ENSC 2A - "Modélisation statistique"

TP 3 : Analyse de la variance (ANOVA) avec le logiciel R

Dans ce TP, on va mettre en œuvre différentes analyses de variance (ANOVA). Nous verrons successivement des ANOVA à un facteur, des ANOVA à deux facteurs, et enfin des ANOVA à trois facteurs.

1 ANOVA à un facteur

1.1 Un premier exemple de données réelles

Le tableau ci-dessous donne le rendement en quintaux par hectare d'une variété de blé cultivée sur un sol donnée et traitée avec les engrais A, B et C.

A	48	49	50	49
B	47	49	48	48
C	49	51	50	50

L'agriculteur se demande si le type d'engrais a un effet sur le rendement moyen en quintaux par hectare de la variété de blé qu'il cultive. On est ici typiquement dans le cadre de l'analyse de la variance à un facteur (le type d'engrais).

1.2 Mise en œuvre avec R

1. Saisie des données dans R.

```
donnees <- data.frame(engraisA=c(48,49,50,49),engraisB=c(47,49,48,48),engraisC=c(49,51,50,50))
donnees
stack(donnees)  # cette fonction R permet d'empiler les colonnes d'un tableau de donnees
rendement <- stack(donnees)$values
engrais <- stack(donnees)$ind
is.factor(engrais)  # on peut verifier que la variable "engrais" est bien de type "facteur"
# ou
class(engrais)     # permet de verifier de quelle type est la variable "engrais"
class(rendement)     # permet de verifier de quelle type est la variable "rendement"
```

2. Une représentation graphique des données.

```
plot(rendement~engrais)   # permet d'obtenir les boxplots paralleles du rendement
                          # par type d'engrais
```

3. Quelques statistiques descriptives par groupe (i.e. par type d'engrais).

```
apply(donnees,2,summary)
# ou de maniere equivalente
tapply(rendement,engrais,summary)
```

4. Utilisation de la fonction `aov` pour obtenir le tableau de l'analyse de variance.

```
res.aov <- aov(rendement~engrais)
summary(res.aov)
```

5. Utilisation de la fonction `lm` pour estimer les paramètres du modèle (et obtenir le tableau de l'analyse de variance).

```
res <- lm(rendement~engrais)
anova(res)    # permet de retrouver le tableau de l'analyse de variance precedent
summary(res)  # permet d'obtenir en plus l'estimation des parametres
              # (calage sur le premier groupe PAR DEFAUT)

# pour se caler sur le troisieme groupe
#-------------------------------
res <- lm(rendement~C(engrais,base=3))
summary(res)

# pour se caler sur la moyenne
#-------------------------------
res <- lm(rendement~C(engrais,sum))
summary(res)
```

6. Etude des résidus (normalité, homoscédasticité, graphique).

```
shapiro.test(res$residuals)  # test de normalite des residus
bartlett.test(rendement,engrais)     # test d'homoscedasticite
plot(res$fitted,res$residuals)     # graphique des valeurs predites versus les residus
```

7. Comparaisons multiples des moyennes.

```
pairwise.t.test(rendement,engrais,p.adjust="bonf")
```

1.3 Autres jeux de données

Il vous faut étudier les différents jeux de données ci-dessous et répondre aux problématiques posées.

Jeu de données 1.1. Pour comparer 6 procédés de fabrication de lampes à rayons ultra-violets, on a pris un échantillon d'ampoules de chaque fabrication et on a observé les durées de vie de ces ampoules.

Les données sont disponibles dans le "data frame" appelé `lampes.Rda` contenant deux variables mesurées sur $n = 150$ ampoules :

- `duree` : variable quantitative correspondant à la durées de vie de l'ampoule,
- `procede` : facteur à 5 niveaux `P1`, ..., `P5` correspondant au procédé de fabrication de l'ampoule.

Le procédé de fabrication a-t-il un effet sur la durée de vie moyenne de ces ampoules ?

Jeu de données 1.2. Les données étudiées ici sont des mesures de rugosité de plaques suivant une durée de broyage de la matière première. Elles sont disponibles dans le "data frame" appelé `plaques.Rda` contenant deux variables mesurées sur $n = 60$ plaques :

- `rugosite` : variable quantitative correspondant à la rugosité de la plaque,
- `duree` : facteur à 4 niveaux `D1`, ..., `D4` correspondant à la durée de broyage avant la fabrication de la plaque.

La durée du broyage a-t-elle un effet sur la mesure moyenne de rugosité des plaques ?

2 ANOVA à deux facteurs

2.1 Un premier exemple de données réelles

Le tableau ci-dessous donne le rendement de quatre types différents de plants cultivés sur des terrains traités avec trois types d'engrais.

	Plant I	Plant II	Plant III	Plant IV
Engrais A	15 14 17	18 17 18	12 13 12	14 15 15
Engrais B	16 19 18	23 24 25	15 14 14	12 11 10
Engrais C	18 17 17	20 21 21	17 19 17	12 13 12

On désire étudier les effets du type d'engrais (A, B ou C) et du type de plant (I, II, III ou IV) sur le rendement à l'hectare. On est ici typiquement dans le cadre de l'analyse de la variance à deux facteurs (le type d'engrais et le type de plant). On désire savoir si les facteurs "engrais" et "plant" ont un effet en moyenne sur le rendement à l'hectare.

2.2 Mise en œuvre avec R

1. Saisie des données dans R.

```
rendement<-c(15,14,17,18,17,18,12,13,12,14,15,15,
             16,19,18,23,24,25,15,14,14,12,11,10,
             18,17,17,20,21,21,17,19,17,12,13,12)
plant<-as.factor(rep(rep(c("Plant I","Plant II","Plant III","Plant IV"),each=3),3))
engrais<-as.factor(rep(c("Engrais A","Engrais B","Engrais C"),each=12))
donnees<-data.frame(rendement,plant,engrais)

levels(engrais)   # donne les differents niveaux du facteur "engrais"
levels(plant)   # donne les differents niveaux du facteur "plant"
table(engrais)   # indique les nombres d'observations par niveau du facteur "engrais"
table(plant)   # indique les nombres d'observations par niveau du facteur "plant"
```

2. Quelques représentations graphiques des données.

```
boxplot(rendement~engrais*plant)
boxplot(rendement~plant)
boxplot(rendement~engrais)

interaction.plot(plant,engrais,rendement)
# ou
interaction.plot(engrais,plant,rendement)
```

3. Quelques statistiques descriptives par groupe (i.e. croisement d'un type de plant et d'un engrais).

```
tapply(rendement,list(engrais,plant),mean)
tapply(rendement,list(engrais,plant),min)
tapply(rendement,list(engrais,plant),var)
```

4. Utilisation de la fonction `aov` pour obtenir le tableau de l'analyse de variance.

```
res.aov <- aov(rendement~engrais*plant) # modele avec interaction
summary(res.aov)
# ou
res.aov <- aov(rendement~engrais+plant+engrais:plant) # modele avec interaction
summary(res.aov)

res.aov <- aov(rendement~engrais+plant) # modele sans interaction
summary(res.aov)
```

5. Utilisation de la fonction `lm` pour estimer les paramètres du modèle (et obtenir le tableau de l'analyse de variance).

```
res<-lm(rendement~engrais*plant)  # calage sur la case "Type I" et "Engrais A"
summary(res)

# pour se caler sur la moyenne
res <- lm(rendement~C(engrais,sum)+C(plant,sum)+C(engrais,sum):C(plant,sum))
summary(res)
```

6. Etude des résidus (normalité, homoscédasticité, graphique).

```
shapiro.test(res$residuals)  # test de normalite des residus
bartlett.test(rendement,engrais,plant)     # test d'homoscedasticite
fligner.test(rendement,engrais,plant)    # autre test d'homoscedasticite
plot(res$fitted,res$residuals)     # graphique des valeurs predites versus les residus
```

2.3 Autres jeux de données

Jeu de données 2.1. On s'intéresse à l'influence du type d'alimentation sur le rendement laitier. On a relevé les $n = 40$ observations suivantes :

	Paille	Foin	Herbe	Aliments ensilés
Dose faible	8 11 11 10 7	12 13 14 11 10	10 12 12 13 14	17 13 17 14 13
Dose forte	8 9 8 10 9	10 7 10 12 11	11 9 11 11 12	17 19 17 16 21

Ces données sont disponibles dans le "data frame" appelé `lait.Rda` contenant trois variables suivantes :

- `rdt.laitier` : variable quantitative correspondant au rendement laitier de la vache,
- `aliment` : facteur à 4 niveaux correspondant au type d'alimentation donnée à la vache,
- `aliment` : facteur à 2 niveaux correspondant à la dose donnée à la vache.

Faire une analyse de la variance à deux facteurs (type d'alimentation et dose) pour étudier l'influence sur le rendement laitier moyen du type d'alimentation et de la dose.

Jeu de données 2.2. On s'intéresse à l'influence du pourcentage de dopants ajoutés à la composition de base (noté S_1, ..., S_4) et/ou de la durée de broyage (D_1 et D_2) sur une mesure quantitative de rugosité. Les données sont disponibles dans le "data frame" appelé `plaques2.Rda` contenant trois variables suivantes :

- `rugosite` : variable quantitative correspondant à la mesure de rugosité de la plaque,
- `dopant` : facteur à 4 niveaux correspondant au pourcentage de produits dopants utilisés dans la fabrication de la plaque,
- `duree` : facteur à 2 niveaux correspondant à la durée de broyage lors de la fabrication de la plaque.

Faire une analyse de la variance à deux facteurs (durée de broyage et pourcentage de produits dopants ajoutés) pour étudier l'influence sur la rugosité de la durée de broyage et du pourcentage de produits dopants ajoutés.

Jeu de données 2.3. On se propose d'étudier les effets de deux paramètres, la température et le pourcentage de molybdène, sur la dureté de l'acier. Pour cela, on a fabriqué deux éprouvettes pour chaque température considérée ($400C^o$, $450C^o$, $500C^o$) et chaque pourcentage considéré (0.25%, 0.50%, 0.75%) et on a ensuite mesuré la dureté de chaque éprouvette. Les données sont disponibles dans le "data frame" appelé `acier.Rda` contenant trois variables suivantes :

- `durete` : variable quantitative correspondant à la mesure de la dureté de l'acier,
- `temperature` : facteur à 4 niveaux correspondant à la température considérée dans la fabrication de l'acier,
- `duree` : facteur à 3 niveaux correspondant au pourcentage de molybdène utilisé lors de la fabrication de l'acier.

Faire l'analyse de variance à deux facteurs qui peut permettre de répondre au problème posé.

3 ANOVA à trois facteurs

3.1 Un premier exemple de données réelles

On s'intéresse à des données concernant la rapidité de lecture suivant le caractère (concret ou abstrait) des mots, le sexe et l'âge des enfants (petit ou grand). Voici ci-dessous les mesures (en secondes) obtenues pour les garçons puis pour les filles lors de la lecture de deux textes (l'un avec des mots concrets et l'autre avec des mots abstraits) :

garçons :

	concret	abstrait
petit	1450 1495 1668	1445 1433 1702
grand	966 1168 590	1104 1027 788

filles :

	concret	abstrait
petit	1224 1286 1098	1344 1182 1524
grand	765 840 1183	1092 893 1150

Au moyen d'une analyse de la variance à trois facteurs, étudier les effets de l'âge, du sexe des enfants ainsi que du caractère des mots sur la rapidité de lecture.

3.2 Un autre jeu de données

Etudier les effets de la variété d'orge (3 variétés), de l'année (2017, 2018) et du lieu (4 lieux différents) sur le rendement moyen à l'hectare. On dispose des données contenues dans le "data frame" appelé `rendement3.Rda` contenant quatre variables suivantes :

- `rdt` : variable quantitative correspondant au rendement à l'hectare de la parcelle,
- `annee` : facteur à 2 niveaux correspondant à l'année de la récolte de la parcelle,
- `variete` : facteur à 3 niveaux correspondant à la variété utilisée pour ensemencer la parcelle,
- `lieu` : facteur à 4 niveaux correspondant au lieu où se situe la parcelle.

Faire l'analyse de la variance à 3 facteurs répondant au problème.