TD4 : Analyse de la variance à un et deux facteurs controlés

BIO5XX - BIOSTATISTIQUE L3

12 octobre 2016

Objectifs de la séance
Tester l'effet d'une variable discrète sur une variable continue par la méthode d'analyse de la variance à un ou deux facteurs contrôlés

L'analyse de la variance à un facteur peut être utilisée pour tester l'effet d'une variable discrète sur une variable continue ce qui est équivalent à comparer simultanément l'égalité entre elles de plusieurs moyennes.

Si l'on considère l'analyse de la variance à un facteur comme une façons de comparer simultanement la moyenne de plusieurs échantillons, les hypothèses testées sont :

— H_0 Toutes les moyennes sont égales : $\mu_1 = \mu_2 = ... = \mu_n$
— H_1 Au moins une des moyennes testées est différente des autres. Attention cela ne signifie pas que $\mu_1 \neq \mu_2 \neq ... \neq \mu_n$

1 Culture comparée des bactéries du BCG

1.1 Les données expérimentales

L'influence du milieu de culture sur la croissance du bacille de Calmette et Guérin à été testé. Il s'agit de vérifier si l'utilisation d'au moins un des cinq milieux : A, B, C, D, E augmente ou diminue cette croissance. Pour chacun des cinq milieux dix cultures sont réalisées dans des conditions équivalentes. À la fin de celles-ci un aliquote de chacune des cinquante cultures est prélevé et étalé sur une boite de milieu solide afin de dénombrer le nombre de bactéries présentes.

Exemple R - 1 :

Construction de la feuille de données à partir du tableau 1.

```
A <- c(10,12,8,10,6,13,9,10,8,9)
B <- c(11,18,12,15,13,8,15,16,9,13)
C <- c(7,14,10,11,9,10,0,11,7,9)
D <- c(12,9,11,10,7,8,13,14,10,11)
E <- c(7,6,10,7,7,5,6,7,9,6)
BCG <- data.frame(A,B,C,D,E)
BCG
```

A	B	C	D	E
10	11	7	12	7
12	18	14	9	6
8	12	10	11	10
10	15	11	10	7
6	13	9	7	7
13	8	10	8	5
9	15	0	13	6
10	16	11	14	7
8	9	7	10	9
9	13	9	11	6

TABLE 1 – Analyse de la croissance du BCG dans 5 milieux de culture différents. 10 tubes par milieu de culture sont ensemencés à partir d'une même suspension de BCG. Le tableau ci-dessous donne le nombre de colonies obtenues sur boite de Pétri après étalement d'un aliquote de chaque culture.

```
##     A  B  C  D  E
## 1  10 11  7 12  7
## 2  12 18 14  9  6
## 3   8 12 10 11 10
## 4  10 15 11 10  7
## 5   6 13  9  7  7
## 6  13  8 10  8  5
## 7   9 15  0 13  6
## 8  10 16 11 14  7
## 9   8  9  7 10  9
## 10  9 13  9 11  6
```

1.2 Description des échantillons

À la vue de la manipulation 2, il semblerait que toutes les moyennes ne sont pas égales. Mais du fait de la faible taille des échantillons testé (dix cultures par milieu) il est important de déterminer si cette fluctuation est statistiquement significative ou si elle reflète juste le biais d'échantillonnage.

Plutôt que de tester l'égalité entre chaque paire de moyenne $\mu_A = \mu_B$, $\mu_A = \mu_C$, ..., $\mu_D = \mu_E$ qui conduirai à analyser $n(n+1)/2$ tests (avec $n = 5$ le nombre de conditions). Nous allons tester simultanément l'égalité de toutes les moyennes.

Exemple R - 2 :

Analyse descriptive des données BCG. La deuxième commande retourne la matrice de covariance pour chacun des milieux testés. Nous ne sommes dans ce cas intéressés que par la variance observée dans chacun des milieux. Ces variances correspondent à la diagonale de cette matrice qu'il est possible d'extraire par la commande *diag*

```
colMeans(BCG)
```

```
##    A    B    C    D    E
##  9.5 13.0  8.8 10.5  7.0
```

```
var(BCG)
```

```
##             A          B          C           D          E
## A  4.05555556  0.3333333  2.5555556  0.05555556 -1.7777778
## B  0.33333333  9.7777778  2.5555556  2.00000000 -0.8888889
## C  2.55555556  2.5555556 13.7333333 -3.11111111  0.1111111
## D  0.05555556  2.0000000 -3.1111111  4.72222222  0.4444444
## E -1.77777778 -0.8888889  0.1111111  0.44444444  2.2222222
```

```
diag(var(BCG))
```

```
##         A         B         C         D         E
##  4.055556  9.777778 13.733333  4.722222  2.222222
```

```
summary(BCG)
```

```
##        A               B               C               D
##  Min.   : 6.00   Min.   : 8.00   Min.   : 0.00   Min.   : 7.00
##  1st Qu.: 8.25   1st Qu.:11.25   1st Qu.: 7.50   1st Qu.: 9.25
##  Median : 9.50   Median :13.00   Median : 9.50   Median :10.50
##  Mean   : 9.50   Mean   :13.00   Mean   : 8.80   Mean   :10.50
##  3rd Qu.:10.00   3rd Qu.:15.00   3rd Qu.:10.75   3rd Qu.:11.75
##  Max.   :13.00   Max.   :18.00   Max.   :14.00   Max.   :14.00
##        E
##  Min.   : 5
##  1st Qu.: 6
##  Median : 7
##  Mean   : 7
##  3rd Qu.: 7
##  Max.   :10
```

1.3 Test des conditions préalables à la réalisation de l'analyse

Le test simultané de l'égalité des moyennes de plusieurs échantillons différents entre eux selon un critère contrôlé (ici le milieu de culture) peut être réalisé par la méthode d'analyse de la variance à un facteur contrôlé. Pour pouvoir être réalisé ce test impose certaines conditions sur les données.

— Chacun des échantillons testés doit se distribuer selon une loi normale.
— La variance de tous les échantillons doivent être égales.

1.3.1 test de la normalité des échantillons

Nous allons tester la normalité des échantillons de manière graphique et par le test de normalité de *Shapiro-Wilk*.

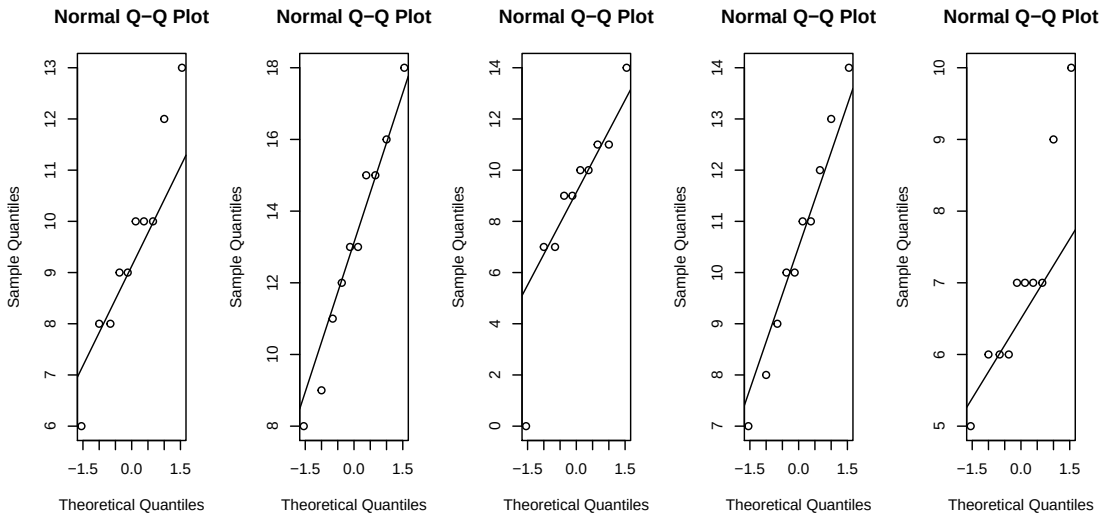
> Réaliser ces deux tests à la main pour le milieu *A*. puis regardez la manipulation 3 pour réaliser ce test sur les 5 milieux.

Exemple R - 3 :

> Test de la normalité des données de chaque échantillon. Le test doit être réalisé sur chacune des colonnes du dataframe *BCG*. C'est pour nous l'occasion d'utiliser la commande *apply* qui permet d'appliquer une fonction soit aux lignes, soit aux colonnes d'un tableau. Le premier argument indique le

tableau à analyser, le deuxième si l'on travaille par ligne (1) ou par colonne (2). Le dernier indique la fonction à appliquer. Successivement nous utilisons *apply* pour tracer les 5 qq-plots et pour calculer les 5 tests de *Shapiro-Wilk*

```
par(mfrow=c(1,5))
apply(BCG,2,function(x) {qqnorm(x);qqline(x)})
```




```
## NULL

apply(BCG,2,function(x) shapiro.test(x)$p.value)

##          A          B          C          D          E
## 0.80056903 0.93740852 0.09924677 0.98289022 0.12676445
```

L'observation des 5 p_{values} du test de *Shapiro-Wilk* montre qu'il n'est possible de rejeter l'hypothèse de normalité pour aucun des échantillons. Nous continuerons donc l'analyse malgré l'aspect des *qq-plot*. Rappelez vous cependant du manque de puissance des tests de normalité pour des échantillons de petite taille.

1.3.2 Test de l'égalité des variances de chacun des échantillons

Pour comparer deux variances, il est habituel d'utiliser le test de *Fisher*. Lorsque l'on veut tester l'égalité simultanée de plus de 2 variances, il existe comme dans le cas de la comparaison de moyenne un test plus approprié ne nécessitant pas la réalisation de toutes les comparaisons deux à deux. Il s'agit du test de *Bartlett* d'homogénéité des variances.

Ces hypothèses de travail sont :

— H_0 Toutes les variances sont égales : $V_1 = V_2 = ... = V_n$

— H_1 Au moins une des variances est différentes des autres.

Exemple R - 4 :

Test de l'homogénéité des variances des échantillons

```
bartlett.test(BCG)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  BCG
## Bartlett's K-squared = 8.6894, df = 4, p-value = 0.06935
```

1.4 Analyse de la variance

1.4.1 La variance totale de l'expérience

Dans ce cas nous considérons l'ensemble des 50 cultures de manière homogène sans tenir compte de la variation du milieu. Nous allons calculer la somme des carrés des écarts à la moyenne globale SCT.

$$SCT = \sum (x - \mu_x)^2 \tag{1}$$

cette variance total possède $n-1$ degré de liberté avec n le nombre total d'expériences, 50 dans notre cas. V_T la variance total est donc égale à

$$V_T = \frac{SCT}{n-1} \tag{2}$$

Avec μ_x la moyenne globale de toutes les mesures

Exemple R - 5 :

Calcul de SCT selon l'équation 2. La première commande a pour but de regrouper toutes les colonnes du tableau *BCG* en un seul vecteur

```
x <- rapply(BCG,c)
x
```

```
##  A1  A2  A3  A4  A5  A6  A7  A8  A9 A10  B1  B2  B3  B4  B5  B6  B7  B8
##  10  12   8  10   6  13   9  10   8   9  11  18  12  15  13   8  15  16
##  B9 B10  C1  C2  C3  C4  C5  C6  C7  C8  C9 C10  D1  D2  D3  D4  D5  D6
##   9  13   7  14  10  11   9  10   0  11   7   9  12   9  11  10   7   8
##  D7  D8  D9 D10  E1  E2  E3  E4  E5  E6  E7  E8  E9 E10
##  13  14  10  11   7   6  10   7   7   5   6   7   9   6
```

```
SCT = sum((x - mean(x))**2)
SCT
```

```
## [1] 507.12
```

```
VT = SCT/(length(x)-1)
VT
```

```
## [1] 10.34939
```

1.4.2 La variance intra-groupe de l'expérience

Il est aussi possible de définir la variance intra-groupe qui ce définit à partir de la somme des écarts de chaque expérience à la moyenne de son groupe au carré (SCE)

$$SCE = \sum_{j=1}^{k} \sum_{i=1}^{e} (x_{ij} - \mu_j)^2 \tag{3}$$

avec k le nombre de conditions et e le nombre d'expérience pour chaque condition. Dans notre car $k = 5$ et $e = 10$.

Le nombre de degrés de liberté associé à cette variance est $n - k$ avec n le nombre total d'expériences. Chaque classe retirant un degré de liberté du fait de l'utilisation de sa moyenne dans la formule

Exemple R - 6 :

Calcul de SCE selon l'équation 3. La fonction *sweep* permet de retrancher le résultat de $mean(BCG)$ à chacune des colonnes du tableau.

```
colMeans(BCG)
```

```
##    A    B    C    D    E
##  9.5 13.0  8.8 10.5  7.0
```

```
sweep(BCG,2,colMeans(BCG),'-')**2
```

```
##            A  B     C     D E
##  [1,]  0.25  4  3.24  2.25 0
##  [2,]  6.25 25 27.04  2.25 1
##  [3,]  2.25  1  1.44  0.25 9
##  [4,]  0.25  4  4.84  0.25 0
##  [5,] 12.25  0  0.04 12.25 0
##  [6,] 12.25 25  1.44  6.25 4
##  [7,]  0.25  4 77.44  6.25 1
##  [8,]  0.25  9  4.84 12.25 0
##  [9,]  2.25 16  3.24  0.25 4
## [10,]  0.25  0  0.04  0.25 1
```

```
SCE <- sum(sweep(BCG,2,colMeans(BCG),'-')**2)
SCE
```

```
## [1] 310.6
```

```
k=length(BCG)
VE=SCE/(length(x)-k)
VE
```

```
## [1] 6.902222
```

1.4.3 La variance inter-groupe de l'expérience

De manière similaire il est possible de calculer la variance inter-groupe. Elle se calcule à partir de la somme des carrés des écarts entre la moyenne de chaque groupe à la moyenne totale, pondéré par la taille

des groupes (SCI).

$$SCI = \sum_{j=1}^{k} n_j(\mu_j - \mu_x)^2 \tag{4}$$

avec k le nombre de groupe, ici $k = 5$, μ_j la moyenne des expériences pour le groupe j, μ_x la moyenne globale des expériences et n_j le nombre d'expérience dans le groupe j. Dans notre cas tous les n_j sont égaux à 10.

Le nombre de degrés de liberté de cette variance est $k - 1$ du fait des k variables utilisés pour la calculer, moins un pour l'utilisation de la moyenne globale qui est une relation les unissant.

Exemple R - 7 :

Calcule de SCI selon l'équation 4.

```
apply(BCG,2,length)
```

```
##  A  B  C  D  E
## 10 10 10 10 10
```

```
mean(x)
```

```
## [1] 9.76
```

```
colMeans(BCG)
```

```
##    A    B    C    D    E
##  9.5 13.0  8.8 10.5  7.0
```

```
(colMeans(BCG) - mean(x)) ** 2 * apply(BCG,2,length)
```

```
##       A       B       C       D       E
##   0.676 104.976   9.216   5.476  76.176
```

```
SCI = sum((colMeans(BCG) - mean(x)) ** 2 * apply(BCG,2,length))
SCI
```

```
## [1] 196.52
```

```
VI <- SCI / (length(BCG) -1)
VI
```

```
## [1] 49.13
```

Il existe une relation entre SCT, SCI et SCI telle que :

$$SCT = SCI + SCE \tag{5}$$

1.5 Test de l'égalité des moyennes

Si les k groupes testés se distribuent normalement, si leur variance sont égales. En posant l'hypothèse H_0 ou toutes les moyennes sont égales on peut montrer que la variance intra-groupe et la variance intergroupe sont deux estimations de la variance totale.

Donc si l'on démontre l'égalité des deux variances V_{intra} et V_{inter} alors on démontre l'égalité des moyennes des différents échantillons.

Dans le cas de l'hypothèse alternative H_1 où au moins une des moyennes est différente, la variance intergroupe devient plus grande que la variance intragroupe. Nous avons donc

$$F_c = \frac{V_{inter}}{V_{intra}} \tag{6}$$

avec $F_c = 1$ sous H_0 et $F_c > 1$ sous H_1

Ici la variable F_c suit une loi de *Fischer* à $k-1$ et $n-k$ degrés de liberté. Comme il s'agit d'un test unilatéral car $V_{inter} \geq V_{intra}$ la p_{value} du test égale

$$p_{value} = 1 - F[> F_c, (k-1, n-k)] \tag{7}$$

avec $F[> F_c, (k-1, n-k)]$ la probabilité d'observé une valeur de $F > F_c$ pour $k-1$ et $n-k$ degrés de liberté.

Exemple R - 8 :

Calcul de la p_{value} associé au test d'égalité des variances intra-groupe et intergroupe.

```
Fc = VI/VE
Fc
```

```
## [1] 7.117997
```

```
1-pf(Fc,k-1,length(x)-k)
```

```
## [1] 0.0001574611
```

1.6 Utilisation de la fonction d'ANOVA intégré à R

1.6.1 Reformatage des données

L'ANOVA intégrée dans *R* impose que toutes les données soient dans un même vecteur et qu'un second vecteur contienne pour chaque valeur sa catégorie d'appartenance.

Exemple R - 9 :

Formatage des données pour l'ANOVA de R.

```
BCG_Group <- data.frame(UFC    = rapply(BCG,c),
                        milieu= rep(names(BCG),
                                    apply(BCG,2,length)))
BCG_Group
```

```
##     UFC milieu
## A1   10      A
## A2   12      A
## A3    8      A
## A4   10      A
## A5    6      A
```

```
## A6   13      A
## A7    9      A
## A8   10      A
## A9    8      A
## A10   9      A
## B1   11      B
## B2   18      B
## B3   12      B
## B4   15      B
## B5   13      B
## B6    8      B
## B7   15      B
## B8   16      B
## B9    9      B
## B10  13      B
## C1    7      C
## C2   14      C
## C3   10      C
## C4   11      C
## C5    9      C
## C6   10      C
## C7    0      C
## C8   11      C
## C9    7      C
## C10   9      C
## D1   12      D
## D2    9      D
## D3   11      D
## D4   10      D
## D5    7      D
## D6    8      D
## D7   13      D
## D8   14      D
## D9   10      D
## D10  11      D
## E1    7      E
## E2    6      E
## E3   10      E
## E4    7      E
## E5    7      E
## E6    5      E
## E7    6      E
## E8    7      E
## E9    9      E
## E10   6      E
```

1.6.2 Réalisation de l'ANOVA

L'ANOVA se réalise ensuite facilement par l'utilisation de la fonction *R aov*

Exemple R - 10 :

Réalisation du test.

```
a <- aov(BCG_Group$UFC ~ BCG_Group$milieu)
summary(a)

##                  Df Sum Sq Mean Sq F value   Pr(>F)
## BCG_Group$milieu  4  196.5   49.13   7.118 0.000157 ***
## Residuals        45  310.6    6.90
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La p_{value} obtenue permet de rejeter l'hypothèse d'égalité des moyennes pour toutes les catégories. La fonction *pairwise.t.test* permet de réaliser tous les test t deux à deux afin d'identifier les moyennes différentes des deux autres.

Exemple R - 11 :

Réalisation de la série de test t grâce à la fonction *pairwise.t.test*.

```
pairwise.t.test(BCG_Group$UFC,
                BCG_Group$milieu,
                p.adjust.method="bonferroni",
                var.equal=T)

##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  BCG_Group$UFC and BCG_Group$milieu
##
##   A      B       C      D
## B 0.0465 -       -      -
## C 1.0000 0.0085  -      -
## D 1.0000 0.3887  1.0000 -
## E 0.3887 6.4e-05 1.0000 0.0465
##
## P value adjustment method: bonferroni
```

L'analyse de variance à deux facteurs a pour objectif de tester l'effet de 2 facteurs sur une variable réponse. Dans le cas d'un plan équilibré complet, le fait d'ajouter 2 facteurs implique également la présence d'une interaction.

2 Analyse de variance à deux facteurs

Rendement laitier de 40 vaches selon leur alimentation

— Facteur A : nature de l'aliment : paille/foin/herbe/autre p=4
— Facteur B : dose
— Plan complet équilibré.

Dose=faible				Dose=Forte			
paille	foin	herbe	autre	paille	foin	herbe	autre
8	12	10	17	8	10	11	17
11	13	12	13	9	7	9	19
11	14	12	17	8	10	11	17
10	11	13	15	10	12	11	16
7	10	14	13	9	11	12	21

TABLE 2 – Rendement laitier de 40 vaches selon leur alimentation

Exemple R - 12 :

Préparation des données.

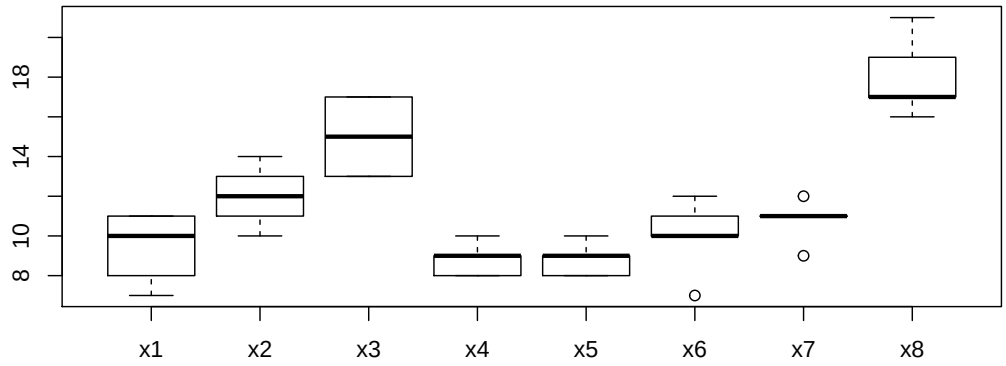
```
x1<-c(8,11,11,10,7)
x2<-c(12,13,14,11,10)
x3<-c(17,13,17,15,13)
x4<-c(8,9,8,10,9)
x5<-c(8,9,8,10,9)
x6<-c(10,7,10,12,11)
x7<-c(11,9,11,11,12)
x8<-c(17,19,17,16,21)
rendement<-data.frame(x1,x2,x3,x4,x5,x6,x7,x8)
colMeans(rendement)
```

```
##   x1   x2   x3   x4   x5   x6   x7   x8
##  9.4 12.0 15.0  8.8  8.8 10.0 10.8 18.0
```

```
diag(var(rendement))
```

```
##  x1  x2  x3  x4  x5  x6  x7  x8
## 3.3 2.5 4.0 0.7 0.7 3.5 1.2 4.0
```

```
boxplot(rendement)
```



Les moyennes semblent différentes. Mais du fait de la petite taille des échantillons, il est nécessaire de regarder si cette fluctuation est statistiquement significative, ou si elle reflète juste le biais d'échantillonnage. Comme dans le TD précédent, nous allons tester simultanément l'égalité de toutes les moyennes. Du fait de la présence de 2 facteurs, on énonce les hypothèse de mani-re un peu différente.

On va tester les hypothèses suivantes

H0 il n'y a pas d'effet du type d'alimentation et de la dose

H1 il y a un effet de l'alimentation

H1' il y a un effet de la dose

H1" il y a un effet de l'interaction

Avant d'appliquer le test d'analyse de variance, on doit vérifier

1. La normalité de chaque échantillon.
2. l'égalité des variances

Exemple R - 13 :

Test de Shapiro sur chaque échantillon.

```
apply(rendement,2,function(x) shapiro.test(x)$p.value)
```

```
##        x1        x2        x3        x4        x5        x6        x7
## 0.2538465 0.9671739 0.1185099 0.3140396 0.3140396 0.4531606 0.1350226
##        x8
## 0.4399399
```

Les valeurs de p_{value} données par le test sont toutes plus élevées que le risque α qu'on pourrait se donner et en particulier que 5%. On accepte donc l'hypothèse de normalité.

Vérifier également la normalité de manière graphique (cf TD précédent).

Exemple R - 14 :

Test de Bartlett pour vérifier l'égalité des variances.

```
bartlett.test(rendement)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  rendement
## Bartlett's K-squared = 6.1648, df = 7, p-value = 0.5206
```

Donc la p_{value} de 0.5 est supérieure aux risques habituels. On accepte l'hypothèse d'égalité des variances des différents échantillons. On peut donc appliquer le test d'analyse de variance pour tester l'effet des différents facteurs.

Exemple R - 15 :

Mise en place de l'analyse de variance.

```
rdt<-c(x1,x2,x3,x4,x5,x6,x7,x8)
z<-rep(c(1,2,3,4),c(5,5,5,5))
alim<-rep(z,2)
dose<-rep(c(0,1),c(20,20))
rendement_aov<-data.frame(rdt,alim,dose)
aov1<-aov(rdt~alim+dose+alim*dose,data=rendement_aov)
summary(aov1)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## alim         1 109.52  109.52  15.401 0.000376 ***
## dose         1   3.60    3.60   0.506 0.481352
## alim:dose    1  92.48   92.48  13.005 0.000934 ***
## Residuals   36 256.00    7.11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interprétation ensemble au tableau de chaque F.

3 Faire l'exercice seul sous R

	Fertilisant 1		Fertilisant 2		Fertilisant 3	
Blé 1	14.3	14.5	18.1	17.6	17.6	18.2
	11.5	13.6	17.1	17.6	18.9	18.2
Blé 2	12.6	11.2	10.5	12.8	15.7	17.5
	11.0	12.1	8.3	9.1	16.7	16.6

TABLE 3 – Rendement par hectare

Description des données (voir table 3) : Deux types de blé différents sont récoltés après avoir été semés sur trois parcelles traitées par trois fertilisants différents. On a répliqué chaque expérimentation quatre fois .

1. Le type de blé est-il influent sur le rendement ?
2. Le fertilisant est-il influent sur le rendement ?
3. L'interaction entre le type de blé et le fertilisant est-il influent ?