Exercices sur le cours d'Analyse de la Variance

Chapitre 3 - Analyse de la variance à un facteur

Exercice 1

1. On note SC_F la somme des carrés des écarts, et S_F^2 les carrés moyens. On sait que la somme des carrés des écats totale, notée SC_T, est égale à la somme des carrés expliquée par le modèle (SC_F) plus la somme des carrés des écarts résiduels, que l'on note SC_R. On a donc :

$$SC_T = SC_F + SC_R\,.$$

Par conséquent, $SC_R = 125,35 - 72,25 = 53,10$.
Le nombre total d'observations n est égal au nombre de traitements ($I = 6$) multiplié par le nombre de répétitions, c'est-à-dire : $n = 6 \times 4 = 24$.
Le reste du tableau de l'analyse de la variance se complète ensuite aisément :

Source de variabilité	**Somme des carrés des écarts**	**Degrés de liberté**	**Carrés moyens**	f_{obs}
Facteur	72,25	$I - 1 = 5$	14,45	4,90
Résiduelle	53,10	$n - I = 18$	2.95	
Totale	125,35	$n - 1 = 23$		

2. Le pourcentage de variabilité expliqué par le traitement est :

$$\frac{SC_F}{SC_T} = \frac{72,25}{125,35} = 0,576$$

Donc, $57,6\,\%$ de la variabilité du rendement de blé est expliqué par le facteur traitement.

Exercice 2

1. Commençons par une brève analyse descriptive de ces données, pour tenter de voir si certaines tendances se dégagent :

```
> X<-data.frame(Placebo=c(5,8,7,7,10,8),T2=c(4,6,6,3,5,6),
+ T3=c(6,4,4,5,4,3),T4=c(7,4,6,6,3,5),T5=c(9,3,5,7,7,6))
> delai <- stack(X)$values # stack() permet d'opérer un
                           # empilement.
> traitement <- stack(X)$ind
> tapply(delai,traitement,summary)
> moy<- tapply(delai,traitement,mean)
> moy
> plot(delai~traitement,col="green")
```

Sur la boîte à moustaches des délais de cicatrisation pour chaque traitement, le placebo semble différent des quatre autres traitements.

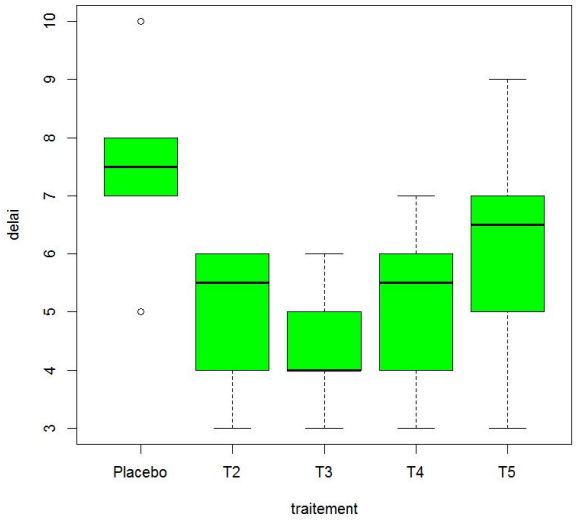

FIGURE 1 – Boîte à moustaches des délais de cicatrisation pour chaque traitement

La table d'analyse de variance est la suivante, après les commandes suivantes :

```
> mon.aov <- aov(delai~traitement)
> summary(mon.aov)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
traitement    4  36.47   9.117   3.896 0.01359 *
Residuals    25  58.50   2.340
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comme l'ANOVA est un modèle linéaire, il est possible d'effectuer une analyse de variance du modèle linéaire sous-jacent :

```
> modele <- lm(delai~traitement)
> anova(modele)
Analysis of Variance Table
Response: delai
Response: delai
          Df Sum Sq Mean Sq F value Pr(>F)
traitement 4 36.467  9.1167  3.896  0.01359 *
Residuals 25 58.500  2.3400
Signif . codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1
```

La valeur de la probabilité critique vaut 0,01359, et permet donc de conclure que les effets d'au moins deux traitements diffèrent.

2. Les estimations sont fournies grâce à la fonction summary pour le modèle

```
> modele <- lm(delai~traitement)
> summary(modele)
```

. Rappelons ici encore que la contrainte imposée par $\boldsymbol{R}$ est : $\alpha_1 = 0$.

```
 Call:
lm(formula = delai ~ traitement)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1667 -0.8750 -0.0833  0.8333  2.8333

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    7.5000     0.6245  12.010 7.06e-12 ***
traitementT2  -2.5000     0.8832  -2.831  0.00903 **
traitementT3  -3.1667     0.8832  -3.586  0.00142 **
traitementT4  -2.3333     0.8832  -2.642  0.01401 *
traitementT5  -1.3333     0.8832  -1.510  0.14366
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.53 on 25 degrees of freedom
Multiple R-squared:  0.384,     Adjusted R-squared:  0.2854
F-statistic: 3.896 on 4 and 25 DF,  p-value: 0.01359
```

L'*intercept* correspond ici à l'estimation du délai moyen du placebo (le traitement 1 est pris comme référence). L'estimation associée à la variable $\boldsymbol{T_2}$ correspond à l'effet différentiel entre le placebo de le traitement $\boldsymbol{T_2}$. Les tests bilatéraux effectués dans ce modèle sont résumés ci-dessous :

$\boldsymbol{H_1}$	
Intercept	$\mu_1 \neq 0$
Traitement T_2	$\alpha_2 \neq 0 \Leftrightarrow \mu_1 \neq \mu_2$
Traitement T_3	$\alpha_3 \neq 0 \Leftrightarrow \mu_1 \neq \mu_3$
Traitement T_4	$\alpha_4 \neq 0 \Leftrightarrow \mu_1 \neq \mu_4$
Traitement T_5	$\alpha_5 \neq 0 \Leftrightarrow \mu_1 \neq \mu_5$

Les résultats fournis par $\boldsymbol{R}$ nous indiquent qu'il existe une différence significative entre le placebo et les traitements 2,3 et 4. Dans ce cas de comparaison vis-à-vis du placebo, il était logique de prendre le placebo comme référence.

Il est possible de choisir une autre référence ou une autre contrainte linéaire au moyen de l'instruction C() comme le montre l'exemple ci-dessous :

```
> summary(lm(delai~C(traitement,base=2)))

Call:
lm(formula = delai ~ C(traitement, base = 2))
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.1667 -0.8750 -0.0833  0.8333  2.8333

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                 5.0000     0.6245   8.006 2.32e-08 ***
C(traitement, base = 2)1    2.5000     0.8832   2.831  0.00903 **
C(traitement, base = 2)3   -0.6667     0.8832  -0.755  0.45739
C(traitement, base = 2)4    0.1667     0.8832   0.189  0.85184
C(traitement, base = 2)5    1.1667     0.8832   1.321  0.19847
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.53 on 25 degrees of freedom
Multiple R-squared:  0.384,     Adjusted R-squared:  0.2854
F-statistic: 3.896 on 4 and 25 DF,  p-value: 0.01359
```

Les estimations et les tests de Student diffèrent. Les résultats montrent que le traitement 2 ne diffère pas des traitements 3,4 et 5, mais on retrouve que le test de Student est significatif pour la comparaison de traiement 2 vis-à-vis du placebo.

À noter que pour obtenir la contrainte : $\sum_{i=1}^{I} \alpha_i = 0$, il faut utiliser la commande : C(traitement,sum).

3. Ce modèle d'ANOVA correspond à un modèle avec une seule variable explicative. Les hypothèses du modèle peuvent être validées grâce à l'analyse des résidus sortie du modèle de régression. Le graphique des résidus est obtenu grâce aux commandes $\boldsymbol{R}$:

```
> par(mfrow=c(2,2))
> plot(modele,col.smooth="red")
```

On peut aussi tester l'égalité des variances pour savoir si oui ou non l'hypothèse d'homoscédasticité est admissible. Le test de Bartlett (sous condition de normalité dans les sous-populations) est obtenu via la commande :

```
> bartlett.test(delai~traitement)

        Bartlett test of homogeneity of variances

data:  delai by traitement
Bartlett's K-squared = 2.4197, df = 4, p-value = 0.6591
```

On ne rejette pas l'hypothèse d'homogénéité des variances.

Comme ce test n'est pas robuste face à la non-normalité, on peut aussi utiliser le test de Levene :

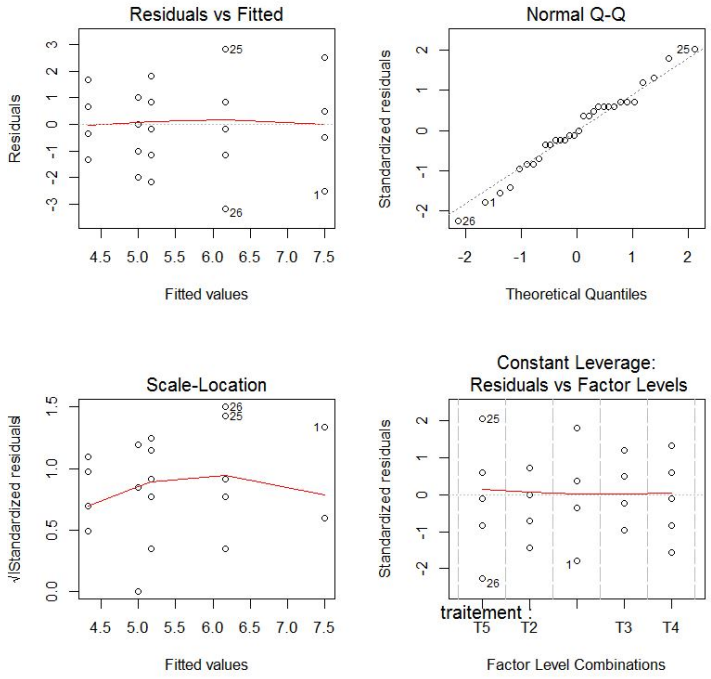

FIGURE 2 – Analyse des résidus

```
> levene.test(delai~traitement)
Levene's Test for Homog. of Var.(center = median)  Df    F      value Pr(>F)
group                                              4  0.5851    0.6763
```

On en arrive à la même conclusion que ci-dessus.

Exercice 3

Les commandes ***R*** sont les suivantes :

```
alim<-rep(1 :3,c(4,5,5))
poids<-c(42.1,37.7,45.1,43.2,45.2,54.2,38.1,48.3,55.1,48.3,44.1,56.9,42.2,54.0)
alim<-factor(alim)
don<-data.frame(alim,poids)
moy<-tapply(don$poids,don$alim,mean)
moy.g<-mean(don$poids)
plot(don$alim,don$poids,col="green")
points(1 :3,moy,pch="@")
abline(h=moy.g)
modele<-aov(poids alim,data=don)
summary(modele)
```

La boîte à moustaches donne une idée des moyennes de gains de poids des veaux suivant l'alimentation. La table d'analyse de variance (ANOVA à un facteur fixe) est la suivante :

```
            Df Sum Sq Mean Sq F value Pr(>F)
alim         2  127.1   63.57   1.829  0.206
Residuals   11  382.3   34.75
```

Les différences entre alimentations ne sont pas significatives car la probabilité critique vaut 0,206. Bien entendu, les 15 veaux ont été considérés comme constituant un échantillon aléatoire simple et leur répartition en trois groupes ayant dû être réalisée de manière complètement aléatoire.

Exercice 4

1. Les commandes ***R*** sont les suivantes :

```
> sal<-rep(1:3,c(6,5,4))
> ventes<-c(425,507,450,483,466,492,420,448,437,432,444,430,492,470,501)
> sal<-factor(sal)
> don<-data.frame(sal,ventes)
> moy.g<-mean(don$ventes)
> moy<-tapply(don$ventes,don$sal,mean)
> moy
```

Les trois moyennes sont données en sortie :

```
      1      2      3
470.50 436.20 473.25
```

Les trois écarts-types sont donnés par les commandes suivantes :

```
> ecart<-tapply(don$ventes,don$sal,sd)
> ecart
```

Cela donne :

```
        1        2        3
29.87139 10.96358 31.63727
```

Pour tracer les boîtes à moustaches, il faut les commandes suivantes :

```
> plot(don$sal,don$ventes,col="green")
> points(1:3,moy,pch="@")
> abline(h=moy.g)
```

2. La commande pour l'analyse de variance est :

```
> modele<-aov(ventes~sal,data=don)
> summary(modele)
```

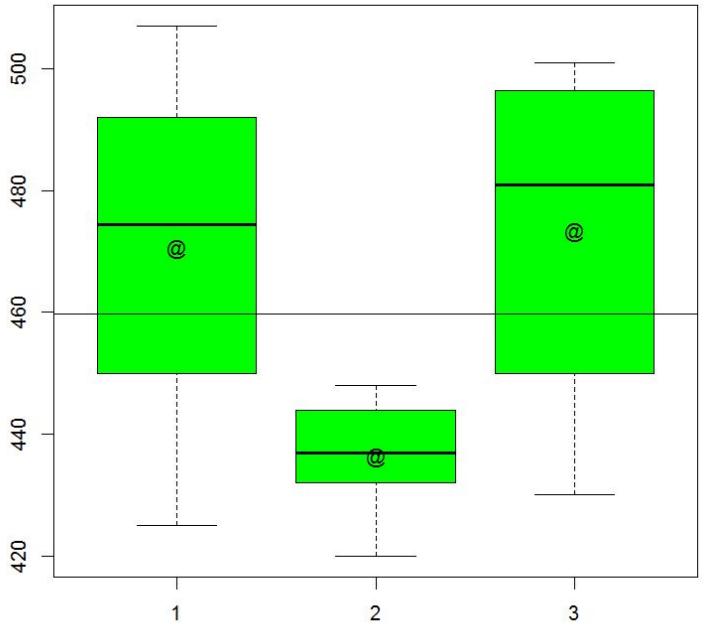

FIGURE 3 – Boîtes à moustaches pour les trois catégories

La table d'analyse de la variance est la suivante :

```
            Df   Sum  Sq Mean  Sq F      value Pr(>F)
sal          2   4195  2097.7  3.168     0.0786
Residuals   12   7945   662.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

On ne rejette donc pas H_0. On peut donc considérer qu'il n'y a significativement pas de différences de moyennes de ventes suivant les trois catégories.

3. L'intervalle de confiance à 90 % pour m_3 est :

$$\bar{x}_3 \pm t_{1-\alpha/2}\sqrt{\frac{CM_R}{n-p}} = 473.25 \pm 1,782 \times \sqrt{\frac{2097.7}{12}} = [449,69\,;\,496,81]\,.$$

Exercice 5

1. Le code ***R*** complet est :

```
> com<-rep(1:3,c(8,8,8))
> sal<-c(43.5,49.5,38.0,66.5,57.5,32.0,67.5,71.5,
+ 73.5,62.0,47.5,36.5,44.5,56.0,68.0,63.5,
+ 45.5,65.4,49.4,58.7,67.4,64.8,69.4,70.5)
```

```
> com<-factor(com)
> don<-data.frame(com,sal)
> moy<-tapply(don$sal,don$com,mean)
> moy
> moy.g<-mean(don$sal)
> moy.g
> ecart<-tapply(don$sal,don$com,sd)
> ecart
> ecart.g<-sd(don$sal)
> ecart.g
> plot(don$com,don$sal,col="green")
> points(1:3,moy,pch="@")
> abline(h=moy.g)
> modele<-aov(sal~com,data=don)
> summary(modele)
```

Nous reportons ici uniquement le tableau de l'analyse de la variance :

```
             Df SuSq   Mean Sq  F value Pr(>F)
com           2    269   134.5    0.866   0.435
Residuals    21   3263   155.4
```

Au vu de la valeur de la probabilité critique (0,435), il n'y a aucune différence significative entre les moyennes des salaires dans les trois communautés.

2. L'intervalle de confiance considéré est de la forme : $\bar{x}_1 - \bar{x}_2 \pm t_{1-\alpha/2}\sqrt{CM_R(1/n_1 + 1/n_2)}$. Ici, dans une table de Student, on peut lire $t_{21;0,975} = 2.08$, le nombre de degrés de liberté étant $n - p = 24 - 3 = 21$. On obtient alors :

$$53,25 - 56,4375 \pm 2,08 \times \sqrt{155,4\left(\frac{1}{8} + \frac{1}{8}\right)} = [-16,1521\,;\,9,7771]\,,$$

intervalle qui contient l'origine, et qui implique donc que les deux moyennes des salaires des deux premières communautés ne sont pas siginificativement différentes, ce que l'on savait déjà.

Exercice 6

1. Il s'agit d'une expérience aléatoire complètement randomisée car les pièces de boeufs ont été choisies au hasard.
2. Le code ***R*** complet est :

```
> spm<-rep(1:4,c(4,4,4,4))
> gras<-c(22,20,23,25,25,27,24,24,30,20,23,27,18,20,17,17)
> spm<-factor(spm)
```

```
> don<-data.frame(spm,gras)
> moy<-tapply(don$gras,don$spm,mean)
> moy
> moy.g<-mean(don$gras)
> moy.g
> ecart<-tapply(don$gras,don$spm,sd)
> ecart
> ecart.g<-sd(don$gras)
> ecart.g
> plot(don$spm,don$gras,col="green")
> points(1:4,moy,pch="@")
> abline(h=moy.g)
> modele<-aov(gras~spm,data=don)
> summary(modele)
```

Le tableau d'analyse de la variance obtenu est :

```
            Df Sum Sq Mean Sq F value  Pr(>F)
spm          3  130.8   43.58   6.301 0.00821 **
Residuals   12   83.0    6.92
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La probabilité critique vaut ici 0,00821, et est donc très inférieure au seuil $\alpha = 5\,\%$. On rejette donc l'hypothèse nulle que les moyennes de gras sont significativement les mêmes dans les 4 supermarchés. On en conclut donc qu'il y a au moins deux supermarchés où les pourcentages de gras dans les pièces de boeufs sont différents.

Exercice 7

1. Le code ***R*** pour cette première question est :

```
> Acidité<-read.table("chap3ex7.csv", header = TRUE, sep=";")
> Acidité[,"Bière"] <- as.factor(Acidité[,"Bière"])
> moy<-tapply(Acidité$Note,Acidité$Bière,mean)
> moy
> moy.g<-mean(Acidité$Note)
> moy.g
> plot(Acidité$Bière,Acidité$Note,col="green")
> points(1:4,moy,pch="@")
> abline(h=moy.g)
```

Le lien entre l'acidité et la bière blanche est visualisé ci-dessous dans les boîtes à moustache (voir figure 4).

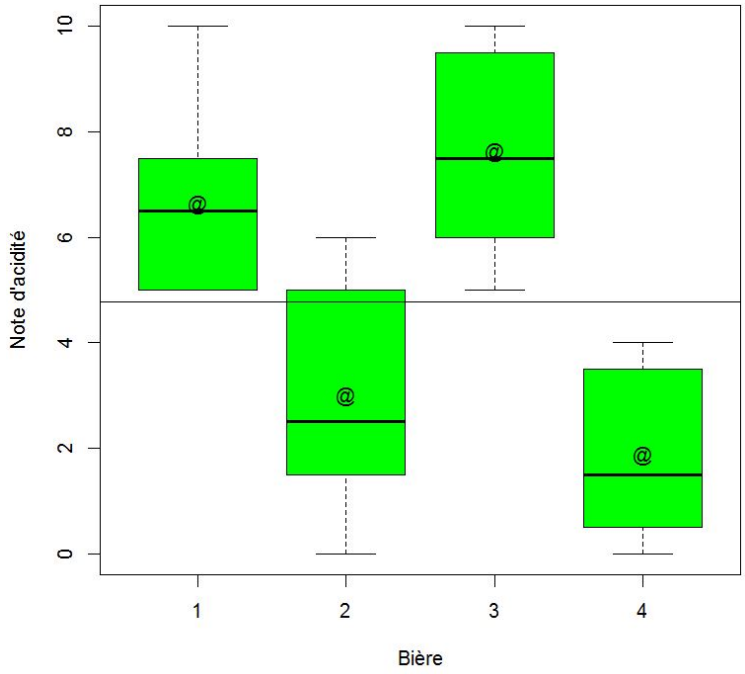

FIGURE 4 – Boîtes à moustaches pour les 4 catégories de bières

2. Le problème revient à comparer les acidités moyennes des quatre bières. Plutôt que de réaliser l'ensemble des tests de comparaison 2 à 2 (procédure critiquée car elle conduit à réaliser beaucoup de tests non indépendants), on choisit de construire un modèle d'analyse de la variance à un facteur (le facteur *bière*).
3. Le modèle s'écrit :

$$Y_{i,j} = \mu + \alpha_i + \varepsilon_{i,j} \qquad i = 1, \cdots, I = 4\,,\; j = 1, \cdots, J = 8$$

avec les hypothèses usuelles sur les résidus :

$$\mathcal{L}(\varepsilon_{i,j}) = \mathcal{N}(0, \sigma^2) \text{ et } cov(\varepsilon_{i,j}, \varepsilon_{k,l}) = 0\,,\; \forall\,(i,j) \neq (k,l)\,.$$

4. Voici le code pour obtenir le tableau d'analyse de la variance :

```
> summary(aov(Note~Bière,data=Acidité))
```

Le tableau obtenu est le suivant :

```
             Df Sum Sq Mean Sq F value   Pr(>F)
Bière         3  184.8   61.61   17.14 1.62e-06 ***
Residuals    28  100.6    3.59
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5. Pour tester globalement l'effet bière, on pose les hypothèses suivantes :

$$H_0 \,:\, \alpha_i = 0 \quad , \quad \forall\, i = 1, \cdots, I = 4$$

contre l'hypothèse alternative :

$$H_1 : \text{ il existe au moins un indice } i_0 \in \{1, \cdots, I\} \text{ tel que } \alpha_{i_0} \neq 0\,.$$

La statistique de test est : $F = \dfrac{S_F^2}{S_R^2}$.

Si l'hypothèse H_0 est vraie, la statistique de test F suit une loi de Fisher-Snedecor à 3 et 28 degrés de liberté.

La décision est prise à l'observation de la valeur de F : $f_{obs} = 17,14$. Et cette valeur est bien supérieure à la valeur trouvée dans la table de Fisher-Snedecor : $f_{3,28}(0,99) = 4,568$. Donc on rejette l'hypothèse H_0 au niveau de confiance 99 %. On considère donc qu'il y a des différences significatives entre ces 4 bières. Autrement dit, au moins une bière est plus (ou moins) acide que les autres.

6. Le pourcentage de variabilité de la note expliquée par le facteur *bière* est : $\dfrac{SC_F}{SC_T} = 0,6475$, soit environ 65 % d'explication.