TP Analyse de la variance (ANOVA) à un facteur

Sujet

Mathilde Boissel

25/10/2021

Table of Contents

Ce TP est prévu sous le logiciel R.

Étude de cas 1 : Étude du rendement en jus de 3 variétés de pommes

Pour chaque variété, 4 arbres sont échantillonnés.
On se pose la question suivante : **Existe-t-il une différence significative entre ces 3 variétés quant à la moyenne des rendements ?**

Lire les données

On peut construire le tableau de données (`data.frame`) à la main.

- Pour ce faire, recopier les commandes suivantes dans R.

```
## data en groupe
pommes_by_group <- data.frame(
  Golden = c(48,46,52,50),
  Delicious = c(52,50,49,49),
  Jonagold = c(53,51,55,57)
)
## data en long
pommes <- data.frame(
  rendement = c(48,46,52,50,52,50,49,49,53,51,55,57),
  variete = factor(rep(c("Golden","Delicious","Jonagold"), rep(4,3)))
)
```

La construction de ce `data.frame` peut passer par bien d'autres procédures.
Par exemple il existe la commande `gl`, pour la construction de facteur, qui peut être utilisé comme suit :

```
variete_facteur <- gl(n = 3, k = 4, label = c("Golden","Delicious","Jonagold"
))
```

Il y a 3 modalités pour le facteur "variete", 4 répétitions et on donne des noms aux modalités du facteur (pour plus d'informations, voir `help("gl")`).

- Noter la différence de structure en affichant les données.

Visualisation format Groupe

Golden	Delicious	Jonagold
48	52	53
46	50	51
52	49	55
50	49	57

Visualisation format Long

rendement	variete
48	Golden
46	Golden
52	Golden
50	Golden
52	Delicious
50	Delicious
49	Delicious
49	Delicious
53	Jonagold
51	Jonagold
55	Jonagold
57	Jonagold

Les données sont souvent collectées par groupe, alors que pour les traiter dans R nous aurons besoin d'une seule observation par ligne. Il faut alors ranger les données avec chaque variable (ici "rendement" et "variete") en colonne et une ligne par observation.

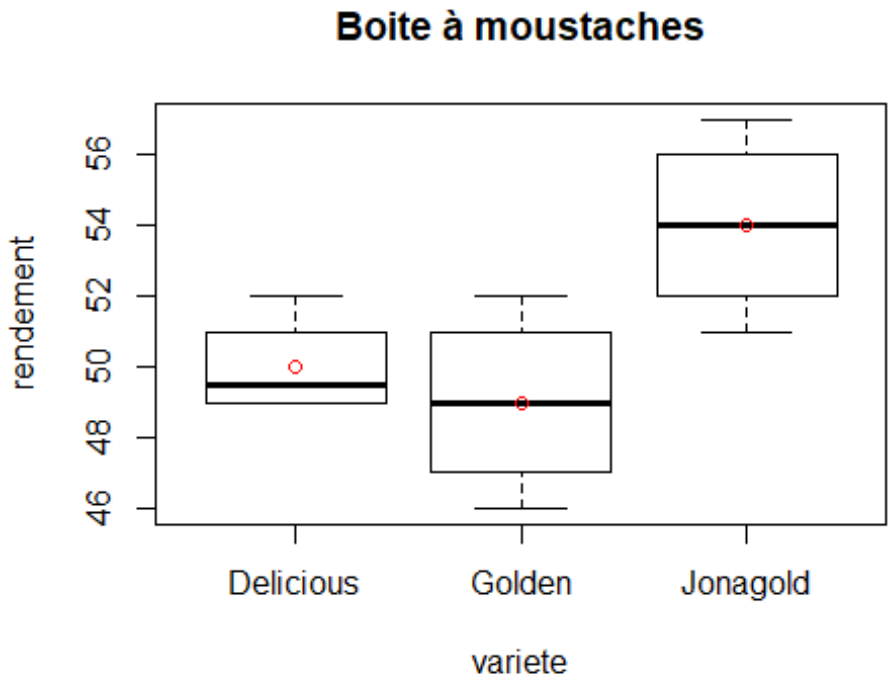
Pour la suite nous utiliserons le jeu de données nommé "pommes", au format **Long**.

- S'assurer que la variable à expliquer ("rendement") est numérique et que la variable explicative "variete" est bien un facteur.

`str` procède à l'affichage compact de la structure interne d'un objet R.
Les fonctions `is.[type]` testent le `[type]` d'un objet R et retourne une valeur booléenne.
La fonction `class` retourne la classe (i.e. le type) d'un objet R.

Visualiser et résumer les données

- Visualiser les données avec le graphique adéquat. `?boxplot`



On affiche les boxplots des observations correspondant à chaque niveau (modalité) du facteur "variete". Si les boxplots sont décalés, on peut soupçonner un effet du facteur.

- Tester les commandes suivantes pour avoir un résumé numérique des données.

```r
summary(pommes$rendement)

summary(pommes$variete) # summary.factor(pommes$variete)

table(pommes$variete)
```

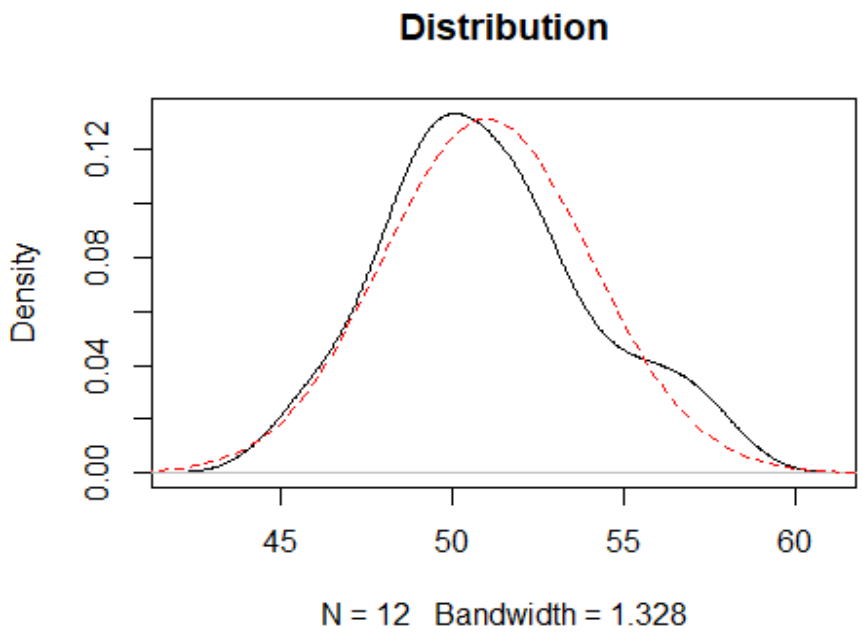
`summary` est une fonction générique qui va invoquer la méthode `summary.[type]` adapté à l'object R d'un certain `[type]`. Ici pour "variete" qui est un facteur, la fonction `summary(pommes$variete)` applique en fait `summary.factor(pommes$variete)` sans qu'on ait besoin de faire la nuance nous-même.
Si on sait à l'avance que l'on souhaite compter des effectifs, on peut aussi directement choisir la fonction `table`.

- Ces commandes indique-t-elle s'il y a des valeurs manquantes ?

- Résumer les données pour chaque groupe :
  effectif, moyenne, ecart-type, données manquantes.
  `?aggregate ?tapply`

- Visuellement, vérifier la normalité de la variable d'intérêt.

```r
## With a density plot
plot(
  density(…),
  col = "black",
  lty = 1,
  main = "Distribution"
)
mr = mean(…, na.rm = TRUE)
sdr = sd(…, na.rm = TRUE)
x_norm = seq(-4,4,length=100) * sdr + mr
lines(
  x = x_norm,
  y = quelle_fonction_ici(x = x_norm, mean = mr, sd = sdr),
  col = "red", lty = 2
)

# Completer les « … »
# quelle_fonction_ici est la fonction de densité de la loi Normal, à trouver.
```



Si la validation visuelle n'est pas concluante, on peut utiliser le test de Normalité de Shapiro-Wilk comme suit :

```
shapiro.test(x = pommes$rendement)
##  Shapiro-Wilk normality test
##
## data:  pommes$rendement
## W = 0.97643, p-value = 0.9653
```

On rappelle l'hypothèse nulle $H0$: la variable est normalement distribuée. Si la p-value est inférieure à un niveau alpha choisi (par exemple 0.05), Alors l'hypothèse nulle est rejetée. Pour supposer la normalité des résidus, il est donc nécessaire d'obtenir une p-value > 0.05. (Ici p-value = 0.96, on ne rejette pas $H0$)

Test ANOVA

Formule

La valeur $rendement_{ik}$ est la valeur du rendement pour le k-ème arbre de la varieté i et μ_i est la moyenne inconnue des rendements pour la variété i.

L'ANOVA est un modèle régression linéaire qui fait l'hypothèse d'une moyenne par modalité du facteur étudié. Son modèle peut s'écrire :

$$rendement_{ik} = \mu_i + \varepsilon_{ik}$$

$$\Leftrightarrow$$

$$rendement_{ik} = \mu + \alpha_i + \varepsilon_{ik}$$

avec

- μ moyenne générale inconnue de tous les rendements,
- α_i effet principal additif par rapport à μ dû à la i-ème modalité du facteur,
- $\varepsilon_{ik} \sim \mathcal{N}(0, \sigma^2)$ les résidus (les écarts entre les observations et les moyennes des groupes auxquels elles sont relatives),
- avec $\sum_{i=1}^{n} \alpha_i = 0$.

Dans R cette formule est écrite ainsi : `rendement ~ variete`.

Réalisation de l'analyse de la variance sous R

- Lire la documentation de la fonction `aov` dans R : `?stats::aov`
- Lire la documentation de la fonction `lm` dans R : `?stats::lm`
- Lire la documentation de la fonction `anova` dans R : `?stats::anova`
- Expliquer la différence et noter ce qui nous intéresse pour répondre à la question initiale.
- Réaliser l'ANOVA qui permet de tester l'effet du facteur "variete" sur la mesure "rendement", afficher le résumé statistique et conclure.

On obtient le résultat suivant

```
my_anova <- fonction_a_choisir(…)
my_anova
```

```
## Call:
##    fonction_a_choisir(…)
##
## Terms:
##                  variete Residuals
## Sum of Squares       56        46
## Deg. of Freedom       2         9
##
## Residual standard error: 2.260777
## Estimated effects may be unbalanced
```

Pour faire le parallèle avec le tableau ANOVA (cf. tableau du cours), l'objet "my_anova" affiche un résultat de la forme :

	Facteur A	Residuelle
SCE	SCE_A	SCE_R
d.d.l.	ddl_A	ddl_R

Et l'estimation de l'écart-type résiduel : $\hat{\sigma} = \sqrt{CM_R}$.

En réalité, "my_anova" est une liste avec 13 composantes dont on trouve les noms en faisant `names(my_anova)`.

Pour aller plus loin dans le tableau d'analyse de la variance, il suffit d'afficher le résumé statistique comme suit :

```
summary(my_anova)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## variete      2     56  28.000   5.478 0.0278 *
## Residuals    9     46   5.111
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

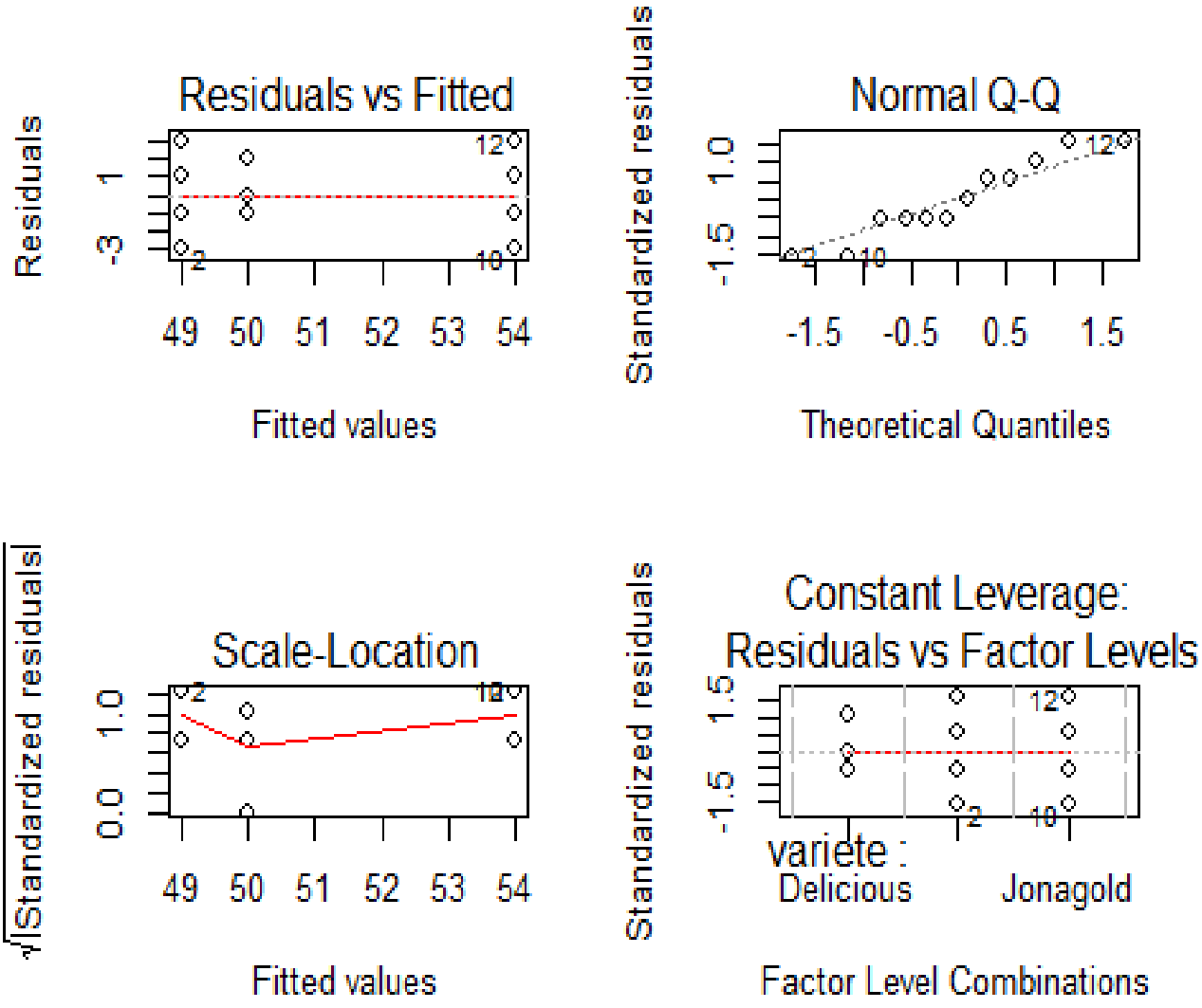
Ici les informations sont sous la forme suivante :

	d.d.l	SCE	CM	Statistique F	p.value
Facteur A	ddl_A	SCE_A	CM_A	$F_{\{obs\}}$	$P(F > F_{\{obs\}})$
Résiduelle	ddl_R	SCE_R	CM_R		

avec bien sûr $F \sim \mathcal{F}(ddl_A, ddl_R)$

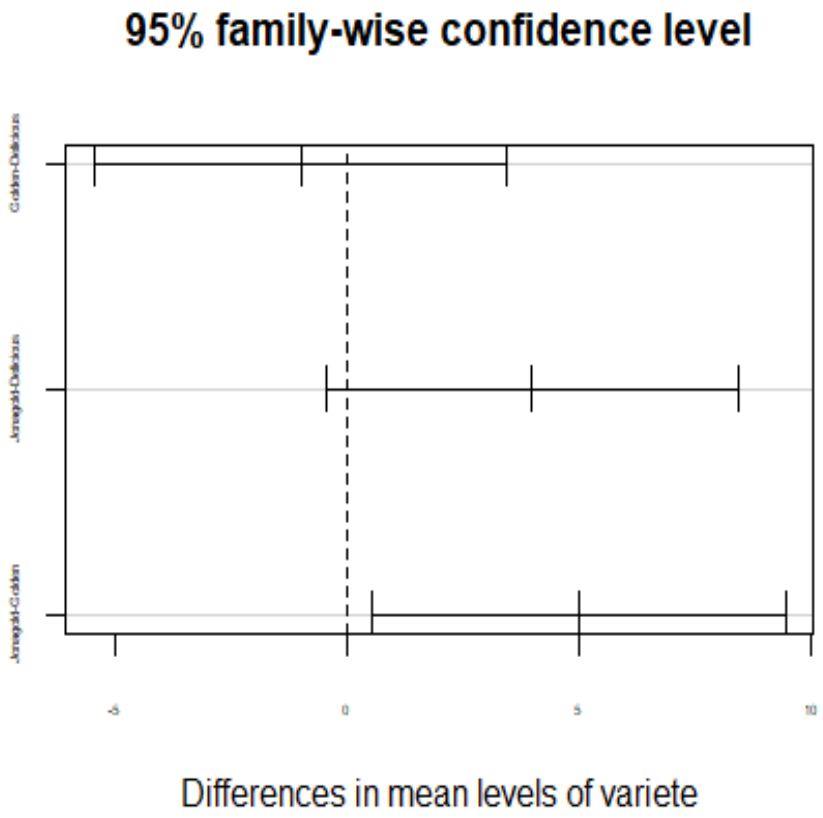
Diagnostique

- Réaliser 4 graphiques diagnostiques de l'anova et commenter.



Test post-hoc

- Lire la documentation du test HSD de Tukey dans R avec `?TukeyHSD`.
- Réaliser ce test post-hoc.



- Quelle est la particularité des p-valeurs retournées ?
- Quelle est la conclusion du test post-hoc ?
- Pour l'exercice, tester ces autres commandes.

N.B. : IMPORTANT

Dans la pratique, UNE SEULE méthode, et UNE SEULE correction, serait à choisir A PRIORI (de même que le seuil alpha). Attention faire une multitude de tests, et choisir A POSTERIORI le meilleur (celui qui nous arrange) serait du "p-hacking" (comprendre manipuler les données ou les résultats pour avoir une bonne p-valeur).

Étude de cas 2 : Étude de l'IMC

Pour plusieurs pays, des centres hospitaliers ont recrutés des patients pour étudier leur poids, la prise en charge de leur diabète et leur profil génétique.
Afin de réaliser une étude au niveau européen, les données sont réunies dans une cohorte pour être mise en commun.
Pour réaliser une étude équilibrée, on souhaite vérifier si les critères de recrutement des patients soient homogènes entre les pays. On se pose la question suivante :
Existe-t-il une différence significative entre ces pays quant à la moyenne des IMC ?

- Lire les données. `?read.csv()`

- Formater les données (avec le bon type) et faire en sorte que la modalité "France" soit la référence.

Quand un modèle utilise une variable factorielle, il est important de connaitre la référence. Par défaut ce sera la première modalité (ordre alphanumérique).

N.B. : Quand on réalise une analyse "cas vs control", si on a un phénotype codé 1/0, tout va bien, les contrôles (0) seront bien pris en référence. Mais on voit le problème si les phénotypes étaient indiqués avec le code "cas" et "control", ici l'ordre alphabétique ne donnera pas la modalité "control" en référence.

- Résumer les données pour en voir un aperçu numérique.

```
summary(cohort$bmi)

table(cohort$Pays, useNA = "always")
```

- D'où vient le problème avec les données renseignées dans la colonne "Pays"?

- Tester les commandes suivantes et expliquer les sorties.

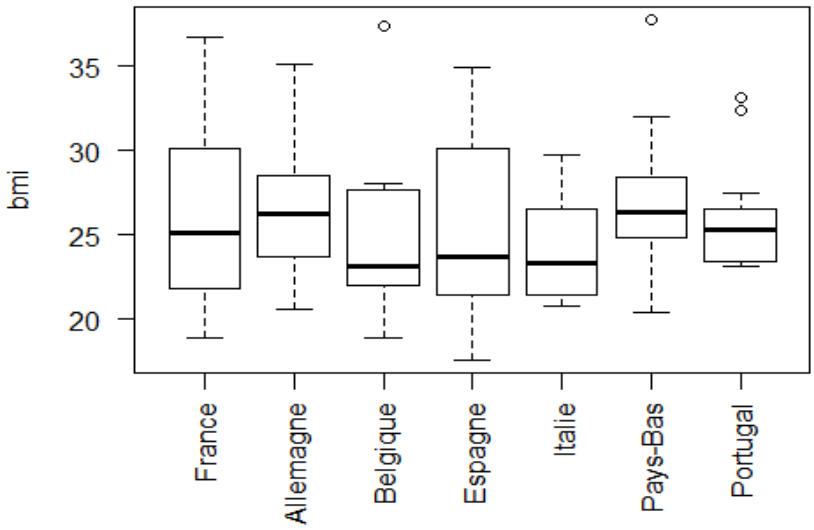
```
sum(is.na(cohort$Pays))

## [1] 1

sum(cohort$Pays=="")

## [1] NA

sum(cohort$Pays%in%"")

## [1] 1
```

- Traiter les valeurs manquantes

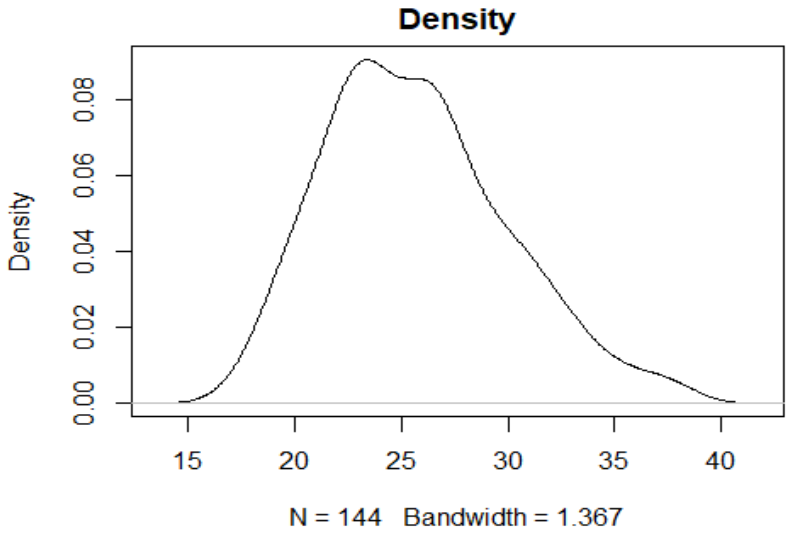
Noter que le fait d'enlever des observations n'impacte pas les niveaux de facteurs définis au moment du formatage. On peut définir explicitement les niveaux à considérer avec le paramètre "levels" de la fonction "factor" (ainsi que la façon de les afficher avec "labels").

N.B. : Notre choix ici est de retirer les individus ayant des données manquantes. Remarquer qu'une autre possibilité pourrait être d'imputer ces données (l'imputation est par exemple souvent utilisée pour les données génotypiques par exemple).

- Visualiser les données (avec tous les noms de pays lisible).



- Contrôler la normalité de la variable d'intérêt.



- Que faire si notre variable d'intérêt n'est pas normalement distribuée ?

Appliquer transformation peut résoudre notre problème : ici on va étudier log(bmi).

- Réaliser l'ANOVA

On rappelle les hypothèses
$H0$: L'égalité des moyennes entre tous les pays.
contre
$H1$: Au moins un des pays à une moyenne différente des autres.

- Donner une réponse à la question initiale.

- Comment se comporte l'ANOVA, dans R, si notre jeu de données possède des données manquantes ?

Sources

- Le contenu de ce TP s'est basé sur un extrait du support écrit par Christophe Chesneau.
- le livre R Cookbook, 2nd Edition, James (JD) Long, Paul Teetor, 2019-09-26
- les pages suivantes :

https://statistique-et-logiciel-r.com/anova-a-un-facteur-partie-1/

https://statistique-et-logiciel-r.com/anova-a-un-facteur-partie-2-la-pratique/

Pour aller plus loin :

- Pas utilisé ici, mais il existe aussi la fonction `Anova` provenant du package car : https://www.rdocumentation.org/packages/car/versions/3.0-10/topics/Anova
- Pas utilisé ici, mais il existe aussi ce package `DescTools` qui propose la fonction `PostHocTest` pour réaliser les tests post-hoc : https://www.rdocumentation.org/packages/DescTools/versions/0.99.38/topics/PostHocTest