TD DE MODÈLES LINÉAIRES I - Série 6

Méthodes d'analyse de la variance.

Partie 1

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

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

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 $i = 1, ..., k$, 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

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

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

$$V_1 = (1, ..., 1, 0, ..., 0, 0, ..., 0),$$

$$...$$

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

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

Démontrer que sous l'hypothèse H_0, la v.a. $\dfrac{nk-k}{k-1} Z$ 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 $m \in \mathbb{R}^N$, $m = (m_1, \ldots, m_N)$ 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) :

$$\mathbf{H_0} : m \in H \qquad \text{contre} \qquad \mathbf{H_1} : m \notin H.$$

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.*

- *La v.a.* $Z = \dfrac{\frac{\|X_E - X_H\|^2}{k-r}}{\frac{\|X - X_E\|^2}{N-k}} = \frac{N-k}{k-r} \dfrac{\|X_E - X_H\|^2}{\|X - X_E\|^2}$ *suit une loi de Fischer décentrée* $F\left(k-r, N-k, \frac{\|m - m_H\|^2}{\sigma^2}\right)$ *où m_H est la projection orthogonale de m 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 α, 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 α 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. α_i et β_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 α_i sont nuls" contre H_1 = "les α_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 β_j sont nuls" contre H_1' = "les β_j ne sont pas tous nuls".

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

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 $i = 1, \ldots, k$ et $j = 1, \ldots, n$:

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

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

$$\begin{aligned} (X_E)_{ij} &= \bar{X} + \Big(\bar{X}_{i.} - \bar{X}\Big) + \Big(\bar{X}_{.j} - \bar{X}\Big), \\ \text{et } (X_H)_{ij} &= \bar{X} + \Big(\bar{X}_{.j} - \bar{X}\Big). \end{aligned}$$

Question 3-c. Démontrer que

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

Question 3-d. Application numérique.

Question 3-e. Faire la même chose pour tester H_0' contre H_1'.