Chapitre 3: Analyse de la variance

1) Analyse de la variance à un facteur $AV(1)$:

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

y_{ij} : $i^{ème}$ obs du $j^{ème}$ niveau
n_j : taille de l'échantillon y_{*j}
$\overline{y_j}$: moyenne de la classe j
n : taille de l'échantillon y
$\overline{y}$: moyenne totalle

A_1	...	A_j	...	A_p
y_{11}	...	y_{1j}	...	y_{1p}
:	:	:	:	:
$y_{n_1 1}$	...	$y_{n_j j}$	...	$y_{n_p p}$
$\overline{y_1}$	...	$\overline{y_j}$	...	$\overline{y_p}$

On a

$$n = \sum_{j=1}^{p} n_j, \qquad \overline{y_j} = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij}, \qquad \overline{y} = \frac{1}{n}\sum_{j=1}^{p} n_j \overline{y_j} = \frac{1}{n}\sum_{j=1}^{p}\sum_{i=1}^{n_j} y_{ij}.$$

1.1) Modèle d'$AV(1)$:

Soit $\mu = \frac{1}{n}\sum_{j=1}^{p} n_j \mu_j, \quad \left(\mu_j = E[Y_j] : \text{ moyenne théorique}\right)$. En terme d'observations, on a le modèle:

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

et théoriquement, on a

$$y_{ij} - \mu = \left(y_{ij} - \mu_j\right) + \left(\mu_j - \mu\right).$$

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

$$\underbrace{y_{ij}}_{\text{modèle}} = \underbrace{\mu}_{\text{moyenne générale}} + \underbrace{\alpha_j}_{\text{effet principale}} + \underbrace{\varepsilon_{ij}}_{\text{effet résiduel}}, \tag{2}$$

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

1.2) Equation d'$AV(1)$:

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

$$\sum_{j=1}^{p}\sum_{i=1}^{n_j}(y_{ij} - \overline{y})^2 = \sum_{j=1}^{p}\sum_{i=1}^{n_j}(y_{ij} - \overline{y_j})^2 + \sum_{j=1}^{p}\sum_{i=1}^{n_j}(\overline{y_j} - \overline{y})^2$$

$$\underbrace{\sum_{j=1}^{p}\sum_{i=1}^{n_j} y_{ij}^2 - n\overline{y}^2}_{SCT} = \underbrace{\sum_{j=1}^{p}\sum_{i=1}^{n_j}(y_{ij} - \overline{y_j})^2}_{SCR} + \underbrace{\sum_{j=1}^{p} n_j \overline{y_j}^2 - n\overline{y}^2}_{SCA}.$$

Tableau d'$AV(1)$

Variation	ddl	SC	MC	F
Factorielle	$p-1$	SCA	MCA	**F**
Résiduelle	$n-p$	SCR	MCR	
Total	$n-1$	SCT		

1.3) Test d'égalité des moyennes:

Si le facteur A n'a pas d'influance sur Y. Alors, l'hypothèse à tester est

$$H_0 : \mu_1 = \mu_2 = ... = \mu_p \Longleftrightarrow \alpha_1 = \alpha_2 = ... = \alpha_p = 0.$$

Sous l'hypothèse de normalité des ε, on a

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

Nous acceptons l'hypothèse H_0 d'égalité des moyenne si $F \leq f_{1-\alpha}(p-1, n-p)$. Le rejette de H_0 implique qu'au moins deux moyennes sont différentes $(\exists \mu_k, \mu_j | \mu_k \neq \mu_j)$. Dans ce cas nous utilisons le test de Benfarouni basé sur les comparaisons deux à deux des couples (μ_k, μ_j) :

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

On a donc, $m = \complement_p^2$ tests de Student à faire. On accepte $H_0\,(\mu_k = \mu_j)$ si $|T| < t_{1-\frac{\alpha}{2m}}(n-p)$, où

$$T = \frac{\overline{y_k} - \overline{y_j}}{\sqrt{\left(\frac{1}{n_k} + \frac{1}{n_j}\right) MCR}},$$

t_* est le fractile d'order $*$ du Stedent à $n-p$ degré de liberté.

1.4) Exemple d'AV (1) : En veux étudier l'influence de trois types d'essence A_1, A_2, A_3 sur les distances parcourés en (km). Le type d'éssence 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	1550	775	21.46
Résiduelle	9	325	36.11	
Total	11	1875		

$\overline{y_1} = 247, \ \overline{y_2} = 265.6, \ \ \overline{y_3} = 240.$

Au niveau 95% de confiance,$F = 21.46 > f_{0.95}(2, 9) = 19.4$. Alors il y une influence (non égalté des moyennes : H_0 rejetté). Il y'a $m = \complement_3^2 = 3$ tests de comparaison $(\mu_1 \ vs \ \mu_2)$, $(\mu_1 \ vs \ \mu_3)$, $(\mu_2 \ vs \ \mu_3)$:
1^{er} *cas*, $(\mu_1 \ vs \ \mu_2)$: on a

$$T_1 = \frac{\overline{y_1} - \overline{y_2}}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) MCR}} = \frac{247 - 265.6}{\sqrt{\left(\frac{1}{4} + \frac{1}{5}\right) 58.36}} = -3.53$$

avec, $t_{1-\frac{0.05}{6}}(9) = 2.28 < |T| \Longrightarrow H_0$ rejetté $(\mu_1 \neq \mu_2)$ à 95%. (de même pour les autres cas, on a $(\mu_1 = \mu_3)$ et $(\mu_2 \neq \mu_3)$). On dit dans ce cas que les type A_1 et A_3 ont le même effet qu'est différent de celui de A_2.

2) Analyse de la variance à deux facteur AV (2) :

Dans ce cas les données sont regroupées selon deux catégories ou facteurs: A, B. Notons $A_1, ..., A_p$ les p niveaux du facteur A et $B_1, ..., B_q$ les q niveaux du facteur B, de telle sort que chaque case contient r mesures de la variable d'interêt Y $(y_{**1}, ..., y_{**r})$:

	A_1	...	A_i	...	A_p
B_1	$y_{111}, ..., y_{11r}$	...	$y_{i11}, ..., y_{i1r}$	...	$y_{p11}, ..., y_{p1r}$
:	:	:	:	:	:
B_j	$y_{1j1}, ..., y_{1jr}$	...	$y_{ij1}, ..., y_{ijr}$	...	$y_{pj1}, ..., y_{pjr}$
:	:	...	:	...	:
B_q	$y_{1q1}, ..., y_{1qr}$	...	$y_{iq1}, ..., y_{iqr}$	...	$y_{pq1}, ..., y_{pqr}$

2.1) Modèle d'$AV\,(2)$:

$$y_{ijk} = \underbrace{\mu}_{\text{moyenne générale}} + \underbrace{\alpha_i}_{\text{effet de } A} + \underbrace{\beta_j}_{\text{effet de } B} + \underbrace{\gamma_{ij}}_{\text{effet d'interaction}} + \underbrace{\varepsilon_{ijk}}_{\text{effet résiduel}},$$

$$\begin{cases} \alpha_i = \mu_{i*} - \mu & \text{effet de } A \text{ au niveau } i = 1,...,p \\ \beta_j = \mu_{*j} - \mu & \text{effet de } B \text{ au niveau } j = 1,...,q \\ \gamma_{ij} = \mu_{ij} - \mu_{i*} - \mu_{*j} + \mu & \text{effet d'interaction entre } A \text{ et } B \\ \varepsilon_{ijk} = y_{ijk} - \mu_{ij} & \text{erreur } \ (k = 1,...,r)\,. \end{cases}$$

Les hypothèses d'interêt dans l'$AV\,(2)$ sont:

$$\begin{cases} 1) \quad H_0 : \mu_{ij} = \mu_{\ell m}, \quad H_1 : \exists!\,(i,j,\ell,m) \mid \mu_{ij} \neq \mu_{\ell m} \\ 2) \quad H_0 : \alpha_i = 0, \qquad H_1 : \alpha_i \neq 0, \qquad i = 1,...,p \\ 3) \quad H_0 : \beta_j = 0, \qquad H_1 : \beta_j \neq 0, \qquad j = 1,...,q \\ 4) \quad H_0 : \gamma_{ij} = 0, \qquad H_1 : \gamma_{ij} \neq 0, \qquad i = 1,...,p, \quad j = 1,...,q \end{cases}$$

2.2) Equation d'$AV\,(2)$:

Notons

$$\begin{aligned} &\overline{y_{i**}} = \tfrac{1}{qr}\sum_{j=1}^{q}\sum_{k=1}^{r} y_{ijk} && \text{moyenne de la colonne } i \\ &\overline{y_{*j*}} = \tfrac{1}{pr}\sum_{i=1}^{p}\sum_{k=1}^{r} y_{ijk} && \text{moyenne de la ligne } j \\ &\overline{y_{ij*}} = \tfrac{1}{r}\sum_{k=1}^{r} y_{ijk} && \text{moyenne de la case } \ (i,j) \\ &\overline{y} = \overline{y_{***}} = \tfrac{1}{pqr}\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{r} y_{ijk} && \text{moyenne total } \ (n = pqr)\,. \end{aligned}$$

L'equation d'analyse de la variance dans ce cas s'écrit

$$\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{r}(y_{ijk} - \overline{y})^2 = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{r}(y_{ijk} - \overline{y_{ij*}})^2 + qr\sum_{i=1}^{p}(\overline{y_{i**}} - \overline{y})^2 + pr\sum_{j=1}^{q}(\overline{y_{*j*}} - \overline{y})^2 + r\sum_{i=1}^{p}\sum_{j=1}^{q}(\overline{y_{ij*}} - \overline{y_{i**}} - \overline{y_{*j*}} + \overline{y})^2$$

en d'autre terme:

$$\begin{aligned} \underbrace{\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{r} y_{ijk}^2 - n\overline{y}^2}_{SCT} \quad = \quad & \underbrace{\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{r}(y_{ijk} - \overline{y_{ij*}})^2}_{SCR} + \underbrace{qr\sum_{i=1}^{p}\overline{y_{i**}}^2 - n\overline{y}^2}_{SCA} \\ & + \underbrace{pr\sum_{j=1}^{q}\overline{y_{*j*}} - n\overline{y}^2}_{SCB} + \underbrace{r\sum_{i=1}^{p}\sum_{j=1}^{q}(\overline{y_{ij*}} - \overline{y_{i**}} - \overline{y_{*j*}} + \overline{y})^2}_{SCAB} \end{aligned}$$

L'$AV\,(2)$ permet de tester l'absence ou l'existence des effets de A et/ou de B ou d'interaction enter A et B. Les résultats d'$AV\,(2)$ sont résumés dans le tableau suivant:

Tableau d'$AV\,(2)$

Variation	ddl	SC	MC	F
Fac. A	$p-1$	SCA	MCA	$\mathbf{F}_A$
Fac. B	$q-1$	SCB	MCB	$\mathbf{F}_B$
Fac. A, B	$(p-1)\,(q-1)$	$SCAB$	$MCAB$	$\mathbf{F}_{AB}$
Résiduelle	$pq\,(r-1)$	SCR	MCR	
Total	$n-1 = pqr-1$	SCT		

$$\begin{cases} F_A > f_{1-\alpha}\left(p-1, pq\left(r-1\right)\right) & \Rightarrow H_1 : \text{effet de } A \\ F_B > f_{1-\alpha}\left(q-1, pq\left(r-1\right)\right) & \Rightarrow H_1 : \text{effet de } B \\ F_{AB} > f_{1-\alpha}\left(\left(p-1\right)\left(q-1\right), pq\left(r-1\right)\right) & \Rightarrow H_1 : \text{effet d'interaction.} \end{cases}$$

2.3) Modèle d'$AV\,(2)$ sans répétitions:

Supposons dans ce cas que chaque case contien une seulle observation $(r = 1)$. On dit aussi qu'une seulle mesure de la vraible Y pour chaque couple (A_i, B_j) :

	A_1	...	A_i	...	A_p
B_1	y_{11}	...	y_{i1}	...	y_{p1}
:	:	:	:	:	:
B_j	y_{1j}	...	y_{ij}	...	y_{pj}
:	:	...	:	...	:
B_q	y_{1q}	...	y_{iq}	...	y_{pq}

Tableau d'$AV\,(2)$

Variation	ddl	SC	MC	F
Fac. A	$p-1$	SCA	MCA	$\mathbf{F}_A$
Fac. B	$q-1$	SCB	MCB	$\mathbf{F}_B$
Résiduelle	$(p-1)(q-1)$	SCR	MCR	
Total	$pq-1$	SCT		

Le modèle d'$AV\,(2)$ est

$$y_{ij} = \underbrace{\mu}_{\text{moyenne générale}} + \underbrace{\alpha_i}_{\text{effet de } A} + \underbrace{\beta_j}_{\text{effet de } B} + \underbrace{\varepsilon_{ij}}_{\text{effet résiduel}}.$$

et l'equation d'$AV\,(2)$ est

$$\underbrace{\sum_{i=1}^{p}\sum_{j=1}^{q}\left(y_{ijk}-\overline{y}\right)^2}_{SCT} = \underbrace{\sum_{i=1}^{p}\sum_{j=1}^{q}\left(y_{ij}-\overline{y_{ij}}\right)^2}_{SCR} + \underbrace{q\sum_{i=1}^{p}\left(\overline{y_{i*}}-n\overline{y}\right)^2}_{SCA} + \underbrace{p\sum_{j=1}^{q}\left(\overline{y_{*j}}-n\overline{y}\right)^2}_{SCB}.$$

2.4) Exemple: considérons les résultats d'$AV\,(2)$:

Variation	ddl	SC	MC	F
Fac. A	2	105	52.5	17.5
Fac. B	4	225	112.5	37.5
Fac. A, B	8	130	65	21.67
Résiduelle	30	90	3	
Total	44	550		

$p = 3$
$q = 5$
$r = 3$
$n = 45$

au niveau 95% de confiance, nous avons:

$$\begin{cases} F_A = 17.5 < f_{0.05}\left(2, 30\right) = 19.5 & \Rightarrow H_0 : \text{absence d'effet de } A \\ F_B = 37.5 > f_{0.05}\left(4, 30\right) = 5.75 & \Rightarrow H_1 : \text{effet de } B \\ F_{AB} = 21.67 > f_{0.05}\left(8, 30\right) = 3.08 & \Rightarrow H_1 : \text{effet d'interaction.} \end{cases}$$

2.5) Exercice: La quantité d'oxygène consommé par deux espèces de patelle : Acmaea Scabra et Acmaea Digitalis a été analysée pour différentes conditions halines : pourcentages d'eau; les résultats suivants ont été obtenues:

% d'eau	A. Scabra	A. Digitalis
100%	7.16, 6.78, 13.6, 8.93, 8.26	6.14, 3.86, 10.4, 5.49, 6.14
75%	5.2, 5.2, 7.18, 6.37, 13.2	4.47, 9.9, 5.75, 11.8, 4.95
50%	11.11, 9.74, 18.8, 9.74, 10.5	9.63, 6.38, 13.4, 14.5, 14.5

Analyser les résultats obtenus au cours de cette expérience au niveaux : $\alpha = 90\%, 95\%, 99\%$.