TP4_Taina_KERRIOU.R

taina

2022-05-24

```
##TP4-Basice of Probability

##Exercise: Foshing competition

#The two categroy is not paired
#H0: The

library(readxl)
ANNEX_CONCOURS <- read_excel("C:/Users/taina/Downloads/ANNEX_CONCOURS.xlsx")
View(ANNEX_CONCOURS)
#H0: The mean of ASTICOT and TEIGNE is the same
#H1: The mean of ASTICOT and TEIGNE is different
ASTICOT= ANNEX_CONCOURS$SCORE_POIDS[ANNEX_CONCOURS$APPAT=="ASTICOT"]
ASTICOT

##  [1] 1473 1589 1959 1249 1434 1582 1280 2150 1309 1558 1303 1687 1517 1693
1454
## [16] 1594  850 1402  960 2189 1680 1466 1716 1326 1467 1933 1900 1508 1366
1663
## [31] 1278 1745

TEIGNE= ANNEX_CONCOURS$SCORE_POIDS[ANNEX_CONCOURS$APPAT=="TEIGNE"]
TEIGNE

##  [1] 2005 1542 1446 2784 1796 2702 2910 1131 1670  550 2675  996 2457 2795
2350
## [16] 2407 2483 1800 1227 2364 1518 1450  750 1149 2025 2050 1894 1967

#1 Test Shapiro
#H0:Les données suivent une loi normale
#H1:Les données ne suivent pas une loi normale
shapiro.test(ANNEX_CONCOURS$SCORE_POIDS[ANNEX_CONCOURS$APPAT=="ASTICOT"])

##
##  Shapiro-Wilk normality test
##
## data:  ANNEX_CONCOURS$SCORE_POIDS[ANNEX_CONCOURS$APPAT == "ASTICOT"]
## W = 0.97152, p-value = 0.5424

#p-value>0,05 donc on admet H0 les données ASTICOT suivent une loi nomale
shapiro.test(ANNEX_CONCOURS$SCORE_POIDS[ANNEX_CONCOURS$APPAT=="TEIGNE"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ANNEX_CONCOURS$SCORE_POIDS[ANNEX_CONCOURS$APPAT == "TEIGNE"]
## W = 0.96707, p-value = 0.5045
```

```
#p-value>0,05 donc on admet H0 les données TEIGNE suivent une loi nomale

# Test fisher
#H0:Les variances sont égales
#H1:Les variances sont différentes
```

```
var.test(ASTICOT,TEIGNE)
```

```
##
##  F test to compare two variances
##
## data:  ASTICOT and TEIGNE
## F = 0.20262, num df = 31, denom df = 27, p-value = 3.62e-05
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.0953345 0.4224209
## sample estimates:
## ratio of variances
##          0.2026175
```

```
#p-value<0,05 Les variances sont différentes

#2
t.test(ASTICOT,TEIGNE,alternative="two.sided",paired=FALSE,var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  ASTICOT and TEIGNE
## t = -2.6027, df = 36.425, p-value = 0.0133
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -620.90378  -77.16765
## sample estimates:
## mean of x mean of y
##  1540.000  1889.036
```

```
# p-value= 0.0133 donc pvalue<0,05 On rejette H0 et on admet H1 au risque de
5%

#a
mean(ASTICOT)
```

```
## [1] 1540
```

```r
mean(TEIGNE)
```

```
## [1] 1889.036
```

```r
t.test(TEIGNE,ASTICOT,alternative="greater",paired=FALSE,var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  TEIGNE and ASTICOT
## t = 2.6027, df = 36.425, p-value = 0.006648
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  122.6957      Inf
## sample estimates:
## mean of x mean of y
##  1889.036  1540.000
```

```r
#p-value<2xalpha On rejette H0 et on admet H1 au risque de 10%

#b
t.test(TEIGNE,ASTICOT,alternative="less",paired=FALSE,var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  TEIGNE and ASTICOT
## t = 2.6027, df = 36.425, p-value = 0.9934
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf 575.3757
## sample estimates:
## mean of x mean of y
##  1889.036  1540.000
```

```r
#pvalue=0.9934>0,10 si on obtenait ce resultat on pourra admettre que les
deux moyennes sont pareils

##Exercice2 Study: Length of stay depending on the holiday site

#H0:site 1 et site 2 sont indépendants
#H1:site 1 et site 2 sont liées statistiquement

site1=c(42,31,87,120)
site2=c(58,49,72,101)
time=c("1 or 2 jours","Half day","Day","More than one day")
holiday=data.frame(site1,site2,row.names=time)

khideux1<-chisq.test(holiday)
khideux1
```

```
##
##  Pearson's Chi-squared test
##
## data:  holiday
## X-squared = 9.6586, df = 3, p-value = 0.0217
```

```
#p-value<0.05 on rejette H0 et on peut admettre H1 au risque 5%
khideux1$residuals
```

```
##                         site1      site2
## 1 or 2 jours      -1.1313708  1.1313708
## Half day          -1.4230249  1.4230249
## Day                0.8411582 -0.8411582
## More than one day  0.9037378 -0.9037378
```

```
#Il y a une correlation entre toutes les valeurs entre les deux sites. Nous pouvons cependant dire que le site 1 et 2 sont plus indépendant sur day et more than one day que 1 ou 2 jours et half day

##Exercice Clubs and baits used, during the fishing competition
tableau=table(ANNEX_CONCOURS$CLUB,ANNEX_CONCOURS$APPAT)
#H0:site 1 et site 2 sont indépendants
#H1:site 1 et site 2 sont liées statistiquement

khideux2<-chisq.test(tableau)
khideux2
```

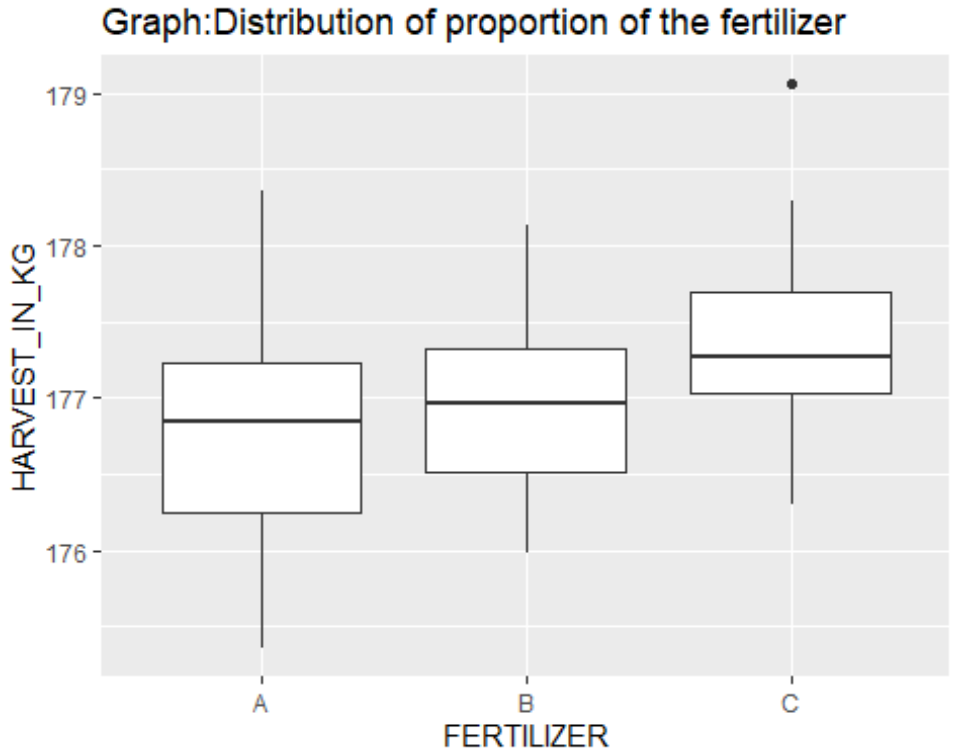
```
##
##  Pearson's Chi-squared test
##
## data:  tableau
## X-squared = 6.4286, df = 2, p-value = 0.04018
```

```
#p-value<0.05 on rejette H0 et on peut admettre H1 au risque 5%
khideux2$residuals
```

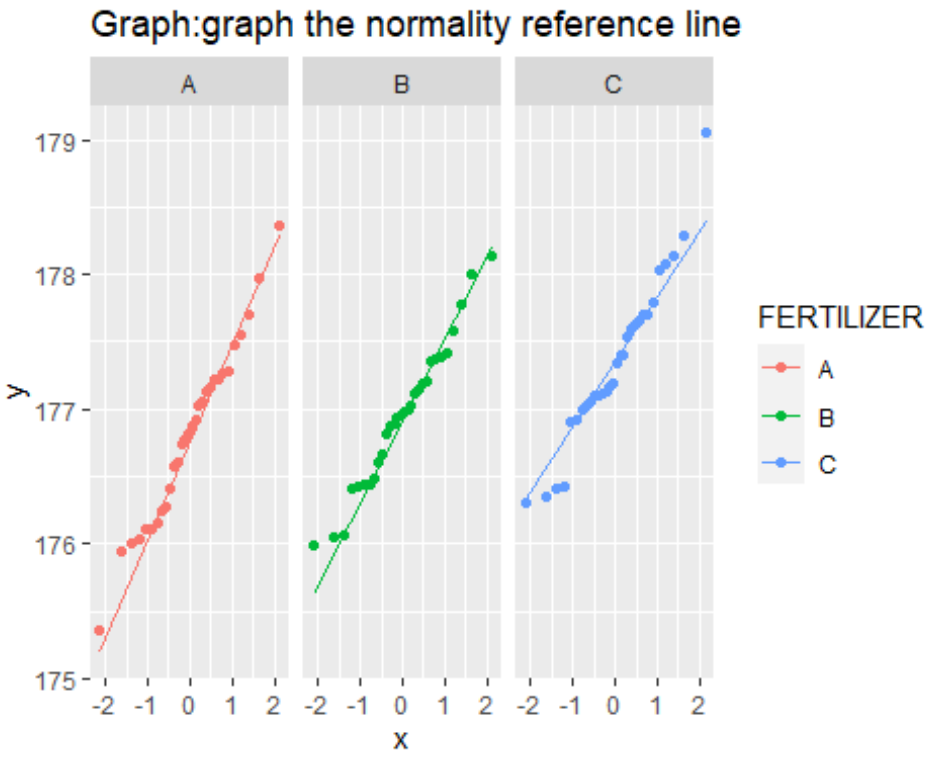
```
##
##                       ASTICOT     TEIGNE
##   AAPPMA_GARDOIS   -1.3416408  1.4342743
##   CLERMONT_FISHING  0.8944272 -0.9561829
##   MTP_PECHE         0.6324555 -0.6761234
```

```
library(readxl)
Annex_WHEAT <- read_excel("C:/Users/taina/Downloads/Annex_WHEAT.xlsx")
View(Annex_WHEAT)

library(ggplot2)


levels(as.factor(Annex_WHEAT$FERTILIZER))
```

```
## [1] "A" "B" "C"

#Independence of observations: OK
#No significant outliers:
ggplot(Annex_WHEAT,aes(y=HARVEST_IN_KG,x=FERTILIZER))+
  geom_boxplot()+
  labs(title="Graph:Distribution of proportion of the fertilizer")
```




```
#Normality
ggplot(Annex_WHEAT,aes(sample=HARVEST_IN_KG,fill=FERTILIZER,color=FERTILIZER)
)+
  geom_qq()+
  geom_qq_line()+
  facet_wrap(~FERTILIZER)+scale_fill_manual(values=c("red","yellow","blue"))+
  labs(title="Graph:graph the normality reference line")
```




```
#Each fertilizer are aproximatly distributed in an normal way
#Homogeneity of variances
#H0 : the variances of all groups are equal
#H1: At least one group has a variance that differs from the others
bartlett.test(HARVEST_IN_KG ~ FERTILIZER, data=Annex_WHEAT)
```

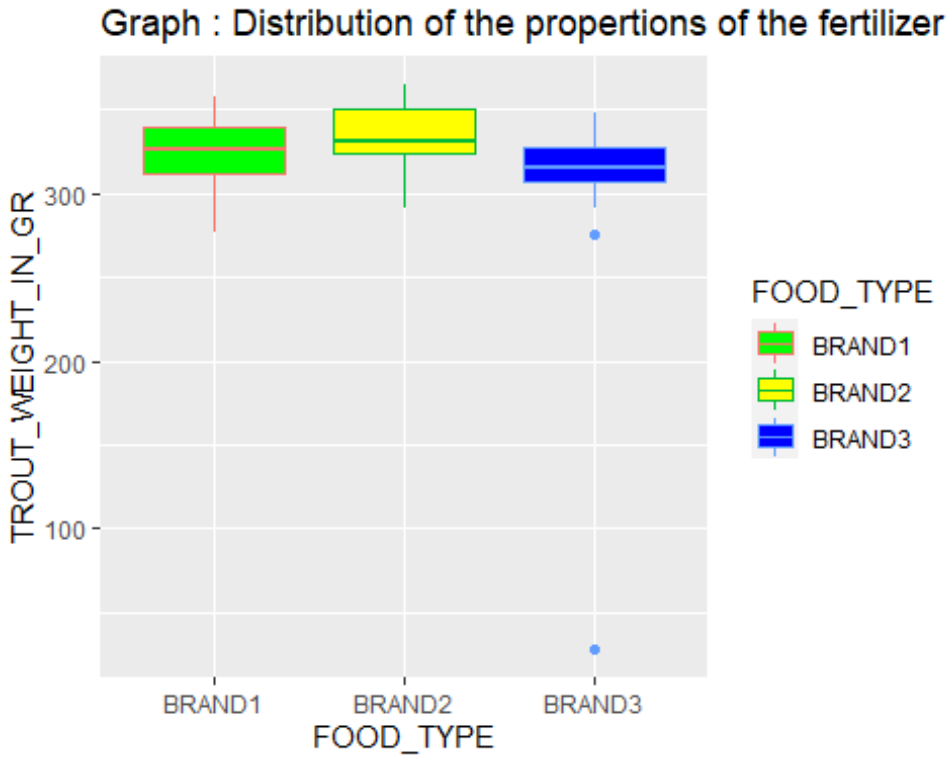
```
##
##  Bartlett test of homogeneity of variances
##
## data:  HARVEST_IN_KG by FERTILIZER
## Bartlett's K-squared = 1.2144, df = 2, p-value = 0.5449
```

```
#p-value=0.5339>0.05 so the test is not significative. We accept H0 with a
risk 5%
#The hypotheses are all verified and validated in this case we can do an
ANOVA test

#H0: Means between groups are similar
#H1:There is at least one group whose the mean is different from the others
My_AOV_model= aov(HARVEST_IN_KG ~ FERTILIZER, data=Annex_WHEAT)
My_AOV_model
```

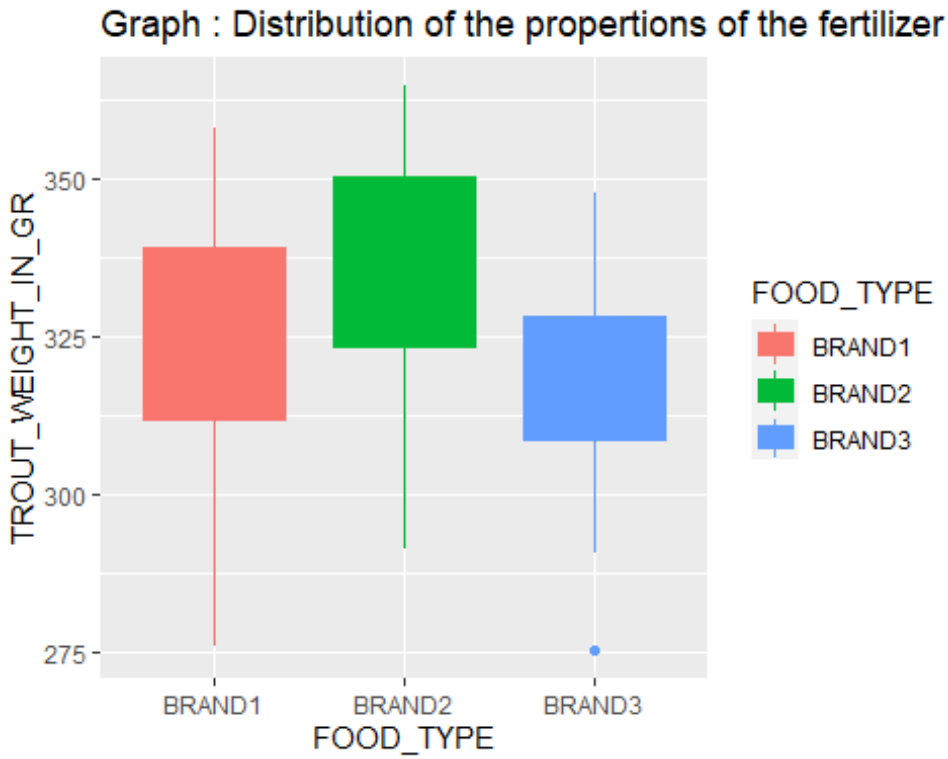
```
## Call:
##    aov(formula = HARVEST_IN_KG ~ FERTILIZER, data = Annex_WHEAT)
##
## Terms:
##                 FERTILIZER Residuals
```

```
## Sum of Squares     4.68106  32.59891
## Deg. of Freedom          2        87
##
## Residual standard error: 0.6121275
## Estimated effects may be unbalanced

summary(My_AOV_model)

##             Df Sum Sq Mean Sq F value  Pr(>F)
## FERTILIZER   2   4.68  2.3405   6.246 0.00292 **
## Residuals   87  32.60  0.3747
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#Pr(>F)=0.00292 so it's < alpha risk at 1%
#test significatif: we accept H1

#Conclusion
#So there is at least one group where the mean is different from the other

##Multiple comparisons (post-hoc)

TukeyHSD(My_AOV_model)

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = HARVEST_IN_KG ~ FERTILIZER, data = Annex_WHEAT)
##
## $FERTILIZER
##          diff         lwr       upr     p adj
## B-A 0.1454667 -0.23140188 0.5223352 0.6289440
## C-A 0.5398333  0.16296479 0.9167019 0.0027617
## C-B 0.3943667  0.01749812 0.7712352 0.0381134

#H0: There is no difference between the means of the two groups
#H1: There is a difference between the means of the two groups
#We choose a risk of 5%
#The group C-A and C-B are the only one where p-value<0.05 so for these two
groups there is a difference between their means


##Exercise: Feeding study in a trout farm

library(readxl)
Annex_TROUTS <- read_excel("C:/Users/taina/Downloads/Annex_TROUTS.xlsx")
View(Annex_TROUTS)

levels(as.factor(Annex_TROUTS$FOOD_TYPE))#give the level of modality of
different type of food
```

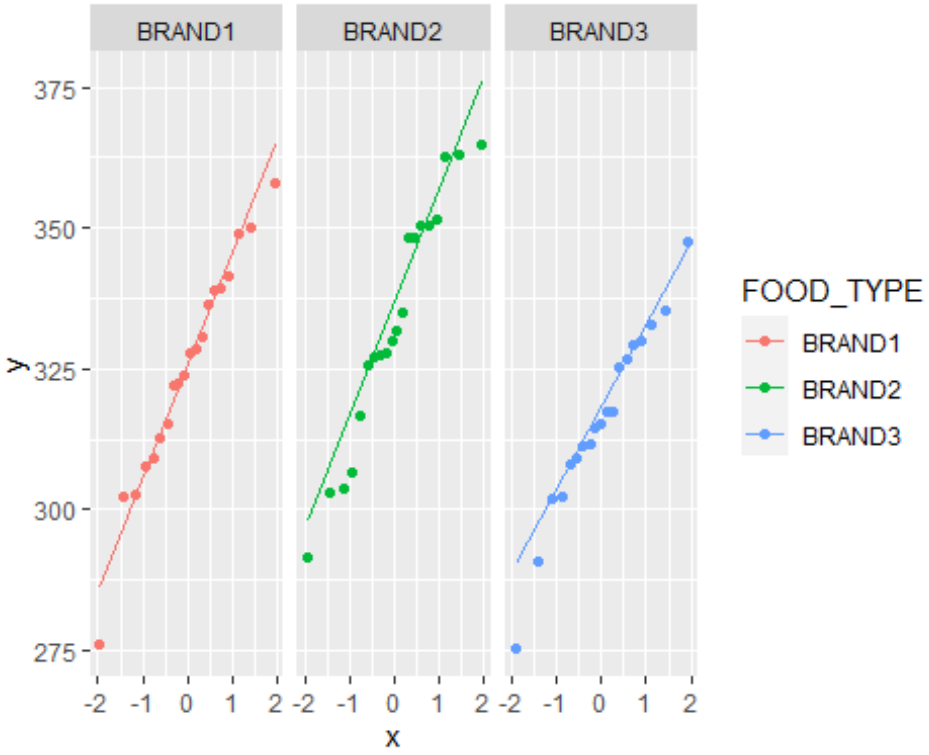
```
## [1] "BRAND1" "BRAND2" "BRAND3"

# H0: Means between groups are similar
# H1: There is at least one group whose the mean is different from the others

# ANOVA of 1 facteur

# independance of observations : ok (no test to do)
# each observation are owned by a group

# to do an anova test we must verify those conditions :
# No significant outliers :
ggplot(Annex_TROUTS, aes(y=TROUT_WEIGHT_IN_GR, x=FOOD_TYPE,
fill=FOOD_TYPE,color=FOOD_TYPE)) +
  geom_boxplot() +
  labs(title="Graph : Distribution of the propertions of the fertilizer",
       x="FOOD_TYPE", y= "TROUT_WEIGHT_IN_GR") +
  scale_fill_manual(values=c("green","yellow","blue"))
```




```
# we delete the outliers
tab_bis=Annex_TROUTS[-53,]
# we plot again the fonction
ggplot(tab_bis, aes(y=TROUT_WEIGHT_IN_GR, x=FOOD_TYPE,
fill=FOOD_TYPE,color=FOOD_TYPE)) +
  geom_boxplot() +
  labs(title="Graph : Distribution of the propertions of the fertilizer",
       x="FOOD_TYPE", y= "TROUT_WEIGHT_IN_GR")
```




```
# Normality with qqplot :

ggplot(tab_bis, aes(sample = TROUT_WEIGHT_IN_GR, fill=FOOD_TYPE, color=
FOOD_TYPE))+
  geom_qq()+
  geom_qq_line()+
  facet_wrap(~FOOD_TYPE)+
  scale_fill_manual(values=c("red","yellow","blue"))
```




```
# each brand food are distributed in an approximately normal way

# Homogeneity of variances :
# H0: the variances of all groups are equal
# H1: At least one group has a variance that differs from the others
bartlett.test(TROUT_WEIGHT_IN_GR ~ FOOD_TYPE, data= tab_bis)

##
##  Bartlett test of homogeneity of variances
##
## data:  TROUT_WEIGHT_IN_GR by FOOD_TYPE
## Bartlett's K-squared = 1.0812, df = 2, p-value = 0.5824

# p-value = 1.751e-08 < 5% risk so the test is non significative
# we accept H0 at the risk beta

# The hypotheses are all verified and validated in this case so we can do an
ANOVA test

My_AOV_model= aov(TROUT_WEIGHT_IN_GR ~ FOOD_TYPE, data= tab_bis)
summary(My_AOV_model)

##             Df Sum Sq Mean Sq F value Pr(>F)
## FOOD_TYPE    2   2935  1467.3   3.877 0.0265 *
## Residuals   56  21195   378.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Pr( > F) = 0.0.0538 so < alpha risk at 1%
# test significatif : we accept H1

# conclusion :
# at least one groupe where the mean is different from the others

# Multiple comparisons (post-hoc)

# we set a risk alpha = 5%
TukeyHSD(My_AOV_model)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = TROUT_WEIGHT_IN_GR ~ FOOD_TYPE, data = tab_bis)
##
## $FOOD_TYPE
##                      diff       lwr       upr     p adj
## BRAND2-BRAND1   8.555000  -6.25638 23.366380 0.3526560
## BRAND3-BRAND1  -8.798947 -23.80395  6.206053 0.3417921
## BRAND3-BRAND2 -17.353947 -32.35895 -2.348947 0.0196752
```

```r
# Tukey's method will compare all possible pair combinations to study if
there is a significant difference
# in their means!

#The hypotheses for the Tukey method are:
#H0: There is no difference between the means of the two groups
#H1: There is a difference between the means of the two groups

# the group 3-2 is the only groups where p-value < 0.05,
# it is then a significatif test, so we accept H1 with an alpha risk of 0.05

# So we can conclued that the brand 2 is the one with a different bigger mean
because the difference is positive
# the brand 2 is the most efficient brand
```