

**TD DE MODÈLES LINÉAIRES - FEUILLE 8**  
**Analyse de variance - partie 1**

**EXERCICE 1** On dispose de  $k$   $n$ -échantillons  $(Y_{i1}, \dots, Y_{in})$ ,  $i = 1, \dots, k$ , les  $n$ -échantillons étant indépendants les uns des autres. Pour l'échantillon  $i = 1, \dots, k$ , les variables  $Y_{i1}, \dots, Y_{in} \sim \mathcal{N}(m_i, \sigma^2)$ . On veut tester l'homogénéité des moyennes :

$$H_0 : m_1 = \dots = m_k \quad \text{contre} \quad H_1 : \exists i \neq j, \quad m_i \neq m_j.$$

On utilise les notations suivantes :

- pour la moyenne empirique :  $\bar{Y} = \frac{1}{nk} \sum_{i=1}^k \sum_{j=1}^n Y_{ij}$ ,
- dans l'échantillon  $i = 1, \dots, k$ , la moyenne empirique des variables est notée  $\bar{Y}_i = \frac{1}{n} \sum_{j=1}^n Y_{ij}$ ,
- pour la variabilité totale de l'échantillon :  $(nk - 1)S^2 = \sum_{i=1}^k \sum_{j=1}^n (Y_{ij} - \bar{Y})^2$ .
- la variabilité intra-groupe est  $\sum_{i=1}^k \sum_{j=1}^n (Y_{ij} - \bar{Y}_i)^2$ .
- la variabilité inter-groupe est  $n \sum_{i=1}^k (\bar{Y}_i - \bar{Y})^2$ .

1. Montrer que la variabilité de l'échantillon s'écrit comme la somme des variabilités intra et inter-groupes.
2. Considérons les vecteurs

$$\begin{aligned} Y &= (Y_{11}, \dots, Y_{1n}, \dots, Y_{k1}, \dots, Y_{kn}), \\ \Upsilon &= (\bar{Y}_1, \dots, \bar{Y}_1, \dots, \bar{Y}_k, \dots, \bar{Y}_k). \end{aligned}$$

Montrer que  $\Upsilon$  est la projection orthogonale de  $Y$  sur le sous espace vectoriel  $E$  (de  $\mathbb{R}^{nk}$ ) de dimension  $k$  engendré par les vecteurs indicateurs de bloc

$$v_1 = (1, \dots, 1, 0, \dots, 0, 0, \dots, 0)^T, \quad \dots, \quad v_k = (0, \dots, 0, 0, \dots, 0, 1, \dots, 1)^T.$$

3. Soit la statistique  $Z = \frac{n \sum_{i=1}^k (\bar{Y}_i - \bar{Y})^2}{\sum_{i=1}^k \sum_{j=1}^n (Y_{ij} - \bar{Y}_i)^2}$ . Démontrer que sous l'hypothèse  $H_0$ , la v.a.

$\frac{nk - k}{k - 1} Z$  suit une loi de Fisher  $\mathcal{F}(k - 1, nk - k)$ .

Indication : on pourra utiliser le théorème des 3 perpendiculaires.

4. Application numérique : On a relevé les scores des étudiants de 4 écoles à un concours. Comparer les performances des écoles. Les différences observées sont-elles significatives au risque 5 % ? Comparer deux à deux les échantillons.

$E_1$	$E_2$	$E_3$	$E_4$
73	84	69	65
57	95	80	58
95	96	73	82
78	62	62	86
86	80	50	35
61	87	71	52
80	100	84	70
98	74	66	79
64	85	52	43
78	77	73	60

$$\bar{Y}_1 = 77, \bar{Y}_2 = 84, \bar{Y}_3 = 68, \bar{Y}_4 = 63.$$

**EXERCICE 2** Considérons le modèle

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, 1 \leq i \leq I, 1 \leq j \leq n_i,$$

où les  $\epsilon_{ij}$  sont indépendantes de loi  $\mathcal{N}(0, \sigma^2)$ ,  $\sum_{i=1}^I n_i \alpha_i = 0$  et  $\sum_{i=1}^I n_i = n$ .

1. Montrer que  $Y = A\theta + \varepsilon$ , où  $A = [\mathbf{1}_n a_1 \dots a_I]$  avec  $\mathbf{1}_n, a_1, \dots, a_I$ , éléments de  $\mathbb{R}^{IJ}$  que l'on précisera.
2. Soit  $F_0 = \{\mu \mathbf{1}_n, \mu \in \mathbb{R}\}$ ,  $F_1 = \left\{ \sum_{i=1}^I \alpha_i a_i, \alpha_i \in \mathbb{R}, \sum_{i=1}^I n_i \alpha_i = 0 \right\}$ .  
Montrer que  $F_0$  et  $F_1$  sont des sous-espaces vectoriels deux-à-deux orthogonaux. En déduire qu'il existe un espace  $G$  tel que

$$\mathbb{R}^{IJ} = F_0 \oplus F_1 \oplus G.$$

3. Calculer  $P_0, P_0 + P_1$  ( $P_0$  étant la projection orthogonale sur  $F_0$ ). En déduire  $P_1$ . Calculer  $P_G$ .
4. Construire la table d'analyse de la variance.
5. Tester l'hypothèse  $H_0$  : "tous les  $\alpha_i$  sont nuls" contre  $H_1$  : "tous les  $\alpha_i$  ne sont pas nuls".
6. **Application.** Les mesures de teneur en octane sur des échantillons de carburant prélevés dans quatre régions du nord-est des États Unis durant l'été 1953, sont reproduites dans le tableau ci-après.

Notant  $Y_{ij}$  la jème mesure effectuée dans la région  $i$ , on donne les quantités suivantes :

Région	A	B	C	D
$n_i$	16	13	18	22
$\bar{Y}_i$	83.875	82.846	83.22	83.009
$\sum_j (Y_{ij} - \bar{Y}_i)^2$	17.81	2.67	28.97	33.04

Peut-on conclure que la teneur en octane est différente suivant les régions ?

Région A	Région B	Région C	Région D
84.0	82.4	83.2	80.2
83.5	82.4	82.8	82.9
84.0	83.4	83.4	84.6
85.0	83.3	80.2	84.2
83.1	83.1	82.7	82.8
83.5	83.3	83.0	83.0
81.7	82.4	85.0	82.9
85.4	83.3	83.0	83.4
84.1	82.6	85.0	83.1
83.0	82.0	83.7	83.5
85.8	83.2	83.6	83.6
84.0	83.1	83.3	86.7
84.2	82.5	83.8	82.6
82.2		85.1	82.4
83.6		83.1	83.4
84.9		84.2	82.7
		80.6	82.9
		82.3	83.7
			81.5
			81.9
			81.7
			82.5