

Travail Pratique 6

Analyse de variance à un critère de classification : ANOVA

STT1039
BIOLOGIE QUANTITATIVE

Réalisé par
Marc Pépino



Département des sciences de l'environnement

1. Objectifs du TP6

Le travail pratique 6 (TP 6) a pour objectif général d'initier les étudiants à l'analyse de variance à un critère de classification (ANOVA) réalisée à l'aide du logiciel R. À la fin du TP 6, l'étudiant doit savoir : comparer les moyennes de plus de deux échantillons, vérifier les conditions d'application de l'ANOVA, effectuer un test post-hoc de comparaisons multiples à l'aide des tests H.S.D. de Tukey.

Note : Tout ce qui est mis en rouge et en gras dans les exemples (encadrés) sont des noms donnés par l'utilisateur. Ces noms sont à modifier en conséquence si les exemples sont utilisés avec d'autres bases de données.

Les tests statistiques présentés dans les exemples seront réalisés à partir de la base de données *TP6_Exemples.csv* disponible sur le portail du cours :

```
> #Importation de la base de données
> tab <- read.csv("D:/bureau/TP6_Exemples.csv",sep=";",dec=".",h=T)
>
> #Visualisation de la structure du tableau de données
> str(tab)
'data.frame': 100 obs. of 3 variables:
 $ x : num 197 210 193 187 192 ...
 $ y : num 754 648 593 621 534 ...
 $ f : Factor w/ 3 levels "A","B","C": 1 1 1 1 1 1 1 1 1 1 ...
```

OPTIONNEL :

```
#Une base de données similaire peut être obtenue à partir du code suivant:
#Construction des variables
x <- c(rnorm(60,200,10),rnorm(40,210,7))
f <- c(rep("A",25),rep("B",35),rep("C",40))
y <- rep(NA,100)
y[1:60] <- 55 + 3*x[1:60] + rnorm(60,0,50)
y[61:100] <- 60 + 3*x[61:100] + rnorm(40,0,60)

#Construction du tableau de données
tab <- as.data.frame(x)
tab$y <- y
tab$f <- as.factor(f)
```

Pour plus de détails sur les tests statistiques avec R, vous pouvez consulter, entre autres, **Crawley (2015)**, **Millot (2018)**, ou **Bertrand et Maumy-Bertrand (2010)**.

2. Analyse de variance à un critère de classification : ANOVA

La fonction `aov()` permet de faire directement la comparaison des moyennes de plus de deux échantillons indépendants. La fonction `summary()` permet d'obtenir le tableau de sortie de l'analyse.

```
> #Exemple 1
> #test ANOVA
> a <- aov(tab$x~tab$f)
> summary(a)
          Df Sum Sq Mean Sq  F value    Pr(>F)
tab$f      2  4602.3   2301.15   27.035 4.658e-10 ***
Residuals 97  8256.3    85.12
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> #INTERPRETATION :
> #Dispersion totale ou variation totale
> SST <- sum((tab$x-mean(tab$x))^2)
> SST
[1] 12858.62
>
> #Dispersion intragroupe ou variation résiduelle
> A <- sum((tab[tab$f=="A",]$x-mean(tab[tab$f=="A",]$x))^2)
> B <- sum((tab[tab$f=="B",]$x-mean(tab[tab$f=="B",]$x))^2)
> C <- sum((tab[tab$f=="C",]$x-mean(tab[tab$f=="C",]$x))^2)
> SCE <- A+B+C
> SCE
[1] 8256.316
>
> #Dispersion intergroupe ou variation due au facteur
> SCI <- SST - SCE
> SCI
[1] 4602.308
>
> #valeur de F
> F = (SCI/2) / (SCE/97)
> F
[1] 27.0353
```

Note : le symbole « ~ » signifie « en fonction de ». On compare les moyennes de x en fonction des catégories de f .

Important : la variable explicative ou indépendante (f dans l'exemple) doit absolument être lue comme un facteur par le logiciel (utiliser `as.factor(tab$f)` si f est numérique).

Interprétation de la sortie de l'analyse :

La première ligne du tableau correspond à la **dispersion intergroupe ou variation due au facteur** et la deuxième ligne correspond à la **dispersion intragroupe ou variation résiduelle**. La colonne **Df** correspond au nombre de degré de liberté. Le nombre de degré de liberté est égal à $k - 1$ pour la dispersion intergroupe et à $n - k$ pour la dispersion intragroupe, où n et k sont respectivement la taille de l'échantillon et le nombre de groupes à être comparés. La colonne **Sum Sq** correspond à la somme des moindres carrés pour la dispersion intergroupe (première ligne) et intragroupe (deuxième ligne). La colonne **Mean Sq** est la division de la colonne **Sum Sq** par la colonne **Df**. La colonne **F value** est la valeur de F de la loi Fischer-Snedecor correspondant au rapport de la variance intergroupe sur la variance intragroupe. La colonne **Pr(>F)** correspond à la probabilité d'avoir une variance intergroupe supérieure à la variance intragroupe : il s'agit de la valeur de p du test. Ici (exemple 1), la valeur de p nous indique que le test est significatif au seuil de signification de 0.1% ($p < 0.001$).

Il est également possible d'effectuer le même test à partir du tableau de données et en utilisant l'argument **data=** à l'intérieur de la fonction **aov()**.

```
> #Exemple 2
> #test ANOVA
> aa <- aov(x~f,data=tab)
> summary(aa)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
tab\$f	2	4602.3	2301.15	27.035	4.658e-10 ***
Residuals	97	8256.3	85.12		

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Dans les deux exemples précédents, l'analyse est réalisée à partir d'un tableau de données (tab). L'exemple 3 ci-dessous montre comment obtenir les mêmes résultats à partir de variables se présentant sous forme de vecteur :

```
> #Exemple 3
> #test ANOVA
> f <- tab$f
> x <- tab$x
> aaa <- aov(x~f)
> summary(aaa)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
tab\$f	2	4602.3	2301.15	27.035	4.658e-10 ***
Residuals	97	8256.3	85.12		

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note : les exemples 1 à 3 sont trois codes de programmation différents pour faire la même analyse.

3. Condition d'applications du test ANOVA

3.1. Condition d'indépendance des échantillons

Il n'existe pas de tests statistiques simples permettant de vérifier cette condition. Seule l'expérience permettra de vérifier si cette condition est respectée.

3.2. Condition d'homogénéité des variances

La fonction **var.test()**, vue au TP5, permet de faire directement la comparaison des variances de deux échantillons indépendants (test de F). Se référer au TP5 pour plus de détails sur cette fonction. Dans le cas de la comparaison des variances de plus de deux échantillons, il est possible de comparer la plus grande variance à la plus petite et de vérifier si elles sont significativement différentes en utilisant cette même fonction (test de F). Il est toutefois conseillé de faire des validations graphiques de cette condition, par exemple en regardant la distribution des résidus en fonction des valeurs prédites (voir plus de détails dans le cours).

```
> #Exemple 4
> #Savoir quelle est la plus grande et la plus petite variance
> aggregate(tab$x,list(tab$f),var)
  Group.1      x
1     A 94.04135
2     B 111.12926
3     C  56.94688
>
> #La plus grande variance est-elle significativement différente de la plus petite?
> var.test(tab[tab$f=="B",]$x,tab[tab$f=="C",]$x)

F test to compare two variances

data:  tab[tab$f == "B", ]$x and tab[tab$f == "C", ]$x
F = 1.9515, num df = 34, denom df = 39, p-value = 0.04491
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.015527 3.806936
sample estimates:
ratio of variances
 1.951455
```

Informations complémentaires :

La fonction **bartlett.test()** permet de faire la comparaison de plus de deux variances simultanément. Le test de Bartlett est toutefois très sensible à la condition de normalité. Le test non paramétrique de Levene permet également de comparer plus de deux variances lorsque la condition de normalité n'est pas respectée. La fonction **leveneTest()** de la bibliothèque **car**

permet de faire un tel test. Dans la littérature scientifique, le test de F (exemple 4) est toutefois préféré à ces deux derniers tests.

```
> #Exemple 5
> #Comparaison des variances avec le test de Bartlett
> bartlett.test(tab$x~tab$f)

Bartlett test of homogeneity of variances

data: tab$x by tab$f
Bartlett's K-squared = 4.1545, df = 2, p-value = 0.1253

>
> #Comparaison des variances avec le test de Levene
> library(car)
> leveneTest(tab$x~tab$f)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  2   1.297  0.2780
      97
```

3.3. Condition de normalité

L'ANOVA suppose que chaque échantillon se distribue normalement. Tout comme la régression linéaire, la visualisation des résidus sous forme d'histogramme (voir TP2) ou avec le diagramme quantile-quantile (« Normal QQ-plot ») permet de juger de cette condition. Le test de Shapiro-Wilk peut être utilisé pour appuyer l'interprétation graphique. La fonction **residuals()** permet de récupérer les résidus de l'ANOVA.

```
#Exemple 6
a <- aov(tab$x~tab$f)
par(mfrow=c(1,2))
hist(residuals(a),col="grey",main="")
qqnorm(residuals(a),main="")
qqline(residuals(a))
shapiro.test(residuals(a))
```

3.4. Violation des conditions d'application : test non paramétrique de Kruskal-Wallis

L'ANOVA est une analyse robuste aux violations des conditions d'applications. Toutefois, lorsque les conditions d'application sont fortement violées, il est possible de procéder au test non paramétrique de Kruskal-Wallis pour comparer plusieurs moyennes. La fonction **kruskal.test()** permet de réaliser un tel test.

```
> #Exemple 7
> #test de Kruskal-Wallis
> kruskal.test(tab$x~tab$f)
```

Kruskal-Wallis rank sum test

data: tab\$x by tab\$f

Kruskal-Wallis chi-squared = 37.4182, df = 2, p-value = 7.495e-09

4. Comparaisons multiples : tests H.S.D. de Tukey

Lorsque l'hypothèse nulle de l'ANOVA est rejetée, il est nécessaire de faire un test post-hoc afin de connaître quelles moyennes sont significativement différentes. La fonction **TukeyHSD()** correspond aux tests H.S.D. de Tukey et compare deux à deux toutes les moyennes des échantillons.

```
> #Exemple 8
> #test ANOVA
> mod <- aov(tab$x~tab$f)
> #test post-hoc de Tukey
> TukeyHSD(mod)
```

Tukey multiple comparisons of means
95% family-wise confidence level

Fit: aov(formula = tab\$x ~ tab\$f)

\$`tab\$f`

	diff	lwr	upr	p adj
B-A	0.5679492	-5.182426	6.318324	0.9700024
C-A	14.1720714	8.573448	19.770695	0.0000001
C-B	13.6041222	8.521461	18.686784	0.0000000

Note : La fonction *TukeyHSD()* est appliquée à l'objet *mod* qui contient les résultats de l'ANOVA.

Interprétation de la sortie de l'analyse :

Dans l'exemple 8 précédent, la moyenne de la catégorie (ou de l'échantillon) « C » diffère significativement des catégories (ou des échantillons) « A » et « B » (voir dernière colonne, *p adj*). La sortie présente les différences entre chacune des catégories (*diff*), ainsi que les bornes inférieures (*lwr*) et supérieures (*upr*) des intervalles de confiance à 95% de ces différences.

5. Exercices

Répondez aux exercices dans un fichier texte « Nom_Prenom_TP6.txt ». Au début du fichier, identifiez-vous par votre nom et prénom, ajouter la date et le numéro du TP, le tout en utilisant des commentaires. Utilisez des commentaires pour structurer vos réponses.

Exercice 1. Considérer les deux variables suivantes qui représentent les données de longueur obtenues sur quatre lacs :

```
#Variables Long et Lac
Long <- c(rnorm(27,220,5),rnorm(29,210,5),rnorm(22,250,5),rnorm(19,250,5))
Lac <- as.factor(c(rep("A",27),rep("B",29),rep("C",22),rep("D",19)))
```

Les longueurs moyennes des quatre lacs sont-elles significativement différentes (on suppose l'homogénéité des variances et la normalité des données pour chaque lac) ? **Optionnel :** Retrouver de façon analytique la colonne **Sum Sq** de la sortie du test ainsi que la valeur F du test.

Exercice 2. Vérifier les conditions d'applications du test effectué à l'exercice 1.

Exercice 3. Si le test de l'exercice 1 est significatif, effectuer le test post-hoc H.S.D. de Tukey. Visualiser les données des longueurs pour chacun des lacs sous forme d'un diagramme à moustache (« boxplot ») et interpréter la sortie du test H.S.D. de Tukey.

Bibliographie

- Bertrand, F. et Maumy-Bertrand, M. 2010. Initiation à la statistique avec R. Dunod, Paris, France.
- Crawley, M. J. 2013. The R book (2^{ème} éd.). John Wiley and Sons, Chichester, England.
- Crawley, M. J. 2015. Statistics: an introduction using R (2^{ème} éd.). John Wiley and Sons, Chichester, England.
- Legendre, P. 2009. Introduction au langage statistique R.
http://biol09.biol.umontreal.ca/BIO6077/Introduction_R.pdf.
- Millot, G. 2018. Comprendre et réaliser les tests statistiques à l'aide de R : manuel pour les débutants (4^{ème} éd.). De Boeck, Bruxelles, Belgique.
- Murrell, P. 2011. R graphics (2^{ème} éd.). Chapman & Hall/CRC The R series, Boca Raton, Floride.
- Paradis, E. 2005. R pour les débutants.
http://cran.r-project.org/doc/contrib/Paradis-rdebuts_fr.pdf.
- Short, T. 2004. R Reference Card. <http://cran.r-project.org/doc/contrib/Short-refcard.pdf>.

Liens utiles :

- Site officiel de R.
<http://www.r-project.org>
- Ateliers R du CSBQ.
<https://qcbs.ca/wiki/r>
- Quick R.
<http://www.statmethods.net/>
- Apprentissage en ligne du logiciel R.
<https://www.datacamp.com/>
- Beginners statistics introduction.
http://influentialpoints.com/Training/basic_statistics_introduction.htm
- R Graphics Cookbook.
<http://www.cookbook-r.com/Graphs/>
- Liste des bases de données disponibles sous R.
<https://vincentarelbundock.github.io/Rdatasets/datasets.html>