# STT- 2300

# Cours d'Analyse de la Variance

Professeur Michel Carbon

Département de Mathématiques et Statistique

Université de Laval

Hiver 2015

 

# Table des matières

 

 

# Chapitre 1

# Fondements mathématiques et statistiques

Ce chapitre sert à présenter brièvement quelques rappels sur la loi normale, la loi du khi-deux, la loi $t$ de Student et la loi $F$ de Fisher. La connaissance de quelques propriétés fondamentales de ces lois est indispensable à un traitement statistique rigoureux de l'analyse de la variance, qui est l'objet de ce cours.

## 1.1 La loi normale

### 1.1.1 Loi normale centrée réduite

**Définition 1.1.1**
*La loi normale (ou gaussienne) centrée réduite, notée $\mathcal{N}(0,1)$, est une loi à densité. Cette densité de probabilité a pour expression :*

$$f(x) = \frac{1}{2\pi} e^{-x^2/2}\,.$$

C'est une fonction paire, positive, qui a une forme en cloche, et telle que $\lim_{x\to\pm\infty} f(x) = 0$, comme on peut le voir sur la figure (1.2). C'est bien une densité de probabilité car :

1. $f(x) > 0$ pour tout $x \in \mathbb{R}$,
2. $\displaystyle\int_{-\infty}^{+\infty} f(x)\,dx = 1\,.$

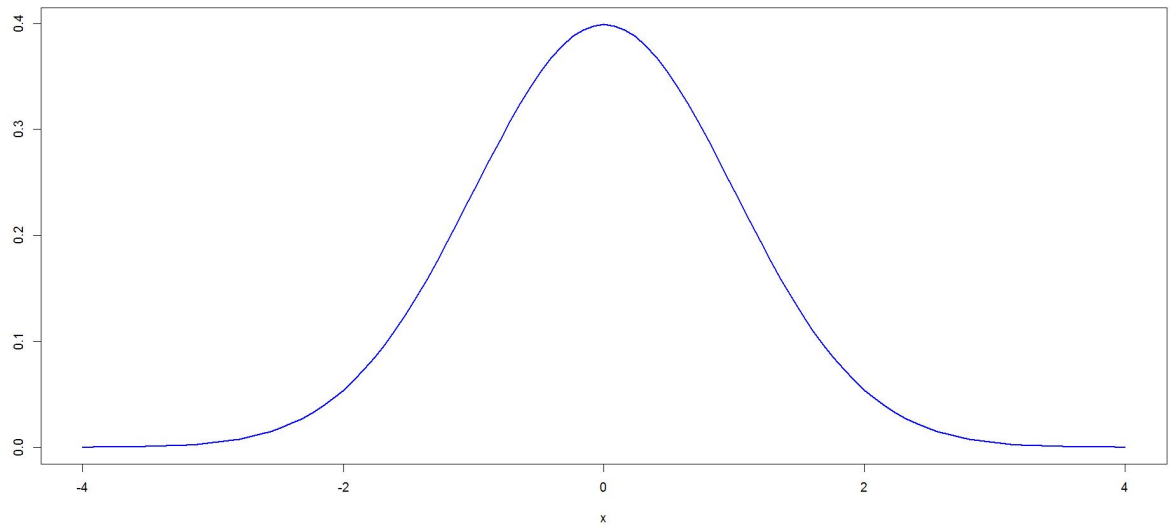


FIGURE 1.1 – Densité de la loi normale centrée réduite

Le premier point se démontre trivialement. Quant au second, on peut le prouver comme suit :

$$\begin{aligned}
\int_{-\infty}^{\infty} f(x)\,dx &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,dx \\
&= \sqrt{(\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,dx)(\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\,dy)} \\
&= \sqrt{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{2\pi} e^{-(x^2+y^2)/2}\,dxdy} \\
&= \sqrt{\int_{0}^{2\pi}\int_{0}^{\infty} \frac{1}{2\pi} e^{-r^2/2} r\,drd\theta} \\
&= \sqrt{\int_{0}^{2\pi} \frac{1}{2\pi}\,d\theta \int_{0}^{\infty} re^{-r^2/2}\,dr} \\
&= \sqrt{1} = 1.
\end{aligned}$$

Le passage de la ligne 3 à la ligne 4 se fait par un changement de variables polaires :

$$\begin{cases} x &= r\cos\theta \\ y &= r\sin\theta \end{cases}$$

en remarquant que, les domaines s'échangent de $\mathbb{R}\times\mathbb{R} \hookrightarrow \mathbb{R}^+ \times [0, 2\pi[$, puis que : $x^2+y^2 = r^2$ et en calculant le jacobien de la transformation $J = r$.

### 1.1.2 Loi normale quelconque

**Définition 1.1.2**
*Soient $\mu \in \mathbb{R}$ et $\sigma > 0$. La loi normale (ou gaussienne) univariée de moyenne $\mu$ et de variance $\sigma^2$, notée $\mathcal{N}(\mu, \sigma^2)$ est la loi de probabilité admettant pour densité :*

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad x \in \mathbb{R}$$

Cette fonction est bel et bien une fonction de densité puisqu'on a :

1. $f(x) > 0$ pour tout $x \in \mathbb{R}$ ;
2. $\displaystyle\int_{-\infty}^{\infty} f(x)dx = 1.$

Le premier point est trivial. On peut vérifier le deuxième grâce au changement de variable $y = (x - m)/\sigma$ dans le calcul aisé ci-dessous :

$$\begin{aligned} \int_{-\infty}^{\infty} f(x)dx &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy \\ &= 1\,, \end{aligned}$$

grâce à la démonstration faite dans le cas de la variable normale centrée réduite.

On rappelle le résultat suivant, très utile en pratique pour le calcul de probabilités concernant les lois gaussiennes.

**Théorème 1.1.1**
*On a :*

$$X \sim \mathcal{N}(\mu, \sigma^2) \quad \Longleftrightarrow \quad \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1)\,.$$

La proposition ci-dessous, laissée en exercice, se démontre aisément.

**Proposition 1.1.1**
*Si $X$ est une variable aléatoire de loi $\mathcal{N}(\mu, \sigma^2)$, alors on :*

$$E(X) = \mu \qquad et \qquad V(X) = \sigma^2\,.$$

Cela justifie a posteriori l'appellation de normale centrée réduite, et notée $\mathcal{N}(0, 1)$.

**Théorème 1.1.2**
*Si $X \sim \mathcal{N}(\mu, \sigma^2)$, alors la fonction génératrice des moments (f.g.m) de la variable aléatoire $X$ est égale à :*

$$M_X(t) = e^{t\mu + t^2\sigma^2/2}\,.$$

 

**Démonstration**

$$\begin{aligned} M_X(t) &= E[e^{tX}] \\ &= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx \\ &= \int_{-\infty}^{\infty} e^{t(\mu+\sigma y)} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy \\ &= e^{t\mu} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(y^2-2t\sigma y+t^2\sigma^2-t^2\sigma^2)/2} dy \\ &= e^{t\mu+t^2\sigma^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(y-t\sigma)^2/2} dy = e^{t\mu+t^2\sigma^2/2} \end{aligned}$$

On passe de la ligne 2 à la ligne 3 en faisant le changement de variable $y = (x - \mu)/\sigma$, et en remarquant que l'intégrale de la ligne 5 vaut 1, car c'est l'intégrale sur $\mathbb{R}$ de la densité de la loi normale de moyenne $t\sigma$ et de variance 1.

On trouvera ci-dessous les graphes de trois densités de lois normales de variance égale à 1, et avec trois moyennes différentes (0, puis 0.5, puis 1) :

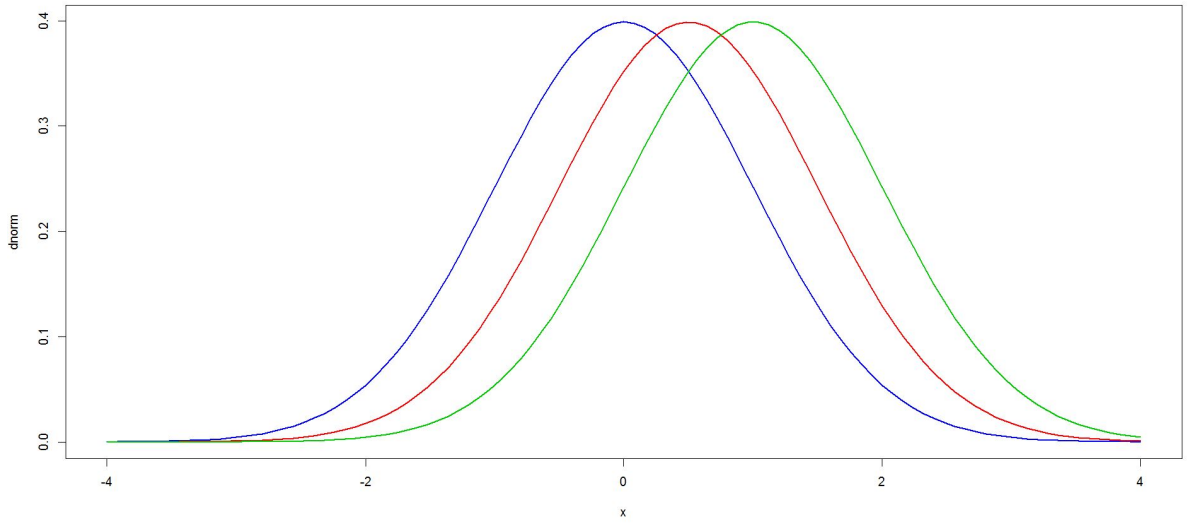


FIGURE 1.2 – Densités de lois normales réduites

On remarquera que les graphes sont juste translatés l'un de l'autre.

On trouvera ci-dessous les graphes de trois densités de lois normales de moyenne nulle et d'écarts-types différents ($\sigma = 1$, puis 1.2, puis 1.5) :

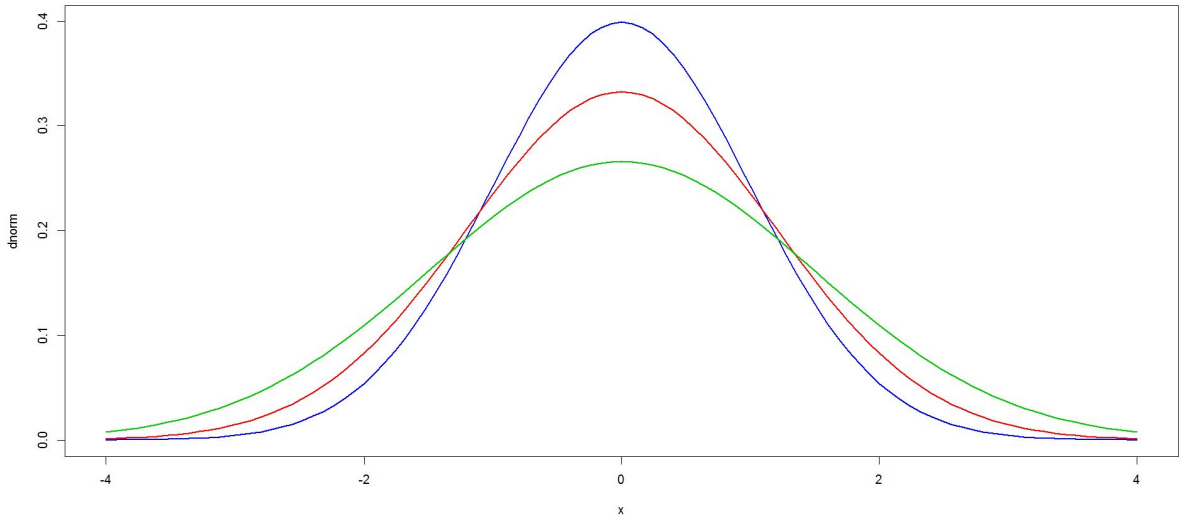


FIGURE 1.3 – Densités de lois normales centrées

**Théorème 1.1.3**
*Soient $X_1, X_2, \cdots, X_n$ $n$ variables indépendantes telles que $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ pour $i = 1, \cdots, n$. Soit $Y = a_0 + \sum_{i=1}^{n} a_i X_i$ où $a_0, a_2, \cdots, a_n$ sont des constantes. On a alors :*

$$Y \sim \mathcal{N}(a_0 + \sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2)\,.$$

Calculons la f.g.m. de $Y$ :

$$\begin{aligned} M_Y(t) &= E[e^{tY}] \\ &= E[e^{t(a_0 + \sum_{i=1}^{n} a_i X_i)}] \\ &= e^{ta_0} \prod_{i=1}^{n} E[e^{ta_i X_i}] = e^{ta_0} \prod_{i=1}^{n} M_{X_i}(ta_i) \\ &= e^{ta_0} \prod_{i=1}^{n} e^{ta_i\mu_i + t^2 a_i^2 \sigma_i^2/2} \\ &= e^{t\{a_0 + \sum_{i=1}^{n} a_i \mu_i\} + t^2 \{\sum_{i=1}^{n} a_i^2 \sigma_i^2\}/2} \end{aligned}$$

On reconnaît la f.g.m. d'une loi normale de moyenne $a_0 + \sum_{i=1}^{n} a_i \mu_i$ et de variance $\sum_{i=1}^{n} a_i^2 \sigma_i^2$.

**Exemple 1.1.1**
*Soit $X \sim \mathcal{N}(\mu, \sigma^2)$. Appliquons le théorème précédent avec $n = 1$, $a_0 = -\mu/\sigma$ et $a_1 = 1/\sigma$. On obtient alors $Z = \dfrac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$, la loi normale standard.*
*Cette loi joue un rôle primordial en calcul de probabilité. Notons sa densité et sa fonction de répartition respectivement par : $\phi(\cdot)$ et $\Phi(\cdot)$. Elles sont respectivement définies par : $\phi(z) = e^{-z^2/2}/\sqrt{2\pi}$ et $\Phi(z) = \displaystyle\int_{-\infty}^{z} \phi(t)dt$.*
*Notons qu'il n'y a pas d'expressions explicites pour $\Phi(\cdot)$. Ses valeurs numériques sont données dans des tables qu'on trouve dans presque tous les livres de statistique et probabilité, des logiciels d'analyses statistiques et certaines calculatrices scientifiques.*
*Ainsi, on peut calculer numériquement des probabilités du type $P(a < X \leq b)$, $-\infty \leq a \leq b \leq \infty$, comme suit :*

$$\begin{aligned} P(a < X \leq b) &= P(\frac{a-\mu}{\sigma} < Z \leq \frac{b-\mu}{\sigma}) \\ &= \Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma}) \end{aligned}$$

**Exemple 1.1.2**
*Soit $\{X_1, X_2, \cdots, X_n\}$ un échantillon issu d'une loi normale $\mathcal{N}(\mu, \sigma^2)$, c'est-à-dire $n$ variables aléatoires i.i.d.*
*Appliquons le théorème précédent avec $a_0 = 0$ et $a_i = 1/n$ pour $i = 1, \cdots, n$.*
*On obtient alors :*

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}(\mu, \sigma^2/n)\,.$$

## 1.2 La loi du khi-deux

Dans cette section, on définit et on étudie les propriétes élementaires de la loi du khi-deux.

**Définition 1.2.1**
*Soit $k$ un entier strictement positif. On dit qu'une variable aléatoire continue $X$ suit une loi de khi-deux et on écrit $X \sim \chi^2_k$ si et seulement si la densité de $X$ s'écrit :*

$$f(x) = \begin{cases} \dfrac{1}{\Gamma(k/2)2^{k/2}} x^{(k/2)-1} e^{-x/2} & si\ x \geq 0 \\ \\ 0 & sinon \end{cases} \tag{1.2.1}$$

*où : $\Gamma(\alpha) = \displaystyle\int_0^{\infty} y^{\alpha-1}e^{-y}\,dy$ (appelée fonction gamma). $k$ est appelé le nombre de degrés de liberté de cette loi.*

Vérifions que la fonction définie par (1.2.1) est bel et bien une densité. Il est évident que cette fonction est positive. Reste à vérifier que son intégrale vaut 1.

Par définition, pour $r > 0$, on a : $\Gamma(r/2) = \displaystyle\int_0^\infty y^{(r/2)-1} e^{-y}\, dy$. Effectuons le changement de variable $y = x/2$. Cette dernière intégrale est alors égale à :

$$\begin{aligned}\Gamma(r/2) &= \int_0^\infty (\frac{x}{2})^{(r/2)-1} e^{-x/2}\, \frac{dx}{2} \\ &= \int_0^\infty \frac{1}{2^{r/2}} x^{(r/2)-1} e^{-x/2}\, dx\,.\end{aligned}$$

On en déduit que :

$$\int_0^\infty \frac{1}{2^{r/2}\Gamma(r/2)} x^{(r/2)-1} e^{-x/2}\, dx = 1\,,$$

ce qu'il fallait démontrer.

À toutes fins utiles, on rappelle quelques propriétés de la fonction $\Gamma$. On a :

$$\Gamma(p) = (p-1)\Gamma(p-1) \quad \text{pour} \quad p > 0$$

$$\Gamma(p) = (p-1)! \quad \text{pour} \quad p \in \mathbb{N}^*$$

$$\Gamma(1/2) = \sqrt{\pi} \qquad .$$

**Proposition 1.2.1**
*La f.g.m. de la loi du khi-deux à $r$ degrés de liberté est donnée par :*

$$M(t) = (1-2t)^{-r/2} \text{ pour } t < 1/2\,.$$

**Démonstration** :
En effet, on a :

$$\begin{aligned}M(t) &= E[e^{Xt}] \\ &= \int_0^\infty e^{xt} \frac{1}{\Gamma(r/2)2^{r/2}} x^{(r/2)-1} e^{-x/2}\, dx \\ &= \int_0^\infty \frac{e^{-\frac{x}{2}(1-2t)} x^{(r/2)-1}}{\Gamma(r/2)2^{r/2}}\, dx \\ &= \int_0^\infty \frac{e^{-\frac{y}{2}} y^{(r/2)-1}}{\Gamma(r/2)2^{r/2}(1-2t)^{r/2}}\, dy = (1-2t)^{-r/2}\end{aligned}$$

Le passage de la troisième à la quatrième ligne s'est effectué à l'aide du changement de variable $y = (1-2t)x$. Cette intégrale ne converge que si $1 - 2t > 0$, i.e. si $t < 1/2$.

**Proposition 1.2.2**
*Soient $Z \sim \mathcal{N}(0,1)$ et $X = Z^2$. Alors $X \sim \chi^2_1$.*

**Démonstration** :
Calculons la densité $f_X(x)$ pour $x > 0$. Elle est égale à :

$$\begin{aligned} f_X(x) &= \frac{d}{dx} F_X(x) \\ &= \frac{d}{dx} P(X \leq x) \\ &= \frac{d}{dx} P(Z^2 \leq x) \\ &= \frac{d}{dx} P(-\sqrt{x} \leq Z \leq \sqrt{x}) \\ &= \frac{d}{dx} \{\Phi(\sqrt{x}) - \Phi(-\sqrt{x})\} \\ &= \frac{d}{dx} \{2\Phi(\sqrt{x}) - 1\} \\ &= 2\phi(\sqrt{x}) \frac{d}{dx} \{\sqrt{x}\} \\ &= \frac{1}{\sqrt{2\pi x}} e^{-x/2} \,. \end{aligned}$$

On reconnaît alors la densité d'une loi khi-deux à 1 degré de liberté.

On représente dans la figure (1.4) ci-dessous quelques densités de lois du khi-deux.

**Théorème 1.2.1**
*Soient $U$ et $V$ deux variables aléatoires indépendantes telles que $U \sim \chi^2_u$ et $V \sim \chi^2_v$. Alors, on a : $W = U + V \sim \chi^2_{u+v}$.*

Calculons la f.g.m de $W$ :

$$\begin{aligned} M_W(t) &= E[e^{Wt}] = E[e^{(U+V)t}] \\ &= M_U(t) M_V(t) \text{ grâce à l'indépendance de } U \text{ et de } V \\ &= (1-2t)^{-u/2} (1-2t)^{-v/2} = (1-2t)^{-(u+v)/2} \,. \end{aligned}$$

On reconnaît la f.g.m de la loi du khi-deux à $u + v$ degrés de liberté.

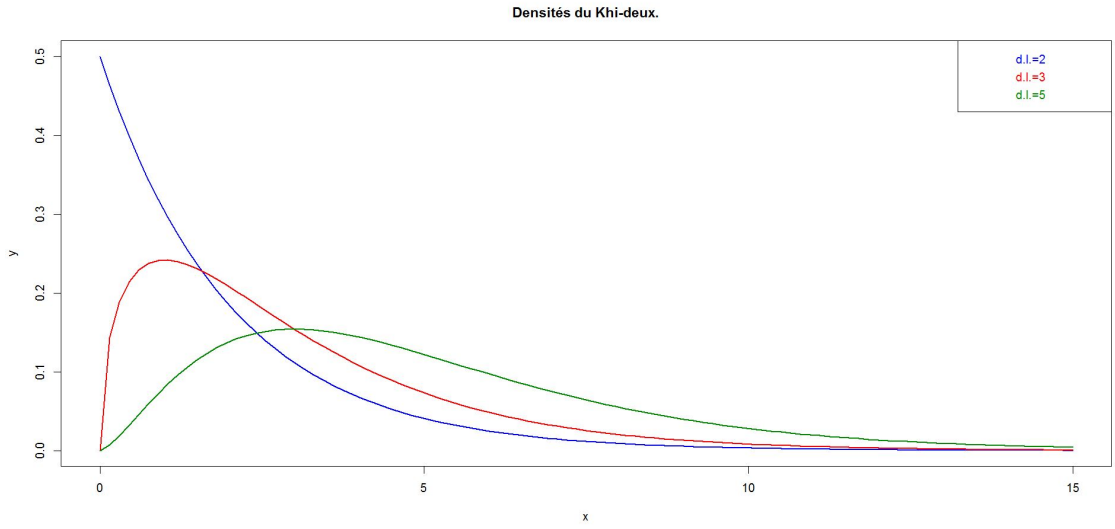


FIGURE 1.4 – Densités de lois du khi-deux

**Exemple 1.2.1**
*Soit* $\{X_1, X_2, \cdots, X_n\}$ *un échantillon issu d'une loi normale* $N(\mu, \sigma^2)$. *Posons :*

$$S_*^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 .$$

*D'après la proposition (1.2.2), on a* $((X_i - \mu)/\sigma)^2 \sim \chi_1^2$ *pour* $i = 1, 2, \cdots, n$.
*Et d'après la proposition (1.2.1), on en déduit immédiatement que :*

$$n\frac{S_*^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi_n^2 .$$

## 1.3 La loi $t$ de Student

**Définition 1.3.1**
*La loi du* $t$ *de Student est la loi continue de densité donnée par :*

$$f(t) = \frac{1}{\sqrt{r\pi}}\frac{\Gamma((r+1)/2)}{\Gamma(r/2)}\frac{1}{(1 + t^2/r)^{(r+1)/2}} \qquad t \in \mathbb{R},$$

*où* $r$ *est un entier strictement positif, appelé nombre de degrés de liberté. On écrit* $T \sim t_r$.

Si $r = 1$, l'espérance de $T$ n'existe pas. Si $r > 1$, on a $E[T] = 0$.
Si $r \leq 2$, la variance de $T$ n'existe pas. Si $r > 2$, on a $Var[T] = r/(r-2)$.

Tout comme la loi normale standard $\mathcal{N}(0,1)$, la loi de student $t_r$ est symétrique, centrée et son graphe est en forme de cloche. Lorsque $r$ devient grand, la loi $t$ du Student converge vers la loi normale standard.

Ci-dessous (voir figure (1.5)) sont tracées quelques densités de lois de Student :

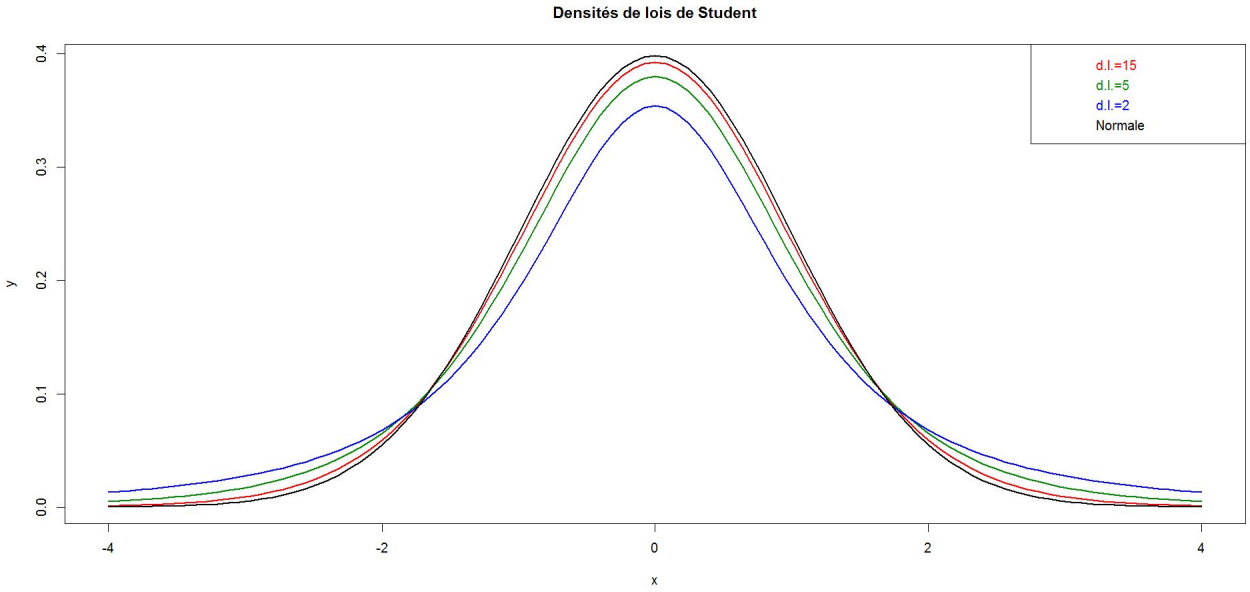


FIGURE 1.5 – Densités de lois de Student

On observera qu'elles sont toutes centrées, symétriques par rapport à l'axe des $y$.

**Proposition 1.3.1**
*Soient $W$ et $V$ deux variables aléatoires indépendantes telles que $W \sim \mathcal{N}(0,1)$ et $V \sim \chi^2_r$. Posons : $T = W/\sqrt{V/r}$. Alors, on a : $T \sim t_r$*

Pour démontrer cette proposition, on effectue le changement de variable :

$$\begin{cases} t &= w/\sqrt{v/r} \\ u &= v\,. \end{cases}$$

Ce qui nous donne $w = t\sqrt{u/r}$ et $v = u$. La densité bivariée de $T$ et de $U$ s'écrit alors :

$$\begin{aligned} g(t,u) &= \phi\left(t\sqrt{\frac{u}{r}}\right) f_{\chi^2_r}(u)|J| \\ &= \begin{cases} \dfrac{1}{\sqrt{2\pi}\Gamma(r/2)2^{r/2}} u^{r/2-1} e^{-\frac{u}{2}(1+\frac{t^2}{r})} \dfrac{\sqrt{u}}{\sqrt{r}} & \text{si } u \geq 0 \\ 0 & \text{sinon} \end{cases} \end{aligned}$$

La densité marginale de $T$ s'obtient alors en intégrant la fonction de densité bivariée :

$$\begin{aligned} f_T(t) &= \int_0^\infty g(t,u)\,du \\ &= \int_0^\infty \frac{1}{\sqrt{2\pi r}\Gamma(r/2)2^{r/2}} u^{((r+1)/2)-1} e^{-\frac{u}{2}(1+\frac{t^2}{r})}\,du \\ &= \int_0^\infty \frac{1}{\sqrt{2\pi r}\Gamma(r/2)2^{r/2}} \left(\frac{2z}{1+t^2/r}\right)^{((r+1)/2)-1} e^{-z} \left(\frac{2}{1+t^2/r}\right) dz \\ &= \frac{\Gamma[(r+1)/2]}{\sqrt{\pi r}\Gamma[r/2]} \frac{1}{(1+t^2/r)^{(r+1)/2}}, \end{aligned}$$

où le passage de la ligne 2 à la ligne 3 se fait par un changement de variable : $z = \dfrac{u}{2}\left(1+\dfrac{t^2}{r}\right)$.
On reconnaît alors l'expression de la densité d'une loi $t_r$ de Student.

## 1.4 La loi $F$ de Fisher

**Définition 1.4.1**
*La loi $F$ de Fisher est la loi continue de densité donnée par :*

$$f(w) = \begin{cases} \dfrac{\Gamma((n+m)/2)(n/m)^{n/2}}{\Gamma(n/2)\Gamma(m/2)} \dfrac{w^{(n/2)-1}}{(1+nw/m)^{(n+m)/2}} & si\ w \geq 0 \\ \\ 0 & sinon \end{cases}$$

*où $n$ et $m$ sont des entiers strictement positifs appelés nombres de degrés de liberté. On écrit alors que : $W \sim F_{n,m}$.*

Si $m \leq 2$, l'espérance de $W$ n'existe pas. Si $m > 2$, on a $E[W] = m/(m-2)$.

Si $m \leq 3$, la variance de $W$ n'existe pas. Si $m > 4$, on a $Var[W] = \dfrac{2m^2(n+m-2)}{n(m-2)^2(m-4)}$.

On trouvera ci-dessous quelques densités de lois de Fisher (voir figure (1.6)) et figure (1.7)).

**Proposition 1.4.1**
*Soient $U$ et $V$ deux variables aléatoires indépendantes telles que $U \sim \chi^2_n$ et $V \sim \chi^2_m$. Posons $W = (U/n)/(V/m)$. Alors, on a : $W \sim F_{n,m}$.*

Cette proposition se démontre de la même manière que la proposition 1.3.1.
À partir de la proposition (1.4.1), on déduit que si $W \sim F_{n,m}$, alors $1/W \sim F_{m,n}$.

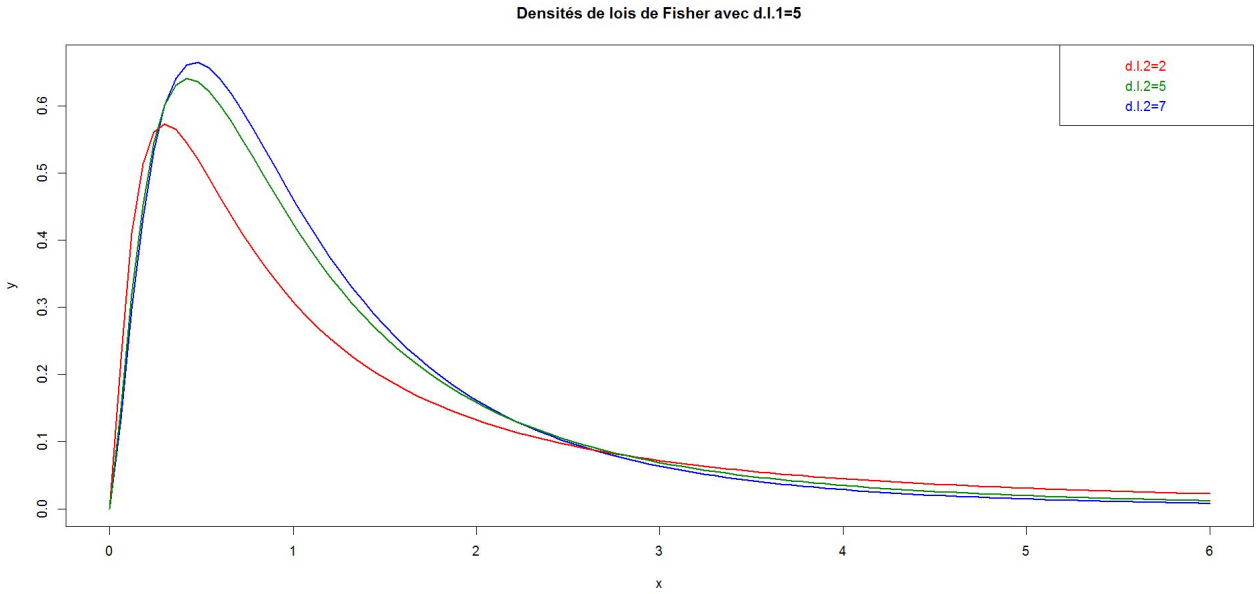


FIGURE 1.6 – Densités de lois de Fisher avec d.l.1=5

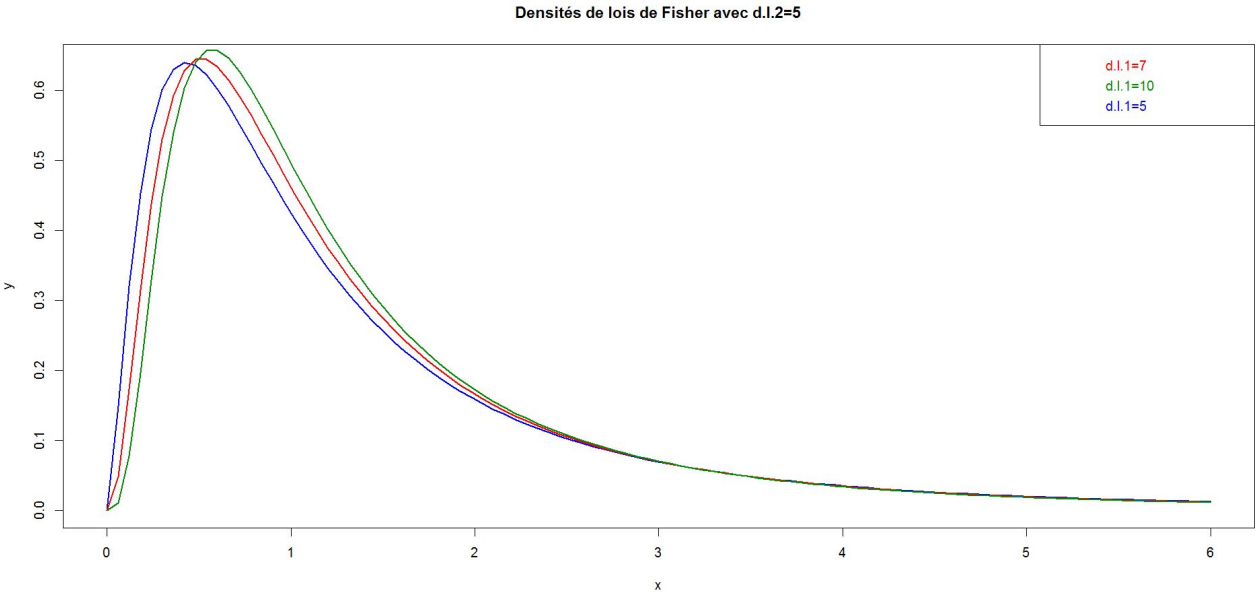


FIGURE 1.7 – Densités de lois de Fisher avec d.l.2=5

**Exemple 1.4.1**

*Soient* $\{X_1, X_2, \cdots X_n\}$ *et* $\{Y_1, Y_2, \cdots Y_m\}$ *deux échantillons issus respectivement de lois* $\mathcal{N}(\mu_X, \sigma_X^2)$ *et* $\mathcal{N}(\mu_Y, \sigma_Y^2)$*. On suppose que les moyennes théoriques* $\mu_X$ *et* $\mu_Y$ *sont connues.*

*D'après l'exemple (1.2.1),* $nS_{X,*}^2/\sigma_X^2 \sim \chi_n^2$ *et* $mS_{Y,*}^2/\sigma_Y^2 \sim \chi_m^2$*.*

*Les échantillons étant indépendants, on a donc :*

$$(S^2_{X,*}/\sigma^2_X)/(S^2_{Y,*}/\sigma^2_Y) \sim F_{n,m}\,,$$

*d'après la proposition (1.4.1).*
*Cette propriété est utilisée pour trouver un intervalle de confiance au niveau* $1-\alpha$ *pour le rapport de variances* $\sigma^2_X/\sigma^2_Y$*, lorsque les moyennes théoriques* $\mu_X$ *et* $\mu_Y$ *sont connues.*
*Celui-ci s'écrit alors :*

$$\left[\frac{1}{\mathcal{F}_{\alpha/2,n,m}}\frac{S^2_{X,*}}{S^2_{Y,*}}, \frac{1}{\mathcal{F}_{1-\alpha/2,n,m}}\frac{S^2_{X,*}}{S^2_{Y,*}}\right],$$

*où* $\mathcal{F}_{\gamma,r,s}$ *est le quantile d'ordre* $1-\gamma$ *de la loi* $F_{r,s}$*.*
*Donc si* $W \sim F_{r,s}$*,* $\mathcal{F}_{\gamma,r,s}$ *est la valeur telle que* $P\{W > \mathcal{F}_{\gamma,r,s}\} = \gamma$*.*
*Cette même propriété est utilisée pour tester l'égalité de variances de deux échantillons indépendants lorsque les moyennes théoriques sont connues. On rejette l'hypothèse* $H_0 : \sigma^2_X = \sigma^2_Y$ *si et seulement si* 1 *n'appartient pas à l'intervalle de confiance ci-dessus, c'est à dire si :*

$$\frac{S^2_{X,*}}{S^2_{Y,*}} \geq \mathcal{F}_{\alpha/2,n,m},$$

*ou*

$$\frac{S^2_{X,*}}{S^2_{Y,*}} \leq \mathcal{F}_{1-\alpha/2,n,m}.$$

## 1.5 Les lois non centrées

Dans cette section, on introduit les lois du khi-deux et de Fisher non centrées. En analyse de la variance, certaines statistiques étudiées suivent ces distributions sous l'hypothèse alternative, c'est-à-dire lorsque les moyennes ne sont pas toutes égales. Ces distributions servent alors pour le calcul de la puissance des tests.

**Définition 1.5.1**
*Soient* $n$ *variables aléatoires indépendantes* $\{X_1, X_2, \cdots, X_n\}$ *telles que* $X_i \sim \mathcal{N}(\mu_i, 1)$ *pour* $i = 1, \cdots, n$*. Posons* $Y = \sum_{i=1}^{n} X_i^2$*. Alors la loi de* $Y$ *ne dépend que de* $n$ *et* $\delta = \sum_{i=1}^{n} \mu_i^2$*. Sa f.g.m. est égale à* $M_Y(t) = (1-2t)^{-n/2} e^{\frac{t\delta}{1-2t}}$ *pour* $t < 1/2$*. On dit que* $Y$ *suit une loi non centrée du khi-deux à* $n$ *degrés de liberté et de paramètre de non-centralité* $\delta$*.*

Calculons la f.g.m de $Y$. Grâce à l'indépendance des $X_i$, on a :

$$M_Y(t) = E[e^{Yt}] = \prod_{i=1}^{n} E[e^{X_i^2 t}]$$

Pour $i = 1, \cdots, n$ et $t < 1/2$, on a :

$$\begin{aligned} E[e^{X_i^2 t}] &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{x^2 t} e^{-\frac{(x-\mu_i)^2}{2}}\, dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(1-2t)(x-\frac{\mu_i}{1-2t})^2 + \frac{t\mu_i^2}{1-2t}}\, dx \\ &= (1-2t)^{-1/2} e^{\frac{t\mu_i^2}{1-2t}} \int_{-\infty}^{\infty} \sqrt{\frac{1-2t}{2\pi}} e^{-\frac{1}{2}(1-2t)(x-\frac{\mu_i}{1-2t})^2}\, dx \\ &= (1-2t)^{-1/2} e^{\frac{t\mu_i^2}{1-2t}}. \end{aligned}$$

La dernière intégrale ci-dessus est égale à 1 puisqu'on intègre la densité d'une loi normale de moyenne $\mu_i(1-2t)^{-1}$ et de variance $(1-2t)^{-1}$ sur l'ensemble $\mathbb{R}$. Finalement, on obtient :

$$M_Y(t) = (1-2t)^{-n/2} e^{\frac{t\delta}{1-2t}}$$

Le cas $\delta = 0$ correspond à la loi du khi-deux présentée à la section (1.2). Un calcul de dérivation nous donne :

$$M_Y'(0) = n + \delta \quad \text{et} \quad M_Y''(0) = n^2 + 2n(\delta+1) + \delta(\delta+4)\,.$$

On peut alors annoncer la proposition suivante :

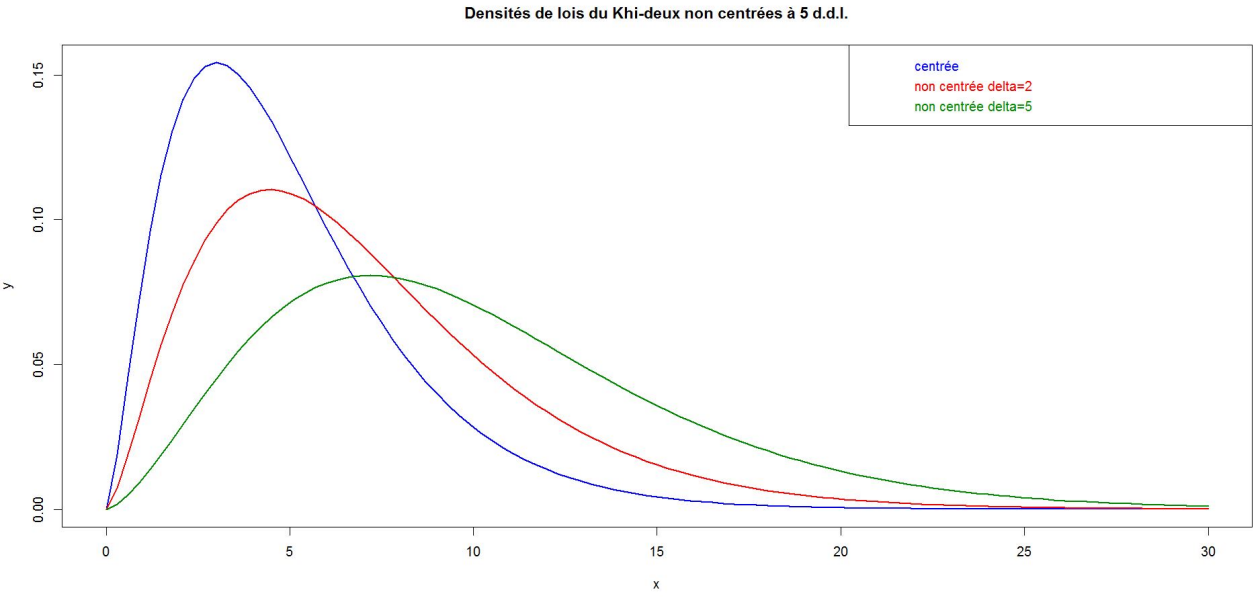


FIGURE 1.8 – Densités de lois du khi-deux non centrées avec d.d.l=5

**Proposition 1.5.1**
*La moyenne et la variance d'une loi du khi-deux non-centrée à $n$ degrés de liberté et de paramètre de non centralité $\delta$ sont égales respectivement à $n+\delta$ et $2(2\delta+n)$.*

La figure (1.8) précédente représente les densités de différentes lois du khi-deux non-centrées à 5 degrés de liberté avec des paramètres de non centralité respectives : $\delta = 0$, 2 et 5. À degrés de liberté constants, la figure (1.8) montre que la moyenne et la variabilité augmentent avec le paramètre $\delta$.

**Théorème 1.5.1**
*Soient $U$ et $V$ deux variables aléatoires telles que $U \sim \chi^2_n(\delta)$ et $V \sim \chi^2_m$.*
*Posons : $W = (U/n)/(V/m)$. Alors $W$ suit une loi $F$ de Fisher décentrée à $n$ et $m$ degrés de liberté et de paramètre de non centralité $\delta$. On écrit alors $W \sim F_{n,m}(\delta)$.*

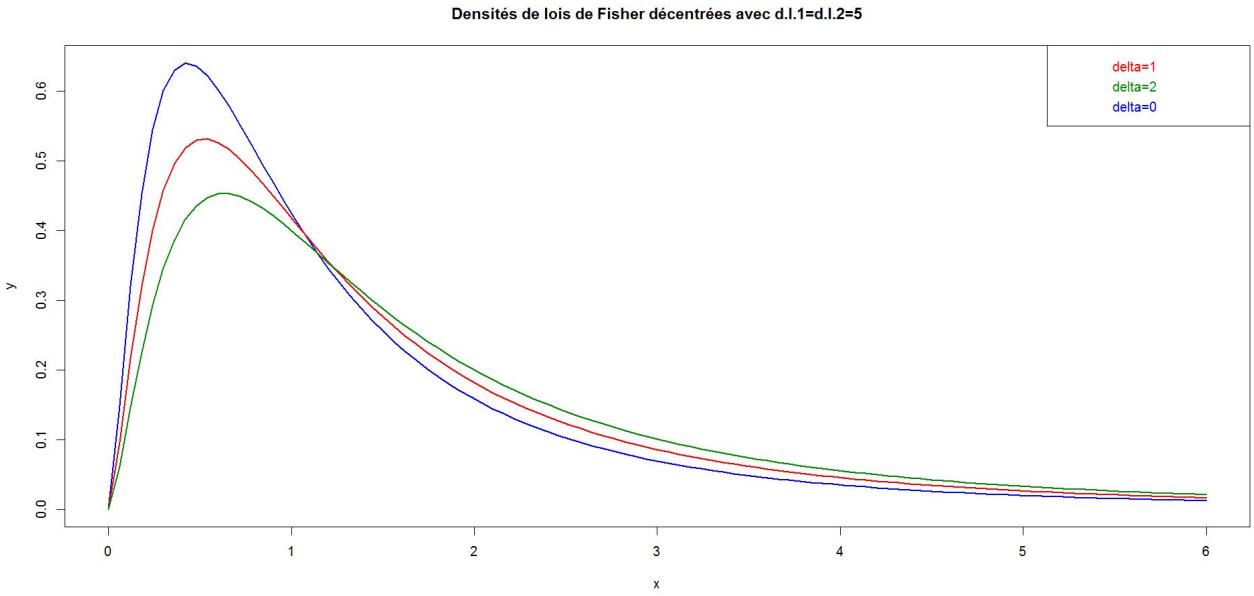


FIGURE 1.9 – Densités de lois de Fisher décentrées avec d.l.1=d.l.2=5

## 1.6 Formes quadratiques

Dans cette section, on introduit les formes quadratiques aléatoires et on étudie leurs propriétés. En effet, plusieurs statistiques qui interviennent lors de l'analyse de la variance peuvent s'écrire comme une forme quadratique aléatoire et les propriétés qui seront présentées dans cette section nous serviront à identifier les lois et à montrer l'indépendance de ces statistiques.

**Définition 1.6.1**
*Soient $n$ variables indépendantes $\mathbf{X} = \{X_1, X_2, \cdots, X_n\}$. Une forme quadratique en $\mathbf{X}$ est une variable aléatoire $Q$ qui peut s'écrire sous la forme* $Q = \sum_{i=1}^{n} a_{ii} X_i^2 + 2 \sum_{1=i<j=n} a_{ij} X_i X_j$.

En écriture matricielle, soit $\mathbf{X} = (X_1, X_2, \cdots, X_n)^T$ où l'opérateur $(\cdot)^T$ désigne le transposé. On a alors $Q = \mathbf{X}^T \mathbf{A} \mathbf{X}$ où $\mathbf{A}$ est la matrice symétrique de taille $n$ ayant le terme $a_{ij}$ à la $i^{\text{ème}}$ ligne et $j^{\text{ème}}$ colonne.

**Exemple 1.6.1**
*Soit $\mathbf{X} = \{X_1, X_2, \cdots, X_n\}$ un échantillon de taille $n$. Le carré de la moyenne arithmétique $\bar{X}^2$ est une forme quadratique aléatoire en $\mathbf{X}$. En effet, on a :*

$$\bar{X}^2 = \left\{ \frac{1}{n} \sum_{i=1}^{n} X_i \right\}^2 = \sum_{i=1}^{n} \frac{1}{n^2} X_i^2 + \frac{2}{n^2} \sum_{1=i<j=n} X_i X_j \,.$$

*Tous les éléments de la matrice $\mathbf{A}$ correspondante sont égaux à $\frac{1}{n^2}$. La matrice $\mathbf{A}$ s'écrit alors : $\mathbf{A} = \frac{1}{n^2} \mathbf{1_n} \mathbf{1_n}^T$ où $\mathbf{1_n}$ est le vecteur de taille $n$ dont tous les éléments sont égaux à 1.*

**Exemple 1.6.2**
*Considérons maintenant la variance de l'échantillon définie par :*

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

*La statistique $(n-1)S^2$ est une forme quadratique aléatoire en $\mathbf{X}$. En effet, on a :*

$$\begin{aligned} (n-1)S^2 &= \sum_{i=1}^{n} \left( X_i - \frac{X_1 + X_2 + \cdots + X_n}{n} \right)^2 \\ &= \frac{n-1}{n} (X_1^2 + X_2^2 + \cdots + X_n^2) - \frac{2}{n} (X_1 X_2 + X_1 X_3 + \cdots + X_{n-1} X_n) \,. \end{aligned}$$

*Soit $\mathbf{B}$ la matrice associée à cette forme quadratique. Tous les élements de la diagonale de cette matrice sont égaux à $(n-1)/n = 1 - 1/n$ et tous les élements hors diagonale sont égaux à $-1/n$. On a alors : $\mathbf{B} = \mathbf{I_n} - \frac{1}{n} \mathbf{1_n} \mathbf{1_n}^T$ où $\mathbf{I_n}$ est la matrice identité de taille $n$.*

Le théorème suivant énonce les conditions nécessaires et suffisantes pour qu'une forme quadratique aléatoire suive une loi du Khi-deux.

**Théorème 1.6.1**
*Soit* $\mathbf{X} = (X_1, X_2, \cdots, X_n)^T$ *un vecteur aléatoire de variables indépendantes telles que* $X_i \sim \mathcal{N}(\mu_i, \sigma^2)$ *pour* $i = 1, \cdots, n$. *(de même variance). Soit* $\mathbf{A}$ *une matrice symétrique de taille* $n$ *et de rang* $d$, $0 < d \leq n$. *Posons* $Q = \mathbf{X}^T\mathbf{A}\mathbf{X}$. *Alors :* $Q/\sigma^2 \sim \chi^2_d(\delta)$ *où* $\delta = \mu^T A\mu/\sigma^2$ *et* $\mu = (\mu_1, \cdots, \mu_n)^T$ *si et seulement si la matrice* $\mathbf{A}$ *est idempotente, i.e. si et seulement si* $\mathbf{A}^2 = \mathbf{A}$.

Dans l'exemple qui suit, on utilise le dernier théorème pour démontrer un résultat classique de la théorie d'échantillonnage, à partir d'une population normale. Ce résultat nous sera très utile pour la suite de ce cours.

**Exemple 1.6.3**
*Reconsidérons l'exemple (1.6.2) dans le cas où le vecteur* $\mathbf{X}$ *est un échantillon de taille* $n$ *issu d'une loi normale de moyenne* $\mu$ *et de variance* $\sigma^2$. *Dans ce cas, on a* $E[\mathbf{X}] = \mu\mathbf{1_n}$. *On a vu que la statistique* $Q = (n-1)S^2$ *est une forme quadratique et que sa matrice correspondante est égale à* $\mathbf{B} = \mathbf{I_n} - \dfrac{1}{n}\mathbf{1_n}\mathbf{1_n^T}$. *On a ici :*

$$\mathbf{B^2} = \mathbf{I_n} - \frac{1}{n}\mathbf{1_n}\mathbf{1_n}^T + \frac{1}{n^2}\mathbf{1_n}\mathbf{1_n}^T\mathbf{1_n}\mathbf{1_n}^T - \frac{1}{n}\mathbf{1_n}\mathbf{1_n}^T.$$

*Or on a* $\mathbf{1_n^T}\mathbf{1_n} = n$, *donc les deux derniers termes du membre droit de la dernière équation s'annulent et ainsi :* $\mathbf{B^2} = \mathbf{B}$.
*D'après le théorème précédent,* $Q/\sigma^2 = (n-1)S^2/\sigma^2$ *suit une loi du khi-deux non centrée à* $d$ *degrés de liberté et de paramètre* $\delta = (\mu^2/\sigma^2)\mathbf{1_n}^T\mathbf{B}\mathbf{1_n}$ *où* $d$ *est égal au rang de la matrice* $\mathbf{B}$. *La somme de toutes les colonnes* $\mathbf{1_n}^T\mathbf{B}$ *de la matrice* $\mathbf{B}$ *est nulle, on a donc* $\delta = 0$ *et* $d \leq n-1$. *Les* $n-1$ *premières colonnes de* $\mathbf{B}$ *sont indépendantes donc* $d = n-1$.

**Exemple 1.6.4**
*Soient* $\{X_1, X_2, \cdots, X_n\}$ *et* $\{Y_1, Y_2, \cdots, Y_m\}$ *deux échantillons issus respectivement de lois* $\mathcal{N}(\mu_X, \sigma^2_X)$ *et* $\mathcal{N}(\mu_Y, \sigma^2_Y)$.
*D'après l'exemple (1.6.3), on a* $(n-1)S^2_X/\sigma^2_X \sim \chi^2_{n-1}$ *et* $(m-1)S^2_Y/\sigma^2_Y \sim \chi^2_{m-1}$.
*Les deux échantillons sont indépendants donc* $(S^2_X/\sigma^2_X)/(S^2_Y/\sigma^2_Y) \sim F_{n-1,m-1}$ *d'après la proposition (1.4.1). Cette propriété est utilisée pour trouver un intervalle de confiance au niveau* $1-\alpha$ *pour le rapport de variances* $\sigma^2_X/\sigma^2_Y$ *lorsque les moyennes théoriques* $\mu_X$ *et* $\mu_Y$ *sont inconnues. Celui-ci s'écrit alors :*

$$\left[\frac{1}{\mathcal{F}_{\alpha/2,n-1,m-1}}\frac{S^2_X}{S^2_Y}, \frac{1}{\mathcal{F}_{1-\alpha/2,n-1,m-1}}\frac{S^2_X}{S^2_Y}\right]$$

*Cette même propriété est utilisée pour tester l'égalité de variances de deux échantillons indépendants lorsque les moyennes théoriques sont inconnues. On rejette l'hypothèse* $H_0 : \sigma^2_X =$

*$\sigma_Y^2$ si et seulement si 1 n'appartient pas à l'intervalle de confiance ci-dessus, c'est-à-dire si et seulement si :*

$$\frac{S_X^2}{S_Y^2} \geq \mathcal{F}_{\alpha/2,n-1,m-1}$$

*ou*

$$\frac{S_X^2}{S_Y^2} \leq \mathcal{F}_{1-\alpha/2,n-1,m-1}$$

**Théorème 1.6.2**
*Soit $\mathbf{X} = (X_1, X_2, \cdots, X_n)^T$ un vecteur aléatoire de variables indépendantes telles que $X_i \sim \mathcal{N}(\mu_i, \sigma^2)$ pour $i = 1, \cdots, n$. (de même variance). Soient $\mathbf{A}$ et $\mathbf{B}$ deux matrices symétriques de taille $n$ et soient $Q_1 = \mathbf{X}^T\mathbf{A}\mathbf{X}$ et $Q_2 = \mathbf{X}^T\mathbf{B}\mathbf{X}$ leurs formes quadratiques associées.*
*Alors : $Q_1$ et $Q_2$ sont indépendantes si et seulement si $\mathbf{AB} = \mathbf{BA} = \mathbf{0}$.*

**Exemple 1.6.5**
*Reconsidérons les deux exemples (1.6.1) et (1.6.2) dans le cas où le vecteur $\mathbf{X}$ est un échantillon de taille $n$ issu d'une loi normale de moyenne $\mu$ et de variance $\sigma^2$. On a vu que $\bar{X}^2$ et $(n-1)S^2$ sont des formes quadratiques de matrices associées respectives $\mathbf{A} = \frac{1}{n^2}\mathbf{1_n}\mathbf{1_n}^T$ et $\mathbf{B} = \mathbf{I_n} - \frac{1}{n}\mathbf{1_n}\mathbf{1_n}^T$. Calculons le produit matriciel $\mathbf{AB}$. On a :*

$$\begin{aligned}\mathbf{AB} &= \frac{1}{n^2}\mathbf{1_n}\mathbf{1_n}^T(\mathbf{I_n} - \frac{1}{n}\mathbf{1_n}\mathbf{1_n}^T) \\ &= \frac{1}{n^2}\mathbf{1_n}\mathbf{1_n}^T - \frac{1}{n^3}\mathbf{1_n}\mathbf{1_n}^T\mathbf{1_n}\mathbf{1_n}^T.\end{aligned}$$

*Or $\mathbf{1_n^T}\mathbf{1_n} = n$, donc on a : $\mathbf{AB} = \mathbf{0}$. Ainsi, d'après le théorème précédent, les statistiques $S^2$ et $\bar{X}$ sont indépendantes.*

*On vient donc de démontrer un résultat classique en théorie de l'échantillonnage d'une loi normale. On peut résumer les résultats antérieurs dans la proposition suivante :*

***Proposition 1.6.1***
*Soit $\{X_1, X_2, \cdots, X_n\}$ un échantillon issu d'une loi normale $\mathcal{N}(\mu, \sigma^2)$.*
*Soient $\bar{X}$ et $S^2$ la moyenne et la variance de l'échantillon, respectivement définies par $\bar{X} = \sum_{i=1}^{n} X_i/n$ et $S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$. On a alors :*

1. *$\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$ (exemple (1.1.2))*
2. *$(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$ (exemple (1.6.3))*
3. *$\bar{X}$ et $S^2$ sont indépendantes (exemple (1.6.5))*
4. *$(\bar{X} - \mu)/\sqrt{S^2/n} \sim t_{n-1}$ (proposition (1.3.1))*

# Exercices

**Exercice 1**

La taille moyenne de 500 élèves des petites classes d'un lycée est 1,51 m et l'écart-type est 0,15 m. On suppose que la taille suit une loi normale.

1. Combien d'élèves ont une taille comprise entre 1,2 m et 1,55 m ?
2. Combien d'élèves mesurent au moins 1,85 m ?
3. Combien d'élèves ont une taille inférieure à 1,28 m ?

**Exercice 2**

Le diamètre intérieur moyen d'un échantillon de 200 rondelles produites par une machine est égal à 1,275 cm et l'écart-type à 0,013 cm. L'usage que l'on fait des rondelles nécessite que le diamètre varie entre des bornes de tolérance de 1,26 cm et 1,29 cm, sinon les rondelles sont considérées comme défectueuses. Déterminez le pourcentage de rondelles défectueuses produites par la machine, en supposant que ces diamètres sont de loi normale.

**Exercice 3**

Calculez : $\displaystyle\int_3^5 e^{-3(x-4)^2}\,dx$ .

**Exercice 4**

Soit $X$ une variable aléatoire de f.g.m. $M_X(t) = e^{3t+8t^2}$, définie au voisinage de l'origine. Calculez : $P(-1 < X < 8)$.

**Exercice 5**

On suppose que $Y$ suit une loi normale $\mathcal{N}(\mu, \sigma^2)$, et que $Y = \ln X$. Calculez $E(X)$ et $V(X)$.

**Exercice 6**

On suppose que $X$ est une variable aléatoire de loi gaussienne $\mathcal{N}(1, 4)$. Calculez la probabilité suivante : $P(1 < X^2 < 9)$.

**Exercice 7**

1. Calculez l'espérance et la variance d'une variable aléatoire $X$ de loi du khi-deux ayant pour nombre de degrés de liberté $n$.

 

2. Calculez l'espérance et la variance d'une variable aléatoire $Y$ de loi uniforme sur $[a, b]$.
3. Déterminez $a$ et $b$ tels que l'espérance et la variance d'une variable aléatoire de loi uniforme sur $[a, b]$ coïncident respectivement avec l'espérance et la variance d'une variable aléatoire d'une loi du khi-deux à 5 degrés de liberté.

**Exercice 8**

Les statistiques $\overline{X}$ et $S^2$ désignent les estimateurs usuels de $m$ et $\sigma^2$ pour un échantillon $(X_1, X_2, \cdots, X_n)$ d'une loi normale $\mathcal{N}(m, \sigma)$.

1. Calculer le coeficient de corrélation $\rho$ entre $\overline{X}$ et la statistique de Student : $T = \dfrac{\sqrt{n}(\overline{X} - m)}{S}$.
2. Sachant que, lorsque $k \to +\infty$, on a :

$$\frac{\Gamma\left(\dfrac{k+1}{2}\right)}{\sqrt{\dfrac{k}{2}}\,\Gamma\left(\dfrac{k}{2}\right)} \simeq 1 - \frac{1}{4k}\,,$$

donnez une approximation à l'ordre 1 de $\rho$.

**Exercice 9**

Soient $X_1$ et $X_2$ deux variables aléatoires indépendantes telles que : $X_1$ et $Y = X_1 + X_2$ aient pour lois respectives des lois du $\chi^2_{r_1}$ et $\chi^2_r$ (avec $r_1 < r$).

Montrez que la loi de $X_2$ est une loi du $\chi^2_{r-r_1}$.

**Exercice 10**

Soit $F$ une variable aléatoire de loi $F_{r_1,r_2}$. Calculez $E[F^k]$ pour tout $k < \dfrac{r_2}{2}$.

**Exercice 11**

Soient $X_1$ et $X_2$ deux variables aléatoires i.i.d. de loi à densité exponentielle de paramètre 1. Quelle est la loi de la variable aléatoire $V = X_1/X_2$ ?

**Exercice 12**

On considère trois variables aléatoires indépendantes $X_1, X_2$ et $X_3$ toutes de lois du khi-deux à, respectivement, $r_1, r_2$ et $r_3$ de degrés de liberté.

1. Montrez que $Y_1 = X_1/X_2$ et $Y_2 = X_1 + X_2$ sont indépendantes et que $Y_2 \sim \chi^2_{r_1+r_2}$.

 

2. En déduire que les variables aléatoires $(X_1/r_1)/(X_2/r_2)$ et $(X_3/r_3)/((X_1+X_2)/(r_1+r_2))$ sont indépendantes et suivent chacune la loi de Fisher.

**Exercice 13**

On considère trois variables aléatoires indépendantes $X_i$, pour $i = 1, 2, 3$ et telles que $X_i \sim \mathcal{N}(i, i^2)$. À partir de ces trois variables, construire des statistiques ayant pour lois respectives :

1. $\chi^2_3$
2. $t_2$
3. $F_{1,2}$ .

 

 

# Chapitre 2

# Comparaison de deux moyennes

## 2.1 Comparaison des moyennes de deux échantillons indépendants

On considère deux échantillons indépendants $\{X_{1,1}, X_{1,2}, \cdots, X_{1,n_1}\}$ et $\{X_{2,1}, X_{2,2}, \cdots, X_{2,n_2}\}$ formés respectivement de $n_1$ d'observations indépendantes d'une loi normale $\mathcal{N}(m_1, \sigma_1^2)$ et $n_2$ observations indépendantes d'une loi normale $\mathcal{N}(m_1, \sigma_1^2)$.

### 2.1.1 Variances égales

Dans cette première partie, on suppose que les deux populations possèdent la même variance théorique $\sigma^2$.

Dans ce contexte, on veut tester l'hypothèse $H_0 : m_1 = m_2$ contre une des hypothèses alternatives $H_1 : m_1 \neq m_2$ ou $H_1 : m_1 > m_2$ ou $H_1 : m_1 < m_2$.

On pose :

$$\overline{X}_{1,\bullet} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1,i} \qquad S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_{1,i} - \overline{X}_{1,\bullet})^2$$

$$\overline{X}_{2,\bullet} = \frac{1}{n_2}\sum_{i=1}^{n_2} X_{2,i} \qquad S_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(X_{2,i} - \overline{X}_{2,\bullet})^2$$

où $\overline{X}_{1,\bullet}$ et $\overline{X}_{2,\bullet}$ sont les moyennes théoriques des deux échantillons indépendants, où $S_1^2$ et $S_2^2$ sont des estimateurs non biaisés de $\sigma_1^2$ et $\sigma_2^2$ respectivement.

Grâce aux résultats du chapitre précédent, on a successivement :

$$\overline{X}_{1,\bullet} \sim \mathcal{N}(m_1, \frac{\sigma^2}{n_1}) \tag{2.1.1}$$

$$\overline{X}_{1,\bullet} \sim \mathcal{N}(m_2, \frac{\sigma^2}{n_2}) \quad , \tag{2.1.2}$$

et

$$\frac{n_1 - 1}{\sigma^2} S_1^2 \sim \chi^2_{n_1 - 1} \tag{2.1.3}$$

$$\frac{n_2 - 1}{\sigma^2} S_2^2 \sim \chi^2_{n_2 - 1} \quad , \tag{2.1.4}$$

et ces quatre statistiques sont indépendantes. De plus, comme les échantillons sont indépendants, grâce au chapitre 1, on a également :

$$\overline{X}_{1,\bullet} - \overline{X}_{2,\bullet} \sim \mathcal{N}\left(m_1 - m_2, \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right) \tag{2.1.5}$$

$$\frac{n_1 + n_2 - 2}{\sigma^2} S_p^2 = \frac{1}{\sigma^2}\left\{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2\right\} \sim \chi^2_{n_1 + n_2 - 2}\,. \tag{2.1.6}$$

$S_p^2$ est l'estimateur global de la variance calculée en utilisant les deux échantillons globalement :

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \tag{2.1.7}$$

On rappelle qu'on a supposé que les deux variances $\sigma_1^2$ et $\sigma_2^2$ sont égales.

### Variance connue

On suppose que la variance théorique commune aux deux populations est connue. Sous ces hypothèses, on a :

$$\overline{X}_{1,\bullet} - \overline{X}_{2,\bullet} \sim \mathcal{N}\left(m_1 - m_2, \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

Ainsi, sous $H_0$, on a :

$$\overline{X}_{1,\bullet} - \overline{X}_{2,\bullet} \sim \mathcal{N}\left(0, \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)\,,$$

ce qui s'écrit encore :

$$Z = \frac{\overline{X}_{1,\bullet} - \overline{X}_{2,\bullet}}{\sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim \mathcal{N}(0, 1)\,.$$

Donc, on rejette $H_0$ contre $H_1 : m_1 \neq m_2$ au seuil $(1-\alpha)$ si $|Z| > Z_{\alpha/2}$.

On rejette $H_0$ contre $H_1 : m_1 > m_2$ au seuil $(1-\alpha)$ si $Z > Z_\alpha$.

On rejette $H_0$ contre $H_1 : m_1 < m_2$ au seuil $(1-\alpha)$ si $Z < Z_\alpha$.

Cela nous permet également de construire un intervalle de confiance au niveau $(1-\alpha)$ pour $m_1 - m_2$ :

$$\left[ \left(\overline{X}_{1,\bullet} - \overline{X}_{2,\bullet}\right) - Z_{\alpha/2}\sqrt{\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\,, \left(\overline{X}_{1,\bullet} - \overline{X}_{2,\bullet}\right) + Z_{\alpha/2}\sqrt{\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \right]$$

**Variance inconnue**

En pratique, dans la plupart des cas, on ignore la variance théorique et on doit donc l'estimer.

La statistique $S_p^2$, définie en (2.1.7), est un estimateur de $\sigma^2$, tout comme $S_1^2$ et $S_2^2$. En effet, on a immédiatement :

$$E[S_1^2] = E[S_2^2] = E[S_p^2]\,.$$

Intuitivement, $S_p^2$ est un meilleur estimateur que $S_1^2$ et $S_2^2$, car il utilise toute l'information disponible dans les deux échantillons. On peut aussi vérifier que c'est le meilleur estimateur sans biais de $\sigma^2$ parmi toutes les combinaisons linéaires de $S_1^2$ et de $S_2^2$. Une telle combinaison linéaire s'écrit sous la forme :

$$\hat{\sigma}^2 = aS_1^2 + bS_2^2\,.$$

Cet estimateur est non biaisé, et donc :

$$E[\hat{\sigma}^2] = aE[S_1^2] + bE[S_2^2]\,,$$

et ainsi :

$$a + b = 1 \Longleftrightarrow b = 1 - a\,.$$

D'autre part, on a :

$$\begin{aligned} Var[\hat{\sigma}^2] &= a^2 Var[S_1^2] + b^2 Var[S_2^2] \\ &= a^2\frac{2\sigma^4}{n_1-1} + (1-a)^2\frac{2\sigma^4}{n_2-1}\,, \end{aligned}$$

en utilisant (2.1.3) et (2.1.4) et le fait qu'une v.a.r. $U$ de loi du khi-deux à $n$ degrés de liberté a pour variance $2n$.

On en déduit que :

$$Var[\hat{\sigma}^2] = 2\sigma^4\left\{\frac{a^2}{n_1-1} + \frac{(1-a)^2}{n_2-1}\right\}\,.$$

Il est alors aisé, en étudiant la fonction :

$$a \mapsto \frac{a^2}{n_1 - 1} + \frac{(1-a)^2}{n_2 - 1}\,,$$

de voir qu'elle est minimisée pour $a = \dfrac{n_1 - 1}{n_1 + n_2 - 2}$, ce qui correspond à $b = \dfrac{n_2 - 1}{n_1 + n_2 - 2}$. Le $\hat{\sigma}^2$ optimal est donc égal à $S_p^2$.

Grâce aux résultats du chapitre 1, en utilisant (2.1.5) et (2.1.6), on a :

$$T = \frac{\overline{X}_{1,\bullet} - \overline{X}_{2,\bullet}}{\sqrt{S_p^2 \left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}\,. \tag{2.1.8}$$

Donc, on rejette $H_0$ contre $H_1 : m_1 \neq m_2$ au seuil $(1-\alpha)$ si $|T| > t_{n_1+n_2-2,\alpha/2}$.

On rejette $H_0$ contre $H_1 : m_1 > m_2$ au seuil $(1-\alpha)$ si $T > t_{n_1+n_2-2,\alpha}$.

On rejette $H_0$ contre $H_1 : m_1 < m_2$ au seuil $(1-\alpha)$ si $T < t_{n_1+n_2-2,\alpha}$.

On peut aussi en déduire un intervalle de confiance au niveau $(1-\alpha)$ pour $m_1 - m_2$ :

$$\left[\overline{X}_{1,\bullet} - \overline{X}_{2,\bullet} - t_{n_1+n_2-2,\alpha/2}\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\,,\ \overline{X}_{1,\bullet} - \overline{X}_{2,\bullet} + t_{n_1+n_2-2,\alpha/2}\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right]$$

**Exemple 2.1.1** *On a remarqué que la mobilité de travailleurs américains et japonais était différente dans l'industrie des fabricants de climatiseurs. Les résultats en pourcentages sont les suivants :*

| *Américains* | *Japonais* |
|---|---|
| *7,11* | *3,52* |
| *6,06* | *2,02* |
| *8,00* | *4,91* |
| *6,87* | *3,22* |
| *4,77* | *1,92* |

*La mobilité entre les Américains et les Japonais est-elle égale à 3,1 comme l'a prétendu une étude antérieure ? (faire un test avec $\alpha = 0,1$)*

*On a :*

$$\overline{x}_{1,\bullet} = \frac{32,81}{5} = 6,562 \quad , \quad s_1^2 = \frac{221,2255 - \dfrac{32,81^2}{5}}{5-1} = 1,48157\,,$$

*et*

$$\overline{x}_{2,\bullet} = \frac{15,99}{5} = 3,118 \quad , \quad s2^2 = \frac{54,6337 - \dfrac{15,59^2}{5}}{5-1} = 1,50602\,.$$

*On a aussi :*

$$n_1 = n_2 = 5 \quad , \quad s_p^2 = \frac{s_1^2 + s_2^2}{2} = \frac{1,48157 + 1,50602}{2} = 1,493795\,.$$

*On cherche à tester :*

$$H_0 \,:\, m_1 - m_2 = 3,1 \quad \textit{contre} \quad H_1 \,:\, m_1 - m_2 \neq 3,1\,.$$

*La statistique associée est :*

$$t_{obs} = \frac{\overline{x}_{1,\bullet} - \overline{x}_{2,\bullet} - 3,1}{\sqrt{s_p^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} = \frac{6,562 - 3,118 - 3,1}{\sqrt{1,493795\left(\dfrac{1}{5} + \dfrac{1}{5}\right)}} = 0,445\,.$$

*Ona ici :* $n_1 + n_2 - 2 = 8$ *;* $\alpha/2 = 0,05$*. Une table de loi de Student donne :* $t_{8;0,05} = 1,860$*. On rejette donc si :* $t_{obs} < -1,86$ *ou* $t_{obs} > 1,86$*.*

*Ici,* $H_0$ *n'est pas rejeté.*

*N'oublions pas, dans cet exemple, qu'on a supposé que les lois de la mobilité dans chacun des pays sont gaussiennes, que les variances sont égales, et que les échantillons sont indépendants.*

### Calcul de puissance

Sous l'hypothèse alternative $H_1 \,:\, m_1 \neq m_2$, la statistique suivante suit une loi normale :

$$\frac{\overline{X}_{1,\bullet} - \overline{X}_{2,\bullet}}{\sqrt{S_p^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} \sim \mathcal{N}\left(\frac{m_1 - m_2}{\sqrt{S_p^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}\,,\, 1\right)$$

La statistique $T$ définie en (2.1.8) suit alors une loi de Student $t$ à $n_1 + n_2 - 2$ degrés de liberté avec un paramètre de non centralité égal à :

$$\nu = \frac{m_1 - m_2}{\sqrt{S_p^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$$

C'est cette dernière loi qui est utilisée pour calculer la puissance du test.

**Exemple 2.1.2**

*On a divisé un ensemble de 20 souris en deux groupes de 10 souris. Chacun de ces deux sous-groupes a été soumis à une diète différente.*

*Les données sont les gains de poids des 20 souris après 3 semaines.*

*On veut tester l'hypothèse nulle que le gain de poids est le même dans chacun des deux groupes.*

*Après une transformation logarithmique pour rendre les données normales, on obtient un test non significatif (p-value=0,1877), avec une différence de moyennes* $m_1 - m_2 = -0,32$ *et une variance modifiée : 0,2715.*

*Quelle taille* $n$ *doit avoir chaque groupe (supposés de même effectif) pour que le test d'égalité des moyennes bialtéral au seuil* $5\%$ *ait une puissance de* $90\%$ *?*

*On va estimer la différence des moyennes par* $-0,32$ *la variance modifiée par* $0,2715$*. Pour un* $n$ *quelconque, la paramètre de non centralité* $\nu \simeq -0,44 \times \sqrt{n}$*. La puissance du test est donc :*

$$\gamma(n) = P(T_{2(n-1)}(\nu) < -t_{2(n-1),0.975}) + P(T_{2(n-1)}(\nu) > t_{2(n-1),0.975}),$$

*où* $T_n(\nu)$ *est une variable aléatoire avec une loi t non centrée de paramètre de non centralité* $\nu$ *et de* $n$ *degrés de liberté.*

*Tout ceci est obtenu via les lignes de code R suivantes pour faire des tests et tracer le graphe de la fonction puissance :*

*# Données :*

```
grp1 <- c(4,14,7,9,11,7,13,14,12,8)
grp2<- c(5,21,16,23,4,16,13,19,9,21)
```

*# Calcul de statistiques descriptives :*

```
c(mean(grp1),mean(grp2),var(grp1),var(grp2))
c(mean(log(grp1)),mean(log(grp2)),var(log(grp1)),var(log(grp2)))
```

*# Représentations graphiques :*

```
{\color{blue}boxplot(as.data.frame(cbind(grp1,grp2)))
boxplot(as.data.frame(cbind(log(grp1),log(grp2))))
```

*# Test t avec l'échelle logarithmique :*

```
t.test(log(grp1),log(grp2),var.equal=TRUE)
```

```
# Calcul de l'estimation combinée :

varm<-(var(log(grp1))+var(log(grp2)))/2


# Calcul de la puissance et son graphe :

nu=(mean(log(grp1))-mean(log(grp2)))/sqrt(2*varm)
pui<-rep(0,51)
for(i in (1:51)){
 n<- 9+i
 nu<- nu*sqrt(n)
 pui[i]<- pt(qt(0.025,df=2*(n-1)),df=2*(n-1),ncp=nu)
 + 1-pt(qt(0.975,df=2*(n-1)),df=2*(n-1),ncp=nu)}
n<-10:60
plot(n,pui,type="l",ylab="Puissance",xlab="taille échantillon n",col="blue")
```

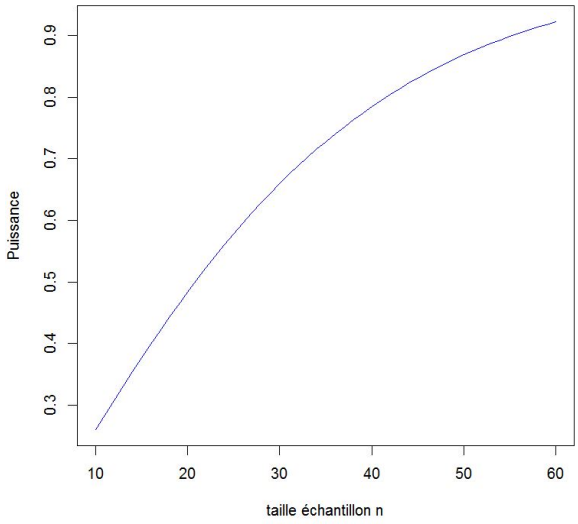


FIGURE 2.1 – Puissance du test en fonction de la taille $n$ de l'échantillon

*Le graphique (figure(2.1)) de la puissance en fonction de la taille $n$ de l'échantillon montre qu'une taille $n = 56$ est nécessaire pour obtenir une puissance de $90\%$ sous les spécifications proposées.*

### 2.1.2 Variances inégales

Lorsque les variances des deux échantillons ne sont pas égales, et si les effectifs de ces deux échantillons sont aussi inégaux, la statistique (2.1.8) définie précédemment doit être

modifiée par l'expression suivante :

$$T' = \frac{\overline{X}_{1,\bullet} - \overline{X}_{2,\bullet}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}\,. \tag{2.1.9}$$

Contrairement au test $t$ de Student précédent, le dénominateur n'est pas basé sur une estimation de la variance.

Le calcul du nombre de degré de liberté $n_1 + n_2 - 2$ doit être remplacé par une valeur approchée, définie comme suit :

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{S_1^4}{n_1^2(n_1-1)} + \frac{S_2^4}{n_2^2(n_2-1)}}$$

Cette variante du test $t$ de Student est appelé test de Welch. Le principe de la méthode de Welch est de tenir compte intégralement du nombre de degrés de liberté de la variance la plus élevée, et de ne faire internevnir que partiellement le nombre de degrés de liberté de la variance la plus petite.

On remarquera aussi que, pour des échantillons de même taille, $t$ et $t'$ sont strictement équivalents.

Comme dans le test $t$ de Student, on rejette $H_0$ contre $H_1 : m_1 \neq m_2$ au seuil $(1-\alpha)$ si $|T'| > t_{1-\alpha/2}$.

On rejette $H_0$ contre $H_1 : m_1 > m_2$ au seuil $(1-\alpha)$ si $T' > t_{1-\alpha}$.

On rejette $H_0$ contre $H_1 : m_1 < m_2$ au seuil $(1-\alpha)$ si $T' < t_{1-\alpha}$.

Un intervalle de confiance au niveau $(1-\alpha)$ pour $m_1 - m_2$ :

$$\left[\left(\overline{X}_{1,\bullet} - \overline{X}_{2,\bullet}\right) - t_{1-\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\,,\, \left(\overline{X}_{1,\bullet} - \overline{X}_{2,\bullet}\right) + t_{1-\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right]$$

**Exemple 2.1.3**

*On s'intéresse à la comparaison de diverses méthodes d'échantillonnage de sols forestiers. Pour cela, on a analysé, d'une part, 20 échantillons de terre prélevés individuellement, et d'autre part, 10 échantillons moyens obtenus chacun en mélangeant 25 échantillons individuels. Tous les prélèvements ont été réalisés au hasard et indépendamment les uns des autres. Les résultats relatifs à la teneur en $K_2O$ exprimée en ppm (parts par million ou, ici, milligrammes de $K_2O$ par kilogramme de terre sèche)*

| *Échantillons individuels* | | *Échantillons moyens* |
|---|---|---|
| *8,0* | *12,8* | *9,6* |
| *8,4* | *14,0* | *10,0* |
| *8,8* | *14,8* | *10,4* |
| *8,8* | *14,8* | *10,4* |
| *9,2* | *14,8* | *10,8* |
| *9,2* | *15,2* | *10,8* |
| *10,0* | *15,6* | *10,8* |
| *10,4* | *18,8* | *11,6* |
| *12,0* | *19,2* | *12,0* |
| *12,4* | *22,0* | *12,8* |

*On peut alors aisément calculer les valeurs suivantes :*

$$\overline{x}_{1,\bullet} = 12,96 \quad et \quad \overline{x}_{2,\bullet} = 10,92$$

$$s_1^2 = 15,9395 \quad et \quad s_2^2 = 0,9262$$

*On remarque que $s_1^2$ et $s_2^2$ sont des estimations respectives des variances pour chaque échantillon, et qu'elles sont visiblement différentes. Le test de Welch s'applique alors :*

$$t' = \frac{\overline{x}_{1,\bullet} - \overline{x}_{2,\bullet}}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} = \frac{12,96 - 10,92}{\sqrt{\dfrac{15,9395}{20} + \dfrac{0,9262}{10}}} = \frac{2,04}{0,9432} = 2,1628\,,$$

*avec un nombre de degrés de liberté égal à :*

$$\nu = \frac{\left(\dfrac{15,9395}{20} + \dfrac{0,926}{10}\right)^2}{\dfrac{15,9395^2}{7600} + \dfrac{0,9262^2}{900}} = \frac{0,79134}{0,03343 + 0,000953} = 23\,.$$

*et $P(|T| \geq 2,16) = 0,041$ au niveau $\alpha = 0,95$ avec 23 degrés de liberté. Au niveau 5%, la différence est juste significative.*

### 2.1.3 Méthodes non paramétriques

Dans le cas d'échantillons indépendants, sans faire l'hypothèse gaussienne, le test classique des rangs ou de la somme des rangs, encore appelé test de Mann-Whitney ou test de Wilcoxon.

L'idée est de classer l'ensemble des observations des deux échantillons par ordre croissant, puis de déterminer les rangs de chacune d'entre elles dans cet ensemble, et enfin de calculer la somme des rangs relative par exemple au premier échantillon, qui sera noté $X_{1,\bullet}$.

Pour $n_1 + n_2 \geq 30$, on peut démontrer que :

$$U = X_{1,\bullet} - \frac{n_1(n_1 + n_2 + 1)}{2} \Big/ \sqrt{n_1 n_2 (n_1 + n_2 + 1)/12}$$

est approximativement, en valeur absolue, une variable gaussienne centrée réduite.

Pour un test bilatéral au niveau $\alpha$, on rejettera l'identité des lois des deux échantillons quand :

$$P(|U| \geq u_{obs}) \leq \alpha \Longleftrightarrow u_{obs} \geq u_{1-\alpha/2}\,.$$

Pour des effectifs plus restreints, $X_{1,\bullet}$ doit être comparé à des valeurs critiques particulières qu'on peut trouver dans des tables statistiques.

Une correction doit être apportée à chaque fois qu'il y a présence d'ex-æquo communs aux deux échantillons. Ces ex-æquo sont alors affectés d'un rang égal à la moyenne des rangs qui reviennent normalement aux différentes valeurs.

D'autres tests existent, comme le test des médianes, qui a pour principe de comparer entre elles les proportions de valeurs observées des deux échantillons qui sont inférieures ou supérieures à la médiane de l'ensemble des observations.

## 2.2 Comparaison des moyennes de deux échantillons non indépendants

### 2.2.1 Principe général

D'une manière générale, les tests relatifs aux échantillons gaussiens non indépendants, ou associés par paires ou couples, sont basées sur le calcul des différences entre les paires ou couples d'observations.

### 2.2.2 Le test $t$ par paires

Il s'agit ici encore de tester la même hypothèse nulle :

$$H_0 \,:\, m_1 = m_2\,.$$

Le test $t$ par paires ou par couples est réalisé en calculant les différences :

$$d_1 = x_{1,1} - x_{2,1}, \cdots, d_n = x_{1,n} - x_{2,n}\,.$$

Notons $SCE_d = \sum_{i=1}^{n}(d_i - \overline{d})^2$ et désignons par $t_{obs}$ la quantité :

$$t_{obs} = \frac{|\overline{d}|}{\sqrt{\dfrac{SCE_d}{n(n-1)}}} = \frac{\overline{x}_{1,\bullet} - \overline{x}_{2,\bullet}}{\sqrt{\dfrac{SCE_d}{n(n-1)}}}$$

L'hypothèse d'égalité des moyennes doit être rejeté pour un test bilatéral de niveau $\alpha$ lorsque :

$$P(|t| \geq t_{obs}) \leq \alpha \quad \text{ou} \quad t_{obs} \geq t_{1-\alpha/2}\,,$$

où $t$ suit une loi de Student à $n-1$ degrés de liberté.

Cette méthode requiert uniquement que les $n$ couples d'observations constituent un échantillon aléatoire simple et que la population des différences soit gaussienne.

**Exemple 2.2.1**

*Dans une étude relative à l'alimentation du mouton, on a comparé deux méthodes d'analyse des matières fécales par spectrométrie. Pour cela, on a examiné 30 échantillons de matières fécales en appliquant sur chacune les deux méthodes d'analyse. Les résultats ci-dessous sont exprimés en teneurs de lutécium observées.*

***Table des teneurs en lutécium par deux méthodes d'analyse***

| *Échant.* | *Méth. 1* | *Méth. 2* | *Diff.* | *Échant.* | *Méth. 1* | *Méth. 2* | *Diff.* |
|---|---|---|---|---|---|---|---|
| *1* | *133* | *129* | *4* | *16* | *153* | *150* | *3* |
| *2* | *131* | *132* | *-1* | *17* | *125* | *123* | *2* |
| *3* | *119* | *121* | *-2* | *18* | *124* | *120* | *4* |
| *4* | *124* | *124* | *0* | *19* | *127* | *125* | *2* |
| *5* | *123* | *124* | *-1* | *20* | *136* | *132* | *4* |
| *6* | *122* | *122* | *0* | *21* | *131* | *130* | *1* |
| *7* | *127* | *131* | *-4* | *22* | *136* | *136* | *0* |
| *8* | *116* | *116* | *0* | *23* | *123* | *120* | *3* |
| *9* | *116* | *118* | *-2* | *24* | *123* | *117* | *6* |
| *10* | *104* | *101* | *3* | *25* | *122* | *118* | *4* |
| *11* | *101* | *104* | *-3* | *26* | *119* | *117* | *2* |
| *12* | *96* | *97* | *-1* | *27* | *126* | *120* | *6* |
| *13* | *96* | *93* | *3* | *28* | *108* | *106* | *2* |
| *14* | *100* | *97* | *3* | *29* | *124* | *122* | *2* |
| *15* | *103* | *99* | *4* | *30* | *137* | *136* | *1* |

*La différence $d_i$ entre les deux séries d'observations figure également dans le tableau ci-dessus. Nous comparons les deux méthodes, il y a lieu d'effectuer un test $t'$ par paires, et non par un test $t$ standard relatif aux échantillons indépendants.*

*Les moyennes utilisées sont :*

$$\overline{x}_{1,\bullet} = 120,83; \quad \overline{x}_{2,\bullet} = 119,33; \quad \overline{d} = 1,50$$

*Le test $t'$ par paires donne les résultats suivants :*

$$t'_{obs} = \frac{1,50}{\sqrt{187,5/(30 \times 29)}} = 3,23 \quad et \quad P(|t'| \geq 3,23) = 0,0031,$$

*avec 29 degrés de liberté. La différence entre les deux méthodes d'analyse s'avère donc hautement significative, bien que les différences observées $d_i$ soient relativement petites.*

*Il était bien entendu hors de question d'appliquer le test classique $t$ des échantillons indépendants car la corrélation entre les deux séries est ici de $\rho = 0,982$.*

### 2.2.3 Méthodes non paramétriques

Le test non paramétrique, dans le cas d'échantillons non indépendants, pour des données continues, est encore ici le test de Wilcoxon, ou test des rangs par paires ou test des rangs et des signes.

Pour cela, il faut calculer les différences entre les couples d'observations, déterminer les rangs de ces différences en valeur absolue, et calculer la somme des rangs relatifs aux différences négatives ou aux différences positives.

Appelons $X_-$ la somme des rangs correspondant aux différences négatives. Pour $n \geq 25$, on peut démontrer que :

$$U = X_- - \frac{n(n-1)}{4} \Big/ \sqrt{n(n+1)(2n+1)/24}$$

soit une loi normale centrée réduite.

Le rejet de l'hypothèse nulle intervient dans les mêmes conditions que pour les échantillons indépendants.

**Exemple 2.2.2**

*Reprenons l'exemple des teneurs en lutécium radioactif. La somme des rangs relatifs aux différences négatives est :*

$$X_- = 64$$

*On en déduit que :*

$$u_{obs} = |64 - (26\,.\,27)/4) \Big/ \sqrt{26\,.\,27\,.\,53/24}| = 2,83$$

*et*

$$P(|U| \geq 2,83) = 0,0047$$

*en considérant que l'effectif est égal à $n = 26$ après élimination des quatre valeurs nulles.*

*La différence de moyennes est ici hautement significative, conclusion identique à celle du test $t'$ par paires.*

## 2.3 Exemple (traitement informatique de la comparaison de deux moyennes)

Sous le logiciel $R$, les étapes seront les suivantes :

1. Importation des données ;
2. Comparaison graphique des deux sous-populations ;
3. Estimation des statistiques de base (moyenne, écart-type, quantiles) par sous-population ;
4. Test de la normalité des données dans chaque population ;

5. Test de l'égalité des variances ;
6. Test de l'égalité des moyennes.

Nous allons comparer les poids, exprimés en grammes, de poulpes mâles et femelles pêchés au large des côtes mauritaniennes. On a ici 15 poulpes mâles et 13 poulpes femelles. Le fichier se nomme "poulpe.cvs".

### 2.3.1 Importation des données

La ligne de commande d'importation est la suivante :

```
{\color{blue}don<-read.table("poulpe.csv",header=TRUE,sep=";")}
```

Le résumé du jeu de données s'obtient via la commande :

```
summary(don)}
```

On obtient des statistiques de base :

```
     Poids           Sexe
 Min.   : 300   Femelle:13
 1st Qu.:1480   Mâle   :15
 Median :1800
 Mean   :2099
 3rd Qu.:2750
 Max.   :5400
```

On voit que la variable *Poids* est bien quantitative et que la variable *Sexe* est bien qualitative.

### 2.3.2 Comparaison graphique des deux sous-populations

Au départ de toute analyse de ce type, il est toujours intéressant de visualiser les données. Les boîtes à moustaches permettent de comparer la distribution des *Poids* dans chaque modalité de la variable *Sexe* :

La ligne de commande pour obtenir la boîte à moustaches est :

```
boxplot(Poids~Sexe,ylab="Poids",xlab="Sexe",data=don, col="lightblue")
```

La figure (2.2) montre que les mâles sont en général plus lourds que les femelles puisque médianes et quartiles de poids sont supérieurs chez les mâles.

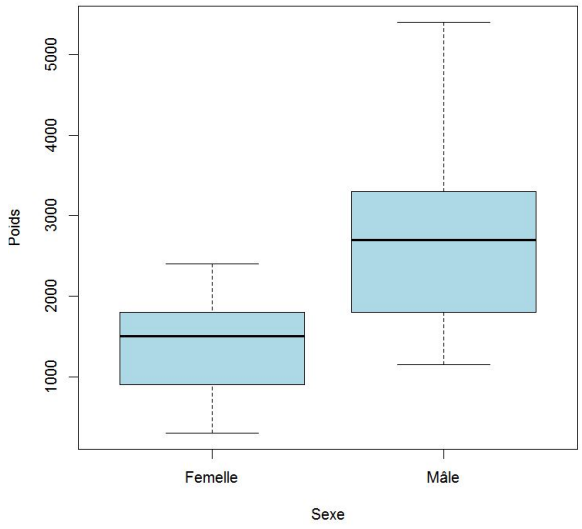


FIGURE 2.2 – Boîtes à moustaches pour les deux sous-populations

### 2.3.3 Estimation des statistiques de base par sous-population

On estime la moyenne, l'écart-type et les quartiles par *Sexe*.

On remarque ici que l'argument *na.rm=TRUE* est en fait inutile ici, puisqu'il n'y a pas de données manquantes :

```
tapply(don[,"Poids"],don[,"Sexe"],mean,na.rm=TRUE)
```

```
 Femelle     Mâle
1405.385   2700.000
```

```
{\color{blue}tapply(don[,"Poids"],don[,"Sexe"],sd,na.rm=TRUE)
```

```
 Femelle      Mâle
 621.9943   1158.3547
```

```
tapply(don[,"Poids"],don[,"Sexe"],quantile,na.rm=TRUE)
```

```
$Femelle
  0%  25%  50%  75% 100%
 300  900 1500 1800 2400

$Mâle
  0%  25%  50%  75% 100%
1150 1800 2700 3300 5400
```

### 2.3.4 Test de la normalité des données dans chaque population

Pour construire le test de comparaison de moyennes, on fait souvent l'hypothèse que l'estimateur de la moyenne, dans chaque sous-population, suit une loi normale. Cela est vrai si la distribution des données suit une loi normale, ou si la taille de l'échantillon est suffisamment grande (en pratique $n > 30$) grâce au théorème central limite.

Ici les effectifs sont inférieurs à 30. Il faut donc tester la normalité des données dans chaque sous-population.

On peut utiliser le test de Shapiro-Wilk. Pour tester la normalité des mâles seuls, on sélectionne les poids des mâles en imposant que la variable *Sexe* prenne la modalité " Mâle ". On effectue une sélection des lignes en construisant le vecteur logique *select. males*. Les composantes de ce vecteur sont *TRUE* pour un mâle et *FALSE* sinon. On construit le test de Shapiro-Wilk sur les individus de cette sélection :

```
select.males<-don[,"Sexe"]=="Mâle"}
shapiro.test(don[select.males,"Poids"])}
```

```
Shapiro-Wilk normality test

data:  don[select.males, "Poids"]
W = 0.935, p-value = 0.3238
```

La probabilité critique étant supérieure à 5%, on accepte la normalité des poids des mâles.

Pour les femelles, la conclusion est identique.

Quand l'hypothèse de normalité est rejetée, le test d'égalité des moyennes peut être réalisé avec le test de Wilcoxon (*Wilcox.test*) ou celui de Kruskal et Wallis (*kruskal.test*).

### 2.3.5 Test de l'égalité des variances

Pour comparer les moyennes des deux sous-populations mâle et femelle, il existe deux types de test : un quand les variances sont inconnues, et l'autre lorsqu'elles sont connues.

Nous devons donc avant tout tester l'égalité des variances :

$$H_0 \,:\, \sigma_1^2 = \sigma_2^2 \quad \text{contre} \quad H_1 \,:\, \sigma_1^2 \neq \sigma_2^2 \,.$$

La ligne de commandes est la suivante :

```
var.test(Poids~Sexe,conf.level=0.95,data=don)}
```

```
        F test to compare two variances

data:  Poids by Sexe
F = 0.2883, num df = 12, denom df = 14, p-value = 0.03713
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.09452959 0.92444666
sample estimates:
ratio of variances
         0.2883299
```

La probabilité critique vaut $0,037$. On rejette donc $H_0$, et on peut considérer les variances significativement différentes.

### 2.3.6 Test de l'égalité des moyennes

On utilise pour cela la commande *t.test*. Comme les variances sont différentes, c'est le test de Welch qui est ici utilisé. On précise que les variances sont égales grâce à l'argument *var.equal=FALSE*.

Si les variances avaient été égales, on aurait utilisé un test de Student avec l'argument *var.equal=TRUE*.

Le test construit par défaut est bilatéral (*alternative='two.sided'*), mais l'hypothèse alternative $H_1$ pourrait être que les mâles sont plus légers (*alternative='less'*) ou plus lourds (*alternative='greater'*).

On considère ici que la modalité de référence est *Femelle* (première modalité de la variable *Sexe*) et on teste l'autre modalité par rapport à celle-ci :

```
t.test(Poids~Sexe,alternative='two.sided',conf.level=0.95,var.equal=FALSE,data=don)
```

```
        Welch Two Sample t-test

data:  Poids by Sexe
t = -3.7496, df = 22.021, p-value = 0.001107
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2010.624  -578.607
sample estimates:
mean in group Femelle    mean in group Mâle
             1405.385              2700.000
```

La probabilité critique $0,001$ indique que les moyennes sont très significativement différentes.

# Exercices

**Exercice 1**

Deux échantillons indépendants ont été sélectionnés, 130 d'une population 1 et 170 d'une population 2. Les moyennes calculées sont respectivement : $\overline{x}_{1,\bullet} = 534$ et $\overline{x}_{2,\bullet} = 615$. Les écarts-types correspondants sont connus et valent respectivement : $\sigma_1 = 25$ et $\sigma_2 = 30$..

1. On suppose dans cette question que : $m_1 - m_2 = -70$. Que dire de la loi de $\overline{X}_{1,\bullet} - \overline{X}_{2,\bullet}$ ?
2. Testez $H_0 : m_1 - m_2 = -70$ contre $H_1 : m_1 - m_2 < -70$ au niveau $\alpha = 0.01$.Interprétez le résultat de votre test.

**Exercice 2**

Dans une université, on a mené une étude afin de comparer le nombre moyen d'heures d'étude passées chaque semaine par des étudiants, en mettant l'accent entre les étudiants athlètes et les étudiants qui ne le sont pas. Pour cela, un échantillon de 55 étudiants athlètes a donné une moyenne hebdomadaire d'étude de 20,6 heures et un écart-type de 5,3 heures. Un second échantillon de 200 non athlètes a, quant à lui, donné une moyenne hebdomadaire d'étude de 23,5 heures et un écart-type de 4,1 heures.

1. À partir des échantillons observés, peut-on affirmer qu'il y a une différence significative entre les nombres d'heures d'étude hebdomadaire efectués par les athlètes et ceux qui ne le sont pas ? (Faites un test au niveau $\alpha = 0.01$)
2. Construisez un intervalle de confiance à 99% pour $m_1 - m_2$.
3. Est-ce qu'un intervalle de confiance à 95% sera plus étroit ou plus large qu'un intervalle de confiance à 99% ?

**Exercice 3**

Un nouveau type de broches a été conçu par un laboratoire dentaire pour les enfants qui doivent porter des appareils. Les nouvelles broches sont censées être plus confortables, d'un aspect esthétique amélioré, et censées accélérer le processus de réalignement des dents.

Une expérience a été menée pour comparer les temps nécessaires au bon réalignement des dents avec les anciennes broches et avec les nouvelles. Une centaine d'enfants a été choisie au hasard, 50 dans chaque groupe. Un résumé des résultats obtenus est fourni dans le tableau suivant :

| | **Vieilles broches** | **Nouvelles broches** |
|---|---|---|
| $\bar{x}$ | 410 jours | 380 jours |
| $s$ | 45 jours | 60 jours |

1. Peut-on conclure, au niveau $\alpha = 0,01$, que les vieilles broches doivent être significativement portées plus longtemps que les nouvelles ?
2. Trouvez un intervalle de confiance à 95% pour la différence des temps où les deux types de broches ont été portés.

**Exercice 4**

Une chaîne de grands supermarchés s'intéresse à savoir s'il existe une différence de durée de vie, en jours, entre deux marques de pain : $A$ et $B$. Deux échantillons de 50 pains de chaque marque fraîchement cuits ont donné les résultats suivants :

| **Marque** $A$ | **Marque** $B$ |
|---|---|
| $\bar{x}_1$=4.1 | $\bar{x}_2$=5,2 |
| $s_1$=1,2 | $s_2$=1,4 |

1. Déterminez les hypothèses $H_0$ et $H_1$ nécessaires pour étudier le problème posé.
2. Le test a été mené en utilisant un logiciel statistique. Interprétez les résultats ci-dessous :

$$z = -4,218 \qquad P - value = 0,0000$$

3. Pensez-vous que la vraie probabilité critique ci-dessus est vraiment nulle ?
4. Si $m_1$ et $m_2$ sont les durées de vie des pains de marque respectivement $A$ et $B$, construisez un intervalle de confiance à 90 % pour $m_1 - m_2$.

**Exercice 5**

Un fabricant d'amortisseurs d'automobiles s'intéresse à la comparaison de ses amortisseurs vis-à-vis de ceux de son plus grand concurrent. Pour cela, chaque fabricant choisit 6 voitures au hasard et des amortisseurs sont montés sur les voitures concernées. Après que les automobiles aient parcouru 30 000 kms, la résistance aux chocs des amortisseurs a été mesurée, codée et enregistrée. Les résultats sont les suivants :

| **Voiture n°** | **Fabricant** | **Concurrent** |
|---|---|---|
| 1 | 8,8 | 8,4 |
| 2 | 10,5 | 10,1 |
| 3 | 12,5 | 12,0 |
| 4 | 9,7 | 9,3 |
| 5 | 9,6 | 9,0 |
| 6 | 13,2 | 13,0 |

1. Peut-on conclure à une différence de résistance pour les amortisseurs entre les deux fabricants après 30 000 kms d'utilisation ? ($\alpha = 0,05$)
2. Quelles hypothèses ont été nécessaires pour cette comparaison ?
3. Construire un intervalle de confiance à 95 % de la différence $m_1 - m_2$. Interprétez.

# Chapitre 3

# Analyse de la variance à un facteur

## 3.1 Introduction

### 3.1.1 Exemple introductif

Un exemple de reproductibilité pour étudier les performances de trois laboratoires relativement à la détermination de la quantité de sodium de lasalocide dans de la nourriture pour de la volaille.

Une portion de nourriture contenant la dose nominale de 85 mg/kg de sodium de lasalocide a été envoyée à chacun des laboratoires à qui il a été demandé de procéder à 10 réplications de l'analyse.

Les mesures de sodium de lasalocide obtenues sont exprimées en mg/kg. Elles sont reproduites dans le tableau suivant :

| | **Lab. A** | **Lab. B** | **Lab. C** |
|---|---|---|---|
| 1 | 87 | 88 | 85 |
| 2 | 88 | 93 | 84 |
| 3 | 84 | 88 | 79 |
| 4 | 84 | 89 | 86 |
| 5 | 87 | 85 | 81 |
| 6 | 81 | 87 | 86 |
| 7 | 86 | 86 | 88 |
| 8 | 84 | 89 | 83 |
| 9 | 88 | 88 | 83 |
| 10 | 86 | 93 | 83 |

Cette écriture du tableau est dite *désempilée.*, Nous pouvons l'écrire sous forme standard (*empilée*), c'est-à-dire avec deux colonnes, une pour la laboratoire et une pour la valeur de la teneur en sodium de lasalocide mesurée, et trente lignes pour chacune des observations réalisées.

| Essai | Laboratoire | Lasalocide |
|---|---|---|
| 1 | Laboratoire A | 87 |
| 2 | Laboratoire A | 88 |
| 3 | Laboratoire A | 84 |
| 4 | Laboratoire A | 84 |
| 5 | Laboratoire A | 87 |
| 6 | Laboratoire A | 81 |
| 7 | Laboratoire A | 86 |
| 8 | Laboratoire A | 84 |
| 9 | Laboratoire A | 88 |
| 10 | Laboratoire A | 86 |

| Essai | Laboratoire | Lasalocide |
|---|---|---|
| 11 | Laboratoire B | 88 |
| 12 | Laboratoire B | 93 |
| 13 | Laboratoire B | 88 |
| 14 | Laboratoire B | 89 |
| 15 | Laboratoire B | 85 |
| 16 | Laboratoire B | 87 |
| 17 | Laboratoire B | 86 |
| 18 | Laboratoire B | 89 |
| 19 | Laboratoire B | 88 |
| 20 | Laboratoire B | 93 |

| Essai | Laboratoire | Lasalocide |
|---|---|---|
| 21 | Laboratoire C | 85 |
| 22 | Laboratoire C | 84 |
| 23 | Laboratoire C | 79 |
| 24 | Laboratoire C | 86 |
| 25 | Laboratoire C | 81 |
| 26 | Laboratoire C | 86 |
| 27 | Laboratoire C | 88 |
| 28 | Laboratoire C | 83 |
| 29 | Laboratoire C | 83 |
| 30 | Laboratoire C | 83 |

**Remarque 3.1.1** *Dans la plupart des logiciels, c'est sous cette dernière forme que sont saisies et traitées les données. Dans les deux tableaux, nous avons omis les unités de la mesure réalisée pour abréger l'écriture. Mais en principe, cela doit être indiqué entre parenthèses à côté de la mesure.*

**Remarque 3.1.2** *Il va de soi que, lorsque vous rentrez les données dans un logiciel, vous n'indiquerez pas le mot "Laboratoire" à côté des lettres* $(A, B, C)$*. Il est juste là pour vous faciliter la compréhension du tableau.*

**Définition 3.1.1** *Sur* **chaque essai**, *on observe* **deux variables**.

1. *Le laboratoire. Il est totalement contrôlé. La variable "Laboratoire" est considérée comme qualitative avec trois modalités bien déterminées : A, B, et C. Nous l'appelons le* **facteur**. *Ici, le facteur "Laboratoire" est à* **effets fixes**.

2. *La quantité de sodium de lasalocide. La variable "Lasalocide" est considérée comme quantitative comme généralement tous les résultats obtenus par une mesure. Nous l'appelons* **la variable réponse**.

La variable mesurée dans un tel schéma expérimental sera notée $Y$.

Pour les observations, nous utilisons deux indices :

1. le premier indice indique le numéro du groupe dans la population ("Laboratoire") ;
2. le second indice indique le numéro de l'observation dans l'échantillon ("Essai").

**Pour le premier indice**, nous utiliserons en général l'indice $i$.

**Pour le second indice**, nous utiliserons en général l'indice $j$.

Ainsi, les observations seront notées en général :

$$y_{i,j} \qquad i = 1, \cdots, I \quad ; \quad j = 1, \cdots, J(I).$$

**Définition 3.1.2** *Lorsque les échantillons sont de même taille, à savoir* $J(i) = I$ *et ce, quel que soit* $i$*, nous disons alors que l'expérience est* **équilibrée**.

**Remarque 3.1.3** *Si les tailles des échantillons sont différentes, alors elles sont notées par :*

$$n_i \qquad où \quad i = 1, \cdots, I\,.$$

*Mais ce plan expérimental est à éviter, si possible, parce que les différences qu'il est alors possible de détecter, sont supérieures à celles du schéma équilibré.*

**Définition 3.1.3** *En se plaçant dans le* **cas équilibré**, *nous notons les moyennes de chaque échantillon par :*

$$\bar{y}_{i,\bullet} = \frac{1}{J}\sum_{j=1}^{J} y_{i,j} \qquad i = 1, \cdots, I\,,$$

*et les variances de chaque échantillon par :*

$$s_i^2(y) = \frac{1}{J}\sum_{j=1}^{J} (y_{i,j} - \bar{y}_{i,\bullet})^2 \qquad i = 1, \cdots, I\,.$$

**Remarque 3.1.4** *Cette dernière formule exprime la variance non corrigée. Très souvent, dans les ouvrages ou logiciels, c'est la variance corrigée qui est utilisée : au lieu d'être divisée par $J$, la somme est divisée par $J-1$.*

Après calculs avec le logiciel ***R***, nous avons :

$$\bar{y}_{1,\bullet} = 85,5 \qquad \bar{y}_{2,\bullet} = 88,6 \qquad \bar{y}_{3,\bullet} = 83,8$$

et

$$s_{1,c} = 2,224 \qquad s_{2,c} = 2,633 \qquad s_{3,c} = 2,616\,.$$

Le nombre total d'observations est égal à :

$$n = I \times J = 3 \times 10 = 30\,.$$

On aimerait savoir, et c'est l'objet de l'analyse de variance, s'il y a une différence pour les teneurs en sodium de lasalocide entre les trois laboratoires.

Y a-t-il un effet *laboratoire* sur les teneurs en sodium de lasalocide ?

Voilà la question qui nous intéresse dans ce chapitre, et qui est l'objet principal de l'analyse de la variance à un facteur.

### 3.1.2 Objectifs

L'analyse de la variance (ANOVA) est une méthode statistique qui permet d'étudier la modification de la moyenne $\mu$ d'une quantité $Y$ (variable réponse quantitative) selon l'influence éventuelle d'un ou de plusieurs facteurs d'expérience qualitatifs (traitements ... ). Dans le cas où la moyenne n'est influencée que par un seul facteur (noté facteur $A$) , il s'agit d'une analyse de la variance à un seul facteur ("one way ANOVA"), objet de ce chapitre. Un facteur $A$ est souvent une variable qualitative présentant un nombre restreint de modalités. Le nombre de modalités (c'est-à-dire de niveaux) du facteur $A$ sera noté $I$. On suppose que $Y$ suit une loi normale $\mathcal{N}(\mu_i, \sigma^2)$ sur chaque sous-population $i$ définie par les modalités de $A$. L'objectif est ici de tester l'égalité des moyennes de ces populations, à savoir de tester l'hypothèse nulle :

$$H_0 \,:\, \mu_1 = \mu_2 = \cdots = \mu_I\,.$$

contre l'hypothèse alternative :

$$H_1 \,:\, \exists\, i_0 \neq i' \text{ tel que } \mu_{i_0} \neq \mu_{i'} \quad \text{(il existe au moins deux moyennes différentes)}\,.$$

Pour chaque population $i$ (ou modalité $i$ du facteur $A$), on dispose d'un échantillon $\boldsymbol{y}$ de $n_i$ observations de la variable réponse $Y$ :

$$y_{i,1},\, y_{i,2}, \cdots, y_{i,n_i}\,.$$

 

Le modèle s'écrit alors :

$$Y_{i,j} = \mu_i + \varepsilon_{i,j} \quad , \quad i = 1, \cdots, I \, , \, j = 1, \cdots, n_i \, ,$$

où les erreurs $\varepsilon_{i,j}$ sont des variables aléatoires indépendantes de loi normale $\mathcal{N}(0, \sigma^2)$.

On peut également écrire : $\mu_i = \mu + \alpha_i$ pour tout $i = 1, \cdots, I$. Sous cette forme, $\mu$ est alors appelé effet moyen du facteur $A$, et $\alpha_i$ est appelé effet différentiel de la modalité (ou du niveau) $i$ du facteur $A$. Le modèle précédent peut alors s'écrire comme suit :

$$Y_{i,j} = \mu + \alpha_i + \varepsilon_{i,j} \quad , \quad i = 1, \cdots, I \, , \, j = 1, \cdots, n_i \, ,$$

Ce modèle n'est pas identifiable et il est alors nécessaire d'appliquer une contrainte (linéaire) pour le rendre identifiable. Par défaut, le logiciel ***R*** propose d'imposer $\alpha_1 = 0$. Les comparaisons des moyennes sont alors faites par rapport à la moyenne $\mu_1$. La classe de référence est donc le niveau 1 du facteur $A$. D'autres contraintes sont envisageables, comme la contrainte : $\sum_{i=1}^{I} \alpha_i = 0$, qui correspond alors à prendre l'effet moyen $\mu$ comme référence.

### 3.1.3 Un peu de vocabulaire

1. On appelle réponse une variable dont nous cherchons à comprendre le comportement. Dans ce cours, nous examinerons le cas des réponses quantitatives continues. Une réponse sera généralement notée $Y$.
2. On appelle facteur toute variable que nous voulons utiliser pour analyser les variations de la variable réponse. On notera souvent les différents facteurs par $A, B, C, \cdots$. Dans le contexte de l'analyse de variance, les facteurs seront considérés comme des variables qualitatives. Les niveaux ou modalités de chaque facteur seront notés en indexant la variable correspondante : $A_1, \cdots, A_I$ si le facteur $A$ possède $I$ modalités.
3. Un modèle statistique est une équation reliant les différentes mesures $Y_{i,j,\cdots}$ de la variable réponse $Y$ aux effets des niveaux $A_1, \cdots, A_I, B_1, \cdots, B_J, \cdots$ des facteurs $A, B, \cdots$ pour lesquels ont été réalisées ces différentes mesures via une relation fonctionnelle $f$, et en modélisant les fluctuations expérimentales à l'aide de variables aléatoires d'erreurs notées généralement $\varepsilon_{i,j,\cdots}$ sur lesquelles on fait un certain nombre d'hypothèses :

$$Y_{i,j,\cdots} = f(A_i, B_j, \cdots) + \varepsilon_{i,j,\cdots}$$

4. Les modèles associés à l'analyse de variance sont des modèles linéaires. Donc la relation fonctionnelle $f$ sera tout simplement une somme de termes. Par exemple, si on souhaite étudier une réponse continue $Y$ à l'aide d'un facteur qualitatif $A$ à effets fixes, avec un nombre identique $J$ de répétitions effectuées pour chacun des niveaux du facteur. $\mu$ est l'effet moyen sur toute la population du facteur $A$.

   On introduit alors le modèle :

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

sous la contrainte $\sum_{i=1}^{I} \alpha_i = 0$, où $Y_{i,j}$ est la valeur prise par la variable réponse $Y$ dans la condition $A_i$ lors de la $j$-ème répétition.

Les hypothèses classiques pour les erreurs sont :

(a) $\varepsilon_{i,j}$ et $\varepsilon_{k,l}$ sont indépendantes si $(i,j) \neq (k,l)$ avec $1 \leq i,k \leq I$ et $1 \leq j,l \leq J$.

(b) $\mathcal{L}(\varepsilon_{i,j}) = \mathcal{N}(0,\sigma^2)$.

Les $\mu, \alpha_1, \cdots, \alpha_I$ sont les paramètres du modèle. $\varepsilon_{i,j}$ est parfois appelé terme d'erreur du modèle. Les hypothèses faites sur les erreurs font partie intégrante de la définition du modèle. Il faudra donc toujours examiner soigneusement ces conditions, car la validité des résultats dépend fortement des conditions d'application.

### 3.1.4 Plans équilibrés - Plans déséquilibrés

Un plan équilibré est un dispositif expérimental comportant au moins un facteur et ayant un nombre identique de répétitions dans chacune des modalités des facteurs.

Nous conseillons donc, lors de la phase de planification expérimentale, de prévoir d'utiliser des plans équilibrés pour l'analyse des effets ou interactions d'un ou plusieurs facteurs qualitatifs sur une réponse expérimentale modélisée par une variable continue. Ainsi, dans le cas d'un plan équilibré :

1. Les estimations des paramètres d'un modèle où les facteurs sont à effets fixes seront aisées à calculer, car on se trouve alors dans une situation dénommée en statistique : plan d'expérience orthogonal. Cela signifie que, pour estimer l'effet d'un des termes du modèle, il n'est pas nécessaire d'estimer les effets des autres termes de ce modèle.
2. Lorsqu'on ne dispose d'aucune information a priori sur les résultats possibles de l'expérience, la situation la plus commune consiste à tenter de mettre en évidence une différence, si elle existe, entre les différents effets.

Il est toujours possible d'utiliser tous les modèles exposés dans ce qui suit dans le contexte particulier des plans équilibrés lorsqu'en fait le plan est déséquilibré, c'est-à-dire lorsque le nombre de répétitions varie pour chaque modalité des facteurs.

En aucun cas, il n'est légitime de supprimer des observations pour se ramener à une situation de plan équilibré.

Deux faits doivent toujours être à l'esprit lorsqu'on se sert de ce type de modèles avec un plan déséquilibré :

1. L'effet des estimations des différents niveaux des facteurs n'est plus aussi simple que précédemment, excepté le cas d'un modèle d'analyse de variance à un facteur quand le plan est dit à effectifs proportionnels. En effet, ces estimations ne sont plus indépendantes.

2. De plus, les lois statistiques des tests que l'on utilise pour tester les différentes hypothèses présentes dans le tableau d'analyse de variance ne sont, dans la plupart des cas, connues qu'approximativement.

### 3.1.5 Quelques remarques liminaires

De manière générale, l'analyse de variance (en anglais *analysis of variance*, ANOVA) a comme objectif de comparer des ensembles de plus de deux moyennes, en tentant d'identifier les sources de variation qui peuvent expliquer les différences entre elles.

Pour l'approche inférentielle, l'analyse de variance s'applique dans les mêmes conditions que le test $t$ de Student, à savoir des populations normales, de même variance, et des échantillons aléatoires simples et indépendants.

On peut également faire les mêmes remarques que pour le test de Student : l'analyse de variance est en effet assez peu sensible à la non-normalité des populations parents et, pour des échantillons de mêmes effectifs, à l'inégalité des variances.

Une réserve cependant est à formuler. Si l'analyse de variance est peu sensible à une éventuelle inégalité des variances dans le cas d'échantillons de mêmes effectifs, il n'en est pas de même pour le cas d'échantillons n'ayant pas les mêmes effectifs.

On considérera deux types de modèles :

1. Le modèle à effets fixes, qui est le plus classique, a pour objet la comparaison d'un nombre limité $I$ de populations, pour chacune desquelles peut être prélevé un échantillon.
2. Le modèle à effets aléatoires a trait, à l'inverse, à la comparaison d'une infinité ou d'un très grand nombre de populations, pour toutes lesquelles il n'est pas possible, en pratique, de prélever chaque fois un échantillon. Dans ce cas, l'échantillonnage est un échantillonnage à deux degrés. D'abord, on choisit de façon complètement aléatoire un nombre réduit $I$ de populations, et ensuite, on choisit un échantillon à l'intérieur de chacune de ces seules $I$ populations. Bien entendu, on doit être attentif au fait que ces $I$ populations prises en considération ne sont que l'image de l'ensemble plus vaste de toutes les populations au sujet desquelles on s'interroge.

## 3.2 Modèle à effets fixes

Supposons qu'un facteur contrôlé $A$ possède $I$ modalités, chacune d'entre elles étant notée $A_i$. Pour chacune des modalités, on effectue $J$ ($J \geq 2$) mesures d'une variable réponse $Y$ supposée continue. Le nombre total de mesures est donc $I \times J$.

On introduit le modèle :

$$Y_{i,j} = \mu + \alpha_i + \varepsilon_{i,j} \qquad i = 1, \cdots, I\,;\; j = 1, \cdots, J\,, \tag{3.2.1}$$

 

sous la contrainte que : $\sum_{i=1}^{I} \alpha_i = 0$, où $Y_{i,j}$ est la valeur prise par la variable réponse $Y$ dans la condition $A_i$ lors de la $j$-ème répétition, et où $\varepsilon_{i,j}$ est le résidu du modèle. Un individu statistique est donc défini par le couple $(i, j)$. L'analyse de variance revient alors à tester l'égalité des moyennes dans chaque modalité, c'est-à-dire tester l'égalité des $\alpha_i$ à zéro.

On fera les hypothèses importantes suivantes :

1. $\varepsilon_{i,j}$ et $\varepsilon_{k,l}$ sont indépendantes si $(i, j) \neq (k, l)$ avec $1 \leq i, k \leq I$ et $1 \leq j, l \leq J$.
2. $\mathcal{L}(\varepsilon_{i,j}) = \mathcal{N}(0, \sigma^2)$.

Nous supposerons que ces conditions d'utilisation sont bien remplies. Nous regroupons les valeurs que peut prendre la réponse $Y$ dans les conditions $A_i$ lors des $J$ répétitions dans le tableau suivant :

| **Facteur $A$** | $Y$ |
|---|---|
| $A_1$ | $Y_{1,1}, \cdots, Y_{1,J}$ |
| $\vdots$ | $\vdots$ |
| $A_i$ | $Y_{i,1}, \cdots, Y_{i,J}$ |
| $\vdots$ | $\vdots$ |
| $A_I$ | $Y_{I,1}, \cdots, Y_{I,J}$ |

On notera également $\mu_i = \mu + \alpha_i$. Clairement, on a : $Y_{i,j} \sim \mathcal{N}(\mu_i, \sigma^2)$ pour tout $i = 1, \cdots, I$ et $j = 1, \cdots, J$.

### 3.2.1 Notations

On a observé $n = I \times J$ valeurs de la variable $Y$ indexée par deux indices $i$ et $j$. La moyenne de ces valeurs par rapport à l'indice $i$ est notée $Y_{\bullet,j}$. Il s'agit simplement de la moyenne de valeurs de la $j$-ème colonne du tableau :

$$Y_{\bullet,j} = \frac{1}{I} \sum_{i=1}^{I} Y_{i,j} \, .$$

La moyenne de ces valeurs par rapport à l'indice $j$ est notée $Y_{i,\bullet}$. Il s'agit simplement de la moyenne de valeurs de la $i$-ème ligne du tableau :

$$Y_{i,\bullet} = \frac{1}{J} \sum_{j=1}^{J} Y_{i,j} \, .$$

La moyenne globale par rapport aux indices $i$ et $j$ est notée $Y_{\bullet,\bullet}$. Il s'agit simplement de la moyenne globale du tableau :

$$Y_{\bullet,\bullet} = \frac{1}{I\,J} \sum_{i=1}^{I} \sum_{j=1}^{J} Y_{i,j} \, .$$

 

Là aussi, il est aisé d'avoir les lois des éléments ci-dessus. Par exemple : $Y_{i,\bullet} \sim \mathcal{N}(\mu_i, \sigma^2/J)$ et $Y_{\bullet,\bullet} \sim \mathcal{N}(\mu, \sigma^2/n)$ où $n = I \times J$.

### 3.2.2 Équation de l'analyse de variance et tests

Alors, la variation totale théorique, ou somme totale des carrés des écarts est égale à :

$$SC_{TOT} = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{i,j} - Y_{\bullet,\bullet})^2 \,. \tag{3.2.2}$$

On appelle variation théorique due au facteur $A$, la quantité :

$$SC_F = J\sum_{i=1}^{I}(Y_{i,\bullet} - Y_{\bullet,\bullet})^2 \,, \tag{3.2.3}$$

et variation résiduelle, la quantité :

$$SC_R = \sum_{i=1}^{I}\left(\sum_{j=1}^{J}(Y_{i,j} - Y_{i,\bullet})^2\right) \tag{3.2.4}$$

**Proposition 3.2.1**
*On a la décomposition fondamentale de l'analyse de variance suivante :*

$$SC_{TOT} = SC_F + SC_R \,. \tag{3.2.5}$$

Remarquons que : $Y_{i,j} - Y_{\bullet,\bullet} = Y_{i,j} - Y_{i,\bullet} + Y_{i,\bullet} - Y_{\bullet,\bullet}$. D'où :

$$\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{i,j} - Y_{\bullet,\bullet})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{i,j} - Y_{i,\bullet})^2 + \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{i,\bullet} - Y_{\bullet,\bullet})^2 \,, \tag{3.2.6}$$

car le double produit, dans le développement de la somme au carré, vaut zéro. En effet :

$$2\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{i,j} - Y_{i,\bullet})(Y_{i,\bullet} - Y_{\bullet,\bullet}) = 2\sum_{i=1}^{I}(Y_{i,\bullet} - Y_{\bullet,\bullet})\left[\sum_{j=1}^{J}(Y_{i,j} - Y_{i,\bullet})\right] \,.$$

D'autre part, on remarque que : $\displaystyle\sum_{j=1}^{J} Y_{i,j} = JY_{i,\bullet}$ par définition de $Y_{i,\bullet}$. D'où :

$$\sum_{j=1}^{J}(Y_{i,j} - Y_{i,\bullet}) = JY_{i,\bullet} - JY_{i,\bullet} = 0 \,.$$

L'égalité (3.2.6), en utilisant (3.2.2), (3.2.3) et (3.2.4) peut alors s'écrire sous la forme de la relation fondamentale (3.2.5) de l'ANOVA.

**Proposition 3.2.2**
*Sous les hypothèses de normalité et d'égalité des variances, on a :*

$$\frac{SC_F}{\sigma^2} \sim \chi_{I-1}(\delta)$$

*avec* $\delta = J\sum_{i=1}^{I}(\mu_i - \mu)^2 = J\sum_{i=1}^{I}\alpha_i^2$.

Il est facile de voir que $SC_F = \sum_{i=1}^{I} JY_{i,\bullet}{}^2 - nY_{\bullet,\bullet}{}^2$. Posons $Z_i = \sqrt{J}Y_{i,\bullet}$ pour $i = 1, 2, \cdots, I$. On a alors :

$$Z_i \sim \mathcal{N}(\sqrt{J}\mu_i, \sigma^2) \text{ et } Y_{\bullet,\bullet} = \frac{1}{n}\sum_{i=1}^{I}\sqrt{J}Z_i\,.$$

On en déduit que :

$$\begin{aligned} SC_F &= \sum_{i=1}^{I} Z_i^2 - (\frac{1}{\sqrt{n}}\sum_{i=1}^{I}\sqrt{J}Z_i)^2 \\ &= \sum_{i=1}^{I}(1-\frac{J}{n})Z_i^2 - 2\sum_{i<j}\frac{\sqrt{J}\sqrt{J}}{n}Z_iZ_j \end{aligned}$$

$SC_F$ s'exprime alors comme une forme quadratique en $\mathbf{Z} = (Z_1, Z_2, \cdots, Z_I)^T$.

La matrice $A$ associée à cette forme quadratique s'exprime sous la forme $\mathbf{A} = \mathbf{I_I} - \dfrac{\mathbf{1}}{\mathbf{n}}\nu\nu^{\mathbf{T}}$ où $\nu = (\sqrt{\mathbf{J}}, \sqrt{\mathbf{J}}, \cdots, \sqrt{\mathbf{J}})^{\mathbf{T}}$.

Calculons $\mathbf{A^2}$. On a ici :

$$\begin{aligned} \mathbf{A^2} &= (\mathbf{I_I} - \frac{\mathbf{1}}{\mathbf{n}}\nu\nu^{\mathbf{T}}) \times (\mathbf{I_I} - \frac{\mathbf{1}}{\mathbf{n}}\nu\nu^{\mathbf{T}}) \\ &= \mathbf{I_I} - \frac{\mathbf{1}}{\mathbf{n}}\nu\nu^{\mathbf{T}} - \frac{\mathbf{1}}{\mathbf{n}}\nu\nu^{\mathbf{T}} + \frac{\mathbf{1}}{\mathbf{n^2}}\nu\nu^{\mathbf{T}}\nu\nu^{\mathbf{T}} \end{aligned}$$

On peut remarquer que $\nu^{\mathbf{T}}\nu = \sum_{i=1}^{I}\sqrt{J}\sqrt{J} = n$.

Donc les deux derniers termes de la partie droite de la dernière équation s'annulent et on a $\mathbf{A^2} = \mathbf{A}$.

D'après un théorème du chapitre 1, $SC_F/\sigma^2$ suit alors une loi du khi-deux $\chi^2_d(\delta)$ où $d$ est l'ordre de la matrice $\mathbf{A}$, $\delta = \xi^{\mathbf{T}}\mathbf{A}\xi/\sigma^2$ et $\xi = E[\mathbf{Z}] = (\sqrt{\mathbf{J}}\mu_{\mathbf{1}}, \sqrt{\mathbf{J}}\mu_{\mathbf{2}}, \cdots, \sqrt{\mathbf{J}}\mu_{\mathbf{I}})^{\mathbf{T}}$.

 

L'ordre de la matrice $\mathbf{A}$ est égal à $I-1$. D'autre part, on a :

$$\begin{aligned}\xi^{\mathbf{T}}\mathbf{A}\xi &= \xi^{\mathbf{T}}\xi - \frac{1}{n}\xi^{\mathbf{T}}\mu\mu^{\mathbf{T}}\xi \\ &= \sum_{i=1}^{I} J\mu_i^2 - \frac{1}{n}\left(\sum_{i=1}^{I} J\mu_i\right)^2 \\ &= \sum_{i=1}^{I} J\mu_i^2 - n\mu^2 \\ &= \sum_{i=1}^{I} J(\mu_i-\mu)^2\end{aligned}$$

D'où $\delta = \sum_{i=1}^{I} J(\mu_i-\mu)^2/\sigma^2$.

**Proposition 3.2.3**
*Sous les hypothèses usuelles de l'analyse de la variance, on a :*

$$\frac{SC_R}{\sigma^2} \sim \chi^2_{n-I}\,.$$

En effet, posons :

$$S_i^2 = \frac{1}{J-1}\sum_{j=1}^{J}(Y_{i,j}-Y_{i,\bullet})^2\,.$$

Pour $i=1,2,\cdots,I$, d'après une proposition du chapitre 1, on a $(J-1)S_i^2/\sigma^2 \sim \chi^2_{J-1}$. Par indépendance des échantillons, $SC_R/\sigma^2 = \sum_{i=1}^{I}(J-1)S_i^2/\sigma^2 \sim \chi^2_{n-I}$ puisque $\sum_{i=1}^{I}(J-1) = n-I$.

**Proposition 3.2.4**
*Sous les hypothèses habituelles de l'analyse de la variance, les statistiques $SC_F$ et $SC_R$ sont indépendantes et on a :*

$$\frac{SC_T}{\sigma^2} \sim \chi^2_{n-1}\left(\sum_{i=1}^{I} J(\mu_i-\mu)^2/\sigma^2\right) = \chi^2_{n-1}\left(J\sum_{i=1}^{I}\alpha_i^2/\sigma^2\right)$$

Pour $i=1,2,\cdots,I$, les statistiques $Y_{i,\bullet}$ et $S_i^2$ sont indépendantes d'après une proposition du chapitre 1. Les statistiques $SC_F$ et $SC_R$ sont donc indépendantes puisque la première est une fonction de $\{Y_{1,\bullet}, Y_{2,\bullet}, \cdots, Y_{I,\bullet}\}$ et la deuxième est une fonction de $\{S_1^2, S_2^2, \cdots, S_I^2\}$. D'après un exercice du chapitre 1 , $SC_T = SC_F + SC_R \sim \chi^2_{n-1}\left(\sum_{i=1}^{I} J(\mu_i-\mu)^2/\sigma^2\right)$.

**Proposition 3.2.5**
*Sous les hypothèses habituelles de l'analyse de la variance, posons :*

$$F = \frac{SC_F/(I-1)}{SC_R/(n-I)}\,.$$

*On a alors :* $F \sim \mathcal{F}_{I-1,n-I}\left(J\sum_{i=1}^{I}\alpha_i^2/\sigma^2\right)$.

Cette dernière proposition est une conséquence directe de la définition d'une loi de Fisher non centrée. On en déduit alors que

$$E\left[\frac{SC_F/(I-1)}{SC_R/(n-I)}\right] = \frac{I-1+\sum_{i=1}^{I}J(\mu_i-\mu)^2/\sigma^2}{I-1}\,\frac{n-I}{n-I-2}$$

d'après un exercice du chapitre 1.

On peut aussi remarquer que sous l'hypothèse $H_0 : \mu_1 = \cdots = \mu_I = \mu$ ou, de manière équivalente $H_0 : \alpha_1 = \cdots = \alpha_i = 0$, on a :

$$\sum_{i=1}^{I}J(\mu_i-\mu)^2/\sigma^2 = J\sum_{i=1}^{I}\alpha_i^2/\sigma^2 = 0\,.$$

Sous cette hypothèse, les trois statistiques $SC_T$, $SC_F$ et $SC_R$ suivent des lois du khi-deux centrées à respectivement $n-1$, $I-1$ et $n-I$ degrés de liberté.

**Test d'égalité des moyennes avec variance connue :**

Dans cette section, on élabore le test d'égalité des moyennes

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_I$$

contre

$$H_1 : \text{ il existe } i \neq j \text{ tels que } \mu_i \neq \mu_j\,,$$

ce qui est équivalent au test d'hypothèses suivant :

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_I = 0$$

contre

$$H_1 : \text{Il existe } i_0 \in \{1, 2, \cdots, I\} \text{ tel que } \alpha_{i_0} \neq 0\,.$$

La construction de ce test est basée sur la méthode du rapport des maximums de vraisemblance. Lorsque la variance commune $\sigma^2$ est connue, la vraisemblance globale s'écrit sous la forme :

$$\begin{aligned} L(\mu_1, \mu_2, \cdots, \mu_I) &= \prod_{i=1}^{I}\prod_{j=1}^{n_i}\{\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(Y_{ij}-\mu_i)^2}{2\sigma^2}}\} \\ &= (2\pi\sigma^2)^{-\frac{N}{2}}\exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij}-\mu_i)^2) \end{aligned}$$

L'estimateur du maximum de vraisemblance du vecteur $(\mu_1, \mu_2, \cdots, \mu_I)$ est $(Y_{1,\bullet}, Y_{2,\bullet}, \cdots, Y_{I,\bullet})$. Sous $H_0$, cet estimateur devient $(Y_{\bullet,\bullet}, \bar{Y}_{\bullet,\bullet}, \cdots, Y_{\bullet,\bullet})$. Le rapport des vraisemblances s'écrit alors :

$$\begin{aligned} \Lambda &= \frac{L(Y_{\bullet,\bullet}, Y_{\bullet,\bullet}, \cdots, Y_{\bullet,\bullet})}{L(Y_{1,\bullet}, Y_{2,\bullet}, \cdots, Y_{I,\bullet})} \\ &= \frac{(2\pi\sigma^2)^{-\frac{N}{2}}\exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij}-Y_{\bullet,\bullet})^2)}{(2\pi\sigma^2)^{-\frac{N}{2}}\exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij}-Y_{i,\bullet})^2)} \\ &= \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij}-Y_{\bullet,\bullet})^2 - \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij}-Y_{i,\bullet})^2\} \\ &= \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{I} n_i(Y_{i,\bullet}-Y_{\bullet,\bullet})^2\} \end{aligned}$$

Le passage de l'avant dernière ligne à la dernière ligne se fait en remarquant que :

$$\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij}-Y_{i,\bullet})^2 = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij}-Y_{\bullet,\bullet})^2 - \sum_{i=1}^{I} n_i(Y_{i,\bullet}-Y_{\bullet,\bullet})^2$$

Cette égalité a été établie lors de la décomposition de $SC_T$ en somme $SC_T = SC_F + SC_R$.

Donc on rejette $H_0$ si le rapport $\Lambda = \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{I} n_i(Y_{i,\bullet}-Y_{\bullet,\bullet})^2\}$ est petit, c'est à dire si $\sum_{i=1}^{I} n_i(Y_{i,\bullet}-Y_{\bullet,\bullet})^2/\sigma^2$ est grand. On reconnaît l'expression de $SC_F/\sigma^2$.

**Proposition 3.2.6**
*Lorsque la variance est connue, on rejette $H_0$ au seuil $1-\alpha$ si et seulement si*

$$\frac{SC_F}{\sigma^2} > \chi^2_{I-1,1-\alpha}$$

En effet, d'après la section précédente, on a vu que $SC_F/\sigma^2 \sim \chi^2_{I-1}$ sous $H_0$.

**Test d'égalité des moyennes avec variance inconnue :**

Dans cette section, on élabore le test d'égalité des moyennes

$$H_0 \,:\, \mu_1 = \mu_2 = \cdots = \mu_I$$

contre

$$H_1 \,:\, \text{ il existe } i \neq j \text{ tels que } \mu_i \neq \mu_j \,,$$

ce qui est équivalent au test d'hypothèses suivant :

$$H_0 \,:\, \alpha_1 = \alpha_2 = \cdots = \alpha_I = 0$$

contre

$$H_1 \,:\, \text{Il existe } i_0 \in \{1, 2, \cdots, I\} \text{ tel que } \alpha_{i_0} \neq 0 \,.$$

. Sous $H_0$, la variance est supposée commune à tous les échantillons, mais inconnue. Dans le paragraphe précédent, si la variance est connue, on a vu qu'on rejette $H_0$ lorsque $SC_F/\sigma^2$ est grande . Si $\hat{\sigma}^2$ est un estimateur convenable de $\sigma^2$, il est alors naturel de rejeter $H_0$ lorsque $SC_F/\hat{\sigma}^2$ est grande lorsque la variance est inconnue.

Cherchons alors le "meilleur" estimateur possible pour $\sigma^2$. On sait que pour $i = 1, 2, \cdots, I$, $(n_i - 1)S_i^2 \sim \chi^2_{n_i - 1}$ et par conséquent $E[S_i^2] = \sigma^2$. Chacun des $S_1^2, S_2^2, \cdots, S_I^2$ est un estimateur sans biais de $\sigma^2$.

Il est donc naturel de chercher le meilleur estimateur sans biais de $\sigma^2$ parmi les combinaisons linéaires de $S_1^2, S_2^2, \cdots, S_I^2$. Un tel estimateur s'écrit sous la forme $\tilde{\sigma}^2 = \sum_{i=1}^{I} a_i S_i^2$. On a alors ; $\sigma^2 = E[\tilde{\sigma}^2] = \sum_{i=1}^{I} a_i \sigma^2$. On en déduit que : $\sum_{i=1}^{I} a_i = 1$ qu'on peut encore écrire :

$$a_I = 1 - (a_1 + a_2 + \cdots + a_{I-1}) = 1 - \sum_{i=1}^{I-1} a_i.$$

D'autre part $var[\tilde{\sigma}^2] = \sum_{i=1}^{I} a_i^2 var[S_i^2]$.

Or, pour $i = 1, 2, \cdots, I$, on a :

$$var[(n_i - 1)\frac{S_i^2}{\sigma^2}] = var[\chi^2_{n_i - 1}] = 2(n_i - 1).$$

On en déduit que $var[S_i^2] = 2\sigma^4/(n_i - 1)$ et que :

$$\begin{aligned} var[\tilde{\sigma}^2] &= \sum_{i=1}^{I} 2a_i^2 \frac{\sigma^4}{n_i - 1} \\ &= 2\sigma^4 \left\{ \sum_{i=1}^{I-1} \frac{a_i^2}{n_i - 1} + \frac{(1 - (a_1 + a_2 + \cdots + a_{I-1}))^2}{n_I - 1} \right\} \end{aligned}$$

Pour minimiser $var[\tilde{\sigma}^2]$, on la dérive par rapport à $(a_1, a_2, \cdots, a_{I-1})$.
Pour $i = 1, 2, \cdots, I$, on a :

$$\frac{\partial}{\partial a_i} var[\tilde{\sigma}^2] = 2\sigma^4 \left\{ \frac{a_i}{n_i - 1} - \frac{1 - (a_1 + a_2 + \cdots + a_{I-1})}{n_I - 1} \right\}$$

En égalant cette dérivée à 0 : $\partial \mathrm{Var}[\tilde{\sigma}^2]/\partial a_i = 0$, on obtient :

$$\frac{a_1}{n_1 - 1} = \frac{a_2}{n_2 - 1} = \cdots = \frac{a_I}{n_I - 1}\,.$$

Et comme on doit avoir $a_1 + a_2 + \cdots + a_I = 1$, on en déduit que $a_i = (n_i - 1)/(N - I)$.

Le meilleur estimateur sans biais s'écrit alors

$$\tilde{\sigma}^2 = \sum_{i=1}^{n} \frac{n_i - 1}{N - I} S_i^2 = \frac{SC_R}{N - I} = S_F^2.$$

On rejette alors $H_0$ si $SC_F/SC_R$ est grand.

**Proposition 3.2.7**
*Lorsque la variance est inconnue, on rejette $H_0$ au seuil $1 - \alpha$ si et seulement si :*

$$F = \frac{SC_F/(I-1)}{SC_R/(n-I)} = \frac{S_F^2}{S_R^2} > F_{I-1,n-I,1-\alpha}$$

Ce dernier résultat est une conséquence directe de la proposition 3.2.5.

Les statistiques $S_F^2$ et $S_R^2$ définies respectivement par $S_F^2 = SC_F/(I-1)$ et $S_R^2 = SC_R/(n-I)$.

La méthode du rapport de vraisemblances avec variance inconnue donne le même test que la proposition 3.2.7.

**Table de l'analyse de variance :**

La relation (3.2.5) s'écrit parfois, de manière équivalente, en utilisant les variances au lieu des carrés des écarts ; il suffit pour cela de diviser tous les termes par $I \times J$ :

$$\frac{1}{I \times J} \sum_{i=1}^{I} \sum_{j=1}^{J} (Y_{i,j} - Y_{\bullet,\bullet})^2 = \frac{1}{I} \sum_{i=1}^{I} \sum_{j=1}^{J} (Y_{i,j} - Y_{i,\bullet})^2 + \frac{1}{I \times J} \sum_{i=1}^{I} \sum_{j=1}^{J} (Y_{i,\bullet} - Y_{\bullet,\bullet})^2\,, \quad (3.2.7)$$

qu'on peut noter :

$$S^2 = S_F^2 + S_R^2\,, \quad (3.2.8)$$

appelée formule de l'analyse de la variance, où $S_F^2$ représente la variance due au facteur $A$, où $S_R^2$ représente la variance résiduelle, et $S^2$ la variance globale.

Les diverses quantités (3.2.2), (3.2.3) et (3.2.4) suivent, au facteur multiplicatif $\sigma^2$ près, des lois du khi-deux avec des nombres de degrés de liberté respectifs $IJ-1$, $I-1$ et $IJ-I$.

Ceci est résumé dans le tableau suivant :

| **Source** | **Variations** | **Degrés de liberté** | **Carrés moyens** | $\boldsymbol{F}$ |
|---|---|---|---|---|
| Facteur | $SC_F$ | $I-1$ | $S_F^2 = \dfrac{SC_F}{I-1}$ | $F = \dfrac{S_F^2}{S_R^2}$ |
| Résidu | $SC_R$ | $IJ-I = I(J-1)$ | $S_R^2 = \dfrac{SC_R}{I(J-1)}$ | |
| Total | $SC_{TOT}$ | $IJ-1$ | $S_T^2 = \dfrac{SC_{TOT}}{IJ-1}$ | |

On a défini ci-dessus les carrés moyens :

$$S_F^2 = \frac{SC_F}{I-1} \quad ; \quad S_R^2 = \frac{SC_R}{I(J-1)} \quad ; \quad S_T^2 = \frac{SC_{TOT}}{IJ-1}\,,$$

qui constituent eux aussi des mesures globales de variations.

La liste $\boldsymbol{y}$ des données expérimentales $y_{1,1}, \cdots, y_{1,J}, y_{2,1}, \cdots, y_{2,J}, \cdots, y_{I,J}$ permet de construire une réalisation du tableau précédent :

| **Facteur $\boldsymbol{A}$** | $\boldsymbol{y}$ |
|---|---|
| $A_1$ | $y_{1,1}, \cdots, y_{1,J}$ |
| $\vdots$ | $\vdots$ |
| $A_i$ | $y_{i,1}, \cdots, y_{i,J}$ |
| $\vdots$ | $\vdots$ |
| $A_I$ | $y_{I,1}, \cdots, y_{I,J}$ |

La variation due au facteur $A$ observée sur la liste de données $\boldsymbol{y}$ est définie par :

$$sc_F = J \sum_{i=1}^{I} (y_{i,\bullet} - y_{\bullet,\bullet})^2\,.$$

La variation résiduelle observée sur la liste de données $\boldsymbol{y}$ est définie par :

$$sc_R = \sum_{i=1}^{I} \left( \sum_{j=1}^{J} (y_{i,j} - y_{i,\bullet})^2 \right)$$

Enfin, la variation totale observée sur la liste de données $\boldsymbol{y}$ est égale par :

$$sc_{TOT} = \sum_{i=1}^{I} \sum_{j=1}^{J} (y_{i,j} - y_{\bullet,\bullet})^2\,.$$

La relation fondamentale de l'ANOVA reste valable lorsqu'elle est évaluée sur la liste de données $\boldsymbol{y}$ :

$$sc_{TOT} = sc_F + sc_R\,.$$

On peut résumer toutes ces informations dans le tableau d'analyse de variance suivant :

| Source | Variations | Degrés de liberté | Carrés moyens | $F$ | Décision |
|---|---|---|---|---|---|
| Facteur | $sc_F$ | $I-1$ | $s_F^2 = \dfrac{sc_F}{I-1}$ | $f = \dfrac{s_F^2}{s_R^2}$ | $H_1$ ou $H_1$ |
| Résidu | $sc_R$ | $n-I = I(J-1)$ | $s_R^2 = \dfrac{sc_R}{n-I}$ | | |
| Total | $sc_{TOT}$ | $n-1 = IJ-1$ | | | |

On souhaite effectuer le test d'hypothèses suivant :

$$H_0 \,:\, \alpha_1 = \alpha_2 = \cdots = \alpha_I = 0$$

contre

$$H_1 \,:\, \text{Il existe } i_0 \in \{1, 2, \cdots, I\} \text{ tel que } \alpha_{i_0} \neq 0\,.$$

On rappelle que nous sommes dans le cas gaussien et que les variances sont supposées égales. On montre alors que :

$$E\,(S_T^2) \;=\; \sigma^2 + \frac{J}{IJ-1}\sum_{i=1}^{I}\alpha_i^2$$

$$E\,(S_F^2) \;=\; \sigma^2 + \frac{J}{I-1}\sum_{i=1}^{I}\alpha_i^2$$

$$E\,(S_R^2) \;=\; \sigma^2\,.$$

Clairement, sous $H_0$, les variables :

$$\chi^2_{TOT} = S_T^2 = \frac{SC_{TOT}}{\sigma^2} \quad ; \quad \chi^2_F = S_F^2 = \frac{SC_F}{\sigma^2} \quad ; \quad \chi^2_R = S_R^2 = \frac{SC_R}{\sigma^2}$$

sont des variables aléatoires de lois du khi-deux à respectivement $IJ-1$, $I-1$ et $IJ-I$ degrés de liberté, et les deux dernières sont indépendantes.

De plus, le rapport des variables $S_F^2$ et $S_R^2$ divisées par leur nombre de degrés de liberté respectifs :

$$\frac{\chi^2_F}{I-1} \Big/ \frac{\chi^2_R}{IJ-I} = \frac{S_F^2}{S_R^2} \text{ suit une loi de Fisher-Snedecor } \mathcal{F}(I-1, IJ-I)\,.$$

Le test de l'hypothèse nulle nécessite le calcul de la quantité : $f = \dfrac{S_F^2}{S_R^2}$.

Le rejet de l'hypothèse nulle, au niveau $\alpha$, intervient quand cette dernière quantité est trop élevée, c'est-à-dire quand :

$$P(F \geq f) \leq \alpha \quad \text{ou} \quad f \geq F_{1-\alpha}\,,$$

avec une loi de Fischer-Snedecor à $I-1$, $IJ-I$ degrés de liberté. Ce test est unilatéral, car dans tous les cas où $H_0$ est fausse, les valeurs observées $f$ dépassent en moyenne les valeurs que donnent usuellement les lois $F$ de Fisher-Snedecor.

On concluera donc à l'aide de la probabilité critique, et on rejettera $H_0$ si cette probabilité est inférieure ou égale au seuil $\alpha$ du test.

Lorsque $H_0$ est rejetée, on peut alors procéder à des comparaisons multiples des différents effets du niveau du facteur, ce qui sera vu plus loin.

### 3.2.3 Estimations

Les estimateurs $\hat{\mu}, \hat{\alpha}_1, \cdots, \hat{\alpha}_I$ et $\hat{\sigma}^2$ des paramètres respectifs $\mu, \alpha_1, \cdots, \alpha_I$ et $\sigma^2$ du modèle sont données par :

$$\hat{\mu} = Y_{\bullet,\bullet} \quad ; \quad \hat{\alpha}_i = Y_{i,\bullet} - \hat{\mu} \qquad 1 \leq i \leq I$$

$$\hat{\sigma}^2 = \frac{SC_R}{IJ-I} = S_R^2\,.$$

Ce sont des estimateurs sans biais. Les estimations obtenues pour une liste de données $\boldsymbol{y}$, notées $\hat{\mu}(\boldsymbol{y})$, $\hat{\alpha}_1(\boldsymbol{y}), \cdots, \hat{\alpha}_I(\boldsymbol{y})$ et $\hat{\sigma}^2(\boldsymbol{y})$ des paramètres $\mu$, $\alpha_1, \cdots, \alpha_I$ et $\sigma^2$ du modèle se déduisent des formules précédentes.

On peut en outre calculer comme suit des intervalles de confiance pour les moyennes des différentes populations (parmi les $I$ populations) :

$$Y_{i,\bullet} \pm t_{1-\alpha/2}\sqrt{\frac{S_R^2}{n_i}}\,,$$

et aussi pour les différences de moyennes :

$$Y_{i,\bullet} - Y_{i',\bullet} \pm t_{1-\alpha/2}\sqrt{S_R^2\left(\frac{1}{n_i} + \frac{1}{n_{i'}}\right)}\,,$$

la variable de Student $t$ étant une variable à $n-I$ degrés de liberté.

Des limites de confiance relatives à la variance $\sigma^2$ peuvent également être obtenues selon les procédures usuelles vues au chapitre précédent, à partir de la somme des carrés résiduels, ou du carré moyen résiduel, et grâce à la loi du khi-deux à $n-I$ degrés de liberté.

### 3.2.4 Retour à l'exemple initial

On veut comparer les teneurs en sodium de lasalocide dans la nourriture de volaille, et ce, pour trois laboratoires.

On cherche à savoir s'il existe ou non, en moyenne, des différences significatives entre laboratoires.

On suppose que les hypothèses de normalité et d'égalité des variances sont satisfaites.

Il y a 3 modalités, les trois laboratoires.
Après calculs avec le logiciel ***R***, on a déjà calculé les moyennes :

$$\bar{y}_{1,\bullet} = 85,5 \qquad \bar{y}_{2,\bullet} = 88,6 \qquad \bar{y}_{3,\bullet} = 83,8 \qquad \bar{y}_{\bullet,\bullet} = 85.97$$

Appliqué à la première observation du premier laboratoire ($y_{1,1} = 87$), le modèle observé d'analyse de variance s'écrit :

$$(87 - 85,97) = (85,5 - 85,97) + (87 - 85,5) \quad \text{ou} \quad 1,03 = -0,47 + 1,5\,.$$

L'effet positif de 1,03 entre cette observation particulière et la moyenne générale provient, à la fois du fait que cette teneur est mesurée dans un certain laboratoire, dont la moyenne est inférieure de 0,47 par rapport à la moyenne générale, et que cette observation a une teneur supérieure de 1,5 par rapport à la moyennes de toutes les teneurs observées dans ce même laboratoire.

Un calcul similaire peut être effectué pour chaque des 30 observations, et en sommant les carrés des écarts, on aboutit aux trois sommes des carrés des écarts :

$$SC_{TOT} = (1,03)^2 + \cdots\,, \quad SC_F = (-0,47)^2 + \cdots\,, \quad SC_R = (1,5)^2 + \cdots\,.$$

Cela donne le tableau suivant :
On trouvera ci-dessous le tableau d'analyse de variance :

```
            Df Sum Sq Mean Sq F value   Pr(>F)
labo         2  118.5   59.23   9.491 0.000756 ***
Residuals   27  168.5    6.24
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Il y a donc une différence significative des moyennes des teneurs en sodium pour les 3 laboratoires.

### 3.2.5 Généralisation

Pour faire face aux cas pratiques, on peut généraliser au cas où, pour chaque modalité, le nombre d'observations n'est pas nécessairement le même. On notera $n_i$ le nombre d'observations dans la modalité $A_i$.

Le modèle (3.2.1) reste inchangé, mais, cette fois, sous la contrainte :

$$\sum_{i=1}^{I} n_i \alpha_i = 0\,,$$

sous les mêmes hypothèses.

On désignera les $I$ moyennes de chaque modalité par :

$$Y_{i,\bullet} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{i,j}\,.$$

En notant $n = \sum_{i=1}^{I} n_i$, la moyenne globale est :

$$Y_{\bullet,\bullet} = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{n_i} Y_{i,j}\,.$$

La variation théorique totale ou somme totale des carrés des écarts vaut :

$$SC_{TOT} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{i,j} - Y_{\bullet,\bullet})^2\,. \tag{3.2.9}$$

La variation théorique due au facteur $A$ est :

$$SC_F = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{i,\bullet} - Y_{\bullet,\bullet})^2\,. \tag{3.2.10}$$

Enfin, la variation résiduelle vaut :

$$SC_R = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{i,j} - Y_{i,\bullet})^2 \tag{3.2.11}$$

La relation fondamentale de l'ANOVA est toujours la relation (3.2.5). Les remarques faites dans le paragraphe 1.2.2 quant aux lois rencontrées, restent valables ici, mutatis mutandis.

Le tableau d'analyse de variance s'écrit alors :

| **Source** | **Variations** | **Degrés de liberté** | **Carrés moyens** | $\boldsymbol{F}$ |
|---|---|---|---|---|
| Facteur | $SC_F$ | $I-1$ | $S_F^2 = \dfrac{SC_F}{I-1}$ | $F = \dfrac{S_F^2}{S_R^2}$ |
| Résidu | $SC_R$ | $n-I$ | $S_R^2 = \dfrac{SC_R}{n-I}$ | |
| Total | $SC_{TOT}$ | $n-1$ | $S_T^2 = \dfrac{SC_{TOT}}{n-1}$ | |

**Exemple 3.2.1**

*On veut comparer des hauteurs moyennes, exprimées en mètres, des arbres de trois types de hêtraies. On cherche effectivement à savoir s'il existe ou non, en moyenne, des différences significatives de hauteurs d'arbres entre les trois types de forêts. On suppose que les hypothèses de normalité et d'égalité des variances sont satisfaites. Les données sont fournies dans le tableau suivant :*

| *Type 1* | *Type 2* | *Type 3* |
|---|---|---|
| *23,4* | *22,5* | *18,9* |
| *24,4* | *22,9* | *21,1* |
| *24,6* | *23,7* | *21,2* |
| *24,9* | *24,0* | *22,1* |
| *25,0* | *24,4* | *22,5* |
| *26,2* | *24,5* | *23,6* |
| *26,3* | *25,3* | *24,5* |
| *26,8* | *26,0* | *24,6* |
| *26,8* | *26,2* | *26,2* |
| *26,9* | *26,4* | *26,7* |
| *27,0* | *26,7* | |
| *27,6* | *26,9* | |
| *27,7* | *27,4* | |
| | *28,5* | |

*Il y a trois modalités, les trois types de hêtraies. Les valeurs relatives aux 37 endroits où les mesures de hauteur ont été réalisées, produisent les moyennes respectives :*

$$\overline{x}_{1,\bullet} = 25,97 \quad , \quad \overline{x}_{2,\bullet} = 25,39 \quad , \quad \overline{x}_{3,\bullet} = 23,14 \quad \textit{et} \quad \overline{x}_{\bullet,\bullet} = 24,98\,.$$

*Appliqué à la première observation du premier échantillon ($x_{1,1} = 23,4$), le modèle observé d'analyse de variance s'écrit :*

$$(23,4 - 24,98) = (25,97 - 24,98) + (23,4 - 25,97) \quad \textit{ou} \quad -1,58 = 0,99 - 2,57\,.$$

*L'effet négatif de 1,58 m entre cette observation particulière et la moyenne générale provient, à la fois du fait que l'endroit considéré appartient à un certain type de forêt dont la moyenne est supérieure de 0,99 m par rapport à la moyenne générale, et que cet endroit présente une hauteur inférieure de 2,57 m, par rapport à la moyenne de toutes les observations relatives à ce même type de forêt.*

*Un calcul similaire peut être effectué pour chacun des 36 autres arbres, et, en sommant les carrés des écarts ainsi obtenus, on aboutit aux trois sommes des carrés des écarts :*

$$SC_{TOT} = (-1,58)^2 + \cdots\,, \qquad SC_F = (0,99)^2 + \cdots\,, \qquad SC_R = (-2,57)^2 + \cdots\,.$$

*Cette façon de procéder n'est pas celle suivie habituellement, car on le fait souvent informatiquement, mais est utile d'un point de vue didactique pour bien comprendre le mécanisme de l'analyse de variance.*

*Le tableau ci-dessous présente la somme des carrés des écarts obtenue de cette manière :*

| Sources de variation | Variations | Degrés de liberté | Carrés moyens |
|---|---|---|---|
| *Différences entre types de hêtraies* | $SC_F = 48,88$ | *2* | *24,44* |
| *Différences intra-type* | $SC_R = 116,65$ | *34* | *3,431* |
| *Total* | $SC_{TOT} = 165,53$ | *36* | *4,598* |

*À partir du tableau précédent, on obtient alors :*

$$F_{obs} = 24,4/3,431 = 7,12 \quad et \quad P(F \geq 7,12) = 0,0026\,,$$

*avec 2 et 34 degrés de liberté. L'hypothèse d'égalité des hauteurs moyennes des arbres dans les trois types de hêtraies doit donc être rejetée, même au niveau 1% : les différences observées entre les trois types de hêtraies sont hautement significatives.*

*Les limites de confiance des différences sont, pour un degré de confiance de 95%, et pour les deux premiers types de forêts :*

$$25,97 - 25,39 \pm 2,032\sqrt{3,431\left(\frac{1}{13}+\frac{1}{14}\right)} = 0,58 \pm 1,45 = -0,87 \ et\ 2,03m\,,$$

*pour le premier et troisième type de forêts :*

$$25,97 - 23,14 \pm 2,032\sqrt{3,431\left(\frac{1}{13}+\frac{1}{10}\right)} = 2,83 \pm 1,58 = 1,25 \ et\ 4,41m\,,$$

*et pour les deux derniers types de forêts :*

$$25,39 - 23,14 \pm 2,032\sqrt{3,431\left(\frac{1}{14}+\frac{1}{10}\right)} = 2,25 \pm 1,56 = 0,69 \ et\ 3,81m\,.$$

*Le fait que le premier intervalle de confiance contienne zéro, indique qu'il n'y a pas de différence significative entre les deux premiers types d'arbres, ce qui était déjà la conclusion à laquelle nous avions abouti dans un exemple du chapitre précédent.*

*On peut aussi réaliser cette étude sous* ***R****. Les commandes sont les suivantes :*

```
> hetraie<-rep(1:3,c(13,14,10))
>  hauteur<-c(23.4,24.4,24.6,24.9,25.0,26.2,26.3,26.8,26.8,26.9,27.0,27.6,27.7,
+ + 22.5,22.9,23.7,24.0,24.4,24.5,25.3,26.0,26.2,26.4,26.7,26.9,27.4,28.5,
+ + 18.9,21.1,21.2,22.1,22.5,23.6,24.5,24.6,26.2,26.7)
> hetraie<-factor(hetraie)
> arbre<-data.frame(hetraie,hauteur)
> modele1<-aov(hauteur~hetraie,data=arbre)
> summary(modele1)
```

*Le tableau d'analyse de variance fourni est le suivant :*

```
            Df Sum Sq Mean Sq F value  Pr(>F)
hetraie      2  48.88  24.441   7.124 0.00261 **
Residuals   34 116.65   3.431
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

*On conclut bien entendu comme ci-dessus.*

## 3.3 Modèle à effets aléatoires

L'étude en détail exposée pour le modèle à effets fixes l'a été volontairement, pour pouvoir aller un peu plus vite pour le modèle à effets aléatoires, et plus tard, pour l'analyse de variance à deux et plus de deux facteurs.

**Exemple 3.3.1**

*On s'intéresse au niveau en mathématiques des étudiants des cégéps de la région de Québec. On prend alors un échantillon de 20 finissants de chaque cégep de la région de Québec. On leur fait passer une épreuve commune, puis on compare les résultats. C'est une expérience à effets fixes. Les modalités du facteur étudié sont les cégeps de la région de Québec. Ce facteur est fixe.*

*Supposons maintenant qu'on veuille répondre à la question suivante : Est ce que le niveau en mathématiques est variable d'un cégep à l'autre dans la province de Québec ? Si tel était le cas, on aimerait mesurer cette variabilité.*

*Dans un premier temps, on sélectionne un échantillon de cégeps parmi les cégeps de la province ; ensuite on procède comme avant et on tire au hasard 20 étudiants de chaque cégep (il s'agit d'un échantillonnage à deux degrés). On s'intéresse autant aux cégeps échantillonnés qu'à ceux qui ne l'ont pas été, car on veut étudier la variabilité inter-cégeps des compétences en mathématiques. Dans ce dernier contexte, le facteur cégep est aléatoire.*

Comme nous l'avons signalé plus haut, dans le cas du modèle aléatoire, les populations dans lesquelles les observations sont réalisées, sont choisies au hasard au sein d'un ensemble très vaste.

On admettra donc que les effets des $A_i$, à savoir les $\alpha_i$, sont des variables aléatoires de loi normale centrée de variance $\sigma_A^2$.

Le modèle ne dépend plus, cette fois, que de trois paramètres $\mu$, $\sigma^2$ et $\sigma_A^2$. Pour chacune des $i$ modalités, on effectue $n_i$ mesures d'une réponse $Y$ qui est une variable continue. On notera encore $n = \sum_{i=1}^{I} n_i$ le nombre total de mesures ayant été effectuées.

Le modèle est le suivant :

$$Y_{i,j} = \mu + \alpha_i + \varepsilon_{i,j} \qquad , \qquad i = 1, \cdots, I\,;\, j = 1, \cdots, n_i\,, \tag{3.3.1}$$

où $Y_{i,j}$ est la valeur prise par la variable réponse $Y$ dans la condition $A_i$ lors de la $j$-ème répétition.

On suppose que :

$$\mathcal{L}(\alpha_i) = \mathcal{N}(0, \sigma_A^2)\,, \qquad \forall\, i\,,\, 1 \leq i \leq I\,,$$

ainsi que l'indépendance des effets aléatoires :

$$\alpha_i \text{ est indépendant de } \alpha_k \text{ si } i \neq k \text{ et } 1 \leq i, k \leq I\,.$$

On postule également les hypothèses supplémentaires pour les erreurs :

$$\forall\,(i,j),\, 1 \leq i \leq I,\, 1 \leq j \leq n_i\,,\, \mathcal{L}(\varepsilon_{i,j}) = \mathcal{N}(0, \sigma^2)\,,$$

$$\varepsilon_{i,j} \text{ est indépendant de } \varepsilon_{k,l} \text{ si } (i,j) \neq (k,l) \text{ avec } 1 \leq i, k \leq I, \text{ et } 1 \leq j, k \leq n_i\,,$$

 

ainsi que l'indépendance des effets aléatoires et des erreurs :

$$\alpha_i \text{ est indépendant de } \varepsilon_{j,k} \text{ si } 1 \leq i, j \leq I, \text{ et } 1 \leq k \leq n_i \,.$$

Nous supposons que les conditions d'utilisation de ce modèle sont bien remplies ; l'étude de leur vérification sera étudiée plus loin.

Avec ce modèle, on a : $Y_{i,j} \sim \mathcal{N}(0, \sigma_A^2 + \sigma^2)$. On dit alors que $\sigma_A^2$ et $\sigma^2$ sont les composantes de la variance. Une partie de la variabilité de $Y$ est expliquée par la variabilité entre les traitements ($\sigma_A^2$), l'autre par la variabilité résiduelle ($\sigma^2$).

On utilise toujours les mêmes quantités $SC_F$, $SC_R$, $SC_{TOT}$, $sc_F$, $sc_R$ et $sc_{TOT}$ introduites à la section 1.2. La relation fondamentale de l'ANOVA tient toujours :

$$SC_{TOT} = SC_F + SC_R \,.$$

Dans l'analyse de la variance à un facteur fixe, on considère l'hypothèse :

$$H_0 \,:\, \alpha_1 = \alpha_2 = \cdots = \alpha_I = 0 \,.$$

Cette dernière n'a plus de sens dans le contexte d'une analyse de la variance à un facteur aléatoire puisque les modalités sont aléatoires. On cherche à tester si le facteur influence la variabilité de la variable réponse $Y$. Donc, on souhaite, cette fois, faire le test d'hypothèses suivant :

$$H_0 \,:\, \sigma_A^2 = 0 \qquad \text{contre} \qquad H_1 \,:\, \sigma_A^2 \neq 0 \,.$$

La nullité de la variance des effets $A_i$ implique l'égalité des moyennes de toutes les populations considérées, et non pas seulement des moyennes des $I$ populations pour lesquelles on dispose d'observations.

Bien que les deux scénarios soient très différents (entre effets fixes et effets aléatoires), on utilise la même règle de décision dans les deux cas.

On rejettera $H_0$ si $\dfrac{S_F^2}{S_R^2} > F_{I-1,n-I,1-\alpha}$.

Le tableau d'analyse de variance suivant résume les informations nécessaires :

| **Source** | **Variations** | **Degrés de liberté** | **Carrés moyens** | $\boldsymbol{F}$ |
|---|---|---|---|---|
| Facteur | $SC_F$ | $I-1$ | $S_F^2 = \dfrac{SC_F}{I-1}$ | $F = \dfrac{S_F^2}{S_R^2}$ |
| Résidu | $SC_R$ | $n-I$ | $S_R^2 = \dfrac{SC_R}{n-I}$ | |
| Total | $SC_{TOT}$ | $n-1$ | $S_T^2 = \dfrac{SC_{TOT}}{n-1}$ | |

Sous l'hypothèse nulle $H_0$ précédente d'absence d'effet du facteur $A$, et lorsque les conditions de validité du modèle sont pleinement respectées, $F$ est une variable aléatoire qui suit une loi de Fisher-Snedecor à $I-1$ et $n-I$ degrés de liberté. Nous pouvons alors conclure, à partir d'une

réalisation $f$ de $F$, à l'aide de la $p$-valeur. On rejette $H_0$ si $f$ inférieure ou égale au seuil $\alpha$ du test, ou à l'aide d'une table, on rejette $H_0$ si la valeur $f$ est supérieure ou égale à la valeur critique fournie par la table.

On note : $n' = \left( n^2 - \sum_{i=1}^{I} n_i^2 \right) \Big/ [n(I-1)]$.

Les estimateurs $\hat{\mu}$, $\hat{\sigma}_A^2$, $\hat{\sigma}^2$ des paramètres $\mu$, $\sigma_A^2$, $\sigma^2$ du modèle sont fournis par les expressions suivantes :

$$\hat{\mu} = Y_{\bullet,\bullet} \quad , \quad \hat{\sigma}_A^2 = \frac{1}{n'} \left( S_F^2 - S_R^2 \right) \quad , \quad \hat{\sigma}^2 = \frac{SC_R}{n-I} = S_R^2 \, ,$$

où $S_F^2 = \dfrac{SC_F}{I-1}$ et $S_R^2 = \dfrac{SC_R}{n-I}$. Ces estimateurs sont sans biais.

Les estimations obtenues pour la liste de données $\boldsymbol{y}$, sont notées :
$\hat{\mu}(\boldsymbol{y}) = y_{\bullet,\bullet}$ , $\hat{\sigma}_A^2(\boldsymbol{y}) = \dfrac{1}{n'} \left( s_F^2 - s_R^2 \right)$ , $\hat{\sigma}^2(\boldsymbol{y}) = \dfrac{sc_R}{n-I} = s_R^2$ .

**Exemple 3.3.2**

*On s'intéresse à l'ensemble des prairies d'une région donnée, et on souhaite identifier l'importance, absolue ou relative, de la variabilité de la production fourragère, d'une part d'une prairie à l'autre, et d'autre part, d'un endroit à l'autre à l'intérieur des différentes prairies.*

*Pour cela, on a choisi au hasard trois prairies dans l'ensemble du territoire considéré, puis au sein de chacune de ces trois prairies, cinq petites parcelles de 2 m$^2$. En termes d'échantillonnage, c'est un échantillonnage à deux degrés : le choix des trois prairies constitue trois unités du premier degré et les 15 petites parcelles de 2 m$^2$ constituent les 15 unités de second degré.*

*Dans chacune des 15 parcelles, on a mesuré les rendements en matière sèche de fourrage à une date donnée. Les valeurs en tonnes par hectare sont les suivantes :*

| **Parcelles** | **Prairie 1** | **Prairie 2** | **Prairie 3** |
|---|---|---|---|
| *1* | *2,06* | *1,59* | *1,92* |
| *2* | *2,99* | *2,63* | *1,85* |
| *3* | *1,98* | *1,98* | *2,14* |
| *4* | *2,95* | *2,25* | *1,33* |
| *5* | *2,70* | *2,09* | *1,83* |

*Les résultats de l'analyse de variance sont résumés ci-dessous :*

| **Sources de variations** | **Degrés de liberté** | **Variations** | **Carrés moyens** | ***F*** |
|---|---|---|---|---|
| *Différences entre prairies* | *2* | *1,3182* | *0,6591* | *4,23* |
| *Différences entre parcelles* | *12* | *1,8711* | *0,1559* | |
| *Totaux* | *14* | *3,1893* | | |

*La probabilité de dépasser la valeur 4,23 est égale à 0,041, pour une variable $F$ de loi de Fisher-Snedecor à 2 et 12 degrés de liberté. Les différences entre prairies doivent donc être considérées comme juste significatives.*

## 3.4 Méthode non paramétrique

Le test non paramétrique le plus couramment utilisé pour comparer $I$ populations est un test basé sur les rangs. Il est connu sous le nom de test de Kruskal et Wallis.

Le test nécessite le classement de l'ensemble $n$ de toutes les observations par ordre croissant, la détermination des rangs des différentes observations, le calcul de la somme des rangs $X_{i,\bullet}$ relative aux $I$ échantillons, et enfin la détermination de la quantité :

$$\chi^2_{obs} = \frac{12}{n(n+1)} \sum_{i=1}^{I} \frac{X_{i,\bullet}}{n_i} - 3(n+1)\,.$$

Quand l'hypothèse $H_0$ est vraie, cette quantité observée est approximativement une valeur observée d'une variable de loi du khi-deux à $I-1$ degrés de liberté.

$H_0$ sera rejetée, au niveau $\alpha$, si (test unilatéral) :

$$P(\chi^2 \geq \chi^2_{obs}) \leq \alpha \qquad \text{et} \qquad \chi^2_{obs} \geq \chi^2_{1-\alpha}\,.$$

**Exemple 3.4.1**

*On reprend l'exemple de la hauteur des arbres de trois types de hêtraies. Le tableau ci-dessous reprend les données et les rangs sont indiqués à droite :*

| **Hauteurs** | | | **Rangs** | | |
|---|---|---|---|---|---|
| *Type 1* | *Type 2* | *Type 3* | *Type 1* | *Type 2* | *Type 3* |
| *23,4* | *22,5* | *18,9* | *8* | *5,5* | *1* |
| *24,4* | *22,9* | *21,1* | *12,5* | *7* | *2* |
| *24,6* | *23,7* | *21,2* | *16,5* | *10* | *3* |
| *24,9* | *24,0* | *22,1* | *18* | *11* | *4* |
| *25,0* | *24,4* | *22,5* | *19* | *12,5* | *5,5* |
| *26,2* | *24,5* | *23,6* | *23* | *14,5* | *9* |
| *26,3* | *25,3* | *24,5* | *25* | *20* | *14,5* |
| *26,8* | *26,0* | *24,6* | *29,5* | *21* | *16,5* |
| *26,8* | *26,2* | *26,2* | *29,5* | *23* | *23* |
| *26,9* | *26,4* | *26,7* | *31,5* | *26* | *27,5* |
| *27,0* | *26,7* | | *33* | *27,5* | |
| *27,6* | *26,9* | | *35* | *31,5* | |
| *27,7* | *27,4* | | *36* | *34* | |
| | *28,5* | | | *37* | |
| *Total* | | | *316,5* | *280,5* | *106* |

*À partir de la somme des rangs, on obtient :*

$$\chi^2_{obs} = \frac{12}{37 \times 38}\left(\frac{316,5^2}{13} + \frac{280,5^2}{14} + \frac{106^2}{10} - 3 \times 38\right) = 9,32\,,$$

*avec deux degrés de liberté :*

$$P\chi^2 \geq 9,32) = 0,0095\,.$$

*Comme dans l'analyse de variance, le test de Kruskal et Wallis met donc en évidence des différences très significatives de hauteurs d'arbres entre les trois types de hêtraies.*

# Exercices

**Exercice 1**

On cherche à étudier l'effet d'un facteur traitement à 6 modalités sur le rendement de blé. Chaque traitement a été répété sur 4 petites parcelles de 10 mètres carrés.

1. Complétez le tableau d'analyse de la variance suivant :

| **Source de variabilité** | **Somme des carrés des écarts** | **Degrés de liberté** | **Carrés moyens** | $f_{obs}$ |
|---|---|---|---|---|
| **Facteur** | 72,25 | $\cdots$ | $\cdots$ | $\cdots$ |
| **Résiduelle** | $\cdots$ | $\cdots$ | $\cdots$ | |
| **Totale** | 125,35 | $\cdots$ | | |

2. Quel pourcentage d'explication sur le rendement du blé est dû au traitement ?

**Exercice 2**

On considère cinq traitements $T_1, \cdots, T_5$ contre les boutons de fièvre, dont un est un placebo (traitement $T_1$). Ces traitements ont été administrés au hasard sur trente patients ( six patients par groupe de traitement). Le délai, exprimé en jours, entre l'apparition des boutons de fièvre et la cicatrisation complète a été recueilli chez chacun des trente patients, détaillé ci-dessous :

| $T_1$ | $T_2$ | $T_3$ | $T_4$ | $T_5$ |
|---|---|---|---|---|
| 5 | 4 | 6 | 7 | 9 |
| 8 | 6 | 4 | 4 | 3 |
| 7 | 6 | 4 | 6 | 5 |
| 7 | 3 | 5 | 6 | 7 |
| 10 | 5 | 4 | 3 | 7 |
| 8 | 6 | 3 | 5 | 6 |

1. Comparez les moyennes des délais de cicatrisation, délais observés sur cinq échantillons indépendants (groupes de traitement).
2. Estimez les différents paramètres du modèle.

**Exercice 3**

Quinze veaux ont été répartis au hasard en trois lots, les veaux d'un même lot recevant une alimentation particulière. Les gains de poids, observés au cours d'une même période et exprimés en kg, sont présentés ci-dessous, une donnée étant manquante :

Alimentation 1 : 42,1 ; 37,7 ; 45,1 ; 43,1 ;
Alimentation 2 : 45,2 ; 54,2 ; 38,1 ; 48,3 ; 55,1 ;
Alimentation 3 : 48,3 ; 44,1 ; 56,9 ; 42,2 ; 54,0 .

Peut-on considérer que les différences de moyennes constatées entre les alimentations des trois lots sont significatives ?

Dans l'affirmative, estimez ces différences de moyennes et déterminez-en les limites de confiance à 95%.

**Exercice 4**

Une compagnie emploie un grand nombre de représentants, et cherche à savoir lesquels d'entre eux vendent le mieux, parmi les différentes catégories de représentants : ceux payés strictement à la commission, ceux avec un salaire fixe, et ceux qui ont un salaire fixe plus une commission. Une étude des ventes dans cette compagnie, sur le mois précédent, a donné les résultats suivants (résultats des ventes obtenus en milliers de dollars par chaque représentant) :

| **Commissionnés** | **Salaires fixes** | **Commission + salaire** |
|---|---|---|
| 425 | 420 | 430 |
| 507 | 448 | 492 |
| 450 | 437 | 470 |
| 483 | 432 | 501 |
| 466 | 444 | |
| 492 | | |

1. Estimez les moyennes et les écarts-types pour les trois catégories. Faites une boîte à moustache pour une meilleure illustration.
2. Est-ce qu'en moyenne, les ventes diffèrent en fonction des trois différentes catégories ?
3. Déterminez un intervalle à 90% pour les ventes de la catégorie des représentants recevant un salaire plus une commission.

**Exercice 5**

On a relevé les salaires dans trois quartiers d'une grande ville. Ces trois quartiers sont en grande partie occupés par trois communautés $A$, $B$ et $C$ différentes. Le tableau qui suit résume ces salaires (en milliers de dollars) :

| **Communauté** $A$ | **Communauté** $B$ | **Communauté** $C$ |
|---|---|---|
| 43,5 | 73,5 | 45,5 |
| 49,5 | 62,0 | 65.4 |
| 38,0 | 47,5 | 49.4 |
| 66,5 | 36,5 | 58.7 |
| 57,5 | 44,5 | 67.4 |
| 32,0 | 56,0 | 64.8 |
| 67,5 | 68,0 | 69.4 |
| 71,5 | 63,5 | 70.5 |

1. Y a-t-il une différence significative entre les moyennes des salaires dans les trois communautés ?
2. Donnez un intervalle de confiance à 95 % de la différence des moyennes de salaire entre les deux premières communautés ($\alpha = 0,05$).

**Exercice 6**

Pour une étude de santé globale, on s'intéresse à la quantité de gras contenu dans des pièces de viandes de boeuf. Pour cela, on a sélectionné au hasard quatre supermarchés. Dans chacun d'eux, on a choisi aléatoirement 4 pièces de boeuf, d'un même poids d'un kilogramme, pour mesurer le pourcentage de gras dans chacune. Les résultats sont les suivants :

| **Supermarché A** | **Supermarché B** | **Supermarché C** | **Supermarché D** |
|---|---|---|---|
| 22 | 25 | 30 | 18 |
| 20 | 27 | 20 | 20 |
| 23 | 24 | 23 | 17 |
| 25 | 24 | 27 | 17 |

1. De quel type d'échantillonnage s'agit-il ?
2. Peut-on constater, au niveau 5 % une différence significative de pourcentage moyen de gras entre au moins deux supermarchés ?

**Exercice 7**

À l'issue d'un test de dégustation, on a recueilli 8 notes mesurant l'acidité ressentie pour chacune de 4 bières blanches. Ces notes sont rassemblées dans le tableau suivant :

| | **Bière 1** | **Bière 2** | **Bière 3** | **Bière 4** |
|---|---|---|---|---|
| **note 1** | 5 | 0 | 5 | 0 |
| **note 2** | 5 | 1 | 6 | 0 |
| **note 3** | 5 | 2 | 6 | 1 |
| **note 4** | 6 | 2 | 7 | 1 |
| **note 5** | 7 | 3 | 8 | 2 |
| **note 6** | 7 | 4 | 9 | 3 |
| **note 7** | 8 | 6 | 10 | 4 |
| **note 8** | 10 | 6 | 10 | 4 |

On pourra remarquer que chaque note est évaluée sur une échelle allant de 0 à 10. Par exemple, la première note accordée à la bière 4 (note de 0) traduit une absence totale d'acidité pour cette bière La huitième note de la bière 1 (note de 10) traduit au contraire une acidité extrême. Bien entendu, chaque bière est évaluée par un jury indépendant des autres jurys.

1. Faites des boîtes à moustaches pour illustrer le lien entre l'acidité et la bière.
2. Quelle méthode semble adaptée pour savoir si les bières diffèrent par leur acidité ?
3. Écrire le modèle correspondant.
4. Dressez le tableau d'analyse de la variance correspondant.
5. Proposez un test pour comparer globalement ces bières (hypothèse nulle, hypothèse alternative, statistique de test, loi de la statistique sous $H_0$). Prenez une décision au seuil de risque $\alpha = 1\,\%$.
6. Quel pourcentage de variabilité de la note est expliqué par le facteur bière ?

 

 

# Chapitre 4

# Validation des hypothèses d'une ANOVA à un facteur

Dans le modèle standard d'ANOVA vu au chapitre précédent, on a initialement fait quelques hypothèses. Pour que les résultats de l'analyse alors effectuée soient fiables, il est nécessaire que ces diverses hypothèses soient vérifiées. En pratique, il faut valider ces hypothèses à l'aide d'outils statistiques. Dans ce chapitre, nous présentons quelques procédures pratiques pour valider les hypothèses sous-jacentes d'une analyse de variance, qui sont rappelées ci-dessous :

Les résidus $\hat{e}_{i,j}$ sont associés, sans en être des réalisations, aux variables erreurs $\varepsilon_{i,j}$ qui sont inobservables et satisfont aux 3 conditions suivantes :

1. Elles sont indépendantes ;
2. Elles ont même variance $\sigma^2$ inconnue. C'est la condition d'homogénéité ou d'homoscédasticité ;
3. Elles sont de loi gaussienne.

**Remarque 4.0.1**
*Ces trois conditions se transfèrent immédiatement sur les variables aléatoires $Y_{i,j}$.*

Nous étudions les possibilités d'évaluer la validité des trois conditions que nous avons supposées satisfaites.

## 4.1 Conditions d'indépendance

Il n'existe pas, dans un contexte général, de test statistique simple permettant d'étudier l'indépendance. Ce sont les conditions de l'expérience qui nous permettront d'affirmer que nous sommes dans le cas de l'indépendance.

Dans une analyse de variance à un facteur, l'expérience est complètement faite au hasard, au sens où les unités expérimentales sont réparties au hasard entre les modalités du facteur à l'étude. Souvent une bonne planification de l'expérience fait en sorte que les hypothèses de base sont respectées.

Dans une expérience qui compare deux diètes pour des rats, on suppose que les 20 rats de l'expérience sont tous ensemble dans une grande cage. Dix rats recevront la diète 1 et les dix autres rats la diète 2. Supposons également que l'on dispose de 20 cages individuelles.

**Planification 1** : On pourrait prendre les dix premiers rats de la grosse cage du début, et les mettre dans des cages individuelles pour la diète 1. Les 10 restants seraient alors associés à la deuxième diète. L'effet diète est donc ici confondu avec l'ordre de sortie de la cage de départ. Ce sont peut-être les rats les plus actifs qui sont sortis en premier. Ainsi les 2 échantillons ne sont pas nécessairement identiques au début de l'expérience.

**Planification 2** : On utilise un tirage au hasard. Pour ce faire, on permute au hasard dix "1" et dix "2".

Les instructions **R** pour faire cela sont :

```
sample(c(rep(1,10),rep(2,10)),20,replace=FALSE)
 [1] 2 1 1 1 1 2 2 1 1 2 2 1 1 1 2 2 1 2 2 2
```

Le résultat fournit l'assignation de chacun des rats : le premier tiré reçoit la deuxième diète ; ceux tirés en positions 2 à 5 reçoivent la diète 1 ; les positions 6 et 7 reçoivent la diète 2, etc...

Une bonne planification cherche à faire en sorte que les $I$ échantillons soient le plus semblable possible. Si une expérience est mal planifiée, l'interprétation d'un résultat significatif peut être problématique. Il est peut-être causé par une planification déficiente. Dans l'expérience sur les rats, ceux choisis en premier pourraient être plus en forme. C'est peut-être la raison pour laquelle les deux échantillons ont des moyennes différentes.

Si on soupçonne qu'un facteur auxiliaire a un impact sur le résultat d'une expérience, on peut incorporer ce facteur dans la planification pour s'assurer que les échantillons soient bien "balancés" pour ce facteur. Ce facteur auxiliaire est appelé *bloc*. Le schéma expérimental est appelé un *schéma aléatoire avec blocs*.

Il faut aussi veiller à ce que les $I$ échantillons soient indépendants les uns des autres.

Dans la plupart des situations, la réponse à cette question dépend de la façon avec laquelle on a récolté les données. L'indépendance des échantillons, appelée aussi *indépendance inter-échantillonnale*, est donc une conséquence directe du scénario de l'échantillonnage. Une situation standard dans laquelle cette hypothèse est violée, est le cas de données appariées, c'est-à-dire lorsque chaque observation dans un échantillon est reliée à une observation dans chacun des autres échantillons.

**Exemple 4.1.1** *Un chercheur en sciences médicales veut comparer deux médicaments pour réduire le taux de glycémie chez les personnes âgées. Il prend des couples de personnes âgées et administre à chacun des deux membres du couple un des deux médicaments. Les données ainsi récoltées ne sont clairement pas indépendantes puisque les données d'un couple sont reliées entre elles. En effet, le couple partage le quotidien, et il se peut qu'un couple fasse très attention à son alimentation alors qu'un autre couple non, ou peu d'attention.*

Les observations sont-elles identiquement distribuées à l'intérieur de chaque échantillon ?

Ici encore, c'est le plan d'expérience qui permet de répondre à cette question. Une situation standard pour laquelle cette hypothèse n'est pas vérifiée est par exemple lorsque les données sont obtenues séquentiellement dans le temps : d'abord $Y_{i1}$, puis $Y_{i2}$, ensuite $Y_{i3}$, etc... Lorsque la loi des $Y_{ij}$ évolue dans le temps, les données ne sont généralement pas identiquement distribuées. Pour détecter cette situation, on peut effectuer un graphe de $\{Y_{i1}, Y_{i2}, \cdots, Y_{in_i}\}$ en fonction de $j = 1, 2, \cdots, n_i$ pour

$i = 1, 2, \cdots, I$. Si ce graphe montre une tendance quelconque, on peut penser que cette hypothèse n'est pas vérifiée.

Est-ce que les observations sont indépendantes les unes des autres à l'intérieur de chaque échantillon ?

Encore une fois, c'est le schéma expérimental qui rend cette hypothèse raisonnable. Le cas où les données sont récoltées séquentiellement soulève un doute concernant la véracité de cette hypothèse. En effet, il se peut que les données soient autocorrélées, c'est-à-dire que $Y_{ij}$ soit corrélée avec $Y_{i(j+1)}$. On peut détecter cette situation en traçant le nuage de points $(Y_{ij}, Y_{i,j+1})$ pour $j = 1, 2, \cdots, n_i - 1$, ou en calculant les coefficients d'autocorrélation. Pour pouvoir répondre positivement à la question, le nuage de points ne doit montrer aucune tendance et les autocorrélations ne doivent pas être significativement différentes de 0.

## 4.2 Condition de normalité

L'hypothèse de normalité est cruciale pour l'analyse de la variance. En pratique, la validation de cette hypothèse est une étape importante lors de l'analyse.

Nous ne pouvons pas, en général, la tester pour chaque échantillon. En effet le nombre d'observations est souvent très limité pour chaque échantillon. Nous allons donc la tester sur l'ensemble des données. D'où la nécessité de ramener toutes les observations à la même échelle pour avoir une population homogène sur laquelle on va effectuer les différents tests de normalité.

Notons, pour tout $i \in I$ :

$$\mu_i = \mu + \alpha_i \, .$$

Pour $i = 1, 2, \cdots, I$, on a $Y_{ij} \sim \mathcal{N}(\mu_i, \sigma^2)$. D'où :

$$\varepsilon_{i,j} = Y_{i,j} - \mu_i \sim \mathcal{N}(0, \sigma^2)$$

Donc la loi des $\varepsilon_{i,j}$ est identique pour toutes les unités.

Les moyennes $\mu_i$ sont inconnues. On va alors les estimer par les estimateurs :

$$Y_{i,\bullet} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{i,j} \qquad \text{pour tout } i \in I \, .$$

Nous obtenons alors les estimations $y_{i,\bullet}$. On en déduit les résidus, notés $\hat{e}_{i,j}$ . Les résidus s'expriment par :

$$\hat{e}_{i,j} = y_{i,j} - y_{i,\bullet} \qquad i = 1, \cdots I \quad ; \quad j = 1, \cdots, n_i \, ..$$

Les résidus peuvent s'interpréter comme des estimations des erreurs de mesure.

Définissons les résidus $e_{ij}$ par : $Y_{ij} - \mu_i$. On a alors $e_{ij} \sim \mathcal{N}(0, \sigma^2)$, résidus estimés ci-dessus.

### 4.2.1 Les cœfficients d'asymétrie et d'aplatissement

On peut déjà examiner les coefficients d'asymétrie et d'aplatissement.

Le cœfficient d'asymétrie (skewness) de l'échantillon $\{X_1, \ldots, X_n\}$ est donné par :

$$g_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right)^{3/2}}.$$

Certains logiciels calculent plutôt un estimateur corrigé pour le biais :

$$G_1 = \frac{\sqrt{n\,(n-1)}}{n-2}\, g_1\,.$$

Le cœfficient d'aplatissement (kurtosis) est donné par :

$$g_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right)^{2}} - 3.$$

Certains logiciels calculent un estimateur corrigé pour le biais. La valeur théorique de ces deux statistiques est nulle lorsque les données sont normales.

### 4.2.2 La droite d'Henry

Ce test purement visuel est basé sur le nuage de $n$ points $(\Phi^{-1}(i/(n+1)), X_{(i)})$, où $\{X_{(1)}, \cdots, X_{(n)}\}$ est l'échantillon de statistiques d'ordre obtenu en ordonnant l'échantillon initial.

En effet, si l'échantillon $\{X_1, X_2, \cdots, X_n\}$ provenait d'une loi normale, on aurait $F(\cdot) = \Phi(\cdot)$ et $\Phi(X_{(i)}) \simeq \hat{F}(X_{(i)}) = i/n$ qu'on peut écrire encore $\Phi^{-1}[i/n] \simeq X_{(i)}$.

Le nuage de points sera aligné globalement alors sur une droite. On utilise $i/(n+1)$ à la place de $i/n$ pour éviter $\Phi^{-1}(0)$, qui n'existe pas.

### 4.2.3 Le test de Shapiro et Wilk

Ce test est une approche plus approfondie du test précédent. Supposons que le nuage de points $\{X_{(i)}, \Phi^{-1}[i/(n+1)]\}$ soit aligné sur une droite, alors le cœfficient de corrélation défini par

$$r = \frac{\sum_{i=1}^{n}(u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\sum_{i=1}^{n}(u_i - \bar{u})^2 \sum_{i=1}^{n}(v_i - \bar{v})^2}}$$

 

où $u_i = X_{(i)}$ et $v_i = \Phi^{-1}[i/(n+1)]$, ne sera pas loin de 1.

Ceci équivaut à dire que $r^2$ ne sera pas loin de 1. On rejette alors la normalité si $r^2$ est loin de 1. Il existe des tables pour la distribution de $r^2$ sous $H_0$. Ces tables nous servent à calculer la valeur critique à un seuil donné, par exemple à 5% et à calculer la valeur critique $(p-value)$ associée à un jeu de données.

### 4.2.4 Le test de Kolmogorov-Smirnov

Le test de Kolmogorov-Smirnov est basé sur une distance entre la fonction de répartition empirique $\hat{F}(\cdot)$ et la fonction de répartition qu'on veut tester, ici $\Phi(\cdot)$.

Si l'échantillon $\{X_1, X_2, \cdots, X_n\}$ provient d'une loi normale, on devrait avoir $\hat{F}(t) \simeq \Phi(t)$ pour tout réel $t$.

En particulier, la statistique $D$ définie par

$$D = \sup_{t \in \mathcal{R}} |\hat{F}_n(t) - \Phi(t)|$$

doit être petite.

Le test de Kolmogorov-Smirnov consiste donc à rejeter la normalité si la statistique $D$ est trop grande. Il existe des tables pour la loi $D$ sous $H_0$. Ces tables nous servent à calculer la valeur critique à un seuil donné, par exemple à 5% et à calculer la valeur critique $(p-value)$ associée à un jeu de données.

## 4.3 Condition d'homogénéité des variances

La vérification de l'hypothèse d'homogénéité des variances est une étape importante lors de la réalisation d'une ANOVA. Il existe dans la littérature plusieurs procédures pour effectuer le test $H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_I^2$ contre $H_1$ : les variances ne sont pas toutes égales.

Cependant, plusieurs de ces tests requièrent la normalité ou l'égalité des tailles des échantillons $(n_1 = n_2 = \cdots = n_I)$. Dans ce chapitre, on présente les tests les plus utilisés dans la pratique.

### 4.3.1 Le test de Levene

Le test de Levene date du début des années soixante. Il a pour principe de calculer séparément pour les différents échantillons, les écarts par rapport aux moyennes, et de soumettre les valeurs absolues de ces écarts à l'analyse de la variance à un facteur. L'hypothèse d'égalité des moyennes des valeurs absolues des écarts, qui est testée par l'analyse de la variance, est alors considérée comme équivalent à l'hypothèse d'égalité des variances. Cette méthode a l'avantage d'être plus robuste que les tests de Bartlett et Hartley. Elle n'est toutefois qu'approchée, du fait que, d'une part, les écarts par rapport aux moyennes ne sont pas indépendants les uns des autres, en particulier, dans le cas de très petits échantillons, et d'autre part, les valeurs absolues des écarts ne possèdent pas, elles-mêmes, des lois normales, ce que suppose l'analyse de variance.

On effectue donc une analyse de la variance sur des données transformées. Pour $i = 1, 2, \cdots, I$ et $j = 1, 2, \cdots, n_i$, définissons $Z_{i,j}$ par $Z_{i,j} = |Y_{i,j} - Y_{i,\bullet}|$. Le test de Levene consiste à effectuer une ANOVA sur les variables transformées $Z_{i,j}$.

Ainsi, on rejette l'hypothèse d'homogénéité des variances au seuil $\alpha$ si $F_{obs} > F_{\alpha,I-1,N-I}$ où $F_{obs}$ est défini par :

$$F_{obs} = \frac{\sum_{i=1}^{I} n_i (Z_{i,\bullet} - Z_{\bullet,\bullet})^2/(I-1)}{\sum_{i=1}^{I}\sum_{j=1}^{n_i} (Z_{i,j} - Z_{i,\bullet})^2/(n-I)}.$$

### 4.3.2 Le test de Brown et Forsythe

Ce test est une variante du test précédent. Ici, on définit les $Z_{i,j}$ par $|Y_{i,j} - \tilde{Y}_{i,\bullet}|$ où $\tilde{Y}_{i,\bullet}$ est la médiane de l'échantillon $\{Y_{i1}, Y_{i2}, \cdots, Y_{in_i}\}$. Le reste de la procédure demeure inchangée.

### 4.3.3 Le test de Bartlett

Le test de Bartlett nécessite des calculs relativement longs, mais s'applique indifféremment à des échantillons d'effectifs égaux ou inégaux. Pour $I$ populations, l'hypothèse nulle s'écrit :

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_I^2 .$$

Le test de Bartlett, considéré comme un test de rapport de vraisemblance est basé sur la statistique $L$ défini par :

$$L = \frac{(S_1^2)^{\frac{n_1-1}{n-I}} (S_2^2)^{\frac{n_2-1}{n-I}} \cdots (S_I^2)^{\frac{n_I-1}{n-I}}}{\frac{n_1-1}{n-I}S_1^2 + \frac{n_2-1}{n-I}S_2^2 + \cdots + \frac{n_I-1}{n-I}S_I^2}.$$

Le dénominateur et le numérateur de la statistique $L$ définis par l'équation ci-dessus sont les moyennes arithmétiques et géométriques respectives de $\{S_1^2, S_2^2, \cdots, S_I^2\}$ pondérées par $w_1 = (n_1 - 1)/(n-I), w_2 = (n_2-1)/(n-I), \cdots, w_I = (n_I-1)/(n-I)$. Ces poids vérifient : $w_1+w_2+\cdots+w_I = 1$.

On rejette l'hypothèse d'homogénéité des variances si $L$ est trop grand. Il existe des tables pour la distribution exacte de $L$. Néanmoins, en pratique, on utilise l'approximation qui suit. Posons :

$$B = \frac{-(n-I)\log(L)}{c}$$

avec

$$c = 1 + \frac{(\sum_{i=1}^{I} \frac{1}{n_i - 1}) - \frac{1}{n-I}}{3(I-1)}.$$

Sous $H_0$, lorsque les tailles des échantillons $n_1, n_2, \cdot, n_I$ tendent vers l'infini, on obtient asymptotiquement :

$$B \sim \chi^2_{I-1}.$$

On rejette donc $H_0$ si $B > \chi^2_{I-1,\alpha}$.

 

## 4.4 Résumé et commentaires

Il y a donc trois hypothèses à vérifier :

1. L'indépendance intra et inter échantillons ;
2. La normalité des erreurs ;
3. L'égalité des variances.

Pour la première hypothèse on cherche à vérifier la bonne planification de l'expérience. S'agit-il d'une expérience complètement aléatoire ? Quelles sont les unités expérimentales ? Peut-être que l'ordre dans lequel les données ont été récoltées est associé à leur valeur.

L'hypothèse de normalité n'est pas cruciale pour la validité du test $F$ d'homogénéité des moyennes. Si les tailles d'échantillons $n_i$ sont grandes, la loi de la statistique $F$ de la table ANOVA est approximativement une loi de Fisher-Snedecor $F_{I-1,\sum n_i}$, même si les données ne sont pas normales. La non normalité des données compromet cependant la puissance du test.

Si cette hypothèse de normalité est violée, le test non paramétrique de Kruskal-Wallis, basé sur les rangs, est souvent plus puissant que le test $F$ de la table ANOVA. C'est le cas lorsque les données contiennent des valeurs extrêmes (outliers). On peut d'ailleurs s'assurer que quelques valeurs extrêmes n'ont pas une influence sur les résultats en refaisant les analyses après avoir exclu ces données.

L'égalité des variances non plus n'est pas vraiment cruciale pour la validité du test $F$ d'homogénéité. Si l'expérience est à peu près balancée et si les tailles d'échantillons sont grandes, on peut montrer que le test $F$ est valide même si les variances sont inégales. Comme on l'a vu, le test de Welch tient compte des variances inégales.

Dans une analyse de variance, la variabilité des données ne doit pas être associée à leurs valeurs moyennes. Par exemple, si les données sont des dénombrements avec une loi de Poisson, alors la variance est à peu près égale à la moyenne. Ainsi, il y a un lien moyenne-variance, et les hypothèses sous-jacentes à l'ANOVA sont violées. Dans ce cas, deux solutions sont possibles. On peut faire une transformation pour chercher à stabiliser la variance et traiter les données avec une ANOVA (pour la loi de Poisson, la transformation racine carrée s'avère utile).

On peut également utiliser un modèle linéaire généralisé construit spécifiquement pour la loi de Poisson.

Le mode de variation de certaines variables, que l'on décrit en termes relatifs plutôt qu'en termes absolus, peut aussi suggérer une transformation. Un modèle ANOVA postule des variations absolues additives. Des variations relatives sont en fait multiplicatives. Une transformation logarithmique peut alors s'imposer pour que les hypothèses du modèles ANOVA soient vérifiées. Mais une transformation complique l'analyse, car c'est souvent sur l'échelle originale que les résultats doivent être interprétés.

## 4.5 Un exemple détaillé

On considère le fichier de données "ozone" disponible sur le site web du cours. Il s'agit de la pollution de l'air. De nombreuses études épidémiologiques ont mis en évidence l'influence sur la

santé de certains composants chimiques, comme le dioxyde soufre ($SO_2$), le dioxyde d'azote ($NO_2$), l'ozone($O_3$) et quelques autres particules flottant dans l'air.

Des stations de surveillance enregistrent les conditions météorologiques comme la température, la nébulosité, le vent, etc... Nous allons analyser la relation existant entre le maximum journalier de la concentration en ozone (en $\mu g/m^3$) et la direction du vent (classée en quatre secteurs : Nord, Sud, Est, Ouest). La variable **vent** possède donc 4 modalités. Pour cette étude, le fichier "ozone" dispose de 112 données relevées durant l'été 2001 à Rennes en France. On utilise le logiciel *R*.

Les différentes étapes sont les suivantes :

1. Importation des données :
   On importe le jeu de données et on va résumer les variables d'intérêts :
   ozone<-read.table("ozone.txt",header=T)
   attach(ozone)
   summary(ozone[,c("maxO3","vent")])

```
      maxO3           vent
 Min.   : 42.00   Est  :10
 1st Qu.: 70.75   Nord :31
 Median : 81.50   Ouest:50
 Mean   : 90.30   Sud  :21
 3rd Qu.:106.00
 Max.   :166.00
```

   Pendant l'été, le vent dominant est le vent d'Ouest, et il y a peu de journées avec un vent d'Est.

2. Représentation des données : On va tracer ci-dessous les boîtes à moustaches pour chacune des modalités de la variable qualitative, c'est-à-dire qu'on représente la dispersion de la variable **maxO3** en fonction de la direction du vent :
   plot(maxO3 vent, data=ozone, pch=15,cex=0.5,col="green")
   En examinant le graphique ci-dessus, il semble bien qu'il y ait un effet **vent**.

3. Analyse de l'homogénéité :
   On cherche à tester l'égalité des variances, à savoir :
   $H_0$ : $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$ contre $H_1$ : il existe au moins deux variances non égales.
   Le test de Levene consiste à réaliser une analyse de variance sur les valeurs absolues des résidus :

```
model<-aov(maxO3~vent)

 summary(aov(abs(model$res) ~ vent))
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
vent          3    671   223.6   0.705  0.551
Residuals   108  34231   317.0
```

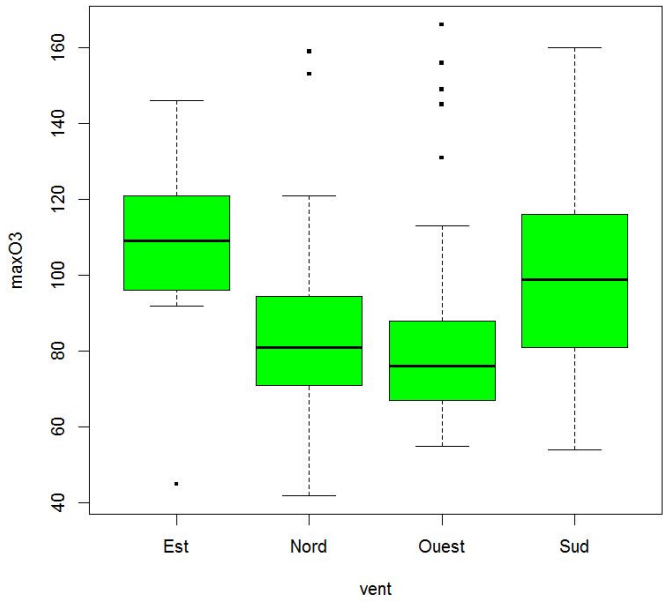


FIGURE 4.1 – Boîtes à moustaches de **maxO3** selon les modalités de la variable **vent**

On ne rejette donc pas l'hypothèse de l'égalité des variances, et on conclut significativement à l'homogénéité des variances.

On peut aussi utiliser le test de Bartlett :

```
bartlett.test(model$res ~ vent)

        Bartlett test of homogeneity of variances

data:  model$res by vent
Bartlett's K-squared = 0.9981, df = 3, p-value = 0.8017
```

On conclut là aussi à l'homogénéité significative des variances.

4. Analyse de la normalité :
On peut tester la normalité de chaque sous-population grâce au test de Shapiro-Wilk. On l'illustre ici sur la sous-population de "**vent**=Est" :

```
select.est <-ozone[,"vent"]=="Est"

shapiro.test(ozone[select.est,"maxO3"])

        Shapiro-Wilk normality test

data:  ozone[select.est, "maxO3"]
W = 0.9184, p-value = 0.344
```

La probabilité critique étant supérieure à 5%, on accepte l'hypothèse de normalité de la sous-population du taux d'ozone par vent d'Est.

5. Analyse de la variance :

   On peut lancer l'analyse de la variance pour tester la significativité du facteur vent :

```
ozone.aov <- aov(maxO3~vent)

summary(ozone.aov)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
vent          3   7586  2528.7   3.388 0.0207 *
Residuals   108  80606   746.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

   La première colonne fournit les degrés de liberté associés au facteur, la seconde, la somme des carrés, la troisième, le carré moyen (qui est la somme des carrés divisée par les degrés de liberté). La quatrième colonne indique la valeur observée de la statistique de test. La cinquième colonne donne la probabilité critique, c'est-à-dire la probabilité pour la statistique de test sous $H_0$ de dépasser la valeur estimée.

   Ici, la probabilité critique vaut 0,0207 qui est inférieure à 5%, et donc, on rejette l'hypothèse nulle d'égalité des moyennes, c'est-à-dire qu'on rejette $H_0$. On conclut donc à la significativité du facteur **vent** au niveau 5% pour expliquer le taux d'ozone. Il existe donc au moins une direction de vent pour laquelle le maximum d'ozone est significativement différent des autres.

6. Analyse des résidus :

   Les résidus sont disponibles en sortie de la fonction **aov**, mais ces derniers ne sont pas de même variance. Nous allons donc les studentiser :

   res.ozone <- rstudent(ozone.aov)

   Pour tracer les graphes des résidus selon les quatre modalités de la variable **vent**, on utilise le package *lattice* :

   >library(lattice)

   > monpanel <- function (...)

   + panel.xyplot(...)

   + panel.abline(h=c(-2,0,2),lty=c(3,2,3),...)

   > trellis.par.set(list(fontsize=list(point=5,text=8)))

   > xyplot(res.ozone I(1 :112)|vent,data=ozone,pch="+",ylim=c(-3,3),panel=monpanel, ylab="Résidus", xlab="")

   Théoriquement, 95% des résidus studentisés se trouvent dans l'intervalle $[-2, 2]$. Ici, neuf résidus se situent à l'extérieur de l'intervalle, soit environ 8% du total, ce qui est acceptable.

7. Interprétation des coefficients :

   Pour aller plus loin encore, après avoir constaté qu'il y a un effet du vent sur la quantité d'ozone, on aimerait préciser comment la direction du vent influe sur le maximum d'ozone. Nous utilisons cette fois la fonction **lm** :

```
summary(lm(maxO3~C(vent,sum),data=ozone))
```

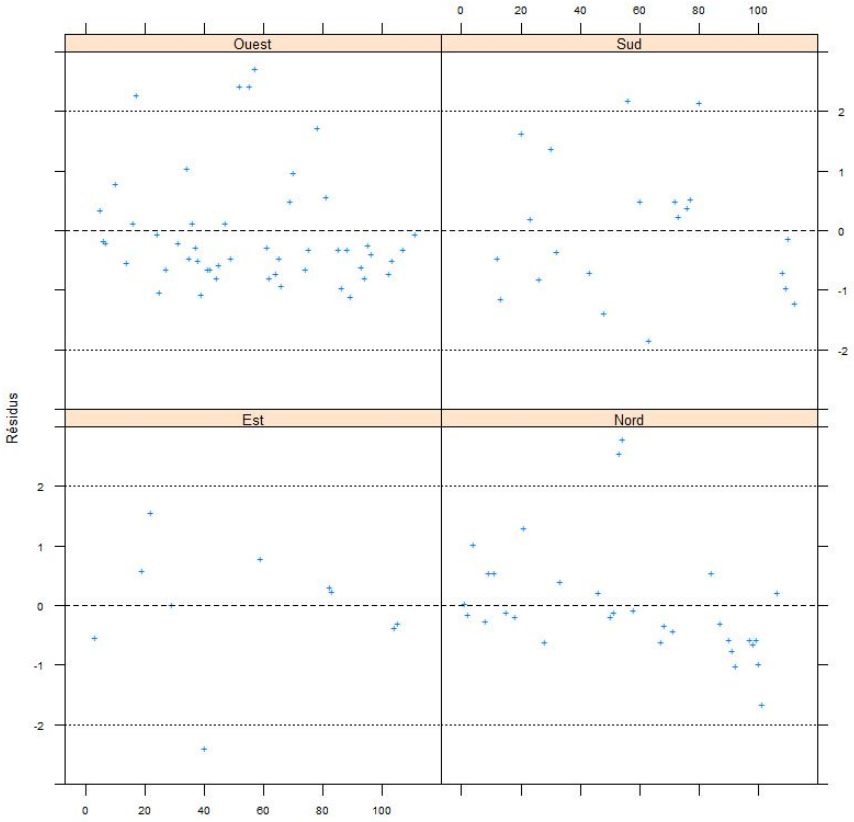


FIGURE 4.2 – Représentation des résidus selon les modalités de la variable **vent**

```
Call:
lm(formula = maxO3 ~ C(vent, sum), data = ozone)

Residuals:
    Min      1Q  Median      3Q     Max
-60.600 -16.807  -7.365  11.478  81.300

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     94.738      3.053  31.027   <2e-16 ***
C(vent, sum)1   10.862      6.829   1.590   0.1147
C(vent, sum)2   -8.609      4.622  -1.863   0.0652 .
C(vent, sum)3  -10.038      4.097  -2.450   0.0159 *
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 27.32 on 108 degrees of freedom
Multiple R-squared:  0.08602,   Adjusted R-squared:  0.06063
```

```
F-statistic: 3.388 on 3 and 108 DF,  p-value: 0.02074
```

Le logiciel $R$ nous fournit les valeurs de $\hat{\mu}, \hat{\alpha}_1, \hat{\alpha}_2$ et $\hat{\alpha}_3$ (les vents sont numérotés suivant l'ordre alphabétique). Or, comme $\sum_{i=1}^{4} \alpha_i = 0$, pour trouver le coefficient $\hat{\alpha}_4$, le coefficient associé au vent Sud, il faut calculer :

$$\hat{\alpha}_4 = -\hat{\alpha}_1 - \hat{\alpha}_2 - \hat{\alpha}_3 = 7,785\,.$$

La valeur de $\hat{\mu}$, notée ici *Intercept*, est la moyenne globale de la concentration en **maxO3**. Les autres valeurs sont les écarts à cette moyenne pour la modalité de vent considéré. Le vent d'ouest est significativement différent de la moyenne globale. C'est donc à cause de lui qu'on a une différence significative entre les taux d'ozone suivant la direction du vent.

# Exercices

**Exercice 1**

Validez les hypothèses du modèle dans l'exercice 2 du chapitre 3.

**Exercice 2**

Validez les hypothèses du modèle dans l'exercice 3 du chapitre 3.

**Exercice 3**

Validez les hypothèses du modèle dans l'exercice 4 du chapitre 3.

**Exercice 4**

Validez les hypothèses du modèle dans l'exercice 5 du chapitre 3.

**Exercice 5**

Validez les hypothèses du modèle dans l'exercice 6 du chapitre 3.

**Exercice 6**

Validez les hypothèses du modèle dans l'exercice 7 du chapitre 3.

**Exercice 7**

Nous voulons tester quatre types de carburateurs : A1, A2, A3 et A4. Pour chaque type de carburateur, nous disposons de six pièces qui sont montées successivement en parallèle sur quatre voitures que nous supposons avoir des caractéristiques parfaitement identiques. Le tableau ci-dessous indique pour chacun des essais la valeur d'un paramètre lié à la consommation :

| **Essai** | $A_1$ | $A_2$ | $A_3$ | $A_4$ |
|---|---|---|---|---|
| 1 | 21 | 23 | 18 | 20 |
| 2 | 24 | 23 | 19 | 21 |
| 3 | 25 | 32 | 28 | 25 |
| 4 | 20 | 23 | 19 | 15 |
| 5 | 34 | 32 | 24 | 29 |
| 6 | 17 | 15 | 14 | 9 |

1. Faites une boîte à moustaches de la consommation par carburateurs.
2. Proposez une méthode statistique permettant d'étudier l'influence des modalités du facteur *carburateur* sur la *consommation*. Énoncez le modèle et les hypothèses nécessaires au modèle que vous projetez d'utiliser. Ce modèle comporte-t-il des répétitions ?
3. Validez les hypothèses du modèle.
4. Y a-t-il des différences entre les carburateurs ? Quelles sont les estimations des coefficients du modèle ? Si nécessaire, comparez les différents niveaux du facteur *carburateur*.

 

# Chapitre 5

# Comparaisons multiples

Si, après avoir effectué une analyse de variance, on rejette l'hypothèse d'égalité des moyennes relatives à un facteur $A$ à $I$ modalités, une question intéressante est de savoir quelles sont les moyennes qui diffèrent significativement des autres.

Reprenons l'exercice 1 du chapitre 3 sur les boutons de fièvre. On aimerait savoir quel est le traitement le plus efficace, à savoir celui qui permet d'obtenir une cicatrisation la plus rapide.

Le test individuel de Student dans le modèle linéaire est parfaitement valide pour comparer deux traitements choisis a priori. Par contre, il n'est pas du tout utilisable pour comparer par exemple le traitement qui donne en apparence les résultats les meilleurs avec celui qui donne en apparence les résultats les plus mauvais. Cela revient en effet à comparer tous les traitements deux à deux. Chaque test a alors une probabilité $\alpha$ (niveau du test) de déclarer présente une différence qui n'existe pas. Au total, sur les $I(I-1)/2$ comparaisons possibles, la probabilité d'en déclarer une significative "par hasard" devient importante. Pour contrôler un risque global sur les $I(I-1)/2$ comparaisons deux à deux, il existe diverses méthodes.

## 5.1 Contrastes

### 5.1.1 Définition

Pour introduire la notion de contraste, on considère le cas d'un modèle d'analyse de la variance pour un facteur $A$ à effets fixes. Nous noterons $A_i$, pour $i = 1, \cdots, I$, les modalités contrôlées du facteur $A$, et $\alpha_i$ les effets de ces différentes modalités.

Reprenons le modèle :

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

avec la contrainte supplémentaire : $\displaystyle\sum_{i=1}^{I} \alpha_i = 0 \, ,$

où $Y_{i,j}$ est la valeur prise par la variable réponse $Y$ dans la condition $A_i$ lors de la $j$-ème répétition.

Nous postulerons les hypothèses classiques suivantes pour les erreurs :

1. $\varepsilon_{i,j}$ et $\varepsilon_{k,l}$ sont indépendantes si $(i,j) \neq (k,l)$ avec $1 \leq i,k \leq I$ et $1 \leq j,l \leq J$.
2. $\forall\,(i,j)$, $1 \leq i \leq I$, $1 \leq j \leq J$, $\mathcal{L}(\varepsilon_{i,j}) = \mathcal{N}(0,\sigma^2)$.

**Définition 5.1.1**

*Nous appelons contraste $L$ des $I$ moyennes $\mu_1, \cdots, \mu_I$ la quantité :*

$$L = l_1\mu_1 + l_2\mu_2 + \cdots, +l_I\mu_I\,,$$

*où $l_1, \cdots, l_I$ sont $I$ nombres réels tels que : $\sum_{i=1}^{I} l_i = 0$ et $\mu_1 = \mu + \alpha_1, \cdots, \mu_I = \mu + \alpha_I$ sont tels que : $\alpha_1, \cdots, \alpha_I$ sont les $I$ différents effets des $I$ niveaux du facteur $A$.*

Les expressions suivantes sont des exemples de contrastes :
$\mu_1 - \mu_3$ est un contraste qui permettra par exemple de comparer $\mu_1$ et $\mu_3$.
$\mu_1 - 2\mu_2 + \mu_3$ est un contraste qui permettra par exemple de comparer $\mu_1 + \mu_3$ et $2\mu_2$.

### 5.1.2 Orthogonalité

Considérons deux contrastes $L_1$ et $L_2$ définis par :

$$L_1 = l_1\mu_1 + l_2\mu_2 + \cdots, +l_I\mu_I\,,$$

et

$$L_2 = l'_1\mu_1 + l'_2\mu_2 + \cdots, +l'_I\mu_I\,,$$

Nous avons bien entendu : $\sum_{i=1}^{I} l_i = 0$ et $\sum_{i=1}^{I} l'_i = 0$ et $\mu_1 = \mu + \alpha_1, \cdots, \mu_I = \mu + \alpha_I$ sont tels que : $\alpha_1, \cdots, \alpha_I$ sont les $I$ différents effets des $I$ niveaux du facteur $A$.

**Définition 5.1.2**

*$L_1$ et $L_2$ sont des contrastes dits orthogonaux si et seulement si la relation suivante est vérifiée :*

$$l_l l'_1 + l_2 l'_2 + \cdots + l_I l'_I = 0\,.$$

Par exemple les deux contrastes suivants sont des contrastes orthogonaux :

$$L_1 = \mu_1 - \mu_2 \quad \text{et} \quad L_2 = \mu_1 + \mu_2 - \mu_3 - \mu_4\,.$$

### 5.1.3 Estimation

Soit $L$ un contraste. Un estimateur sans biais $\hat{L}$ de ce contraste $L$ est obtenu de la manière suivante :

$$\hat{L} = l_1\hat{\mu}_1 + l_2\hat{\mu}_2 + \cdots, +l_I\hat{\mu}_I\,,$$

où $\hat{\mu}_i = \widehat{\mu + \alpha_i} = \hat{\mu} + \hat{\alpha}_i$, avec $1 \leq i \leq I$.

 

La somme des carrés associés à l'estimateur $\hat{L}$ du contraste $L$ est définie par :

$$SC_{\hat{L}} = J \frac{\left(\sum_{i=1}^{I} l_i \hat{\mu}_i\right)^2}{\sum_{i=1}^{I} l_i^2} .$$

Si le nombre de répétitions diffère d'un niveau $A_i$ du facteur $A$ à l'autre, en notant $n_i$ le nombre de répétitions pour la modalité $A_i$, on obtient la formule modifiée suivante :

$$SC_{\hat{L}} = \frac{\left(\sum_{i=1}^{I} l_i \hat{\mu}_i\right)^2}{\sum_{i=1}^{I} \frac{l_i^2}{n_i}} .$$

### 5.1.4 Test d'une hypothèse impliquant un contraste

Pour le modèle d'analyse de la variance à un facteur, les estimateurs $Y_{i,\bullet}$ des $\mu_i$ sont indépendants, même si le plan n'est pas équilibré. Donc, la variance de l'estimateur $\hat{L}$ du contraste $L$ vaut :

$$var\left[\hat{L}\right] = \sum_{i=1}^{I} \left(l_i^2 var\left[Y_{i,\bullet}\right]\right)^2 = \sigma^2 \sum_{i=1}^{I} \frac{l_i^2}{n_i} .$$

Un estimateur sans biais de cette variance est donc :

$$\widehat{var\left[\hat{L}\right]} = s_R^2 \sum_{i=1}^{I} \frac{l_i^2}{n_i} .$$

Les hypothèses du modèle utilisé impliquent alors que, puisque $\hat{L}$ est une combinaison linéaire de variables aléatoires indépendantes qui suivent une loi normale, l'estimateur $\hat{L}$ suit aussi une loi normale. Donc :

$$\frac{\hat{L} - L}{\sqrt{\widehat{var\left[\hat{L}\right]}}} \sim t_{n-I} ,$$

où $t_{n-I}$ est la loi de Student à $n - I$ degrés de liberté.

**Remarque 5.1.1**

*Le résultat sur la loi de* $\dfrac{\hat{L} - L}{\sqrt{\widehat{var\left[\hat{L}\right]}}}$ *permet de déterminer un intervalle de confiance de niveau* $100(1-\alpha)\%$ *pour la valeur du contraste* $L$.

Donnons-nous désormais un réel $L_0$. Nous désirons tester l'hypothèse :

$$H_0 \, : \, L = L_0$$

contre l'hypothèse alternative :

$$H_1 \, : \, L \neq L_0 \, .$$

Sous $H_0$, et les hypothèses du modèle,

$$l = \frac{\hat{L}(\boldsymbol{y}) - L_0}{\sqrt{s_R^2 \sum_{i=1}^{I} \frac{l_i^2}{n_i}}}$$

est une réalisation d'une variable aléatoire suivant une loi de Student à $n - I$ degrés de liberté. En comparant la valeur $l$ calculée à partir d'un échantillon à la valeur critique au seuil $\alpha$ pour une loi de Student à $n - I$ degrés de liberté, nous pouvons décider de la significativité du test. Certains logiciels fournissent directement la probabilité critique associée au test d'un contraste, ce qui permet aussi de conclure quant à la significativité du test.

## 5.2 Comparaisons multiples sous l'hypothèse d'homoscédasticité

Dans le premier paragraphe, nous avons étudié le test d'une seule hypothèse concernant les effets des niveaux du facteur $A$. Le contexte des tests de comparaisons multiples est totalement différent car on cherche alors à comparer tous les effets des niveaux $A_i$ du facteur $A$ entre eux ou avec un niveau de référence dit de contrôle.

On doit donc réaliser $I(I-1)/2$ comparaisons dans la première situation ou $I-1$ dans la seconde situation où nous comparons les effets à un niveau de contrôle fixé a priori.

Tester l'égalité des effets de deux niveaux $A_i$ et $A_j$, pour $i \neq j$, d'un facteur $A$ revient à tester la nullité du contraste $L = \mu_i - \mu_j$. Nous allons détailler dans ce qui suit les procédures de tests simultanés de plusieurs contrastes en gardant à l'esprit que nous appliquerons principalement les résultats dans le cas où ces contrastes sont des différences de moyennes.

On rappelle cependant que nous n'utiliserons l'un des tests de comparaisons multiples que si le facteur étudié est à effets fixes et quand nous avons rejeté l'hypothèse nulle d'absence d'effet de ce facteur sur la réponse $Y$.

Nous détaillons ici la théorie des comparaisons multiples pour le cas d'un modèle à un facteur à effets fixes. Plus généralement, il est possible de comparer les effets des différents niveaux d'un facteur si ceux-ci sont à effets fixes. À noter qu'il n'est généralement intéressant de comparer les effets des différents niveaux d'un facteur que si aucun des termes d'interaction mettant en jeu ce facteur n'a un effet significatif au seuil $\alpha$.

### 5.2.1 La méthode de Tukey

Cette méthode n'est valable que si le nombre de répétitions $J_i$ d'une modalité à l'autre du facteur $A$ est constant. Ce nombre commun de répétitions est alors noté $J$. Pour une version de la méthode de Tukey adaptée au cas où le plan n'est pas équilibré, on pourrait voir le paragraphe suivant sur la méthode de Tukey-Kramer.

Soit $L$ un contraste dont un estimateur est $\hat{L}$. Un intervalle de confiance de niveau simultané $100(1-\alpha)\,\%$ pour tous les contrastes considérés est donné par la formule :

$$\hat{L}(\boldsymbol{y}) - T\sqrt{\frac{s_R^2}{J}}\left(\frac{1}{2}\sum_{i=1}^{I}|l_i|\right) < L < \hat{L}(\boldsymbol{y}) + T\sqrt{\frac{s_R^2}{J}}\left(\frac{1}{2}\sum_{i=1}^{I}|l_i|\right)\,,$$

où $T = q(I, I(J-1); 1-\alpha)$ est le $100(1-\alpha)$ quantile de la loi de l'étendue studentisée à $I$ et $I(J-1)$ degrés de liberté. Si l'intervalle de confiance obtenu contient la valeur 0, nous décidons que le contraste n'est pas significativement différent de 0 au seuil $\alpha$. A contrario, si l'intervalle de confiance ne contient pas 0, alors on décide que le contraste est significativement différent de 0 au seuil $\alpha$.

L'intérêt de ce procédé est que, si nous fixons un seuil $\alpha$, les intervalles définis ci-dessus sont valables simultanément pour tous les contrastes qu'il est possible de construire !

Nous désirons tester l'hypothèse :

$$H_0 \,:\, L = 0$$

contre l'hypothèse alternative :

$$H_0 \,:\, L \neq 0\,.$$

Le test est significatif au seuil $\alpha$ et alors nous décidons de rejeter l'hypothèse nulle $H_0$ en faveur de l'hypothèse alternative $H_1$ si :

$$\frac{\left|\hat{L}(\boldsymbol{y})\right|}{T\sqrt{\frac{s_R^2}{J}}\left(\frac{1}{2}\sum_{i=1}^{I}|l_i|\right)} \geq q(I, I(J-1); 1-\alpha)\,.$$

Le test n'est pas significatif au seuil $\alpha$ et alors nous décidons de conserver par défaut l'hypothèse nulle $H_0$ si :

$$\frac{\left|\hat{L}(\boldsymbol{y})\right|}{T\sqrt{\frac{s_R^2}{J}}\left(\frac{1}{2}\sum_{i=1}^{I}|l_i|\right)} < q(I, I(J-1); 1-\alpha)\,.$$

Appliqué au contexte des comparaisons multiples ( procédure souvent appelée "Tukey'HSD" pour "Tukey's Honestly Significance Difference"), l'intervalle de confiance ci-dessus se transforme de la manière suivante puisque les contrastes étudiés sont du type : $L = \mu_i - \mu_j$ avec $i \neq j$ :

$$\frac{1}{2}\sum_{i=1}^{I}|l_i| = \frac{1}{2}(|1| + |-1|) = \frac{1}{2}(1+1) = 1\,.$$

Les intervalles de confiance se simplifient alors en :

$$\hat{\mu}_i(\boldsymbol{y}) - \hat{\mu}_{i'}(\boldsymbol{y}) - T\sqrt{\frac{s_R^2}{J}} < \mu_i - \mu_{i'} < \hat{\mu}_i(\boldsymbol{y}) - \hat{\mu}_{i'}(\boldsymbol{y}) + T\sqrt{\frac{s_R^2}{J}}\,.$$

Les hypothèses et les statistiques des tests se transforment, quant à elles, en :

$$H_0 \,:\, \mu_i = \mu_{i'}$$

contre l'hypothèse alternative :

$$H_0 \,:\, \mu_i \neq \mu_{i'}\,.$$

Rappelons que ces hypothèses sont équivalentes aux suivantes :

$$H_0 \,:\, \alpha_i = \alpha_{i'}$$

contre l'hypothèse alternative :

$$H_1 \,:\, \alpha_i \neq \alpha_{i'}\,.$$

Le test est significatif au seuil $\alpha$ et on décide de rejeter l'hypothèse nulle $H_0$ en faveur de l'hypothèse alternative $H_1$ si :

$$\frac{|\hat{\mu}_i(\boldsymbol{y}) - \hat{\mu}_{i'}(\boldsymbol{y})|}{\sqrt{\frac{s_R^2}{J}}} \geq q(I, I(J-1); 1-\alpha)\,.$$

Le test n'est pas significatif au seuil $\alpha$ et on décide de conserver par défaut l'hypothèse nulle $H_0$ en faveur de l'hypothèse alternative $H_1$ si :

$$\frac{|\hat{\mu}_i(\boldsymbol{y}) - \hat{\mu}_{i'}(\boldsymbol{y})|}{\sqrt{\frac{s_R^2}{J}}} < q(I, I(J-1); 1-\alpha)\,.$$

En utilisant ces intervalles de confiance pour décider simultanément de la significativité des $I(I-1)/2$ différences entre les effets des modalités du facteur $A$, nous sommes assurés que la probabilité qu'aucune des différences n'est significative est exactement de valeur $1-\alpha$.

### 5.2.2 La méthode de Tukey-Kramer

On rappelle tout d'abord que la moyenne harmonique de deux réels $x$ et $y$ est :

$$Harm(x,y) = \frac{1}{\frac{1}{2}\left(\frac{1}{x} + \frac{1}{y}\right)}$$

Il s'agit d'une adaptation de la méthode de Tukey au cas où le plan expérimental n'est pas équilibré. Nous désirons comparer deux moyennes $\mu_i$ et $\mu_{i'}$, et nous remplaçons alors simplement la valeur $J$ correspondant au nombre total constant d'essais réalisés dans les conditions des modalités $A_i$ du facteur $A$ par la moyenne harmonique du nombre de répétitions effectuées dans la modalité $A_i$ et dans la modalité $A_{i'}$.

Les intervalles de confiance précédents se modifient en conséquence :

$$\hat{\mu}_i(\boldsymbol{y}) - \hat{\mu}_{i'}(\boldsymbol{y}) - T\sqrt{\frac{s_R^2}{2}\left(\frac{1}{n_i}+\frac{1}{n_{i'}}\right)} < \mu_i - \mu_{i'} < \hat{\mu}_i(\boldsymbol{y}) - \hat{\mu}_{i'}(\boldsymbol{y}) + T\sqrt{\frac{s_R^2}{2}\left(\frac{1}{n_i}+\frac{1}{n_{i'}}\right)},$$

où $T = q(I, n-I; 1-\alpha)$ est le $100(1-\alpha)$ quantile de la loi de l'étendue studentisée à $I$ et $n-I$ degrés de liberté.

Les hypothèses sont toujours : Les hypothèses et les statistiques des tests se transforment, quant à elles, en :

$$H_0 \,:\, \mu_i = \mu_{i'}$$

contre l'hypothèse alternative :

$$H_0 \,:\, \mu_i \neq \mu_{i'}\,.$$

Rappelons que ces hypothèses sont équivalentes aux suivantes :

$$H_0 \,:\, \alpha_i = \alpha_{i'}$$

contre l'hypothèse alternative :

$$H_1 \,:\, \alpha_i \neq \alpha_{i'}\,.$$

Le test est significatif au seuil $\alpha$ et on décide de rejeter l'hypothèse nulle $H_0$ en faveur de l'hypothèse alternative $H_1$ si :

$$\frac{|\hat{\mu}_i(\boldsymbol{y}) - \hat{\mu}_{i'}(\boldsymbol{y})|}{\sqrt{\frac{s_R^2}{2}\left(\frac{1}{n_i}+\frac{1}{n_{i'}}\right)}} \geq q(I, n-I; 1-\alpha)\,.$$

Le test n'est pas significatif au seuil $\alpha$ et on décide de conserver par défaut l'hypothèse nulle $H_0$ en faveur de l'hypothèse alternative $H_1$ si :

$$\frac{|\hat{\mu}_i(\boldsymbol{y}) - \hat{\mu}_{i'}(\boldsymbol{y})|}{\sqrt{\frac{s_R^2}{2}\left(\frac{1}{n_i}+\frac{1}{n_{i'}}\right)}} < q(I, n-I; 1-\alpha)\,.$$

### 5.2.3 La méthode de Scheffé

Contrairement à la méthode de Tukey, la méthode de Scheffé est valide même si le plan est déséquilibré. Notons $n_i$ le nombre de répétitions effectuées pour la modalité $A_i$ du facteur $A$.

Si $L$ est un contraste quelconque, estimé par $\hat{L}$, un intervalle de confiance au niveau $100(1-\alpha)\,\%$ est donné par :

$$\hat{L}(\boldsymbol{y}) - S\sqrt{(I-1)s_R^2\left(\sum_{i=1}^{I}\frac{l_i^2}{n_i}\right)} < L < \hat{L}(\boldsymbol{y}) + S\sqrt{(I-1)s_R^2\left(\sum_{i=1}^{I}\frac{l_i^2}{n_i}\right)},$$

où $S^2 = \mathcal{F}(I-1, n-I; 1-\alpha)$ est le $100(1-\alpha)$ quantile de la loi de Fisher- Snedecor à $I-1$ et $n-I$ degrés de liberté.

Si l'intervalle de confiance obtenu contient la valeur 0, on décide que le contraste n'est pas significativement différent de 0 au seuil $\alpha$. A contrario, si l'intervalle de confiance ne contient pas 0, alors on décide que le contraste est significativement différent de 0 au seuil $\alpha$.

Nous désirons tester l'hypothèse :

$$H_0 \,:\, L = 0$$

contre l'hypothèse alternative :

$$H_0 \,:\, L \neq 0\,.$$

Le test est significatif au seuil $\alpha$ et alors nous décidons de rejeter l'hypothèse nulle $H_0$ en faveur de l'hypothèse alternative $H_1$ si :

$$\frac{\left|\hat{L}(\boldsymbol{y})\right|}{\sqrt{(I-1)s_R^2\left(\sum_{i=1}^{I}\frac{l_i^2}{n_i}\right)}} \geq \sqrt{\mathcal{F}(I-1,n-I;1-\alpha)}\,.$$

Le test n'est pas significatif au seuil $\alpha$ et on décide de conserver par défaut l'hypothèse nulle $H_0$ en faveur de l'hypothèse alternative $H_1$ si :

$$\frac{\left|\hat{L}(\boldsymbol{y})\right|}{\sqrt{(I-1)s_R^2\left(\sum_{i=1}^{I}\frac{l_i^2}{n_i}\right)}} < \sqrt{\mathcal{F}(I-1,n-I;1-\alpha)}\,.$$

Appliqué au contexte des comparaisons multiples, l'intervalle de confiance ci-dessus se transforme de la manière suivante, puisque les contrastes étudiés sont du type : $L = \mu_i - \mu_j$, avec $i \neq j$ :

$$\sum_{i=1}^{I}\frac{l_i^2}{n_i} = \frac{1}{n_i} + \frac{(-1)^2}{n_{i'}} = \frac{1}{n_i} + \frac{1}{n_{i'}}\,.$$

Donc, dans le cas des comparaisons multiples, les intervalles de confiance ci-dessus se simplifient en :

$$\hat{\mu}_i(\boldsymbol{y}) - \hat{\mu}_{i'}(\boldsymbol{y}) - S\sqrt{(I-1)s_R^2\left(\frac{1}{n_i}+\frac{1}{n_{i'}}\right)} < \mu_i - \mu_{i'} < \hat{\mu}_i(\boldsymbol{y}) - \hat{\mu}_{i'}(\boldsymbol{y}) + S\sqrt{(I-1s_R^2\left(\frac{1}{n_i}+\frac{1}{n_{i'}}\right)}\,.$$

Les hypothèses et les tests statistiques se transforment, quant à eux, en :

$$H_0 \,:\, \mu_i = \mu_{i'}$$

contre l'hypothèse alternative :

$$H_0 \,:\, \mu_i \neq \mu_{i'}\,.$$

Rappelons que ces hypothèses sont équivalentes aux suivantes :

$$H_0 \,:\, \alpha_i = \alpha_{i'}$$

contre l'hypothèse alternative :

$$H_1 \,:\, \alpha_i \neq \alpha_{i'}\,.$$

Le test est significatif au seuil $\alpha$ et alors nous décidons de rejeter l'hypothèse nulle $H_0$ en faveur de l'hypothèse alternative $H_1$ si :

$$\frac{|\hat{\mu}_i(\boldsymbol{y}) - \hat{\mu}_{i'}(\boldsymbol{y})|}{\sqrt{(I-1)s_R^2\left(\frac{1}{n_i}+\frac{1}{n_{i'}}\right)}} \geq \sqrt{\mathcal{F}(I-1, n-I; 1-\alpha)}\,.$$

Le test n'est pas significatif au seuil $\alpha$ et on décide de conserver par défaut l'hypothèse nulle $H_0$ en faveur de l'hypothèse alternative $H_1$ si :

$$\frac{|\hat{\mu}_i(\boldsymbol{y}) - \hat{\mu}_{i'}(\boldsymbol{y})|}{\sqrt{(I-1)s_R^2\left(\frac{1}{n_i}+\frac{1}{n_{i'}}\right)}} < \sqrt{\mathcal{F}(I-1, n-I; 1-\alpha)}\,.$$

En utilisant ces intervalles de confiance pour décider simultanément de la significativité des $I(I-1)/2$ différences entre les effets des modalités du facteur $A$, nous sommes assurés que la probabilité qu'aucune des différences n'est significative est exactement de valeur $1-\alpha$.

Dans le cas où le plan est équilibré, on obtient l'intervalle de confiance de niveau $100(1-\alpha)$ suivant :

$$\hat{\mu}_i(\boldsymbol{y}) - \hat{\mu}_{i'}(\boldsymbol{y}) - S\sqrt{\frac{2(I-1)s_R^2}{J}} < \mu_i - \mu_{i'} < \hat{\mu}_i(\boldsymbol{y}) - \hat{\mu}_{i'}(\boldsymbol{y}) + S\sqrt{\frac{2(I-1)s_R^2}{J}}\,.$$

### 5.2.4 La méthode de Bonferroni

On suppose que nous souhaitons réaliser $k$ comparaisons de moyennes ou $k$ tests de contrastes et prendre une décision conjointe au risque de première espèce $\alpha$.

Soit $E_i$ l'événement associé au rejet de l'absence de différence significative au seuil $\alpha_{ind}$ lors de la $i$-ème comparaison du $i$-ème contraste. La méthode de Bonferroni est basée sur l'inégalité suivante :

$$\begin{aligned}\alpha &= P[E_1 \cup E_2 \cup \cdots \cup E_k] \\ &\leq P[E_1] + P[E_2] + \cdots + P[E_k] \\ &\leq k \times \alpha_{ind}\,.\end{aligned}$$

Donc, si nous voulons être sûr que le risque de première espèce $\alpha$ associé globalement à la prise simultanée de toutes les décisions lors des $k$ comparaisons ou des $k$ tests de contrastes est plus petit qu'une valeur $\alpha_0$ fixée à l'avance, il suffit de choisir :

$$\alpha_{ind} \leq \alpha_0/k\,.$$

Nous pouvons alors procéder à des comparaisons des moyennes deux à deux avec un test $t$ de Student de seuil $\alpha_0/k$ ou à un test de chacun des contrastes exposés précédemment au seuil $\alpha_0/k$. Cette procédure s'applique que le plan soit équilibré ou pas.

Les intervalles de confiance pour $k$ comparaisons de deux moyennes $\mu_i$ et $\mu_{i'}$ de deux groupes d'effectifs respectifs $n_i$ et $n_{i'}$ sont :

$$\hat{\mu}_i(\boldsymbol{y}) - \hat{\mu}_{i'}(\boldsymbol{y}) - t_B\sqrt{s_R^2\left(\frac{1}{n_i}+\frac{1}{n_{i'}}\right)} < \mu_i - \mu_{i'} < \hat{\mu}_i(\boldsymbol{y}) - \hat{\mu}_{i'}(\boldsymbol{y}) + t_B\sqrt{s_R^2\left(\frac{1}{n_i}+\frac{1}{n_{i'}}\right)},$$

où $t_B = t\left(n-I; 1-\dfrac{\alpha}{2k}\right)$ est le $100\left(1-\dfrac{\alpha}{2k}\right)$ quantile de la loi de Student à $n-I$ degrés de liberté.

**Remarque 5.2.1**

*On pourra retenir les conseils suivants pour utiliser la méthode de Bonferroni :*

1. *Le nombre de comparaisons k n'est pas très élevé. La procédure est trop conservatrice si k est élevé.*
2. *On préfère la méthode de Bonferroni à celle de Scheffé si le nombre de comparaisons est strictement inférieur à $I^2$.*
3. *On préfère la méthode de Bonferroni à celle de Tukey si le nombre de comparaisons est strictement inférieur à $I(I-1)/2$ ou si on souhaite tester en plus un petit nombre de comparaisons autres que celles des effets pricipaux des modalités $A_i$ du facteur $A$.*

### 5.2.5 La méthode de rejet séquentiel de Bonferroni et Holm

C'est une méthode dérivée de la méthode de Bonferroni exposée précédemment. Elle a été introduite pour augmenter la puissance de la méthode de Bonferroni. Elle consiste simplement à modifier le seuil des tests de comparaisons de la manière suivante : si l'un des tests de comparaison est significatif au seuil $\alpha_{ind} = \alpha/k$, alors nous effectuons le test suivant au niveau $\alpha_{ind} = \alpha/(k-1)$, et ainsi de suite ...

Les conseils d'utilisation sont identiques à ceux exposés dans la remarque précédente et on peut utiliser cette méthode aussi bien pour des plans équilibrés que pour des plans déséquilibrés.

## 5.3 Un exemple détaillé

On reprend l'exercice 1 du chapitre 3, où on a essayé 5 traitements contre les boutons de fièvre, le traitement 1 étant un placebo. Ces traitements ont été administrés au hasard sur trente patients ( six patients par groupe de traitement). Le délai, exprimé en jours, entre l'apparition des boutons de fièvre et la cicatrisation complète a été recueilli chez chacun des trente patients, détaillé ci-dessous :

| $T_1$ | $T_2$ | $T_3$ | $T_4$ | $T_5$ |
|---|---|---|---|---|
| 5 | 4 | 6 | 7 | 9 |
| 8 | 6 | 4 | 4 | 3 |
| 7 | 6 | 4 | 6 | 5 |
| 7 | 3 | 5 | 6 | 7 |
| 10 | 5 | 4 | 3 | 7 |
| 8 | 6 | 3 | 5 | 6 |

Il faut d'abord entrer les données :

```
> X<-data.frame(T1=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
```

On peut tracer des boîtes à moustaches (voir dessin ci-dessous) pour examiner, traitement par traitement des délais de cicatrisation pour chaque traitement, grâce à la commande :

> plot(delai traitement,col="green")

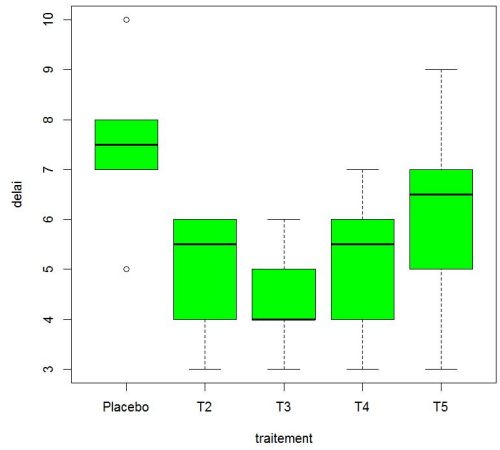


FIGURE 5.1 – Boîte à moustaches des délais de cicatrisation pour chaque traitement

On remarque aisément que la moyenne du traitement 1 (le placebo) est différente des autres. La table d'ANOVA est appelée via la commande :

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

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. 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 et le traitement $\boldsymbol{T_2}$. Les tests bilatéraux effectués dans ce modèle sont résumés ci-dessous :

| | $H_1$ |
|---|---|
| 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 ***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 des précédents. 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 traitement 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).

Supposons que nous voulions comparer les traitements 2 et 3. Il convient alors d'utiliser le contraste : $L_1 = \boldsymbol{\lambda}^t \boldsymbol{\mu}$ avec $\boldsymbol{\lambda} = (0, 1, -1, 0, 0)^t$ et $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_5)^t$, et d'effectuer le test $H_0 : L_1 = 0$ contre $H_1 : L_1 \neq 0$. Pour cela, on peut utiliser la fonction fit.contrast( ) disponible dans le package gregmisc :

```
> require(gregmisc)
> cmat <- rbind(" : 2 versus 3"=c(0,1,-1,0,0))
> fit.contrast(mon.aov,traitement,cmat)

                        Estimate Std. Error   t value  Pr(>|t|)
traitement : 2 versus 3 0.6666667  0.8831761 0.7548514 0.4573908
```

Il n'y a pas de différence significative entre les traitements 2 et 3.

La fonction pairwise.t.test() permet d'effectuer toutes les comparaisons deux à deux en proposant plusieurs méthodes de correction du risque afin de tenir compte du problème de la multiplicité des tests.

```
> pairwise.t.test(delai,traitement,p.adjust="bonf")

        Pairwise comparisons using t tests with pooled SD
data:  delai and traitement

   Placebo T2    T3    T4
T2 0.090   -     -     -
T3 0.014   1.000 -     -
T4 0.140   1.000 1.000 -
T5 1.000   1.000 0.483 1.000

P value adjustment method: bonferroni
```

Le logiciel ***R*** donne les valeurs ajustées suivant la correction de Bonferroni, c'est-à-dire que les valeurs corrigées sont obtenues en multipliant les valeurs des tests de Student par le nombre de tests effectués. Au vu des résultats ci-dessus, on trouve une valeur de probabilité critique de 0,014 entre le traitement 1 (placebo) et le traitement 3. Il existe donc une différence significative entre le traitement 1 (placebo) et le traitement 3 au risque de 5 %.

On peut remarquer que la comparaison entre le traitement 1 et le traitement 3 avait déjà été effectué auparavant. La valeur de la probabilité critique pour ce test individuel était de 0,0014. Comme dix comparaisons ont été effectuées, cette dernière valeur a été multipliée par 10 par la méthode de Bonferroni.

Comme cela a été détaillé ci-dessus, beaucoup d'autres méthodes sont possibles. Cependant, dans le cas d'une analyse de variance à un facteur avec le même nombre d'observations par groupe, la méthode de Tukey est la plus précise. Elle fournit des intervalles de confiance simultanés pour les différences entre les paramètres $\mu_i - \mu_j$ où $1 \leq i < j \leq I$.

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

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

Fit: aov(formula = delai ~ traitement)
```

```
$traitement
                 diff        lwr         upr     p adj
T2-Placebo -2.5000000 -5.0937744  0.09377442 0.0627671
T3-Placebo -3.1666667 -5.7604411 -0.57289224 0.0113209
T4-Placebo -2.3333333 -4.9271078  0.26044109 0.0927171
T5-Placebo -1.3333333 -3.9271078  1.26044109 0.5660002
T3-T2      -0.6666667 -3.2604411  1.92710776 0.9410027
T4-T2       0.1666667 -2.4271078  2.76044109 0.9996956
T5-T2       1.1666667 -1.4271078  3.76044109 0.6811222
T4-T3       0.8333333 -1.7604411  3.42710776 0.8770466
T5-T3       1.8333333 -0.7604411  4.42710776 0.2614661
T5-T4       1.0000000 -1.5937744  3.59377442 0.7881333
```

On peut illustrer cela sur le graphique ci-dessous :

```
> par(las=1) # Écriture horizontale des étiquettes.
> plot(TukeyHSD(mon.aov))
```

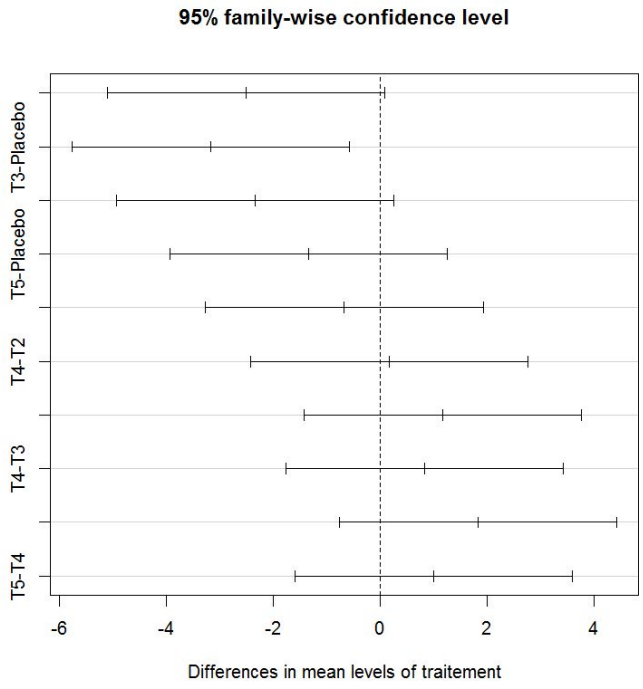


FIGURE 5.2 – Intervalles de confiance des différences de moyennes

La méthode de Tukey va dans le même sens que les résultats obtenus via la méthode de Bonferroni. En effet, seul l'intervalle de confiance ne contenant pas la valeur 0 est celui concernant la différence entre le traitement 3 et le traitement 1 (placebo). Ainsi, le délai de cicatrisation étant plus court pour le traitement 3, nous proposons d'utiliser ce traitement.

# Exercices

**Exercice 1**

On reprend l'exercice 3 du chapitre 3.

1. Y a-t-il une différence significative entre la prise de poids des veaux par l'alimentation 1 et la prise de poids des veaux par chacune des deux autres alimentations ?
2. Utilisez la méthode de Bonferroni pour effectuer toutes les comparaisons deux à deux des prises de poids selon les alimentations.
3. Reprendre ces comparaisons en utilisant la méthode de Tukey.
4. Le fabricant d'alimentation 3 profère que son alimentation permet d'augmenter de 50 % la prise de poids des veaux par rapport à l'alimentation 1. Que pensez-vous de cette affirmation ?

**Exercice 2**

On reprend l'exercice 7 du chapitre 3 sur l'acidité des bières.

1. Les acidités des bières 1 et 3 sont-elles 2 fois supérieures à celles des bières 2 et 4 ?
2. Comparez, en utilisant la méthode de Bonferroni, deux à deux les quatre bières quant à leur acidité moyenne.
3. Utilisez la méthode de Tukey pour poursuivre cette comparaison.

**Exercice 3**

On veut comparer la perception de la saveur amère de 30 cidres bruts et de 30 cidres mi-secs. Une note d'amertume variant entre 0 et 10 a été affectée pour chaque cidre par un jury de 20 experts. Les résultats de cette étude se trouvent dans le fichier chap5.ex3.csv fourni.

Dans la suite, on notera $Y_b$ et $Y_m$ les variables *amertume* des cidres bruts et des cidres mi-secs respectivement. On supposera que les notes d'amertume pour un même type de cidre (brut ou mi-sec) suivent une loi normale. On supposera que la variance des notes est la même pour les cidres bruts et les cidres mi-secs. On peut donc écrire : $Y_b \sim \mathcal{N}(\mu_b, \sigma^2)$ et $Y_m \sim \mathcal{N}(\mu_m, \sigma^2)$ avec $\mu_b$, $\mu_m$ et $\sigma^2$ des paramètres inconnus.

On souhaite dans un premier temps tester l'égalité des paramètres $\mu_b$ et $\mu_m$.

1. Avant toute chose, testez d'abord l'égalité des variances, et testez la normalité des deux variables $Y_b$ et $Y_m$. Donnez aussi une estimation de $\sigma^2$.
2. En vue de ce que l'on veut faire, quelle est l'hypothèse que l'on cherche à tester ? Quelle est l'hypothèse alternative ? Quelles méthodes statistiques permettent de tester cette hypothèse ?
3. À partir de la question précédente, proposez deux stratégies de test de comparaison des deux populations (cidres bruts et cidres demi-secs). Quelle est la région critique associé au risque $\alpha$ de première espèce ?
4. Que décidez-vous de faire ici en effectuant ces deux tests ? Interprétez.
5. Quel lien existe-t-il entre les valeurs des deux statistiques de test utilisées dans ces deux tests ?

6. Démontrez que, dans le cas d'un facteur à deux modalités, la statistique de test calculée par l'analyse de la variance est toujours égale au carré de la statistique de test pour le test de Student correspondant.
7. Si on voulait comparer l'amertume de trois types de cidres (brut,demi-sec et doux), laquelle des deux stratégies pourrait-on utiliser ?
8. Dans toute la suite, on s'intéresse à la puissance du test de comparaison de deux moyennes dans le cas où la variance $\sigma^2$ est supposée connue et égale à 1,83. Proposez une nouvelle stratégie de décision pour tester l'égalité des paramètres $\mu_b$ et $\mu_m$, tenant compte de la connaissance de $\sigma^2$. Construisez ce test.
9. Montrez que, sous $H_1$, la loi de la statistique de test ne dépend que de l'écart entre les deux moyennes (noté $\delta$) : $\delta = \mu_b - \mu_m$.
10. Dans le cas où l'échantillon contient autant de cidres bruts que de cidres demi-secs ($n_b = n_m = n$), calculez la puissance du test en fonction de $\delta$ et $n$.
11. Calculez la taille de l'échantillon minimale pour détecter une différence entre $\mu_b$ et $\mu_m$ de l'ordre de 0,5 point avec une probabilité valant 0,9 lorsque le test est construit avec un niveau de confiance de 95 %.
12. Tracez, pour une taille d'échantillon de $n_b = n_m = 30$, la courbe de puissance du test, c'est-à-dire tracez, en fonction de l'écart $\delta$, la puissance du test d'égalité des paramètres $\mu_b$ et $\mu_m$.

 

# Chapitre 6

# Analyse de la variance à deux facteurs

L'analyse de variance à deux facteurs peut être considérée comme une généralisation de l'analyse de variance à un facteur, permettant de tenir compte simultanément de deux facteurs. Les deux facteurs peuvent être placés soit sur un pied d'égalité, soit subordonnés l'un à l'autre. Dans le premier cas, les modèles d'analyse de variance sont dits croisés, et, dans le second cas, ils sont appelés hiérarchisés ou multi-niveaux.

Là encore, on distinguera entre modèles fixes, modèles aléatoires et modèles mixtes. Une distinction importante sera faite entre le cas des effectifs égaux, souvent qualifié de plan équilibré ou orthogonal, et le cas des effectifs inégaux, souvent qualifié de plan non équilibré ou non orthogonal.

Globalement, les conditions d'application de l'analyse de variance à deux facteurs sont de la même nature que pour un seul facteur : populations normales, de même variance, et échantillons simples et indépendants.

Nous irons plus rapidement dans la description de ce modèle, renvoyant au chapitre 3 pour les démonstrations. Par exemple, la décomposition de la variation totale se fait de la même façon dans le cas de l'analyse de la variance à un facteur.

## 6.1 Modèles à effets fixes

### 6.1.1 Modèles sans répétition

On suppose qu'un facteur contrôlé $A$ possède $I$ modalités, chacune d'elle étant notée $A_i$. De même, on suppose qu'un facteur contrôlé $B$ possède $J$ modalités, chacune d'elle étant notée $B_j$. Pour chaque couple de modalités $(A_i, B_j)$, nous effectuons **une seule mesure** de la variable réponse $Y$, toujours supposée continue, ce qui est assez fréquemment le cas. Nous noterons $n$ le nombre total de mesures effectuées : $n = I \times J$.

Pour cela, on introduit le modèle :

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

avec les contraintes supplémentaires : $\sum_{i=1}^{I} \alpha_i = 0$ et $\sum_{j=1}^{J} \beta_j = 0$.

$Y_{i,j}$ est la valeur prise par la variable réponse $Y$ dans les conditions $(A_i, B_j)$. On supposera toujours réalisées les hypothèses standards suivantes :

1. $\varepsilon_{i,j}$ et $\varepsilon_{k,l}$ sont indépendantes si $(i,j) \neq (k,l)$ avec $1 \leq i,k \leq I$ et $1 \leq j,l \leq J$.
2. $\forall\,(i,j)$, $i = 1, \cdots, I\,;\, j = 1, \cdots, J\,, \mathcal{L}(\varepsilon_{i,j}) = \mathcal{N}(0, \sigma^2)$.

L'étude de la vérification des hypothèses ci-dessus a été faite dans le précédent chapitre.

Nous regroupons les valeurs prises par la variable réponse $Y$ dans les conditions $(A_i, B_j)$ dans le tableau ci-dessous :

| **Facteur $A$** | **Facteur $B$** | | | | |
|---|---|---|---|---|---|
| | $\boldsymbol{B_1}$ | $\cdots$ | $\boldsymbol{B_j}$ | $\cdots$ | $\boldsymbol{B_J}$ |
| $\boldsymbol{A_1}$ | $Y_{1,1}$ | $\cdots$ | $Y_{1,j}$ | $\cdots$ | $Y_{1,J}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $\boldsymbol{A_i}$ | $Y_{i,1}$ | $\cdot$ | $Y_{i,j}$ | $\cdots$ | $Y_{i,J}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $\boldsymbol{A_I}$ | $Y_{I,1}$ | $\cdots$ | $Y_{I,j}$ | $\cdots$ | $Y_{I,J}$ |

TABLE 6.1 – Tableau des valeurs de la variable réponse $Y$

La variation théorique totale est définie par :

$$SC_{TOT} = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{i,j} - Y_{\bullet,\bullet})^2\,. \tag{6.1.1}$$

On appelle variation théorique due au facteur $A$, la quantité :

$$SC_A = J\sum_{i=1}^{I}(Y_{i,\bullet} - Y_{\bullet,\bullet})^2\,. \tag{6.1.2}$$

De la même façon, la variation théorique due au facteur $B$ est :

$$SC_B = I\sum_{j=1}^{J}(Y_{\bullet,j} - Y_{\bullet,\bullet})^2\,. \tag{6.1.3}$$

La variance résiduelle est définie par :

$$SC_R = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{i,j} - Y_{i,\bullet} - Y_{\bullet,j} + Y_{\bullet,\bullet})^2\,. \tag{6.1.4}$$

On montre alors aisément la relation fondamentale de l'analyse de variance à deux facteurs sans répétition :

$$SC_{TOT} = SC_A + SC_B + SC_R\,. \tag{6.1.5}$$

| Source | Degrés de liberté |
|---|---|
| **Facteur $A$** | $n_A = I - 1$ |
| **Facteur $B$** | $n_B = J - 1$ |
| **Résiduelle** | $n_R = (I-1)(J-1)$ |
| **Totale** | $n_{TOT} = IJ - 1$ |

TABLE 6.2 – Table des degrés de liberté

Aux différentes sommes des carrés des écarts peuvent être associés des nombres de degrés de liberté :

On notera, comme dans le chapitre d'analyse de variance à un facteur, les carrés moyens théoriques par :

$$S_A^2 = \frac{SC_A}{n_A} \quad ; \quad S_B^2 = \frac{SC_B}{n_B} \quad ; \quad S_R^2 = \frac{SC_R}{n_R} \quad ; \quad S_T^2 = \frac{SC_{TOT}}{n_{TOT}}\,,$$

qui constituent eux aussi des mesures globales de variations.

Notons $\boldsymbol{y}$ des données expérimentales $y_{1,1}, \cdots, y_{1,J}, y_{2,1}, \cdots, j_{2,J}, \cdots, y_{I,J}$ permettant une réalisation du tableau précédent (6.1) :

| **Facteur $A$** | **Facteur $B$** | | | | |
|---|---|---|---|---|---|
| | $\boldsymbol{B_1}$ | $\cdots$ | $\boldsymbol{B_j}$ | $\cdots$ | $\boldsymbol{B_J}$ |
| $\boldsymbol{A_1}$ | $y_{1,1}$ | $\cdots$ | $y_{1,j}$ | $\cdots$ | $y_{1,J}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $\boldsymbol{A_i}$ | $y_{i,1}$ | $\cdot$ | $y_{i,j}$ | $\cdots$ | $y_{i,J}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $\boldsymbol{A_I}$ | $y_{I,1}$ | $\cdots$ | $y_{I,j}$ | $\cdots$ | $y_{I,J}$ |

TABLE 6.3 – Tableau des données expérimentales $(y_{i,j})$ de la variable réponse

La variation totale observée sur la liste $\boldsymbol{y}$ de données expérimentales est définie par :

$$sc_{TOT} = \sum_{i=1}^{I}\sum_{j=1}^{J}(y_{i,j} - y_{\bullet,\bullet})^2\,. \tag{6.1.6}$$

La variation due au facteur $A$ observée sur la liste $\boldsymbol{y}$ de données expérimentales est définie par :

$$sc_A = J\sum_{i=1}^{I}(y_{i,\bullet} - y_{\bullet,\bullet})^2\,. \tag{6.1.7}$$

La variation due au facteur $B$ observée sur la liste $\boldsymbol{y}$ de données expérimentales est définie par :

$$sc_B = I \sum_{j=1}^{J} (y_{\bullet,j} - y_{\bullet,\bullet})^2 \,. \tag{6.1.8}$$

La variation résiduelle observée sur la liste $\boldsymbol{y}$ de données expérimentales est quant à elle égale à :

$$sc_R = \sum_{i=1}^{I} \sum_{j=1}^{J} (y_{i,j} - y_{i,\bullet} - y_{\bullet,j} + y_{\bullet,\bullet})^2 \,. \tag{6.1.9}$$

La relation fondamentale de l'analyse de variance reste valable lorsqu'elle est évaluée sur la liste $\boldsymbol{y}$ de données expérimentales :

$$sc_{TOT} = sc_A + sc_B + sc_R \,. \tag{6.1.10}$$

On désire faire les tests d'hypothèses suivantes :

$$H_0' \,;\, \alpha_1 = \alpha_2 = \cdots = \alpha_I = 0$$

contre

$$H_1' \,:\, \text{Il existe } i_0 \in \{1, 2, \cdots, I\} \text{ tel que } \alpha_{i_0} \neq 0 \,.$$

On notera, comme dans le chapitre d'analyse de variance à un facteur, les carrés moyens observés par :

$$s_A^2 = \frac{sc_A}{n_A} \quad ; \quad s_B^2 = \frac{sc_B}{n_B} \quad ; \quad s_R^2 = \frac{sc_R}{n_R} \quad ; \quad s_T^2 = \frac{sc_{TOT}}{n_{TOT}} \,,$$

qui constituent eux aussi des mesures globales de variations.

Sous l'hypothèse nulle $H_0'$ d'absence d'effet du facteur $A$ et lorsque les conditions de validité du modèle sont respectées, $f_A$ est la réalisation de la variable aléatoire $S_A^2/S_R^2$ qui suit une loi de Fisher-Snedecor à $n_A = I - 1$ et $n_R = (I-1)(J-1)$ degrés de liberté.

On peut alors conclure grâce à la valeur critique, et on rejette l'hypothèse nulle si elle est inférieure ou égale au seuil $\alpha$ du test, ou à l'aide d'une table. Il y a rejet si $f_A$ est supérieure ou égale à la valeur critique issue de la table. Si l'hypothèse $H_0'$ est rejetée, on pourra procéder à des comparaisons multiples des différents effets des niveaux du facteur, ce qui sera vu dans un chapitre ultérieur dédié.

Nous pouvons répéter tout ce qui précède pour le facteur $B$. On peut souhaiter tester les hyphothèses :

$$H_0'' \,;\, \beta_1 = \beta_2 = \cdots = \beta_I = 0$$

contre

$$H_1'' \,:\, \text{Il existe } j_0 \in \{1, 2, \cdots, J\} \text{ tel que } \beta_{j_0} \neq 0 \,.$$

Sous l'hypothèse nulle $H_0''$ d'absence d'effet du facteur $B$ et lorsque les conditions de validité du modèle sont respectées, $f_B$ est la réalisation de la variable aléatoire $S_B^2/S_R^2$ qui suit une loi de Fisher-Snedecor à $n_B = J - 1$ et $n_R = (I-1)(J-1)$ degrés de liberté.

On peut alors conclure grâce à la valeur critique, et on rejette l'hypothèse nulle si elle est inférieure ou égale au seuil $\alpha$ du test, ou à l'aide d'une table. Il y a rejet si $f_B$ est supérieure ou égale à la valeur critique issue de la table. Si l'hypothèse $H_0''$ est rejetée, on pourra procéder à des comparaisons multiples des différents effets des niveaux du facteur, ce qui sera vu dans un chapitre ultérieur dédié.

Le tableau d'analyse de la variance à deux facteurs résume les choses ci-dessous :

| Source | Variations | d.d.l. | Carrés moyens | $F$ | Décision |
|---|---|---|---|---|---|
| **Facteur $A$** | $sc_A$ | $n_A$ | $s_A^2 = \dfrac{sc_A}{n_A}$ | $f_A = \dfrac{s_A^2}{s_R^2}$ | $H_0'$ ou $H_1'$ |
| **Facteur $B$** | $sc_B$ | $n_B$ | $s_B^2 = \dfrac{sc_B}{n_B}$ | $f_B = \dfrac{s_B^2}{s_R^2}$ | $H_0''$ ou $H_1''$ |
| **Résiduelle** | $sc_R$ | $n_R$ | $s_R^2 = \dfrac{sc_R}{n_R}$ | | |
| **Totale** | $sc_{TOT}$ | $n_{TOT}$ | | | |

TABLE 6.4 – Tableau de l'analyse de variance à deux facteurs

Les estimateurs $\hat{\mu}, \hat{\alpha}_1, \cdots, \hat{\alpha}_I, \hat{\beta}_1, \cdots, \hat{\beta}_J$ et $\hat{\sigma}^2$ des paramètres respectifs $\mu, \alpha_1, \cdots, \alpha_I, \beta_1, \cdots, \beta_J$ et $\sigma^2$ du modèle sont données par :

$$\hat{\mu} = Y_{\bullet,\bullet} \quad ; \quad \hat{\alpha}_i = Y_{i,\bullet} - \hat{\mu} \qquad 1 \leq i \leq I \quad ; \quad \hat{\beta}_j = Y_{\bullet,j} - \hat{\mu} \qquad 1 \leq j \leq J$$

$$\hat{\sigma}^2 = \frac{SC_R}{(I-1)(J-1)} = S_R^2 \,.$$

Ce sont des estimateurs sans biais. Les estimations obtenues pour une liste de données expérimentales $\boldsymbol{y}$, notées $\hat{\mu}(\boldsymbol{y}), \hat{\alpha}_1(\boldsymbol{y}), \cdots, \hat{\alpha}_I(\boldsymbol{y}), \hat{\beta}_1(\boldsymbol{y}), \cdots, \hat{\beta}_J(\boldsymbol{y})$ et $\hat{\sigma}^2(\boldsymbol{y})$ des paramètres $\mu, \alpha_1, \cdots, \alpha_I, \beta_1, \cdots, \beta_J$ et $\sigma^2$ du modèle se déduisent mutatis mutandis des formules précédentes.

**Exemple 6.1.1**

*L'influence d'un traitement grossissant, à base de vitamines, est étudiée sur des animaux de races différentes. Pour cela nous disposons d'animaux de trois races, notées $R_i$ , pour $i = 1, 2, 3$, et nous avons effectué trois traitements, notés $D_j$ , pour $j = 1, 23$, utilisant respectivement 5, 10 et 15$\mu$g de vitamines B12 par cm3. Le gain moyen de poids par jour est mesuré, à l'issue d'un traitement de 50 jours dans chaque cas. Un seul animal est utilisé pour chaque couple « race-traitement ».*

| Race / Traitement | $R_1$ | $R_2$ | $R_3$ |
|---|---|---|---|
| $D_1$ | *1,26* | *1,21* | *1,19* |
| $D_2$ | *1,29* | *1,23* | *1,23* |
| $D_3$ | *1,38* | *1,27* | *1,22* |

*L'objectif est d'effectuer une analyse de la variance à deux facteurs sans répétition (il y a en effet une seule observation par « case »). Les facteurs, contrôlés, à effets fixes, sont la race et la dose, tous les deux à 3 modalités. La réponse est le gain moyen de poids.*

*Nous désirons tester les hypothèses suivantes :*

$$\begin{cases} H_0^R : & \textit{Les races n'ont pas d'effet sur la prise de poids} \\ \textit{contre} & \\ H_1^R : & \textit{Les races ont un effet sur la prise de poids} \end{cases}$$

*puis*

$$\begin{cases} H_0^D : & \textit{Les doses n'ont pas d'effet sur la prise de poids} \\ \textit{contre} & \\ H_1^D : & \textit{Les doses ont un effet sur la prise de poids} \end{cases}$$

*Le tableau d'analyse de la variance correspondant est :*

```
            Df   Sum Sq  Mean Sq F value Pr(>F)
races        2 0.015267 0.007633   9.745 0.0290 *
doses        2 0.007400 0.003700   4.723 0.0885 .
Residuals    4 0.003133 0.000783
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

*Nous décidons alors que :*

1. *$H_1^R$ est vraie : il y a un effet de la race sur le gain de poids ($p = 0,0290$)*
2. *$H_0^D$ est vraie, il n'y a pas d'effet de la dose sur le gain de poids ($p = 0,0885$)*

*Nous avons bien sûr supposé l'indépendance des observations.*

*Pour vérifier l'hypothèse de normalité des résidus, nous pouvons effectuer un test de Shapiro-Wilk :*

```
        Shapiro-Wilk normality test

data:  mod$res
W = 0.9798, p-value = 0.9632
```

*Nous décidons donc que l'hypothèse de normalité est vérifiée, c'est-à-dire que nous décidons que la normalité de l'erreur théorique est acceptée.*

*Il ne nous reste plus qu'à vérifier l'égalité des variances des résidus, encore appelé l'homogénéité des variances. Remarquons tout d'abord que nous ne pouvons pas tester l'égalité des variances : en effet, nous n'avons qu'une observation par « case ».*

*Cependant, à titre indicatif, nous pouvons tester : l'égalité des variances des gains selon les races, c'est-à-dire :*

$$\begin{cases} H_0 : & \textit{les variances des races sont égales} \\ H_1 : & \textit{les variances des races ne sont pas égales} \end{cases}$$

*Nous effectuons pour cela le test de Bartlett, qui donne :*

```
        Bartlett test of homogeneity of variances

data:  mod$res by races
Bartlett's K-squared = 3.2583, df = 2, p-value = 0.1961
```

*Nous décidons donc que l'hypothèse d'homogénéité est vérifiée, c'est-à-dire que nous décidons que les variances théoriques des gains des trois races sont égales.*

*Nous pouvons tester aussi l'égalité des variances des gains selon les doses, c'est-à-dire :*

$$\begin{cases} H_0 : & \textit{les variances des doses sont égales} \\ H_1 : & \textit{les variances des doses ne sont pas égales} \end{cases}$$

*Le test de Bartlett donne alors :*

```
        Bartlett test of homogeneity of variances

data:  mod$res by doses
Bartlett's K-squared = 1.0819, df = 2, p-value = 0.5822
```

*Nous décidons donc que l'hypothèse d'homogénéité est vérifiée,c'est-à-dire que nous décidons que les variances théoriques des gains de poids selon les trois doses sont égales.*

### 6.1.2 Modèles avec répétition

On suppose qu'un facteur contrôlé $A$ possède $I$ modalités, chacune d'elle étant notée $A_i$. De même, on suppose qu'un facteur contrôlé $B$ possède $J$ modalités, chacune d'elle étant notée $B_j$. Pour chaque couple de modalités $(A_i, B_j)$, nous effectuons $K \geq 2$ mesures d'une variable réponse $Y$ qui est supposée être une variable continue. Nous noterons $n$ le nombre total de mesures effectuées : $n = I \times J \times K$.

Pour cela, on introduit le modèle :

$$Y_{i,j,k} = \mu + \alpha_i + \beta_j + \gamma_{i,j} + \varepsilon_{i,j,k} \qquad , \qquad i = 1, \cdots, I \, ; \, j = 1, \cdots, J \, ; \, k = 1, \cdots, K \, .$$

avec les contraintes supplémentaires : $\sum_{i=1}^{I} \alpha_i = 0$ et $\sum_{j=1}^{J} \beta_j = 0$

$$\sum_{i=1}^{I} \gamma_{i,j} = 0\,,\ \forall\, j \in \{1, \cdots, J\} \quad \text{et} \quad \sum_{j=1}^{J} \gamma_{i,j} = 0\,,\ \forall\, i \in \{1, \cdots, I\}\,.$$

On a introduit ici un terme d'interaction $\gamma_{i,j}$ qui représente l'interaction entre les deux facteurs. $Y_{i,j,k}$ est la valeur prise par la variable réponse $Y$ dans les conditions $(A_i, B_j)$ lors du $k$-ième essai. On supposera toujours réalisées les hypothèses standards suivantes :

1. $\varepsilon_{i,j,k}$ et $\varepsilon_{l,m,n}$ sont indépendantes si $(i,j,k) \neq (l,m,n)$ avec $1 \leq i,l \leq I$ , $1 \leq j,m \leq J$ et $1 \leq k,n \leq K$.

2. $\forall\,(i,j,k)$, $i = 1, \cdots, I$ ; $j = 1, \cdots, J$ ; $k = 1, \cdots, K$ ; $\mathcal{L}(\varepsilon_{i,j,k}) = \mathcal{N}(0, \sigma^2)$.

L'étude de la vérification des hypothèses ci-dessus a été faite dans le précédent chapitre.

Nous regroupons les valeurs prises par la variable réponse $Y$ dans les conditions $(A_i, B_j)$ lors des $K$ répétitions dans le tableau ci-dessous :

| **Facteur $A$** | **Facteur $B$** | | | | |
|---|---|---|---|---|---|
| | $\boldsymbol{B_1}$ | $\cdots$ | $\boldsymbol{B_j}$ | $\cdots$ | $\boldsymbol{B_J}$ |
| $\boldsymbol{A_1}$ | $Y_{1,1,1} \cdots Y_{1,1,K}$ | $\cdots$ | $Y_{1,j,1} \cdots Y_{1,j,K}$ | $\cdots$ | $Y_{1,J,1} \cdots Y_{1,J,K}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $\boldsymbol{A_i}$ | $Y_{i,1,1} \cdots Y_{i,1,K}$ | $\cdot$ | $Y_{i,j,1} \cdots Y_{i,j,K}$ | $\cdots$ | $Y_{i,J,1} \cdots Y_{i,J,K}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $\boldsymbol{A_I}$ | $Y_{I,1,1} \cdots Y_{I,1,K}$ | $\cdots$ | $Y_{I,j,1} \cdots Y_{I,j,K}$ | $\cdots$ | $Y_{I,J,1} \cdots Y_{I,J,K}$ |

TABLE 6.5 – Tableau des valeurs de la variable réponse $Y$

Il y a donc $I \times J \times K$ variables aléatoires $Y_{i,j,k}$. On peut, comme au chapitre 3, définir les

différentes moyennes suivantes :

$$\begin{aligned}
Y_{\bullet,j,k} &= \frac{1}{I}\sum_{i=1}^{I} Y_{i,j,k} \\
Y_{i,\bullet,k} &= \frac{1}{J}\sum_{j=1}^{J} Y_{i,j,k} \\
Y_{i,j,\bullet} &= \frac{1}{K}\sum_{k=1}^{K} Y_{i,j,k} \\
Y_{\bullet,\bullet,k} &= \frac{1}{I\times J}\sum_{i=1}^{I}\sum_{j=1}^{J} Y_{i,j,k} \\
Y_{\bullet,j,\bullet} &= \frac{1}{I\times K}\sum_{i=1}^{I}\sum_{k=1}^{K} Y_{i,j,k} \\
Y_{i,\bullet,\bullet} &= \frac{1}{J\times K}\sum_{j=1}^{J}\sum_{k=1}^{K} Y_{i,j,k} \\
Y_{\bullet,\bullet,\bullet} &= \frac{1}{I\times J\times K}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} Y_{i,j,k}
\end{aligned}$$

Comme dans le chapitre 3, la variation théorique due au facteur $A$ est définie par :

$$SC_A = JK\sum_{i=1}^{I}(Y_{i,\bullet,\bullet} - Y_{\bullet,\bullet,\bullet})^2 .$$

De même la variation théorique due au facteur $B$ est définie par :

$$SC_B = IK\sum_{j=1}^{J}(Y_{\bullet,j,\bullet} - Y_{\bullet,\bullet,\bullet})^2 .$$

La variation théorique due à l'interaction des facteurs $A$ et $B$ est définie par :

$$SC_{AB} = K\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{i,j,\bullet} - Y_{i,\bullet,\bullet} - Y_{\bullet,j,\bullet} + Y_{\bullet,\bullet,\bullet})^2 .$$

La variation résiduelle théorique, quant à elle, est définie par :

$$SC_R = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{i,j,k} - Y_{i,j,\bullet})^2 .$$

Enfin, la variation totale théorique est égale à :

$$SC_{TOT} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{i,j,k} - Y_{\bullet,\bullet,\bullet})^2 .$$

On peut subdiviser les écarts par rapport à la moyenne générale $Y_{\bullet,\bullet,\bullet}$ en deux, puis quatre composantes :

$$\begin{aligned} Y_{i,j,k} - Y_{\bullet,\bullet,\bullet} &= (Y_{i,j,\bullet} - Y_{\bullet,\bullet,\bullet}) + (Y_{i,j,k} - Y_{i,j,\bullet}) \\ &= (Y_{i,\bullet,\bullet} - Y_{\bullet,\bullet,\bullet}) + (Y_{\bullet,j,\bullet)} - Y_{\bullet,\bullet,\bullet}) + (Y_{i,j,\bullet} - Y_{i,\bullet,\bullet} - Y_{\bullet,j,\bullet} + Y_{\bullet,\bullet,\bullet}) + (Y_{i,j,k} - Y_{i,j,\bullet}) \,. \end{aligned}$$

La première décomposition est identique à celle réalisée pour l'analyse de la variance à un facteur. La seconde décomposition fait, quant à elle, apparaître deux termes de variations des facteurs, relatifs à l'un et à l'autre des facteurs, un terme dit d'interaction, et un terme de variation résiduelle.

Par élévation au carré et sommation pour les $I \times J \times K$ observations, on obtient l'équation d'analyse de la variance à deux facteurs :

$$\begin{aligned} \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{i,j,k} - Y_{\bullet,\bullet,\bullet})^2 = JK\sum_{i=1}^{I}(Y_{i,\bullet,\bullet} - Y_{\bullet,\bullet,\bullet})^2 + IK\sum_{j=1}^{J}(Y_{\bullet,j,\bullet})^2 \\ +K\sum_{i=1}^{I}{}_{j=1}^{J}(Y_{i,j,\bullet} - Y_{i,\bullet,\bullet} - Y_{\bullet,j,\bullet} + Y_{\bullet,\bullet,\bullet})^2 + \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{i,j,k} - Y_{i,j,\bullet})^2 \,. \end{aligned}$$

Les deux premiers facteurs sont des sommes de carrés dûs au deux facteurs, la troisième est une somme de carrés liés à l'interaction, et la quatrième est une somme de carrés d'écarts résiduelle. En utilisant les définitions ci-dessus, l'équation de l'analyse de variance à deux facteurs peut encore s'écrire sous la forme :

$$SC_{TOT} = SC_A + SC_B + SC_{AB} + SC_R \,.$$

Aux différentes sommes des carrés des écarts peuvent ici encore être asssociés des nombres de degrés de liberté vérifiant la relation :

$$IJK = (I-1) + (J-1) + (I-1)(J-1) + IJ(K-1) \,.$$

Il s'agit de $IJK-1$ degrés de liberté pour la somme totale, puisqu'elle fait intervenir globalement $IJK$ observations individuelles, $I-1$ et $J-1$ degrés de liberté pour les deux sommes de chacun des deux facteurs, car elles sont calculées respectivement à partir de $I$ et $J$ moyennes, $IJ(K-1)$ degrés de liberté pour la somme résiduelle puisqu'elle fait intervenir $IJ$ échantillons de $K$ observations, et, par différence, $(I-1)(J-1)$ degrés de liberté pour la somme des carrés des écarts de l'interaction.

Comme dans le cas de l'analyse de variance à un facteur, en divisant les différentes sommes des carrés des écarts, on obtient les carrés moyens : $S_A^2$, $S_B^2$,$S_{AB}^2$, $S_R^2$ et $S_T^2$.

L'ensemble de ces résultats peut alors être présenté sous la forme d'un tableau de l'analyse de variance :

| Sources de variation | Degrés de liberté | Variations | Carrés moyens |
|---|---|---|---|
| Facteur $A$ | $n_A = I - 1$ | $SC_A$ | $S_A^2 = \dfrac{SC_A}{I-1}$ |
| Facteur $B$ | $n_B = J - 1$ | $SC_B$ | $S_B^2 = \dfrac{SC_B}{J-1}$ |
| Interaction | $n_{AB} = (I-1)(J-1)$ | $SC_{AB}$ | $S_{AB}^2 = \dfrac{SC_{AB}}{(I-1)(J-1)}$ |
| Résidus | $n_R = IJ(K-1)$ | $SC_R$ | $S_R^2 = \dfrac{SC_R}{IJ(K-1)}$ |
| Total | $n_{TOT} = IJK - 1$ | $SC_{TOT}$ | $S_T^2 = \dfrac{SC_{TOT}}{IJK-1}$ |

L'interaction $Y_{i,j,\bullet} - Y_{i,\bullet,\bullet} - Y_{\bullet,j,\bullet} + Y_{\bullet,\bullet,\bullet}$ apparaît naturellement dans le modèle d'analyse de variance à deux facteurs lorsqu'on veut équilibrer le modèle après y avoir fait figurer les deux termes dus aux deux facteurs : $Y_{i,\bullet,\bullet} - Y_{\bullet,\bullet,\bullet}$, $Y_{\bullet,j,\bullet} - Y_{\bullet,\bullet,\bullet}$ et le terme résiduel : $Y_{i,j,k} - Y_{i,j,\bullet}$.

Ces termes d'interaction sont nuls quand les différences liées à l'action d'un des deux facteurs ne dépendent pas de l'autre facteur, c'est-à-dire quand, par exemple, les écarts $Y_{i,j,\bullet} - Y_{\bullet,j\bullet}$ relatifs au premier facteur sont indépendants des modalités $j$ du second facteur.

En effet, quand ces écarts ne dépendent pas de $j$, ils sont tous égaux entre eux, pour chaque valeur de $i$, et donc égaux aussi à leur moyenne :

$$Y_{i,1,\bullet} - Y_{\bullet,1,\bullet} = \cdots = Y_{i,j,\bullet} - Y_{\bullet,j,\bullet} = \cdots = Y_{i,J,\bullet} - Y_{\bullet,J,\bullet} = Y_{i,\bullet,\bullet} - Y_{\bullet,\bullet,\bullet}\,.$$

On a alors, pour tout $i$ et tout $j$ :

$$Y_{i,j,\bullet} - Y_{i,\bullet,\bullet} - Y_{\bullet,j,\bullet} + Y_{\bullet,\bullet,\bullet} = 0\,.$$

De même, les termes d'interaction sont nuls quand les écarts $Y_{i,j\bullet} - Y_{i,\bullet,\bullet}$ relatifs au second facteur sont indépendants de $i$, c'est-à-dire du premier facteur. De plus, ces deux conditions de nullité des termes d'interaction sont strictement équivalentes.

Notons $\boldsymbol{y}$ des données expérimentales $y_{1,1,1}, \cdots, y_{1,1,K}, y_{1,2,1}, \cdots, j_{1,2,K}, \cdots, y_{I,J,K}$ permettant une réalisation du tableau précédent (6.5) :

| **Facteur $A$** | **Facteur $B$** | | | | |
|---|---|---|---|---|---|
| | $\boldsymbol{B_1}$ | $\cdots$ | $\boldsymbol{B_j}$ | $\cdots$ | $\boldsymbol{B_J}$ |
| $\boldsymbol{A_1}$ | $y_{1,1,1} \cdots y_{1,1,K}$ | $\cdots$ | $y_{1,j,1} \cdots y_{1,j,K}$ | $\cdots$ | $y_{1,J,1} \cdots y_{1,J,K}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $\boldsymbol{A_i}$ | $y_{i,1,1} \cdots y_{i,1,K}$ | $\cdot$ | $y_{i,j,1} \cdots y_{i,j,K}$ | $\cdots$ | $y_{i,J,1} \cdots y_{i,J,K}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $\boldsymbol{A_I}$ | $y_{I,1,1} \cdots y_{I,1,K}$ | $\cdots$ | $y_{I,j,1} \cdots y_{I,j,K}$ | $\cdots$ | $y_{I,J,1} \cdots y_{I,J,K}$ |

TABLE 6.6 – Tableau des valeurs de la variable réponse $Y$

La variation due au facteur $A$ observée sur la liste $\boldsymbol{y}$ de données expérimentales est définie par :

$$sc_A = JK \sum_{i=1}^{I} (y_{i,\bullet,\bullet} - y_{\bullet,\bullet,\bullet})^2 \,.$$

La variation due au facteur $B$ observée sur la liste $\boldsymbol{y}$ de données expérimentales est définie par :

$$sc_B = IK \sum_{j=1}^{J} (y_{\bullet,j,\bullet} - y_{\bullet,\bullet,\bullet})^2 \,.$$

La variation due à l'interaction des facteurs $A$ et $B$, observée sur la liste $\boldsymbol{y}$ de données expérimentales est définie par :

$$sc_{AB} = K \sum_{i=1}^{I} \sum_{j=1}^{J} (y_{i,j,\bullet} - y_{i,\bullet,\bullet} - y_{\bullet,j,\bullet} + y_{\bullet,\bullet,\bullet})^2 \,.$$

La variation résiduelle observée sur la liste $\boldsymbol{y}$ de données expérimentales est définie par :

$$sc_R = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} (y_{i,j,k} - y_{i,j,\bullet})^2 \,.$$

Enfin, la variation totale observée sur la liste $\boldsymbol{y}$ de données expérimentales est égale par :

$$sc_{TOT} = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} (y_{i,j,k} - y_{\bullet,\bullet,\bullet})^2 \,.$$

La relation fondamentale de l'analyse de variance reste valable lorsqu'elle est évaluée sur la liste $\boldsymbol{y}$ de données expérimentales :

$$sc_{TOT} = sc_A + sc_B + sc_{AB} + sc_R \,.$$

On désire faire les tests d'hypothèses suivantes :

$$H_0' \,;\, \alpha_1 = \alpha_2 = \cdots = \alpha_I = 0$$

contre

$$H_1' \text{ : Il existe } i_0 \in \{1, 2, \cdots, I\} \text{ tel que } \alpha_{i_0} \neq 0\,.$$

Sous l'hypothèse nulle $H_0'$ d'absence d'effet du facteur $A$ et lorsque les conditions de validité du modèle sont respectées, $f_A$ est la réalisation de la variable aléatoire $S_A^2/S_R^2$ qui suit une loi de Fisher-Snedecor à $n_A = I - 1$ et $n_R = IJ(K-1)$ degrés de liberté.

On peut alors conclure grâce à la valeur critique, et on rejette l'hypothèse nulle si elle est inférieure ou égale au seuil $\alpha$ du test, ou à l'aide d'une table. Il y a rejet si $f_A$ est supérieure ou égale à la valeur critique issue de la table. Si l'hypothèse $H_0'$ est rejetée, on pourra procéder à des comparaisons multiples des différents effets des niveaux du facteur, ce qui sera vu dans un chapitre ultérieur dédié.

Nous pouvons répéter tout ce qui précède pour le facteur $B$. On peut souhaiter tester les hyphothèses :

$$H_0'' \,;\, \beta_1 = \beta_2 = \cdots = \beta_I = 0$$

contre

$$H_1'' \text{ : Il existe } j_0 \in \{1, 2, \cdots, J\} \text{ tel que } \beta_{j_0} \neq 0\,.$$

Sous l'hypothèse nulle $H_0''$ d'absence d'effet du facteur $B$ et lorsque les conditions de validité du modèle sont respectées, $f_B$ est la réalisation de la variable aléatoire $S_B^2/S_R^2$ qui suit une loi de Fisher-Snedecor à $n_B = J - 1$ et $n_R = IJ(K-1)$ degrés de liberté.

On peut alors conclure grâce à la valeur critique, et on rejette l'hypothèse nulle si elle est inférieure ou égale au seuil $\alpha$ du test, ou à l'aide d'une table. Il y a rejet si $f_B$ est supérieure ou égale à la valeur critique issue de la table. Si l'hypothèse $H_0''$ est rejetée, on pourra procéder à des comparaisons multiples des différents effets des niveaux du facteur, ce qui sera vu dans un chapitre ultérieur dédié.

Nous pouvons également faire des tests d'hypothèses sur l'absence ou la présence d'interaction entre les facteurs $A$ et $B$ :

$$H_0''' \,;\, \gamma_{1,1} = \gamma_{1,2} = \cdots = \gamma_{1,J} = \gamma_{2,1} = \cdots = \gamma_{I,J} = 0$$

contre

$$H_1''' \text{ : Il existe } (i_0, j_0) \in \{1, 2, \cdots, I\} \times \{1, 2, \cdots, J\} \text{ tel que } \gamma_{i_0,j_0} \neq 0\,.$$

Sous l'hypothèse nulle $H_0'''$ d'absence d'effet de l'interaction des facteurs $A$ et $B$ et lorsque les conditions de validité du modèle sont respectées, $f_{AB}$ est la réalisation de la variable aléatoire $S_{AB}^2/S_R^2$ qui suit une loi de Fisher-Snedecor à $nAB_=(I-1)(J-1)$ et $n_R = IJ(K-1)$ degrés de liberté.

On peut alors conclure grâce à la valeur critique, et on rejette l'hypothèse nulle si elle est inférieure ou égale au seuil $\alpha$ du test, ou à l'aide d'une table. Il y a rejet si $f_{AB}$ est supérieure ou égale à la valeur critique issue de la table.

On résume toutes ces informations dans le tableau d'analyse de variance suivant :

| **Sources de variation** | **Degrés de liberté** | **Variations** | **Carrés moyens** | **Statistique** F | **Décision** |
|---|---|---|---|---|---|
| Facteur $A$ | $n_A = I-1$ | $sc_A$ | $s_A^2 = \dfrac{sc_A}{n_A}$ | $f_A = \dfrac{s_A^2}{s_R^2}$ | $H_0'$ ou $H_1'$ |
| Facteur $B$ | $n_B = J-1$ | $sc_B$ | $s_B^2 = \dfrac{sc_B}{n_B}$ | $f_B = \dfrac{s_B^2}{s_R^2}$ | $H_0''$ ou $H_1''$ |
| Interaction | $n_{AB} = (I-1)(J-1)$ | $sc_{AB}$ | $s_{AB}^2 = \dfrac{sc_{AB}}{n_{AB}}$ | $f_{AB} = \dfrac{s_{AB}^2}{s_R^2}$ | $H_0'''$ ou $H_1'''$ |
| Résidus | $n_R = IJ(K-1)$ | $sc_R$ | $s_R^2 = \dfrac{sc_R}{n_R}$ | | |
| Total | $n_{TOT} = IJK$-1 | $sc_{TOT}$ | $s_T^2 = \dfrac{sc_{TOT}}{n_{TOT}}$ | | |

Les estimateurs $\hat{\mu}$, $\hat{\alpha}_1, \cdots, \hat{\alpha}_I, \hat{\beta}_1, \cdots, \hat{\beta}_J$, $\hat{\gamma}_{1,1}, \hat{\gamma}_{1,2}, \cdots, \hat{\gamma}_{1,J}$, $\hat{\gamma}_{2,1}, \cdots, \hat{\gamma}_{I,J}$ et $\hat{\sigma}^2$ des paramètres respectifs $\mu$, $\alpha_1, \cdots, \alpha_I, \beta_1, \cdots, \beta_J, \gamma_{1,1}, \cdots, \gamma_{I,J}$ et $\sigma^2$ du modèle sont donnés par :

$$\hat{\mu} = Y_{\bullet,\bullet,\bullet} \quad ; \quad \hat{\alpha}_i = Y_{i,\bullet,\bullet} - \hat{\mu} \qquad 1 \leq i \leq I \quad ; \quad \hat{\beta}_j = Y_{\bullet,j,\bullet} - \hat{\mu} \qquad 1 \leq j \leq J$$

$$\hat{\gamma}_{i,j} = Y_{i,j,\bullet} - Y_{i,\bullet,\bullet} - Y_{\bullet,j,\bullet} + Y_{\bullet,\bullet,\bullet}$$

$$\hat{\sigma}^2 = \frac{SC_R}{IJ(K-1)} = S_R^2 \,.$$

Ce sont des estimateurs sans biais.

Les estimations obtenues pour une liste de données expérimentales $\boldsymbol{y}$, notées $\hat{\mu}(\boldsymbol{y})$, $\hat{\alpha}_1(\boldsymbol{y}), \cdots, \hat{\alpha}_I(\boldsymbol{y})$, $\hat{\beta}_1(\boldsymbol{y}), \cdots, \hat{\beta}_J(\boldsymbol{y})$, $\hat{\gamma}_{1,1}(\boldsymbol{y}), \cdots, \hat{\gamma}_{I,J}(\boldsymbol{y})$ et $\hat{\sigma}^2(\boldsymbol{y})$ des paramètres $\mu$, $\alpha_1, \cdots, \alpha_I, \beta_1, \cdots, \beta_J, \gamma_{1,1}, \cdots, \gamma_{I,J}$ et $\sigma^2$ du modèle se déduisent, mutatis mutandis, des formules précédentes.

**Exemple 6.1.2**

*Nous nous proposons d'analyser l'influence du temps et de trois espèces ligneuses d'arbres sur la décomposition de la masse d'une litière constituée de feuilles de lierre. Pour ce faire, 24 sachets d'une masse identique de feuilles de lierre ont été constitués, sachets permettant une décomposition naturelle. Puis une première série de 8 sachets, choisis au hasard, a été déposée sous un chêne, une deuxième sous un peuplier, et la dernière série sous un frêne. Après 2, 7, 10 et 16 semaines*

*respectivement, deux sachets sont prélevés au hasard sous chaque arbre et la masse résiduelle est déterminée pour chacun d'eux. Cette masse est exprimée en pourcentage de la masse initiale.*

*Les valeurs observées sont les suivantes :*

| Semaine | Chêne | Peuplier | Frêne |
|---|---|---|---|
| *2* | *85,10* | *85,20* | *84,30* |
| | *87,60* | *84,90* | *85,75* |
| *7* | *75,90* | *73,00* | *72,80* |
| | *72,85* | *75,70* | *70,80* |
| *10* | *71,60* | *74,15* | *67,10* |
| | *66,95* | *71,85* | *64,95* |
| *16* | *62,10* | *67,25* | *58,75* |
| | *64,30* | *60,25* | *59,00* |

*Nous observons trois variables :*

1. *Deux d'entre elles sont des variables contrôlées, l'espèce d'arbre, qualitative à trois modalités, et la semaine qui peut être considérée comme qualitative à quatre modalités.*
2. *La troisième variable est une réponse quantitative.*

*Donc l'analyse de la variance à deux facteurs (semaine et espèce d'arbre) croisés, avec interaction, peut convenir, entre autres méthodes d'analyse de ces données.*

*Le résultat de l'analyse de variance à deux facteurs donne les résultats ci-dessous :*

```
Analysis of Variance Table

Response: masse
               Df  Sum Sq Mean Sq  F value    Pr(>F)
semaine         3 1741.31  580.44 121.6927 3.004e-09 ***
espece          2   58.08   29.04   6.0881   0.01495 *
semaine:espece  6   30.22    5.04   1.0559   0.43853
Residuals      12   57.24    4.77
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

*La probabilité critique ($0,43853$) est supérieure à 5%, donc on accepte $H_0$, et on conclut à la non-significativité de l'interaction. On peut tracer un graphe d'interaction :*

*Le parallélisme des courbes indique là aussi un manque d'interaction.*

*On va donc estimer à nouveau le modèle sans interaction. Cela donne :*

```
Analysis of Variance Table

Response: masse
          Df  Sum Sq Mean Sq  F value    Pr(>F)
semaine    3 1741.31  580.44 119.4657 4.509e-12 ***
espece     2   58.08   29.04   5.9767   0.01022 *
```

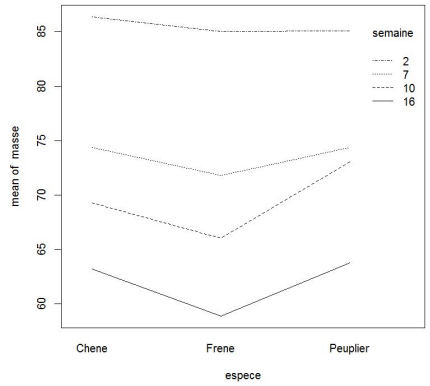


FIGURE 6.1 – Représentation des masses moyennes observées en fonction des deux facteurs considérés

```
Residuals 18   87.45    4.86
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

*Les deux facteurs sont significatifs ; il y a donc un effet semaine et un effet espèce sur la masse résiduelle de lierre.*

*Nous pouvons estimer les différents coefficients $\alpha_i$ et $\beta_j$.*

```
Call:
lm(formula = masse ~ C(semaine, sum) + C(espece, sum), data = don)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1937 -1.4573 -0.3625  1.4516  3.8604

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       72.5896     0.4499 161.333  < 2e-16 ***
C(semaine, sum)1  12.8854     0.7793  16.534  2.5e-12 ***
C(semaine, sum)2   0.9188     0.7793   1.179 0.253777
C(semaine, sum)3  -3.1562     0.7793  -4.050 0.000751 ***
C(espece, sum)1    0.7104     0.6363   1.116 0.278902
C(espece, sum)2   -2.1583     0.6363  -3.392 0.003249 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.204 on 18 degrees of freedom
Multiple R-squared:  0.9537,    Adjusted R-squared:  0.9408
F-statistic: 74.07 on 5 and 18 DF,  p-value: 2.285e-11
```

*On obtient une matrice de "Coefficients" qui comporte pour chaque paramètre (chaque ligne) 4 colonnes : son estimation, son écart-type estimé ("Std.Error"), la valeur observée de la statistique de test considérée ; enfin, la probabilité critique ($Pr(>|t|)$) donne pour la statistique de test sous $H_0$, la probabilité de dépasser la valeur estimée.*

*La valeur de $\mu$, notée ici "Intercept" correspond à l'effet moyen. L'effet de la semaine 16 n'est pas donné dans le listing de sortie, mais comme la somme des $\alpha_i$ est nulle, on estime $\alpha_4$ par : $\alpha_4 = -\alpha_1 - \alpha_2 - \alpha_3 = -12,8854 - 0,9188 + 3,1562 = -10,648$. De la même façon, on estime $\beta_3$ par : $\beta_3 = -\beta_1 - \beta_2 = -0,7104 + 2,1583 = 1,4479$.*

## 6.2 Modèles à effets aléatoires

Il se peut que les effets d'un facteur ne puissent être modélisés par des effets fixes. Dans certains cas, en particulier quand les modalités sont choisies au hasard, le fait de supposer que les effets sont fixes n'est pas adapté. Par conséquent, nous pouvons être confrontés à deux autres types de modèles, modèles avec deux facteurs à effets aléatoires, avec ou sans répétitions. Ces deux modèles sont appelés modèles à effets aléatoires.

### 6.2.1 Modèles à effets aléatoires sans répétition

Les termes $A_i$ représentent un échantillon de taille $I$ prélevé dans une population importante. Nous admettrons que les effets des $A_i$ sont distribués suivant une loi normale centrée de variance $\sigma_A^2$.

Les termes $B_j$ représentent un échantillon de taille $J$ prélevé dans une population importante. Nous admettrons que les effets des $B_j$ sont distribués suivant une loi normale centrée de variance $\sigma_B^2$.

Pour chacun des couples de modalités $(A_i; B_j)$, nous effectuons une mesure d'une réponse $Y$ qui est une variable continue. Nous notons $n = I \times J$ le nombre total de mesures ayant été effectuées.

Nous introduisons alors le modèle suivant :

$$Y_{i,j} = \mu + \alpha_i + \beta_j + \varepsilon_{i,j}\,, \quad i = 1, \cdots, I\,, \quad j = 1, \cdots, J\,,$$

où $Y_{i,j}$ est la valeur prise par la variable réponse $Y$ dans les conditions $(A_i, B_j)$. Nous supposons aussi que :

$$\begin{array}{rcll} \mathcal{L}(\alpha_i) & = & \mathcal{N}(0, \sigma_A^2) & \forall\, i = 1, \cdots, I \\ \mathcal{L}(\beta_i) & = & \mathcal{N}(0, \sigma_B^2) & \forall\, j = 1, \cdots, J \end{array}$$

ainsi que l'indépendance des effets aléatoires :

$$\begin{array}{ll} \alpha_i \text{ indépendante de } \alpha_k & \text{si } i \neq k \text{ et } 1 \leq i, k \leq I \\ \beta_j \text{ indépendante de } \beta_k & \text{si } j \neq k \text{ et } 1 \leq j, k \leq J \\ \alpha_i \text{ indépendante de } \beta_j & \text{si } 1 \leq i \leq I \text{ et } 1 \leq j \leq J \end{array} \quad .$$

Nous postulons les hypothèses classiques suivantes :

$$\forall\,(i,j),\, 1 \leq i \leq I,\, 1 \leq j \leq J,\; \mathcal{L}(\varepsilon_{i,j}) = \mathcal{N}(0, \sigma^2)\,,$$
$$\varepsilon_{i,j} \text{ indépendante de } \varepsilon_{k,l} \text{ si } (i,j) \neq (k,l) \text{ avec } 1 \leq i, k \leq I \text{ et } 1 \leq j, l \leq J\,,$$

 

ainsi que l'indépendance des effets aléatoires et des erreurs :

$$\alpha_i \text{ indépendante de } \varepsilon_{j,k} \text{ si } 1 \leq i, j \leq I \text{ et } 1 \leq k \leq J\,,$$
$$\beta_j \text{ indépendante de } \varepsilon_{l,k} \text{ si } 1 \leq l \leq I \text{ et } 1 \leq j, k \leq J\,.$$

Nous supposons que les conditions d'application de ce modèle sont bien remplies. Nous utilisons ici les quantités $SC_A$, $SC_B$, $SC_R$, $SC_{TOT}$, $sc_A$, $sc_B$, $sc_R$, $sc_{TOT}$ introduites à la section 5.1.1.

La relation fondamentale de l'analyse de variance tient toujours :

$$SC_{TOT} = SC_A + SC_B + SC_R\,.$$

On introduit une fois encore le nombre de degrés de liberté associés à chaque ligne du tableau de l'analyse de variance :

| Source | Degrés de liberté |
|---|---|
| **Facteur *A*** | $n_A = I - 1$ |
| **Facteur *B*** | $n_B = J - 1$ |
| **Résiduelle** | $n_R = (I-1)(J-1)$ |
| **Totale** | $n_{TOT} = IJ - 1$ |

Nous pouvons résumer toutes ces informations dans le tableau d'analyse de variance suivant :

| Source | Variations | d.d.l. | Carrés moyens | $F$ | Décision |
|---|---|---|---|---|---|
| **Facteur *A*** | $sc_A$ | $n_A$ | $s_A^2 = \dfrac{sc_A}{n_A}$ | $f_A = \dfrac{s_A^2}{s_R^2}$ | $H_0'$ ou $H_1'$ |
| **Facteur *B*** | $sc_B$ | $n_B$ | $s_B^2 = \dfrac{sc_B}{n_B}$ | $f_B = \dfrac{s_B^2}{s_R^2}$ | $H_0''$ ou $H_1''$ |
| **Résiduelle** | $sc_R$ | $n_R$ | $s_R^2 = \dfrac{sc_R}{n_R}$ | | |
| **Totale** | $sc_{TOT}$ | $n_{TOT}$ | | | |

TABLE 6.7 – Tableau de l'analyse de variance à deux facteurs

L'analyse de la variance à deux facteurs aléatoires sans répétition permet deux tests de Fisher. Le premier test concernant le facteur $A$ est le suivant :

$$H_0' \, : \, \sigma_A^2 = 0$$

contre

$$H_1' \, : \, \sigma_A^2 \neq 0 \, .$$

Sous l'hypothèse nulle $(H_0')$ précédente, d'absence d'effet du facteur $A$, et lorsque les conditions de validité du modèle sont respectées, $f_A$ est la réalisation d'une variable aléatoire qui suit une loi de Fisher à $I-1$ et $(I-1)(J-1)$ degrés de liberté.

Le second test concernant le second facteur $B$ est le suivant :

$$H_0'' \, : \, \sigma_B^2 = 0$$

contre

$$H_1'' \, : \, \sigma_B^2 \neq 0 \, .$$

Sous l'hypothèse nulle $(H_0^{'prime})$ précédente, d'absence d'effet du facteur $B$, et lorsque les conditions de validité du modèle sont respectées, $f_B$ est la réalisation d'une variable aléatoire qui suit une loi de Fisher à $J-1$ et $(I-1)(J-1)$ degrés de liberté.

Les estimateurs $\hat{\mu}$, $\hat{\sigma}_A^2$, $\hat{\sigma}_B^2$, $\hat{\sigma}^2$ des paramètres $\mu$, $\sigma_A^2$, $\sigma_B^2$ et $\sigma^2$ du modèle sont donnés par les formules suivantes :

$$\begin{aligned}
\hat{\mu} &= Y_{\bullet,\bullet} \, , \\
\hat{\sigma}_A^2 &= \frac{1}{J} \left( S_A^2 - S_R^2 \right) \quad ; \quad \hat{\sigma}_B^2 = \frac{1}{I} \left( S_B^2 - S_R^2 \right) \, , \\
\hat{\sigma}^2 &= \frac{SC_R}{(I-1)(J-1)} = S_R^2 \, ,
\end{aligned}$$

où $S_A^2 = \dfrac{SC_A}{n_A}$, $S_B^2 = \dfrac{SC_B}{n_B}$ et $S_R^2 = \dfrac{SC_R}{n_R}$. Ces estimateurs sont non biaisés.

Les estimations, obtenues pour une liste de données expérimentales $\boldsymbol{y}$, notées $\hat{\mu}(\boldsymbol{y})$, $\hat{\sigma}_A^2(\boldsymbol{y})$, $\hat{\sigma}_B^2(\boldsymbol{y})$, $\hat{\sigma}^2(\boldsymbol{y})$ des paramètres $\mu$, $\sigma_A^2$, $\sigma_B^2$ et $\sigma^2$ du modèle, se déduisent immédiatement des formules ci-dessus :

$$\begin{aligned}
\hat{\mu}(\boldsymbol{y}) &= y_{\bullet,\bullet} \, , \\
\hat{\sigma}_A^2(\boldsymbol{y}) &= \frac{1}{J} \left( s_A^2 - s_R^2 \right) \quad ; \quad \hat{\sigma}_B^2(\boldsymbol{y}) = \frac{1}{I} \left( s_B^2 - s_R^2 \right) \, , \\
\hat{\sigma}^2(\boldsymbol{y}) &= \frac{sc_R}{(I-1)(J-1)} = s_R^2 \, .
\end{aligned}$$

**Exemple 6.2.1**

*Nous étudions la dissolution du principe actif contenu dans un type donné de comprimé issu de lots de production distincts. Pour cela, six lots ont été sélectionnés au hasard parmi toute la production et la dissolution de quatre comprimés pris au hasard dans chacun des lots est observée. Après 15, 30, 45 et 60 minutes, un comprimé de chaque lot est sélectionné et le pourcentage de principe actif dissous, par rapport à la valeur titre, est déterminé. Ces valeurs sont données dans le*

*tableau qui va suivre. Il est à noter que les temps d'observation à savoir, 15, 30, 45 et 60 minutes sont des temps qui ont été choisis aléatoirement par l'expérimentateur qui n'avait pas de connaissance a priori sur ces 24 comprimés.*

| **Temps** / **Lots** | **15 min.** | **30 min.** | **45 min.** | **60 min.** |
|---|---|---|---|---|
| **Lot 1** | *66* | *87* | *93* | *90* |
| **Lot 2** | *60* | *91* | *99* | *98* |
| **Lot 3** | *69* | *91* | *93* | *92* |
| **Lot 4** | *61* | *97* | *97* | *101* |
| **Lot 5** | *61* | *84* | *106* | *103* |
| **Lot 6** | *57* | *88* | *94* | *99* |

*L'expérimentateur se demande à partir de quel instant peut-on admettre qu'un comprimé est entièrement dissous ?*

*Le tableau d'analyse de variance est le suivant :*

```
Analysis of Variance Table

Response: principe
          Df Sum Sq Mean Sq F value    Pr(>F)
temps      3 4908.5 1636.15 66.6382 6.694e-09 ***
lots       5   83.2   16.64  0.6778     0.647
Residuals 15  368.3   24.55
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

*Pour le second test, la valeur critique "P-value" = 0,647, nous décidons donc de ne pas refuser l'hypothèse nulle ($H_0$). Par conséquent, nous n'avons pas réussi à mettre en évidence d'effet du facteur aléatoire « Lot ». Le risque associé à cette décision est un risque de deuxième espèce. Pour l'évaluer, il resterait à calculer la puissance de ce test.*

*Pour le premier test, la valeur critique "P-value" = 6.694e-09 , nous décidons donc de refuser l'hypothèse nulle ($H_0$). Par conséquent, nous pouvons dire, au seuil 5%, qu'il y a un effet significatif du facteur aléatoire « Temps ».*

*Nous ne sommes pas capables de répondre à la question de l'expérimentateur, à savoir : « à partir de quel instant pouvons-nous admettre qu'un comprimé est entièrement dissous ? », puisque nous ne pouvons pas faire de tests de comparaisons multiples, étant donné que le facteur « Temps » est à effets aléatoires.*

*Bien sûr, nous ne pouvons faire cette analyse des résultats, qu'en supposant avoir auparavant vérifié que les conditions du modèle soient bien remplies.*

### 6.2.2 Modèles à effets aléatoires avec répétition

Les termes $A_i$ représentent un échantillon de taille $I$ prélevé dans une population importante. Nous admettrons que les effets des $A_i$ sont distribués suivant une loi normale centrée de variance $\sigma_A^2$.

Les termes $B_j$ représentent un échantillon de taille $J$ prélevé dans une population importante. Nous admettrons que les effets des $B_j$ sont distribués suivant une loi normale centrée de variance $\sigma_B^2$.

Pour chacun des couples de modalités $(A_i; B_j)$, nous effectuons $K \geq 2$ d'une réponse $Y$ qui est une variable continue. Nous notons $n = I \times J \times K$ le nombre total de mesures ayant été effectuées.

Nous introduisons alors le modèle suivant :

$$Y_{i,j,k} = \mu + \alpha_i + \beta_j + \gamma_{i,j} + \varepsilon_{i,j,k}\,, \quad i = 1, \cdots, I\,, \quad j = 1, \cdots, J\,, \quad k = 1, \cdots, K\,,$$

où $Y_{i,j,k}$ est la valeur prise par la variable réponse $Y$ dans les conditions $(A_i, B_j)$ lors du $k$-ième essai. Nous supposons aussi que :

$$\begin{array}{rcll} \mathcal{L}(\alpha_i) & = & \mathcal{N}(0, \sigma_A^2) & \forall\, i = 1, \cdots, I\,, \\ \mathcal{L}(\beta_i) & = & \mathcal{N}(0, \sigma_B^2) & \forall\, j = 1, \cdots, J\,, \\ \mathcal{L}(\gamma_{i,j}) & = & \mathcal{N}(0, \sigma_{AB}^2) & \forall\, i = 1, \cdots, I \quad j = 1, \cdots, J\,, \end{array}$$

ainsi que l'indépendance des effets aléatoires :

$$\begin{array}{rl} \alpha_i \text{ indépendante de } \alpha_k & \text{si } i \neq k \text{ et } 1 \leq i, k \leq I\,, \\ \beta_j \text{ indépendante de } \beta_k & \text{si } j \neq k \text{ et } 1 \leq j, k \leq J\,, \\ \gamma_{i,j} \text{ indépendante de } \gamma_{k,l} & \text{si } (i,j) \neq (k,l) \text{ avec } 1 \leq i, k \leq I \text{ et } 1 \leq j, l \leq J\,, \\ \alpha_i \text{ indépendante de } \beta_j & \text{si } 1 \leq i \leq I \text{ et } 1 \leq j \leq J\,, \\ \alpha_i \text{ indépendante de } \gamma_{j,k} & \text{si } 1 \leq i, j \leq I \text{ et } 1 \leq k \leq J\,, \\ \beta_i \text{ indépendante de } \gamma_{j,k} & \text{si } 1 \leq j \leq I \text{ et } 1 \leq i, k \leq J\,. \end{array}$$

Nous postulons aussi les hypothèses classiques suivantes pour les erreurs :

$$\forall\,(i,j,k),\, 1 \leq i \leq I,\, 1 \leq j \leq J,\, 1 \leq k \leq K \;\; \mathcal{L}(\varepsilon_{i,j,k}) = \mathcal{N}(0, \sigma^2)\,,$$
$$\varepsilon_{i,j,k} \text{ indépendante de } \varepsilon_{l,m,n} \text{ si } (i,j,k) \neq (l,m,n) \text{ avec } 1 \leq i, l \leq I\,, 1 \leq j, m \leq J \text{ et } 1 \leq k, n \leq K\,,$$

ainsi que l'indépendance des effets aléatoires et des erreurs :

$$\begin{array}{l} \alpha_i \text{ indépendante de } \varepsilon_{j,k,l} \text{ si } 1 \leq i, j \leq I\,, 1 \leq k \leq J\,, \text{ et } 1 \leq l \leq K\,, \\ \beta_j \text{ indépendante de } \varepsilon_{j,k,l} \text{ si } 1 \leq j \leq I\,,\, 1 \leq i, k \leq J \text{ et } 1 \leq l \leq K\,, \\ \gamma_{i,j} \text{ indépendante de } \varepsilon_{k,l,m} \text{ si } 1 \leq i, k \leq I\,, 1 \leq j, l \leq J \text{ et } 1 \leq m \leq K\,. \end{array}$$

Nous supposons que les conditions d'utilisation de ce modèle sont satisfaites. Une fois encore, nous utilisons les quantités $SC_A$, $SC_B$, $SC_{AB}$, $SC_R$, $SC_{TOT}$, $sc_A$, $sc_B$, $sc_{AB}$, $sc_R$ et $sc_{TOT}$ introduites dans la section 6.1.2.

La relation fondamentale de l'analyse de variance s'écrit ici aussi :

$$SC_{TOT} = SC_A + SC_B + SC_{AB} + SC_R\,.$$

Nous avons les degrés de liberté suivants, associés à chaque ligne du tableau de l'analyse de variance :

Nous résumons toutes ces informations dans le tableau de l'analyse de variance suivant :

| Source | Degrés de liberté |
|---|---|
| **Facteur $A$** | $n_A = I - 1$ |
| **Facteur $B$** | $n_B = J - 1$ |
| **Interaction $AB$** | $n_{AB} = (I-1)(J-1)$ |
| **Résiduelle** | $n_R = IJ(K-1)$ |
| **Totale** | $n_{TOT} = IJK - 1$ |

L'analyse de la variance à deux facteurs aléatoires avec répétitions permet trois tests de Fisher. Le premier test concernant le facteur $A$ est le suivant :

$$H'_0 \, : \, \sigma_A^2 = 0$$

contre

$$H'_1 \, : \, \sigma_A^2 \neq 0 \, .$$

Sous l'hypothèse nulle $(H'_0)$ précédente, d'absence d'effet du facteur $A$, et lorsque les conditions de validité du modèle sont respectées, $f_A$ est la réalisation d'une variable aléatoire qui suit une loi de Fisher à $I-1$ et $(I-1)(J-1)$ degrés de liberté.

Le second test concernant le second facteur $B$ est le suivant :

$$H''_0 \, : \, \sigma_B^2 = 0$$

contre

$$H''_1 \, : \, \sigma_B^2 \neq 0 \, .$$

Sous l'hypothèse nulle $(H''_0)$ précédente, d'absence d'effet du facteur $B$, et lorsque les conditions de validité du modèle sont respectées, $f_B$ est la réalisation d'une variable aléatoire qui suit une loi de Fisher à $J-1$ et $(I-1)(J-1)$ degrés de liberté.

Enfin, le troisième test concernant l'interaction entre les facteurs $A$ et $B$

$$H'''_0 \, : \, \sigma_{AB}^2 = 0$$

contre

$$H'''_1 \, : \, \sigma_{AB}^2 \neq 0 \, .$$

Sous l'hypothèse nulle $(H'''_0)$ précédente, d'absence d'effet de l'interaction entre les facteurs $A$ et $B$, et lorsque les conditions de validité du modèle sont respectées, $f_{AB}$ est la réalisation d'une variable aléatoire qui suit une loi de Fisher à $(I-1)(J-1)$ et $IJ(K-1)$ degrés de liberté.

Les estimateurs $\hat{\mu}$, $\hat{\sigma}_A^2$, $\hat{\sigma}_B^2$, $\hat{\sigma}_{AB}^2$ et $\hat{\sigma}^2$ des paramètres $\mu$, $\sigma_A^2$, $\sigma_B^2$, $\sigma_{AB}^2$ et $\sigma^2$ du modèle sont

| Source | Variations | d.d.l. | Carrés moyens | $F$ | Décision |
|---|---|---|---|---|---|
| **Facteur $A$** | $sc_A$ | $n_A$ | $s_A^2 = \dfrac{sc_A}{n_A}$ | $f_A = \dfrac{s_A^2}{s_{AB}^2}$ | $H_0'$ ou $H_1'$ |
| **Facteur $B$** | $sc_B$ | $n_B$ | $s_B^2 = \dfrac{sc_B}{n_B}$ | $f_B = \dfrac{s_B^2}{s_{AB}^2}$ | $H_0''$ ou $H_1''$ |
| **Interaction** | $sc_{AB}$ | $n_{AB}$ | $s_{AB}^2 = \dfrac{sc_{AB}}{n_{AB}}$ | $f_{AB} = \dfrac{s_{AB}^2}{s_R^2}$ | $H_0'''$ ou $H_1'''$ |
| **Résiduelle** | $sc_R$ | $n_R$ | $s_R^2 = \dfrac{sc_R}{n_R}$ | | |
| **Totale** | $sc_{TOT}$ | $n_{TOT}$ | | | |

donnés par les formules suivantes :

$$\begin{aligned}
\hat{\mu} &= Y_{\bullet,\bullet,\bullet}\,, \\
\hat{\sigma}_A^2 &= \frac{1}{JK}\left(S_A^2 - S_{AB}^2\right) \quad ; \quad \hat{\sigma}_B^2 = \frac{1}{IK}\left(S_B^2 - S_{AB}^2\right)\,, \\
\sigma_{AB}^2 &= \frac{1}{K}\left(S_{AB}^2 - S_R^2\right) \\
\hat{\sigma}^2 &= \frac{SC_R}{(I-1)(J-1)} = S_R^2\,,
\end{aligned}$$

où $S_A^2 = \dfrac{SC_A}{n_A}$, $S_B^2 = \dfrac{SC_B}{n_B}$, $S_{AB}^2 = \dfrac{SC_{AB}}{n_{AB}}$ et $S_R^2 = \dfrac{SC_R}{n_R}$. Ces estimateurs sont non biaisés.

Les estimations, obtenues pour une liste de données expérimentales $\boldsymbol{y}$, notées $\hat{\mu}(\boldsymbol{y})$, $\hat{\sigma}_A^2(\boldsymbol{y})$, $\hat{\sigma}_B^2(\boldsymbol{y})$, $\hat{\sigma}_{AB}^2(\boldsymbol{y})$, $\hat{\sigma}^2(\boldsymbol{y})$ des paramètres $\mu$, $\sigma_A^2$, $\sigma_B^2$, $\sigma_{AB}^2$ et $\sigma^2$ du modèle, se déduisent immédiatement des

 

formules ci-dessus :

$$\begin{aligned}
\hat{\mu}(\boldsymbol{y}) &= y_{\bullet,\bullet,\bullet}\,, \\
\hat{\sigma}^2_A(\boldsymbol{y}) &= \frac{1}{JK}\left(s^2_A - s^2_{AB}\right) \quad ; \quad \hat{\sigma}^2_B(\boldsymbol{y}) = \frac{1}{IK}\left(s^2_B - s^2_{AB}\right)\,, \\
\hat{\sigma}^2_{AB}(\boldsymbol{y}) &= \frac{1}{K}\left(s^2_{AB} - s^2_R\right) \\
\hat{\sigma}^2(\boldsymbol{y}) &= \frac{sc_R}{(I-1)(J-1)} = s^2_R\,.
\end{aligned}$$

**Exemple 6.2.2**

*Les responsables d'un laboratoire d'analyse chimique par spectrométrie dans le proche infrarouge se sont intéressés à la variabilité des résultats qu'ils obtenaient pour les mesures des teneurs en protéines du blé. En particulier, ils se sont interrogés sur l'importance des différences qui pouvaient découler des étapes successives de préparation des matières à analyser. Nous considérons ici le problème du broyage, en examinant les résultats obtenus à l'aide de trois moulins différents.*

*Cinq échantillons de grains de blé ont été prélevés au hasard dans un arrivage relativement important et divisés chacun en trois sous-échantillons. Pour chacun des échantillons, les sous-échantillons ont ensuite été affectés au hasard pour trois moulins choisis au hasard dans une production de moulins et deux analyses chimiques ont été effectuées dans chaque cas.Le tableau ci-dessous présente les teneurs en protéines, exprimées en pourcentages de la matière sèche :*

| **Échantillons** / **Moulins** | **Ech. 1** | **Ech. 2** | **Ech. 3** | **Ech. 4** | **Ech. 5** |
|---|---|---|---|---|---|
| **Moul. 1** | *13,33* | *13,62* | *13,53* | *13,60* | *13,97* |
| | *13,43* | *13,33* | *13,75* | *13,44* | *13,32* |
| **Moul. 2** | *13,04* | *13,26* | *13,49* | *13,05* | *13,28* |
| | *13,34* | *13,49* | *13,59* | *13,44* | *13,67* |
| **Moul. 3** | *13,24* | *13,33* | *13,07* | *13,47* | *13,46* |
| | *13,25* | *13,46* | *13,33* | *13,04* | *13,32* |

*On cherche à analyser l'homogénéité des moulins, au sens où ils donnent les mêmes teneurs en protéines après broyage.*

*Il s'agit bien, dans ce cas de figure, d'une analyse de la variance à deux facteurs aléatoires avec répétitions.*

*Nous supposons que les conditions du modèle sont bien remplies.*

*L'étude de la variabilité des résultats en spectrométrie infrarouge est donnée ci-dessous dans le tableau d'analyse de la variance associé :*

| **Source** | **Variations** | **D.l.l.** | **Carré moyen** | ***F*** | **Proba critique** |
|---|---|---|---|---|---|
| *Moulin* | *0,29246* | *2* | *0,14623* | *8,70* | *0,010* |
| *Echant* | *0,20731* | *4* | *0,05183* | *3,08* | *0,082* |
| *Moul*Echant* | *0,13451* | *8* | *0,01681* | *0,38* | *0,917* |
| *Erreur* | *0,66840* | *15* | *0,04456* | | |
| *Total* | *1,30268* | *29* | | | |

*Pour le premier test, $P = 0,010$,et nous décidons donc, au seuil $\alpha = 0,05$, de refuser l'hypothèse nulle $H_0'$. Par conséquent, nous pouvons affirmer qu'il y a un effet significatif du facteur aléatoire "moulin".*

*Pour le second test, $P = 0,082$, et nous décidons au seuil $\alpha = 0,05$, de ne pas refuser l'hypothèse nulle $H_0''$. Par conséquent, nous n'avons pas réussi à mettre en évidence d'effet du facteur aléatoire "échantillon".*

*Pour le troisième test, $P = 0,917$, et nous décidons au seuil $\alpha = 0,05$, de ne pas refuser l'hypothèse nulle $H_0'''$. Par conséquent, nous n'avons pas réussi à mettre en évidence d'effet du facteur aléatoire "interaction".*

## 6.3 Modèles à effets mixtes

### 6.3.1 Modèles à effets mixtes sans répétition

Un facteur contrôlé $A$, donc à effets fixes, se présente sous la forme de $I$ modalités, chacune d'elles étant notée $A_i$. Nous admettrons que les effets des $B_j$, les $\beta_j$, représentent un échantillon de taille $J$ prélevé dans une population importante, et que les $\beta_j$ ont une loi normale centrée de variance $\sigma_B^2$.

Pour chacun des couples de modalités $(A_i, B_j)$, on effectue une unique mesure d'une réponse $Y$ qui est une variable continue. Nous noterons, là encore, $n = I \times J$ le nombre total de mesures ayant été effectuées.

On introduit le modèle suivant :

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

avec la contrainte supplémentaire : $\sum_{i=1}^{I} \alpha_i = 0$, où $Y_{i,j}$ est la valeur prise par la variable réponse $Y$ dans les conditions $(A_i, B_j)$.

On suppose aussi également que :

$$\mathcal{L}(\beta_j) = \mathcal{N}(0, \sigma_B^2) \, , \quad \forall \, j \, : \, 1 \leq j \leq J \, ,$$

ainsi que l'indépendance des effets aléatoires :

$$\beta_i \text{ indépendant de } \beta_j \text{ si } i \neq j \text{ et } 1 \leq i, j \leq J \, .$$

Nous postulons les hypothèses classiques suivantes pour les erreurs :

$$\forall \, (i,j) \, , \, 1 \leq i \leq I, \, 1 \leq j \leq J, \; \mathcal{L}(\varepsilon_{i,j}) = \mathcal{N}(0, \sigma^2) \, ,$$

$$\varepsilon_{i,j} \text{ indépendant de } \varepsilon_{k,l} \text{ si } (i,j) \neq (k,l) \text{ avec } 1 \leq i, k \leq I \text{ et } 1 \leq j, l \leq J$$

ainsi que l'indépendance des effets aléatoires et des erreurs :

$$\beta_i \text{ indépendant de } \varepsilon_{j,k} \text{ si } 1 \leq j \leq I \text{ et } 1 \leq i, k \leq J \, .$$

 

Nous supposons que les conditions d'utilisation de ce modèle sont satisfaites. Une fois encore, nous utilisons les quantités $SC_A$, $SC_B$, $SC_R$, $SC_{TOT}$, $sc_A$, $sc_B$, $sc_R$ et $sc_{TOT}$ introduites dans la section 6.1.2.

La relation fondamentale de l'analyse de variance s'écrit ici aussi :

$$SC_{TOT} = SC_A + SC_B + SC_R\,.$$

Nous avons les degrés de liberté suivants, associés à chaque ligne du tableau de l'analyse de variance :

| **Source** | **Degrés de liberté** |
|---|---|
| **Facteur *A*** | $n_A = I - 1$ |
| **Facteur *B*** | $n_B = J - 1$ |
| **Résiduelle** | $n_R = (I-1)(J-1)$ |
| **Totale** | $n_{TOT} = IJ - 1$ |

Nous résumons toutes ces informations dans le tableau de l'analyse de variance ci-dessous.

| **Source** | **Variations** | **d.d.l.** | **Carrés moyens** | ***F*** | **Décision** |
|---|---|---|---|---|---|
| **Facteur *A*** | $sc_A$ | $n_A$ | $s_A^2 = \dfrac{sc_A}{n_A}$ | $f_A = \dfrac{s_A^2}{s_R^2}$ | $H_0'$ ou $H_1'$ |
| **Facteur *B*** | $sc_B$ | $n_B$ | $s_B^2 = \dfrac{sc_B}{n_B}$ | $f_B = \dfrac{s_B^2}{s_R^2}$ | $H_0''$ ou $H_1''$ |
| **Résiduelle** | $sc_R$ | $n_R$ | $s_R^2 = \dfrac{sc_R}{n_R}$ | | |
| **Totale** | $sc_{TOT}$ | $n_{TOT}$ | | | |

L'analyse de la variance d'un modèle à effets mixtes facteurs aléatoires sans répétition permet deux tests de Fisher. Le premier test concernant le facteur fixe $A$ est le suivant :

$$H_0' : \alpha_1 = \alpha_2 = \cdots = \alpha_I = 0$$

contre

$$H_1' \ : \ \text{Il existe } i_0 \in \{1, 2, \cdots, I\} \text{ tel que } \alpha_{i_0} \neq 0.$$

Sous l'hypothèse nulle $(H_0')$ précédente, d'absence d'effet du facteur fixe $A$, et lorsque les conditions de validité du modèle sont respectées, $f_A$ est la réalisation d'une variable aléatoire qui suit une loi de Fisher à $I-1$ et $(I-1)(J-1)$ degrés de liberté. Nous concluons alors à l'aide de la valeur critique ("p-value"), et on rejette si elle est inférieure au seuil $\alpha$ du test. Lorsque l'hypothèse nulle $(H_0')$ est rejetée, on peut alors procéder à des comparaisons multiples des différents effets des niveaux du facteur.

Le second test concernant le second facteur aléatoire $B$ est le suivant :

$$H_0'' \ : \ \sigma_B^2 = 0$$

contre

$$H_1'' \ : \ \sigma_B^2 \neq 0\,.$$

Sous l'hypothèse nulle $(H_0'')$ précédente, d'absence d'effet du facteur $B$, et lorsque les conditions de validité du modèle sont respectées, $f_B$ est la réalisation d'une variable aléatoire qui suit une loi de Fisher à $J-1$ et $(I-1)(J-1)$ degrés de liberté.

Les estimateurs $\hat{\mu}$, $\hat{\alpha}_1, \cdots, \hat{\alpha}_I$ , $\hat{\sigma}_B^2$, $\hat{\sigma}^2$ des paramètres $\mu$, $\alpha_1, \cdots, \alpha_I$, $\sigma_B^2$, et $\sigma^2$ du modèle sont donnés par les formules suivantes :

$$\hat{\mu} = Y_{\bullet,\bullet,\bullet}\,,$$

$$\hat{\alpha}_i = Y_{i,\bullet,\bullet} - \hat{\mu}\,, 1 \leq i \leq I\,,$$

$$\hat{\sigma}_B^2 = \frac{1}{I}\left(S_B^2 - S_R^2\right)\,,$$

$$\hat{\sigma}^2 = \frac{SC_R}{(I-1)(J-1)} = S_R^2\,,$$

où $S_B^2 = \dfrac{SC_B}{n_B}$ et $S_R^2 = \dfrac{SC_R}{n_R}$. Ces estimateurs sont non biaisés.

Les estimations, obtenues pour une liste de données expérimentales $\boldsymbol{y}$, notées $\hat{\mu}(\boldsymbol{y})$, $\hat{\alpha}_1(\boldsymbol{y})$,$\cdots$, $\hat{\alpha}_I(\boldsymbol{y})$, $\hat{\sigma}_B^2(\boldsymbol{y})$ et $\hat{\sigma}^2(\boldsymbol{y})$ des paramètres $\mu$, $\alpha_1, \cdots, \alpha_I$, $\sigma_B^2$ et $\sigma^2$ du modèle, se déduisent immédiatement des formules ci-dessus :

$$\begin{aligned}
\hat{\mu}(\boldsymbol{y}) &= y_{\bullet,\bullet,\bullet}\,, \quad \hat{\alpha}_i = y_{i,\bullet,\bullet} - \hat{\mu}(\boldsymbol{y})\, 1 \leq i \leq I\,,\\
\hat{\sigma}_B^2(\boldsymbol{y}) &= \frac{1}{I}\left(s_B^2 - s_R^2\right)\,,\\
\hat{\sigma}^2(\boldsymbol{y}) &= \frac{sc_R}{(I-1)(J-1)} = s_R^2\,.
\end{aligned}$$

**Exemple 6.3.1**

*Nous reprenons les données de l'exemple 6.2.1 que nous avions étudié dans le cas de l'analyse à deux facteurs aléatoires sans répétition. Mais cette fois-ci, nous allons considérer le facteur « Temps » comme un facteur fixe. Par contre le facteur « Comprimé » reste toujours un facteur aléatoire.*

*Le modèle statistique s'écrit de la façon suivante :*

$$\boldsymbol{Y}_{i,j} = \mu + \alpha_i + \boldsymbol{B}_j + \boldsymbol{\varepsilon}_{i,j}$$

*où* $i = 1, \cdots, I$ *et* $j = 1, \cdots, J$, *avec la contrainte supplémentaire :* $\sum_{i=1}^{I} = \alpha_i = 0$,
*où* $\boldsymbol{Y}_{i,j}$ *est la valeur prise par la réponse* $\boldsymbol{Y}$ *dans les conditions* $(\alpha_i, \boldsymbol{B}_i)$.
*Notons* $n = I \times J$ *le nombre total de mesures ayant été effectuées.*

*Le tableau d'analyse de la variance pour Principe actif dissous est le suivant :*

| Source | D.l.l. | Variations | Carré moyen | $F$ | Proba critique |
|---|---|---|---|---|---|
| *Comprimé* | *5* | *83,21* | *16,64* | *0,68* | *0,647* |
| *Temps* | *3* | *4908,46* | *1636,15* | *66,6* | *0,000* |
| *Erreur* | *15* | *368,29* | *24,55* | | |
| *Total* | *23* | *5359,96* | | | |

*Pour le premier test, la probabilité critique vaut 0,647, et nous décidons de ne pas refuser l'hypothèse nulle* ($H_0$). *Par conséquent, nous n'avons pas réussi à mettre en évidence d'effet du facteur aléatoire « Comprimé ».*

*Pour le deuxième test, la probabilité critique vaut 0,000, nous décidons de refuser l'hypothèse nulle* ($H_0$). *Par conséquent, nous pouvons dire, au seuil* $\alpha = 5\%$, *qu'il y a un effet significatif du facteur fixe « Temps ».*

### 6.3.2 Modèles à effets mixtes avec répétitions

Un facteur contrôlé $A$, donc à effets fixes, se présente sous la forme de $I$ modalités, chacune d'elles étant notée $A_i$. Nous admettrons que les effets des $B_j$, les $\beta_j$, représentent un échantillon de taille $J$ prélevé dans une population importante, et que les $\beta_j$ ont une loi normale centrée de variance $\sigma_B^2$.

Pour chacun des couples de modalités $(A_i, B_j)$, on effectue $K$ mesures d'une réponse $Y$ qui est une variable continue. Nous noterons $n = I \times J \times K$ le nombre total de mesures ayant été effectuées.

On introduit le modèle, dit restreint :

$$Y_{i,j,k} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{i,j} + \varepsilon_{i,j,k} \quad i = 1, \cdots, I\,,\ j = 1, \cdots, J\,,\ k = 1, \cdots, K\,,$$

avec les contraintes supplémentaires : $\sum_{i=1}^{I} \alpha_i = 0$, et $\sum_{i=1}^{I} (\alpha\beta)_{i,j} = 0\,, \forall\, j \in \{1, \cdots, J\}$,

 

où $Y_{i,j,k}$ est la valeur prise par la réponse $Y$ dans les conditions $(A_i, B_j)$ lors du $k$-ième essai. Nous supposerons de plus que :

$$\mathcal{L}(\beta_j) = \mathcal{N}(0, \sigma_B^2)\,,\ \forall\, j\,, 1 \leq j \leq J\,,$$
$$\mathcal{L}((\alpha\beta)_{i,j}) = \mathcal{N}(0, \sigma_{AB}^2)\,,\ \forall\,(i,j)\,, 1 \leq i \leq I\,,\ 1 \leq j \leq J\,,$$

ainsi que l'indépendance des effets aléatoires :

$$\beta_i \text{ indépendant de } \beta_j \text{ si } i \neq j \text{ et } 1 \leq i, j \leq J\,,$$
$$\beta_i \text{ indépendant de } (\alpha\beta)_{j,k} \text{ si } 1 \leq j \leq I \text{ et } 1 \leq i, k \leq J\,.$$

On postule les hypothèses classiques suivantes pour les erreurs :

$$\forall\,(i,j,k)\,,\ 1 \leq i \leq I\,,\ 1 \leq j \leq J\,,\ 1 \leq k \leq K\,,\ \mathcal{L}(\varepsilon_{i,j,k}) = \mathcal{N}(0, \sigma^2)\,,$$
$$\text{et } \varepsilon_{i,j,k} \text{ indépendant de } \varepsilon_{l,m,n} \text{ si } (i,j,k) \neq (l,m,n) \text{ avec}$$
$$1 \leq i, l \leq I\,,\ 1 \leq j, m \leq J \text{ et } 1 \leq k, n \leq K\,,$$

ainsi que l'indépendance des effets aléatoires et des erreurs :

$$\beta_i \text{ indépendant de } \varepsilon_{j,k,l}\,,\ 1 \leq j \leq I\,,\ 1 \leq i, k \leq J \text{ et } 1 \leq l \leq K\,,$$
$$(\alpha\beta)_{i,j} \text{ indépendant de } \varepsilon_{k,l,m} \text{ si } 1 \leq i, k \leq I\,,\ 1 \leq j, l \leq J \text{ et } \ \leq m \leq K\,.$$

Dans un modèle mixte restreint, les effets aléatoires croisant des facteurs à effets fixes et à effets aléatoires, ici les $(\alpha\beta)_{i,j}$, ne sont pas mutuellement indépendants à cause des contraintes portant sur leur somme : $\displaystyle\sum_{i=1}^{I}(\alpha\beta)_{i,j} = 0\,, \forall\, j \in \{1, \cdots, J\}$. Par contre, ils le sont dès qu'on ne les considère pas tous en même temps.

Nous introduisons aussi le modèle, dit non restreint, suivant :

$$Y_{i,j,k} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{i,j} + \varepsilon_{i,j,k} \quad i = 1, \cdots, I\,,\ j = 1, \cdots, J\,,\ k = 1, \cdots, K\,,$$

avec la contrainte supplémentaire : $\displaystyle\sum_{i=1}^{I}\alpha_i = 0,$

où $Y_{i,j,k}$ est la valeur prise par la réponse $Y$ dans les conditions $(A_i, B_j)$ lors du $k$-ième essai. Nous supposerons de plus que :

$$\mathcal{L}(\beta_j) = \mathcal{N}(0, \sigma_B^2)\,,\ \forall\, j\,, 1 \leq j \leq J\,,$$
$$\mathcal{L}((\alpha\beta)_{i,j}) = \mathcal{N}(0, \sigma_{AB}^2)\,,\ \forall\,(i,j)\,, 1 \leq i \leq I\,,\ 1 \leq j \leq J\,,$$

ainsi que l'indépendance des effets aléatoires :

$$\beta_i \text{ indépendant de } \beta_j \text{ si } i \neq j \text{ et } 1 \leq i, j \leq J\,,$$
$$(\alpha\beta)_{i,j} \text{ indépendant de } (\alpha\beta)_{k,l} \text{ si } (i,j) \neq (k,l) \text{ avec } 1 \leq i, k \leq I \text{ et } 1 \leq j, l \leq J\,,$$
$$\beta_i \text{ indépendant de } (\alpha\beta)_{j,k} \text{ si } 1 \leq j \leq I \text{ et } 1 \leq i, k \leq J\,.$$

Nous postulons là aussi les hypothèses classiques suivantes pour les erreurs :

$$\forall\,(i,j,k)\,,\ 1 \leq i \leq I\,,\ 1 \leq j \leq J\,,\ 1 \leq k \leq K\,,\ \mathcal{L}(\varepsilon_{i,j,k}) = \mathcal{N}(0, \sigma^2)\,,$$
$$\text{et } \varepsilon_{i,j,k} \text{ indépendant de } \varepsilon_{l,m,n} \text{ si } (i,j,k) \neq (l,m,n) \text{ avec}$$
$$1 \leq i, l \leq I\,,\ 1 \leq j, m \leq J \text{ et } 1 \leq k, n \leq K\,,$$

 

ainsi que l'indépendance des effets aléatoires et des erreurs :

$$\beta_i \text{ indépendant de } \varepsilon_{j,k,l}\,,\ 1 \leq j \leq I\,,\ 1 \leq i,k \leq J \text{ et } 1 \leq l \leq K\,,$$
$$(\alpha\beta)_{i,j} \text{ indépendant de } \varepsilon_{k,l,m} \text{ si } 1 \leq i,k \leq I\,,\ 1 \leq j,l \leq J \text{ et } \ \leq m \leq K\,.$$

Dans un modèle mixte non restreint, les effets aléatoires croisant des facteurs à effets fixes et à effets aléatoires, ici les $(\alpha\beta)_{i,j}$, sont mutuellement indépendants. Il n'existe aucun consensus sur une raison statistique quelconque qui permettrait de privilégier l'une ou l'autre de ces approches. Nous utiliserons plutôt des modèles restreints.

Nous supposons que les conditions d'utilisation de ce modèle sont satisfaites. Une fois encore, nous utilisons les quantités $SC_A$, $SC_B$, $SC_{AB}$, $SC_R$, $SC_{TOT}$, $sc_A$, $sc_B$, $sc_{AB}$, $sc_R$ et $sc_{TOT}$ introduites dans la section 6.1.2.

La relation fondamentale de l'analyse de variance s'écrit ici aussi :

$$SC_{TOT} = SC_A + SC_B + SC_{AB} + SC_R\,.$$

On introduit une fois encore le nombre de degrés de liberté associés à chaque ligne du tableau de l'analyse de variance (voir tableau ci-dessous) Nous résumons enfin ces informations dans le tableau

| **Source** | **Degrés de liberté** |
|---|---|
| **Facteur $A$** | $n_A = I - 1$ |
| **Facteur $B$** | $n_B = J - 1$ |
| **Interaction $AB$** | $n_{AB} = (I-1)(J-1)$ |
| **Résiduelle** | $n_R = IJ(K-1)$ |
| **Totale** | $n_{TOT} = IJK - 1$ |

d'analyse de variance ci-dessus.

| **Sources de variation** | **Degrés de liberté** | **Variations** | **Carrés moyens** | **Statistique** F | **Décision** |
|---|---|---|---|---|---|
| Facteur $A$ | $n_A = I - 1$ | $sc_A$ | $s_A^2 = \dfrac{sc_A}{n_A}$ | $f_A = \dfrac{s_A^2}{s_R^2}$ | $H_0'$ ou $H_1'$ |
| Facteur $B$ | $n_B = J - 1$ | $sc_B$ | $s_B^2 = \dfrac{sc_B}{n_B}$ | $f_B = \dfrac{s_B^2}{s_{AB}^2}$ | $H_0''$ ou $H_1''$ |
| Interaction | $n_{AB} = (I-1)(J-1)$ | $sc_{AB}$ | $s_{AB}^2 = \dfrac{sc_{AB}}{n_{AB}}$ | $f_{AB} = \dfrac{s_{AB}^2}{s_R^2}$ | $H_0'''$ ou $H_1'''$ |
| Résidus | $n_R = IJ(K-1)$ | $sc_R$ | $s_R^2 = \dfrac{sc_R}{n_R}$ | | |
| Total | $n_{TOT} = IJK - 1$ | $sc_{TOT}$ | $s_T^2 = \dfrac{sc_{TOT}}{n_{TOT}}$ | | |

Nous désirons faire les trois tests d'hypothèse suivants :

$$H_0' \,:\, \alpha_1 = \alpha_2 = \cdots = \alpha_I = 0$$

contre

$$H_1' \,:\, \text{ Il existe } i_0 \in \{1, 2, \cdots, I\} \text{ tel que } \alpha_{i_0} \neq 0.$$

Sous l'hypothèse nulle $(H_0')$ précédente, d'absence d'effet du facteur fixe $A$, et lorsque les conditions de validité du modèle sont respectées, $f_A$ est la réalisation d'une variable aléatoire qui suit une loi de Fisher à $I - 1$ et $IJ(K-1)$ degrés de liberté. Nous concluons alors à l'aide de la valeur critique ("p-value"), et on rejette si elle est inférieure au seuil $\alpha$ du test. Lorsque l'hypothèse nulle $(H_0')$ est rejetée, on peut alors procéder à des comparaisons multiples des différents effets des niveaux du facteur.

Le second test concernant le second facteur aléatoire $B$ est le suivant :

$$H_0'' \,:\, \sigma_B^2 = 0$$

contre

$$H_1'' \,:\, \sigma_B^2 \neq 0\,.$$

Sous l'hypothèse nulle $(H_0'')$ précédente, d'absence d'effet du facteur $B$, et lorsque les conditions de validité du modèle sont respectées, $f_B$ est la réalisation d'une variable aléatoire qui suit une loi de Fisher à $J - 1$ et $(I-1)(J-1)$ degrés de liberté.

Le troisième test concerne l'interaction entre les facteurs $A$ et $B$ :

$$H_0''' : \sigma^2_{AB} = 0$$

contre

$$H_1''' : \sigma^2_{AB} \neq 0\,.$$

Sous l'hypothèse nulle ($H_0'''$) précédente, d'absence d'effet de l'interaction entre les facteurs $A$ et $B$, et lorsque les conditions de validité du modèle sont respectées, $f_{AB}$ est la réalisation d'une variable aléatoire qui suit une loi de Fisher à $(I-1)(J-1)$ et $IJ(K-1)$ degrés de liberté.

**Exemple 6.3.2**

*Eysenck (1974) a mené une étude consacrée à la rétention de matériel verbal en fonction du niveau de traitement. Elle faisait varier aussi bien l'âge que la condition de rétention. Le modèle de la mémorisation proposé par Craik et Lockhart (1972) stipule que le degré auquel un sujet se rappelle un matériel verbal est fonction du degré auquel ce matériel a été traité lors de sa présentation initiale. Ainsi, si l'on essaie de mémoriser une liste de mots, répéter simplement un mot pour soi-même (un niveau de traitement très bas) ne permet pas de le mémoriser aussi bien que si l'on y réfléchit en tentant de former des associations entre ce mot et un autre. Eysenck (1974) voulait tester ce modèle et, plus important encore, examiner s'il pouvait contribuer à expliquer certaines différences relevées entre des sujets jeunes et âgés concernant leur aptitude à se rappeler du matériel verbal. Eysenck a réparti aléatoirement 50 sujets âgés de 55 à 65 ans dans cinq groupes ; les quatre premiers impliquaient un apprentissage involontaire et le dernier un apprentissage intentionnel (l'apprentissage involontaire se caractérisait par le fait que le sujet ne savait pas qu'il devrait plus tard se rappeler le matériel appris).*

*Le premier groupe (addition) devait lire une liste de mots et se contenter de compter le nombre de lettres de chacun d'eux. Il s'agissait du niveau de traitement le plus bas, puisqu'il n'était pas nécessaire de mémoriser chaque mot autrement que comme une suite de lettres.*

*Le deuxième groupe (rimes) devait lire chaque mot et lui trouver une rime. Cette tâche impliquait de considérer la consonance de chaque mot, mais pas sa signification.*

*Le troisième groupe (adjectifs) devait donner un adjectif qui aurait pu être utilisé pour modifier chaque mot de la liste.*

*Le quatrième groupe (images) devait essayer de se former une image précise de chaque mot. Cette dernière tâche était supposée nécessiter le niveau de traitement le plus élevé parmi les quatre groupes d'apprentissage involontaire.*

*Aucun de ces groupes ne savait qu'il faudrait se rappeler les mots ultérieurement.*

*Enfin, le groupe d'apprentissage intentionnel devait lire la liste et mémoriser tous les mots. Après avoir passé trois fois en revue la liste de 27 mots, les sujets devaient retranscrire tous les mots dont ils se souvenaient.*

*Si l'apprentissage n'impliquait rien de plus qu'une exposition au matériel (soit la façon dont la plupart d'entre nous lisent le journal ou, pis encore, un devoir), les cinq groupes devaient obtenir des résultats identiques ; après tout, ils avaient tous vu tous les mots. Si le niveau de traitement était important, on devait constater des différences sensibles entre les moyennes des groupes.*

*L'étude incluait 50 participants dont l'âge se situait entre 18 et 30 ans, ainsi que 50 participants compris dans la tranche d'âge 55-65 ans. Pour plus de facilité, nous avons regroupé les 50 participants dont l'âge se situait entre 18 et 30 ans dans une classe que nous appellerons « sujets jeunes » et les 50 participants dont l'âge se situait entre 55 et 65 ans dans une classe que nous allons appeler « sujets âgés ».*

*Les données sont présentées dans le tableau suivant :*

| Addition | Rimes | Adjectifs | Images | Intentionnel |
|---|---|---|---|---|
| 8 | 10 | 14 | 20 | 21 |
| 6 | 7 | 11 | 16 | 19 |
| 4 | 8 | 18 | 16 | 17 |
| 6 | 10 | 14 | 15 | 15 |
| 7 | 4 | 13 | 18 | 22 |
| 6 | 7 | 22 | 16 | 16 |
| 5 | 10 | 17 | 20 | 22 |
| 7 | 6 | 16 | 22 | 22 |
| 9 | 7 | 12 | 14 | 18 |
| 7 | 7 | 11 | 19 | 21 |
| 9 | 7 | 11 | 12 | 10 |
| 8 | 9 | 13 | 11 | 19 |
| 6 | 6 | 8 | 16 | 14 |
| 8 | 6 | 6 | 11 | 5 |
| 10 | 6 | 14 | 9 | 10 |
| 4 | 11 | 11 | 23 | 11 |
| 6 | 6 | 13 | 12 | 14 |
| 5 | 3 | 13 | 10 | 15 |
| 7 | 8 | 10 | 19 | 11 |
| 7 | 7 | 11 | 11 | 11 |

*Les résultats de l'analyse de la variance se trouvent dans le tableau ci-dessous.*

| Source | D.l.l. | Variations | Carré moyen | $F$ | Proba critique |
|---|---|---|---|---|---|
| *Âge* | 1 | 240,25 | 240,25 | 5,05 | 0,027 |
| *Méthode* | 4 | 1541,94 | 378,73 | 7,96 | 1,53e-05 |
| *Âge * Méthode* | 4 | 190,30 | 47,57 | 5,93 | 0,000 |
| *Erreur* | 90 | 722,30 | 8,03 | | |
| *Total* | 99 | 2667,79 | | | |

*Analysons les résultats :*

1. *Pour le premier test, la probabilité critique vaut 0,027, nous décidons de refuser l'hypothèse nulle ($H_0$). Par conséquent, nous avons réussi à mettre en évidence un effet du facteur aléatoire « Âge ». Le risque associé à cette décision est un risque de deuxième espèce. Pour l'évaluer, il resterait à calculer la puissance de ce test.*

2. *Pour le deuxième test, la probabilité critique vaut 0,001, nous décidons de refuser l'hypothèse nulle ($H_0$). Par conséquent, nous pouvons dire, au seuil $\alpha = 5\%$, qu'il y a un effet significatif du facteur fixe « Méthode ».*
3. *Pour le troisième test, la probabilité critique vaut 0,000, nous décidons de refuser l'hypothèse nulle ($H_0$). Par conséquent, nous pouvons dire, au seuil $\alpha = 5\%$, qu'il y a un effet significatif du facteur aléatoire « Interaction ».*

# Exercices

**Exercice 1**

Trois cafés ont été dégustés par 6 juges. Le tableau ci-dessous fournit les notes d'acidité accordées par les 6 juges aux différents cafés :

| **Juges** | **Café 1** | **Café 2** | **Café 3** |
|---|---|---|---|
| Juge 1 | 0 | 3 | 4 |
| Juge 2 | 2 | 3 | 6 |
| Juge 3 | 3 | 5 | 7 |
| Juge 4 | 3 | 6 | 7 |
| Juge 5 | 5 | 6 | 8 |
| Juge 6 | 6 | 8 | 10 |

Les notes sont attribuées sur la base d'une échelle allant de 0 (café très peu acide) à 10 (café très acide).

1. On se pose la question de savoir si, en moyenne, certains cafés sont perçus plus acides que d'autres. Dans cette perspective, on réalise dans un premier temps une analyse de variance à un facteur, le facteur *café* (à trois modalités). Écrire le modèle correspondant et complétez le tableau d'analyse de variance suivant :

| **Variabilité** | **Variations** | **D.l.l.** | **Carrés moyens** | ***F*** |
|---|---|---|---|---|
| due au type de café | 44,111 | ... | ... | ... |
| due au résidu | 61,667 | ... | ... | |
| totale | ... | ... | | |

2. Testez l'hypothèse selon laquelle les 3 cafés présentent en moyenne une acidité identique. Donnez vos conclusions au risque 5 % (puis 1 %).
3. Quelle interprétation concrète donner à l'effet *juge* ? Est-il vraiment intéressant, lorsqu'on s'intéresse seulement à l'effet *café*, de prendre en compte l'effet *juge* dans le modèle d'analyse de variance ci-dessus ?

4. Écrire le modèle d'analyse de la variance comportant juste les deux effets *café* et *juge*, et concluez.
5. Commentez les façons de noter des juges 1 et 3. Quel café achèteriez-vous si vous préférez les cafés peu acides ?

**Exercice 2**

Lors d'une évaluation sensorielle, 31 personnes ont jugé 6 compotes de pomme sur la base de critères relatifs à l'odeur, l'aspect, la texture et la saveur. À l'issue du test, chacun attribue à chaque produit une note, dite note hédonique, allant de 0 (je n'aime pas du tout) à 10 (j'aime énormément). Les données se trouvent dans le fichier "chap6_ex2.csv" contenant $31 \times 6 = 186$ données.

1. Tracez les boîtes à moustaches en fonction de chaque compote.
2. Proposez une méthode statistique permettant d'étudier l'influence de la compote sur la note hédonique.
3. Faites une analyse de variance à un facteur en fonction du seul facteur *compote*. Quel modèle utilisez-vous ? Y a-t-il des différences entre compotes ? Quelle compote est la plus appréciée ?
4. Reprendre l'étude en intégrant en plus l'effet *juge*. Qu'en conclure ?

**Exercice 3**

On fait une étude sur l'évolution de la viscosité de crème liquide dans le temps. On mesure la viscosité à $J + 5$ (5 jours après la date de fabrication) et à $J + 30$. Plus la valeur de la mesure est élevée et plus la crème est visqueuse. Les crèmes sont des crèmes UHT, et donc la viscosité ne doit pas trop évoluer dans le temps. On suppose que la variance des viscosités est la même aux deux dates.

Pour cette étude, on réalise une expérience dans laquelle on mesure la viscosité de 22 crèmes à $J + 5$ et de 22 autres à $J + 30$. Les données se trouvent dans le fichier "chap6_ex3.csv".

1. Quel test d'hypothèse formalise la question suivante : "y a-t-il une différence de viscosité des crèmes entre les deux dates ?"
2. Décrire complètement la procédure de test pour conclure éventuellement à l'existence d'une différence ou non. Commentez les résultats obtenus. Quelle décision prendre sur cette base ?
3. À partir de maintenant, on suggère qu'il serait plus judicieux de comparer les viscosités à $J + 5$ et à $J + 30$ à partir de crèmes provenant d'un même lot, car les crèmes d'un même lot sont homogènes. On décide alors de faire une nouvelle expérience en choisissant 22 lots de production et en effectuant, pour chaque lot, une mesure à $J + 5$ et une à $J + 30$. On est donc confronté à un problème de comparaison de moyennes dites *appariées*.

   Quels facteurs peut-on prendre en compte avec cette nouvelle structure des données ? Rappeler l'équation qui décompose la variabilité de la viscosité en fonction de ces deux facteurs. En déduire des procédures de test de ces deux facteurs. Quel est, a priori, l'intérêt de cette nouvelle procédure par rapport à la précédente ?
4. Quelle conclusion tirez-vous à l'issue du test ?

 

5. Comparez les résultats obtenus vis-à-vis des résultats obtenus à la question 2.
6. Cette procédure peut-elle être facilement étendue au cas de la comparaison entre plus de deux dates ?

**Exercice 4**

On évalue l'efficacité d'un nouveau traitement ayant pour objet d'améliorer le développement global des enfants atteints de trisomie 21. Pour cela, une étude a été menée auprès de 12 enfants trisomiques. Six d'entre eux ont reçu un produit actif alors que les six autres ont reçu un placebo, et ce durant 6 mois. Un indice de développement global de chaque enfant a été calculé avant et après le début de l'étude par un même psychologue. Cet indice de développement global résume l'ensemble des capacités en terme de coordination, posture, langage et sociabilité. La nature du traitement donné n'est connu ni de la famille du patient, ni du psychologue. Deux psychologues ont participé à l'étude. Les données se trouvent dans le fichier "chap6_ex4.csv".

1. Proposez un modèle permettant de mettre en évidence un éventuel effet *traitement*.
2. Peut-on considérer que le nouveau traitement est efficace ?

**Exercice 5**

Le tableau ci-dessous présente le rendement de deux variétés de plantes lorsque trois types de fongicides ont été appliqués.

| | **Fongicide 1** | **Fongicide 2** | **Fongicide 3** | $\bar{y}_{i,\bullet,\bullet}$ |
|---|---|---|---|---|
| **Variété 1** | 1 | 1 | 2 | 2 |
| | 3 | 1 | 4 | |
| **Variété 2** | 2 | 4 | 4 | 4 |
| | 2 | 6 | 6 | |
| $\bar{y}_{\bullet,j,\bullet}$ | 2 | 3 | 4 | 3 |

1. Ecrire le modèle permettant d'étudier l'influence de la variété, du fongicide et de leurs interactions sur le traitement.
2. Après avoir noté que le plan d'expérience est complet et équirépété, donnez une estimation des paramètres du modèle.
3. Donnez les 12 valeurs du rendement prédites par le modèle. En déduire une estimation de $\sigma^2$.

**Exercice 6**

Pour étudier les facteurs influençant le rendement en blé, on a comparé trois variétés (L, N et NF) de blé et deux apports d'engrais azotés ( un apport "normal", la dose 1, et un apport "intensif", la dose 2). Trois répétitions pour chaque couple(variété, dose d'engrais) ont été effectuées et le rendement (en quintal par hectare) a été mesuré. On s'intéresse principalement aux différences qui pourraient exister d'une variété à l'autre, et aux interactions éventuelles des variétés avec les apports azotés. Les données se trouvent dans le fichier "chap6_ex6.csv".

1. Écrivez le modèle relatif à cette étude.
2. La personne en charge de cette étude hésite entre les deux méthodes suivantes :
   (a) Méthode 1 : conserver toutes les données ;
   (b) Méthode 2 : substituer aux trois valeurs observées pour un même couple (variété, dose d'engrais) leur valeur moyenne.

   Pour chacune des deux méthodes, donnez les degrés de liberté des différentes sources de variabilité présentes dans la table d'analyse de la variance. Quelle méthode utiliseriez-vous ? Pourquoi ?
3. Faites les calculs et déterminez les effets significatifs, en prenant soin de construire les tests en posant bien les hypothèses que vous voulez tester, la statistique de test sous $H_0$ et la décision que vous prenez. Quel modèle retenez-vous ?
4. Quelle variété et quelle dose d'azote conseilleriez-vous ?

**Exercice 7**

Lors d'un test hédonique, on s'intéresse à l'appréciation globale de trois chocolats. Pour cela, 45 juges ont participé à cette évaluation qui a eu lieu durant 2 jours (on dispose de 15 échantillons par

chocolat). Chaque juge n'a évalué qu'un chocolat. Comme chacun choisit son jour de dégustation et le chocolat qu'il évalue, le nombre de données et la répartition des chocolats évalués ne sont pas les mêmes d'un jour à l'autre.

On souhaite d'une part vérifier qu'il y a bien un effet *chocolat*, s'il y a un effet *jour* (les chocolats pouvant être plus ou moins appréciés lors du premier ou du second jour), et un effet interaction entre *chocolat* et *jour*.

| | **Chocolat 1** | | | **Chocolat 2** | | | **Chocolat 3** | | | $\bar{y}_{i,\bullet,\bullet}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| **Jour 1** | 5,2<br>5,2 | 6<br>6,6 | 5,4 | 4,2<br>4,4<br>4,4 | 5,2<br>5,4<br>5,6 | 5<br>4,8<br>6 | 4,6<br>4,8<br>5<br>5,4 | 5,6<br>5,8<br>6,2<br>6,4 | 5,2<br>6,4<br>5,2 | 5,36 |
| **Jour 2** | 3,2<br>4,2<br>4<br>4 | 4<br>3,8<br>4,4 | 3,6<br>4,2<br>4,6 | 3,2<br>3,4 | 3,6<br>4 | 4,2<br>3,6 | 3 | 3,8<br>4,4 | 3,4 | 3,83 |
| | 4,56 | | | 4,47 | | | 5,01 | | | |

1. Écrivez le modèle permettant de répondre à la problématique. Que signifie l'interaction entre les facteurs *chocolat* et *jour* ?
2. Quelles sont les estimations des différents paramètres de ce modèle ?
3. Y a-t-il un effet *jour*, un effet *chocolat* et un effet de l'interaction ?
4. Refaites un modèle d'analyse de la variance sans interaction. Commentez ce nouveau tableau d'analyse de variance et expliquez les résultats. Comparez les différences par rapport aux résultats obtenus ci-dessus.
5. Par chocolat, calculez la moyenne des notes et comparez la à la moyenne ajustée $(\hat{\mu} + \hat{\alpha}_i)$ que l'on peut calculer à l'aide des résultats du tableau de la question 4. Qu'en pensez-vous ? Quel est le chocolat préféré ?
6. Testez l'hypothèse $H_0$ : l'effet du chocolat 1 est nul (précisez les hypothèses du test, la statistique de test sous $H_0$, et la décision).

# Chapitre 7

# Analyse de la variance à deux facteurs emboîtés

Les modèles emboîtés d'analyse de la variance à deux facteurs correspondent à des situations où un des critères est subordonné à l'autre.

Ainsi, par exemple, quand on compare les productions laitières d'une même race bovine dans deux ou plusieurs régions, en choisissant au hasard et indépendamment plusieurs exploitations agricoles dans chaque région, et en mesurant dans chacune d'elles les productions laitières de plusieurs bêtes, elles aussi choisies au hasard et indépendamment les unes des autres, le facteur **exploitation** est alors subordonné au facteur **région**, puisque le choix des exploitations est réalisé à l'intérieur de chacune des régions, sans qu'il n'y ait aucune correspondance entre les différentes exploitations des différentes régions.

Dans ces conditions, il ne se justifie pas de calculer $\bar{x}_{\bullet,1,\bullet}$ qui serait relative aux premières exploitations des différentes régions. Par contre, il se justifie toujours de calculer les moyennes relatives à l'intérieur de chaque région, c'est-à-dire les moyennes $\bar{x}_{i,\bullet,\bullet}$ relatives au premier critère de classification.

Un point très important est que nous ne pourrons nous servir de modèles où les facteurs sont emboîtés que si nous disposons de répétitions. Dans le cas contraire où les mesures ne seraient pas répétées, le modèle que nous devrons alors utiliser pour analyser les données sera l'un de ceux déjà exposés au chapitre 6.

## 7.1 Modèles à effets fixes

On suppose qu'un facteur contrôlé $A$ possède $I$ modalités, chacune d'elle étant notée $A_i$. De même, on suppose qu'un facteur contrôlé $B$ possède $J$ modalités, chacune d'elle dépendant du niveau $A_i$ du facteur $A$, et étant notée $B_{j(i)}$. Pour chaque couple de modalités $(A_i, B_{j(i)})$, nous effectuons $K \geq 2$ mesures d'une variable réponse $Y$ qui est supposée être une variable continue. Nous noterons $n$ le nombre total de mesures effectuées : $n = I \times J \times K$.

Pour cela, on introduit le modèle :

$$Y_{i,j,k} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{i,j,k} \qquad , \qquad i = 1, \cdots, I \, ; \, j = 1, \cdots, J \, ; \, k = 1, \cdots, K \, .$$

avec les contraintes supplémentaires : $\sum_{i=1}^{I} \alpha_i = 0$ et $\sum_{j=1}^{J} \beta_{j(i)} = 0\,,\ \forall\, i \in \{1, \cdots, I\}$, où $Y_{i,j,k}$ est la valeur prise par la réponse $Y$ dans les conditions $(A_i, B_{j(i)})$ lors de la $k$-ème mesure.

On supposera toujours réalisées les hypothèses standards suivantes :

1. $\varepsilon_{i,j,k}$ et $\varepsilon_{l,m,n}$ sont indépendantes si $(i,j,k) \neq (l,m,n)$ avec $1 \leq i,l \leq I$ , $1 \leq j,m \leq J$ et $1 \leq k,n \leq K$.
2. $\forall\,(i,j,k)$, $i = 1, \cdots, I\,;\ j = 1, \cdots, J\,;\ k = 1, \cdots, K\ ;\ \mathcal{L}(\varepsilon_{i,j,k}) = \mathcal{N}(0, \sigma^2)$.

Nous supposerons que les conditions d'utilisation de ce modèle sont bien remplies.

Nous regroupons les valeurs prises par la variable réponse $Y$ dans les conditions $(A_i, B_j)$ lors des $K$ répétitions dans le tableau ci-dessous :

| $\boldsymbol{A_1}$ | | | $\cdots$ | | | $\boldsymbol{A_I}$ | | |
|---|---|---|---|---|---|---|---|---|
| $\boldsymbol{B_{1(1)}}$ | $\cdots$ | $\boldsymbol{B_{J(1)}}$ | $\cdots$ | $\cdots$ | $\cdots$ | $\boldsymbol{B_{1(I)}}$ | $\cdots$ | $\boldsymbol{B_{J(1)}}$ |
| $Y_{1,1,1}$ | $\cdots$ | $Y_{1,J,1}$ | $\cdots$ | $\cdots$ | $\cdots$ | $Y_{I,1,1}$ | $\cdots$ | $Y_{I,J,1}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\cdots$ | $\cdots$ | $\cdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $Y_{1,1,K}$ | $\cdots$ | $Y_{1,J,K}$ | $\cdots$ | $\cdots$ | $\cdots$ | $Y_{I,1,K}$ | $\cdots$ | $Y_{I,J,K}$ |

Il y a donc $I \times J \times K$ variables aléatoires $Y_{i,j,k}$. On peut, comme au chapitre 6, définir les différentes moyennes suivantes :

$$
\begin{aligned}
Y_{\bullet,j,k} &= \frac{1}{I}\sum_{i=1}^{I} Y_{i,j,k} \\
Y_{i,\bullet,k} &= \frac{1}{J}\sum_{j=1}^{J} Y_{i,j,k} \\
Y_{i,j,\bullet} &= \frac{1}{K}\sum_{k=1}^{K} Y_{i,j,k} \\
Y_{\bullet,\bullet,k} &= \frac{1}{I \times J}\sum_{i=1}^{I}\sum_{j=1}^{J} Y_{i,j,k} \\
Y_{\bullet,j,\bullet} &= \frac{1}{I \times K}\sum_{i=1}^{I}\sum_{k=1}^{K} Y_{i,j,k} \\
Y_{i,\bullet,\bullet} &= \frac{1}{J \times K}\sum_{j=1}^{J}\sum_{k=1}^{K} Y_{i,j,k} \\
Y_{\bullet,\bullet,\bullet} &= \frac{1}{I \times J \times K}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} Y_{i,j,k}
\end{aligned}
$$

Nous rappelons que la variation théorique due au facteur $A$ est définie par :

$$SC_A = JK\sum_{i=1}^{I}(Y_{i,\bullet,\bullet} - Y_{\bullet,\bullet,\bullet})^2\,.$$

 

La variation théorique du facteur $B$ dans le facteur $A$ est définie par :

$$SC_{B|A} = K \sum_{i=1}^{I} \sum_{j=1}^{J} (Y_{i,j,\bullet} - Y_{i,\bullet,\bullet})^2 .$$

La variation résiduelle théorique est quant à elle définie par :

$$SC_R = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1} K(Y_{i,j,k} - Y_{i,j,\bullet})^2 .$$

Enfin, la variation totale est égale à :

$$SC_{TOT} = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} (Y_{i,j,k} - Y_{\bullet,\bullet,\bullet})^2 .$$

Nous rappelons la relation fondamentale de l'ANOVA :

$$SC_{TOT} = SC_A + SC_{B|A} + SC_R .$$

La liste $\boldsymbol{y}$ des données expérimentales $y_{1,1,1}, \cdots, y_{1,1,K}, y_{1,2,1}, \cdots, j_{1,2,K}, \cdots, y_{I,J,K}$ permet de construire une réalisation du tableau précédent :

| $\boldsymbol{A_1}$ | | | $\cdots$ | | | $\boldsymbol{A_I}$ | | |
|---|---|---|---|---|---|---|---|---|
| $\boldsymbol{B_{1(1)}}$ | $\cdots$ | $\boldsymbol{B_{J(1)}}$ | $\cdots$ | $\cdots$ | $\cdots$ | $\boldsymbol{B_{1(I)}}$ | $\cdots$ | $\boldsymbol{B_{J(1)}}$ |
| $y_{1,1,1}$ | $\cdots$ | $y_{1,J,1}$ | $\cdots$ | $\cdots$ | $\cdots$ | $y_{I,1,1}$ | $\cdots$ | $y_{I,J,1}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\cdots$ | $\cdots$ | $\cdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $y_{1,1,K}$ | $\cdots$ | $y_{1,J,K}$ | $\cdots$ | $\cdots$ | $\cdots$ | $y_{I,1,K}$ | $\cdots$ | $y_{I,J,K}$ |

La variation due au facteur $A$ observée sur la liste $\boldsymbol{y}$ de données expérimentales est définie par :

$$sc_A = JK \sum_{i=1}^{I} (y_{i,\bullet,\bullet} - y_{\bullet,\bullet,\bullet})^2 .$$

La variation due au facteur $B$ dans le facteur $A$ observée sur la liste $\boldsymbol{y}$ de données expérimentales est définie par :

$$sc_{B|A} = K \sum_{i=1}^{I} \sum_{j=1}^{J} (y_{i,j,\bullet} - y_{i,\bullet,\bullet})^2 .$$

La variation résiduelle observée sur la liste $\boldsymbol{y}$ de données expérimentales est définie par :

$$sc_R = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} (y_{i,j,k} - y_{i,j,\bullet})^2 .$$

 

Enfin, la variation totale observée sur la liste $\boldsymbol{y}$ de données expérimentales est égale par :

$$sc_{TOT} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{i,j,k} - y_{\bullet,\bullet,\bullet})^2 \,.$$

La relation fondamentale de l'analyse de variance reste valable lorsqu'elle est évaluée sur la liste $\boldsymbol{y}$ de données expérimentales :

$$sc_{TOT} = sc_A + sc_{B|A} + sc_R \,.$$

On reconnaît parmi les quantités définies ci-dessus des quantités similaires à celles introduites dans les chapitres 3 et 6.

Nous remarquons que les nouvelles quantités : $SC_{B|A}$ et $sc_{B|A}$ sont liées aux relations précédentes par les relations :

$$SC_{B|A} = SC_B + SC_{AB}$$

$$sc_{B|A} = sc_B + sc_{AB}$$

On introduit les degrés de liberté associés à chaque ligne du tableau de l'ANOVA :

| **Source de variations** | de d.d.l. |
|---|---|
| Facteur $A$ | $n_A = I - 1$ |
| Facteur $B$ dans $A$ | $n_{B\|A} = I(J-1)$ |
| Résiduelle | $n_R = IJ(K-1)$ |
| Totale | $n_{TOT} = IJK - 1$ |

Nous résumons toutes ces informations dans le tableau d'ANOVA ci-dessous :

| **Sources de variation** | **Degrés de liberté** | **Variations** | **Carrés moyens** | **Statistique** F | **Décision** |
|---|---|---|---|---|---|
| Facteur $A$ | $n_A = I - 1$ | $sc_A$ | $s_A^2 = \dfrac{sc_A}{n_A}$ | $f_A = \dfrac{s_A^2}{s_R^2}$ | $H_0'$ ou $H_1'$ |
| Facteur $B$ dans facteur $A$ | $n_{B\|A} = I(J-1)$ | $sc_{B\|A}$ | $s_{B\|A}^2 = \dfrac{sc_{B\|A}}{n_{B\|A}}$ | $f_{B\|A} = \dfrac{s_{B\|A}^2}{s_R^2}$ | $H_0''$ ou $H_1''$ |
| Résiduelle | $n_R = IJ(K-1)$ | $sc_R$ | $s_R^2 = \dfrac{sc_R}{n_R}$ | | |
| Total | $n_{TOT} = IJK$-1 | $sc_{TOT}$ | $s_T^2 = \dfrac{sc_{TOT}}{n_{TOT}}$ | | |

Nous souhaitons faire les tests d'hypothèses suivants :

$$H'_0\,;\,\alpha_1=\alpha_2=\cdots=\alpha_I=0$$

contre

$$H'_1\,:\,\text{Il existe } i_0\in\{1,2,\cdots,I\} \text{ tel que } \alpha_{i_0}\neq 0\,.$$

Sous l'hypothèse nulle $H'_0$ d'absence d'effet du facteur $A$ et lorsque les conditions de validité du modèle sont respectées, $f_A$ est la réalisation de la variable aléatoire $S_A^2/S_R^2$ qui suit une loi de Fisher-Snedecor à $n_A=I-1$ et $n_R=IJ(K-1)$ degrés de liberté.

On peut alors conclure grâce à la valeur critique, et on rejette l'hypothèse nulle si elle est inférieure ou égale au seuil $\alpha$ du test, ou à l'aide d'une table. Il y a rejet si $f_A$ est supérieure ou égale à la valeur critique issue de la table. Si l'hypothèse $H'_0$ est rejetée, on pourra procéder à des comparaisons multiples des différents effets des niveaux du facteur.

Nous pouvons répéter ce qui précède pour le facteur $B$ :

$$H''_0\,;\,\beta_{1(1)}=\beta_{2(1)}=\cdots=\beta_{J(1)}=\cdots=\beta_{J(I)}=0$$

contre

$$H''_1\,:\,\text{Il existe } (i_0,j_0)\in\{1,2,\cdots,I\}\times\{1,2,\cdots,J\} \text{ tel que } \beta_{j_0(i_o)}\neq 0\,.$$

Sous l'hypothèse nulle $H''_0$ d'absence d'effet du facteur $B$ dans le facteur $A$ et lorsque les conditions de validité du modèle sont respectées, $f_{B|A}$ est la réalisation de la variable aléatoire $S_{B|A}^2/S_R^2$ qui suit une loi de Fisher-Snedecor à $n_B=I(J-1)$ et $n_R=IJ(K-1)$ degrés de liberté.

On peut alors conclure grâce à la valeur critique, et on rejette l'hypothèse nulle si elle est inférieure ou égale au seuil $\alpha$ du test, ou à l'aide d'une table. Il y a rejet si $f_B$ est supérieure ou égale à la valeur critique issue de la table. Si l'hypothèse $H''_0$ est rejetée, on pourra procéder à des comparaisons multiples des différents effets des niveaux du facteur.

Les estimateurs $\hat\mu$, $\hat\alpha_1,\cdots,\hat\alpha_I,\hat\beta_{1(1)},\hat\beta_{2(1)},\cdots,\hat\beta_{J(1)},\cdots,\hat\beta_{J(I)}$, et $\hat\sigma^2$ des paramètres respectifs $\mu$, $\alpha_1,\cdots,\alpha_I,\beta_{1(1)},\beta_{2(1)},\cdots,\beta_{J(1)},\cdots,\beta_{J(I)}$ et $\sigma^2$ du modèle sont donnés par :

$$\hat\mu=Y_{\bullet,\bullet,\bullet}\quad;\quad \hat\alpha_i=Y_{i,\bullet,\bullet}-\hat\mu,\qquad 1\leq i\leq I$$

$$\hat\beta_{j(i)}=Y_{i,j,\bullet}-Y_{i,\bullet,\bullet}\qquad 1\leq i\leq I,\quad 1\leq j\leq J$$

$$\hat\sigma^2=\frac{SC_R}{IJ(K-1)}=S_R^2\,.$$

Ce sont des estimateurs sans biais.

Les estimations obtenues pour une liste de données expérimentales $\boldsymbol{y}$, notées $\hat\mu(\boldsymbol{y})$, $\hat\alpha_1(\boldsymbol{y}),\cdots$, $\hat\alpha_I(\boldsymbol{y})$, $\hat\beta_{1(1)},(\boldsymbol{y}),\hat\beta_{2(1)},(\boldsymbol{y}),\cdots$, $\hat\beta_{J(1)}(\boldsymbol{y})$, $\cdots,\hat\beta_{J(I)}(\boldsymbol{y})$ et $\hat\sigma^2(\boldsymbol{y})$ des paramètres $\mu$, $\alpha_1$, $\cdots$, $\alpha_I$, $\beta_{1(1)}$, $\beta_{2(1)}$, $\cdots$, $\beta_{J(1)}$, $\cdots,\beta_{J(I)}$ et $\sigma^2$ du modèle se déduisent, mutatis mutandis, des formules précédentes.

**Exemple 7.1.1**

*L'expérience consiste à évaluer le gain de masse, en grammes, entre la dixième et la vingtième semaine, de poulets soumis à quatre régimes alimentaires obtenus en combinant des niveaux faibles*

*ou élevés de calcium et de lysine. Deux enclos de six poulets ont été utilisés pour chacun des quatre traitements.*

*Les deux facteurs,* **régime** *et* **enclos**, *sont contrôlés par l'expérimentateur.*

*Les données sont fournies dans le tableau ci-dessous :*

| | **Régime** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **LoCaLoL** | | **LoCaHiL** | | **HiCaLoL** | | **HiCaHiL** | |
| **Enclos** | *1* | *2* | *1* | *2* | *1* | *2* | *1* | *2* |
| | *573* | *1041* | *618* | *943* | *731* | *416* | *518* | *416* |
| **Gain** | *636* | *814* | *926* | *640* | *845* | *729* | *782* | *729* |
| **de** | *883* | *498* | *717* | *373* | *866* | *590* | *938* | *590* |
| **masse** | *550* | *890* | *677* | *907* | *729* | *552* | *755* | *552* |
| **en g.** | *613* | *636* | *659* | *734* | *770* | *776* | *672* | *776* |
| | *901* | *685* | *817* | *1050* | *787* | *657* | *576* | *657* |

*La signification des sigles ci-dessus est la suivante : par exemple, "LoCaLoL" signifie faible dose en calcium et faible dose en lysine, "HiCaLoL" signifie haute dose en calcium et faible dose en lysine, etc.*

*Le tableau de l'analyse de variance est le suivant :*

```
Analysis of Variance Table

Response: masse
              Df Sum Sq Mean Sq F value Pr(>F)
regime         3  53943   17981  0.7319 0.5391
regime:enclos  4 125688   31422  1.2791 0.2943
Residuals     40 982654   24566
```

*Nous supposons bien sûr que les conditions du modèle sont bien remplies.*

*Analysons les résultats :*

1. *Pour le premier test, la valeur critique vaut 0,5391 et nous décidons de ne pas refuser l'hypothèse nulle $H_0$. Par conséquent, nous n'avons pas réussi à mettre en évidence d'effet du facteur à effets fixes* **régime**. *Le risque associé à cette décision est un risque de seconde espèce, et pour l'évaluer, il resterait à calculer la puissance de ce test.*
2. *Pour le second test, la valeur critique vaut 0,2943 et nous décidons de ne pas refuser l'hypothèse nulle $H_0$. Par conséquent, nous n'avons pas réussi à mettre en évidence d'effet du facteur à effets fixes* **enclos** *dans le facteur* **régime**. *Le risque associé à cette décision est un risque de seconde espèce, et pour l'évaluer, il resterait à calculer la puissance de ce test.*

## 7.2 Modèles à effets aléatoires

Cette fois, les deux facteurs sont considérés aléatoires. Les termes $A_i$ représentent un échantillon de taille $I$ prélevé dans une population importante. Nous admettrons que les effets des $A_i$ sont de loi normale centrée de variance $\sigma_A^2$.

Les termes $B_{j(i)}$ représentent un échantillon de taille $J$ prélevé dans une population importante dépendant du niveau $A_i$ du facteur $A$. Nous admettrons que les effets des $B_{j(i)}$ sont de loi normale centrée de variance $\sigma^2_{B|A}$.

Pour chacun des couples $(A_i, B_{j(i)})$, nous effectuons $K \geq 2$ mesures d'une réponse $Y$ qui est une variable continue. Nous noterons $n = I \times J \times K$ le nombre total de mesures ayant été effectuées.

On introduit le modèle :

$$Y_{i,j,k} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{i,j,k} \qquad , \qquad i = 1, \cdots, I \, ; \, j = 1, \cdots, J \, ; \, k = 1, \cdots, K \, ,$$

où $Y_{i,j,k}$ est la valeur prise par la réponse $Y$ dans les conditions $(A_i, B_{j(i)})$ lors de la $k$-ème mesure.

Nous supposons que :

$$\begin{aligned} \mathcal{L}(\alpha_i) &= \mathcal{N}(0, \sigma^2_A), \quad \forall \, i \, , 1 \leq i \leq I \, , \\ \mathcal{L}(\beta_{j(i)}) &= \mathcal{N}(0, \sigma^2_{B|A}), \quad \forall \, (i,j) \, , 1 \leq i \leq I, \, 1 \leq j \leq J \, , \end{aligned}$$

ainsi que l'indépendance des effets aléatoires :

$$\alpha_i \text{ et } \alpha_j \text{ sont indépendants si } i \neq j \text{ et } 1 \leq i, j \leq I,$$

$$\beta_{j(i)} \text{ et } \beta_{l(k)} \text{ sont indépendants si } (i,j) \neq (k,l) \text{ avec } 1 \leq i, k \leq I \text{ et } 1 \leq j, l \leq J,$$

$$\alpha_i \text{ et } \beta_{k(j)} \text{ sont indépendants si } 1 \leq i, j \leq I \text{ et } 1 \leq k \leq J.$$

On supposera toujours réalisées les hypothèses standards suivantes :

1. $\varepsilon_{i,j,k}$ et $\varepsilon_{l,m,n}$ sont indépendantes si $(i,j,k) \neq (l,m,n)$ avec $1 \leq i, l \leq I$ , $1 \leq j, m \leq J$ et $1 \leq k, n \leq K$.
2. $\forall \, (i,j,k)$, $i = 1, \cdots, I \, ; \, j = 1, \cdots, J \, ; \, k = 1, \cdots, K$ ; $\mathcal{L}(\varepsilon_{i,j,k}) = \mathcal{N}(0, \sigma^2)$.

Nous supposerons que les conditions d'utilisation de ce modèle sont bien remplies.

Nous utilisons les quantités $SC_A$, $SC_{B|A}$, $SC_R$, $SC_{TOT}$,$sc_A$, $sc_{B|A}$, $sc_R$ et $sc_{TOT}$ introduites à la section précédente.

Nous rappelons la relation fondamentale de l'ANOVA :

$$SC_{TOT} = SC_A + SC_{B|A} + SC_R \, .$$

On introduit les degrés de liberté correspondant à chaque ligne du tableau de l'ANOVA :

| **Source** | **Nombre de d.d.l.** |
|---|---|
| **Facteur $A$** | $n_A = I - 1$ |
| **Facteur $B$ dans $A$** | $n_{B\|A} = I(J-1)$ |
| **Résiduelle** | $n_R = IJ(K-1)$ |
| **Totale** | $n_{TOT} = IJK - 1$ |

On peut résumer toutes ces informations dans le tableau de l'ANOVA ci-dessous :

| **Sources de variation** | **Degrés de liberté** | **Variations** | **Carrés moyens** | **Statistique** F | **Décision** |
|---|---|---|---|---|---|
| Facteur $A$ | $n_A = I - 1$ | $sc_A$ | $s_A^2 = \dfrac{sc_A}{n_A}$ | $f_A = \dfrac{s_A^2}{s_{B\|A}^2}$ | $H_0'$ ou $H_1'$ |
| Facteur $B$ dans facteur $A$ | $n_{B\|A} = I(J-1)$ | $sc_{B\|A}$ | $s_{B\|A}^2 = \dfrac{sc_{B\|A}}{n_{B\|A}}$ | $f_{B\|A} = \dfrac{s_{B\|A}^2}{s_R^2}$ | $H_0''$ ou $H_1''$ |
| Résiduelle | $n_R = IJ(K-1)$ | $sc_R$ | $s_R^2 = \dfrac{sc_R}{n_R}$ | | |
| Total | $n_{TOT} = IJK$-1 | $sc_{TOT}$ | $s_T^2 = \dfrac{sc_{TOT}}{n_{TOT}}$ | | |

L'analyse de variance à deux facteurs emboîtés permet deux tests de Fisher.

Le premier test concernant le facteur $A$ est le suivant :

$$H_0' \,:\, \sigma_A^2 = 0$$

contre

$$H_1' \,:\, \sigma_A^2 \neq 0\,.$$

Sous l'hypothèse nulle $(H_0')$ précédente, d'absence d'effet du facteur $A$, et lorsque les conditions de validité du modèle sont respectées, $f_A$ est la réalisation d'une variable aléatoire qui suit une loi de Fisher à $I-1$ et $(I-1)(J-1)$ degrés de liberté.

Le second test concernant le second facteur $B$ est le suivant :

$$H_0'' \,:\, \sigma_{B|A}^2 = 0$$

contre

$$H_1'' \,:\, \sigma_{B|A}^2 \neq 0\,.$$

Sous l'hypothèse nulle $(H_0'')$ précédente, d'absence d'effet du facteur $B$ dans le facteur $A$, et lorsque les conditions de validité du modèle sont respectées, $f_{B|A}$ est la réalisation d'une variable aléatoire qui suit une loi de Fisher à $I(J-1)$ et $IJ(K-1)$ degrés de liberté.

Les estimateurs $\hat{\mu}$, $\hat{\sigma}^2_A$, $\hat{\sigma}^2_{B|A}$, $\hat{\sigma}^2$ des paramètres $\mu$, $\sigma^2_A$, $\sigma^2_{B|A}$ et $\sigma^2$ du modèle sont donnés par les formules suivantes :

$$\begin{aligned}
\hat{\mu} &= Y_{\bullet,\bullet,\bullet}\,,\\
\hat{\sigma}^2_A &= \frac{1}{JK}\left(S^2_A - S^2_{B|A}\right) \quad ; \quad \hat{\sigma}^2_{B|A} = \frac{1}{K}\left(S^2_{B|A} - S^2_R\right)\,,\\
\hat{\sigma}^2 &= \frac{SC_R}{(I-1)(J-1)} = S^2_R\,,
\end{aligned}$$

où $S^2_A = \dfrac{SC_A}{n_A}$, $S^2_{B|A} = \dfrac{SC_{B|A}}{n_{B|A}}$ et $S^2_R = \dfrac{SC_R}{n_R}$. Ces estimateurs sont non biaisés.

Les estimations obtenues pour une liste de données expérimentales $\boldsymbol{y}$, notées $\hat{\mu}(\boldsymbol{y})$, $\hat{\sigma}^2_A(\boldsymbol{y})$, $\hat{\sigma}^2_{B|A}(\boldsymbol{y})$, $\hat{\sigma}^2(\boldsymbol{y})$ des paramètres $\mu$, $\sigma^2_A$, $\sigma^2_{B|A}$ et $\sigma^2$ du modèle, se déduisent immédiatement des formules ci-dessus :

$$\begin{aligned}
\hat{\mu}(\boldsymbol{y}) &= y_{\bullet,\bullet,\bullet}\,,\\
\hat{\sigma}^2_A(\boldsymbol{y}) &= \frac{1}{JK}\left(s^2_A - s^2_{B|A}\right) \quad ; \quad \hat{\sigma}^2_{B|A}(\boldsymbol{y}) = \frac{1}{K}\left(s^2_{B|A} - s^2_R\right)\,,\\
\hat{\sigma}^2(\boldsymbol{y}) &= \frac{sc_R}{(I-1)(J-1)} = s^2_R\,.
\end{aligned}$$

**Exemple 7.2.1**

*On a récolté des données d'une expérience conçue pour estimer la moisissure contenue dans une pâte de piment produite par une entreprise agro-alimentaire. Pour cela, quinze lots de pots de pâte de piment ont été sélectionnés au hasard dans la production de l'entreprise et dans chacun de ces lots, deux pots de pâte ont été à nouveau sélectionnés au hasard. Deux prélèvements distincts de pâte ont été analysés pour chacun de ces pots.*

*Remarquons que les deux facteurs,* **lot** *et* **échantillon***, sont tous les deux considérés comme des facteurs à effets aléatoires.*

*Les données sont fournies dans le tableau ci-dessous :*

| **Lot** | **1** | | **2** | | **3** | | **4** | | **5** | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Échant.** | **1** | **2** | **1** | **2** | **1** | **2** | **1** | **2** | **1** | **2** |
| **Analyses** | *40* | *30* | *26* | *25* | *29* | *14* | *30* | *24* | *19* | *17* |
| | *39* | *30* | *28* | *26* | *28* | *15* | *31* | *24* | *20* | *17* |
| **Lot** | **6** | | **7** | | **8** | | **9** | | **10** | |
| **Échant.** | **1** | **2** | **1** | **2** | **1** | **2** | **1** | **2** | **1** | **2** |
| **Analyses** | *33* | *26* | *23* | *32* | *34* | *29* | *27* | *31* | *13* | *27* |
| | *32* | *24* | *24* | *33* | *34* | *29* | *27* | *31* | *16* | *24* |
| **Lot** | **11** | | **12** | | **13** | | **14** | | **15** | |
| **Échant.** | **1** | **2** | **1** | **2** | **1** | **2** | **1** | **2** | **1** | **2** |
| **Analyses** | *25* | *25* | *29* | *31* | *19* | *29* | *23* | *25* | *39* | *26* |
| | *23* | *27* | *29* | *32* | *20* | *30* | *24* | *25* | *37* | *28* |

*Nous supposons les conditions du modèle bien remplies. Le tableau d'analyse de variance est alors le suivant :*

```
Analysis of Variance Table

Response: mesure
           Df  Sum Sq Mean Sq F value    Pr(>F)
lot        14 1210.93  86.495  1.4917   0.2256179
lot:echant 15  869.75  57.983  63.255  < 2.2e-16 ***
Residuals  30   27.50   0.917
```

*Analysons les résultats :*

1. *Pour le premier test, la probabilité critique vaut 0.2256179 et nous décidons de ne pas refuser l'hypothèse nulle $H_0$. Par conséquent, nous n'avons pas réussi à mettre en évidence d'effet du facteur à effets aléatoires* **lot**. *Le risque associé à cette décision est un risque de seconde espèce, et pour l'évaluer, il resterait à calculer la puissance de ce test.*
2. *Pour le second test, la probabilité critique vaut quasiment zéro, et nous décidons, au seuil $\alpha = 5\,\%$, de refuser l'hypothèse nulle $H_0$. Par conséquent, nous pouvons dire qu'il y a un effet significatif du facteur à effets aléatoires* **échantillon** *dans le facteur à effets aléatoires* **lot**. *Le risque associé à cette décision est un risque de première espèce qui vaut $5\,\%$.*

## 7.3 Modèles à effets mixtes

Pour la majorité des auteurs sur l'analyse de la variance à ce sujet, un facteur emboîté dans un facteur aléatoire doit être considéré comme aléatoire. Ainsi, le seul modèle mixte possible est le cas où le facteur $A$ est fixe et le facteur que nous emboîtons dans $A$, le facteur $B$, est aléatoire.

Un facteur contrôlé $A$ se présente sous $I$ modalités, chacune d'elle étant notée $A_i$.

Les termes $B_{j(i)}$ représentent un échantillon de taille $J$ prélevé dans une population importante dépendant du niveau $A_i$ du facteur $A$. Nous admettrons que les effets des $B_{j(i)}$ sont de loi normale centrée de variance $\sigma^2_{B|A}$.

Pour chacun des couples $(A_i, B_{j(i)})$, nous effectuons $K \geq 2$ mesures d'une réponse $Y$ qui est une variable continue. Nous noterons $n = I \times J \times K$ le nombre total de mesures ayant été effectuées.

On introduit le modèle :

$$Y_{i,j,k} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{i,j,k} \qquad , \qquad i = 1, \cdots, I \,;\, j = 1, \cdots, J \,;\, k = 1, \cdots, K \,.$$

avec les contraintes supplémentaires $\displaystyle\sum_{i=1}^{I} \alpha_i = 0 \,,$

où $Y_{i,j,k}$ est la valeur prise par la réponse $Y$ dans les conditions $(A_i, B_{j(i)})$ lors de la $k$-ème mesure.Nous supposons que :

$$\mathcal{L}(\beta_{j(i)}) = \mathcal{N}(0, \sigma^2_{B|A}), \quad \forall\,(i,j)\,, 1 \leq i \leq I,\, 1 \leq j \leq J\,,$$

ainsi que l'indépendance des effets aléatoires :

$\beta_{j(i)}$ et $\beta_{l(k)}$ sont indépendants si $(i,j) \neq (k,l)$ avec $1 \leq i,k \leq I$ et $1 \leq j,l \leq J$,

On supposera toujours réalisées les hypothèses standards suivantes :

1. $\varepsilon_{i,j,k}$ et $\varepsilon_{l,m,n}$ sont indépendantes si $(i,j,k) \neq (l,m,n)$ avec $1 \leq i,l \leq I$ , $1 \leq j,m \leq J$ et $1 \leq k,n \leq K$.
2. $\forall\,(i,j,k)$, $i = 1,\cdots,I\,;\, j = 1,\cdots,J\,;\, k = 1,\cdots,K\,;\; \mathcal{L}(\varepsilon_{i,j,k}) = \mathcal{N}(0,\sigma^2)$.

ainsi que l'indépendance des effets aléatoires et des erreurs :

$$\beta_{j(i)} \text{ est indépendant de } \varepsilon_{k,l,m} \text{ si } 1 \leq i,k \leq I\,;\, 1 \leq j,l \leq J\,;\, 1 \leq m \leq K\,.$$

Nous supposons que les conditions d'utilisation de ce modèle sont bien remplies.

Nous utilisons les quantités $SC_A$, $SC_{B|A}$, $SC_R$, $SC_{TOT}$, $sc_A$, $sc_{B|A}$, $sc_R$ et $sc_{TOT}$ introduites à la première section .

Nous rappelons la relation fondamentale de l'ANOVA :

$$SC_{TOT} = SC_A + SC_{B|A} + SC_R\,.$$

On introduit les degrés de liberté correspondant à chaque ligne du tableau de l'ANOVA :

| **Source** | **Nombre de d.d.l.** |
|---|---|
| **Facteur $A$** | $n_A = I - 1$ |
| **Facteur $B$ dans $A$** | $n_{B\|A} = I(J-1)$ |
| **Résiduelle** | $n_R = IJ(K-1)$ |
| **Totale** | $n_{TOT} = IJK - 1$ |

On peut résumer toutes ces informations dans le tableau de l'ANOVA ci-dessous :

| **Sources de variation** | **Degrés de liberté** | **Variations** | **Carrés moyens** | **Statistique** F | **Décision** |
|---|---|---|---|---|---|
| Facteur $A$ | $n_A = I - 1$ | $sc_A$ | $s_A^2 = \dfrac{sc_A}{n_A}$ | $f_A = \dfrac{s_A^2}{s_{B\|A}^2}$ | $H_0'$ ou $H_1'$ |
| Facteur $B$ dans facteur $A$ | $n_{B\|A} = I(J-1)$ | $sc_{B\|A}$ | $s_{B\|A}^2 = \dfrac{sc_{B\|A}}{n_{B\|A}}$ | $f_{B\|A} = \dfrac{s_{B\|A}^2}{s_R^2}$ | $H_0''$ ou $H_1''$ |
| Résiduelle | $n_R = IJ(K-1)$ | $sc_R$ | $s_R^2 = \dfrac{sc_R}{n_R}$ | | |
| Total | $n_{TOT} = IJK$-1 | $sc_{TOT}$ | $s_T^2 = \dfrac{sc_{TOT}}{n_{TOT}}$ | | |

Nous souhaitons faire les tests d'hypothèses suivants :

$$H_0' \, ; \, \alpha_1 = \alpha_2 = \cdots = \alpha_I = 0$$

contre

$$H_1' \, : \text{ Il existe } i_0 \in \{1, 2, \cdots, I\} \text{ tel que } \alpha_{i_0} \neq 0 \, .$$

Sous l'hypothèse nulle $H_0'$ d'absence d'effet du facteur $A$ et lorsque les conditions de validité du modèle sont respectées, $f_A$ est la réalisation de la variable aléatoire $S_A^2/S_{B|A}^2$ qui suit une loi de Fisher-Snedecor à $n_A = I - 1$ et $n_{B|A} = I(J-1)$ degrés de liberté.

On peut alors conclure grâce à la valeur critique, et on rejette l'hypothèse nulle si elle est inférieure ou égale au seuil $\alpha$ du test, ou à l'aide d'une table. Il y a rejet si $f_A$ est supérieure ou égale à la valeur critique issue de la table. Si l'hypothèse $H_0'$ est rejetée, on pourra procéder à des comparaisons multiples des différents effets des niveaux du facteur.

Le second test concernant le second facteur $B$ est le suivant :

$$H_0'' \, : \, \sigma_{B|A}^2 = 0$$

contre

$$H_1'' \, : \, \sigma_{B|A}^2 \neq 0 \, .$$

Sous l'hypothèse nulle $(H_0'')$ précédente, d'absence d'effet du facteur $B$ dans le facteur $A$, et lorsque les conditions de validité du modèle sont respectées, $f_{B|A}$ est la réalisation d'une variable aléatoire qui suit une loi de Fisher à $I(J-1)$ et $IJ(K-1)$ degrés de liberté.

Les estimateurs $\hat{\mu}$, $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\cdots$, $\hat{\alpha}_I$, $\hat{\sigma}_{B|A}^2$, $\hat{\sigma}^2$ des paramètres $\mu$, $\alpha_1$, $\alpha_2$, $\cdots$, $\alpha_I$, $\sigma_{B|A}^2$ et $\sigma^2$ du modèle sont donnés par les formules suivantes :

$$\begin{aligned}
\hat{\mu} &= Y_{\bullet,\bullet,\bullet} \, , \\
\hat{\alpha}_i &= Y_{i,\bullet,\bullet} - \hat{\mu}, \qquad 1 \leq i \leq I \, , \\
\hat{\sigma}_{B|A}^2 &= \frac{1}{K} \left( S_{B|A}^2 - S_R^2 \right) , \\
\hat{\sigma}^2 &= \frac{SC_R}{(I-1)(J-1)} = S_R^2 \, ,
\end{aligned}$$

où $S_{B|A}^2 = \dfrac{SC_{B|A}}{n_{B|A}}$ et $S_R^2 = \dfrac{SC_R}{n_R}$. Ces estimateurs sont non biaisés.

Les estimations obtenues pour une liste de données expérimentales $\boldsymbol{y}$, notées $\hat{\mu}(\boldsymbol{y})$, $\hat{\alpha}_1(\boldsymbol{y})$, $\cdots$, $\hat{\alpha}_I(\boldsymbol{y})$, $\hat{\sigma}_{B|A}^2(\boldsymbol{y})$, $\hat{\sigma}^2(\boldsymbol{y})$ des paramètres $\mu$, $\alpha_1$, $\cdots$, $\alpha_I$, $\sigma_{B|A}^2$ et $\sigma^2$ du modèle, se déduisent immédia-

tement des formules précédentes :

$$\begin{aligned}
\hat{\mu}(\boldsymbol{y}) &= y_{\bullet,\bullet,\bullet}\,, \\
\hat{\alpha}_i(\boldsymbol{y}) &= y_{i,\bullet,\bullet} - \hat{\mu}(\boldsymbol{y}), \qquad 1 \leq i \leq I\,, \\
\hat{\sigma}^2_{B|A}(\boldsymbol{y}) &= \frac{1}{K}\left(s^2_{B|A} - s^2_R\right)\,, \\
\hat{\sigma}^2(\boldsymbol{y}) &= \frac{sc_R}{(I-1)(J-1)} = s^2_R\,.
\end{aligned}$$

**Exemple 7.3.1**

*L'expérience porte sur la prise de poids quotidienne de jeunes cochons au cours de leur phase de croissance. L'objectif de l'expérience est de déterminer l'influence du patrimoine génétique de cinq pères sur leurs descendants. Pour cela, ces cinq mâles ont eu une portée avec deux mères différentes et choisies au hasard. Dans chacune de ces portées, deux animaux ont été sélectionnés et leur masse mesurée en grammes.*

*On peut remarquer que le facteur* **père** *est considéré comme un facteur à effets fixes et le facteur* **mère** *comme un facteur à effets aléatoires.*

*Les données sont consignées ci-dessous :*

| **Père** | **1** | | **2** | | **3** | |
|---|---|---|---|---|---|---|
| **Mère** | **1** | **2** | **1** | **2** | **1** | **2** |
| **Gain de masse** | *2,77* | *2,58* | *2,28* | *3,01* | *2,36* | *2,72* |
| | *2,38* | *2,94* | *2,22* | *2,61* | *2,71* | *2,74* |
| **Père** | **4** | | **5** | | | |
| **Mère** | **1** | **2** | **1** | **2** | | |
| **Gain de masse** | *2,87* | *2,31* | *2,74* | *2,50* | | |
| | *2,46* | *2,24* | *2,56* | *2,48* | | |

*Nous supposons les conditions du modèle bien remplies. Le tableau d'analyse de variance est alors le suivant :*

```
Analysis of Variance Table

Response: poids
          Df  Sum Sq  Mean Sq F value  Pr(>F)
pere       4 0.09973 0.024932  0.2212 0.91553
pere:mere  5 0.56355 0.112710  2.9124 0.07067 .
Residuals 10 0.38700 0.038700
```

*Analysons les résultats :*

1. *Pour le premier test, la probabilité critique vaut 0,91553 et nous décidons donc de ne pas refuser l'hypothèse nulle $H_0$. Par conséquent, nous n'avons pas réussi à mettre en évidence d'effet du facteur à effets fixes* **père**. *Le risque associé à cette décision est un risque de seconde espèce. Pour l'évaluer, il resterait à calculer la puissance de ce test.*
2. *Pour le second test, la probabilité critique est 0,07067 et nous décidons de ne pas refuser l'hypothèse nulle $H_0$. Par conséquent, nous n'avons pas réussi à mettre en évidence d'effet du facteur à effets aléatoires* **mère** *dans le facteur à effets fixes* **père**. *Le risque associé à cette décision est un risque de seconde espèce. Pour l'évaluer, il resterait à calculer la puissance de ce test.*