Analyse de la variance à un facteur

Y : variable quantitative

X : variable qualitative "facteur"

objectif

Etudier l'effet d'une variable qualitative sur une variable quantitative

Modèle

$$Y_{ij} = \underbrace{\mu + \alpha_i}_{\mu_i} + \varepsilon_{ij}$$

μ : Effet moyen général

Y_{ij} : variable à expliquer

α_i : effet du $i^{ème}$ niveau du facteur

ε_{ij} : Erreur

- ε_{ij} indépendants
- $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma)$ $\forall i, j$.

→ Facteur à k niveaux.

↳ Pour chaque niveau i

↳ n_i : taille de l'échantillon i

↳ Y_{ij} = les valeurs de Y

niveau

observation

$(1 \leq i \leq k)$ $(1 \leq j \leq n_i)$

$$\boxed{Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}}$$

$$n = n_1 + n_2 + \dots + n_k$$
$$= \sum_{i=1}^{k} n_i$$: (taille totale)

Conditions d'application : les variables doivent être normalement distribuées indépendantes et de même variance.

Test d'Homogénéité des variances

Comparaison des variances.

Test d'hypothèses :

$$\begin{cases} H_0 : \sigma_1 = \sigma_2 = \dots = \sigma_k. \\ H_1 : \exists\ i \neq l \text{ tq } \sigma_i \neq \sigma_l. \end{cases}$$

$$\begin{cases} H_0 : \text{variances égales} \\ H_1 : \text{il existe au moins deux variances qui ne sont pas égales.} \end{cases}$$

Statistique du test :

χ^2_{k-1} Khi-deux $(k-1)$ ddl

→ **Commande R** : bartlett.test( ).

Test d'égalité des moyennes

"test d'effet"

$$\begin{cases} H_0 : \mu_1 = \dots = \mu_k = \mu \\ H_1 : \exists i \text{ tq } \mu_i \neq \mu \end{cases} \iff \begin{cases} H_0 : \alpha_1 = \dots = \alpha_k = 0 \\ H_1 : \exists i \text{ tq } \alpha_i \neq 0 \end{cases}$$

H_0 : Tous les niveaux ont le même effet que l'effet moyen général.

H_1 : Il existe au moins un niveau avec un effet différent de l'effet moyen général

$\iff$ H_0 : L'effet individuel de chaque niveau est nul

H_1 : Il existe au moins un niveau avec un effet individuel non nul.

$\iff$ H_0 : Absence d'effet du facteur

H_1 : Le facteur a un effet sur la variable.

Estimation

Sous H_0 : $\mu_1 = \dots = \mu_k = \mu$
$(\alpha_1 = \dots = \alpha_k = 0)$.

Modèle réduit :

$$Y_{ij} = \mu + \varepsilon_{ij}$$

Estimation

$$\hat{\mu} = \overline{Y}$$

Résidus

$$e_{ij} = Y_{ij} - \overline{Y}$$

$$\overline{Y} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}$$

moyenne générale

$$\overline{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$$

moyenne pour chaque niveau

Équation de l'analyse de la variance

$$\boxed{SCT = SCE + SCR}$$

somme des carrés totale	Sommes des carres factorielle [Inter-groupe]	somme des carrés résiduelle [Intra-groupes]
•) Variation totale	•) variation due au facteur	•) Variation résiduelle.
Dispersion des données autour de la moyenne générale	Dispersion des moyennes de chaque niveau autour de la moyenne générale	dispersion des données à l'intérieur de chaque échantillon autour de sa moyenne

$$\boxed{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}\right)^2 = \sum_{i=1}^{k} n_i\left(\bar{Y}_i-\bar{Y}\right)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_i\right)^2}$$

Tableau d'analyse de la variance

Sources de variation	ddl	Somme des carrés	Carrés moyens	Statistique
Inter-groupe. "Variation factorielle"	$k-1$	$SCE = \sum_{i=1}^{k} n_i \left(\overline{Y}_{i.} - \overline{Y}\right)^2$	$\frac{SCE}{k-1} = \frac{1}{k-1} \sum_{i=1}^{k} n_i \left(\overline{Y}_i - \overline{Y}\right)^2$	$F = \dfrac{\frac{SCE}{k-1}}{\frac{SCR}{n-k}}$ Fisher $(k-1)\,;\,(n-k)$ degrés de libertés.
Intra-groupe. "Variation résiduelle"	$n-k$	$SCR = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij.} - \overline{Y}_i\right)^2$	$\frac{SCR}{n-k} = \frac{1}{n-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij.} - \overline{Y}_i\right)^2$	
Totale "Variation totale"	$n-1$	$SCT = SCE + SCR$		

En se basant sur la statistique F et le Pvalue affichée on prend la décision