Chapitre I
Modélisation statistique

El Methni M.

I-1. Modèle statistique

Les données sont constituées d'une famille de N valeurs $y_1, y_2, \ldots y_N$ 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-2. Décomposition de la variance associée à un facteur : (rappel)

Dans toute la suite *Y* désigne la variable dépendante et *G* un facteur à *r* modalités. Les *N* observations sont réparties en *r* classes $g_1, g_2, \ldots g_r$ contenant respectivement $n_1, n_2, \ldots n_r$ observations. (On a donc $N = n_1 + n_2, + \ldots + n_r$). Notons $y_{s(i)}$ l'observation associée au sujet *s* dans la classe 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	$N = \sum_{i=1}^{r} n_i$
Total	T_1	T_2	…	T_i	…	T_r	$T = \sum_{i=1}^{r} T_i$
Moyenne	$\bar{y}_1$	$\bar{y}_2$	…	$\bar{y}_i$	…	$\bar{y}_r$	$\bar{y} = \frac{1}{N}\sum_{i=1}^{r} n_i \bar{y}_i$

Les observations de la variable dépendante *Y* peuvent différer sous l'effet de deux sources de variations :

- La variation à l'intérieur des classes (ou dans les classes) : <u>variation intra</u>
- La variation entre les classes : <u>variation inter</u>.

Le facteur étant *constant* à l'intérieur des classes, seule la variation inter peut être imputée à l'effet du facteur.

La variation intra est une variation *individuelle* correspondant à la variation des observations dans une même classe.

La variation inter est une variation *systématique* correspondant à la variation des classes. Elle exprime l'effet (éventuel) du facteur (VI).

La variation des observations sera mesurée en termes de dispersion autour de la moyenne à l'aide de la variance (ou de la somme des carrées des écarts à la moyenne).

La <u>moyenne générale</u> des observations est donnée par :

$$\bar{y} = \frac{1}{N}\sum_{s}\sum_{i} y_{s(i)} \quad \bar{y} = \frac{\mathrm{T}}{N} \quad \bar{y} = \frac{1}{N}(n_1\bar{y}_1 + n_2\bar{y}_2 + \cdots + n_r\bar{y}_r)$$

La <u>variation **T**otale</u> des observations est mesurée par la **S**omme des **C**arrés des écarts entre chacune des *N* observations et la moyenne générale :

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

En notant $\bar{y}_i$ la moyenne des observations de la classe g_i on a : $\bar{y}_i = \frac{1}{n_i}\sum_{s=1}^{n_i} y_{s(i)}$ $\bar{y}_i = \frac{T_i}{n_i}$

La variation à l'**I**ntérieur de la classe g_i est mesurée par la **S**omme des **C**arrés des écarts entre chaque observation de la classe et la moyenne de cette classe :

$$\mathbf{SCIntra_i} = \sum_{s=1}^{n_i} (y_{s(i)} - \overline{y}_i)^2 = n_i \times Variance(g_i)$$

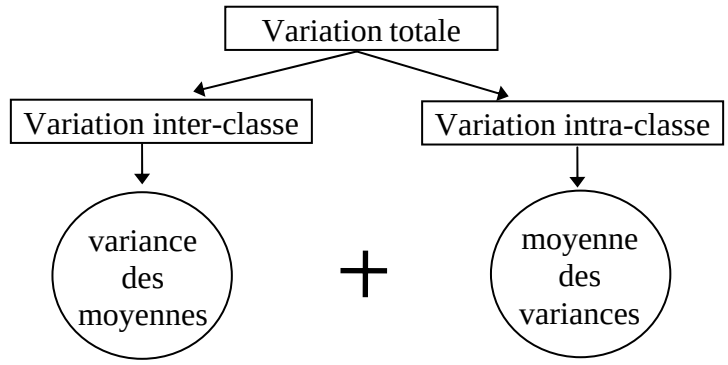
La variation à l'intérieur des classes est définie par la somme, pour toutes les classes, des sommes des carrés intra : $\mathbf{SCIntra} = \sum_{i=1}^{r} \mathbf{SCIntra_i}$

Pour calculer la variation-inter on se place dans le cas de la situation fictive où toutes les observations d'une même classe sont égales à la moyenne $\overline{y}_i$ de cette classe et on calcule la **S**omme des **C**arrés **I**nter : $\mathbf{SCInter} = \sum_{i=1}^{r} n_i (\overline{y}_i - \overline{y})^2 = N \times \text{Variance}(\overline{y}_i)$

On démontre la relation fondamentale suivante : **SCT** = **SCI**nter + **SCI**ntra

Remarque : En divisant par l'effectif total *N* on a une relation analogue entre les variances :

Variance totale = Var Inter + Var Intra



Remarque : On retrouve l'indice de liaison PRE, appelé rapport de corrélation (cf. cours de L1) en considérant le rapport : $\eta^2_{Y/G} = \dfrac{\text{Variation Inter}}{\text{Variation Totale}}$

$\eta^2_{Y/G}$ mesure la liaison entre la variable réponse *Y* et le facteur G :

- $\eta^2_{Y/G} = 0$ indique une absence de liaison (la réponse du sujet ne dépend pas du groupe auquel il appartient).
- $\eta^2_{Y/G} = 1$ indique une liaison parfaite (la réponse du sujet est entièrement déterminée par son appartenance à un groupe).

Ainsi $\eta^2_{Y/G}$ mesure l'intensité de l'effet de G sur la réponse *Y*.

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 χ_m^2 et χ_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 $\frac{\frac{1}{m\sigma^2}\sum_{i=1}^{m} U_i^2}{\frac{1}{n\sigma^2}\sum_{i=1}^{n} V_i^2} = \frac{\frac{1}{m}\sum_{i=1}^{m} U_i^2}{\frac{1}{n}\sum_{i=1}^{n} V_i^2}$ suit une loi de Fisher-Snedecor $\mathcal{F}(m, n)$ 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 $\frac{\frac{1}{m-1}\sum_{i=1}^{m}(U_i - \bar{U})^2}{\frac{1}{n-1}\sum_{i=1}^{n}(V_i - \bar{V})^2}$ suit une loi de Fisher-Snedecor $\mathcal{F}(m\text{-}1, n\text{-}1)$.

$\bar{U} = \frac{\sum_{i=1}^{m} U_i}{m}$ et $\bar{V} = \frac{\sum_{i=1}^{n} V_i}{n}$ Sont les estimateurs respectifs de μ_1 et μ_2

I-2-1 Exemple :

Pour étudier l'influence du facteur « intensité du bruit environnant » sur la capacité d'un sujet à résoudre un problème, l'expérimentateur construit l'expérience suivante : 24 écoliers sont répartis de façon aléatoire dans quatre pièces. Des bruits de la rue ont été enregistrés et sont diffusés dans chaque pièce avec un niveau sonore particulier. Les enfants doivent résoudre une série de problèmes. La variable réponse est la note finale obtenue à la série d'épreuves.

	Niveau sonore				
	1	2	3	4	
	62 60 63 59	56 62 60 61 63 64 63 59	63 67 71 64 65 66	68 66 71 67 68 68	
n_j	4	8	6	6	$\boldsymbol{N}$ = 24
$\overline{y}_i$	61	61	66	68	$\overline{y}$ = 64
Variance	10/4	48/8	40/6	14/6	SCIntra= **SCR** = 112
$n_i\left(\overline{y}_i-\overline{y}\right)^2$	36	72	24	96	SCInter= $\sum_{i=1}^{r} n_i\left(\overline{y}_i-\overline{y}\right)^2$ =228

SCT = N×variance totale = 340 **SCI**ntra = 228 **SCI**nter= 112 $\eta^2_{Y/G} = 0,6706$