République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITÉ MOHAMED KHIDER BISKRA

Faculté des Sciences Exactes et Sciences de la Nature et de la Vie
Département de Mathématiques

Première Année Master

Notes de Cours

Analyse de Données

Chapitre 3 : Analyse de la variance
(Séance 7)

Auteur des notes :

Dr. Sana BENAMEUR

Année universitaire : 2021-2022

Chapitre 3

Analyse de la Variance

3.1 Analyse de la variance à un facteur AV(1)

L'analyse de la variance est une technique qui sert à tester l'influence d'un (ou de plusieurs) facteur (s) qualitative (s) sur une variable quantitative. Notons A le facteur et $A_1, \ldots, A_p$ ses p modalités ou niveaux. Soient Y la variable étudier et $y_1, y_2, \ldots, y_n$ ses observations. Pour chaque niveau A_j du facteur est associer n_j mesures de Y : $y_{1j}, y_{2j}, \ldots, y_{n_j j}$.

y_{ij} : $i^{ème}$ observation du $j^{ème}$ niveau

n_j : la taille de l'échantillon y_{*j}

$\bar{y}_j$: la moyenne de la classe j

n : taille de l'échantillon y

$\bar{y}$: la moyenne totale

A_1	$\cdots$	A_j	$\cdots$	A_p
y_{11}	$\cdots$	y_{1j}	$\cdots$	y_{1p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$y_{n_1 1}$	$\cdots$	$y_{n_j j}$	$\cdots$	$y_{n_p p}$
$\bar{y}_1$	$\cdots$	$\bar{y}_j$	$\cdots$	$\bar{y}_p$

On a

$$n = \sum_{j=1}^{p} n_j, \ \bar{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}, \ \bar{y} = \frac{1}{n} \sum_{j=1}^{p} n_j \bar{y}_j = \frac{1}{n} \sum_{j=1}^{p} \sum_{i=1}^{n_j} y_{ij}.$$

3.1.1 Modèle d'AV(1)

Soit $\mu = \frac{1}{n} \sum_{j=1}^{p} n_j \mu_j$, ($\mu_j = \mathbb{E}[y_j]$: moyenne théorique). En terme d'observation, on a le modèle :

$$\underbrace{(y_{ij} - \bar{y})}_{\text{écart total}} = \underbrace{(y_{ij} - \bar{y}_j)}_{\text{écart résiduel}} + \underbrace{(\bar{y}_j - \bar{y})}_{\text{écart factoriel}}, \tag{1}$$

et théoriquement, on a

$$y_{ij} - \mu = (y_{ij} - \mu_j) + (\mu_j - \mu). \tag{2}$$

Posons $\mu_j - \mu = \alpha_j$ et $y_{ij} - \mu_j = \varepsilon_{ij}$. Nous obtenons le modèle d'AV(1) :

$$\underbrace{y_{ij}}_{\text{modèle}} = \underbrace{\mu}_{\text{moyenne générale}} + \underbrace{\alpha_j}_{\text{effet principale}} + \underbrace{\varepsilon_{ij}}_{\text{effet résiduel}},$$

avec $\varepsilon \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right)$. Il est claire que $\bar{\alpha} = 0$.

3.1.2 Equation d'AV(1)

Le modèle observé (1), permet d'obtenir l'équation d'AV(1)

$$\begin{aligned}
\sum_{j=1}^{p}\sum_{i=1}^{n_j}\left(y_{ij}-\bar{y}\right)^2 &= \sum_{j=1}^{p}\sum_{i=1}^{n_j}\left(y_{ij}-\bar{y}_j\right)^2 + \sum_{j=1}^{p}\sum_{i=1}^{n_j}\left(\bar{y}_j-\bar{y}\right)^2 \\
\underbrace{\sum_{j=1}^{p}\sum_{i=1}^{n_j} y_{ij}^2 - n\bar{y}^2}_{SCT} &= \underbrace{\sum_{j=1}^{p}\sum_{i=1}^{n_j}\left(y_{ij}-\bar{y}_j\right)^2}_{SCR} + \underbrace{\sum_{j=1}^{p} n_j\bar{y}_j^2 - n\bar{y}^2}_{SCA}.
\end{aligned}$$

Variation	**ddl**	SC	MC	F
Factorielle	$p-1$	SCA	MCA	$\frac{MCA}{MCR}$
Résiduelle	$n-p$	SCR	MCR	
Totale	$n-1$	SCT		

TAB. 3.1. Tableau d'AV(1)

3.1.3 Test d'égalité des moyennes

Si le facteur A n'as pas d'influence sur Y, alors l'hypothèse à tester est

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_p \Leftrightarrow \alpha_1 = \alpha_2 = \cdots = \alpha_p = 0.$$

Sous l'hypothèse de normalité des ε, on a

$$MCA = \frac{SCA}{p-1} \sim \mathcal{X}_{p-1}^2, MCR = \frac{SCR}{n-p} \sim \mathcal{X}_{n-p}^2 \text{ et } F = \frac{MCA}{MCR} \sim \mathcal{F}\left(p-1, n-p\right).$$

Nous acceptons l'hypothèse H_0 d'égalité des moyenne si $F \leq f_{1-\alpha}\left(p-1, n-p\right)$. Le rejet de H_0 implique qu'au moins deux moyennes sont différentes $\left(\exists \mu_k, \mu_j \left| \mu_k \neq \mu_j\right.\right)$. Dans ce cas nous utilisons la méthode de Bonferrouni basée sur la comparaison deux à deux des couple $\left(\mu_k, \mu_j\right)$:

$$\begin{cases} H_0 : \mu_k - \mu_j = 0 \\ H_1 : \mu_k - \mu_j \neq 0 \end{cases}, k \neq j$$

On a donc, $m = C_p^2$ tests de Student à faire. on accepte H_0 $\left(\mu_k = \mu_j\right)$ si :

$|T| < t_{1-\frac{\alpha}{2m}}(n-p)$, où

$$T = \frac{\bar{y}_k - \bar{y}_j}{\sqrt{\left(\frac{1}{n_k} + \frac{1}{n_j}\right) MCR}},$$

t_* est le fractile d'ordre $*$ de la loi de Student à $n-p$ degré de liberté.

3.1.4 Exemple d'AV(1)

On veut étudier l'influence de trois types d'essence A_1, A_2, A_3 sur les distances parcourues en (km). Le type d'essence influe sur les distances ? Localiser les différences si oui :

A_1	240	250	243	255	
A_2	253	265	264	270	276
A_3	233	240	247		

Solution

Le tableau d'$AV(1)$ est donné par

Variation	**ddl**	SC	MC	F
Factorielle	2	1444.8	722.4	12.38
Résiduelle	9	525.2	58.4	
Totale	11	1970		

$p = 3$,
$n_1 = 4, n_2 = 5, n_3 = 3, n = 12$
$\bar{y}_1 = 247, \bar{y}_2 = 265.6, \bar{y}_3 = 240$,
$\bar{y} = 253$

Au niveau de confiance 95%, $F = 12.38 > f = 4.26$. Alors il y a une influence (non égalité des moyenne : H_0 est rejeté). Il y'a $m = C_3^2 = 3$ tests de comparaison $(\mu_1 \text{ vs } \mu_2), (\mu_1 \text{ vs } \mu_3)$ et $(\mu_2 \text{ vs } \mu_3)$:
1^{ier} cas $(\mu_1 \text{ vs } \mu_2)$: on a

$$T = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) MCR}} = \frac{247 - 265.6}{\sqrt{\left(\frac{1}{4} + \frac{1}{5}\right) 58.4}} = -3.63,$$

avec $t_{1-\frac{\alpha}{2m}}(n-p) = t_{1-\frac{0.05}{6}}(9) = 2.82 < |T| \Rightarrow H_0$ est rejeté $(\mu_1 \neq \mu_2)$ à 95%. De même pour les autres cas, on $(\mu_1 = \mu_3)$ et $(\mu_2 \neq \mu_3)$.On dit dans ce cas que les types A_1 et A_3 ont le même effet qu'est différent de celui de A_2.