

FICHE N° 19

ANALYSE DE VARIANCE
UNE DIMENSION

k échantillons indépendants
Variable numérique.

But : Etudier l'influence d'un facteur sur les résultats d'observations ou de mesures.

Méthode : a) Données et notations

Un facteur peut présenter les modalités M1, M2, M3, ..., Mi, ..., Mk.

Les observations peuvent être alors réparties en k classes correspondant chacune à une modalité du facteur.

Soit x_{ij} la jème observation de la ième classe, n_i le nombre d'observations de cette classe et m_i leur moyenne.

Soit enfin N le nombre de toutes les observations et m leur moyenne

$$N = \sum_i n_i \quad m = \frac{\sum_{ij} x_{ij}}{N}$$

b) Principe de la méthode :

La variance totale est égale à $\frac{\sum_{ij} (x_{ij} - m)^2}{N-1}$

$$\text{Or } x_{ij} - m = (x_{ij} - m_i) + (m_i - m)$$

$$(x_{ij} - m)^2 = (x_{ij} - m_i)^2 + (m_i - m)^2 + 2(x_{ij} - m_i)(m_i - m)$$

$$\sum_j (x_{ij} - m)^2 = \sum_j (x_{ij} - m_i)^2 + \sum_j (m_i - m)^2 + 2 \sum_j (x_{ij} - m_i)(m_i - m)$$

$$m_i - m \text{ ne dépend pas de } j \text{ et } \sum_j (x_{ij} - m_i) = \sum_j x_{ij} - n_i m_i =$$

$$n_i m_i - n_i m_i = 0$$

$$\text{D'où } \sum_j (x_{ij} - m)^2 = \sum_j (x_{ij} - m_i)^2 + n_i (m_i - m)^2$$

$$\text{et } \sum_{ij} (x_{ij} - m)^2 = \sum_{ij} (x_{ij} - m_i)^2 + \sum_i n_i (m_i - m)^2 \quad (1)$$

Si l'on suppose que la variance σ^2 est la même pour l'ensemble des classes

et à l'intérieur de chacune des classes, le 1° membre de l'égalité (1)

est égal à $(N-1) \sigma^2$,

$$\text{D'autre part } \sum_j (x_{ij} - m_i)^2 = (n_i - 1) \sigma^2$$

.../...

Le 1° terme du 2° membre de (1) est alors égal à

$$(n_1-1)\sigma^2 = ((n_1-1)+(n_2-1)+(n_3-1)+\dots+(n_k-1))\sigma^2 \\ = (N-k)\sigma^2$$

Le 2° terme du 2° membre est donc égal à

$$(N-1)\sigma^2 - (N-k)\sigma^2 = (k-1)\sigma^2$$

On a donc trois estimations de la variance.

$$s^2 = \frac{\sum_{ij} (x_{ij} - m)^2}{N-1} = \frac{S_T}{N-1} \quad \text{Variance totale} \\ s_1^2 = \frac{\sum_{ij} (x_{ij} - m_i)^2}{N-k} = \frac{S_1}{N-k} \quad \text{Variance intraclasse (ou résiduelle)} \\ s_2^2 = \frac{\sum_i n_i (m_i - m)^2}{k-1} = \frac{S_2}{k-1} \quad \text{Variance interclasse (due au facteur)}$$

On démontre que les deux dernières estimations sont indépendantes

On forme le rapport
$$F = \frac{s_2^2}{s_1^2}$$

Si la distribution des observations est normale, on peut alors appliquer le test de Snédécour avec $k-1$ et $N-k$ degrés de liberté. Si la valeur de F est supérieure à la valeur lue dans la table au seuil considéré, on peut dire que le facteur est influent (au seuil considéré).

Remarques : 1/ Si $F \ll 1$, il est clair que le facteur n'a pas d'influence.

2/ Si toutes les classes ont le même effectif n

$$N=kn, N-k = k(n-1), \quad \frac{\sum_i n_i (m_i - m)^2}{k-1} = \frac{n \sum_i (m_i - m)^2}{k-1}$$

Calculs : a) $\sum_{ij} (x_{ij} - m)^2 = \sum_{ij} x_{ij}^2 - N m^2$

soit $T = \sum_{ij} x_{ij} = Nm$

La formule devient $S_T = \sum_{ij} x_{ij}^2 - \frac{T^2}{N}$

b) $\sum_i n_i (m_i - m)^2 = \sum_i n_i m_i^2 - Nm^2 = \sum_i n_i m_i^2 - \frac{T^2}{N} = S_2$

dans le cas où les effectifs des classes sont égaux : $n \sum_i m_i^2 - \frac{T^2}{N}$

c) $\sum_{ij} (x_{ij} - m_i)^2$ se calcule par différence en tenant compte de (1)

$$S_1 = S_T - S_2$$

d) On calcule les variances $s_2^2 = \frac{S_2}{k-1}$ $s_1^2 = \frac{S_1}{N-k}$, puis F

.../...

Exemple : Une série de dix opérations arithmétiques a été proposée à des enfants. Certains ont composé dans la matinée avant la récréation (moment 1), d'autres dans la matinée après la récréation (moment 2) d'autres l'après-midi avant la récréation (moment 3) et enfin les derniers l'après-midi après la récréation (moment 4). On se propose d'étudier l'influence du moment de la journée sur les résultats obtenus.

Données : le tableau 1 donne par élève, le nombre d'opérations exactes, les élèves étant classés suivant le moment de la journée où ils ont composé.

	moment 1	moment 2	moment 3	moment 4
	6	8	9	3
	9	7	5	6
	5	6	4	5
	6	5	2	2
	8	1	7	2
	5	3	6	4
	5	4	6	8
	3	4	7	4
	6	6	5	4
	8	6	7	4
			6	1
				2
total	61	50	64	45
effectif	10	10	11	12
Moyenne	6,1	5	5,8	3,75

Tableau I

Calculs : les effectifs n_1 , n_2 , n_3 , et n_4 sont donnés dans le tableau ainsi que les moyennes m_i .

$$N = 10 + 10 + 11 + 12 = 43$$

$$T = 61 + 50 + 64 + 45 = 220$$

Tableau des carrés : tableau 2

moment 1	moment 2	moment 3	moment 4
36	64	81	9
81	49	25	36
25	36	16	25
36	25	4	4
64	1	49	4
25	9	36	16
25	16	36	64
9	16	49	16
36	36	25	16
64	36	49	16
		36	1
			4
401	288	406	211

tableau 2

somme des carrés : 1306.

$$S_T = 1306 - \frac{220^2}{43} = 1306 - 1126 = 180$$

$$n_1 m_1^2 = 10 \times 6,1 \times 6,1 = 372,1$$

$$n_2 m_2^2 = 10 \times 5 \times 5 = 250$$

$$n_3 m_3^2 = 11 \times 5,8 \times 5,8 = 370$$

$$n_4 m_4^2 = 12 \times 3,75 \times 3,75 = 168,75$$

$$S_2 = 372,1 + 250 + 370 + 168,75 - 1126 = 34,85$$

$$S_1 = S_T - S_2 = 146,15$$

$$s_2^2 = \frac{S_2}{4-1} = \frac{34,85}{3} = 11,62$$

$$s_1^2 = \frac{146,15}{43-4} = \frac{146,15}{39} = 3,75$$

$$F = \frac{11,62}{3,75} = 3,1$$

Test : Hypothèse nulle : Le moment n'a pas d'influence sur la distribution des résultats.

Table de Snédécour pour $v_1 = 3$ $v_2 = 39$ (voir fiche n° 9 bis)

à .05 on lit $F = 2,92$ pour $v_2 = 30$ } $v_1 = 3$
 $F = 2,84$ pour $v_2 = 40$ }

Le F calculé 3,1 leur est supérieur.

L'hypothèse nulle peut être rejetée au seuil .05

à .01, on lit $F = 4,51$ pour $v_2 = 30$ } $v_1 = 3$
 $F = 4,31$ pour $v_2 = 40$ }

L'hypothèse nulle ne peut être rejetée au seuil .01.

FICHE N° 20

ANALYSE DE VARIANCE
DEUX DIMENSIONS ET
INTERACTION

k échantillons indépendants

Variable numérique

BUT : Etudier l'influence de deux facteurs et leur interaction sur les résultats
d'observations ou de mesures.

METHODE : 1) Données et notations

Un premier facteur peut présenter les modalités $M_1, M_2, \dots, M_p$.

Un deuxième facteur peut présenter les modalités $M'_1, M'_2, \dots, M'_q$.

Les observations peuvent être réparties en pq classes correspondant chacune à un élément du produit cartésien. La classe correspondant aux modalités M_i et M'_j sera désignée par C_{ij} . La k ème observation de cette classe sera notée x_{ij}^k , le nombre d'éléments de cette classe n_{ij} , la moyenne des observations m_{ij} et on a $n_{ij} m_{ij} = \sum_k x_{ij}^k$ (1).

L'effectif des observations correspondant à la modalité M_i du 1° facteur sera notée n_i et leur moyenne m_i . Nous avons

$$n_i = \sum_j n_{ij} \quad (2)$$

$$n_i m_i = \sum_{kj} x_{ij}^k = \sum_j n_{ij} m_{ij} \quad (3)$$

De même pour la modalité M'_j du second facteur :

$$n_j = \sum_i n_{ij} \quad (4)$$

$$n_j m_j = \sum_{ki} x_{ij}^k = \sum_i n_{ij} m_{ij} \quad (5)$$

Enfin, soit N le nombre total des observations et m leur moyenne

$$N = \sum_{ij} n_{ij} = \sum_i n_i = \sum_j n_j \quad (6)$$

$$Nm = \sum_{kij} x_{ij}^k = \sum_{ij} n_{ij} m_{ij} = \sum_i n_i m_i = \sum_j n_j m_j \quad (7)$$

Remarque : si toutes les classes C_{ij} ont le même effectif n , les formules ci-dessus deviennent :

$$nm_{ij} = \sum_k x_{ij}^k \quad (1') \quad n_i = nq \quad (2')$$

$$nqm_i = \sum_{kj} x_{ij}^k = n \sum_j m_{ij}$$

$$\text{D'où : } qm_i = \sum_j m_{ij} \quad (3')$$

De même $n_j = np$ (4') et $pm_j = \sum_i m_{ij}$ (5')

Et enfin $N = npq$ (6')

$$pqm = \sum_{ij} m_{ij} = q \sum_i m_i = p \sum_j m_j \quad (7')$$

2) Principe de la méthode (Cf fiches 9 bis et 19)

Estimation de la variance totale : $\frac{\sum_{kij} (x_{ij}^k - m)^2}{N-1}$

$$\text{Or } (x_{ij}^k - m) = \underbrace{(x_{ij}^k - m_{ij})}_{A_1} + \underbrace{(m_{ij} - m_i)}_{A_2} + \underbrace{(m_{ij} - m_j)}_{A_3} + \underbrace{(m_{ij} - m_i - m_j + m)}_{A_4} \quad (8)$$

On démontre que quels que soient s et t avec $s \neq t$ $\sum_{kij} A_s A_t = 0$
 D'où $\sum_{kij} (x_{ij}^k - m)^2 = \sum_{kij} (x_{ij}^k - m_{ij})^2 + \sum_{kij} (m_{ij} - m_i)^2 + \sum_{kij} (m_{ij} - m_j)^2 + \sum_{kij} (m_{ij} - m_i - m_j + m)^2$ (9)

$$= \sum_{kij} (x_{ij}^k - m_{ij})^2 + \sum_i n_i (m_i - m)^2 + \sum_j n_j (m_j - m)^2 + \sum_{ij} n_{ij} (m_{ij} - m_i - m_j + m)^2$$

Degrés de liberté : Par un raisonnement analogue à celui de la fiche 19 on montre que les degrés de liberté des cinq termes figurant dans l'égalité (9) sont respectivement (N-1), N-pq, p-1, q-1, (p-1)(q-1).

Nous avons alors cinq estimations de la variance de la population :

$$s^2 = \frac{\sum_{kij} (x_{ij}^k - m)^2}{N-1} \quad \text{variance totale}$$

$$s_1^2 = \frac{\sum_{kij} (x_{ij}^k - m_{ij})^2}{N-pq} \quad \text{variance résiduelle ou intraclasse}$$

$$s_2^2 = \frac{\sum_i (m_i - m)^2}{p-1} \quad \text{variance due au premier facteur}$$

$$s_3^2 = \frac{\sum_j (m_j - m)^2}{q-1} \quad \text{variance due au 2ème facteur}$$

$$s_4^2 = \frac{\sum_{ij} n_{ij} (m_{ij} - m_i - m_j + m)^2}{(p-1)(q-1)} \quad \text{variance due à l'interaction des 2 facteurs}$$

On démontre que les 4 dernières estimations sont indépendantes.

On forme les rapports $F_1 = \frac{s_1^2}{s_1^2}$, $F_2 = \frac{s_3^2}{s_1^2}$, $F_3 = \frac{s_4^2}{s_1^2}$

Si la population est normale on peut alors appliquer le test de Snédécour

.../...

(Fiche 9 bis) F_1 pour l'influence du 1er facteur, F_2 pour le second facteur F_3 pour l'interaction des deux facteurs.

Si toutes les classes ont le même effectif n , les formules (10) deviennent

$$\left. \begin{aligned} s_1^2 &= \frac{\sum_{ijk} (x_{ij}^k - m_{ij})^2}{pq(n-1)} & s_2^2 &= \frac{nq \sum_i (m_i - m)^2}{p-1} \\ s_3^2 &= \frac{np \sum_j (m_j - m)^2}{q-1} & s_4^2 &= \frac{n \sum_{ij} (m_{ij} - m_i - m_j + m)^2}{(p-1)(q-1)} \end{aligned} \right\} \quad (10')$$

Calculs 1) Disposition des données (tableau 1) : les données seront inscrites dans un tableau de p lignes et de q colonnes. Dans la case ij , figureront toutes les observations correspondant aux modalités M_i et M'_j des facteurs.

2) Expressions des carrés.

Nous désignerons par T_{ij} la somme de tous les nombres de C_{ij} , par T_i la somme de tous les nombres de la ligne i , par T_j la somme de tous les nombres de la colonne j et enfin par T la somme de toutes les données.

D'après les formules (10) et (10') nous avons :

$$S = (N-1) s^2 = \sum_{kij} (x_{ij}^k - m)^2 = \sum_{kij} (x_{ij}^k)^2 - \frac{T^2}{N} = \sum_{kij} (x_{ij}^k)^2 - \frac{T^2}{npq}$$

$$S_1 = (N - pq) s_1^2 = \sum_{kij} (x_{ij}^k - m_{ij})^2 = \sum_{kij} (x_{ij}^k)^2 - \sum_{ij} \frac{T_{ij}^2}{n_{ij}} = \sum_{ijk} (x_{ij}^k)^2 - \frac{\sum_{ij} T_{ij}^2}{n}$$

$$S_2 = (p-1) s_2^2 = \sum_i n_i (m_i - m)^2 = \sum_i \frac{T_i^2}{n_i} - \frac{T^2}{N} = \frac{\sum_i T_i^2}{nq} - \frac{T^2}{npq}$$

$$S_3 = (q-1) s_3^2 = \sum_j n_j (m_j - m)^2 = \sum_j \frac{T_j^2}{n_j} - \frac{T^2}{N} = \frac{\sum_j T_j^2}{np} - \frac{T^2}{npq}$$

$$S_4 = (p-1)(q-1) s_4^2 \text{ se calcule par différence en appliquant (9)}$$

$$S_4 = S - S_1 - S_2 - S_3$$

3) Disposition des calculs :

Il faut donc calculer T_{ij} , T_i , T_j , T , leurs carrés ainsi que la somme des carrés de toutes les données. On suppose pour la suite que toutes les cases contiennent le même nombre n de données.

Dans le tableau II, chacune des cases ij contient T_{ij} , dans deux colonnes supplémentaires sont inscrits la somme par ligne T_i puis les carrés T_i^2 ; dans deux

.../...

lignes supplémentaires seront inscrits la somme par colonne T_j puis les carrés T_j^2 . Figurent en outre dans ce tableau, la somme des T_i^2 , celle des T_j^2 , T et T^2 .

Dans le tableau III chacune des cases ij contient T_{ij}^2 et on inscrit la somme S_{III} de tous les carrés.

Dans le tableau IV, chacune des cases ij contient la somme des carrés des nombres de la case ij du tableau I. On fait leur somme

$$\sum_{kij} (x_{ij}^k)^2 = S_{IV}$$

$$\text{On calcule } \frac{T^2}{npq}, \quad \frac{\sum_{ij} T_{ij}^2}{n} = \frac{S_{III}}{n}, \quad \frac{\sum_i T_i^2}{nq}, \quad \frac{\sum_j T_j^2}{np}$$

On remplit le tableau des résultats (Tableau V)

Tableau des résultats				
Source de variation	Somme des carrés	ddl	Estimation Variance	F
Résidus	$S_1 = S_{IV} - \frac{S_{III}}{n}$	$pq(n-1)$	$s_1^2 = \frac{S_1}{pq(n-1)}$	
1er facteur	$S_2 = \frac{\sum_i T_i^2}{nq} - \frac{T^2}{npq}$	$p-1$	$s_2^2 = \frac{S_2}{p-1}$	$F_1 = \frac{s_2^2}{s_1^2}$
2ème facteur	$S_3 = \frac{\sum_j T_j^2}{np} - \frac{T^2}{npq}$	$q-1$	$s_3^2 = \frac{S_3}{q-1}$	$F_2 = \frac{s_3^2}{s_1^2}$
Interaction	$S_4 = S - S_1 - S_2 - S_3$	$(p-1)(q-1)$	$s_4^2 = \frac{S_4}{(p-1)(q-1)}$	$F_3 = \frac{s_4^2}{s_1^2}$
Total	$S = S_{IV} - \frac{T^2}{npq}$	$npq-1$	$s^2 = \frac{S}{npq-1}$	

Exemple : Dans une expérience sur les probabilités à l'école élémentaire, des groupes d'enfants ont fait des pronostics (ou paris) sur le résultat des tirages.

Après une première série de tirages, chaque groupe a donné dix pronostics pour la seconde série. Cette série de tirages étant effectuée, les enfants ont modifié ou non leurs pronostics et ainsi de suite jusqu'à la cinquième série de pronostics.

On se propose d'étudier si les paris varient

- avec le rang des paris (de 1 à 5) 1° facteur
- avec le groupe d'enfants (5 groupes) 2° facteur
- s'il y a interaction des deux facteurs, c'est-à-dire si l'évolution des paris varie d'un groupe à l'autre.

Les calculs sont faits dans les tableaux I, II, III, IV et le tableau des résultats ci-après.

Conclusions : F_1 calculé : 9,36

F lu pour $v_1 = 4$ et $v_2 = 120$ à .01 : 3,48

On peut rejeter l'hypothèse nulle au seuil .01 : les paris faits par les enfants varient suivant leur rang.

F_2 calculé : 20,92

F lu : le même que ci-dessus

On peut rejeter l'hypothèse nulle au seuil .01

Les paris émis diffèrent significativement suivant les groupes

F_3 calculé : 7,78

F lu pour $v_1 = 12$, $v_2 = 120$ à .01 : 2,34

On peut encore rejeter l'hypothèse nulle au seuil .01.

L'évolution des paris varie avec les groupes.

TABLEAU 1

		Numéro du groupe				
		1	2	3	4	5
rang des paris	1	31,29,29,31,30 33,30,33,36,35	51,54,48,49,54 53,49,48,39,31	44,24,35,15,39 24,24,35,39,39	32,32,41,40,41 40,40,41,41,40	40,41,43,38,38 32,34,33,39,34
	2	38,37,30,32,31 32,38,38,37,30	38,39,37,35,33 34,36,32,40,41	38,37,35,38,34 33,39,38,35,36	38,38,38,39,39 39,41,41,41,32	31,34,41,38,31 34,33,26,39,32
	3	41,38,41,40,42 30,38,39,38,37	38,39,35,40,36 34,36,41,37,41	41,43,42,44,45 47,48,49,41,35	41,41,41,35,35 42,42,36,36,35	42,41,32,43,38 32,26,27,29,39
	4	38,38,38,40,41 41,36,34,35,40	38,33,39,37,38 35,40,38,41,38	23,33,26,38,42 15,32,24,24,23	38,38,38,38,36 36,36,40,40,33	26,14,29,19,26 31,23,24,26,17
	5	40,41,40,38,39 37,39,38,40,41	40,39,40,38,38 39,35,36,40,38	41,40,42,41,42 24,42,20,25,41	45,45,45,41,41 41,22,22,34,35	22,23,30,21,19 40,26,27,29,24

TABLEAU II $T_{ij}, T_i, T_j, T_i^2, T_j^2, \sum_i T_i^2, \sum_j T_j^2, T^2$

numéro des groupes

	1	2	3	4	5	T_i	T_j^2
Rang des paris							
1	317	476	318	388	372	1871	3 500 641
2	343	365	363	386	339	1796	3 225 616
3	384	377	435	384	349	1929	3 721 041
4	381	377	280	373	235	1646	2 709 316
5	393	383	358	371	261	1766	3 118 756
$\sum_j$	1818	1978	1754	1902	1556	$T=9008$	$\sum_i T_i^2 =$
$\sum_j$	3 305 124	3 912 484	3 076 516	3 617 604	2 421 136	$\sum_j T_j^2 =$	16 275 370
							16 332 864

$$T^2 = 81\ 144\ 064$$

TABLEAU IV $\sum_k (x_{ij}^k)^2, \sum_{kij} (x_{ij}^k)^2$

numéro des groupes

	1	2	3	4	5	Total
Rang des paris						
1	10 103	23 134	10 902	15 172	13 964	73 275
2	11 879	13 405	13 213	14 962	11 669	65 128
3	14 848	14 269	19 075	14 838	12 553	75 583
4	14 571	14 261	8 432	13 953	5 781	56 998
5	15 461	14 695	13 536	14 467	7 137	65 296
Total	66 862	79 764	65 158	73 392	51 104	336 280

TABLEAU III : $T_{ij}^2, \sum_{ij} T_{ij}^2 = S_{III}$

-70-

Numéro des groupes

	1	2	3	4	5	Somme T_{ij}^2
1	100 489	226 576	101 124	150 544	138 384	717 117
2	117 649	133 225	131 769	148 996	114 921	646 560
3	147 456	142 129	189 225	147 456	121 801	748 067
4	145 161	142 129	78 400	139 129	55 225	560 044
5	154 449	146 689	128 164	137 641	68 121	635 064
Somme T_{ij}^2	665 204	790 748	628 682	723 766	498 452	$S_{III} =$ 3 306 852