R Notebook

#load necessary libraries  
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5 v purrr 0.3.4  
## v tibble 3.1.5 v dplyr 1.0.7  
## v tidyr 1.1.4 v stringr 1.4.0  
## v readr 2.0.2 v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(dplyr)  
library(readxl)  
library(ggplot2)  
library(ggpubr)  
library(qqplotr)

##   
## Attaching package: 'qqplotr'

## The following objects are masked from 'package:ggplot2':  
##   
## stat\_qq\_line, StatQqLine

#library(car)  
library(e1071)  
library(nortest)  
library(BSDA)

## Loading required package: lattice

##   
## Attaching package: 'BSDA'

## The following object is masked from 'package:datasets':  
##   
## Orange

# options("scipen"=100, "digits"=6)  
library(reshape2)

##   
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':  
##   
## smiths

# 12.34

## Do labels matter?

### A) Solution:

n = c(130,248,174,68)  
x = c(2.93,3.00,3.01,3.39)  
df = data.frame(label = factor(c("Asian American", "Specific ethnicity", "Ethnicity American", "Other")), n = n, x = x)  
df

## label n x  
## 1 Asian American 130 2.93  
## 2 Specific ethnicity 248 3.00  
## 3 Ethnicity American 174 3.01  
## 4 Other 68 3.39

# Total sample size  
N = sum(n)  
# x bar  
xbar =mean(x)  
# number of categories  
cat\_num = length(unique(df$label)) # or nrow(df)  
# Degree of freedom of numerator (groups)  
dfg = cat\_num - 1  
paste("Degree of freedom of numerator (groups) is",dfg)

## [1] "Degree of freedom of numerator (groups) is 3"

# Degree of freedom of denominator (error)  
dfe = N - cat\_num  
paste("Degree of freedom of denominator (error) is",dfe)

## [1] "Degree of freedom of denominator (error) is 616"

### B) Solution:

df$"(x-xbar)^2" = (df$x - xbar)^2  
df$"n(x-xbar)^2" = df$n \* df$`(x-xbar)^2`  
df

## label n x (x-xbar)^2 n(x-xbar)^2  
## 1 Asian American 130 2.93 0.02325625 3.0233125  
## 2 Specific ethnicity 248 3.00 0.00680625 1.6879500  
## 3 Ethnicity American 174 3.01 0.00525625 0.9145875  
## 4 Other 68 3.39 0.09455625 6.4298250

# sum of squares for groups  
ssg = sum(df$`n(x-xbar)^2`)  
ssg

## [1] 12.05568

### C) Solution:

# Here  
sse = 797.25  
# Mean square for groups  
msg = ssg/dfg  
msg

## [1] 4.018558

# Mean square for error  
mse = sse/dfe  
mse

## [1] 1.294237

# F-statistics  
f = msg/mse  
f

## [1] 3.104963

### D) Solution:

# Corresponding P-value for F statistics  
pf(f, dfg, dfe, lower.tail = FALSE)

## [1] 0.02611366

*Null hypothesis:* H0: means are equal *Alternate hypothesis:* H1: means are not equal *Since p value is less than the .05 significance level therefore we can reject the null hypothesis.*

### E) Solution:

*Even though the sample mean are pretty close to each other, due to the fact the sample sizes differ by a good margin, the means are not equal. It supports the conclusion obtained in part D.*

# 12.54

## A dandruff study.

### A) Solution:

file\_name = "ex12-54dandruff.xls"  
df = read\_xls(file\_name)

## New names:  
## \* `` -> ...4  
## \* `` -> ...5

head(df)

## # A tibble: 6 x 8  
## OBS Treatment Flaking ...4 ...5 Pyrl Keto Placebo  
## <dbl> <chr> <dbl> <lgl> <lgl> <dbl> <dbl> <dbl>  
## 1 1 PyrI 17 NA NA 17 17 30  
## 2 2 PyrI 16 NA NA 16 17 29  
## 3 3 PyrI 18 NA NA 18 17 31  
## 4 4 PyrI 17 NA NA 17 15 31  
## 5 5 PyrI 18 NA NA 18 14 29  
## 6 6 PyrI 16 NA NA 16 16 29

summary(df)

## OBS Treatment Flaking ...4   
## Min. : 1.0 Length:355 Min. :14.00 Mode:logical   
## 1st Qu.: 89.5 Class :character 1st Qu.:16.00 NA's:355   
## Median :178.0 Mode :character Median :17.00   
## Mean :178.0 Mean :17.87   
## 3rd Qu.:266.5 3rd Qu.:18.00   
## Max. :355.0 Max. :32.00   
##   
## ...5 Pyrl Keto Placebo   
## Mode:logical Min. :14.0 Min. :14.00 Min. :26.00   
## NA's:355 1st Qu.:16.0 1st Qu.:15.00 1st Qu.:28.00   
## Median :17.0 Median :16.00 Median :29.00   
## Mean :17.3 Mean :16.03 Mean :29.39   
## 3rd Qu.:18.0 3rd Qu.:17.00 3rd Qu.:31.00   
## Max. :21.0 Max. :18.00 Max. :32.00   
## NA's :134 NA's :249 NA's :327

# df = df[-c(4,5)]  
df = df[c(1,2,3)]  
head(df)

## # A tibble: 6 x 3  
## OBS Treatment Flaking  
## <dbl> <chr> <dbl>  
## 1 1 PyrI 17  
## 2 2 PyrI 16  
## 3 3 PyrI 18  
## 4 4 PyrI 17  
## 5 5 PyrI 18  
## 6 6 PyrI 16

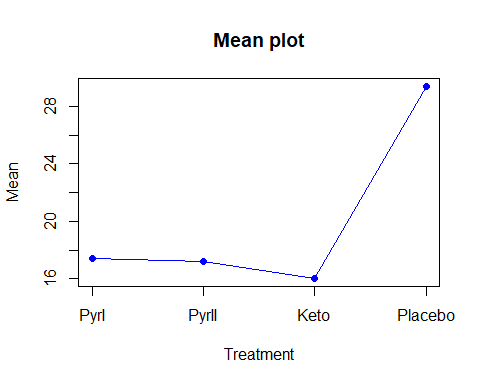
# df = na.omit(df)  
summary(df)

## OBS Treatment Flaking   
## Min. : 1.0 Length:355 Min. :14.00   
## 1st Qu.: 89.5 Class :character 1st Qu.:16.00   
## Median :178.0 Mode :character Median :17.00   
## Mean :178.0 Mean :17.87   
## 3rd Qu.:266.5 3rd Qu.:18.00   
## Max. :355.0 Max. :32.00

report = group\_by(df, Treatment) %>%  
 summarise(  
 count = n(),  
 mean = mean(Flaking, na.rm = TRUE),  
 sd = sd(Flaking, na.rm = TRUE),  
 se = sd(Flaking, na.rm = TRUE)/sqrt(n())  
 ) %>% arrange(desc(count))  
report

## # A tibble: 4 x 5  
## Treatment count mean sd se  
## <chr> <int> <dbl> <dbl> <dbl>  
## 1 PyrI 112 17.4 1.14 0.108   
## 2 PyrII 109 17.2 1.35 0.130   
## 3 Keto 106 16.0 0.931 0.0904  
## 4 Placebo 28 29.4 1.59 0.301

plot(report$mean, xaxt="n", type = "o", main = "Mean plot", col = "blue", pch = 19, xlab = "Treatment", ylab = "Mean")  
axis(1, at = seq(1, 4, 1), labels = report$Treatment)



### B) Solution:

*From the given data, H0 and H1 for ANOVA*

*H0: Population mean scores for four groups are same*

*H1: Population mean scores for four groups are not same*

# one sample anova test  
aov(Flaking~Treatment,df)

## Call:  
## aov(formula = Flaking ~ Treatment, data = df)  
##   
## Terms:  
## Treatment Residuals  
## Sum of Squares 4151.428 501.868  
## Deg. of Freedom 3 351  
##   
## Residual standard error: 1.195752  
## Estimated effects may be unbalanced

# ANOVA table  
anova(aov(Flaking~Treatment,df))

## Analysis of Variance Table  
##   
## Response: Flaking  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Treatment 3 4151.4 1383.81 967.82 < 2.2e-16 \*\*\*  
## Residuals 351 501.9 1.43   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Total degree of freedom  
sum(anova(aov(Flaking~Treatment,df))[1])

## [1] 354

# Total sum of squares  
sum(anova(aov(Flaking~Treatment,df))[2])

## [1] 4653.296

*Test statistics of ANOVA test, F = 967.82 . Degree of freedom for group is 3 and Degree of freedom for error is 351.*

*The P-value is almost zero which indicates a very strong evidence against H0.*

*So, not all means are same.*

# 13.43

## A comparison of plant species under low water conditions.

file\_name = "ex13-43plants1.xls"  
df = read\_xls(file\_name, sheet = 2, range = cell\_cols("A:D"))  
head(df)

## # A tibble: 6 x 4  
## obs species water pctnit  
## <dbl> <dbl> <dbl> <dbl>  
## 1 1 1 1 3.64  
## 2 2 1 1 3.5   
## 3 3 1 1 3.51  
## 4 4 1 1 3.14  
## 5 5 1 1 3.1   
## 6 6 1 1 3.08

df$species = as.factor(df$species)  
df$water = as.factor(df$water)  
summary(df)

## obs species water pctnit   
## Min. : 1.00 1:63 1:36 Min. :0.719   
## 1st Qu.: 63.75 2:63 2:36 1st Qu.:1.643   
## Median :126.50 3:63 3:36 Median :2.626   
## Mean :126.50 4:63 4:36 Mean :2.403   
## 3rd Qu.:189.25 5:36 3rd Qu.:3.103   
## Max. :252.00 6:36 Max. :3.958   
## 7:36

report = group\_by(df, species,water) %>%  
 summarise(  
 # count = n(),  
 mean = mean(pctnit, na.rm = TRUE),  
 sd = sd(pctnit, na.rm = TRUE),  
 # se = sd(pctnit, na.rm = TRUE)/sqrt(n())  
 )# %>% arrange(desc(count))

## `summarise()` has grouped output by 'species'. You can override using the `.groups` argument.

report

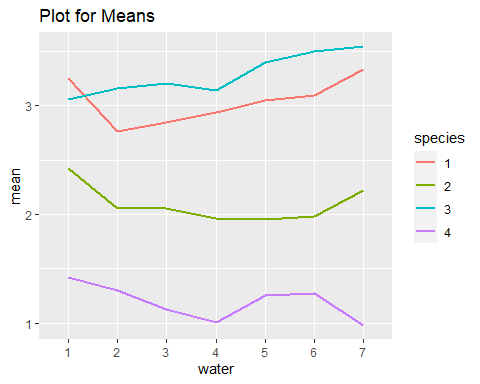
## # A tibble: 28 x 4  
## # Groups: species [4]  
## species water mean sd  
## <fct> <fct> <dbl> <dbl>  
## 1 1 1 3.25 0.229   
## 2 1 2 2.76 0.0666  
## 3 1 3 2.84 0.233   
## 4 1 4 2.94 0.0709  
## 5 1 5 3.05 0.0909  
## 6 1 6 3.10 0.0815  
## 7 1 7 3.33 0.248   
## 8 2 1 2.42 0.165   
## 9 2 2 2.05 0.145   
## 10 2 3 2.05 0.148   
## # ... with 18 more rows

### A) Solution:

# mean for each species by water  
dcast(df, species~water, value.var='pctnit', mean)

## species 1 2 3 4 5 6 7  
## 1 1 3.254333 2.763556 2.842889 2.936222 3.051889 3.096333 3.3334444  
## 2 2 2.421556 2.050222 2.052444 1.967333 1.956000 1.983889 2.2184444  
## 3 3 3.058889 3.154111 3.200333 3.141889 3.395556 3.496111 3.5436667  
## 4 4 1.423000 1.303667 1.125333 1.008667 1.258444 1.271222 0.9787778

ggplot(data = report) +  
 geom\_line(aes(x = water, y = mean, colour = species, group = species), lwd = 1) + ggtitle("Plot for Means")



*Means change across multiple species. So we can say that there is a main effect for species.*

*Means change across multiple water quantities. So we can say that there is a main effect for water*

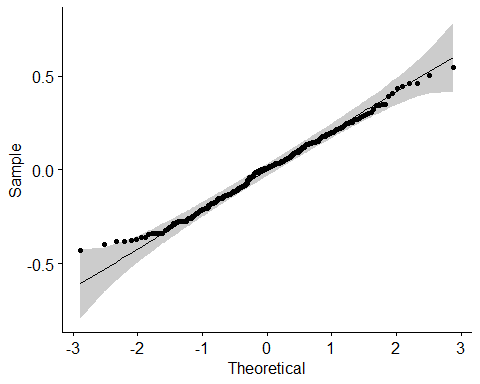
*Plot for means patterns do not seem to be parallel. So we can assume that there is some kind of interaction.*

### B) Solution:

# Standard deviation for each species by water  
dcast(df, species~water, value.var='pctnit', sd)

## species 1 2 3 4 5 6  
## 1 1 0.2286586 0.06657348 0.2332876 0.07093444 0.0908769 0.08153680  
## 2 2 0.1653868 0.14537003 0.1481183 0.22025667 0.1570995 0.28949760  
## 3 3 0.1524946 0.33236330 0.2340951 0.29650567 0.2532642 0.34367261  
## 4 4 0.1737613 0.26607518 0.1230173 0.13097996 0.2488590 0.07948707  
## 7  
## 1 0.2481678  
## 2 0.1237822  
## 3 0.3116220  
## 4 0.2090262

# Build the linear model  
model <- lm(pctnit~species\*water, data = df)  
# Create a QQ plot of residuals  
ggqqplot(residuals(model))



*In the QQ plot, as all the points fall approximately along the reference line, we can assume normality.*

*Also, the sample sizes are equal and sample standard deviations are similar, assumption of equal population variance may be reasonable and the pooled procedure could be used*

###C) Solution:

# two-way ANOVA test without interaction effect (NOT APPLICABLE IN THIS CASE)  
# anova(aov(pctnit~species+water,df))

# Two-way ANOVA with interaction effect  
# anova(aov(pctnit~species\*water, df))  
anova(aov(pctnit~species+water + species:water,df))

## Analysis of Variance Table  
##   
## Response: pctnit  
## Df Sum Sq Mean Sq F value Pr(>F)   
## species 3 172.392 57.464 1301.3166 < 2.2e-16 \*\*\*  
## water 6 2.587 0.431 9.7628 1.486e-09 \*\*\*  
## species:water 18 4.745 0.264 5.9692 1.047e-11 \*\*\*  
## Residuals 224 9.891 0.044   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

*Null hypotheses:*

H0(species): main effect due to species is statistically insignificant.

H0(water): main effect due to water is statistically insignificant.

H0(interaction): interaction effect due to species and water is statistically insignificant.

*As all the P-values are less than .05 level of significance, we can reject all the null hypotheses and conclude main effects due to species and water as well as interaction are statistically significant.*