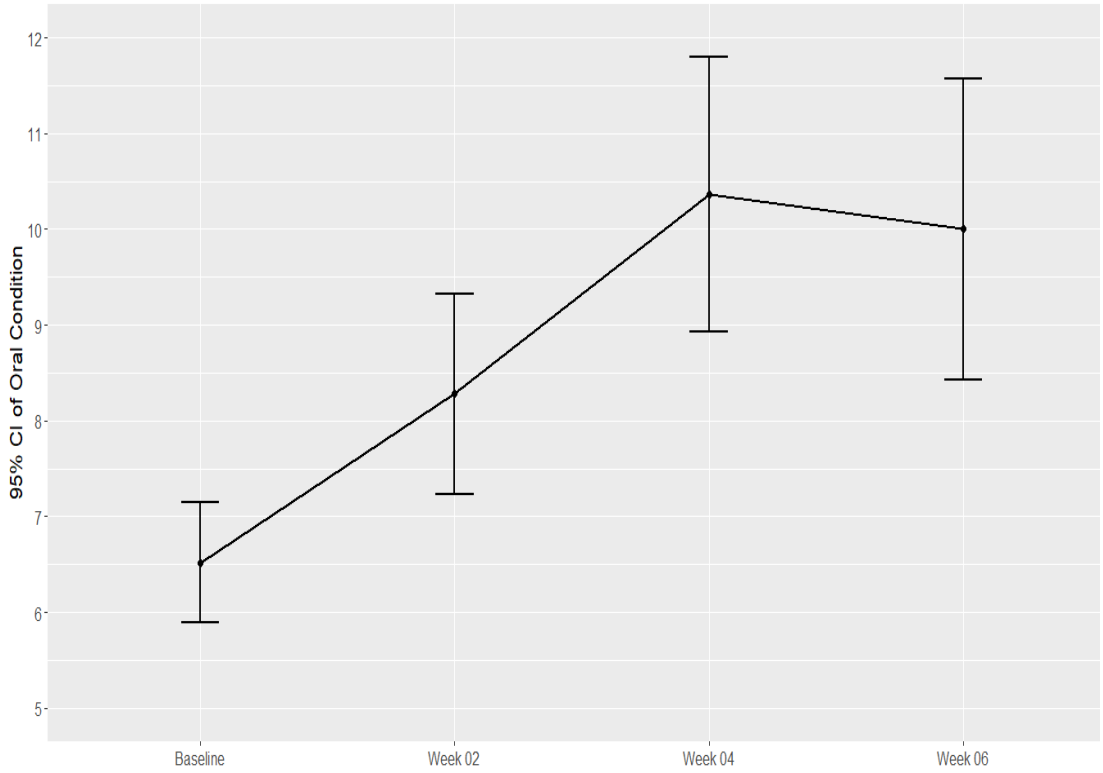
3. Depending on whether the assumptions are violated, apply the appropriate test:

```
> # THIRD Create a linear model and perform an ANOVA (ITT analysis)
> # Option 01: If conditions placed on normality are violated
> # https://www.datanovia.com/en/lessons/friedman-test-in-r/
>
> # Option 02: If conditions are not violated
> library(ez)
> (res1<-ezANOVA(Long,dv=Oral,wid=Patient,within=Time)) # won't work if missing data
Warning: Converting "Patient" to factor for ANOVA.
$ANOVA
  Effect DFn DFd        F            p p<.05       ges
2   Time   3  72 14.49496 1.759854e-07     * 0.2161319

$`Mauchly's Test for Sphericity`
  Effect         W           p p<.05
2   Time 0.4918956 0.006559305     *

$`Sphericity Corrections`
  Effect       GGe        p[GG] p[GG]<.05       HFe        p[HF] p[HF]<.05
2   Time 0.6769597 1.043626e-05         * 0.740008 4.687586e-06         *
```

Main effects plot

4. Plot the means:

```
# FOURTH Main effects plot (ITT analysis)
library(ggpubr)
(mp<-ggline(Long, y ="Oral", x = "Time",add =c("mean_ci"),size=1)+theme_gray())
(g2<-mp+  theme(text = element_text(size=15))+
    labs(y="95% CI of Oral Condition", x="")+
    coord_cartesian(ylim=c(5, 12)) + scale_y_continuous(breaks=seq(5,12,1)))
```



Pairwise comparisons

5.

```
> # FIFTH Pairwise comparisons
> # Tukey
> library(afex)
> # not interested in the aov_car p-values ...
> # ... this is just an approach of doing the pairwise comparisons
> (res2<-aov_car(Oral ~ Time + Error(Patient/Time), data=Long))
Anova Table (Type 3 tests)

Response: Oral
  Effect           df  MSE         F ges p.value
1   Time 2.03, 48.74 7.93 14.49 *** .22  <.0001
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '+' 0.1 ' ' 1

Sphericity correction method: GG
> library(emmeans)
> emmeans(res2, ~Time) %>% pairs(adjust="Tukey")
 contrast            estimate    SE df t.ratio p.value
 Baseline - Week.02     -1.76 0.655 72 -2.685  0.0436
 Baseline - Week.04     -3.84 0.655 72 -5.858  <.0001
 Baseline - Week.06     -3.48 0.655 72 -5.309  <.0001
 Week.02 - Week.04      -2.08 0.655 72 -3.173  0.0116
 Week.02 - Week.06      -1.72 0.655 72 -2.624  0.0508
 Week.04 - Week.06       0.36 0.655 72  0.549  0.9465

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

Effect size

6. Determine the strength of the result. The following table offers a rough guide to the classification of effect size in relation to values of 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,within=Time)) # won't work if missing data
Warning: Converting "Patient" to factor for ANOVA.
$ANOVA
  Effect DFn DFd        F            p p<.05       ges
2   Time   3  72 14.49496 1.759854e-07     * 0.2161319

$`Mauchly's Test for Sphericity`
  Effect         W           p p<.05
2   Time 0.4918956 0.006559305     *

$`Sphericity Corrections`
  Effect       GGe        p[GG] p[GG]<.05       HFe        p[HF] p[HF]<.05
2   Time 0.6769597 1.043626e-05         * 0.740008 4.687586e-06         *
```

Outline

Exercise

- A physician is evaluating a new diet for her patients with a family history of heart disease. To test the effectiveness of this diet, 16 patients are placed on the diet for 4 months. Their weights are measured before, during and after the study.

- Use the dietstudy.xlsx dataset to test the claim of the physician that subject's weight has decreased over the course of the study. If a difference exists, between what time-points does this happen?

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