Chapitre 2 : Analyse de la variance

1. Analyse de la variance à un facteur :

- Dans l'étude des séries statistiques doubles, on peut avoir trois cas :

a) Deux Variables quantitatives⟶ régression, corrélation.

b) Deux Variables qualitatives⟶ khi2.

c) Une variable quantitative et l'autre qualitatives ⟶ ANOVA.

- L' ANOVA : ANalysis Of Variance.
- L'analyse de variance (comme son nom ne l'indique pas) permet de comparer des moyennes de plusieurs échantillons indépendants afin de tester l'influence d'un ou plusieurs facteurs.
- On étudie une variable aléatoire X sur plusieurs populations P_1, P_2,…, P_k.
- μ_i moyenne de X dans P_i $(1 \leq i \leq k)$.
- On extrait de chaque population P_i un échantillon E_i $de\ taille\ n_i$ (les échantillons n'ont pas forcément la même taille).

Tableau des données et notation

	Facteur 1	**Facteur 2**	**. . . . . .**	**Facteur s**	**Moyenne des échantillons**
E_1 (de taille n_1)	x_{11}	x_{12}	. . . . .	x_{1s}	Moyenne de l'éch 1 : $\overline{x_1}$
E_2 (de taille n_2)	x_{21}	x_{22}	. . . . .	x_{2s}	Moyenne de l'éch 2 : $\overline{x_2}$
. . . . .	. . . . .	. . . . .	. . . . .	. . . . .	. . . . .
E_i (de taille n_i)	x_{i1}	x_{i2}	. . . . .	x_{is}	Moyenne de l'éch i : $\overline{x_i}$
. . . .	. . . . .	. . . . .	. . . . .	. . . . .	. . . . .
E_k (de taille n_k)	x_{k1}	x_{k2}	. . . . .	x_{ks}	Moyenne de l'éch k : $\overline{x_k}$

- $N = \sum_{i=1}^{k} n_i$: Taille de tous les échantillons réunis.
- $\bar{X} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{s} x_{ij}$ ou bien $\bar{X} = \frac{1}{N}\sum_{i=1}^{k} n_i \bar{x}_i$ (moyenne des moyennes).
- k : le nombre des échantillons.
- s : le nombre de facteurs.

➢ Se basant sur les k échantillons, on va comparer les moyennes des populations (pas deux à deux mais toutes en même temps).

➢ **<u>Problème posé</u> :** à quoi est due la différence de moyenne entre les échantillons ?

Mise en place du test :

A- Les hypothèses :

H_0: $\mu_1 = \mu_2 = \ \ldots = \mu_k$, (la moyenne des populations est indépendante du facteur étudié)

Contre

H_1: $\exists\ i,j \quad tq \quad \mu_i \neq \mu_j$.

B- Conditions d'application :

- Tous les échantillons sont gaussiens (suivent une loi normale).
- Les populations sont de même variances (homoscédasticité).
- Les échantillons sont indépendants.

C- Calculs :

- A quoi sont dus les écarts à la moyenne ?

On définit 3 écarts :

- Ecart à la moyenne globale : $x_{ij} - \bar{X}$

(La différence entre chaque valeur et la moyenne générale)

- Ecart entre les groupes : $\bar{x_i} - \bar{X}$

(La différence définie par les facteurs : différence expliquée)

- Ecart à l'intérieur des groupes : $x_{ij} - \bar{x_i}$

(La différence non définie : différence inexpliquée résiduelle)

<u>Un petit exemple pour comprendre les calculs</u>

Les échantillons sont des parcelles de terre.

X : On étudie le rendement de pommes de terre dans différents milieux.

	Milieu 1	Milieu 2	Milieu 3	Milieu 4	Moyenne
E_1	3	2	1	2	$\overline{x_1} = \frac{3+2+1+2}{4} = 2$
E_2	5	3	4	3	$\overline{x_2} = \frac{5+3+4+3}{4} = 3{,}75$
E_3	5	7	1	1	$\overline{x_3} = \frac{5+7+1+1}{4} = 3{,}5$

$$\bar{X} = \frac{3+2+1+2+5+3+4+3+5+7+1+1}{12} = 3{,}08$$

Ou bien

$$\bar{X} = \frac{4(2)+4(3{,}75)+4(3{,}5)}{12} = 3{,}08$$

$$x_{ij} - \bar{X} = \left(x_{ij} - \overline{x_\iota}\right) + \left(\overline{x_\iota} - \bar{X}\right)$$

Écart à la moyenne

Écart à l'intérieur du groupe

Écart entre les groupes

En passant au carré et en faisant les sommations adéquates, on obtient l'équation d'analyse de variance

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{X})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x_i})^2 + \sum_{i=1}^{k} n_i(\bar{x_i}-\bar{X})^2$$

SCT : somme des carrés totaux
Exprime la variabilité totale des observations

SCR : somme des carrés résiduels
Exprime la variabilité résiduelle, à savoir la variation que le facteur n'arrive pas à expliquer

SCE : somme des carrés expliqués
Exprime la variabilité expliquée, à savoir la variabilité que le facteur explique

Dans l'exemple :

- SCE : somme des carrés expliqués est la variabilité qu'on explique par la différence des milieux.
- SCR : somme des carrés résiduels est la variabilité non expliquée : pour le même milieu on a différents rendements et on ne connait pas la raison.
- SCT : somme des carrés totaux exprime la variabilité totale des observations.

Pour chaque somme des carrés, on a un degré de liberté associé qu'on résume dans ce tableau :

	Formules	Degrés de liberté
SCR : somme des carrés résiduels	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij}-\overline{x_i}\right)^2$	$ddl_{SCR}=N-k$ (Nbre données total -nbre des échantillons)
SCE : somme des carrés expliqués	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{x_i}-\bar{X})^2 = \sum_{i=1}^{k} n_i\,(\overline{x_i}-\bar{X})^2$	$ddl_{SCE}=k-1$ (nbre des échantillons -1)
SCT : somme des carrés totaux	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij}-\bar{X}\right)^2$	$ddl_{SCT}=N-1$ (Nbre données total -1)

$$ddl_{SCT}=ddl_{SCR}+ddl_{SCE}$$

Interprétation : la source des variations entre les valeurs est due soit aux facteurs (expliqués) ou à des résiduels (inconnus).

Application numérique :

	Milieu 1	Milieu 2	Milieu 3	Milieu 4	Moyenne
E_1	3	2	1	2	$\overline{x_1}=2$
E_2	5	3	4	3	$\overline{x_2}=3{,}75$
E_3	5	7	1	1	$\overline{x_3}=3{,}5$

SCR : somme des carrés résiduels

C'est la somme des carrés de la différence entre chaque point et la moyenne de son propre échantillon.

$SCR=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij}-\overline{x_i}\right)^2$

$$SCR = (3-2)^2 + (2-2)^2 + (1-2)^2 + (2-2)^2$$
$$+ (5-3{,}75)^2 + (3-3{,}75)^2 + (4-3{,}75)^2 + (3-3{,}75)^2$$
$$+(5-3{,}5)^2 + (7-3{,}5)^2 + (1-3{,}5)^2 + (1-3{,}5)^2$$

$$SCR = 31{,}75$$

$$ddl_{SCR} = N - k = 12 - 3 = 9$$

SCE : somme des carrés expliqués

C'est la somme des carrés des écarts entre la moyenne de chaque échantillon et la moyenne générale (répétés pour chaque valeur de l'échantillon)

$$SCE = \sum_{i=1}^{k} n_i\,(\bar{x}_i - \bar{X})^2 \quad = 4(2-3{,}08)^2 + 4(3{,}75-3{,}08)^2$$
$$+ 4(3{,}5-3{,}08)^2$$

$$SCE = 7{,}1667$$

$$ddl_{SCE} = k - 1 = 3 - 1 = 2$$

SCT : somme des carrés totaux

$$SCT = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij} - \bar{X}\right)^2$$

$$= (3-3{,}08)^2 + (2-3{,}08)^2 + (1-3{,}08)^2 + (2-3{,}08)^2 +$$
$$(5-3{,}08)^2 + (3-3{,}08)^2 + (4-3{,}08)^2 + (3-3{,}08)^2 +$$
$$(5-3{,}08)^2 + (7-3{,}08)^2 + (1-3{,}08)^2 + (1-3{,}08)^2$$

$$SCT = 38{,}9167$$

$$ddl_{SCT} = N - 1 = 12 - 1 = 11$$

Vérification

38.9167=SCT= SCE+ SCR=31.75+7.1667 (c'est bon)

$$ddl_{SCT} = ddl_{SCR} + ddl_{SCE}$$

En effet 9+2=11

D- La variable de décision :

Sous H_0, on démontre que

$$f_{calc} = \frac{\frac{SCE}{ddl_{SCE}}}{\frac{SCR}{ddl_{SCR}}} = \frac{\frac{SCE}{k-1}}{\frac{SCR}{N-k}} = \frac{S_F^2}{S_R^2}$$

Suit une loi de Fisher- Snedecor $\mathcal{F}(\text{ddl}_{SCE}; \text{ddl}_{SCR}) = \mathcal{F}(k-1; N-k)$.

Décision

Pour $\propto$ $fix\acute{e}$,on choisit la table de Fisher adéquate, on tire la valeur $\text{f}_{théorique}$ selon les deux d.d.l.s.

- Si $\text{f}_{\text{calc}} < \text{f}_{\text{théorique}}$ H_0 $acceptée$.
- Sinon H_0 $rejetée$.

S_F^2: variance factorielle.

S_R^2: variance résiduelle.

Application numérique :

$$f_{calc} = \frac{\frac{7,1667}{2}}{\frac{31,75}{9}} = 1,016$$

Pour $\alpha = 0,05$; $\text{f}_{\text{théorique}} = \text{F}(2; 9) = 4,26$.

$\text{f}_{\text{calc}} = 1,016 < \text{f}_{\text{théorique}} = 4,26$, donc H_0 acceptée.

Les milieux n'ont pas d'effet sur la production.

Remarque :

Quand H_0 est rejetée, ils existent d'autres tests qui font apparaitre qu'elles sont les moyennes qui différent significativement, quel est le facteur qui fait changer les valeurs. C'est ce qu'on appelle les tests post-hoc.

2. Analyse de la variance à deux facteur :

2.1. Analyse de la variance à deux facteurs (échantillons de plusieurs observations) :

On étudie simultanément deux facteurs A et B

• p modalités pour A (A_1,…, A_p), exemple: (4 différentes doses d'un certain médicament).

• q modalités pour B (B_1,…, B_q) , exemple:(catégorie d'âge :jeune , adultes, âgé).

Pour chaque modalité du couple (A, B), on dispose d'un échantillon de taille n.

	A_1	A_2	…	A_p
B_1	$x_{11,1}; x_{11,2}; \dots ; x_{11,n}$ n observations	$x_{12,1}; x_{12,2}; \dots ; x_{12,n}$ n observations	. . .	$x_{1p,1} ; x_{1p,2} ; \dots ; x_{1p,n}$ n observations
B_2	$x_{21,1}; x_{21,1}; \dots ; x_{21,n}$ n observations	$x_{22,1}; x_{22,2} ; \dots ; x_{22,n}$ n observations	. . .	$x_{2p,1}; x_{2p,2} ; \dots ; x_{2p,n}$ n observations
. . . .	. . . . .	. . . . .	. . .	. . . . .
B_q	xxxxxx n observations	xxxxxx n observations	. . .	xxxxxx n observations

Il s'agit d'un tableau d'ANOVA à deux facteurs croisés avec répétitions.

Conditions d'applications

- Les échantillons sont extraits de populations gaussiennes (loi normale).
- Toutes les populations ont la même variance.
- Tous les échantillons ont la même taille.

Ce qu'on va tester

• L'influence du facteur A seul.

• L'influence du facteur B seul.

• L'influence de l'interaction des deux facteurs.

En fait, ce test va comporter 3 sous tests et par suite on aura 3 hypothèses nulles et 3 hypothèses alternatives.

Les hypothèses

• H_{0A}: le facteur A n'a pas d'influence sur la moyenne des populations c.à.d il n'y a pas de différences entre les moyennes selon la facteur A .
$H_{0A}: \mu_{A_1} = \mu_{A_2} = \cdots = \mu_{A_p}$
• H_{1A} : le facteur A influe sur la moyenne des populations (Il existe au moins deux différentes moyennes selon le facteur A) .

• H_{0B}: le facteur B n'a pas d'influence sur la moyenne des populations C.à.d il n'y a pas de différences entre les moyennes selon la facteur B .
$H_{0B}: \mu_{B_1} = \mu_{B_2} = \cdots = \mu_{B_q}$

• H_{1B}: le facteur B influe sur la moyenne des populations (Il existe au moins deux différentes moyennes selon le facteur B).

• H_{0AB}: il n'y a pas d'interaction entre les deux facteurs,

• H_{1AB}:: il y a interaction entre les deux facteurs.

Les calculs intermédiaires pour déterminer les sources de variations

	A_1	A_2	. . .	A_p	Moyennes lignes
B_1	$\bar{x}_{11}$	$\bar{x}_{12}$	. . .	$\bar{x}_{1p}$	$\bar{x}_{1.} = \frac{1}{p}\sum_{j=1}^{p} \bar{x}_{1j}$
B_2	$\bar{x}_{21}$	$\bar{x}_{22}$	. . .	$\bar{x}_{2p}$	$\bar{x}_{2.} = \frac{1}{p}\sum_{j=1}^{p} \bar{x}_{2j}$
. . .	. . .	. . .	. . .		
B_q	$\bar{x}_{q1}$	$\bar{x}_{q2}$	. . .	$\bar{x}_{qp}$	$\bar{x}_{q.} = \frac{1}{p}\sum_{j=1}^{p} \bar{x}_{qj}$
Moyennes colonnes	$\bar{x}_{.1} = \frac{1}{q}\sum_{i=1}^{q} \bar{x}_{i1}$	$\bar{x}_{.2} = \frac{1}{q}\sum_{i=1}^{q} \bar{x}_{i2}$		$\bar{x}_{.p} = \frac{1}{q}\sum_{i=1}^{q} \bar{x}_{ip}$	$\bar{X} = \frac{1}{npq}\sum_{i=1}^{q}\sum_{j=1}^{p} n\,\bar{x}_{ij}$

$\bar{x}_{ij}$: moyenne de chaque échantillon qui a le caractère B_i et le caractère A_j.
$\bar{x}_{i.}$: moyenne de la ligne i.
$\bar{x}_{.j}$: moyenne de la colonne j.
$\bar{X}$: moyenne générale.

Tableau des variances estimées

	A_1	A_2	...	A_p
B_1	S_{11}^2	S_{12}^2		S_{1p}^2
B_2	S_{21}^2	S_{22}^2		S_{2p}^2
. . .				
B_q	S_{q1}^2	S_{q2}^2		S_{qp}^2

S_{ij}^2: variance estimée de l'échantillon qui a le caractère B_i et le caractère A_j.

La variance résiduelle S_R^2 est la moyenne des variances estimées

$$S_R^2 = \frac{1}{(n-1)pq} \sum_{\substack{1 \le i \le q \\ 1 \le j \le p}} (n-1) S_{ij}^2$$

La variance estimée totale S_T^2 est la variance estimée de tous les échantillons réunis.

Décomposition de la moyenne

$$x_{ijk} - \bar{X} = (\bar{x}_{i.} - \bar{X}) + (\bar{x}_{.j} - \bar{X}) + (\bar{x}_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{X}) + (x_{ijk} - \bar{x}_{ij})$$

Effet du facteur B

Effet du facteur A

Effet de l'interaction

Erreur résiduel (expérimental)

A partir de laquelle on extrait l'équation

SCET = SCEA + SCEB + SCEAB + SCER

Les calculs intermédiaires pour déterminer les sources de variations

• SCEA= somme des carrés des écarts des moyennes des groupes de A par rapport à la moyenne générale.

• SCEB= somme des carrés des écarts des moyennes des groupes de B par rapport à la moyenne générale.

- SCEAB= somme des carrés des écarts dû à l'interaction.
- SCEF : somme des carrés des écarts dû aux facteurs.
- SCER : somme des carrés des écarts résiduels (ni dû à A ni à B).

Somme des carrés des écarts	Formule	Degré de liberté	Carrés moyens
SCEA	$\sum_{j=1}^{p} nq(\bar{x}_{.j} - \bar{X})^2$	$\frac{1}{p-1}$	$S_A^2 = \frac{SCEA}{p-1}$
SCEB	$\sum_{i=1}^{q} np(\bar{x}_{i.} - \bar{X})^2$	$\frac{1}{q-1}$	$S_B^2 = \frac{SCEB}{q-1}$
SCEAB	$\sum_{\substack{1\le i\le q \\ 1\le j\le p}} n(\bar{x}_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{X})^2$	$\frac{1}{(p-1)(q-1)}$	$S_{AB}^2 = \frac{SCEAB}{(p-1)(q-1)}$
SCEF	SCEA+SCEB+SCEAB		On la calcule avec le théorème **de la décomposition de la variance factorielle**
SCER	$\sum_{\substack{1\le i\le q \\ 1\le j\le p}} (n-1){S_{ij}}^2$	$\frac{1}{(n-1)pq}$	$S_R^2 = \frac{SCER}{(n-1)pq}$
SCET	$\sum_{\substack{1\le i\le q \\ 1\le j\le p}} npq(x_{ijk} - \bar{X})^2$	$\frac{1}{npq-1}$	$S_T^2 = \frac{SCET}{npq-1}$

Théorème d'analyse de la variance :

$$(npq-1)S_T^2 = (n-1)pq\ S_R^2 + (pq-1)S_F^2$$

Théorème de la décomposition de la variance factorielle :

$$(pq-1)S_F^2 = (p-1)S_A^2 + (q-1)S_B^2 + (p-1)(q-1)S_{AB}^2$$

Décision

Sous	Variable de décision	Suit une loi	De d.d.l.
$\boldsymbol{H_{0A}}$	$F_A = \frac{S_A^2}{S_R^2} = \frac{SCEA/p-1}{SCER/(n-1)pq}$	Fisher Snedecor	$(p-1;(n-1)pq)$
$\boldsymbol{H_{0B}}$	$F_B = \frac{S_B^2}{S_R^2} = \frac{SCEB/q-1}{SCER/(n-1)pq}$	Fisher Snedecor	$(q-1;(n-1)pq)$
$\boldsymbol{H_{0AB}}$	$F_{AB} = \frac{S_{AB}^2}{S_R^2} = \frac{SCEAB/(p-1)(q-1)}{SCER/(n-1)pq}$	Fisher Snedecor	$((q-1)(p-1);(n-1)pq)$

Pour les 3 hypothèses, on calcule à chaque fois $f(\alpha; d.d.l)$
Si $\boldsymbol{la\ variable\ de\ décision > f(\alpha; d.d.l)}$, $\boldsymbol{H_0}$ est rejetée.

Les étapes de calcul d'une ANOVA à 2 facteurs avec répétition

Après vérifications des conditions d'application et formulation des hypothèses,

Conditions d'application

1. Les variables suivent une loi normale.
2. On a égalité des variances.
3. Tous les échantillons ont la même taille.

On suit les étapes suivantes :

1) Calcul des moyennes de chaque petit échantillon $\bar{x}_{ij} = \frac{1}{n}\sum_{k=1}^{n} x_{ijk}$.

2) Calcul des moyennes lignes $\bar{x}_{i.}$

3) Calcul des moyennes colonnes $\bar{x}_{.j}$

4) Calcul de la moyenne générale $\bar{X}$.

5) Calcul des variances échantillons de chaque échantillon et en déduire les variances estimées $S_{ij}^2 = \frac{n}{n-1} S_{ech\ ij}^2$.

6) En déduire la variance résiduelle :
La variance résiduelle S_R^2 est la moyenne des variances estimées

$$S_R^2 = \frac{1}{(n-1)pq} \sum_{\substack{1 \le i \le q \\ 1 \le j \le p}} (n-1) S_{ij}^2.$$

7) Estimation de la variance totale de la variance totale des échantillons réunis.

8) Calcul de la variance factorielle en utilisant le

Théorème d'analyse de la variance :

$$(npq-1)S_T^2 = (n-1)pq\ S_R^2 + (pq-1)S_F^2$$

9) Calcul de la variance conditionnelle due au facteur A seul

$$S_A^2 = \frac{SCEA}{p-1} = \frac{\sum_{j=1}^{p} nq(\bar{x}_{.j} - \bar{X})^2}{p-1}.$$

10) Calcul de la variance conditionnelle due au facteur B seul

$$S_B^2 = \frac{SCEB}{q-1} = \frac{\sum_{i=1}^{q} np(\bar{x}_{i.} - \bar{X})^2}{q-1}.$$

11) Calcul de la variance conditionnelle due à l'interaction entre A et B, en utilisant le théorème de la décomposition de la variance factorielle :

$$(pq-1)S_F^2 = (p-1)S_A^2 + (q-1)S_B^2 + (p-1)(q-1)S_{AB}^2$$

12) Calcul de variables de décision

Sous H_{0A} (pas d'effet du facteur A)	$F_A = \frac{S_A^2}{S_R^2} \sim \mathcal{F}_\alpha(p-1;(n-1)pq)$
Sous H_{0B} (pas d'effet du facteur B)	$F_B = \frac{S_B^2}{S_R^2} \sim \mathcal{F}_\alpha(q-1;(n-1)pq)$
Sous H_{0AB} (pas d'interaction entre A et B)	$F_{AB} = \frac{S_{AB}^2}{S_R^2} \sim \mathcal{F}_\alpha((p-1)(q-1);(n-1)pq)$

13) Décision :

Si $F > f_\alpha(\nu_1, \nu_2)$ alors H_0 est rejetée.

Exercice : On étudie l'activité d'un enzyme chez des sujets en fonctions de l'âge et du sexe les résultats sont les suivants
L'activité enzymatique moyenne dépend-t-elle de l'âge, du sexe ?

	<12 ans												>12 ans											
Garçon	4,9	2,9	2,7	3,9	4,6	3,3	5,9	4,8	4,1	3,5	7,2	6,1	2,1	2,2	1,1	2,9	5	3,5	2,4	4,4	2,1	3	3,9	5,6
Fille	4,5	6,9	4	5,4	1,9	3,6	4,8	3,3	7,5	5,8	4,4	6	2,4	3,6	4,8	3,9	5,5	5	6,8	2,2	3,1	5	4,1	4,7

Corrigé

On essaye de voir si l'un des facteurs (ou les deux) est (sont) responsable(s) du changement de l'activité de l'enzyme.
Il s'agit donc d'une ANOVA à deux facteurs A : âge, B : sexe.

Conditions d'application

1. Les variables suivent une loi normale.
2. On a égalité des variances.
3. Tous les échantillons ont la même taille.

On notera garçon (G) et fille (F)

Données tirées du tableau

- Les 4 sous échantillons (G, <12), (G,>12), (F<12) et (F>12) ont la même taille $n = 12$, donc la 3ème condition est vérifiée.
- La taille générale de tous les échantillons réunis $N = 48.$
- Le nombre de modalité du caractère A (âge), $p = 2$.
- Le nombre de modalité du caractère B (sexe), $q = 2$.

On supposera que la 1ère et la 2ème condition sont vérifiées.

Les hypothèses

➢ Hypothèses relatives au facteur A

- H_{0A}: le facteur A (âge) n'a pas d'influence sur la moyenne des populations (l'activité enzymatique).

C.à.d il n'y a pas de différences entre les moyennes selon la facteur A.

$$H_{0A}: \mu_{A_1} = \mu_{A_2} = \cdots = \mu_{A_p} .$$

- H_{1A}: le facteur A influe sur la moyenne des populations.
  Il existe au moins deux différentes moyennes selon le facteur A.

➢ Hypothèses relatives au facteur B

- H_{0B}: le facteur B (sexe) n'a pas d'influence sur la moyenne des populations (l'activité enzymatique).

C.à.d il n'y a pas de différences entre les moyennes selon la facteur B.

$$H_{0B}: \mu_{B_1} = \mu_{B_2} = \cdots = \mu_{B_q} .$$

- H_{1B}: le facteur B influe sur la moyenne des populations.
  Il existe au moins deux différentes moyennes selon le facteur B.

➢ Hypothèses relatives à l'interaction entre A et B

- H_{0AB}: il n'y a pas d'interaction entre les deux facteurs (âge et sexe).
- H_{1AB}: il y a interaction entre les deux facteurs.

Calculs de bases

✓ Les différentes moyennes

1) Moyenne de chaque échantillon : $\bar{x}_{ij} = \frac{1}{n}\sum_{k=1}^{n} x_{ijk}$.

	<12 ans												>12 ans											
Garçon	4,9	2,9	2,7	3,9	4,6	3,3	5,9	4,8	4,1	3,5	7,2	6,1	2,1	2,2	1,1	2,9	5	3,5	2,4	4,4	2,1	3	3,9	5,6
Fille	4,5	6,9	4	5,4	1,9	3,6	4,8	3,3	7,5	5,8	4,4	6	2,4	3,6	4,8	3,9	5,5	5	6,8	2,2	3,1	5	4,1	4,7

	<12 ans	>12 ans
Garçon	$\bar{x}_{11} = \frac{4,9 + 2,9 + \cdots + 6,1}{12} = 4,4917$	$\bar{x}_{12} = \frac{2,1 + 2,2 + \cdots + 5,6}{12} = 3,1833$
Fille	$\bar{x}_{21} = \frac{4,5 + 6,9 + \cdots + 6}{12} = 4,8417$	$\bar{x}_{22} = \frac{2,4 + 3,6 + \cdots + 4,7}{12} = 4,2583$

2) Moyennes conditionnelles du facteur A (âge) (moyennes colonnes) :

$\bar{x}_{.1} = \frac{4,4917 + 4,8417}{2} = 4,6667$; $\bar{x}_{.2} = \frac{3,1833 + 4,2583}{2} = 3,7208$.

3) Moyennes conditionnelles du facteur B (sexe) (moyennes lignes) :

$\bar{x}_{1.} = \frac{4,4917 + 3,1833}{2} = 3,8375$; $\bar{x}_{2.} = \frac{4,8417 + 4,2583}{2} = 4,55$.

4) Moyenne générale :

$$\bar{X} = \frac{(4,9 + \cdots + 6,1) + (2,1 + \cdots + 5,6) + (4,5 + \cdots + 6) + (2,4 + \cdots + 4,7)}{48} = 4,1925$$

Ou bien

$$\bar{X} = \frac{\bar{x}_{.1} + \bar{x}_{.2} + \bar{x}_{1.} + \bar{x}_{2.}}{4} = 4,1925 .$$

✓ Les différentes variances

1) Les variances estimées : S_{ij}^2

$$S_{ech11}^2 = \frac{1}{12}(4{,}9^2 + \cdots + 6{,}1^2) - (4{,}4917^2) = 1{,}73576$$

$$\Rightarrow S_{11}^2 = \frac{n}{n-1} S_{ech11}^2 = \frac{12}{11} 1{,}73576 = 1{,}89356.$$

$$S_{ech12}^2 = \frac{1}{12}(2{,}1^2 + \cdots + 5{,}6^2) - (3{,}1833^2) = 1{,}63472$$

$$\Rightarrow S_{12}^2 = \frac{n}{n-1} S_{ech12}^2 = \frac{12}{11} 1{,}63472 = 1{,}78333.$$

$$S_{ech21}^2 = \frac{1}{12}(4{,}5^2 + \cdots + 6^2) - (4{,}8417^2) = 2{,}28909$$

$$\Rightarrow S_{21}^2 = \frac{n}{n-1} S_{ech21}^2 = \frac{12}{11} 2{,}28909 = 2{,}49719.$$

$$S_{ech22}^2 = \frac{1}{12}(2{,}4^2 + \cdots + 4{,}7^2) - (4{,}2583^2) = 1{,}60076$$

$$\Rightarrow S_{22}^2 = \frac{n}{n-1} S_{ech22}^2 = \frac{12}{11} 1{,}60076 = 1{,}74628.$$

2) Variance résiduelle moyenne de variances estimées des petits échantillons : S_R^2

$$S_R^2 = \frac{1}{pq} \sum_{\substack{1 \le i \le 2 \\ 1 \le j \le 2}} S_{ij}^2 = \frac{1}{4}\left(S_{11}^2 + S_{12}^2 + S_{21}^2 + S_{22}^2\right)$$

$$= \frac{1{,}89356 + 1{,}78333 + 2{,}49719 + 1{,}74628}{4}$$

$$S_R^2 = 1{,}98009.$$

3) Variance totale : S_T^2

$$S_{echT}^2 = \frac{(4{,}9^2 + \cdots + 6{,}1^2) + (2{,}1^2 + \cdots + 5{,}6^2) + (4{,}5^2 + \cdots + 6^2) + (2{,}4^2 + \cdots + 4{,}7^2)}{48} - (4{,}1925^2),$$

$$S_{echT}^2 = 2{,}1985.$$

$$S_T^2 = \frac{48}{47} 2{,}1985$$

$$S_T^2 = 2{,}24527.$$

4) Variance totale : S_F^2

D'après le théorème d'analyse de la variance :

$$(npq - 1)S_T^2 = (n-1)pq\, S_R^2 + (pq-1)S_F^2$$

On rappelle que $n = 12, p = 2, q = 2$.

Et qu'on a calculé $S_R^2 = 1{,}98009$ et $S_T^2 = 2{,}24527$.

$47S_T^2 = 44\, S_R^2 + 3S_F^2$ et par suite :

$$S_F^2 = \frac{47(2{,}24527) - 44(1{,}98009)}{3} = 6{,}13457$$

5) Variance conditionnelle due au facteur A seul (âge) : S_A^2

$$S_A^2 = \frac{SCEA}{p-1} = \frac{\sum_{j=1}^{p} nq\left(\bar{x}_{.j} - \bar{X}\right)^2}{p-1} = \frac{12(2)}{1}[(\bar{x}_{.1} - \bar{X})^2 + (\bar{x}_{.2} - \bar{X})^2]$$
$$= 24[(4{,}6667 - 4{,}1925)^2 + (3{,}7208 - 4{,}1925)^2] = 10{,}7368$$

6) Variance conditionnelle due au facteur B seul (sexe) : S_B^2

$$S_B^2 = \frac{SCEB}{q-1} = \frac{\sum_{i=1}^{q} np(\bar{x}_{i.} - \bar{X})^2}{q-1} = \frac{12(2)}{1}[(\bar{x}_{1.} - \bar{X})^2 + (\bar{x}_{2.} - \bar{X})^2]$$
$$= 24[(3{,}8375 - 4{,}1925)^2 + (4{,}55 - 4{,}1925)^2] = 6{,}0919$$

7) Variance conditionnelle due à l'interaction entre A et B :

Théorème de la décomposition de la variance factorielle :

$$(pq-1)S_F^2 = (p-1)S_A^2 + (q-1)S_B^2 + (p-1)(q-1)S_{AB}^2$$

On rappelle que $n = 12, p = 2, q = 2$.

Et qu'on a calculé $S_A^2 = 10{,}7368$, $S_B^2 = 6{,}0919$ et $S_F^2 = 6{,}13457$, en remplaçant, on trouve :

$$3(6{,}13457) = 1(10{,}7368) + 1(6{,}0919) + 1S_{AB}^2$$

$$\Rightarrow S_{AB}^2 = 1{,}57501$$

✓ Les variables de décision

Sous H_{0A} (pas d'effet de l'âge)	$F_A = \frac{S_A^2}{S_R^2} = \frac{10{,}7368}{1{,}98009} = 5{,}422$ $\sim \mathcal{F}_\alpha(p-1;(n-1)pq)$	$\mathcal{F}_{0{,}05}(1;44) = 4{,}05$
Sous H_{0B} (pas d'effet du sexe)	$F_B = \frac{S_B^2}{S_R^2} = \frac{6{,}0919}{1{,}98009} = 3{,}076$ $\sim \mathcal{F}_\alpha(q-1;(n-1)pq)$	$\mathcal{F}_{0{,}05}(1;44) = 4{,}05$
Sous H_{0AB} (pas d'interaction entre âge et sexe)	$F_{AB} = \frac{S_{AB}^2}{S_R^2} = \frac{61{,}57501}{1{,}98009} = 0{,}7954$ $\sim \mathcal{F}_\alpha((p-1)(q-1);(n-1)pq)$	$\mathcal{F}_{0{,}05}(1;44) = 4{,}05$

$\mathcal{F}_{0{,}05}(1;44) = 4{,}05$ approximativement.

✓ Décision

Si $F > f_\alpha(\nu_1, \nu_2)$ alors H_0 est rejetée

$F_A = 5{,}422 > 4{,}05 \Rightarrow H_{0A}$ rejetée	L'âge a une influence significative sur l'action enzymatique au risque de 5%
$F_B = 3{,}076 < 4{,}05 \Rightarrow H_{0B}$ non rejetée	Le sexe n'a pas une influence significative sur l'action enzymatique au risque de 5%
$F_{AB} = 0{,}7954 < 4{,}05 \Rightarrow H_{0AB}$ non rejetée	au risque de 5%, il n'y pas d'interaction entre L'âge et Le sexe.

2.2. Analyse de la variance à deux facteurs (échantillons d'une seule observation) :

Si chaque échantillon ne comporte qu'une seule observation $(n = 1)$, les S_{ij}^2 sont nulles et $S_R^2 = 0$.

Les quotients $\frac{S_A^2}{S_R^2}, \frac{S_B^2}{S_R^2}, \frac{S_{AB}^2}{S_R^2}$ n'ont plus de sens.

Mise en place du test

Théorème d'analyse de la variance devient :

$$(pq-1)S_T^2 = (p-1)\,S_A^2 + (q-1)S_B^2 + (p-1)(q-1)S_{AB}^2$$

et permet d'obtenir S_{AB}^2 à partir de S_A^2 , S_B^2 et S_T^2 .

Décision

Sous	Variable de décision	Suit une loi	De d.d.l.
$\boldsymbol{H_{0A}}$	$F_A = \frac{S_A^2}{S_{AB}^2} = \frac{SCEA/p-1}{SCEAB/(p-1)(q-1)}$	Fisher Snedecor	$(p-1;(p-1)(q-1))$
$\boldsymbol{H_{0B}}$	$F_B = \frac{S_B^2}{S_{AB}^2} = \frac{SCEB/q-1}{SCEAB/(p-1)(q-1)}$	Fisher Snedecor	$(q-1;(p-1)(q-1))$

Pour les 2 hypothèses, on calcule à chaque fois $\boldsymbol{f(\alpha;ddl)}$
Si la variable de décision $\boldsymbol{F > f(\alpha;ddl)}$ alors H0 est rejetée .
Remarque :
H_{0AB} ne peut pas être tester.

Exemple d'application

Cherchant à réaliser une émulsion la plus stable possible, un expérimentateur associe les émulsionnants a, b, c, d aux corps gras α, β, γ. La stabilité des émulsions obtenues avec chacune des 12 associations est notée de 0 à 10 :

	a	b	c	d
α	2	1	3	1
β	2	2	3	2
γ	3	4	5	3

La stabilité est-elle significativement différente, au risque de 2.5%,

- ✓ En fonction du choix du corps gras?
- ✓ En fonction du choix de l'émulsionnant?

Solution

On peut appliquer un test d'Anova à 2 facteurs sans répétition (n=1)

A : corps gras, (p=3 modalités)

B : émulsionnants (q=4 modalités).

✓ Les différentes moyennes

1) Moyennes conditionnelles du facteur A (corps gras) (moyennes lignes) :

$$\bar{x}_{1.} = \frac{2+1+3+1}{4} = 1{,}75 \quad ; \quad \bar{x}_{2.} = \frac{3+2+3+2}{4} = 2{,}5 \quad ; \quad \bar{x}_{3.} = \frac{3+4+5+3}{4} = 3{,}75$$

2) Moyennes conditionnelles du facteur B(émulsionnants) (moyennes colonnes) :

$$\bar{x}_{.1} = \frac{2+3+3}{3} = 2{,}67 \; ; \; \bar{x}_{.2} = \frac{1+2+4}{3} = 2{,}33 \; ; \; \bar{x}_{.3} = \frac{3+3+5}{3} = 3{,}67 \; ;$$

$$\bar{x}_{.4} = \frac{1+2+3}{3} = 2 \; .$$

3) Moyenne générale :

$$\bar{X} = \frac{(2+1+3+1)+(3+2+3+2)+(3+4+5+3)}{12}$$
$$= \frac{1{,}75+2{,}5+3{,}75}{3} = 2{,}67$$

✓ Les différentes variances

1) Variance totale :

$$S^2_{echT} = \frac{(2^2+1^2+\cdots+3^2)}{48} - (2{,}67^2) = 1{,}22,$$

$$S^2_T = \frac{12}{11} S^2_{echT} = 1{,}33.$$

2) Variance conditionnelle due au facteur A seul (corps gras) :

$$\begin{aligned} S^2_A = \frac{SCEA}{p-1} &= \frac{\sum_{i=1}^{p} q(\bar{x}_{i.} - \bar{X})^2}{p-1} \\ &= \frac{4}{2}[(\bar{x}_{1.} - \bar{X})^2 + (\bar{x}_{2.} - \bar{X})^2 + (\bar{x}_{3.} - \bar{X})^2] \\ &= 2[(1{,}75 - 2{,}67)^2 + (2{,}5 - 2{,}67)^2 + (3{,}5 - 2{,}67)^2] \\ &= 4{,}08. \end{aligned}$$

3) Variance conditionnelle due au facteur B seul (émulsionnants) :

$$\begin{aligned} S^2_B = \frac{SCEB}{q-1} &= \frac{\sum_{i=1}^{p} p(\bar{x}_{.j} - \bar{X})^2}{q-1} \\ &= \frac{3}{3}[(\bar{x}_{.1} - \bar{X})^2 + (\bar{x}_{.2} - \bar{X})^2 + (\bar{x}_{.3} - \bar{X})^2 + (\bar{x}_{.4} - \bar{X})^2] \\ &= [(2{,}67 - 2{,}67)^2 + (2{,}33 - 2{,}67)^2 + (3{,}67 - 2{,}67)^2 \\ &+ (2 - 2{,}67)^2] = 1{,}56. \end{aligned}$$

4) Variance conditionnelle due à l'interaction entre A et B :

Théorème d'analyse de la variance devient :

$$(pq-1)S^2_T = (p-1)\,S^2_A + (q-1)S^2_B + (p-1)(q-1)S^2_{AB}$$

et permet d'obtenir S^2_{AB} à partir de S^2_A , S^2_B et S^2_T

$$11(\,1{,}33) = 2(4.08) + 3(1.56) + 6\,S^2_{AB} \;\Rightarrow\; S^2_{AB} = 0{,}31.$$

✓ Variable de décision

H_{0A} (le choix du corps gras n'influe pas la stabilité)	$F_A = \frac{S_A^2}{S_{AB}^2} = \frac{4{,}08}{0{,}31} = 13{,}16$ $\sim \mathcal{F}_\alpha(\boldsymbol{p} - 1; (p-\mathbf{1})(q-\mathbf{1}))$	$\mathcal{F}_{0{,}025}(2;6) = 7{,}26$
$\boldsymbol{H_{0B}}$ (le choix de l'émulsionnant n'influe pas la stabilité)	$F_B = \frac{S_B^2}{S_{AB}^2} = \frac{1{,}56}{0{,}31} = 5{,}03$ $\sim \mathcal{F}_\alpha(q-1; (p-1)(q-1))$	$\mathcal{F}_{0{,}025}(3;6) = 6{,}60$

✓ Décision

Si la variable de décision $\boldsymbol{F > f(\alpha; ddl)}$ alors **H0 est rejetée .**

$F_A = 13{,}16 > 7{,}26 \Rightarrow H_{0A}$ rejetée	Le choix du corps gras influe sur la stabilité significativement au risque de 2.5%
$F_B = 5{,}03 < 6{,}60 \Rightarrow H_{0B}$ non rejetée	L'influence du choix de l'émulsionnant n'est pas significative au risque de 2.5%