Mise à niveau

ANOVA

A. Godichon-Baggioni

I. Analyse de la variance à 1 facteur

Introduction

L'analyse de la variance (ANOVA) désigne un ensemble de techniques statistiques permettant d'apprécier l'effet

- d'une ou plusieurs variables qualitatives, appelées facteurs
- sur une variable quantitative.

L'ANOVA à un facteur, c'est l'étude

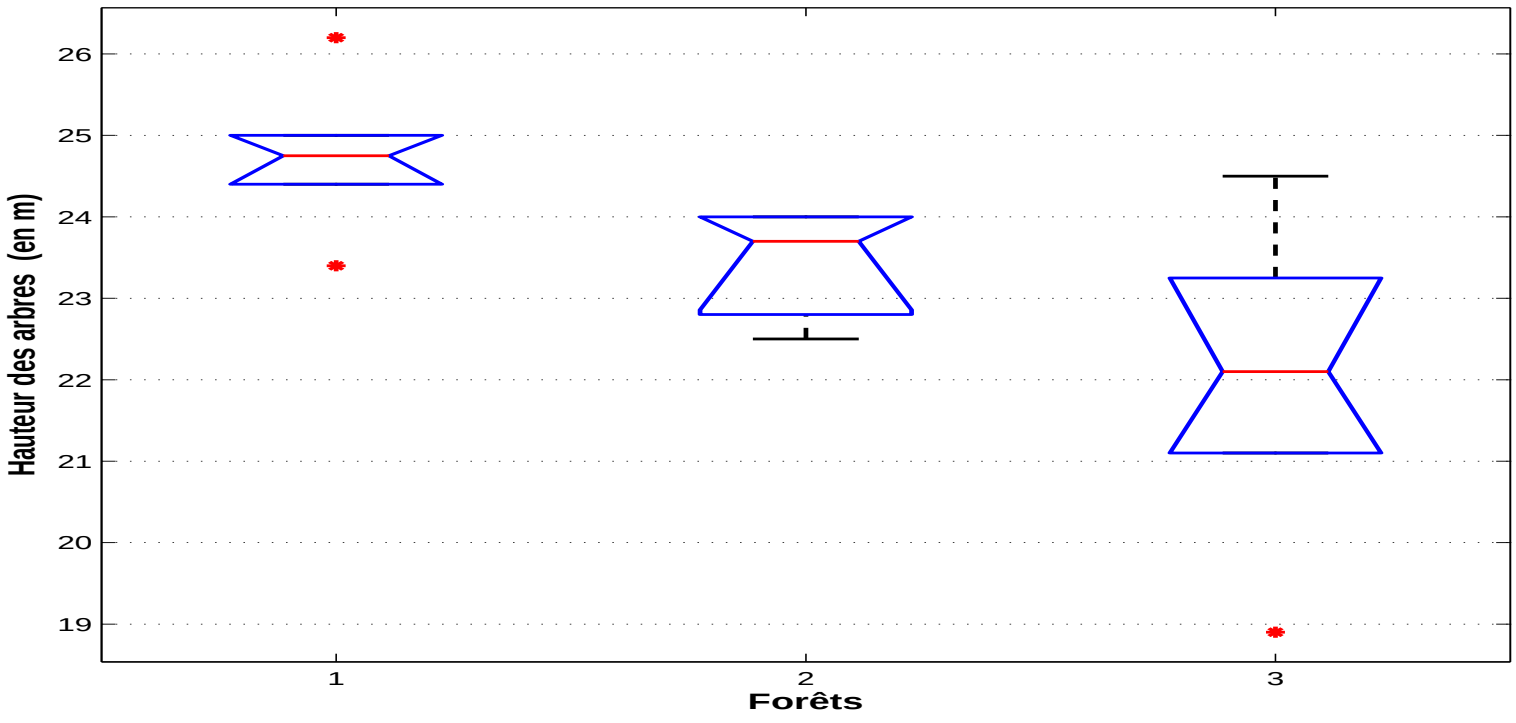
- de l'effet d'un facteur A, que l'on supposera à I niveaux,
- sur une variable quantitative Y.

Exemple

Un forestier s'intéresse aux hauteurs moyennes de 3 forêts. Pour les estimer, il échantillonne un certain nombre d'arbres et mesure leurs hauteurs :

Forêt	1	2	3
	23.4	22.5	18.9
	24.4	22.9	21.1
	24.6	23.7	21.1
	24.9	24.0	22.1
	25.0	24.0	22.5
	26.2		23.5
			24.5
Nombre d'arbres	$n_1 = 6$	$n_2 = 5$	$n_3 = 7$
Moyenne	24.75	23.42	21.96

Exemple



A partir de ces données, le forestier souhaite savoir si la hauteur moyenne des arbres est la même dans les 3 forêts, ou pas.

Le cadre

On considère I échantillons indépendants

$$X_{1,1}, \ldots, X_{1,n_1} \sim \mathcal{N}\left(\mu_1, \sigma^2\right)$$
$$\vdots$$
$$X_{I,1}, \ldots, X_{1,n_I} \sim \mathcal{N}\left(\mu_I, \sigma^2\right)$$

avec $\sigma^2 > 0$. L'objectif est donc de tester

$$H0 : \forall i, j, \mu_i = \mu_j \qquad \text{contre} \qquad H1 : \exists (i, j), \mu_i \neq \mu_j.$$

ANOVA ET MODÈLE LINÉAIRE GAUSSIEN

On peut réécrire notre modèle comme

$$\mathbf{X} = \mathbf{A}\mu + \epsilon$$

avec

$$\mathbf{X} = \begin{pmatrix} X_{1,1} \\ \vdots \\ X_{1,n_1} \\ \vdots \\ X_{I,n_I} \end{pmatrix} \in \mathbb{R}^n, \quad \mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_I \end{pmatrix}, \quad \mathbf{A} = \begin{pmatrix} \mathbf{1}_{n_1} & 0 & \dots & 0 \\ 0 & \mathbf{1}_{n_2} & \vdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \vdots & & \mathbf{1}_{n_I} \end{pmatrix}$$

avec $n = \sum_{i=1}^{I} n_i$, $\mathbf{1}_{n_i} = (1, \dots, 1)^T \in \mathbb{R}^{n_i}$ et $\epsilon \sim \mathcal{N}\left(0, \sigma^2 I_n\right)$.

Estimateur des moindres carrés

L'estimateur des moindres carrés de μ est unique et est donné par

$$\hat{\mu} = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{X} = \begin{pmatrix} \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1,i} \\ \vdots \\ \frac{1}{n_I}\sum_{i=1}^{n_I} X_{I,i} \end{pmatrix} = \begin{pmatrix} \overline{X}_{1,n_1} \\ \vdots \\ \overline{X}_{I,n_I} \end{pmatrix}.$$

L'estimateur non biaisé de σ^2 est donné par

$$\hat{\sigma}^2 = \frac{1}{n-I}\left\|\mathbf{X} - \mathbf{A}\hat{\mu}\right\|^2 = \frac{1}{n-I}\sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(X_{i,j} - \overline{X}_{i,n_i}\right)^2$$

Propriétés des estimateurs

Proposition

On a

1. $\hat{\mu} \sim \mathcal{N}\left(\mu, \sigma^2 Diag\left(n_i^{-1}\right)\right)$
2. $\frac{(n-I)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-I}$
3. $\hat{\mu}$ *et* $\hat{\sigma}^2$ *sont indépendants.*

CONSTRUCTION DE LA STATISTIQUE DE TEST

L'objectif est d'utiliser les tests sur les modèles emboités.

- Soit

$$\mathcal{M}_0 = \text{vect}\{\mathbf{1}_n\} \qquad \text{et} \qquad D = Im(\mathbf{A}) = \left\{\mathbf{A}\alpha, \alpha \in \mathbb{R}^I\right\}.$$

- Notons que $\mathcal{M}_0 \subset D$.
- On réécrit le test comme

$$H0 : \mathbb{E}[\mathbf{X}] \in \mathcal{M}_0 \qquad \text{contre} \qquad H1 : \mathbb{E}[\mathbf{X}] \in D.$$

- On considère la statistique

$$\mathcal{F} = \frac{\dim\left(D^{\perp}\right) \left\|P_{\mathcal{M}_0}\mathbf{X} - P_D\mathbf{X}\right\|^2}{\left(\dim(D) - \dim\left(\mathcal{M}_0\right)\right) \left\|P_{D^{\perp}}\mathbf{X}\right\|^2}$$

Le test

Théorème

On a

$$\mathcal{F} = \frac{(n-I)\sum_{i=1}^{I} n_i \left(\overline{X}_{i,n_i} - \overline{X}_n\right)^2}{(I-1)\sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(X_{i,j} - \overline{X}_{i,n_i}\right)^2} \sim F(I-1, N-I) \quad \textit{sous H0}$$

Un test de niveau α revient donc à considérer la zone de rejet

$$ZR_\alpha = \left\{\mathcal{F} > f_{1-\alpha,I-1,n-I}\right\}.$$

RÉÉCRITURE DU MODÈLE

On note $\mu_i = m + \alpha_i$. On peut réécrire le modèle comme

$$\mathbf{X} = \mathbf{1}_n m + \mathbf{A}\alpha + \epsilon = \tilde{\mathbf{A}} \begin{pmatrix} m \\ \alpha \end{pmatrix} + \epsilon$$

avec $\tilde{\mathbf{A}} = [\mathbf{1}_n, \mathbf{A}] \in \mathbb{R}^{n \times (I+1)}$

Remarque : On a $\text{rang}(\tilde{\mathbf{A}}) = I$ et la matrice n'est donc pas de rang plein, et on a donc pas unicité de l'estimateur des moindres carrés.

CONTRAINTES

Pour rendre l'estimateur des moindres carrés unique, on va considérer la contrainte :

$$\sum_{i=1}^{I} n_i \alpha_i = 0$$

ce qui revient à considérer que

$$m = \frac{1}{n} \sum_{i=1}^{I} n_i \mu_i.$$

Le test se réécrit alors

$$H0 : \forall i, \alpha_i = 0 \qquad \text{contre} \qquad H1 : \exists i, \alpha_i \neq 0.$$

Décomposition de $\mathbb{R}^n$

On peut écrire $\mathbb{R}^n$ comme

$$\mathbb{R}^n = \mathcal{M}_0 \oplus \mathcal{M}_A \oplus D^\perp$$

avec

- $\mathcal{M}_0 = \text{vect}\{\mathbf{1}_n\}$ correspondant à l'effet moyen
- $\mathcal{M}_A = \left\{\mathbf{A}\alpha', \sum_{i=1}^{I} n_i \alpha_i' = 0\right\}$ correspondant à l'effet dû au facteur
- avec $D^\perp$ l'espace des résidus, où $D = \mathcal{M}_0 \oplus \mathcal{M}_A = \left\{\mathbf{A}\alpha', \alpha' \in \mathbb{R}^I\right\}$

TABLEAU D'ANALYSE DE LA VARIANCE

On résume les informations dans le tableau d'analyse de la variance :

Somme des carrés des écarts	DDL	Carré moyen
$\text{SCE}_{\text{total}} = \left\|P_{\mathcal{M}_0^\perp}\mathbf{X}\right\|^2 = \sum_{i,j}\left(X_{i,j} - \overline{X}_n\right)^2$	$n-1$	$\text{CME}_{\text{total}} = \frac{\text{SCE}_{\text{total}}}{n-1}$
$\text{SCE}_{\text{facteur}} = \left\|P_{\mathcal{M}_A}\mathbf{X}\right\|^2 = \sum_{i=1}^{I} n_i\left(\overline{X}_{i,n_i} - \overline{X}_n\right)^2$	$I-1$	$\text{CME}_{\text{facteur}} = \frac{\text{SCE}_{\text{facteur}}}{I-1}$
$\text{SCE}_{\text{résidu}} = \left\|P_{D^\perp}\mathbf{X}\right\|^2 = \sum_{i,j}\left(X_{i,j} - \overline{X}_{i,n_i}\right)^2$	$n-I$	$\text{CME}_{\text{résidu}} = \frac{\text{SCE}_{\text{résidu}}}{n-I}$

QUELQUES RÉSULTATS

Proposition

On a

$$SCE_{total} = SCE_{facteur} + SCE_{résidu}.$$

- $\text{SCE}_{\text{total}}$ représente la variabilité totale dans les données
- $\text{SCE}_{\text{facteur}}$ représente la variabilité dûe au facteur.
- $\text{SCE}_{\text{résidu}}$ représente la variabilité résiduelle.

Remarque : Faire le test revient à considérer la zone de rejet

$$ZR_\alpha = \left\{ \frac{\text{CME}_{\text{facteur}}}{\text{CME}_{\text{résidu}}} > f_{1-\alpha,I-1,n-I} \right\}$$

II. Analyse de la variance à 2 facteurs

Introduction

On généralise le cadre de l'analyse de la variance à un facteur à celui à deux facteurs

Exemple : On veut étudier l'effet de deux facteurs qualitatifs, le niveau de la fertilisation et rotation de la culture, sur le poids des grains de colza. On compare

- $p = 2$ niveaux de fertilisation (1 pour faible et 2 pour fort)
- $q = 3$ types de rotation maïs/blé/colza (A sans enfouissemnt de paille, B avec enfouissement de paille, C avec 4 années de prairie temporaire entre chaque succession sans enfouissement de paille).

On a donc $p \times q = 2 \times 3 = 6$ traitements possibles, un traitement étant une combinaison de niveaux des facteurs *Fertilisation*Rotation* ($1A$, $1B$, $1C$, $2A$, $2B$, $2C$).

Exemple

En notant y_{ijk} le poids des grains sur la $k^{\text{ème}}$ parcelle traitée avec le traitement (ij) (fertilisation i, rotation j), on peut résumer ce tableau en

Rotation / Fertilisation	A ($j=1$)	B ($j=2$)	C ($j=3$)	Total
$i=1$	$\overline{y}_{11\bullet}=24,11$ $n_{11}=10$ $s_{11}=8.61$	$\overline{y}_{12\bullet}=24,00$ $n_{12}=10$ $s_{12}=7.37$	$\overline{y}_{13\bullet}=28,64$ $n_{13}=10$ $s_{13}=5.86$	$\overline{y}_{1\bullet\bullet}=25,58$ $n_{1\bullet}=30$
$i=2$	$\overline{y}_{21\bullet}=15,81$ $n_{11}=10$ $s_{21}=7.44$	$\overline{y}_{22\bullet}=19,84$ $n_{12}=10$ $s_{22}=8.27$	$\overline{y}_{23\bullet}=31,75$ $n_{13}=10$ $s_{23}=7.25$	$\overline{y}_{2\bullet\bullet}=22,47$ $n_{1\bullet}=30$
Total	$\overline{y}_{\bullet 1\bullet}=19,96$ $n_{\bullet 1}=20$	$\overline{y}_{\bullet 2\bullet}=21,92$ $n_{\bullet 2}=20$	$\overline{y}_{\bullet 3\bullet}=30,20$ $n_{\bullet 3}=20$	$\overline{y}=24,03$ $n=60$

Objectif

On cherche donc à étudier

- l'effet de deux variables qualitatives A et B, appelés facteurs,
- sur une variable quantitative Y.

On suppose que

- le facteur A a I niveaux
- et le facteur B a J niveaux.

Pour chaque couple (i, j) de niveaux des facteurs A et B, on dispose de n_{ij} mesures de Y, notées y_{ijk} avec $i = 1, \cdots, I$, $j = 1, \cdots, J$ et $k = 1, \cdots, n_{ij}$.

Analyse de la variance à deux facteurs

On considère des variables aléatoires $X_{i,j,k}$ indépendantes telles que

$$X_{i,j,k} \sim \mathcal{N}\left(\mu_{i,j}, \sigma^2\right)$$

On réécrit le modèle sous la forme

$$X_{i,j,k} = m + \alpha_i + \beta_j + \gamma_{i,j} + \epsilon_{i,j,k}$$

avec les $\epsilon_{i,j,k}$ i.i.d et $\epsilon_{i,j,k} \sim \mathcal{N}\left(0, \sigma^2\right)$.

Pour des raisons d'identifiabilité, on introduit les contraintes

$$\sum_{i=1}^{I} \alpha_i = 0, \quad \sum_{j=1}^{J} \beta_j = 0, \quad \forall i, \sum_{j=1}^{J} \gamma_{i,j} = 0, \quad \forall j, \sum_{i=1}^{I} \gamma_{i,j} = 0.$$

Remarques

- m est la moyenne globale de toutes les observations, et

$$m = \frac{1}{IJ} \sum_{i,j} \mu_{i,j}.$$

- Les α_i décrivent l'effet dû au facteur A.
- Les β_j décrivent l'éffet dû au facteur B.
- Les $\gamma_{i,j}$ décrivent l'effet d'interraction entre A et B.
- Dans ce qui suit, on considère $n_{i,j} = K$ pour tout i, j.

OBJECTIFS

- Tester l'absence d'effet principal du facteur A, i.e tester

$$H0 : \forall i, \alpha_i = 0 \qquad \text{contre} \qquad H1 : \exists i, \alpha_i \neq 0.$$

- Tester l'absence d'effet principal du facteur B, i.e tester

$$H0 : \forall j, \beta_j = 0 \qquad \text{contre} \qquad H1 : \exists j, \beta_j \neq 0.$$

- Tester l'absence d'effet d'interraction, i.e tester

$$H0 : \forall i, j, \gamma_{i,j} = 0 \qquad \text{contre} \qquad H1 : \exists (i, j), \gamma_{ij} \neq 0.$$

RÉÉCRITURE DU MODÈLE

On peut réécrire le modèle comme

$$\mathbf{X} = m\mathbf{1}_n + \mathbf{A}\alpha + \mathbf{B}\beta + \mathbf{C}\gamma + \epsilon$$

$n = IJK$.
$\mathbf{X} = \left(X_{1,1,1}, \ldots, X_{1,1,n_{1,1}}, X_{1,2,1}, \ldots, X_{1,2,n_{1,2}}, \ldots, X_{I,J,1}, \ldots X_{I,J,n_{IJ}}\right)^T$
$\alpha = \left(\alpha_1, \ldots, \alpha_I\right)^T$
$\beta = \left(\beta_1, \ldots, \beta_J\right)^T$
$\gamma = \left(\gamma_{1,1}, \ldots, \gamma_{1,J}, \ldots, \gamma_{I,1}, \ldots, \gamma_{I,J}\right)^T$
$\epsilon = \left(\epsilon_{1,1,1}, \ldots, \epsilon_{1,1,n_{1,1}}, \epsilon_{1,2,1}, \ldots, \epsilon_{1,2,n_{1,2}}, \ldots, \epsilon_{I,J,1}, \ldots \epsilon_{I,J,n_{IJ}}\right)^T \sim \mathcal{N}\left(0, \sigma^2 I_n\right)$

Réécriture du modèle

$$A = \begin{pmatrix} \mathbf{1}_{JK} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \mathbf{1}_{JK} \end{pmatrix} \in \mathbb{R}^{n \times I}, \qquad B = \begin{pmatrix} \mathbf{1}_K & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \mathbf{1}_K \\ \vdots & \vdots & \vdots \\ \mathbf{1}_K & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \mathbf{1}_K \end{pmatrix} \in \mathbb{R}^{n \times J}$$

$$C = \begin{pmatrix} \mathbf{1}_K & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \mathbf{1}_K \end{pmatrix} \in \mathbb{R}^{n \times IJ}$$

ANOVA ET MODÈLES EMBOITÉS

On peut réécrire $\mathbb{R}^n$ comme

$$\mathbb{R}^n = \mathcal{M}_0 \oplus \mathcal{M}_A \oplus \mathcal{M}_B \oplus \mathcal{M}_C \oplus D^\perp$$

avec $\mathcal{M}_0 = \text{vect}\{\mathbf{1}_n\}$ et

$$\mathcal{M}_A = \left\{ \mathbf{A}\alpha', \sum \alpha'_i = 0 \right\}$$
$$\mathcal{M}_B = \left\{ \mathbf{B}\beta'_j, \sum \beta'_j = 0 \right\}$$
$$\mathcal{M}_C = \left\{ \mathbf{C}\gamma', \sum_i \gamma'_{i,j} = 0, \sum_j \gamma'_{i,j} = 0 \right\}$$
$$D = \mathcal{M}_0 \oplus \mathcal{M}_A \oplus \mathcal{M}_B \oplus \mathcal{M}_C$$

RÉÉCRITURE DES TESTS

On peut réécire les tests comme :

- Tester l'absence d'effet principal du facteur A revient à tester

$$H0 : \mathbb{E}[\mathbf{X}] \in \mathcal{M}_0 \oplus \mathcal{M}_B \oplus \mathcal{M}_C \qquad \text{contre} \qquad H1 : \mathbb{E}[\mathbf{X}] \in D$$

- Tester l'absence d'effet principal du facteur B revient à tester

$$H0 : \mathbb{E}[\mathbf{X}] \in \mathcal{M}_0 \oplus \mathcal{M}_A \oplus \mathcal{M}_C \qquad \text{contre} \qquad H1 : \mathbb{E}[\mathbf{X}] \in D$$

- Tester l'absence d'effet d'interaction revient à tester

$$H0 : \mathbb{E}[\mathbf{X}] \in \mathcal{M}_0 \oplus \mathcal{M}_A \oplus \mathcal{M}_B \qquad \text{contre} \qquad H1 : \mathbb{E}[\mathbf{X}] \in D$$

SOMME DES CARRÉS DES ÉCARTS

- $\text{SCE}_{\text{total}} = \left\| P_{\mathcal{M}_0^\perp} \mathbf{X} \right\|^2$ la variabilité totale dans les données
- $\text{SCE}_A = \| P_{\mathcal{M}_A} \mathbf{X} \|^2$ la variabilité dû au facteur A
- $\text{SCE}_B = \| P_{\mathcal{M}_B} \mathbf{X} \|^2$ la variabilité dû au facteur B
- $\text{SCE}_{\text{inter}} = \| P_{\mathcal{M}_C} \mathbf{X} \|^2$ la variabilité dû à l'interraction
- $\text{SCE}_{\text{résidus}} = \| P_{D^\perp} \mathbf{X} \|^2$ la variabilité résiduelle non expliquée par le modèle.

Proposition

On a

$$SCE_{total} = SCE_A + SCE_B + SCE_{inter} + SCE_{résidus}.$$

TABLEAU DE L'ANOVA À DEUX FACTEURS

Somme des carrés des écarts	DDL	Carrés moyens
$\text{SCE}_{\text{total}} = \sum_{i,j,k} \left(X_{i,j,k} - \overline{X}_{.,.,.} \right)^2$	$n-1$	$\text{CME}_{\text{total}} = \frac{\text{SCE}_{\text{total}}}{n-1}$
$\text{SCE}_A = JK \sum_i \left(\overline{X}_{i,.,.} - \overline{X}_{.,.,.} \right)^2$	$I-1$	$\text{CME}_A = \frac{\text{SCE}_A}{I-1}$
$\text{SCE}_B = IK \sum_j \left(\overline{X}_{.,j,.} - \overline{X}_{.,.,.} \right)^2$	$J-1$	$\text{CME}_B = \frac{\text{SCE}_B}{J-1}$
$\text{SCE}_{\text{inter}} = K \sum_{i,j} \left(\overline{X}_{i,j,.} - \overline{X}_{i,.,.} - \overline{X}_{.,j,.} + \overline{X}_{.,.,.} \right)^2$	$(I-1)(J-1)$	$\text{CME}_{\text{inter}} = \frac{\text{SCE}_{\text{inte}}}{(I-1)(J-}$
$\text{SCE}_{\text{residu}} = \sum_{i,j,k} \left(X_{i,j,k} - \overline{X}_{i,j,.} \right)^2$	$IJ(K-1)$	$\text{CME}_{\text{residu}} = \frac{\text{SCE}_{\text{resid}}}{IJ(K-1}$

Test pour l'absence d'effet principal du facteur A

Tester l'absence d'effet principal du facteur A revient à considérer la zone de rejet

$$ZR_\alpha = \left\{ \mathcal{F}_A > f_{1-\alpha, I-1, IJ(K-1)} \right\}$$

avec

$$\begin{aligned} \mathcal{F}_A &= \frac{\dim\left(D^\perp\right) \left\| P_D \mathbf{X} - P_{\mathcal{M}_0 \oplus \mathcal{M}_B \oplus \mathcal{M}_C} \mathbf{X} \right\|^2}{\left(\dim(D) - \dim\left(\mathcal{M}_0 \oplus \mathcal{M}_B \oplus \mathcal{M}_C\right)\right) \left\| P_{D^\perp} \mathbf{X} \right\|^2} \\ &= \frac{\mathrm{CME}_A}{\mathrm{CME}_{\text{residu}}} \sim F_{I-1, IJ(K-1)} \quad \text{sous H0.} \end{aligned}$$

Test pour l'absence d'effet principal du facteur B

Tester l'absence d'effet principal du facteur B revient à considérer la zone de rejet

$$ZR_\alpha = \left\{ \mathcal{F}_B > f_{1-\alpha, J-1, IJ(K-1)} \right\}$$

avec

$$\begin{aligned} \mathcal{F}_B &= \frac{\dim\left(D^\perp\right) \left\| P_D \mathbf{X} - P_{\mathcal{M}_0 \oplus \mathcal{M}_A \oplus \mathcal{M}_C} \mathbf{X} \right\|^2}{\left(\dim(D) - \dim\left(\mathcal{M}_0 \oplus \mathcal{M}_A \oplus \mathcal{M}_C\right)\right) \left\| P_{D^\perp} \mathbf{X} \right\|^2} \\ &= \frac{\mathrm{CME}_B}{\mathrm{CME}_{\text{residu}}} \sim F_{J-1, IJ(K-1)} \quad \text{sous H0.} \end{aligned}$$

TEST POUR L'ABSENCE D'EFFET D'INTERRACTION

Tester l'absence d'effet d'interraction revient à considérer la zone de rejet

$$ZR_\alpha = \left\{ \mathcal{F}_B > f_{1-\alpha,(I-1)(J-1),IJ(K-1)} \right\}$$

avec

$$\begin{aligned} \mathcal{F}_C &= \frac{\dim\left(D^\perp\right) \left\| P_D \mathbf{X} - P_{\mathcal{M}_0 \oplus \mathcal{M}_A \oplus \mathcal{M}_B} \mathbf{X} \right\|^2}{\left(\dim(D) - \dim\left(\mathcal{M}_0 \oplus \mathcal{M}_A \oplus \mathcal{M}_B\right)\right) \left\| P_{D^\perp} \mathbf{X} \right\|^2} \\ &= \frac{\mathrm{CME}_C}{\mathrm{CME}_{\text{residu}}} \sim F_{(I-1)(J-1),IJ(K-1)} \quad \text{sous H0.} \end{aligned}$$