# T. P. n$^{o}$ 5
# Analyse de la variance

**Exercice 1. Mangez des pommes !**

Nous souhaitons comparer la teneur en vitamine C de cinq variétés de pommes notées $V_1$, $V_2$, $V_3$, $V_4$ et $V_5$. Pour chaque variété la teneur en vitamine C, exprimée en $mg/(100g)$ a été mesurée dans cinq pommes prises au hasard. Nous obtenons les données suivantes regroupées dans le tableau ci-dessous :

| $V_1$ | $V_2$ | $V_3$ | $V_4$ | $V_5$ |
|---|---|---|---|---|
| $93,6$ | $95,3$ | $94,5$ | $98,8$ | $94,6$ |
| $95,3$ | $96,9$ | $97,0$ | $98,2$ | $97,8$ |
| $96,0$ | $95,8$ | $97,8$ | $97,8$ | $98,0$ |
| $93,7$ | $97,3$ | $97,0$ | $97,2$ | $95,0$ |
| $96,2$ | $97,7$ | $98,3$ | $97,9$ | $98,9$ |

Qu'est-il possible de conclure ?
Pour répondre à cette question, vous allez procéder à une analyse de la variance à un facteur fixe.

1. Pourquoi faut-il faire ici une analyse de la variance à un facteur fixe et non pas une analyse de la régression linéaire ?
2. Écrire le modèle statistique de l'analyse de la variance à un facteur.
3. Quelles sont les hypothèses que vous faites pour pouvoir appliquer la théorie du modèle linéaire ? Sont-elles vérifiées ?
4. Calculer, à l'aide du logiciel `R`, le tableau de l'ANOVA correspondant à cette étude.
5. Réaliser le test de Fisher au seuil de significativité 5% puis de 1%. Qu'en déduisez-vous ?
6. Donner une estimation de la variance $\sigma^2$.
7. Dans le cas de cette étude, pouvez-vous faire des comparaisons multiples ? Si oui, pourquoi ? Et dans le cas de l'affirmatif, réaliser alors les comparaisons. Que pouvez-vous conclure ?

**Exercice 2. D'après Prum. *Modèle linéaire. Comparaison de groupes et régression.* Les éditions INSERM, 1996.**

Nous souhaitons comparer trois traitements, notés $A$, $B$ et $C$ contre l'asthme : le traitement $B$ est un nouveau traitement, que nous souhaitons mettre en compétition avec les traitements classiques $A$ et $C$. Nous répartissons par tirage au sort les patients venant consulter dans un centre de soin, et nous leur affectons l'un des trois

traitements. Nous mesurons sur chaque patient la durée, en jours, séparant de la prochaine crise d'asthme. Les mesures sont reportées dans le tableau ci-dessous :

| Traitement $A$ | Traitement $B$ | Traitement $C$ |
|---|---|---|
| 26; 27; 35; 36 | 29; 42; 44; 44 | 26; 26; 30; 30 |
| 38; 38; 41; 42 | 45; 48; 48; 52 | 33; 36; 38; 38 |
| 45; 50; 65 | 56; 56; 58; 58 | 39; 46; 47; 51 |
| | 60; 61; 63; 63 | 51; 56; 75 |
| | 69 | |

Pouvons-nous conclure que les traitements ont une efficacité différente pour le critère « temps séparant la prochaine crise ? »
Pour répondre à cette question, vous procéderez à nouveau à une analyse de la variance à un facteur fixe. Reprenez les questions de l'exercice 1 et concluez.
Une remarque : quelle est la différence entre l'exercice 1 et l'exercice 2 ?

**Exercice 3. Et pour finir : génotype.**

Scheffé (1959, pp. 140–141) a reproduit les données d'une expérience destinée à étudier les variations dans la masse, en $g$, de rats femelles hybrides en fonction du génotype de la mère nourricière de la portée et du génotype de la portée. Il s'agit des masses moyennes des individus femelles de 61 portées en fonction du génotype de la mère nourricière de la portée et du génotype de la portée. Précisons que les portées proviennent de mères distinctes et n'ont pas eu les mêmes mères nourricières.

1. Télécharger le fichier `foster` du package `HSAUR` à l'aide de la commande suivante : `data(foster,package="HSAUR")`.
2. Afficher-le à l'écran.
3. Combien de lignes ? Combien de colonnes ? Quel est le type de chaque colonne ?
4. Vous remarquez qu'il y a deux facteurs dans ce `data.frame`. Nous souhaitons étudier le gain de masse en fonction de ces deux facteurs à l'aide d'un modèle de l'analyse de la variance. Il se pose alors la question naturelle de savoir si le plan est équilibré ou déséquilibré ? Quel type d'analyse de la variance pouvez-vous faire sous réserve que les conditions soient vérifiées ? À deux facteurs avec interaction ? À deux facteurs sans interaction ?
5. Exécuter la ligne de commande suivante :
   `plot.design(foster)`
   **Remarque :** Faire un help de plot.design.
6. La figure que vous venez d'obtenir indique que les différences dans les poids pour les quatre niveaux du génotype de la mère sont substancielles. Les différences correspondantes pour le génotype de la lignée sont plus petites.
   Nous allons appliquer une analyse de la variance en utilisant la fontion *aov*, mais dans ce cas présent une complication apparait du au fait que le plan n'est pas équilibré. Ici, le nombre d'observations est différent dans chaque groupe. Par conséquent, il n'est plus possible de partitionner la variation que l'on observe dans le jeu de données en des sommes de carrés orthogonales

représentant les effets principaux et les interactions.
Dans un plan déséquilibré d'analyse de la variance à deux facteurs notés $A$ et $B$, il y a une proportion de la variance de la variable réponse qui peut être attribuée soit au facteur $A$, soit au facteur $B$. La conséquence est que les facteurs $A$ et $B$ ensemble expliquent moins de variation de la variable dépendante que la somme de chacun explique à elle seule.
Nous allons maintenant envisager deux types d'analyse de la variance. Exécuter à la suite les lignes de commande suivantes :

```
>summary(aov(weight∼litgen*motgen,data=foster))
```

et

```
>summary(aov(weight∼motgen*litgen,data=foster))
```

Qu'observez-vous ? Il y a une petite différence dans la somme des carrés pour les deux effets principaux, et par conséquent, dans la statistique de Fisher et dans la $p$-valeur. En fait, cette différence qui apparait est due au type de la somme des carrés qui est calculé. Ici la fonction *aov* calcule une somme de carrés de type I.
Vous pouvez lire cet extrait suivant :
« Le tableau Type I SS est construit en ajoutant les variables une à une dans le modèle, et en évaluant l'impact sur la somme des carrés du modèle. De ce fait, l'ordre dans lequel les variables sont entrées dans le modèle influe sur les résultats obtenus. Le tableau Type III SS est calculé en enlevant ponctuellement chacune des variables du modèle, toutes les autres étant présentes, afin d'évaluer l'impact de la variable supprimée sur le modèle. Ainsi, les valeurs obtenues dans le tableau Type III SS sont indépendantes de l'ordre dans lequel sont sélectionnées les variables. Le tableau Type III SS est souvent préféré pour l'analyse des résultats d'un modèle avec interactions. »
Il ne reste plus qu'à savoir comment calculer les sommes de carrés de Type III. Il est nécessaire d'utiliser la fonction *Anova* (attention avec un A majuscule du package « car »).
La ligne de commande à exécuter est la suivante :

```
>Anova(aov(weight∼motgen*litgen,data=foster), type="III")
```

**7.** La représentation graphique classique pour une analyse de la variance se fait à l'aide des lignes de commande suivantes :

```
>layout(matrix((1 :4), nrow=2, ncol=2, byrow=T))
>plot(aov(weight∼motgen*litgen,data=foster))
```

**8.** Il est évident qu'il faut absolument vérifier les hypothèses classiques d'une analyse de la variance paramétrique. Pour cela, exécuter les lignes de commande suivantes :

```
>residus<-residuals((aov(weight∼litgen*motgen,data=foster)))
>qqnorm(residus)
>qqline(residus)
>shapiro.test(residus)
>library("car")
> bartlett.test(residus∼litgen*motgen,data=foster)
```

**Remarque :** Le test de Bartlett est un test d'égalité des variances construit pour comparer au moins 3 échantillons. Indiquons également que le test de

Fisher d'égalité des variances a pour commande « var.test (échantillon 1, échantillon 2) ».

**9.** Enfin, nous pouvons calculer les estimations des coefficients du modèle en exécutant la ligne de commande suivante :
`>coefficients(aov(weight~motgen*litgen, data=foster))`

**10.** Nous pouvons aussi calculer les valeurs ajustées en exécutant la ligne de commande suivante :
`>fitted.values(aov(weight~motgen*litgen, data=foster))`

**11.** Les « anovas » conduisent à dire qu'il y a un effet principal du facteur « motgen », qui signifie que le génotype de la mère ou encore que le facteur « génotype de la mère » est significatif au seuil $\alpha = 5\%$.
Maintenant, nous souhaitons savoir quelles sont les différences entre les effets des différents niveaux du facteur. Pour cela, nous allons avoir recours à des tests de comparaison multiples. Choisissons un des modèles d'analyse de la variance.
`>foster_aov<-aov(weight~litgen*motgen,data=foster)`
Puis exécuter les lignes de commande suivantes :
`>foster_hsd<-TukeyHSD(foster_aov,"motgen")`
`>foster_hsd`
Qu'observez-vous ? Que concluez-vous ?

### Exercice 4. Quatre modèles de machines à écrire

Une entreprise cherche à tester quatre modèles de machines à écrire. Pour faire ce test, elle demande à cinq secrétaires professionnelles de taper un texte pendant 5 minutes. À la fin du test, on compte le nombre moyen de mots tapés en une minute. On répète l'expérience le lendemain.
Les résultats (nombre moyen de mots par minute) sont présentés dans le tableau au verso.

| | Secrétaires | | | | |
|---|---|---|---|---|---|
| Machines à écrire | 1 | 2 | 3 | 4 | 5 |
| 1 | 33 | 31 | 34 | 34 | 31 |
| | 36 | 31 | 36 | 33 | 31 |
| 2 | 32 | 37 | 40 | 33 | 35 |
| | 35 | 35 | 36 | 36 | 36 |
| 3 | 37 | 35 | 34 | 31 | 37 |
| | 39 | 35 | 37 | 35 | 40 |
| 4 | 29 | 31 | 33 | 31 | 33 |
| | 31 | 33 | 34 | 27 | 33 |

Soit le modèle :

$$y_{i,j,k} = \mu + \alpha_i + \beta_j + \gamma_{i,j} + \epsilon_{i,j,k}$$

avec $i = 1, 2, 3, 4, \quad j = 1, 2, 3, 4, 5$ et $k = 1, 2$ et où $\alpha$ correspond à l'influence de la machine à écrire, $\beta$ l'influence de la secrétaire et $\gamma$ l'interaction entre la machine $i$ et la secrétaire $j$.

1. On considère la forme matricielle du modèle :
$$\mathbf{y} = \mathbf{X}\beta + \epsilon.$$
Écrire le vecteur $\mathbf{y}$ et la matrice $\mathbf{X}$ en tenant compte des contraintes habituelles pour que la matrice soit inversible.
2. Tester les trois hypothèses nulles suivantes :
   (i) Que cherche-t-on à tester avec les deux hypothèses suivantes ?
   $$\mathcal{H}_0 \ : \quad \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$$
   contre
   $$\mathcal{H}_1 \ : \quad \text{les valeurs des } \alpha_i \text{ ne sont pas toutes égales, } i = 1, 2, 3, 4.$$
   (ii) Que cherche-t-on à tester avec les deux hypothèses suivantes ?
   $$\mathcal{H}_0 \ : \quad \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5$$
   contre
   $$\mathcal{H}_1 \ : \quad \text{les valeurs des } \beta_j \text{ ne sont pas toutes égales, } j = 1, 2, 3, 4, 5.$$
   (iii) Que cherche-t-on à tester avec les deux hypothèses suivantes ?
   $$\mathcal{H}_0 \ : \quad \gamma_{1,1} = \gamma_{1,2} = \cdots = \gamma_{1,5} = \gamma_{2,1} = \cdots = \gamma_{4,5}$$
   contre
   $$\mathcal{H}_1 \ : \quad \text{les valeurs des } \gamma_{i,j} \text{ ne sont pas toutes égales, } i = 1, 2, 3, 4 \text{ et } j = 1, 2, 3, 4, 5.$$