

## TD DE MODÈLES LINÉAIRES I - Série 6

### Méthodes d'analyse de la variance.

#### Partie 1

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

$$\mathbf{H}_0 : m_1 = \dots = m_k \quad \text{contre} \quad \mathbf{H}_1 : \exists i, j, m_i \neq m_j.$$

On utilise les notations suivantes :

- pour la moyenne empirique :  $\bar{X} = \frac{1}{nk} \sum_{i=1}^k \sum_{j=1}^n X_{ij}$ ,

- dans l'échantillon  $i = 1, \dots, k$ , la moyenne empirique des variables est notée  $\bar{X}_i = \frac{1}{n} \sum_{j=1}^n X_{ij}$ ,

- pour la variabilité totale de l'échantillon :  $(nk - 1)S^2 = \sum_{i=1}^k \sum_{j=1}^n (X_{ij} - \bar{X})^2$ .

La variabilité intra-groupe est  $\sum_{i=1}^k \sum_{j=1}^n (X_{ij} - \bar{X}_i)^2$  et la variabilité inter-groupe est  $n \sum_{i=1}^k (\bar{X}_i - \bar{X})^2$ .

**Question 1.** Montrer que la variabilité de l'échantillon s'écrit comme la somme des variabilités intra et inter-groupes.

**Question 2.** Considérons les vecteurs

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

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

$$V_1 = (1, \dots, 1, 0, \dots, 0, 0, \dots, 0),$$

...

$$V_k = (0, \dots, 0, 0, \dots, 0, 1, \dots, 1).$$

**Question 3.** Soit la statistique  $Z = \frac{n \sum_{i=1}^k (\bar{X}_i - \bar{X})^2}{\sum_{i=1}^k \sum_{j=1}^n (X_{ij} - \bar{X}_i)^2}$ .

Démontrer que sous l'hypothèse  $H_0$ , la v.a.  $\frac{nk - k}{k - 1} Z$  suit une loi de Fisher  $F(k - 1, nk - k)$ . On pensera à utiliser le théorème des 3 perpendiculaires et à conclure avec Cochran.

**Question 4. Application numérique.** On a relevé les scores des étudiants de 4 écoles un concours. Comparer les performances des écoles.

Est-ce que les différences observées sont 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

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

## Partie 2 - Cadre général de l'analyse de la variance

Soit  $X = m + Y$  où  $m \in \mathbb{R}^N$ ,  $m = (m_1, \dots, m_N)$  et  $Y$  un échantillon de  $\mathcal{N}(0, \sigma^2)$ .  
Soit  $E$  un s.e.v. de  $\mathbb{R}^N$  de dimension  $k$ ,  $k \leq N$ , tel que  $m \in E$ . Soit  $H$  un sous espace vectoriel de  $E$  avec  $\dim(H) = r (\leq k)$ .

On veut tester (problème général de l'analyse de la variance) :

$$\mathbf{H}_0 : m \in H \quad \text{contre} \quad \mathbf{H}_1 : m \notin H.$$

**Question 1.** Montrer que la partie 1 est un cas particulier.

**Théorème 1** Soient  $X_E$  la projection orthogonale de  $X$  sur  $E$  et  $X_H$  la projection orthogonale de  $X$  sur  $H$ .

$$- \text{La v.a. } Z = \frac{\frac{\|X_E - X_H\|^2}{k-r}}{\frac{\|X - X_E\|^2}{N-k}} = \frac{N-k}{k-r} \frac{\|X_E - X_H\|^2}{\|X - X_E\|^2} \text{ suit une loi de Fisher décentrée}$$

$$F\left(k-r, N-k, \frac{\|m - m_H\|^2}{\sigma^2}\right) \text{ où } m_H \text{ est la projection orthogonale de } m \text{ sur } H.$$

- Sous  $H_0$ ,  $Z \sim F(k-r, N-k)$ .

**Corollaire (Test de l'analyse de la variance)** Etant donné le risque d'erreur de première espèce  $\alpha$ , on rejette  $\mathbf{H}_0$  au profit de  $\mathbf{H}_1$  si  $Z > F(k-r, N-k; \alpha)$ , où  $F(k-r, N-k; \alpha)$  est le quantile supérieur d'ordre  $\alpha$  d'une loi de Fisher  $F(k-r, N-k)$ .

**Question 2.** Démontrer le théorème.

**Analyse de la variance à deux facteurs.**

Le problème qui se pose fréquemment en agronomie est l'utilisation de certains engrais suivant la nature du terrain. Par exemple 5 engrais A, B, C, D, E peuvent être utilisés sur 4 natures de sols 1, 2, 3, 4. On dispose de 4 champs correspondant à ces 4 compositions respectives. Chaque champs a été subdivisé en 5 parcelles égales sur lesquelles on a affecté les engrais A, B, C, D, E (par tirage au sort pour diminuer les erreurs systématiques). Les rendements en blé dépendent alors de deux facteurs : nature du sol et type d'engrais. On a observé les résultats suivants

	A	B	C	D	E
1	310	353	366	299	367
2	284	293	335	264	314
3	307	306	339	311	377
4	267	308	312	266	342

On veut pouvoir étudier si les résultats obtenus sont équivalents (i.e. si les différences observées sont dues au hasard) où si l'influence d'un engrais ou d'une nature de sol est prépondérante. On peut formaliser le problème de la manière suivante.

Soit  $X_{ij}$  le rendement du sol  $i$  muni de l'engrais  $j$ . On suppose que  $X_{ij}$  suit la loi  $\mathcal{N}(m_{ij}, \sigma^2)$  avec  $m_{ij}$  de la forme

$$m_{ij} = m + \alpha_i + \beta_j,$$

avec  $\sum_{i=1}^k \alpha_i = 0$  et  $\sum_{j=1}^n \beta_j = 0$ . Ceci revient à dire que  $\frac{1}{k} \sum_{i=1}^k m_{ij} = m + \beta_j$  et  $\frac{1}{n} \sum_{j=1}^n m_{ij} = m + \alpha_i$ .

*Concrètement.*  $\alpha_i$  et  $\beta_j$  traduisent les effets respectifs des deux facteurs  $i$  et  $j$  sur la moyenne  $m_{ij}$  de la v.a.  $X_{ij}$ .

*Problèmes.* Si on veut tester si la nature des sols n'a pas d'influence sur le rendement, on testera l'hypothèse  $H_0 =$  "tous les  $\alpha_i$  sont nuls" contre  $H_1 =$  "les  $\alpha_i$  ne sont pas tous nuls", c'est le premier test.

Pour tester l'influence des engrais, on prendra l'hypothèse  $H'_0 =$  "tous les  $\beta_j$  sont nuls" contre  $H'_1 =$  "les  $\beta_j$  ne sont pas tous nuls".

*Modélisation.* On note  $X = (X_{11}, X_{12}, X_{13}, \dots, X_{21}, X_{22}, \dots, X_{k1}, \dots, X_{kn})' \in \mathbb{R}^{kn}$  et  $M = (m_{11}, m_{12}, \dots, m_{1n}, m_{21}, m_{22}, \dots, m_{k1}, \dots, m_{kn})'$ .

**Question 3-a.** Expliciter le vecteur des moyennes  $M$ , donner la dimension du sous-espace vectoriel défini par ce vecteur. Définir  $H$  et donner sa dimension lorsqu'on teste  $H_0$  contre  $H_1$ .

Pour appliquer le test de l'analyse de la variance pour tester  $H_0$  contre  $H_1$  il faut donc calculer

$$\| X - X_E \|^2 \text{ et } \| X_E - X_H \|^2,$$

i.e. déterminer  $X_E$  et  $X_H$ . Pour cela, on posera, pour  $i = 1, \dots, k$  et  $j = 1, \dots, n$  :

$$\bar{X}_{i.} = \frac{1}{n} \sum_{j=1}^n X_{ij}, \quad \bar{X}_{.j} = \frac{1}{k} \sum_{i=1}^k X_{ij}, \quad \bar{X} = \frac{1}{nk} \sum_{i=1}^k \sum_{j=1}^n X_{ij}.$$

**Question 3-b.** Démontrer les égalités suivantes, pour  $i = 1, \dots, k$  et  $j = 1, \dots, n$  :

$$(X_E)_{ij} = \bar{X} + (\bar{X}_{i.} - \bar{X}) + (\bar{X}_{.j} - \bar{X}),$$

et

$$(X_H)_{ij} = \bar{X} + (\bar{X}_{.j} - \bar{X}).$$

**Question 3-c.** Démontrer que

$$Z = \frac{\frac{\|X_E - X_H\|^2}{\dim E - \dim H}}{\frac{\|X - X_E\|^2}{\dim \mathbb{R}^N - \dim E}} = \frac{n(n-1) \sum_{i=1}^k (\bar{X}_{i.} - \bar{X})^2}{\sum_{i=1}^k \sum_{j=1}^n (X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X})^2}.$$

**Question 3-d.** Application numérique.

**Question 3-e.** Faire la même chose pour tester  $H'_0$  contre  $H'_1$ .