Chp 2/ Plan à un facteur à plusieurs niveaux

I/ Introduction.

Dans ce chapitre, nous allons étudier un plan à un seul facteur X, et X à n niveau notés $x_1, \dots, x_k$.
Pour chaque niveau x_i de X, n_i essais sont realisés la matrice des données est alors donnée sous la forme

niveau / essai	x_1	x_2	---	x_k
1	y_{11}	y_{21}		y_{k1}
.	⋮	⋮		⋮
n_i	y_{1n_1}	y_{2n_2}		y_{kn_k}

n_i pour $i=\overline{1,k}$ = n^{bre} d'essais au niveau x_i.
y_{ij} est la reponse au niveau x_i pour le $j^{ème}$ essai
On note par $n = \sum_{i=1}^{k} n_i$ la ~~taille~~ taille globale de l'echantillon reponse.

Le problème est de verifier si le facteur a un effet sur le phenomène étudié.
L'étude consiste alors de comparer les moyennes des differents niveaux du facteur X.
Le modèle mathematique : $y_{ij} = \mu_i + \varepsilon_{ij}$ $i=\overline{1,k}$ $j=\overline{1,n_i}$
y_{ij} est la reponse du $i^{ème}$ niveau au $j^{ème}$ essai
μ_i est la moyenne du $i^{ème}$ niveau
ε_{ij} est une erreur aleatoire
Rq: En pratique $(\varepsilon_{ij})_{i,j}$ est generalement supposée être independantes, de moyenne nulle et de variance $\sigma^2 < \infty$
et $\varepsilon_{ij} \sim N(0, \sigma)$

II/ Estimation des paramètres du modèle

Pour estimer les paramètres μ_i, $i=\overline{1,k}$ (moyenne), on utilise le critère des moindres carrés (CMC) qui consiste à minimiser la somme des erreurs quadratiques

$$\varepsilon_{ij} = y_{ij} - \mu_i$$

Pb min Δ avec

$$\Delta = \sum_{i=1}^{k}\sum_{j=1}^{n_i} \varepsilon_{ij}^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2$$

On designe par $\alpha_i = \mu_i - \mu$ l'ecart existant entre la moyenne au $i^{ème}$ niveau et et la moyenne globale $\mu = \frac{1}{n}\sum_{i=1}^{k} n_i \mu_i$

α_i est l'effet du facteur au $i^{ème}$ niveau sur la réponse y.

Ainsi $\Delta = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \alpha_i - \mu)^2$

II. 1/ Estimation de la moyenne globale μ

$$\frac{\partial\Delta}{\partial\mu} = 0 \iff -2\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \alpha_i - \mu) = 0$$

$$\iff \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} - \sum_{i=1}^{k} n_i\alpha_i - \mu\sum_{i=1}^{k} n_i = 0$$

On a: $\sum_{i=1}^{k} n_i\alpha_i = \sum_{i=1}^{k} n_i(\mu_i - \mu) = 0$

D'où $\boxed{\hat{\mu} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} = \bar{y}_{..} = \bar{y}}$

II. 2/ Estimation des effets α_i

$$\frac{\partial\Delta}{\partial\alpha_i} = 0 \iff -2\sum_{j=1}^{n_i}(y_{ij} - \alpha_i - \mu) = 0 \text{ pour le } i^{ème} \text{ niveau } x$$

$$\iff \sum_{j=1}^{n_i} y_{ij} - n_i\alpha_i - n_i\mu = 0$$

$$\iff \boxed{\hat{\alpha}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij} - \hat{\mu} = \bar{y}_{i.} - \bar{y}}$$

On a $\mu_i = \alpha_i + \mu \Rightarrow \boxed{\hat{\mu}_i = \hat{\alpha}_i + \hat{\mu} = \bar{y}_{i.}}$

II. 3 / Estimation de la variance σ^2

La variance est estimée par la méthode du maximum de vraisemblance sous l'hypothèse $\varepsilon_{ij} \sim N(0, \sigma^2)$

Donc $y_{ij} \sim N(\mu_i, \sigma)$

La fonction de vraisemblance L est donnée par

$$L(y, \sigma) = \prod_{i=1}^{k} \prod_{j=1}^{n_i} f_y(y_{ij}) \qquad y = (y_{1,1}, \ldots, y_{k n_k})$$

$$\ln L(y,\sigma) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \ln f_y(y_{ij})$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{n_i} \ln \left[\frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{y_{ij}-\mu_i}{\sigma}\right)^2} \right]$$

$$= -n \ln(2\pi\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2$$

$$\frac{\partial \ln(L(y,\sigma))}{\partial \sigma} = 0 \iff -n\sigma^{-1} + \frac{1}{\sigma^3} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2 = 0$$

$$\iff -n\sigma^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2 = 0$$

$$\boxed{\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \hat{\mu}_i)^2}$$

III / Loi des estimateurs

III - 1 / Loi $\hat{\mu} = \bar{y}$

$$y_{ij} \sim N(\mu_i, \sigma) \Rightarrow \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij} \sim N(n\mu, \sqrt{n}\,\sigma)$$

$$\Rightarrow \bar{y} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \quad E(\bar{y}) = \mu \quad \bar{y} \text{ sans biais}$$

III . 2 / Loi de $\hat{\mu}_i = \bar{y}_{i.}$

$$y_{ij} \sim N(\mu_i, \sigma) \qquad \sum_{j=1}^{n_i} y_{ij} \sim N(n_i\mu_i, \sqrt{n_i}\,\sigma)$$

d'où $\bar{y}_{i.} \sim N\left(\mu_i ; \frac{\sigma}{\sqrt{n_i}}\right)$ $E(\bar{y}_{i.}) = \mu_i$ $\bar{y}_{i.}$ sans biais

III . 3 / Loi de $\hat{\sigma}^2$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$$

on a $\frac{n\hat{\sigma}^2}{\sigma^2} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\frac{y_{ij} - \bar{y}_{i.}}{\sigma}\right)^2 \sim \chi^2_{n-k}$

$$E\left(\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left[\frac{y_{ij} - \bar{y}_{i.}}{\sigma}\right]^2\right) = n-k$$

$\Rightarrow \frac{1}{n-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$ est sans biais

IV / Validation du modèle.

On utilise l'analyse de la variance qui consiste à

Test : H_0 : " $\mu_1 = \cdots = \mu_k$ " vs H_1 : " $\exists i \neq j \, / \, \mu_i \neq \mu_j$ "

Test de pas d'effet du facteur contre il y a un effet ce qui revient

Test : H_0 : " $\alpha_i = 0 \quad i = \overline{1,k}$ " vs H_1 : " $\exists i \quad \mu_i \neq 0$ "

IV.1 / Equation de l'analyse de la variance.

On a

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot}+\bar{y}_{i\cdot}-\bar{y})^2$$

$$= \sum_{i}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_{i\cdot}-\bar{y})^2 + 2\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})(\bar{y}_{i\cdot}-\bar{y})$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2 + \sum_{i=1}^{k} n_i(\bar{y}_{i\cdot}-\bar{y})^2 + 2\sum_{i=1}^{k} n_i(\bar{y}_{i\cdot}-\bar{y})\underbrace{\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})}_{=0}$$

D'où

$$\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2 = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2 + \frac{1}{n}\sum_{i=1}^{k} n_i(\bar{y}_{i\cdot}-\bar{y})^2$$

$$S_T^2 \quad = \quad S_R^2 \quad + \quad S_X^2$$

variance totale = variance résiduelle + variance dûe au facteur

IV.2 Calcul de $E(S_T^2)$, $E(S_X^2)$ et $E(S_R^2)$

• On a $\displaystyle\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\frac{y_{ij}-\bar{y}}{\sigma}\right)^2 \sim \chi^2_{n-1}$

$$E\left(\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\frac{y_{ij}-\bar{y}}{\sigma}\right)^2\right) = \frac{1}{n}(n-1)$$

$$\Rightarrow \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2 = \frac{(n-1)\sigma^2}{n} = E(S_T^2)$$

Et

$$\frac{n S_T^2}{\sigma^2} \sim \chi^2_{n-1}$$

• $$E(S_X^2) = \frac{1}{n}\sum_{i=1}^{k} n_i \, E(\bar{y}_{i\cdot}-\bar{y})^2$$

$$\sum_{i=1}^{k} n_i\left(\frac{\bar{y}_{i\cdot}-\bar{y}}{\sigma}\right)^2 \sim \chi^2_{k-1}$$

$$E(S_X^2) = \frac{1}{n}\;\frac{k-1}{\sigma^2}$$

$$\bullet \quad \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\frac{y_{ij}-\bar{y}_{i\circ}}{\sigma}\right)^2 \sim \chi^2_{n-k}$$

$$E(S_R^2) = \frac{1}{n}\sigma^2 E\left(\chi^2_{n-k}\right) = \frac{n-k}{n}\sigma^2$$

$$\text{Et } E(S_X^2) = E(S_T^2) - E(S_R^2)$$

$$= \frac{(n-1)}{n}\sigma^2 - \frac{n-k}{n}\sigma^2 = \frac{\sigma^2}{n}(k-1)$$

Rq : Sous H_0 $S_R^2 \perp S_X^2$

et donc
$$\frac{\sum_{i=1}^{k} n_i \left(\frac{\bar{y}_{i\circ}-\bar{y}_{\circ\circ}}{\sigma}\right)^2 \Big/ (k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\frac{y_{ij}-\bar{y}_{i\circ}}{\sigma}\right)^2 \Big/ (n-k)} = \frac{S_X^2/(k-1)}{S_R^2/(n-k)} \sim F(k-1; n-k)$$

IV.3 / Tableau d'ANOVA

Source de variation	d°L	CM	F_{exp}	F_{Tab}
Variance due au facteur X S_X^2	$k-1$	$S_X^2/(k-1)$	$\dfrac{S_X^2/(k-1)}{S_R^2/(n-k)}$	$k\alpha$ lu sur la table $F_\alpha(k-1, n-k)$
Variance residuelle S_R^2	$n-k$	$S_R^2/(n-k)$		
Variance totale S_T^2	$n-1$			

$$RC = \left[\frac{\frac{S_X^2}{(k-1)}}{\frac{S_R^2}{(n-k)}} > k\alpha\right]$$ α un seuil fixé.