*Chapitre 0*

*Eléments de Méthodologie*

Barumandzadeh T & El Methni M.

## 0-1. Un peu de vocabulaire

Observation, interprétation, expérience, expérimentation, mesure, échelle de mesure, moyenne, variance, hypothèse, test.

## 0-2. Constituants d'une expérience

### 0-2-1 Sujets :

**Définition :** Les unités statistiques sur lesquelles porte l'expérience sont appelées unités expérimentales. L'ensemble des sujets constitue la population (souvent c'est un échantillon). On note $S$ cet ensemble et $s$ un sujet. En psychologie il s'agit souvent d'individus, on les désigne par le terme « sujet ».

### 0-2-2 Variable dépendante :

**Définition :** On appelle variable dépendante (VD) la réponse du système aux différentes conditions expérimentales. Cette réponse est mesurée sur les sujets de l'expérience. Elle peut être uni ou multidimensionnelle. On note $Y$ cette variable et $Y(s)$ la valeur qu'elle prend pour le sujet $s$.

### 0-2-3 Les facteurs de variation :

#### 0-2-3-1 Les variables indépendantes :

**Définition :** Les variables indépendantes (VI) sont les variables représentant les causes postulées des variations de la VD. Le but de l'expérimentation étant de mettre en évidence l'effet de ces variables indépendantes.

**Remarque :** On emploie aussi le terme de facteur de variation ou facteur expérimental ou encore (plus simplement) facteur.

Par usage le terme de « variables indépendantes » est utilisé dans un contexte méthodologique alors que le terme de facteur est utilisé dans un contexte statistique.

**Définition :** On appelle variables indépendantes provoquées les variables indépendantes pouvant être manipulées (contrôlées) par l'expérimentateur.

**Remarque (fondamentale):** Dans le cas des variables indépendantes provoquées l'expérimentateur a la possibilité de répartir aléatoirement les sujets dans différents groupes expérimentaux et ceci avant d'agir sur les VI.

**Définition :** Les variables indépendantes invoquées (ou étiquettes) sont des caractéristiques « naturelles » sur lesquelles l'expérimentateur ne peut agir, mais elles lui servent pour repérer les conditions expérimentales.

**Remarque :** Certains auteurs réservent le mot d'expérimental aux situations d'observations lorsque toutes les variables indépendantes sont provoquées et de quasi-expérimental lorsque le plan d'expérience contient des variables indépendantes invoquées.

**Définition :** On dit que deux variables indépendantes sont confondues si chaque modalité de l'une est associée à une modalité de l'autre.

#### 0-2-3-2 Les variables parasites :

Les variables indépendantes non prises en considération dans l'expérience sont appelées variables parasites. Les variables parasites identifiables sont celles que l'on peut totalement identifier. On cherche alors à les contrôler pour « effacer » leur effet. Les variables parasites non identifiables sont toutes les caractéristiques propres à l'unité statistique. Ces dernières peuvent être contrôlées par randomisation.

## 0-3. Les effets des facteurs

### 0-3-1 Les niveaux des effets des facteurs :

Dans de nombreuses études, les phénomènes sont plus complexes que la simple relation de cause à effet entre les VI et les VD. On peut imaginer tout un réseau de causalités. L'effet d'un facteur peut se traduire alors à plusieurs niveaux :

**a)** effet principal : effet direct sur la VD que l'on aurait mesuré dans une expérience où seul ce facteur serait présent.

**b)** effet d'interaction : effet du facteur sur les effets des autres facteurs. On distingue alors plusieurs niveaux d'interaction :

- Interaction d'ordre 1 : interaction d'un facteur sur l'effet d'un autre facteur.
- Interaction d'ordre 2 : interaction d'un facteur sur l'interaction entre deux autres facteurs
- etc. …

### 0-3-2 Nature des effets des facteurs :

On distingue deux cas :

**a)** Les facteurs à effets fixes: les modalités du facteur qui intéressent le chercheur sont choisies de façon systématique par l'expérimentateur. On veut mesurer l'effet de chaque modalité prise par elle-même.

**b)** Les facteurs à effets aléatoires: toutes les modalités du facteur ne sont pas présentes dans l'expérience. Les modalités retenues ne sont pas choisies de façon systématique mais de façon aléatoire. Le chercheur ne s'intéresse donc pas à chaque modalité prise individuellement mais à l'effet global du facteur. Il lui sera alors possible d'étendre des conclusions à d'autres modalités de la variable indépendante non présentes lors de l'expérimentation.

Dans de nombreux modèles on introduit le facteur de variation « sujet » dont les modalités sont les unités statistiques de l'expérience. Le facteur sujet est par construction même de l'expérience un facteur à effets aléatoires.

## 0-4. Plans d'expérience

### 0-4-1 Généralités :

L'expérimentateur définit un ensemble de $p$ facteurs élémentaires $A_1, A_2, \dots, A_p$ Chaque facteur possède un nombre $n_i$ fini de modalités $A_i = \{a_1^i, a_2^i, \dots, a_{ni}^i\}$.

Chaque observation de la variable dépendante correspond à une combinaison des modalités des facteurs élémentaires $a = \{a_{i1}^1, a_{i2}^2, \dots, a_{ip}^p\}$. $a_{ij}^j \in A_j$

Notons $E$ l'ensemble de toutes les combinaisons des modalités des facteurs élémentaires $E = A_1 \times A_2 \times \dots \times A_p$.

Chaque condition expérimentale envisagée dans l'expérience est décrite par un élément de $E$

Lors de l'expérience il arrive souvent que toutes les combinaisons ne sont pas retenues.

Cependant chaque modalité de chaque facteur est présente au moins une fois dans l'expérience.

L'ensemble des combinaisons retenues définit le plan d'expérience.

**Définition :** On appelle plan d'expérience un sous-ensemble $E^*$ de $E$ tel que chaque modalité de chaque facteur soit présente au moins une fois dans les éléments de $E^*$. Les éléments de $E^*$ sont appelés conditions expérimentales.

### 0-4-2 Relations entre facteurs :

Un plan d'expérience se conçoit en précisant les facteurs élémentaires pris en compte mais aussi les relations entre ces facteurs. Nous étudierons deux relations fondamentales : la relation de croisement et la relation d'emboîtement. Ces deux relations permettent de décrire un ensemble particulier de plans d'expérience (ensemble des plans quasi-complets).

**Définition :** Deux facteurs A et B sont croisés si chaque modalité de l'un apparaît avec chaque modalité de l'autre. On note $A \times B$ (ou $B \times A$).

**Remarque :** La définition précédente se généralise à plusieurs facteurs. *p* facteurs sont croisés dans leur ensemble si chaque modalité de l'un apparaisse avec toutes les modalités de tous les autres facteurs.

**Remarque (importante) :** Des facteurs croisés deux à deux ne sont pas nécessairement croisés dans leur ensemble.

**Définition :** Un facteur A est emboîté dans un facteur B si chaque modalité de A est associée à exactement une modalité de B. A est appelé le facteur emboîté et B le facteur emboîtant. On note A<B>.

**Remarque :** un facteur peut être emboîté dans un facteur lui-même emboîté dans un troisième facteur. Dans ce cas le 1er facteur est emboîté dans le 3ème.

**Définition :** L'emboîtement est dit équilibré (balancé) si et seulement si chaque modalité du facteur emboîtant est associée au même nombre de modalités du facteur emboîté.

**Définition :** Deux facteurs A et B sont confondus si et seulement si A est emboîté dans B et B est emboîté dans A.

**0-4-3 Les plans d'expérience de base:**

**Définition :** Un plan d'expérience est dit complet si et seulement si il est égal au croisement de tous ses facteurs élémentaires ($E^* = E$).Autrement dit il contient toutes les combinaisons des modalités des facteurs élémentaires.

**Définition :** Un plan d'expérience est dit quasi-complet si et seulement si tout couple de facteurs les deux facteurs sont soit croisés soit emboîtés et si tous les facteurs croisés deux à deux sont croisés dans leur ensemble.

Dans toute la suite nous ne nous intéresserons qu'à des plans quasi-complets.

**Définition :** Un plan d'expérience est dit à un facteur et à mesures indépendantes s'il met en jeu un seul facteur A outre les sujets qui ne sont « utilisés » qu'une seule fois (un sujet ne se retrouve que dans un seul groupe expérimental). On a donc S<A>. On parle de plan inter.

**Définition :** Un plan d'expérience est dit à un facteur et à mesures répétées s'il met en jeu un seul facteur A outre les sujets qui sont « utilisés » dans les différentes conditions expérimentales. On a donc S×A. On parle de plan intra.

**Définition :** Un plan d'expérience est dit à deux facteurs et à mesures indépendantes si on a S<A×B>

**Définition :** Un plan d'expérience est dit à deux facteurs et à mesures complètement répétées si les sujets sont utilisés dans l'ensemble des conditions expérimentales. On a donc S×A×B.

**Définition :** Un plan quasi-complet de la forme $S<A>\times B$ est dit mixte.

## I-1. Modèle statistique

Les données sont constituées d'une famille de $N$ valeurs $y_1$, $y_2$, … $y_N$ résultant de l'observation de la variable dépendante dans les différentes conditions expérimentales définies par le plan de l'expérience. On appelle protocole expérimental cette famille et on note $y(e)$ l'observation de la VD dans la condition expérimentale $e$. L'élaboration du modèle qui décrit le processus de génération des données doit prendre en compte d'une part le fait que les mesures sont faites sur un échantillon aléatoire d'unités statistiques (fluctuations d'échantillonnage) et d'autre part qu'interviennent de façon aléatoire au cours de l'expérimentation des facteurs non contrôlés tels que par exemple les erreurs de mesure ou des facteurs individuels. L'observation $y(e)$ est alors considérée comme la réalisation $Y(e)$ d'une variable aléatoire $Y$. Le modèle du score décrit les effets des facteurs sur la variable aléatoire par le modèle algébrique : $Y(e) = Z(e)+\varepsilon(e)$
Où :

- $Z(e)$, appelé modèle structurel, décrit les effets des facteurs du plan
- $\varepsilon(e)$, appelé résidu, mesure l'écart entre la variable réponse et le modèle structurel dû aux effets de facteurs non pris en compte dans le plan de l'expérience.

Ce modèle linéaire est caractérisé par des hypothèses portant sur le résidu

- $\varepsilon(e)$ est une variable aléatoire de moyenne nulle.
- Var $(\varepsilon(e))=\sigma^2$ pour tout $e$ homoscédasticité ou homogénéité des variances des résidus
- Cor$[(\varepsilon(e)$ , $\varepsilon(e'))]=0$ pour tout $e\neq e'$ non corrélation des résidus

Ce modèle linéaire est aussi caractérisé par des hypothèses portant sur le modèle structurel $Z(e) = \mu(e)+A(e)$

- $\mu(e)$ est une fonction déterministe de $e$ qui décrit les effets des modalités des facteurs à effets fixes présentes dans $e$.
- $A(e)$ est une variable aléatoire non corrélée avec $\varepsilon(e)$, qui décrit les effets des facteurs à effets aléatoires
- $\varepsilon(e)$ suit une loi normale $\mathcal{N}(0, \sigma^2)$

Selon la forme du modèle structurel, on distingue trois classes de modèles d'analyse de la variance :

ANOVA modèle 1 : Le modèle ne contient pas de terme aléatoire $A(e)$ : $Y(e) = \mu(e)+\varepsilon(e)$
Ce modèle correspond à des plans où tous les facteurs sont à effets fixes. Leurs effets portent uniquement sur la moyenne de la VD. $Y(e)$ est distribué selon une $\mathcal{N}(\mu(e)\ , \sigma^2)$.

ANOVA modèle 2 : (modèle des composantes de la variance) Le terme déterministe est constant

$$Y(e) = \mu+A(e) +\varepsilon(e)$$

Ce modèle correspond à des plans où tous les facteurs sont à effets aléatoires. Leurs effets portent sur la variance de la VD. $Y(e)$ est distribué selon une $\mathcal{N}(\mu\ , \sigma^2+\sigma^2_{A(e)})$.

ANOVA modèle 3 : (modèle mixte) c'est le modèle qui mélange les deux cas précédents
Ce modèle correspond à des plans où sont présents des facteurs à effets fixes et des facteurs à effets aléatoires. $Y(e)$ est distribué selon une $\mathcal{N}(\mu(e), \sigma^2+\sigma^2_{A(e)})$.

**I-3. Distribution de Fisher-Snedecor**

**Proposition :** Soient $U_1, U_2, \ldots, U_m$ , $V_1, V_2, \ldots, V_n$ $m+n$ variables aléatoires indépendantes de même loi $\mathcal{N}(0, \sigma^2)$ alors les variables $\frac{1}{\sigma^2}\sum_{i=1}^{m} U_i^2$ et $\frac{1}{\sigma^2}\sum_{i=1}^{n} V_i^2$ sont indépendantes et de loi respective $\chi_m^2$ et $\chi_n^2$

**Théorème :** Soient $U_1, U_2, \ldots, U_m$ , $V_1, V_2, \ldots, V_n$ $m+n$ variables aléatoires indépendantes de même loi $\mathcal{N}(0, \sigma^2)$ alors la variable $\dfrac{\frac{1}{m\sigma^2}\sum_{i=1}^{m} U_i^2}{\frac{1}{n\sigma^2}\sum_{i=1}^{n} V_i^2} = \dfrac{\frac{1}{m}\sum_{i=1}^{m} U_i^2}{\frac{1}{n}\sum_{i=1}^{n} V_i^2}$ suit une loi de <u>Fisher-Snedecor</u> $\mathcal{F}(m, n)$ de $(m, n)$ degrés de liberté.

**Remarque :** Si $\mathcal{F}$ est une variable aléatoire de loi $\mathcal{F}(m, n)$ alors la variable aléatoire $1/\mathcal{F}$ suit une loi $\mathcal{F}(n, m)$.

**Corollaire :** Soient $U_1, U_2, \ldots, U_m$ $m$ variables aléatoires indépendantes de même loi $\mathcal{N}(\mu_1, \sigma^2)$ et $V_1, V_2, \ldots, V_n$ $n$ variables aléatoires indépendantes de même loi $\mathcal{N}(\mu_2, \sigma^2)$ indépendantes des $U_i$ alors la variable $\dfrac{\frac{1}{m-1}\sum_{i=1}^{m}(U_i-\overline{U})^2}{\frac{1}{n-1}\sum_{i=1}^{n}(V_i-\overline{V})^2}$ suit une loi de <u>Fisher-Snedecor</u> $\mathcal{F}(m\text{-}1, n\text{-}1)$.

$\overline{U} = \dfrac{\sum_{i=1}^{m} U_i}{m}$ et $\overline{V} = \dfrac{\sum_{i=1}^{n} V_i}{n}$ Sont les estimateurs respectifs de $\mu_1$ et $\mu_2$

## II-1. *G* est un (seul) facteur à effet fixe

### II-1-1 Généralités

*G* est un facteur à *r* modalités (groupes) $g_1, g_2, \ldots g_r$ d'effectifs respectifs $n_1, n_2, \ldots n_r$ . $N=n_1+n_2+ \ldots +n_r$. *Y* est la variable dépendante et $y_{s(i)}$ est le score du sujet *s* dans le groupe $g_i$ (c'est l'observation associée au sujet *s* dans la modalité $g_i$).

| | Facteur *G* (groupe) | | | | | | |
|---|---|---|---|---|---|---|---|
| Sujets | $g_1$ | $g_2$ | … | $g_i$ | … | $g_r$ | |
| *1* | $y_{1(1)}$ | $y_{1(2)}$ | … | $y_{1(i)}$ | … | $y_{1(r)}$ | |
| *2* | $y_{2(1)}$ | $y_{2(2)}$ | … | $y_{2(i)}$ | … | $y_{2(r)}$ | |
| … | … | … | … | … | … | … | |
| *s* | $y_{s(1)}$ | $y_{s(2)}$ | … | $y_{s(i)}$ | … | $y_{s(r)}$ | |
| … | … | … | … | … | … | … | |
| … | | $y_{n_2(2)}$ | … | | … | $y_{n_r(r)}$ | |
| … | $y_{n_1(1)}$ | … | … | $y_{n_i(i)}$ | … | … | |
| Effectif | $n_1$ | $n_2$ | … | $n_i$ | … | $n_r$ | $N=\sum_{i=1}^{r} n_i$ |
| Moyenne | $\overline{y}_1$ | $\overline{y}_2$ | … | $\overline{y}_i$ | … | $\overline{y}_r$ | $\overline{y}=\frac{1}{N}\sum_{i=1}^{r} n_i\overline{y}_i$ |
| Variance | $s'_1$ | $s'_2$ | … | $s'_i$ | … | $s'_r$ | $s'$ |

On veut étudier l'effet du facteur *G* sur la variable réponse *Y*.
Pour cela on testera l'hypothèse nulle : **H₀** : le facteur n'a pas d'effet sur *Y*
contre l'hypothèse alternative : **H₁** : le facteur a de l'effet sur *Y*.

### II-1-2 Décomposition de la variance : SCT = SCG + SCR

$$\textbf{SCT} = \sum_{s}\sum_{i}(y_{s(i)}-\overline{y})^2 = N \times \text{Variance Totale}$$

$$\textbf{SCG} = \sum_{i=1}^{r} n_i(\overline{y}_i-\overline{y})^2 = N \times \text{Variance}(\overline{y}_i) \qquad \textbf{SCR} = \sum_{s=1}^{n_i}(y_{s(i)}-\overline{y}_i)^2$$

### II-1-3 Modèle statistique (Modèle du score, modèle linéaire)

On veut étudier l'effet d'un facteur *G* à *r* modalités sur une variable réponse *Y*. On dispose pour chaque modalité *i* du facteur *G* de $n_i$ observations. On note $y_{s(i)}$ l'observation du sujet *s* dans la modalité *i*. $y_{s(i)}$ est la réalisation de la variable aléatoire $Y_{s(i)}$ décrite par le modèle suivant : $Y_{s(i)} = \mu+\alpha_i+\varepsilon_{s(i)}$ où :

- $\mu$ s'interprète comme un niveau général.
- $\alpha_i$ mesure l'effet de la $i^{\text{ème}}$ modalité du facteur *G*.
- les variables aléatoires $\varepsilon_{s(i)}$ sont indépendantes et de même loi $\mathcal{N}(0, \sigma^2)$
- Pour tout $i$=1, 2,…, *r*. Les variances des *r* groupes sont homogènes (homoscédasticité)
- On rajoute une contrainte d'identifiabilité: $\sum_{i=1}^{r} n_i\alpha_i = 0$ (en moyenne les effets sont nuls)

### II-1-4 Moyennes des carrés et degrés de liberté

$$\textbf{MCT} = \frac{\textbf{SCT}}{N-1} \quad \textbf{MCG} = \frac{\textbf{SCG}}{r-1} \quad \textbf{MCR} = \frac{\textbf{SCR}}{N-r}$$

ddl total = ddl inter + ddl intra $N - 1 = r - 1 + N - r$

**Statistique de test :**

Sous l'hypothèse nulle la statistique $\dfrac{\textbf{SCG}/(r-1)}{\textbf{SCR}/(N-r)} = \dfrac{\textbf{MCG}}{\textbf{MCR}} = \textbf{F}_\textbf{G}$ suit une loi de Fischer à $r$ - 1 et $N$ - $r$ ddl, notée $\mathcal{F}(r - 1, N - r)$.

### II-1-5 Tableau d'analyse de la variance

| Source de variation | Somme des carrés des écarts observés $SC_{obs}$ | ddl | Carrés moyens observés $MC_{obs}$ | F |
|---|---|---|---|---|
| G, Facteur, Inter-groupes, ... | **SCG** | $r$ - 1 | **MCG** | $\textbf{F}_{\textbf{obs}} = \dfrac{\textbf{MCG}}{\textbf{MCR}}$ |
| S<G>, Résiduelle, Intra | **SCR** | $N$ - $r$ | **MCR** | $F_{lu}$ =<br>$\alpha$ = |
| Totale | **SCT** | $N$ - 1 | **MCT** | |

## II-2. *G* est un (seul) facteur à effet aléatoire

### II-2-1 Modèle statistique

On choisit un échantillon de $r$ modalités $g_1$, $g_2$, … $g_r$ du facteur $G$. On dispose pour chaque modalité échantillonnée de $n_i$ observations de la variable dépendante ($N$=$n_1$+$n_2$+ … +$n_r$). On note $y_{s(i)}$ l'observation du sujet $s$ dans le groupe $g_i$. Chaque observation $y_{s(i)}$ est la réalisation d'une variable aléatoire $Y_{s(i)}$ décrite par le modèle : $Y_{s(i)} = \mu + \Gamma_i + \varepsilon_{s(i)}$

Où :

$\mu$ est une constante mesurant le niveau général de la réponse

$\Gamma_i$ est une variable aléatoire qui mesure l'effet aléatoire de $G$.

$\varepsilon_{s(i)}$ est une variable aléatoire représentant le résidu.

On suppose réalisées les trois hypothèses suivantes :

- les $\varepsilon_{s(i)}$ sont des variables aléatoires indépendantes et de même loi $\mathcal{N}(0, \sigma^2)$
- les $\Gamma_i$ ($i$=1, 2, …, $r$) sont des variables aléatoires indépendantes de même loi $\mathcal{N}(0, \sigma_G^2)$
- les $\Gamma_i$ sont indépendantes des $\varepsilon_{s(i)}$.

Le test d'hypothèse s'écrira alors :

$\mathbf{H_0}$ : G n'a pas d'effet $\Leftrightarrow \sigma_G^2 = 0$

$\mathbf{H_1}$ : G a un effet $\Leftrightarrow \sigma_G^2 > 0$

On montre alors les résultats suivants :

### II-2-2 Test statistique

On se ramène donc au cas précédent et on utilisera la même statistique $\textbf{F} = \dfrac{\textbf{MCG}}{\textbf{MCR}}$ pour réaliser le test.

## III-1. Plan $S<A\times B>$ avec *A* et *B* deux facteurs à effet fixe

### III-1-1 Généralités

Un tel plan quasi-complet défini par le croisement de deux facteurs *A* et *B*, est aussi appelé plan factoriel complet d'ordre 2. Chaque condition est répliquée sur des sujets différents. Le facteur sujet est emboîté dans le croisement $A\times B$. L'analyse diffère selon que l'emboîtement est équilibré ou non. Dans toute la suite on ne considère que le cas équilibré.

On se propose d'étudier l'effet de la présence simultanée d'un facteur A à *r* modalités et d'un facteur B à *c* modalités sur une variable réponse *Y*. Le plan étant équilibré, on dispose pour chaque croisement des modalités *i* du facteur A et *j* du facteur B de *n* observations. On note $y_{s(i,j)}$ l'observation du sujet *s* dans le croisement (*i*,*j*).

| | | Facteur *B* | | | | | |
|---|---|---|---|---|---|---|---|
| | | $b_1$ | $b_2$ | … | $b_j$ | … | $B_c$ |
| *A* | $a_1$ | | | | | | |
| | $a_2$ | | | | | | |
| | … | | | | | | |
| | $a_i$ | | | | $y_{1(i,j)}, y_{2(i,j)}, \ldots\ldots y_{n(i,j)}$ | | |
| | … | | | | | | |
| | $a_r$ | | | | | | |

**Remarque :** dans le cas où *n*=1, l'interaction des facteurs *A* et *B* est confondue avec le facteur sujet, on ne peut donc séparer les effets d'interaction des effets résiduels et on ne peut tester leur existence. Un tel plan n'est justifié que lorsque l'on peut modifier a priori la forme de l'interaction entre les facteurs. Dans toute la suite nous développons le cas $n > 1$.

### III-1-2 Analyse descriptive des effets (principaux et interaction) :

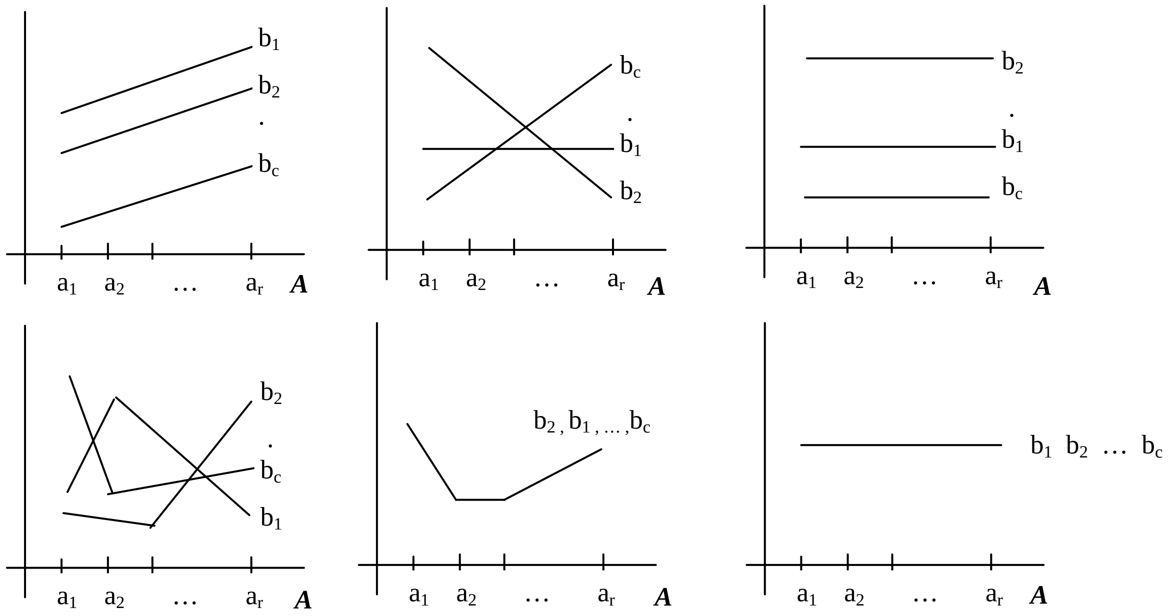
Considérons, pour chaque croisement (*i*,*j*) de la modalité *i* de *A* et de la modalité *j* de *B* la moyenne obtenue :

| | | Facteur *B* | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | $b_1$ | $b_2$ | … | $b_j$ | … | $B_c$ | moyennes |
| *A* | $a_1$ | $\bar{y}_{11}$ | $\bar{y}_{12}$ | … | $\bar{y}_{1j}$ | … | $\bar{y}_{1c}$ | $\bar{y}_{10}$ |
| | $a_2$ | $\bar{y}_{21}$ | $\bar{y}_{22}$ | … | $\bar{y}_{2j}$ | … | $\bar{y}_{2c}$ | $\bar{y}_{20}$ |
| | … | … | … | … | … | … | … | … |
| | $a_i$ | $\bar{y}_{i1}$ | $\bar{y}_{i2}$ | … | $\bar{y}_{ij}$ | … | $\bar{y}_{ic}$ | $\bar{y}_{i0}$ |
| | … | … | … | … | … | … | … | … |
| | $a_r$ | $\bar{y}_{r1}$ | $\bar{y}_{r2}$ | … | $\bar{y}_{rj}$ | … | $\bar{y}_{rc}$ | $\bar{y}_{r0}$ |
| | moyennes | $\bar{y}_{01}$ | $\bar{y}_{02}$ | … | $\bar{y}_{0j}$ | … | $\bar{y}_{0c}$ | $\bar{y}_{00} = \bar{y}$ |

La colonne marginale des moyennes par modalité de *A* (indépendamment de *B*) « traduit » l'effet principal du facteur *A*. De même la ligne marginale des moyennes par modalité de *B* (indépendamment de *A*) « traduit » l'effet principal du facteur *B*. Les différentes moyennes correspondant au croisement des modalités « traduisent » l'interaction des deux facteurs *A* et *B*.

**III-1-3 Représentation géométrique des différents effets :**

Les différentes combinaisons des effets principaux et d'interaction peuvent être illustrés par des graphiques dans lesquels on représente les moyennes en fonction des modalités de l'un des facteurs. Il y a deux graphiques possibles selon le facteur porté en abscisse.




Le <u>parallélisme</u> des lignes traduit le fait que l'effet du facteur *A* (ou *B*) sur la variable réponse *Y* est le même quelque soit la valeur du facteur *B* (ou *A*). Les effets des facteurs *A* et *B* sont <u>additifs</u> et il n'y a pas d'effet de l'interaction entre les deux facteurs *A* et *B* sur *Y*.

Le <u>non parallélisme</u> des lignes traduit le fait que l'effet du facteur *A* (ou du facteur *B*) n'est pas le même en fonction de la modalité du facteur *B* (ou du facteur *A*). Les effets des facteurs *A* et *B* sur la variable *Y* <u>ne sont pas additifs</u>. Aux effets simples des facteurs s'ajoutent l'effet de l'interaction entre les deux facteurs.

Le <u>parallélisme avec l'axe des abscisses</u> traduit l'absence d'effet du facteur *A* (ou *B*). De plus si les <u>différentes lignes</u> correspondant aux modalités de *B* (ou *A*) sont <u>confondues</u> c'est l'absence totale d'effets.

**III-1-4 Le modèle :**

Rappelons que l'on veut étudier l'effet de la présence simultanée d'un facteur *A* à *r* modalités et d'un facteur *B* à *c* modalités sur une variable réponse *Y*. Le plan étant équilibré, on dispose pour chaque modalité *i* du facteur *A* et pour chaque modalité *j* du facteur *B* de *n* observations ($n > 1$). On note $y_{s(i,j)}$ l'observation du sujet *s* dans le croisement $(i,j)$. $y_{s(i,j)}$ est la réalisation de la variable aléatoire $Y_{s(i,j)}$ décrite par le modèle suivant :

$$Y_{s(i,j)} = \mu_{ij} + \varepsilon_{s(i,j)}$$

Où les variables aléatoires $\varepsilon_{s(i,j)}$ sont indépendantes et de même loi $\mathcal{N}(0,\sigma^2)$ et

$$\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} \qquad \text{où}$$

- $\mu$ mesure la moyenne de la population
- les $\alpha_i$ mesurent les effets principaux du facteur *A*
- les $\beta_j$ mesurent les effets principaux du facteur *B*
- les $(\alpha\beta)_{ij}$ mesurent les effets d'interaction.

On rajoute (comme hypothèses) les contraintes d'identifiabilité suivantes :

$$\sum_i \alpha_i = 0 \quad \sum_j \beta_j = 0 \quad \sum_j (\alpha\beta)_{ij} = 0 \quad \sum_i (\alpha\beta)_{ij} = 0$$

### III-1-5 Décomposition de la variation :

**SCT**$_{obs}$ = **SCI**$_{ntercase}$ + **SCI**$_{intracase}$ = **SC**(A×B)$_{obs}$ + **SCR**$_{obs}$

**Définition :** On appelle somme des carrés associée à l'interaction la quantité :

$$\mathbf{SCAB}_{obs} = \mathbf{SC(A\times B)}_{obs} - \mathbf{SCA}_{obs} - \mathbf{SCB}_{obs}$$

**SCT**$_{obs}$ = **SCA** + **SCB** + **SC**(AB) + **SCR**

$\mathbf{SCA}_{obs} = nc\sum_{i}(\bar{y}_{i0} - \bar{y})^2 = ncr\times$variance des moyennes de A

$\mathbf{SCB}_{obs} = nr\sum_{j=1}^{c}(\bar{y}_{0j} - \bar{y})^2 = ncr\times$variance des moyennes de B

$\mathbf{SC(A\times B)}_{obs} = n\sum_{i=1}^{r}\sum_{j=1}^{c}(\bar{y}_{ij} - \bar{y})^2 = ncr\times$variance des moyennes des carrés croisés

**SC(AB)**$_{obs}$ = **SC(A×B)**$_{obs}$ - **SCA**$_{obs}$ - **SCB**$_{obs}$

**SCT**$_{obs}$ = $ncr\times$variance de toutes les données

**SCR**$_{obs}$ = **SCT**$_{obs}$ - **SC(A×B)**$_{obs}$= **SCT**$_{obs}$ - **SC(AB)**$_{obs}$- **SCA**$_{obs}$ - **SCB**$_{obs}$

On ramène toutes les sommes de carrés à des moyennes de carrés en divisant par les degrés de liberté correspondant. $\mathbf{MCA} = \frac{\mathbf{SCA}}{r-1}$ $\mathbf{MCB} = \frac{\mathbf{SCB}}{c-1}$ $\mathbf{MCAB} = \frac{\mathbf{SCAB}}{(r-1)(c-1)}$ $\mathbf{MCR} = \frac{\mathbf{SCR}}{rc(n-1)}$

### III-1-6 Tests et décisions statistiques :

**Statistique dutest 1 :** sous l'hypothèse $\mathbf{H_0}_A$, la statistique $\mathbf{F_A} = \frac{\mathbf{MCA}}{\mathbf{MCR}}$ suit une loi de Fischer à $r$ - 1 et $rc(n$ - 1) degrés de liberté

**Statistique dutest 2 :** sous l'hypothèse $\mathbf{H_0}_B$, la statistique $\mathbf{F_B} = \frac{\mathbf{MCB}}{\mathbf{MCR}}$ suit une loi de Fischer à $c$ - 1 et $rc(n$ - 1) degrés de liberté

**Statistique du test 3 :** sous l'hypothèse $\mathbf{H_0}_{AB}$, la statistique $\mathbf{F_{AB}} = \frac{\mathbf{MCAB}}{\mathbf{MCR}}$ suit une loi de Fischer à ($r$ - 1)($c$ - 1) et $rc(n$ - 1) degrés de liberté

**Tableau d'analyse de la variance :**

| Source | **SC**$_{obs}$ | ddl | **MC**$_{obs}$ | **F**$_{obs}$ |
|---|---|---|---|---|
| *A* | **SCA**$_{obs}$ | $r$ - 1 | **MCA**$_{obs}$ | **F**$_{A\ obs}$ |
| *B* | **SCB**$_{obs}$ | $c$ - 1 | **MCB**$_{obs}$ | **F**$_{B\ obs}$ |
| *AB* | **SCAB**$_{obs}$ | ($r$ - 1)($c$ - 1) | **MCAB**$_{obs}$ | **F**$_{AB\ obs}$ |
| *R* | **SCR**$_{obs}$ | $rc(n$ - 1) | **MCR**$_{obs}$ | |
| Total | **SCT**$_{obs}$ | $rcn$ - 1 | | |

## *Chapitre IV*

**Analyse du plan $S\times 0$**

Barumandzadeh T & El Methni M.

### IV-1. *O* est un (seul) facteur à effet fixe

### IV-1-1 Généralités

$O$ est un facteur à $r$ modalités $o_1$, $o_2$, … $o_r$ (Occasions). S est le facteur sujet (à effets aléatoires) dont les modalités sont les $n$ sujets $s_1$, $s_2$, … $s_n$. Pour chaque modalité du facteur $O$ on « utilise » les $n$ sujets. On a donc des mesures $y_{si}$ répétées pour chaque sujet $s$ dans l'occasion (modalité) $O_i$.
En psychologie et en méthodologie on parle de plan intra, en statistique on parle de plan à mesures répétées.
On dispose donc de $N=n\times r$ mesures $y_{si}$ que l'on présente sous la forme d'un tableau :

| | Facteur $O$ (Occasions) | | | | | |
|---|---|---|---|---|---|---|
| Sujets | $o_1$ | $o_2$ | … | $o_i$ | … | $o_r$ |
| $s_1$ | $y_{11}$ | $y_{12}$ | … | $y_{1i}$ | … | $y_{1r}$ |
| $s_2$ | $y_{21}$ | $y_{22}$ | … | $y_{2i}$ | … | $y_{2r}$ |
| … | … | … | … | | … | … |
| $s_s$ | $y_{s1}$ | $y_{s2}$ | … | $y_{si}$ | … | $y_{sr}$ |
| … | … | … | … | … | … | … |
| $s_n$ | $y_{n1}$ | $y_{n2}$ | | $y_{ni}$ | | $y_{nr}$ |

**Remarque :** Un premier avantage d'un tel plan par rapport au plan $S<O>$ est une « économie » du nombre de sujets. D'autre part avec un tel plan on peut diminuer l'erreur expérimentale (résidu). En effet le contrôle du facteur sujet permet de séparer son effet des effets résiduels. Mais le résidu sera confondu avec l'interaction entre $S$ et $O$.

### IV-1-2 Modèle univarié :

Les données peuvent être d'une part, considérées comme les $N=n\times r$ observations de la variable réponse $Y$ dans les $N=n\times r$ conditions expérimentales décrites par le croisement du facteur sujet et du facteur occasion. Elles peuvent d'autre part, être regardées comme les $n$ observations d'un vecteur de $r$ variables réponses. Ces deux points de vue conduisent à l'élaboration de deux modèles, le modèle mixte univarié dans le premier cas et le modèle multivarié dans le second (voir condition de validation IV-1-5).
Nous étudierons le modèle mixte univarié dans lequel on considère que les modalités du facteur sujet ont été obtenues par échantillonnage, le facteur sujet est donc un facteur aléatoire. Le facteur occasion est lui un facteur à effets fixes.
Chaque donnée $y_{si}$ correspond alors à l'observation d'une variable aléatoire réelle $Y_{si}$ décrite par le modèle suivant : $Y_{si} = \mu_i + \pi_s + \varepsilon_{si}$
Où :

- les $\mu_i$ sont des constantes mesurant les effets fixes des modalités $i$ ($i$=1, 2, …, $r$) du facteur $O$
- les $\pi_s$ sont des variables aléatoires indépendantes de loi $\mathcal{N}(0, \sigma^2_\pi)$ mesurant les effets aléatoires des modalités $s$ ($s$=1, 2, …, $n$) du facteur $S$
- les résidus $\varepsilon_{si}$ sont des variables aléatoires indépendantes de loi $\mathcal{N}(0, \sigma^2)$. On suppose en plus que les $\pi_s$ et les $\varepsilon_{si}$ sont indépendantes.

On peut réécrire l'effet de la modalité $i$ sous la forme : $\mu_i = \mu + \alpha_i$ on a donc : $Y_{si} = \mu + \alpha_i + \pi_s + \varepsilon_{si}$
Où :

- $\mu$ constante s'interprétant comme un niveau général de réponse
- les $\alpha_i$ sont des constantes mesurant les effets fixes des modalités *i* et vérifiant : $\sum_i \alpha_i = 0$ (contrainte d'identifiabilité)

**Conséquence :** D'après ce modèle les $Y_{si}$ sont des variables aléatoires de loi $\mathcal{N}(\mu+\alpha_i, \sigma^2_\pi+\sigma^2)$.
L'effet du facteur fixe se traduit sur la moyenne de la variable réponse tandis que l'effet du facteur aléatoire se traduit sur la variance de la variable réponse.
Ce modèle, l'interaction entre le facteur sujet et le facteur occasion est confondue avec le résidu car pour chaque modalité de l'un et de l'autre on ne dispose que d'une seule observation.

**IV-1-3 Décomposition de la variation :**

$\mathbf{SCT}_{obs} = \mathbf{SC}_{inter\text{-}sujets} + \mathbf{SC}_{intra\text{-}sujets}$ $\quad$ $\mathbf{SC}_{intra\text{-}sujets} = \mathbf{SCO}_{obs} + \mathbf{SCR}_{obs}$

**Conclusion :** Dans le cas du plan complet $S\times O$, la variation totale se décompose de façon additive en : $\mathbf{SCT}_{obs} = \mathbf{SCS}_{obs} + \mathbf{SCO}_{obs} + \mathbf{SCR}_{obs}$

De même on a la décomposition des degrés de liberté : $nr$-1 = ($n$-1) + ($r$-1) + ($n$-1)($r$-1)

$$\mathbf{SCT}_{obs} = \sum_{s=1}^{n}\sum_{i=1}^{r}(y_{si} - \bar{y})^2 = nr \times \text{variance totale}$$

$$\mathbf{SCS}_{inter\text{-}sujet} = \mathbf{SCS}_{obs} = r\sum_{s=1}^{n}(\bar{y}_{s0} - \bar{y})^2 = nr \times \text{variance des moyennes des sujets}$$

$$\mathbf{SCO}_{obs} = n\sum_{i=1}^{r}(\bar{y}_{0i} - \bar{y})^2 = nr \times \text{variance des moyennes par occasion}$$

**SCR** est calculé par : $\mathbf{SCR}_{obs} = \mathbf{SCT}_{obs} - \mathbf{SCS}_{obs} - \mathbf{SCO}_{obs}$

**IV-1-4 Statistiques des tests :**

En considérant les moyennes des carrés : $\mathbf{MCO} = \dfrac{\mathbf{SCO}}{r-1}$ $\quad$ $\mathbf{MCR} = \dfrac{\mathbf{SCR}}{(r-1)(n-1)}$

**Statistique du test :** sous l'hypothèse $\mathbf{H_0}$, la statistique $\mathbf{F_O} = \dfrac{\mathbf{MCA}}{\mathbf{MCR}}$ suit une loi de Fischer à $r$ - 1 et ($n$ - 1)($r$ - 1) degrés de liberté

**Tableau d'analyse de la variance**

| Source de variation | **SC** | ddl | **MC** | $\mathbf{F_O}$ |
|---|---|---|---|---|
| Inter *S* | $\mathbf{SCS}_{obs}$ | $n$ - 1 | $\mathbf{MCS}_{obs}$ | |
| Intra *S* | $\mathbf{SC}_{intra\text{-}sujets}$ | $n(r - 1)$ | | |
| *O* | $\mathbf{SCO}_{obs}$ | $r$ - 1 | $\mathbf{MCO}_{obs}$ | $\mathbf{F_O}_{obs}$ |
| R | $\mathbf{SCR}_{obs}$ | ($n$ - 1)($r$ - 1) | $\mathbf{MCR}_{obs}$ | |
| Total | $\mathbf{SCT}_{obs}$ | $rn$ - 1 = $N$ - 1 | | |

**IV-1-5 Condition de validation :**

Les données peuvent être d'une part, considérées comme les $N$=$n\times r$ observations de la variable réponse $Y$ dans les $N$=$n\times r$ conditions expérimentales décrites par le croisement du facteur sujet et du facteur occasion. Elles peuvent d'autre part, être regardées comme les $n$ observations d'un vecteur de $r$ variables réponses. Ces deux points de vue conduisent à l'élaboration de deux modèles, le <u>modèle mixte univarié</u> dans le premier cas et le <u>modèle multivarié</u> dans le second.

Le fait de mesurer plusieurs fois la variable réponse sur le même sujet introduit des corrélations entre les observations faites sur ce même sujet. Dans le cas du modèle mixte univarié on montre que :

$$\operatorname{cov}(Y_{si}, Y_{s'i'}) = \begin{cases} 0 \text{ si } s \neq s' \\ \sigma_\pi^2 \text{ si } s = s' \text{ } et \text{ } i \neq i' \\ \sigma_\pi^2 + \sigma^2 \text{ si } s = s' \text{ } et \text{ } i = i' \end{cases}$$

$$\text{et } cor(Y_{si}, Y_{s'i'}) = \frac{\sigma_\pi^2}{\sigma_\pi^2 + \sigma^2}$$

Dans le cas du modèle multivarié on considère que les données sont les réalisations de *r* vecteurs aléatoires de dimension *n*. $Y_s=(Y_{s1}, Y_{s2}, \ldots, Y_{sr})$ indépendants et de même loi normale caractérisés par : $E(Y_{si})=\mu_i$ $\quad$ $\operatorname{Var}(Y_{si})=\sigma^2_i$ $\quad$ $\operatorname{cov}(Y_{si}, Y_{si'}) = \operatorname{cov}_{ii'}$

Le modèle mixte univarié est donc un cas particulier du modèle multivarié se caratérisant par un <u>matrice de variance-covariance</u> Σ de la forme :

Une telle matrice où tous les termes sont positifs et où tous les termes diagonaux sont égaux et tous les autres sont égaux est dite <u>matrice circulaire</u>, et cette propriété est appelée la <u>symétrie composée</u> de la matrice.

$$\Sigma = \begin{array}{c} \\ O_1 \\ O_2 \\ \cdots \\ O_i \\ \cdots \\ O_r \end{array} \begin{array}{c} \begin{array}{cccccc} O_1 & O_2 & \cdots & O_i & \cdots & O_r \end{array} \\ \begin{bmatrix} \sigma_\pi^2+\sigma^2 & \sigma_\pi^2 & \cdots & \sigma_\pi^2 & \cdots & \sigma_\pi^2 \\ \sigma_\pi^2 & \sigma_\pi^2+\sigma^2 & \cdots & \sigma_\pi^2 & \cdots & \sigma_\pi^2 \\ \vdots & \vdots & \ddots & \vdots & \cdots & \vdots \\ \sigma_\pi^2 & \sigma_\pi^2 & \cdots & \sigma_\pi^2+\sigma^2 & \cdots & \sigma_\pi^2 \\ \vdots & \vdots & \cdots & \vdots & \ddots & \vdots \\ \sigma_\pi^2 & \sigma_\pi^2 & \cdots & \sigma_\pi^2 & \cdots & \sigma_\pi^2+\sigma^2 \end{bmatrix} \end{array}$$

Sur la diagonale principale de la matrice des variances-covariances Σ se trouvent les variances et les éléments hors diagonale sont des covariances.

L'hypothèse d'homogénéité des variances (*homoscédasticité*) se traduit par l'égalité des éléments de la diagonale de la matrice Σ, Les éléments en dehors de la diagonale principale de la matrice Σ doivent être de même ordre.

De même on peut considérer la matrice des <u>corrélations</u> ρ :

De ce fait, dans le cas du choix de modèle mixte univarié, en calculant les variances-covariances nous obtenons $\hat{\Sigma}$ l'estimation de la matrice des variances-covariances (et aussi l'estimation de la matrice des corrélations).

Si les éléments de la diagonale de la matrice des variances-covariances $\hat{\Sigma}$ sont de même ordre, alors la condition de l'homoscédasticité (homogénéité des variances) est vérifiée.

$$\rho = \begin{array}{c} \\ O_1 \\ O_2 \\ \ldots \\ O_i \\ \ldots \\ O_r \end{array} \begin{array}{c} \begin{array}{cccccc} O_1 & O_2 & \ldots & O_i & \ldots & O_r \end{array} \\ \begin{bmatrix} 1 & \frac{\sigma_\pi^2}{\sigma_\pi^2+\sigma^2} & \ldots & \frac{\sigma_\pi^2}{\sigma_\pi^2+\sigma^2} & \ldots & \frac{\sigma_\pi^2}{\sigma_\pi^2+\sigma^2} \\ \frac{\sigma_\pi^2}{\sigma_\pi^2+\sigma^2} & 1 & \ldots & \frac{\sigma_\pi^2}{\sigma_\pi^2+\sigma^2} & \ldots & \frac{\sigma_\pi^2}{\sigma_\pi^2+\sigma^2} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \frac{\sigma_\pi^2}{\sigma_\pi^2+\sigma^2} & \frac{\sigma_\pi^2}{\sigma_\pi^2+\sigma^2} & \ldots & 1 & \ldots & \frac{\sigma_\pi^2}{\sigma_\pi^2+\sigma^2} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \frac{\sigma_\pi^2}{\sigma_\pi^2+\sigma^2} & \frac{\sigma_\pi^2}{\sigma_\pi^2+\sigma^2} & \ldots & \frac{\sigma_\pi^2}{\sigma_\pi^2+\sigma^2} & \ldots & 1 \end{bmatrix} \end{array}$$

Si les éléments hors diagonale de la matrice des variances-covariance (ou des corrélations) sont de même ordre que les variances, alors le choix du modèle mixte univarié est justifié. C'est-à-dire que la condition de la circularité (de sphéricité) est acquise.

Licence Psychologie – Analyse de la variance

*Chapitre V*

**Analyse du plan $S\times A\times B$**

Barumandzadeh T & El Methni M.

## V-1. *A* et *B* sont deux facteurs à effets fixes

### V-1-1 Généralités

On considère le plan complet $S\times A\times B$ défini par le croisement de trois facteurs $S$, $A$ et $B$.
$S$ est le facteur sujet (aléatoire) à $n$ modalités. $A$ et $B$ sont deux facteurs à effets fixes, $A$ possède $r$ modalités et $B$ $c$ modalités. Chacun des $n$ sujets est observé dans chacun des $r\times c$ croisements des modalités de $A$ et $B$.
On dispose donc de $N=n\times r\times c$ observations $y_{sij}$ que l'on présente sous la forme d'un tableau :

| Sujets | $a_1$ | | | | | … | $a_i$ | | | | | … | $a_r$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $b_1$ | … | $b_j$ | … | $b_c$ | … | $b_1$ | … | $b_j$ | … | $b_c$ | … | $b_1$ | … | $b_j$ | … | $b_c$ |
| 1 | $y_{111}$ | … | $y_{11j}$ | … | $y_{11c}$ | … | $y_{1i1}$ | … | $y_{1ij}$ | … | $y_{1ic}$ | … | $y_{1r1}$ | … | $y_{1rj}$ | … | $y_{1rc}$ |
| 2 | $y_{211}$ | … | $y_{21j}$ | … | $y_{21c}$ | … | $y_{2i1}$ | … | $y_{2ij}$ | … | $y_{2ic}$ | … | $y_{2r1}$ | … | $y_{2rj}$ | … | $y_{2rc}$ |
| … | … | … | … | … | … | … | … | … | … | … | … | … | … | … | … | … | … |
| s | $y_{s11}$ | … | $y_{s1j}$ | … | $y_{s1c}$ | … | $y_{si1}$ | … | $y_{sij}$ | … | $y_{sic}$ | … | $y_{sr1}$ | … | $y_{srj}$ | … | $y_{src}$ |
| … | … | … | … | … | … | … | … | … | … | … | … | … | … | … | … | … | … |
| n | $y_{n11}$ | … | $y_{n1j}$ | … | $y_{n1c}$ | … | $y_{ni1}$ | … | $y_{nij}$ | … | $y_{nic}$ | … | $y_{nr1}$ | … | $y_{nrj}$ | … | $y_{nrc}$ |

### V-1-2 Modèle univarié :

Les données peuvent être d'une part, considérées comme les $N = n\times r\times c$ observations d'une seule variable aléatoire $Y$ dans les $N = n\times r\times c$ conditions expérimentales décrites par le croisement des trois facteurs $S$, $A$ et $B$. Elles peuvent d'autre part, être regardées comme les $n$ observations d'un vecteur de $r\times c$ variables aléatoires. Ces deux points de vue conduisent à l'élaboration de deux modèles, le <u>modèle mixte univarié</u> dans le premier cas, et le <u>modèle multivarié</u> dans le second.
Comme pour le plan $S\times O$, nous allons étudier le modèle mixte univarié dans lequel nous considérons que les modalités du facteur sujet ont été obtenues par échantillonnage, le facteur sujet est donc un facteur aléatoire. Les facteurs $A$ et $B$ sont des facteurs à effets fixes. Chaque donnée $y_{sij}$ correspond alors à l'observation d'une variable aléatoire $Y_{sij}$ et on pose le modèle mixte suivant :

$$Y_{sij} = \mu_{ij} + \pi_s + (\alpha\pi)_{is} + (\beta\pi)_{js} + (\alpha\beta\pi)_{sij} + e_{sij}$$

Pour chaque croisement des modalités $s$, $i$ et $j$ nous ne disposons que d'une seule observation, l'interaction $(\alpha\beta\pi)_{sij}$ sera confondue avec le résidu et on pose : $\varepsilon_{sij}=(\alpha\beta\pi)_{sij} + e_{sij}$

$$\text{On a donc : } Y_{sij} = \mu_{ij} + \pi_s + (\alpha\pi)_{is} + (\beta\pi)_{js} + \varepsilon_{sij}$$

Où
Les $\mu_{ij}$ ($i$ =1, 2, …, $r$, $j$=1, 2, …, $c$) sont des constantes qui mesurent les effets fixes des modalités $(i,j)$ du croisement $A\times B$
Les $\pi_s$ ($s$ = 1, 2, …, $n$) sont des variables aléatoires indépendantes et identiquement distribuées (*iid*) de loi $\mathcal{N}(0\,;\sigma^2_\pi)$ qui mesurent les effets aléatoires des modalités $s$ du facteur $S$
Les $(\alpha\pi)_{si}$ ($s$ = 1, 2, …, $n$, $i$ = 1, 2, …, $r$) sont des variables aléatoires indépendantes et identiquement distribuées (*iid*) de loi $\mathcal{N}(0\,;\sigma^2_{\alpha\pi})$ qui mesurent les effets aléatoires d'interactions entre le facteur $S$ et le facteur $A$

Les $(\beta\pi)_{sj}$ ($s$ = 1, 2, …, $n$, $j$ = 1, 2, …, $c$) sont des variables aléatoires indépendantes et identiquement distribuées ($iid$) de loi $\mathcal{N}(0\,;\sigma^2_{\beta\pi})$ qui mesurent les effets aléatoires d'interactions entre le facteur $S$ et le facteur $B$

Les résidus $\varepsilon_{sij}$ sont des variables aléatoires indépendantes et identiquement distribuées ($iid$) de loi $\mathcal{N}(0\,;\sigma^2)$

De plus, on suppose que les résidus $\varepsilon_{sij}$ sont indépendantes des $\pi_s$, des $(\alpha\pi)_{si}$ et des $(\beta\pi)_{sj}$ .

Nous pouvons réécrire l'effet fixe de la modalité ($i$,$j$) sous la forme suivante :

$$\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

Où :

Le paramètre $\mu$ s'interprète comme un niveau général de réponse commun pour l'ensemble des observations,

Le paramètre $\alpha_i = \mu_i - \mu$ s'interprète comme l'effet de la modalité $i$ du facteur $A$

Le paramètre $\beta_j = \mu_j - \mu$ s'interprète comme l'effet de la modalité $j$ du facteur $B$

Le paramètre $(\alpha\beta)_{ij} = \mu_{ij} - \mu_i - \mu_j + \mu$ s'interprète comme l'effet de la modalité ($i$,$j$) du croisement $A\times B$ (l'effet d'interaction de $A$ et $B$).

Le modèle mixte univarié peut finalement s'écrire :

$$Y_{sij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \pi_s + (\alpha\pi)_{is} + (\beta\pi)_{js} + \varepsilon_{sij}$$

Où : les $Y_{sij}$ sont des variables aléatoires de loi : $\mathcal{N}(\mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}\ , \sigma^2_\pi + \sigma^2_{\alpha\pi} + \sigma^2_{\beta\pi} + \sigma^2)$

Une telle paramétrisation nécessite de rajouter les contraintes d'identifiabilité suivantes :

$$\sum_{i=1}^{r}\alpha_i = 0 \quad \sum_{j=1}^{c}\beta_i = 0 \quad \sum_{i=1}^{r}(\alpha\beta)_{ij} = 0 \quad \sum_{j=1}^{c}(\alpha\beta)_{ij} = 0$$

### V-1-3 Décomposition de la variation :

On commence par décomposer la variation totale en variation inter-sujets et variation intra-sujets :

$\mathbf{SCT}_{\text{obs}} = \mathbf{SC}_{\text{inter-sujets}} + \mathbf{SC}_{\text{intra-sujets}}$

$$\mathbf{SC}_{\text{inter-sujet}} = rc\sum_{s=1}^{n}(\bar{y}_{s00} - \bar{y})^2 = N \times \text{Variance des moyennes des sujets}$$

$$\mathbf{SC}_{\text{intra-sujet}} = \sum_{s,i,j}(y_{sij} - \bar{y}_{s00})^2$$

$\mathbf{SC}_{\text{intra-sujets}} = \mathbf{SC(}A\times B\mathbf{)}_{\text{obs}} + \mathbf{SC(}A\times B\mathbf{)}S_{\text{obs}} + \mathbf{SCR}_{\text{obs}}$

Avec : $\mathbf{SC(}A\times B\mathbf{)}_{\text{obs}} = \mathbf{SC(}A\mathbf{)}_{\text{obs}} + \mathbf{SC(}B\mathbf{)}_{\text{obs}} + \mathbf{SC(}AB\mathbf{)}_{\text{obs}}$

et : $\mathbf{SC(}A\times B\mathbf{)}S_{\text{obs}} = \mathbf{SC(}AS\mathbf{)}_{\text{obs}} + \mathbf{SC(}BS\mathbf{)}_{\text{obs}} + \mathbf{SC(}ABS\mathbf{)}_{\text{obs}}$

$\mathbf{SC(}ABS\mathbf{)}_{\text{obs}} = \mathbf{SCR}$

**Conclusion :** Dans le cas du plan complet $S\times A\times B$, la variation totale se décompose de façon additive en :

$$\boxed{\mathbf{SCT}_{\text{obs}} = \mathbf{SCS}_{\text{obs}} + \mathbf{SC}A_{\text{obs}} + \mathbf{SC}B_{\text{obs}} + \mathbf{SC}AB_{\text{obs}} + \mathbf{SC}AS_{\text{obs}} + \mathbf{SC}BS_{\text{obs}} + \mathbf{SCR}_{\text{obs}}}$$

De même on a la décomposition des degrés de liberté :

$N$-1=$ncr$-1 = ($n$-1) + $n$($rc$-1) = ($n$-1) + $n$($rc$-1) - ($c$-1)($r$-1)($n$-1) + ($c$-1)($r$-1)($n$-1)

$$\boxed{N-1 = ncr-1 = (n-1)+(r-1)+(c-1)+(c-1)(r-1)+(r-1)(n-1)+(c-1)(n-1)}$$

Les différentes sommes de carrées se calculent de façon habituelles par :

$$\mathbf{SCT}_{\text{obs}} = \sum_{s=1}^{n}\sum_{j=1}^{c}\sum_{i=1}^{r}(y_{sij} - \bar{y})^2 = N\times\text{variance de toutes les observations}$$

$$\mathbf{SCS}_{\text{obs}} = rc\sum_{s=1}^{n}(\bar{y}_{s00} - \bar{y})^2 = N\times\text{variance des } n \text{ moyennes des sujets}$$

$\mathbf{SCA}_{\text{obs}} = nc\sum_{i=1}^{r}(\bar{y}_{0i0} - \bar{y})^2$ =$N$×variance des $r$ moyennes par modalités de $A$

$\mathbf{SCB}_{\text{obs}} = nr\sum_{j=1}^{c}(\bar{y}_{00j} - \bar{y})^2$ =$N$×variance des $c$ moyennes par modalités de $B$

$\mathbf{SC(}\mathrm{A}\times\mathrm{B}\mathbf{)}_{\text{obs}} = n\sum_{i=1}^{r}\sum_{j=1}^{c}(\bar{y}_{0ij} - \bar{y})^2$ =$N$×variance des $r$×$c$ moyennes par croisement de $A$ et $B$

$\mathbf{SC}(\mathrm{AB})_{\text{obs}} = \mathbf{SC}(\mathrm{A}\times\mathrm{B})_{\text{obs}}$ **-** $\mathbf{SCA}_{\text{obs}}$ **-** $\mathbf{SCB}_{\text{obs}}$

$\mathbf{SC(}\mathrm{A}\times\mathrm{S}\mathbf{)}_{\text{obs}} = c\sum_{s=1}^{n}\sum_{i=1}^{r}(\bar{y}_{si0} - \bar{y})^2$ =$N$×variance des $r$×$n$ moyennes par croisement de $A$ et $S$

$\mathbf{SC}(\mathrm{AS})_{\text{obs}} = \mathbf{SC}(\mathrm{A}\times\mathrm{S})_{\text{obs}} - \mathbf{SCA}_{\text{obs}}$ **-** $\mathbf{SCS}_{\text{obs}}$

$\mathbf{SC(}\mathrm{B}\times\mathrm{S}\mathbf{)}_{\text{obs}} = r\sum_{s=1}^{n}\sum_{j=1}^{c}(\bar{y}_{s0j} - \bar{y})^2$ =$N$×variance des $n$×$c$ moyennes par croisement de $B$ et $S$

$\mathbf{SC}(\mathrm{BS})_{\text{obs}} = \mathbf{SC}(\mathrm{B}\times\mathrm{S})_{\text{obs}}$ **-** $\mathbf{SCB}_{\text{obs}}$ **-** $\mathbf{SCS}_{\text{obs}}$
$\mathbf{SCR}_{\text{obs}} = \mathbf{SCT}_{\text{obs}}$ **-** $\mathbf{SCS}_{\text{obs}}$ **-** $\mathbf{SCA}_{\text{obs}}$ **-** $\mathbf{SCB}_{\text{obs}}$ **-** $\mathbf{SC}(\mathrm{AB})_{\text{obs}}$ **-** $\mathbf{SC}(\mathrm{AS})_{\text{obs}}$ **-** $\mathbf{SC}(\mathrm{BS})_{\text{obs}}$

On ramène toutes les sommes de carrés à des moyennes de carrés en divisant par les degrés de liberté correspondants.

$\mathbf{MCA} = \dfrac{\mathbf{SCA}}{r-1} \quad \mathbf{MCB} = \dfrac{\mathbf{SCB}}{c-1} \quad \mathbf{MC(AS)} = \dfrac{\mathbf{SC(AS)}}{(r-1)(n-1)} \quad \mathbf{MCAB} = \dfrac{\mathbf{SC(AB)}}{(r-1)(c-1)}$

$\mathbf{MC(BS)} = \dfrac{\mathbf{SC(BS)}}{(c-1)(n-1)} \quad \mathbf{MCR} = \dfrac{\mathbf{SCR}}{(r-1)(c-1)(n-1)}$

**Statistique du test 1 :** sous l'hypothèse $\mathbf{H_0}_{\mathrm{A}}$, la statistique $\mathbf{F_A} = \dfrac{\mathbf{MCA}}{\mathbf{MC(AS)}}$ suit une loi de Fischer à $r$ - 1 et ($n$ - 1)($r$ - 1) degrés de liberté

**Statistique du test 2 :** sous l'hypothèse $\mathbf{H_0}_{\mathrm{B}}$, la statistique $\mathbf{F_B} = \dfrac{\mathbf{MCB}}{\mathbf{MC(BS)}}$ suit une loi de Fischer à $c$ - 1 et ($n$ - 1)($c$ - 1) degrés de liberté.

**Statistique du test 3 :** sous l'hypothèse $\mathbf{H_0}_{\mathrm{AB}}$, la statistique $\mathbf{F_{AB}} = \dfrac{\mathbf{MC(AB)}}{\mathbf{MCR}}$ suit une loi de Fischer à ($r$ - 1)($c$ - 1) et ($n$ - 1)($c$ - 1) ($r$ - 1) degrés de liberté.

**Tableau d'analyse de la variance**

| Source | $\mathbf{SC}_{\text{obs}}$ | ddl | $\mathbf{MC}_{\text{obs}}$ | $\mathbf{F}_{\text{obs}}$ |
|---|---|---|---|---|
| Inter-sujets : S | $\mathbf{SCS}_{\text{obs}}$ | $n$ - 1 | | |
| Intra-sujets : | | | | |
| A | $\mathbf{SCA}_{\text{obs}}$ | $r$ - 1 | $\mathbf{MCA}_{\text{obs}}$ | $\mathbf{F}_{\text{A obs}}$ |
| AS | $\mathbf{SC(AS)}_{\text{obs}}$ | ($r$ - 1)($n$ - 1) | $\mathbf{MC(AS)}_{\text{obs}}$ | |
| B | $\mathbf{SCB}_{\text{obs}}$ | $c$ - 1 | $\mathbf{MCB}_{\text{obs}}$ | $\mathbf{F}_{\text{B obs}}$ |
| BS | $\mathbf{SC(BS)}_{\text{obs}}$ | ($c$ - 1)($n$ - 1) | $\mathbf{MC(BS)}_{\text{obs}}$ | |
| AB | $\mathbf{SC(AB)}_{\text{obs}}$ | ($r$ - 1)($c$ - 1) | $\mathbf{MC(AB)}_{\text{obs}}$ | $\mathbf{F}_{\text{AB obs}}$ |
| R | $\mathbf{SCR}_{\text{obs}}$ | ($r$ - 1)($c$ - 1)($n$ - 1) | $\mathbf{MCR}_{\text{obs}}$ | |
| Total | $\mathbf{SCT}_{\text{obs}}$ | $N$ - 1 | | |

**V-1-5 Condition de validation :**

Le fait de mesurer plusieurs fois la variable réponse sur le même sujet introduit des corrélations entre les observations faites sur ce même sujet. Dans le cas du modèle mixte univarié on montre

$$\mathrm{cov}(Y_{sij}, Y_{s'i'j'}) = \begin{cases} 0 & \text{si } s \neq s' \\ \sigma_\pi^2 & \text{si } s = s' \; et \; i \neq i' \; et \; j \neq j' \\ \sigma_\pi^2 + \sigma_{\alpha\pi}^2 & \text{si } s = s' \; et \; i = i' \; et \; j \neq j' \\ \sigma_\pi^2 + \sigma_{\beta\pi}^2 & \text{si } s = s' \; et \; i \neq i' \; et \; j = j' \\ \sigma_\pi^2 + \sigma_{\alpha\pi}^2 + \sigma_{\beta\pi}^2 & \text{si } s = s' \; et \; i = i' \; et \; j = j' \end{cases}$$

que :

$$cor(Y_{sij}, Y_{s'i'j'}) = \frac{\sigma_\pi^2}{\sigma_\pi^2 + \sigma_{\alpha\pi}^2 + \sigma_{\beta\pi}^2 + \sigma^2}$$

$$cor(Y_{sij}, Y_{s'i'j}) = \frac{\sigma_\pi^2 + \sigma_{\alpha\pi}^2}{\sigma_\pi^2 + \sigma_{\alpha\pi}^2 + \sigma_{\beta\pi}^2 + \sigma^2}$$

$$cor(Y_{sij}, Y_{s'ij'}) = \frac{\sigma_\pi^2 + \sigma_{\beta\pi}^2}{\sigma_\pi^2 + \sigma_{\alpha\pi}^2 + \sigma_{\beta\pi}^2 + \sigma^2}$$

Dans le cas du modèle multivarié on considère que les données sont les réalisations de $n$ vecteurs aléatoires de dimension $rc$. $Y_s=(Y_{s1}, Y_{s2}, \ldots, Y_{src})$ indépendants et de même loi normale caractérisés par : $E(Y_{sk})=\mu_{ij}$ $\quad$ $\mathrm{Var}(Y_{sk})=\sigma^2_{ij}$ $\quad$ $\mathrm{cov}(Y_{sk}, Y_{sk'}) = \mathrm{cov}_{kk'}$

Le modèle mixte univarié est donc un cas particulier du modèle multivarié correspondant aux hypothèses de circularité de la matrice de variance-covariance $\Sigma$ de la forme :

## VI-1. *A* et *B* sont deux facteurs à effets fixes

### VI-1-1 Généralités

On considère le plan quasi-complet $S<A>\times B$ où les sujets sont répartis dans des groupes définis par les $r$ modalités du facteur $A$. Chaque sujet est observé dans les différentes occasions définies par les $c$ modalités du facteur $B$. Il y a $n$ sujets par groupe. On a donc échantillonné $n\times r$ sujets et on dispose donc de $N=n\times r\times c$ observations $y_{s(i)j}$ que l'on présente sous la forme d'un tableau :

| | $a_1$ | | | | | … | $a_i$ | | | | | … | $a_r$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Sujets* | $b_1$ | … | $b_j$ | … | $b_c$ | … | $b_1$ | … | $b_j$ | … | $b_c$ | … | $b_1$ | … | $b_j$ | … | $b_c$ |
| *1* | $y_{1(1)1}$ | … | $y_{1(1)j}$ | … | $y_{1(1)c}$ | … | $y_{1(i)1}$ | … | $y_{1(i)j}$ | … | $Y_{1(i)c}$ | … | $y_{1(r)1}$ | … | $y_{1(r)j}$ | … | $y_{1(r)c}$ |
| *2* | $y_{2(1)1}$ | … | $y_{2(1)j}$ | … | $y_{2(1)c}$ | … | $y_{2(i)1}$ | … | $y_{2(i)j}$ | … | $y_{2(i)c}$ | … | $y_{2(r)1}$ | … | $y_{2(r)j}$ | … | $y_{2(r)c}$ |
| … | … | … | … | … | … | … | … | … | … | … | … | … | … | … | … | … | … |
| *s* | $y_{s(1)1}$ | … | $y_{s(1)j}$ | … | $y_{s(1)c}$ | … | $y_{s(i)1}$ | … | $y_{s(i)j}$ | … | $y_{s(i)c}$ | … | $y_{s(r)1}$ | … | $y_{s(r)j}$ | … | $y_{s(r)c}$ |
| … | … | … | … | … | … | … | … | … | … | … | … | … | … | … | … | … | … |
| *n* | $y_{n(1)1}$ | … | $y_{n(1)j}$ | … | $y_{n(1)c}$ | … | $y_{n(i)1}$ | … | $y_{n(i)j}$ | … | $y_{n(i)c}$ | … | $y_{n(r)1}$ | … | $y_{n(r)j}$ | … | $y_{n(r)c}$ |

### VI-1-2 Modèle univarié :

Les données peuvent être d'une part, considérées comme les $N = n\times r\times c$ observations d'une seule variable aléatoire $Y$ dans les $N = n\times r\times c$ conditions expérimentales décrites par le croisement des trois facteurs $S$ emboîté dans $A$ et $B$. Elles peuvent d'autre part, être regardées comme les $n\times r$ observations d'un vecteur de $c$ variables aléatoires. Ces deux points de vue conduisent à l'élaboration de deux modèles, le <u>modèle mixte univarié</u> dans le premier cas, et le <u>modèle multivarié</u> dans le second.

Comme pour les plans $S\times O$ et $S\times A\times B$ nous allons étudier le modèle mixte univarié dans lequel nous considérons que les modalités du facteur sujet ont été obtenues par échantillonnage, le facteur sujet est donc un facteur aléatoire. Les facteurs $A$ et $B$ sont des facteurs à effets fixes. Chaque donnée $y_{s(i)j}$ correspond alors à l'observation d'une variable aléatoire $Y_{s(i)j}$ et on pose le modèle mixte suivant :

$$Y_{s(i)j} = \mu_{ij} + \pi_{s(i)} + (\pi\beta)_{sj} + e_{s(i)j}$$

Pour chaque croisement des modalités $s$ et $j$ nous ne disposons que d'une seule observation, l'interaction $(\pi\beta)_{sj}$ sera confondue avec le résidu et on pose : $\varepsilon_{s(i)j}=(\pi\beta)_{sj} + e_{s(i)j}$

$$\text{On a donc : } Y_{s(i)j} = \mu_{ij} + \pi_{s(i)} + \varepsilon_{s(i)j}$$

Où

Les $\mu_{ij}$ ($i$ =1, 2, …, $r$, $j$=1, 2, …, $c$) sont des constantes qui mesurent les effets fixes des modalités $(i,j)$ du croisement $A\times B$

Les $\pi_{s(i)}$ ($s$ = 1, 2, …, $n$, $i$ =1, 2, …, $r$) sont des variables aléatoires indépendantes et identiquement distribuées (*iid*) de loi $\mathcal{N}(0\,;\sigma^2_\pi)$ qui mesurent les effets aléatoires des modalités $s$ du facteur $S$

Les résidus $\varepsilon_{s(i)j}$ sont des variables aléatoires indépendantes et identiquement distribuées (*iid*) de loi $\mathcal{N}(0\,;\sigma^2)$

De plus, on suppose que les résidus $\varepsilon_{s(i)j}$ sont indépendantes des $\pi_{s(i)}$.

Nous pouvons réécrire l'effet fixe de la modalité ($i$,$j$) sous la forme suivante :

$$\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

Où :
Le paramètre $\mu$ s'interprète comme un niveau général de réponse commun pour l'ensemble des observations,
Le paramètre $\alpha_i = \mu_i - \mu$ s'interprète comme l'effet de la modalité $i$ du facteur $A$
Le paramètre $\beta_j = \mu_j - \mu$ s'interprète comme l'effet de la modalité $j$ du facteur $B$
Le paramètre $(\alpha\beta)_{ij} = \mu_{ij} - \mu_i - \mu_j + \mu$ s'interprète comme l'effet d'interaction des modalités $i$ et $j$.
Le modèle mixte univarié peut finalement s'écrire :

$$Y_{s(i)j} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \pi_{s(i)} + \varepsilon_{s(i)j}$$

Où : les $Y_{s(i)j}$ sont des variables aléatoires de loi : $\mathcal{N}(\mu + \alpha_i + \beta_j\ , \sigma_\pi^2 + \sigma^2)$

Une telle paramétrisation nécessite de rajouter les contraintes d'identifiabilité suivantes :

$$\sum_{i=1}^{r} \alpha_i = 0 \quad \sum_{j=1}^{c} \beta_i = 0 \quad \sum_{i=1}^{r} (\alpha\beta)_{ij} = 0 \quad \sum_{j=1}^{c} (\alpha\beta)_{ij} = 0$$

**VI-1-3 Décomposition de la variation :**

On note : $\bar{y}_{s(i)0}$ la moyenne des $c$ observations de la modalité $s(i)$ de l'emboîtement $S\langle A\rangle$

$\bar{y}_{0(i)0}$ la moyenne des $nc$ observations de la modalité $i$ de $A$

$\bar{y}_{0(0)j}$ la moyenne des $nr$ observations de la modalité $j$ de $B$

$\bar{y}_{0(i)j}$ la moyenne des $n$ observations de la modalité ($i$,$j$) du croisement $A\times B$ et $\bar{y}$ la moyenne générale

On commence par décomposer la variation totale en variation inter-sujets et variation intra-sujets :
$\mathbf{SCT}_{\text{obs}} = \mathbf{SC}_{\text{inter-sujets}} + \mathbf{SC}_{\text{intra-sujets}}$

$$\mathbf{SC}_{\text{inter-sujet}} = \mathbf{SCS}_{\text{obs}} = c\sum_{i}\sum_{s(i)} (\bar{y}_{s(i)0} - \bar{y})^2 = N \times \text{Variance des moyennes des sujets}$$

$$\mathbf{SC}_{\text{intra-sujet}} = \sum_{i}\sum_{s(i)}\sum_{j} (y_{s(i)j} - \bar{y}_{s(i)0})^2$$

On décompose la variation inter-sujets en variation inter $A$ et variation intra $A$. On obtient :
$\mathbf{SC}_{\text{inter-sujets}} = \mathbf{SCS}_{\text{obs}} = \mathbf{SC}_{\text{inter } A} + \mathbf{SC}_{\text{intra } A}$ où :

$$\mathbf{SC}_{\text{inter } A} = \mathbf{SCA}_{\text{obs}} = nc\sum_{i=1}^{r} (\bar{y}_{0(i)0} - \bar{y})^2$$

$$\mathbf{SC}_{\text{intra } A} = \mathbf{SCS(A)}_{\text{obs}} = c\sum_{i}\sum_{s(i)} (\bar{y}_{s(i)0} - \bar{y}_{0(i)0})^2$$

La variation intra-sujets est due aux effets du facteur $B$ et à l'effet des autres facteurs non contrôlés.
L'effet du facteur $B$ se fait à deux niveaux : effet principal et effet d'interaction avec $A$. On a :

$$\mathbf{SC}_{\text{intra-sujets}} = \mathbf{SC}B_{\text{obs}} + \mathbf{SC}(AB)_{\text{obs}} + \mathbf{SCR}_{\text{obs}}$$

Avec : $\mathbf{SC}(AB)_{\text{obs}} = \mathbf{SC}(A\times B)_{\text{obs}} - \mathbf{SC}(A)_{\text{obs}} - \mathbf{SC}(B)_{\text{obs}}$
**Conclusion :** Dans le cas du plan complet $S\langle A\rangle\times B$, la variation totale se décompose de façon additive en :

$$\boxed{\mathbf{SCT}_{\text{obs}} = \mathbf{SC}A_{\text{obs}} + \mathbf{SC}S(A)_{\text{obs}} + \mathbf{SC}B_{\text{obs}} + \mathbf{SC}AB_{\text{obs}} + \mathbf{SCR}_{\text{obs}}}$$

De même on a la décomposition des degrés de liberté :

$$\boxed{N-1 = ncr-1 = (r-1)+r(n-1)+(c-1)+(r-1)(c-1)+r(c-1)(n-1)}$$

Les différentes sommes de carrées se calculent de façon habituelles par :

$$\mathbf{SCT}_{\text{obs}} = \sum_{s=1}^{n}\sum_{j=1}^{c}\sum_{i=1}^{r}(y_{s(i)j} - \bar{y})^2$$ =$N\times$variance de toutes les observations

$$\mathbf{SCA}_{\text{obs}} = nc\sum_{i=1}^{r}(\bar{y}_{0(i)0} - \bar{y})^2$$ =$N\times$variance des $r$ moyennes par modalités de $A$

$$\mathbf{SCS}_{\text{obs}} = c\sum_{s=1}^{n}(\bar{y}_{s(i)0} - \bar{y})^2$$ =$N\times$variance des $r\times n$ moyennes des sujets

**SC**S(A)$_{\text{obs}}$ = **SCS**$_{\text{obs}}$ - **SCA**$_{\text{obs}}$

$$\mathbf{SCB}_{\text{obs}} = nr\sum_{j=1}^{c}(\bar{y}_{0(0)j} - \bar{y})^2$$ =$N\times$variance des $c$ moyennes par modalités de $B$

$$\mathbf{SC(}\mathrm{A}\times\mathrm{B}\mathbf{)}_{\text{obs}} = n\sum_{i=1}^{r}\sum_{j=1}^{c}(\bar{y}_{0(i)j} - \bar{y})^2$$ =$N\times$variance des $r\times c$ moyennes par croisement de $A$ et $B$

**SC**(AB)$_{\text{obs}}$ = **SC**(A×B)$_{\text{obs}}$ - **SCA**$_{\text{obs}}$ - **SCB**$_{\text{obs}}$
**SCR**$_{\text{obs}}$ = **SCT**$_{\text{obs}}$ - **SCA**$_{\text{obs}}$ - **SCB**$_{\text{obs}}$ - **SC**(AB)$_{\text{obs}}$ - **SC**S(A)$_{\text{obs}}$

On ramène toutes les sommes de carrés à des moyennes de carrés en divisant par les degrés de liberté correspondants.

$$\mathbf{MCA} = \frac{\mathbf{SCA}}{r-1} \quad \mathbf{MCB} = \frac{\mathbf{SCB}}{c-1} \quad \mathbf{MCS(A)} = \frac{\mathbf{SCS(A)}}{r(n-1)} \quad \mathbf{MCAB} = \frac{\mathbf{SC(AB)}}{(r-1)(c-1)} \quad \mathbf{MCR} = \frac{\mathbf{SCR}}{r(n-1)(c-1)}$$

Ceci nous permet de tester l'existence des différents effets des facteurs. On peut construire des tests indépendants sur chacune des sources de variations :

**Statistique du test 1 : :** sous l'hypothèse **H$_{\mathbf{0}}$**$_{\text{A}}$, la statistique $\mathbf{F_A} = \frac{\mathbf{MCA}}{\mathbf{MCS(A)}}$ suit une loi de Fischer à $r$ - 1 et $r$($n$ - 1) degrés de liberté.

**Statistique du test 2 :** sous l'hypothèse **H$_{\mathbf{0}}$**$_{\text{B}}$, la statistique $\mathbf{F_B} = \frac{\mathbf{MCB}}{\mathbf{MCR}}$ suit une loi de Fischer à $c$ - 1 et $r$($n$ - 1)($c$ - 1) degrés de liberté.

**Statistique du test 3 :** sous l'hypothèse **H$_{\mathbf{0}}$**$_{\text{AB}}$, la statistique $\mathbf{F_{AB}} = \frac{\mathbf{MC(AB)}}{\mathbf{MCR}}$ suit une loi de Fischer à ($r$ - 1)($c$ - 1) et $r$($n$ - 1)($c$ - 1) degrés de liberté

**Tableau d'analyse de la variance**

| Source | **SC**$_{\text{obs}}$ | ddl | **MC**$_{\text{obs}}$ | **F**$_{\text{obs}}$ |
|---|---|---|---|---|
| Intra-S : | | | | |
| A | **SCA**$_{\text{obs}}$ | $r$ - 1 | **MCA**$_{\text{obs}}$ | **F**$_{\text{A obs}}$ |
| S(A) | **SCS(A)**$_{\text{obs}}$ | $r$($n$ - 1) | **MCS(A)**$_{\text{obs}}$ | |
| Inter-S : | | | | |
| B | **SCB**$_{\text{obs}}$ | $c$ - 1 | **MCB**$_{\text{obs}}$ | **F**$_{\text{B obs}}$ |
| AB | **SC(AB)**$_{\text{obs}}$ | ($r$ - 1)($c$ - 1) | **MC(AB)**$_{\text{obs}}$ | **F**$_{\text{AB obs}}$ |
| R | **SCR**$_{\text{obs}}$ | $r$($c$ - 1)($n$ - 1) | **MCR**$_{\text{obs}}$ | |
| Total | **SCT**$_{\text{obs}}$ | $N$ - 1 | | |

**VI-1-5 Condition de validation :**

Le fait de mesurer plusieurs fois la variable réponse sur le même sujet introduit des corrélations entre les observations faites sur ce même sujet. Dans le cas du modèle mixte univarié on montre que :

$$\operatorname{cov}(Y_{s(i)j}, Y_{s'(i')j'}) = \begin{cases} 0 & \text{si } s(i) \neq s'(i') \\ \sigma_\pi^2 & \text{si } j \neq j' \\ \sigma_\pi^2 + \sigma^2 & \text{si } j = j' \end{cases}$$

$$cor(Y_{s(i)j}, Y_{s'(i')j'}) = \frac{\sigma_\pi^2}{\sigma_\pi^2 + \sigma^2}$$

Dans le cas du modèle multivarié on considère que les données sont les réalisations de *n* vecteurs aléatoires de dimension *rc*. $Y_s=(Y_{s(i)1}, Y_{s(i)2}, \ldots, Y_{s(i)c})$ indépendants et de même loi normale caractérisés par : $E(Y_{s(i)j})=\mu_{ij}$ $\operatorname{Var}(Y_{s(i)j})=\sigma^2_{ij}$ $\operatorname{cov}(Y_{s(i)j}, Y_{s(i)j'}) = \operatorname{cov}_{(i)jj'}$

Le modèle mixte univarié est donc un cas particulier du modèle multivarié correspondant aux hypothèses de circularité et d'homogénéité des matrices des variances-covariances $\Sigma_\iota$ ($i$ =1, 2, …, $r$) de la forme :

$$\Sigma_i = \begin{bmatrix} \sigma_\pi^2+\sigma^2 & \sigma_\pi^2 & \cdots & \sigma_\pi^2 & \cdots & \sigma_\pi^2 \\ \sigma_\pi^2 & \sigma_\pi^2+\sigma^2 & \cdots & \sigma_\pi^2 & \cdots & \sigma_\pi^2 \\ \vdots & \vdots & \ddots & \vdots & \cdots & \vdots \\ \sigma_\pi^2 & \sigma_\pi^2 & \cdots & \sigma_\pi^2+\sigma^2 & \cdots & \sigma_\pi^2 \\ \vdots & \vdots & \cdots & \vdots & \ddots & \vdots \\ \sigma_\pi^2 & \sigma_\pi^2 & \cdots & \sigma_\pi^2 & \cdots & \sigma_\pi^2+\sigma^2 \end{bmatrix}$$

De même on peut considérer la matrice des corrélations ρ :

$$\rho = \begin{bmatrix} 1 & \frac{\sigma_\pi^2}{\sigma_\pi^2+\sigma^2} & \ldots & \frac{\sigma_\pi^2}{\sigma_\pi^2+\sigma^2} & \ldots & \frac{\sigma_\pi^2}{\sigma_\pi^2+\sigma^2} \\ \frac{\sigma_\pi^2}{\sigma_\pi^2+\sigma^2} & 1 & \ldots & \frac{\sigma_\pi^2}{\sigma_\pi^2+\sigma^2} & \ldots & \frac{\sigma_\pi^2}{\sigma_\pi^2+\sigma^2} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \frac{\sigma_\pi^2}{\sigma_\pi^2+\sigma^2} & \frac{\sigma_\pi^2}{\sigma_\pi^2+\sigma^2} & \ldots & 1 & \ldots & \frac{\sigma_\pi^2}{\sigma_\pi^2+\sigma^2} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \frac{\sigma_\pi^2}{\sigma_\pi^2+\sigma^2} & \frac{\sigma_\pi^2}{\sigma_\pi^2+\sigma^2} & \ldots & \frac{\sigma_\pi^2}{\sigma_\pi^2+\sigma^2} & \ldots & 1 \end{bmatrix}$$

# Chapitre VII

## Analyse des comparaisons

Barumandzadeh T & El Methni M.

### VII-1. Généralités

### VII-1-1 Notion de comparaison

Contrairement à l'analyse de la variance canonique nous nous intéressons aux situations où les hypothèses portent sur des comparaisons entre les effets associés à certaines modalités des facteurs. De ce fait on ne considère que des facteurs à effets fixes. Dans le cas d'un facteur à effets fixes, le modèle pose que les effets du facteur portent sur la moyenne de la VD. Comparer les effets de certaines modalités revient donc à comparer les moyennes qui leur sont associées dans le modèle.

**Définition 1 :** On appelle <u>contraste</u> entre $k$ moyennes $\mu_1, \mu_2, \ldots, \mu_k$ toute combinaison linéaire de la forme : $C = \sum_{i=1}^{k} c_i \mu_i$ où $\sum_{i=1}^{k} c_i = 0$

On appelle <u>comparaison</u> tout ensemble de contrastes.

### VII-1-2 Construction d'un contraste

Une hypothèse portant sur une comparaison entre des moyennes est traduite par la donnée d'un contraste entre les moyennes. Ce contraste est caractérisé par la donnée d'un ensemble de coefficients de somme nulle. Nous donnons un procédé particulier permettant de choisir ces coefficients. Ce procédé repose sur une propriété classique de la moyenne arithmétique pondérée.

**Proposition 1 :** Etant donnée un ensemble de valeurs et une pondération quelconque sur cet ensemble de valeurs, la somme des écarts pondérés à la moyenne pondérée est nulle.

Comparer des moyennes revient à

- Ordonner les moyennes ou les écarts entre les moyennes.
- Associer à chaque moyenne un rang selon l'hypothèse à tester.
- Calculer la moyenne des rangs, puis centrer les rangs par rapport à leur moyenne.
- Les coefficients sont alors calculés à partir des rangs centrés en utilisant la proposition précédente.

**Remarque 1 :** L'usage conduit à considérer des pondérations proportionnelles aux poids des modalités. Ainsi quand les groupes sont équilibrés, on calcule une simple moyenne arithmétique. Par contre, dans le cas d'un plan non équilibré, les effectifs des groupes constituent les poids des modalités.

Certaines moyennes peuvent ne pas être concernées par des hypothèses. Les coefficients relatifs à ces moyennes seront nuls.

### VII-1-3 Tests d'hypothèses sur les comparaisons

Nous avons vu que poser une hypothèse sur certains effets du facteur revient à construire un contraste $C$. Pour définir le test d'hypothèses, nous pouvons écrire les hypothèses de test sur le contraste $C$ : $H_0 : C = 0$ contre $H_1 : C \neq 0$ (respectivement $C>0$ ou $C<0$)

Dans le cas d'une comparaison simple sur deux modalités particulières, le contraste correspondant est écrit sur les moyennes théoriques $\mu_1$ et $\mu_2$ : $C = \mu_1 - \mu_2$

et la statistique de test est définie à partir des observations : $\hat{C} = \overline{Y}_1 - \overline{Y}_2$

La distribution d'échantillonnage de la statistique de test $\hat{C} = \overline{Y}_1 - \overline{Y}_2$ dépend des hypothèses faites sur le modèle, (taille, échantillons, … ) et peut être soit la loi normale soit la loi de Student.

Dans le cas général, la construction des tests de comparaisons est fondée sur l'évaluation des contrastes à partir des observations. Pour effectuer un test d'hypothèses sur le contraste $C = \sum_{i=1}^{k} c_i \mu_i$ où $\sum_{i=1}^{k} c_i = 0$ la statistique de test est définie par : $\hat{C} = \sum_{i=1}^{k} c_i \overline{Y}_i$

On généralise ainsi le test de comparaison entre deux moyennes au test de comparaison entre deux groupes de moyennes.

**VII-1-4 Les erreurs de première espèce**

La question principale qui se pose dans tout examen des procédures de comparaisons multiples est liée à la probabilité de commettre des erreurs de première espèce. La plupart des différences entre les méthodes proposées sont dues au fait qu'on y adopte des approches différentes dans la manière de contrôler les erreurs. Le problème est en partie technique, mais il relève bien plus encore d'une question subjective : comment veut-on définir le taux d'erreur et quelle marge se donne-t-on pour le taux maximal d'erreur possible ?

Étant donné un ensemble de comparaisons, on est amené à distinguer deux types d'erreurs de première espèce :

- l'erreur par comparaison notée $\alpha$(EC) :

$\alpha$(EC) = P(rejeter $H_0$ à tort) dans une comparaison donnée.

Nous avons déjà utilisé l'erreur par comparaison (EC). Il s'agit de la probabilité de commettre une erreur de première espèce dans une comparaison donnée. Si, par exemple, nous procédons à une comparaison en effectuant un test *t* entre deux groupes et rejetons l'hypothèse nulle parce que notre *t* excède $t_{0,05}$, nous travaillons avec une erreur par comparaison de $\alpha$=0,05.

- l'erreur de l'ensemble notée $\alpha$(EE) :

$\alpha$(EE) = P(rejeter au moins une fois $H_0$ à tort) dans l'ensemble des comparaisons.

Lorsque nous aurons effectué un ensemble de comparaisons entre nos moyennes de groupes, nous arrivons à un ensemble (souvent appelé famille) de conclusions. Par exemple, la famille pourrait se composer des affirmations : $\mu_1 < \mu_2$ et $\mu_3 < \mu_4$ et $\mu_1 < (\mu_3 + \mu_4)/2$.

La probabilité de voir cette famille de conclusions contenir au moins une erreur de première espèce est appelée erreur de l'ensemble $\alpha$(EE).

De nombreuses procédures que nous examinons visent de façon spécifique à contrôler l'erreur d'ensemble $\alpha$(EE).

Dans une expérience qui n'implique qu'une seule comparaison on a $\alpha$(EC) = $\alpha$(EE). Par contre, à mesure que le nombre de comparaisons augmente, ces deux probabilités se différencient. On peut donner un calcul approché de la relation entre ces deux probabilités $\alpha$(EC) et $\alpha$(EE).

Considérons une famille de *r* comparaisons $C^1, C^2, \ldots, C^r$ et soit $\alpha$ le seuil de signification pour chaque comparaison simple : $\alpha$(EC) = $\alpha$.

Pour une comparaison donnée, P(ne pas rejeter $H_0$) = 1-$\alpha$.

Pour un ensemble de comparaisons :

P(aucune erreur de première espèce)=P(pas d' erreur de première espèce pour $C^1 \cap$ pas d' erreur de première espèce pour $C^2 \cap \ldots \cap$ pas d' erreur de première espèce pour $C^r$)

Si les comparaisons sont indépendantes alors : P(aucune erreur de première espèce)=P(pas d' erreur de première espèce pour $C^1$ × P(pas d' erreur de première espèce pour $C^2$) × … × P(pas d' erreur de première espèce pour $C^r$)=(1-α)×(1-α)×…×(1-α)=$(1-\alpha)^r$
Et donc : $\alpha(EE) = 1 - (1 - \alpha(EC))^r$.

Par exemple, pour $c = 5$ et α(EC) = 0,05, on obtient α(EE) = 1 - (1 - 0,05)$^5$ = 0,226. On constate ici l'importance de contrôler l'erreur de l'ensemble.

Si on fixe la valeur de α(EE) dans le cas de comparaisons indépendantes, on peut déduire la valeur de α(EC) pour chacune des comparaisons : $\alpha(\mathrm{EC}) = 1 - (1 - \alpha(\mathrm{EE}))^{\frac{1}{r}}$

Par exemple pour cinq comparaisons indépendantes : α(EE) = 10% ⇒ α(EC) = 2,09%
α(EE) = 5% ⇒ α(EC) = 1,02% α(EE) = 1% ⇒ α(EC) = 0,20%

**VII-1-5 Comparaisons a priori et a posteriori**

Une comparaison traduit une hypothèse de l'expérimentateur. On peut distinguer alors deux types de comparaisons :

- les comparaisons a priori définies avant la collecte des données et qui traduisent une hypothèse théorique de l'expérimentateur,
- les comparaisons a posteriori planifiées lorsque l'expérimentateur a rassemblé les données, examiné les moyennes et noté quelles moyennes sont éloignées et quelles moyennes sont proches l'une de l'autre.

La distinction entre deux types d'erreur apparaît bien lorsque l'on compare les procédures de comparaisons a priori et les procédures de comparaisons a posteriori. Considérons le cas où le facteur possède 5 modalités et nous avons 5 moyennes. Dans ce cas, il y a 10 comparaisons possibles impliquant des paires de moyennes (ex. $\overline{Y}_1$ par rapport à $\overline{Y}_2$, $\overline{Y}_1$ par rapport à $\overline{Y}_3$, etc.).

Supposons que l'hypothèse nulle (absence d'effet) soit vérifiée.

Pour effectuer ce test nous sommes amenés à comparer les moyennes empiriques. Supposons que $\overline{Y}_3$ et $\overline{Y}_4$ soient suffisamment éloignées pour nous amener à rejeter l'hypothèse $\mu_3 = \mu_4$, et aucun autre test ne révélant une différence significative entre les moyennes. Si nous devons planifier à l'avance une comparaison unique parmi les 10 comparaisons par paires possibles, nous avons une chance sur dix de faire la comparaison qui implique une erreur de première espèce. Par contre, si nous examinons les données avant de faire le test, nous sommes certains de faire une erreur de première espèce. En effet, au vu des données, même si on veut n'effectuer qu'une seule comparaison, on procède implicitement à l'examen de toutes les comparaisons par paires avant de choisir la plus significative. Cela équivaut alors à la situation où l'on effectue une famille de comparaisons.

Les procédures de comparaisons a posteriori ajustent α(EC) comme s'il s'agissait d'une famille de comparaisons.

Cet exemple montre que si les comparaisons sont planifiées a priori et constituent un sous-ensemble strict de l'ensemble de toutes les comparaisons possibles, alors la probabilité de commettre une erreur de première espèce est plus petite que si les comparaisons interviennent a posteriori.

### VII-2. Analyse des comparaisons a priori dans un plan à un facteur

On étudie des tests portant sur des hypothèses a priori dans le cas d'un plan (inter $S<A>$ ou intra $S\times A$) où $A$ est un facteur à effets fixes. Le facteur $A$ à $k$ modalités et le plan peut être soit équilibré et dans ce cas on dispose de $n$ observations de la variable dépendante $Y$ dans chacune des $k$ conditions expérimentales ($N=kn$), soit un plan non équilibré et on dispose de $n_i$ observations de la variable dépendante $Y$ dans la modalité $i$ du facteur $A$ ( $N = \sum_{i=1}^{k} n_i$ ).

Rappelons que les données sont alors décrites par le modèle : l'observation du sujet $s$ dans la condition $i$ est la réalisation d'une variable aléatoire $Y_{si}$ de loi $\mathcal{N}(\mu_i, \sigma^2_Y)$, ($s=1,\ldots, n_i$ ; $i=1,\ldots, k$)

avec : $\sigma_Y^2 = \begin{cases} \sigma^2 & \text{dans le cas d'un plan inter} \\ \sigma^2 + \sigma_\pi^2 & \text{dans le cas d'un plan intra} \end{cases}$

#### VII-2-1 Analyse d'un contraste

Soit $C$ un contraste entre les moyennes : $C = \sum_{i=1}^{k} c_i \mu_i$ où $\sum_{i=1}^{k} c_i = 0$

On veut tester $H_0$ : $C=0$ contre $H_1$ : $C\neq 0$ (respectivement $C>0$ ou $C<0$) au vu des $N$ observations de la variable dépendante $Y$.

**Proposition 2 :** En estimant les moyennes $\mu_i$ par les $\overline{Y}_i$ on peut construire un estimateur de $C$ par : $\hat{C} = \sum_{i=1}^{k} c_i \overline{Y}_i$ .

Sous les hypothèses du modèle, $\hat{C}$ est une variable aléatoire normalement distribuée d'espérance : $\mathrm{E}(\hat{C}) = C$ et de variance : $\mathrm{Var}(\hat{C}) = \sum_{i=1}^{k} c_i \frac{\sigma^2}{n_i}$

L'estimation $\hat{c} = \sum_{i=1}^{k} c_i \overline{y}_i$ est une réalisation de cette variable aléatoire $\hat{C}$ .

Sous l'hypothèse nulle $H_0$ la moyenne de $\hat{C}$ est nulle et on est donc ramené au test de comparaison à 0 de la moyenne d'une loi normale de variance inconnue. On peut procéder de deux manières.

**Test de Student :**

Pour un plan équilibré on prendra la statistique de test : $T = \dfrac{\hat{C}}{\dfrac{S}{\sqrt{n}}\sqrt{\sum_{i=1}^{k} c_i^2}} = \dfrac{\hat{C}\sqrt{n}}{S\sqrt{\sum_{i=1}^{k} c_i^2}}$

Pour un plan non équilibré on prendra la statistique de test : $T = \dfrac{\hat{C}}{S\sqrt{\sum_{i=1}^{k} \frac{c_i^2}{n_i}}}$

Où $S^2$ est un estimateur de $\sigma^2$. Or on sait que MCR est aussi un estimateur de $\sigma^2$, on a donc :

Pour un plan équilibré on prendra la statistique de test : $T = \dfrac{\hat{C}\sqrt{n}}{\sqrt{\mathrm{MCR}\sum_{i=1}^{k} c_i^2}}$

Pour un plan non équilibré on prendra la statistique de test : $T = \dfrac{\hat{C}}{\sqrt{\mathrm{MCR}\sum_{i=1}^{k} \frac{c_i^2}{n_i}}}$

**Proposition 3 :** Sous l'hypothèse nulle $H_0$ la statistique $T$ suit une loi de Student à $\nu$=ddl(SCR) degrés de liberté.

**Décision statistique :**

Le test $H_0 : C=0$ contre $H_1 : C\neq 0$ de niveau $\alpha$ est défini par la règle de décision : si $|t|=|T_{obs}|\geq\lambda_\alpha$ on rejette H0 avec $\lambda_\alpha$ donné par la table de Student à $\nu$ degrés de liberté : $P(\lambda_\alpha)=1-\alpha/2$

Le test $H_0 : C=0$ contre $H_1 : C>0$ de niveau $\alpha$ est défini par la règle de décision : si $t=T_{obs}\geq\lambda_\alpha$ on rejette H0 avec $\lambda_\alpha$ donné par la table de Student à $\nu$ degrés de liberté : $P(\lambda_\alpha)=1-\alpha$

Le test $H_0 : C=0$ contre $H_1 : C<0$ de niveau $\alpha$ est défini par la règle de décision : si $t=T_{obs}\leq-\lambda_\alpha$ on rejette H0 avec $\lambda_\alpha$ donné par la table de Student à $\nu$ degrés de liberté : $P(\lambda_\alpha)=1-\alpha$

**Test de Fisher :**

On a vu qu'un contraste exprime la différence entre deux groupes de moyennes. En regroupant les moyennes correspondantes à chacun des deux groupes, on se ramène à l'analyse de la variance à un facteur à deux modalités. MCR est calculé par l'analyse de variance canonique. On montre que :

Pour un plan équilibré : $\mathrm{SC}_{\text{contraste}} = \mathrm{SC}_{\text{inter}} = \dfrac{n(\hat{C})^2}{\sum_{i=1}^{k} c_i^2}$

Pour un plan non équilibré : $\mathrm{SC}_{\text{contraste}} = \mathrm{SC}_{\text{inter}} = \dfrac{(\hat{C})^2}{\sum_{i=1}^{k} \frac{c_i^2}{n_i}}$

Cette somme des carrés est appelée <u>somme des carrés du contraste</u> et notée $SC_C$. On peut présenter les calculs sous la forme d'une table d'analyse de la variance :

| Source | SC | ddl | MC | F |
|---|---|---|---|---|
| Contraste C | $SC_C$ | 1 | $MC_C=SC_C$ | $F_C = \dfrac{\mathrm{MC_C}}{\mathrm{MRC}}$ |
| Résidu R | SCR | $\nu$ | MCR | |

**Remarque 2 :** Le degré de liberté d'un contraste est toujours égal à 1 (deux groupes) on a donc $MC_c= SC_c$

Pour un plan équilibré on prendra la statistique de test : $F_C = \dfrac{\mathrm{MC_C}}{\mathrm{MRC}} = \dfrac{n(\hat{C})^2}{\mathrm{MRC}\sum_{i=1}^{k} c_i^2}$

Pour un plan non équilibré on prendra la statistique de test : $F_C = \dfrac{\mathrm{MC_C}}{\mathrm{MRC}} = \dfrac{(\hat{C})^2}{\mathrm{MRC}\sum_{i=1}^{k} \frac{c_i^2}{n_i}}$

**Proposition 4 :** Sous l'hypothèse nulle $H_0$ la statistique $F_C$ suit une loi de Fisher à $\nu_1=1$ et $\nu_2=\nu$=ddl(SCR) degrés de liberté.

**Décision statistique :**

Le test $H_0 : C=0$ contre $H_1 : C\neq 0$ de niveau $\alpha$ est défini par la règle de décision : si $f_C = F_{C\,obs}\geq\lambda_\alpha$ on rejette H0 avec $\lambda_\alpha$ donné par la table de Fisher à 1 et $\nu$ degrés de liberté : $P(\lambda_\alpha)=1-\alpha$

**Remarque 3 :** Les deux méthodes sont équivalentes car $F_C=T^2$

### VII-2-2 Contrastes indépendants Contrastes orthogonaux

**Définition 2 :** étant donnés $r$ contrastes $C^1, C^2, \ldots, C^r$ et $r$ nombres $\lambda_1, \lambda_2, \ldots, \lambda_r$ l'expression $\lambda_1 C^1 + \lambda_2 C^2 + \ldots + \lambda_r C^r$ est une <u>combinaison linéaire</u> des $r$ contrastes.

**Définition 3 :** un contraste est <u>(linéairement) indépendant</u> d'un ensemble de contrastes s'il ne peut s'écrire comme combinaison linéaire de ces contrastes.

**Proposition 5 :** Tout ensemble de contrastes portant sur $k$ moyennes contient au plus $k$-1 contrastes linéairement indépendants.

**Définition 4 :** Considérons deux contrastes

$$C^1 = \sum_{i=1}^{k} c_i^1 \mu_i \quad \text{avec} \quad \sum_{i=1}^{k} c_i^1 = 0 \quad \text{et} \quad C^2 = \sum_{i=1}^{k} c_i^2 \mu_i \quad \text{avec} \quad \sum_{i=1}^{k} c_i^2 = 0$$

Dans le cas d'un plan équilibré les deux contrastes $C^1$ et $C^2$ sont <u>orthogonaux</u> si et seulement

$$\sum_{i=1}^{k} c_i^1 c_i^2 = 0$$

Dans le cas d'un plan non équilibré les deux contrastes $C^1$ et $C^2$ sont <u>orthogonaux</u> si et seulement

$$\sum_{i=1}^{k} \frac{c_i^1 c_i^2}{n_i} = 0$$

**Proposition 6 :** Dans le cas d'un plan inter, si deux contrastes sont orthogonaux alors leurs estimateurs sont non corrélés. (Ce qui équivaut, dans le cas de variables gaussiennes, à l'indépendance).

**Proposition 7 :** Soit une comparaison multiple définie par $r$ contrastes $C^1, C^2, \ldots, C^r$ alors la somme des carrés associée à la comparaison est donnée par : $SC_{comparaison} = SC_{C^1} + SC_{C^2} + \ldots + SC_{C^r}$ et son degré de liberté est égal à $r$.

**Remarque 4 :** Dans le cas de contrastes orthogonaux, l'usage veut de fixer l'erreur de première espèce $\alpha$(EC).

### VII-2-3 Analyse de tendance et contrastes polynomiaux

Lorsque les modalités du facteur sont <u>ordonnées le long d'un continuum</u>, on peut modéliser l'effet du facteur sur la variable dépendante par la donnée d'une fonction $f$ telle que : E(VD)=$f$(facteur)

La fonction $f$ est caractérisée par une expression : $y=f(x)$. Les fonctions élémentaires sont les fonctions polynomiales : $f(x)=a_0+a_1x+a_2x^2+ \ldots +a_px^p$ $p$ est le degré du polynôme $a_0+a_1x+ \ldots +a_px^p$.

Ce polynôme est une somme de monômes de degré $q \leq p$.

$q$=1 : composante linéaire $a_1x$ $q$=2 : composante quadratique $a_2x^2$

$q$=3 : composante cubique $a_3x^3$ $q$=4 : composante d'ordre 4 $a_4x^4$

Les composantes sont appelées <u>tendance</u>.

Tester une forme particulière de liaison entre le facteur et la variable dépendante revient à tester la présence d'une composante d'un certain degré de la fonction $f$ . Ceci revient à tester certains contrastes.

Il existe des procédures très générales permettant de calculer les coefficients des contrastes associés aux différentes tendances. Dans le cas particulier où le plan est équilibré et où les modalités du facteur peuvent être considérées comme régulièrement espacées sur un continuum les coefficients du contraste se déduisent de l'écart entre les rangs des moyennes.

**Proposition 8 :** Etant donnés *k* point il existe un et un seul polynôme de degré *k*-1 passant par ces *k* points. Il suffit, pour un facteur à *k* modalités, de considérer les *k*-1 polynômes orthogonaux de degré ≤ *k*-1.
Tester une tendance revient à tester un contraste. Pour connaître la contribution des tendances dans l'effet globale on mesure son intensité par : $\frac{\mathrm{SC}_{\text{tendance}}}{\mathrm{SC}_{\text{facteur}}}$

**VII-2-4 Comparaisons non orthogonales**
Dans le cas de comparaisons multiples définies a priori par un ensemble de contrastes non orthogonaux on peut contrôler l'erreur de première espèce de l'ensemble α(EE) et ceci par différentes procédures :

**Proposition 9 :** Etant donnée une comparaison multiple définie à partir de *r* contrastes $C^1, C^2, \ldots, C^r$ On a les inégalités :

$$\text{Inégalité de Bonferroni : } \alpha(\mathrm{EE}) \leq \sum_{i=1}^{r} \alpha(\mathrm{EC}^{\mathrm{i}})$$

$$\text{Inégalité de Sidak : } \alpha(\mathrm{EE}) \leq 1 - \prod_{i=1}^{r}\left(1 - \alpha(\mathrm{EC}^{\mathrm{i}})\right)$$

Dans le cas où les erreurs de première espèce par contraste sont toutes égales à α(EC), les inégalités précédentes deviennent :

$$\text{Inégalité de Bonferroni : } \alpha(\mathrm{EE}) \leq r\alpha(\mathrm{EC})$$

$$\text{Inégalité de Sidak : } \alpha(\mathrm{EE}) \leq 1 - (1 - \alpha(\mathrm{EC}))^r$$

$$\text{Et de plus : } \alpha(\mathrm{EE}) \leq 1 - (1 - \alpha(\mathrm{EC}))^r \leq r\alpha(\mathrm{EC})$$

**Remarque 5 :** L'inégalité de Sidak est plus précise mais toutes les deux donnent une borne supérieure à α(EE) et permettent de répartir α(EE) sur chacun des contrastes de façon à contrôler α(EE) afin de ne pas dépasser le seuil de signification α.

**VII-2-4-1 Test de Bonferroni et test de Sidak**

A partir de l'inégalité de Bonferroni α(EE)≤*r*α(EC) il suffit de prendre $\alpha(\mathrm{EC}) \leq \frac{\alpha}{r}$ pour assurer α(EE)≤α. Donc il suffit de tester chaque contraste au niveau de signification : $\alpha' = \frac{\alpha}{r}$

Le test de Bonferroni est un test portant sur un contraste. La statistique de test est $T = \frac{\hat{C}\sqrt{n}}{\sqrt{\mathrm{MCR}\sum_{i=1}^{k} c_i^2}}$ pour un plan équilibré et $T = \frac{\hat{C}}{\sqrt{\mathrm{MCR}\sum_{i=1}^{k} \frac{c_i^2}{n_i}}}$ pour un plan non équilibré.

Sous l'hypothèse $H_0$ cette statistique est distribuée selon une loi de student à ν ddl(MCR) mais les valeurs seront lues dans la table de Bonferroni appropriée.

A partir de l'inégalité de Sidak α(EE)≤1-(1-α(EC))$^r$ il suffit de prendre $\alpha(\mathrm{EC}) \leq 1 - (1 - \alpha)^{\frac{1}{r}}$ pour assurer α(EE)≤α. Donc il suffit de tester chaque contraste au niveau de signification :
$\alpha' = 1 - (1 - \alpha)^{\frac{1}{r}}$

### VII-2-4-2 Le test d'Holm

Holm a proposé une alternative au test de Bonferroni pour augmenter la puissance de test tout en garantissant que $\alpha(EE) \leq \alpha$.

La statistique de test étant la même que pour le test de Bonferroni on calcule les $r$ statistiques $t^{(i)}$ que l'on range par ordre croissant de leur valeur absolue : $|t^{(1)}| \leq |t^{(2)}| \leq \ldots \leq |t^{(r)}|$

Le premier test de signification porte alors sur la significativité de $|t^{(r)}|$ pour un seuil $\alpha' = \dfrac{\alpha}{r}$ Si cette valeur est significative, on teste la significativité de la deuxième plus grande valeur $|t^{(r-1)}|$ pour un seuil $\alpha' = \dfrac{\alpha}{r-1}$. On réitère la procédure jusqu'à ce que le test donne un résultat non significatif.

### VII-2-4-3 Le test de Dunnett

Dunnett a proposé une alternative au test de Bonferroni dans le cas d'une comparaison de toutes les modalités du facteur à une modalité de contrôle ou témoin. Dans un plan équilibré cette variante est plus puissante que le test de Bonferroni.

On connaît la relation exacte qui lie $\alpha(EE)$ et $\alpha(EC)$. Dunnett a établi la table donnant les valeurs critiques $\lambda_{\alpha'}$ pour la statistique $T$ en fonction du nombre $k$ de moyennes, du degré de liberté $\nu$=ddl(MCR) et du seuil $\alpha$ choisi pour $\alpha(EE)$.

Notons $\bar{y}_0$ la moyenne du groupe contrôle et $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_{k-1}$ les $k$-1 moyennes des groupes expérimentaux. Les contrastes correspondants aux différentes comparaisons sont donnés sous la forme : $C^i : \mu_i - \mu_0$ et les tests de Dunnet sont de la forme : $H^i_0 : C^i = 0$ contre $H^i_0 : C^i \neq 0$

La statistique de test est donnée par : $T = \dfrac{|\bar{y}_i - \bar{y}_0|}{\sqrt{\dfrac{2\text{MCR}}{n}}}$

Si $t_{obs} \geq \lambda_{\alpha'}$ lu dans la table de Dunnett pour le degré de liberté $\nu$=ddl(MCR) et $k$ alors $H^i_0$ est rejetée

Une façon équivalente de réaliser ce test consiste à calculer la valeur critique $\lambda_c = \lambda_{\alpha'}\sqrt{\dfrac{2\text{MCR}}{n}}$ et comparer chaque différence $|\bar{y}_i - \bar{y}_0|$ à cette valeur.

# Table des matières