

E. MORICE

Quelques tests non paramétriques

Journal de la société statistique de Paris, tome 99 (1958), p. 254-284

http://www.numdam.org/item?id=JSFS_1958__99__254_0

© Société de statistique de Paris, 1958, tous droits réservés.

L'accès aux archives de la revue « Journal de la société statistique de Paris » (<http://publications-sfds.math.cnrs.fr/index.php/J-SFdS>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
<http://www.numdam.org/>

VIII VARIÉTÉS

Quelques tests non paramétriques

Au cours des quinze dernières années, de nombreuses études ont été publiées, dans les revues statistiques de langue anglaise, sur les tests non paramétriques.

Basés, en général, sur de simples classements et non sur la valeur numérique des observations, leur efficacité est assez réduite.

Cependant dans un test d'homogénéité portant sur deux échantillons — l'hypothèse à tester étant la possibilité pour ces deux échantillons d'appartenir à une même population — ils ont l'avantage de ne faire intervenir aucune hypothèse sur la distribution dans cette population, alors que les tests classiques de comparaison de paramètres impliquent, en particulier pour de petits échantillons, la normalité de cette distribution.

A ce titre, ils peuvent fournir des éléments d'information intéressants.

Leur application dans les problèmes classiques de jugement sur échantillons est sans doute assez limitée; il paraît cependant intéressant de signaler rapidement les plus importants de ces tests et de donner aux lecteurs une bibliographie détaillée qui permettra de se reporter aux travaux originaux.

* * *

PREMIÈRE PARTIE

TESTS BASÉS SUR LA SOMME OU LA MOYENNE DES RANGS

I

Généralités

Considérons N nombres entiers consécutifs $1, 2, \dots, N$, et soit un échantillon de n (tiré sans remise), de somme

$$S = x_1 + \dots + x_i + \dots + x_n = \sum_1^n x_i$$

Le nombre total des échantillons de n nombres pouvant être obtenus est

$$k = C_N^n = \frac{N!}{n!(N-n)!}$$

La valeur moyenne de la somme est :

$$E(S) = \bar{S} = \frac{\sum_1^k \sum_1^n x_i}{k}$$

la notation $\left(\sum_i^k\right)$ indiquant, pour éviter les doubles indices, que la sommation $\sum x_i$ doit être répétée dans les k échantillons possibles.

Étant donné qu'il y a k échantillons possibles de n nombres, chacun contenant une fraction n/N de la population, chaque nombre x_i , pris parmi les N nombres, figurera $k n/N$ fois dans les divers échantillons possibles.

On a donc :

$$E(S) = \bar{S} = \frac{K \frac{n}{N} \sum_1^N x_i}{k} = \frac{n}{N} \sum_1^N x_i$$

$$E(S) = \bar{S} = \frac{n}{N} \frac{N(N+1)}{2} = \frac{n(N+1)}{2}$$

La variance des x dans la population finie des N valeurs de moyenne :

$$m = \frac{1}{N} \sum_1^N x_i$$

est

$$\sigma^2 = \frac{1}{N} \sum_1^N (x - m)^2 = \frac{1}{N} \left[\sum_1^N x^2 - \frac{\left(\sum_1^N x\right)^2}{N} \right]$$

$$\sigma^2 = \frac{1}{N} \left[\frac{N(N+1)(2N+1)}{6} - \frac{1}{N} \left(\frac{N(N+1)}{2} \right)^2 \right]$$

$$\sigma^2 = \frac{N^2 - 1}{12}$$

L'échantillon de n valeurs étant tiré sans remise, la variance de sa moyenne \bar{x} est

$$V(\bar{x}) = \sigma_{\bar{x}}^2 = \frac{N - n}{N - 1} \frac{\sigma^2}{n}$$

et la variance de la somme $S = n\bar{x}$ est

$$V(S) = \sigma_S^2 = n^2 \sigma_{\bar{x}}^2 = n \frac{N - n}{N - 1} \sigma^2 = \frac{n(N+1)(N-n)}{12}$$

II

COMPARAISON DE DEUX ÉCHANTILLONS OU DE PLUSIEURS SÉRIES DE DEUX ÉCHANTILLONS DE MÊME EFFECTIF — TEST DE WILCOXON

A — Comparaison de deux échantillons de n observations

Dans le cas particulier où $n = \frac{N}{2}$, on a

$$E(S) = \bar{S} = \frac{N(N+1)}{4} = \frac{n(2n+1)}{2}$$

$$V(S) = \sigma_S^2 = \frac{N}{2} \frac{2N - N}{2(N-1)} \frac{N^2 - 1}{12} = \frac{N}{12} \frac{N(N+1)}{4}$$

$$V(S) = \sigma_S^2 = \frac{N}{12} \bar{S} = \frac{n}{6} \frac{n(2n+1)}{2} = \frac{n^2}{12} (2n+1)$$

La distribution de la somme S tend rapidement vers la normale de sorte que les sommes correspondant aux seuils 0,05 — 0,02 et 0,01 peuvent être approximativement estimées au moyen des formules

$$S_{.,.5} = \bar{S} \pm 1,96 \sqrt{N \bar{S}/12} = \frac{n}{2} \left(2n + 1 \pm 1,96 \sqrt{\frac{2n + 1}{3}} \right)$$

$$S_{.,02} = \bar{S} \pm 2,326 \sqrt{N \bar{S}/12}$$

$$S_{.,01} = \bar{S} \pm 2,576 \sqrt{N \bar{S}/12}$$

qui doivent s'interpréter comme suit :

$P \sim 0,05$ pour que S soit extérieur à l'intervalle $\bar{S} \pm 1,96 \sigma S$

$P \sim 0,025$ pour que $S < \bar{S} - 1,96 \sigma$, ou $S > \bar{S} + 1,96 \sigma$,

Pour n petit, Wilcoxon a calculé des tables exactes de la distribution de S , que l'on trouvera ci-après.

Application

Considérons deux échantillons de $n \frac{N}{2} =$ observations

$$x_1, x_2, \dots, x_n$$

$$y_1, y_2, \dots, y_n$$

Si l'on veut tester l'hypothèse que les deux échantillons appartiennent à une même population ($m. \sigma$), on est conduit, dans le cas de grandeurs mesurables, à étudier la variable normale réduite.

$$t = \frac{d - 0}{\sigma_d}$$

$$d = \bar{x} - \bar{y}$$

$$\sigma_d^2 = \sigma_x^2 + \sigma_y^2 = \sigma^2 \left(\frac{1}{n} + \frac{1}{n} \right)$$

σ^2 , en général inconnu, étant estimé par

$$s^2 = \frac{\sum (x - \bar{x})^2 + \sum (y - \bar{y})^2}{2n - 2}$$

la distribution de t est celle de Student-Fisher avec $\nu = 2n - 2$ degrés de liberté ou, pour n grand (pratiquement pour $n > 15$), la distribution de Laplace-Gauss).

Le test de Wilcoxon basé sur la somme des rangs permettra d'éviter ces calculs; il s'imposera d'ailleurs si, les $2n$ observations non mesurables peuvent simplement être classées par ordre de préférence avec les rangs

$$1, 2, \dots, 2n$$

Il y a lieu évidemment de noter que dans l'application d'une telle méthode, on remplace la comparaison basée sur les valeurs de grandeurs mesurables

$x_1, \dots, x_n, y_1, y_n$, par l'étude du classement de ces 2 n valeurs, il en résulte une perte d'information qui peut être importante : il s'agit alors simplement d'une approximation rapide de la signification des différences constatées.

Par contre, l'emploi d'un test de comparaison de rangs du type envisagé ne suppose pas la normalité de la distribution de la population d'où l'échantillon est tiré et ceci peut être dans certains cas un avantage important.

Exemple

Deux variétés de blé (X) et (Y) ont été comparées au point de vue rendement à l'aide de deux échantillons de 10 observations portant sur des parcelles de même surface :

	RÉCOLTES		Différence $d_i = x_i - y_i$
	Blé (X) x_i	Blé (Y) y_i	
	980 (4)	900 (1)	80
	1 305 (20)	1 245 (17)	60
	960 (3)	910 (2)	50
	1 140 (13)	1 220 (16)	— 80
	1 130 (12)	1 010 (6)	120
	1 080 (9)	990 (5)	90
	1 090 (10)	1 030 (7)	60
	1 270 (19)	1 170 (14)	100
	1 250 (18)	1 200 (15)	50
	1 100 (11)	1 060 (8)	40
Moyennes:	1 130,5	1 073,5	57

La méthode classique (en admettant l'hypothèse de distributions approximativement normales et de même variance σ^2), nous conduit, si on veut tester l'hypothèse de deux populations de même moyenne, à calculer :

$$\bar{x} = 1130,5 \qquad \bar{y} = 1073,5$$

$$d = \bar{x} - \bar{y} = 57$$

σ^2 estimé par :

$$s^2 = \frac{\Sigma (x - \bar{x})^2 + (\Sigma y - \bar{y})^2}{18} = 14845,8$$

$$s = 121,8$$

σ_d^2 estimé par :

$$s_d^2 = \sqrt{s^2 \left(\frac{1}{10} + \frac{1}{10} \right)}$$

$$s_d = s \sqrt{\frac{1}{5}}$$

$$t = \frac{d}{s_d} = \frac{57}{121,8} \sqrt{5} = 1,05$$

Pour $\nu = 18$ degrés de liberté, la table de Student montre que la probabilité d'un écart égal ou supérieur en valeur absolue à celui qui a été constaté est de l'ordre de 0,30 : la différence constatée n'est pas significative.

Si on utilise la méthode des rangs (le classement global des 20 valeurs est indiqué entre parenthèses dans le tableau ci-dessus), la somme des rangs pour

le blé (Y) est $S = 91$, alors que la valeur moyenne dans l'hypothèse nulle (classement aléatoire) est

$$E(S) = \bar{S} = \frac{n(2n+1)}{2} = 105$$

avec

$$\sigma_s = \frac{n}{2} \sqrt{\frac{2n+1}{3}} = 5\sqrt{7}$$

On a alors

$$|t| = \frac{105 - 91}{5\sqrt{7}} = \frac{14\sqrt{7}}{35} = 1,06$$

qui conduit à la même conclusion que ci-dessus.

On obtient évidemment la même conclusion si l'on considère le blé X pour lequel la somme des rangs (119) diffère de la valeur moyenne de la même quantité (mais en sens opposé que la somme des rangs du blé Y).

Tables de Wilcoxon. — Les tables de *Wilcoxon* donnent pour les petites valeurs de n , les valeurs de $S = S_0$ ayant une probabilité $P(0,01 - 0,02 - 0,05)$, telle que $P(n, S \leq S_0) = \frac{P}{2}$ (voir table I).

TABLE I. — TEST de WILCOXON

Valeurs critiques *inférieures* des sommes de rangs aux seuils
0,05 — 0,02 — 0,01

n	$P = 0,05$	$P = 0,02$	$P = 0,01$
6	18	16	15
7	27	24	23
8	37	34	32
9	49	46	43
10	63	59	56
11	79	74	71
12	97	91	87
13	116	110	105
14	137	130	125
15	160	152	147
16	185	176	170
17	212	202	196
18	241	230	223
19	271	259	252
20	303	291	282
20	338	324	316

Exemple : Pour $n = 12$, $P(S \leq 116) \neq 0,025$

La loi normale donne

$$\bar{S} = \frac{n(2n+1)}{2} = 150$$

$$\sigma_s^2 = \frac{n}{6} \bar{S} = 300 \quad \sigma_s = 10\sqrt{3} \neq 17,3$$

$P \neq 0,95$ pour que $150 - 1,96 \times 17,3 < S < 150 + 1,96 \times 17,3$

soit $116 < S < 184$

Sauf si n est très petit, ces approximations sont excellentes dès que $n + m$ atteint une trentaine d'unités.

Pour les petites valeurs de n et m , l'étude de la distribution de S a été faite par Colin White [6].

Soit :

$W_s(n, m)$ le nombre de manières d'obtenir la somme S lorsqu'il y a n observations dans le groupe pour lequel on calcule S et m observations dans l'autre groupe.

Les valeurs de $W_s(n, m)$ ont été calculées pour $n = 2 (1) \dots 15$, $m = 4 (1) \dots 28$ et pour les valeurs successives de S entre

$$\frac{n(n+1)}{2} \quad \text{et} \quad \frac{n(n+2m+1)}{2}$$

Ainsi, par exemple, pour $n = 5$, $m = 18$, on obtient les résultats suivants :

Somme des rangs S	Valeurs de $W_s(5,18)$	Valeurs cumulées
15	1	1
16	1	2
17	2	4
18	3	7
.....
33	141	769
34	163	932
.....
102	3	33 645
103	2	33 648
104	1	33 647
105	1	33 649 = C_{23}^5

Étant donné que $33\,649 \times 0,025 = 841$ et que

$$769 < 841 < 932$$

769 et 932 correspondant respectivement à 2,29 % et 2,77 %, on peut donc dire que $S = 33$ est à moins d'une unité près la valeur de S telle que

$$P(S \leq 33) = 0,023 \neq 0,025$$

$S = 33$ est la valeur critique inférieure au seuil $P = 0,05$
(Voir tables III, IV et V.)

Afin de limiter les tables, on utilisera dans chaque cas :

Soit la somme S , somme des rangs pour le plus petit groupe, obtenue en donnant le rang 1 à la plus petite des $n + m$ valeurs,

soit la somme conjuguée S' obtenue pour le même groupe en donnant le numéro un à la plus grande des $n + m$ valeurs.

(Il peut arriver que $S' = n(n + m + 1) - S$ soit inférieur à S .)

La distribution des sommes S étant symétrique, la probabilité d'obtenir une certaine valeur de S est la même que celle d'obtenir la valeur conjuguée S' correspondante.

La valeur critique inférieure au seuil 5 %, par exemple, est donc valable soit que l'on considère S , soit que l'on considère S' .

Exemple : $n = k = 7$

$P(S \leq 327) \neq 0,025$

La loi normale donne :

$$\bar{S} = \frac{kn(2n+1)}{2} = 367,5 \quad \sigma_s = \frac{n}{2} \sqrt{\frac{k(2n+1)}{3}} = 20,7$$

$P \neq 0,95$ pour que $367,5 - 1,96 \times 20,7 < S < 367,5 + 1,96 \times 20,7$
 $327 < S < 408$

III

COMPARAISON DE DEUX OU PLUSIEURS ÉCHANTILLONS D'EFFECTIFS DIFFÉRENTS

A — Comparaison de deux échantillons d'effectifs différents

1 — Test de Colin White

Considérons maintenant le cas de deux échantillons d'effectifs différents

$$\begin{array}{ll} x_1, \dots, x_n & n \text{ observations} \\ y_1, \dots, y_m & m \text{ observations} \end{array}$$

(n étant l'effectif du plus petit échantillon)

Les $N = n + m$ observations étant classées par ordre de grandeur et leurs rangs dans cet ensemble étant

$$1, 2, \dots, n + m$$

soit S_x la somme des rangs des n observations de l'échantillon (x).

Cette quantité S_x peut varier entre les valeurs

$$\begin{array}{ll} \frac{n(n+1)}{2} & \text{plus petite valeur} \\ \frac{n(n+2m+1)}{2} & \text{plus grande valeur} \end{array}$$

D'après les calculs présentés au début, la valeur moyenne de S_x , somme des rangs des x , est

$$E(S_x) = \bar{S}_x = \frac{n(n+m+1)}{2}$$

et sa variance est :

$$\begin{aligned} V(S_x) &= \frac{1}{12} n m (n + m + 1) \\ \sigma_s &= \sqrt{\frac{n m (n + m + 1)}{12}} \end{aligned}$$

La distribution de S tend vers la normale lorsque n et m augmentent, les seuils (ou valeurs critiques) correspondant aux probabilités 0,05 — 0,02 — 0,01 étant définis comme ci-dessous :

$$\begin{aligned} S_{,05} &= \bar{S} \pm 1,96 \sigma_s \\ S_{,02} &= \bar{S} \pm 2,326 \sigma_s \\ S_{,01} &= \bar{S} \pm 2,576 \sigma_s \end{aligned}$$

B — Comparaison portant sur plusieurs séries de deux échantillons.

Supposons que l'on dispose de k groupes de deux échantillons de n valeurs x_i, y_i ($i = 1, \dots, n$), non appariées (expériences répétées).

Pour chaque groupe d'observations (non appariées) donnant lieu à un classement de $2n$ rangs, on a pour le total S

$$E(S) = \bar{S} = \frac{n(2n+1)}{2}$$

$$V(S) = n \frac{\bar{S}}{6} = \frac{n^2}{12} (2n+1)$$

et pour le total des k groupes

$$E S_{(k)} = \bar{S}_{(k)} = \frac{kn(2n+1)}{2}$$

$$V [S_{(k)}] = \frac{kn^2}{12} (2n+1)$$

la distribution de $S_{(k)}$ autour de $\bar{S}_{(k)}$ étant approximativement normale.

Tables

F. Wilcoxon [3] (1) a calculé des tables donnant pour diverses valeurs de n (ou de n et k) les valeurs de $S = S_0$ ayant une probabilité $P(0,01 - 0,02 - 0,05)$, telle que

$$P(n, k, S \leq S_0) = P/2$$

(Voir table III.)

TABLE II. — TEST de WILCOXON

*Valeurs critiques inférieures des sommes de rangs
dans le cas de k groupes de n observations*

n	k = 2		k = 3		k = 4		k = 5		k = 6		k = 7	
	T	P	T	P	T	P	T	P	T	P	T	P
2	—	0,01	9	0,01	13	0,01	17	0,01	22	0,01	26	0,01
	—	0,02	10	0,02	14	0,02	18	0,02	23	0,02	27	0,02
	6	0,05	11	0,05	15	0,05	19	0,05	24	0,05	28	0,05
3	13	0,01	21	0,01	30	0,01	39	0,01	49	0,01	58	0,01
	14	0,02	22	0,02	31	0,02	41	0,02	50	0,02	59	0,02
	15	0,05	24	0,05	33	0,05	42	0,05	52	0,05	62	0,05
4	24	0,01	39	0,01	54	0,01	70	0,01	86	0,01	102	0,01
	25	0,02	40	0,02	56	0,02	72	0,02	88	0,02	105	0,02
	26	0,05	42	0,05	58	0,05	75	0,05	91	0,05	108	0,05
5	38	0,01	61	0,01	85	0,01	110	0,01	135	0,01	160	0,01
	39	0,02	63	0,02	88	0,02	112	0,02	138	0,02	163	0,02
	42	0,05	66	0,05	91	0,05	116	0,05	142	0,05	168	0,05
6	55	0,01	89	0,01	124	0,01	159	0,01	195	0,01	231	0,01
	57	0,02	92	0,02	127	0,02	162	0,02	198	0,02	235	0,02
	61	0,05	96	0,05	131	0,05	168	0,05	204	0,05	241	0,05
7	77	0,01	123	0,01	170	0,01	218	0,01	266	0,01	314	0,01
	79	0,02	126	0,02	174	0,02	222	0,02	270	0,02	319	0,02
	83	0,05	131	0,05	179	0,05	228	0,05	277	0,05	327	0,05

(1) Les numéros entre [] correspondent aux numéros de la bibliographie donnée en fin d'article.

Exemple :

$$n = 4 \quad x_i = 25, 33, 43, 45$$

$$m = 14 \quad y_i = 15, 16, 16, 17, 20, 22, 22, 23, 28, 28, 30, 30, 35, 35.$$

Dans le classement global,

$$15, 16, \dots, 43, 45$$

les rangs relatifs aux quatre observations du premier groupe sont :

$$9, 14, 17, 18$$

$$S = 58$$

alors que

$$S' = 4(4 + 14 + 1) - 58 = 18.$$

TABLE III — TEST de COLIN WHITE

Valeurs critiques inférieures des sommes de rangs au seuil 5 %

(Si les deux groupes sont d'effectifs inégaux, n est l'effectif du plus petit).

$m \backslash n$	2	3	4	5	6	7	8	9	10	11	12	13	14	15
4			10											
5		6	11	17										
6		7	12	18	26									
7		7	13	20	27	36								
8	8	8	14	21	29	38	49							
9	8	8	15	22	31	40	51	63						
10	8	9	15	23	32	42	53	65	78					
11	4	9	16	24	34	44	55	68	81	96				
12	4	10	17	26	35	46	58	71	85	99	115			
13	4	10	18	27	37	48	60	73	88	103	119	137		
14	4	11	19	28	38	50	63	76	91	106	123	141	160	
15	4	11	20	29	40	52	65	79	94	110	127	145	164	185
16	4	12	21	31	42	54	67	82	97	114	131	150	169	
17	5	12	21	32	43	56	70	84	100	117	135	154		
18	5	13	22	33	45	58	72	87	103	121	139			
19	5	13	23	34	46	60	74	90	107	124				
20	5	14	24	35	48	62	77	93	110					
21	6	14	25	37	50	64	79	95						
22	6	15	26	38	51	66	82							
23	6	15	27	39	53	68								
24	6	16	28	40	55									
25	6	16	28	42										
26	7	17	29											
27	7	17												
28	7	17												

Exemple : $P \neq 2,5 \%$ pour que $S(n = 8, m = 17) \leq 70$

Au delà de la table, on peut admettre la distribution normale de S autour de :

$$\bar{S} = \frac{1}{2} n (n + m + 1) \quad \text{avec} \quad \sigma_s = \sqrt{\frac{nm(n + m + 1)}{12}}$$

Pour $n = 4$, $m = 14$, les tables montrent que :

- la valeur critique inférieure à 0,05 est 19;
- la valeur critique inférieure à 0,01 est 14 :

$$P(S \leq 19) = 0,025$$

si l'hypothèse nulle est vérifiée (classement dû aux simples circonstances aléatoires et non à une différence significative).

Correction de continuité et correction pour ex æquo

Les valeurs possibles de S étant obligatoirement des entiers compris entre $\frac{n}{2}(n+1)$ et $\frac{n}{2}(2m+n+1)$, il y a lieu, pour plus de précision, dans l'utilisation de la loi normale (loi continue) ci-dessus d'augmenter ou de diminuer S de $\frac{1}{2}$ (selon que S est inférieure ou supérieure à $\frac{n}{2}(n+m+1)$).

De plus, dans le cas de classements ex æquo, si chacun des s entiers $x+1, x+2, \dots, x+s$ est remplacé par $x + \frac{1}{2}(s+1)$, leur somme n'est pas changée, mais la somme des carrés est diminuée de :

$$\sum_{i=1}^s (x+i)^2 - s \left[x + \frac{s+1}{2} \right]^2 = \frac{(s-1)s(s+1)}{12} = \frac{T}{12}$$

La moyenne des carrés, et par conséquent la variance, est diminuée de $\frac{T}{12(m+n)}$.

Cette corrélation est négligeable si le nombre des ex æquo est petit par rapport à $m+n$.

Dans ce cas, les valeurs possibles de S ne sont plus les entiers successifs de :

$$\frac{n(n+1)}{2} \quad \text{à} \quad \frac{n}{2}(2m+n+1)$$

et la correction de continuité n'a qu'une valeur approximative.

Calcul exact

Si n et $m = N-n$ sont petits, on peut calculer exactement la probabilité que S soit inférieur ou égal à la valeur observée S_0 si l'hypothèse nulle est vérifiée.

En effet, il y a $K = \frac{N!}{n!(N-n)!}$ classements possibles, on peut alors écrire et dénombrer ceux, en nombre k , pour lesquels on a $S \leq S_0$. Par symétrie, il y en a aussi k pour lesquels $S \geq n(N+1) - S_0$.

La probabilité d'un désaccord aussi ou plus grand que celui constaté est donc $\frac{2k}{K}$.

Exemple :

$$N = 9, \quad n = 4, \quad m = 5, \quad S_0 = 12$$

1 — L'emploi de la variable normale, sans correction de continuité donne :

$$t = \frac{2 \times 12 - 4 \times 10}{\sqrt{1/3 \times 4 \times 10 \times 5}} = -1,96 \quad P = 0,05$$

2 — Avec la correction de continuité, on a :

$$t = \frac{2 \times 12,5 - 4 \times 10}{\sqrt{1/3 \times 4 \times 10 \times 5}} = -1,84 \quad P = 0,066$$

3 — Enfin, le calcul direct donne :

$$K = \frac{9!}{4!5!} = 126$$

Les classements de 4 observations pour lesquels $S \leq 12$ sont :

1	2	3	4	S = 10
1	2	3	5	11
1	2	3	6	12
1	2	4	5	12
d'où $k = 4$				

$$P = \frac{2 \times 4}{126} = 0,063$$

2 — Test U de Mann et Whitney

Le même problème a été étudié de manière légèrement différente par Mann et Whitney.

Considérons encore les deux échantillons

x_1, \dots, x_n	n observations
y_1, \dots, y_m	m observations

Les $N = n + m$ valeurs étant ordonnées et leurs rangs étant

$$1, 2, \dots, n + m$$

soient :

S_y la somme des rangs des y dans la série ainsi ordonnée

U_{yx} le nombre de fois qu'un y précède un x .

Ainsi, par exemple, dans le classement $x y y x x x y y y y x$, on a :

$$U_{yx} = 4 + 4 + 5 = 13$$

Ces deux quantités sont liées par relation simple

$$U_{yx} = m n + \frac{m(m+1)}{2} - S_y$$

De même, le nombre de fois qu'un x précède un y est

$$U_{xy} = m n + \frac{n(n+1)}{2} - S_x$$

On a d'ailleurs

$$U_{yx} + U_{xy} = mn$$

Les variables S_x , S_y et U_{xy} , U_{yx} ne diffèrent donc que par un changement d'origine — Leurs distributions ne diffèrent que par les valeurs moyennes, les variances étant identiques.

La distribution de U_{yx} , par exemple, est caractérisée par les valeurs suivantes :

$$E(U_{yx}) = \bar{U}_{yx} = \frac{nm}{2}$$

$$V(U_{yx}) = \frac{nm(n+m+1)}{12} = V(S_y)$$

On remarquera que U_{yx} et U_{xy} ont même moyenne et même variance. Pour n et m supérieurs à 8, la distribution de

$$U - 1/2 (nm + 1) \quad (1)$$

est une distribution sensiblement normale de moyenne nulle et d'écart-type

$$\sqrt{\frac{nm(n+m+1)}{12}}$$

Mann et Whitney ont calculé des tables donnant les probabilités d'obtenir une valeur de U inférieure ou égale à celle fournie par la comparaison des échantillons.

Ces tables ne diffèrent de celles de Colin White que par le fait que (pour différentes combinaisons de m et n avec $m \leq n \leq 8$), elles donnent les probabilités en fonction des valeurs de U alors que les tables de Colin White donnent les valeurs de S en fonction de certains seuils de probabilités.

Nota. — On remarquera que les valeurs critiques inférieures fournies pour U , par ces tables, correspondent aux valeurs critiques supérieures des valeurs de S des tests précédents, en raison de la liaison

$$U = mn + \frac{m(m+1)}{2} - S$$

Pour réduire ces tables au minimum, elles ne sont données ci-après que pour les valeurs de U inférieures à sa moyenne $\bar{U} = \frac{mn}{2}$ et pour $m \leq n$.

Ceci suppose que dans les notations ci-dessus, si on calcule U_{yx} (nombre de fois où y précède x), la variable y est celle pour laquelle U est inférieure à sa valeur moyenne, sinon on utilisera U_{xy} d'ailleurs défini par :

$$U_{yx} + U_{xy} = mn$$

Étant donné la relation

$$P_{m,n}(U) = P_{nm}(U)$$

les indices m et n peuvent être permutés dans l'emploi des tables de Mann et Whitney.

(1) $\frac{mn}{2} + \frac{1}{2}$ pour tenir compte de la correction de continuité dans l'approximation d'une loi relative à une variable discrète, par une loi continue.

Exemple : Étant donné les deux échantillons :

(Y) 14 — 25 — 30 — 32 — 40 — 41 — 43 — 45 $m = 8$
 (X) 12 — 16 — 19 — 20 — 24 — 27 — 35 $n = 7$

qui donnent lieu au classement global

12 — 14 — 16 — 19 — 20 — 24 — 25 — 27 — 30 — 32 — 35
 — 40 — 41 — 43 — 45

TABLE VI [1]. — TEST de MANN et WHITNEY

Probabilité d'obtenir, dans la comparaison d'échantillons d'effectif n et m ,
 une valeur de U inférieure ou égale à celle donnée par les tables
 (Probabilités en millièmes)

$m = 3$

$U \backslash n$	1	2	3
0	250	100	050
1	500	200	100
2	750	400	200
3		600	350
4			500
5			650

$m = 4$

$U \backslash n$	1	2	3	4
0	200	087	028	014
1	400	133	057	029
2	600	267	114	057
3		400	200	100
4		600	314	171
5			429	243
6			571	348
7				448
8				557

$m = 5$

$U \backslash n$	1	2	3	4	5
0	167	047	018	008	004
1	333	095	036	016	008
2	500	190	071	032	016
3	667	286	125	056	028
4		429	196	095	048
5		571	286	143	075
6			393	206	111
7			500	278	155
8			607	365	210
9				452	274
10				548	345
11					421
12					500
13					579

$m = 6$

$U \backslash n$	1	2	3	4	5	6
0	143	036	012	005	002	001
1	286	071	024	010	004	002
2	428	143	048	019	009	004
3	571	241	083	033	015	008
4		321	131	057	026	013
5		429	190	086	041	021
6		571	274	129	063	032
7			357	176	089	047
8			452	238	123	066
9			548	305	165	090
10				381	214	120
11				457	268	155
12				545	331	197
13					396	242
14					465	294
15					535	350
16						409
17						469
18						531

(1) Au delà des tables, on peut admettre la distribution normale de U autour de :

$$\bar{U} = 1/2 (m n + 1) \text{ avec}$$

$$\sigma_u = \sqrt{\frac{m n (m + n + 1)}{12}}$$

TABLE VII. — TEST de MANN et WHITNEY

*Probabilité d'obtenir, dans la comparaison d'échantillon d'effectif n et m, une
une valeur de U inférieure ou égale à celle donnée par les tables*

(Probabilités en millièmes)

$m = 7$

U \ n	1	2	3	4	5	6	7
0	125	028	008	003	001	001	000
1	250	056	017	006	003	001	001
2	375	111	033	012	005	002	001
3	500	167	058	021	009	004	002
4	625	250	092	036	015	007	003
5		333	133	055	024	011	006
6		444	192	092	037	017	009
7		556	258	115	053	026	013
8			333	158	074	037	019
9			417	200	101	051	027
10			500	264	134	069	036
11			583	324	172	090	049
12				394	216	117	064
13				464	265	147	082
14				538	319	183	104
15					378	223	130
16					438	267	159
17					500	314	191
18					562	365	228
19						418	267
20						473	310
21						527	355
22							402
23							451
24							500
25							549

TABLE VIII. — TEST de MANN et WHITNEY

*Probabilité d'obtenir dans, la comparaison d'échantillons d'effectif n et m,
une valeur de U inférieure ou égale à celle donnée par les tables*

(Probabilités en millièmes)

$m = 8$

$\frac{n}{U}$	1	2	3	4	5	6	7	8
0	111	022	006	002	001	000	000	000
1	222	044	012	004	002	001	000	000
2	333	089	024	008	003	001	001	000
3	444	133	042	014	005	002	001	001
4	556	200	067	024	009	004	002	001
5		287	097	036	015	006	003	001
6		356	139	055	023	010	005	002
7		444	188	077	033	015	007	003
8		556	248	107	047	021	010	005
9			315	141	064	030	014	007
10			387	184	085	041	020	010
11			461	230	111	054	027	014
12			539	285	142	071	036	019
13				341	177	091	047	025
14				404	217	114	060	032
15				467	262	141	076	041
16				533	311	172	095	052
17					362	207	116	065
18					416	245	140	080
19					472	286	168	097
20					528	331	198	117
21						377	232	139
22						426	268	164
23						475	306	191
24						525	347	221
25							389	253
26							433	287
27							478	323
28							522	360
29								399
30								439
31								480
32								502

Le calcul direct de U_{yx} donne :

$$U_{yx} = 6 + 2 + 2 + 10$$

En effet :

14 précède 6 valeurs de y

25 précède 2 valeurs de y

30 précède 1 valeur de y

32 précède 1 valeur de y

Total : 10

Les rangs de y , en italique ci-dessus, étant

2, 7, 9, 10, 12, 13, 14, 15

on a d'ailleurs $S_y = 82$

d'où

$$U_{yx} = mn + \frac{m(m+1)}{2} - S_y = 56 + \frac{72}{2} - 82 = 10$$

Pour (m,n) ou $(n,m) = (7, 8)$ la table de Mann et Whitney donne :

$$P(U \leq 10) = 0,020$$

$$P(U \leq 11) = 0,027$$

Comparaison avec la table de Colin White

Dans cette table, n représente l'effectif du plus petit échantillon.

Dans l'exemple précédent, pour $n = 7$, $m = 8$ (le plus petit échantillon étant celui des x), la table de Colin White donne :

$$S_x = 38 \text{ soit } S_y = 120 - 38 = 82$$

telle que

$$P = 0,025 \text{ pour que } S_y \leq 82$$

Comparaison avec la loi normale

Dans ce cas, l'utilisation de la loi normale comme loi limite donnerait les résultats suivants :

Pour U , valeurs critiques au seuil 5 %

$$\begin{aligned} U &= \frac{1}{2}(mn + 1) \pm 1,96 \sqrt{\frac{mn(m+n+1)}{12}} \\ &= 28,5 \pm 1,96 \times 8,64 \\ &= 11,5 \text{ et } 45,5 \end{aligned}$$

B — Généralisation : comparaison de k échantillons — Test de Kruskal et Wallis

Considérons maintenant le cas de k échantillons et soit à tester l'hypothèse qu'ils proviennent de populations identiques.

Kruskal et Wallis [5] ont proposé un test généralisant ceux de Wilcoxon et Colin White.

Considérons d'abord comme ci-dessus deux échantillons de n observations x et $m = N - n$ observations y .

On a vu dans les généralités que dans l'hypothèse d'un échantillon de n nombres entiers tirés sans remise de l'ensemble des N premiers nombres entiers, le rang moyen $\bar{R} = \frac{S}{n}$ relatif à cet échantillon était caractérisé par :

$$\begin{aligned} E(\bar{R}) &= \frac{N+1}{2} \\ V(\bar{R}) &= \frac{1}{n} \frac{N-n}{N-1} \frac{N^2-1}{12} = \frac{(N+1)(N-n)}{12n} \end{aligned}$$

Si n n'est pas trop petit, on peut considérer la variable :

$$t = \frac{R - \frac{1}{2}(N+1)}{\sqrt{\frac{1}{12n}(N+1)(N-n)}} = \frac{2S - n(N+1)}{\sqrt{\frac{n}{3}(N+1)(N-n)}}$$

comme une variable normale réduite, son carré t^2 est donc distribué comme χ^2 avec $\nu = 1$ degré de liberté.

Si on considère k échantillons, dont les moyennes de rangs dans l'ensemble ordonne des $N = n_1 + \dots + n_k$ observations sont liées par la relation

$$n_1 \bar{R}_1 + \dots + n_k \bar{R}_k = \frac{1}{2} N (N + 1)$$

les rangs moyens de $k - 1$ d'entre-eux — pourvu qu'aucun n_i ne soit trop petit — sont distribués approximativement selon une distribution normale à $k - 1$ variables dont l'exposant multiplié par $- 2$, indépendant des $k - 1$ échantillons considérés, est approximativement distribué comme χ^2 ($\nu = k - 1$).

La valeur de cette expression est, dans le cas où il n'y a pas d'ex-æquo :

$$H = \frac{12}{N(N+1)} \sum_{i=1}^k \frac{S_i^2}{n_i} - 3(N+1)$$

$$= \frac{N-1}{N} \sum_{i=1}^k n_i \left[\bar{R}_i - \frac{1}{2}(N+1) \right]^2$$

$$\frac{1/12 (N^2 - 1)}{}$$

avec

k nombre des échantillons

n_i nombre d'observations de l'échantillon (i)

$N = n_1 + n_2 + \dots + n_k$

S_i la somme des rangs de l'échantillon (i)

\bar{R}_i la moyenne de ces n_i rangs

H étant distribué comme χ^2 ($\nu = k - 1$)

Lorsqu'il y a des observations ex æquo, on donne à chacune de ces observations un rang égal à la moyenne de leurs rangs.

Dans ce cas, il y a lieu de diviser H donné ci-dessus par

$$1 - \frac{\sum T}{N(N^2 - 1)}$$

(Pour chaque groupe de s rangs ex æquo, on a $T = (s - 1) s (s + 1)$)

Kruskal et Wallis donnent l'exemple suivant [5] :

Dans une usine, trois machines fabriquent des capsules de bouteilles, le tableau ci-après donne les productions de ces trois machines pour diverses observations de même durée :

	Machines					
	1		2		3	
	Production	Rang	Production	Rang	Production	Rang
	340	5	339	4	347	10
	345	9	333	2	343	7
	330	1	344	8	349	11
	342	6			355	12
	338	3				
Sommes	$\sum n$	5	8	4		
12	$\sum S$	24	14	40		
78	$\sum S^2$					
580,533	$\frac{\sum S^2}{n}$	115,2	65,333	400		

$$H = \frac{12 \times 580,533}{12 \times 13} - 3 \times 13 = 5,656$$

$$P [\chi^2 > 5,656 \text{ v } = 2] = 0,059$$

L'analyse classique de variance appliquée aux productions, à l'aide du test de Fischer-Snedecor donnerait, pour le rapport des variances entre classes et intra-classes, $F (v_1 = 2, v_2 = 9) = 4,2884$ correspondant à une probabilité $P = 0,05$ d'être dépassée.

Dans le cas de trois échantillons, Kruskal et Wallis ont (pour les valeurs de n_1, n_2, n_3 inférieures à 6), calculé les valeurs de H correspondant à des probabilités voisines de 0,10, 0,05 et 0,01.

Pour l'exemple précédent, ces tables donnent :

$$P = 0,049 \text{ au lieu de } P = 0,059$$

donné ci-dessus par la table de χ^2 .

Pour $n_1 = n_2 = n_3$, l'approximation donnée par χ^2 donne pour P une erreur par excès inférieure à 0,01.

DEUXIÈME PARTIE

TESTS BASÉS SUR LE NOMBRE ET LA LONGUEUR DES SUITES HOMOGÈNES

Considérons l'un particulier des rangements de N éléments dont n appartiennent à l'espèce (A) et m à l'espèce (B) : étude d'une alternative. Soit, par exemple, le rangement suivant de 8 (A) et 7 (B) :

b a b b b b a b a a b a a a

Étant donné une séquence d'éléments (A), (B), nous appellerons suite homogène toute séquence partielle constituée par un ou plusieurs éléments de la même espèce.

Ainsi, la séquence ci-dessus comprend 8 suites homogènes : une de 1 (B), une de 1 (A), une de 4 (B), ..., et enfin une de 4 (A).

I

TESTS BASÉS SUR LE NOMBRE DES SUITES HOMOGÈNES

Il apparaît que le nombre total de suites homogènes, que l'on peut trouver dans un ensemble de cette espèce, peut donner une bonne indication sur le caractère plus ou moins aléatoire du rangement.

On peut donc envisager un test basé sur le fait que le nombre total de suites homogènes que l'on constate a une probabilité petite d'être inférieur (ou supérieur) à celui que l'on pourrait raisonnablement espérer si le rangement était simplement aléatoire.

Soit R le nombre de suites homogènes constaté dans le rangement de $N = n + m$ éléments (A) et (B).

La distribution du nombre de suites homogènes a été étudiée par divers auteurs : Mood, Mosteller, Sved et Eisenhart, Stevens, Wald et Wolfowitz (Cf références dans la bibliographie annexée).

Les résultats essentiels de ces travaux sont les suivants :

Si m (ou n) est supérieur à 20, n (ou m), n'étant pas trop petit (au moins 10), la distribution de R est approximativement normale, avec les paramètres :

$$\text{Moyenne} \quad \frac{2 m n}{m + n} + 1$$

$$\text{Variance} \quad \frac{2 m n (2 m n - m - n)}{(m + n)^2 (m + n - 1)}$$

Pour les petites valeurs de m et n , Sved et Eisenhart ont calculé des tables (voir Tables IX et X) donnant les valeurs critiques de R pour les différents seuils.

On trouvera ci-après un extrait de ces tables pour $R_{0,025}$ et $R_{0,975}$.

TABLE IX

Valeurs critiques $R_{0,025}$ du nombre total de suites homogènes

$P [R \leq R_{0,025}] \neq 0,025$

$m \backslash n$	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
5	2	3	3	3	3	3	4	4	4	4	4	4	4	5	5	5
6	3	3	3	3	4	4	4	4	5	5	5	5	5	6	6	6
7	3	3	3	4	4	5	5	5	5	6	6	6	6	7	7	7
8	3	3	4	4	5	5	5	6	6	6	6	6	7	7	7	7
9	3	4	4	5	5	5	6	6	6	7	7	7	7	8	8	8
10	3	4	5	5	5	6	6	7	7	7	7	8	8	8	8	9
11	4	4	5	5	6	6	7	7	7	8	8	8	9	9	9	9
12	4	4	5	6	6	7	7	7	8	8	8	9	9	9	10	10
13	4	5	5	6	6	7	7	8	8	9	9	9	10	10	10	10
14	4	5	5	6	7	7	8	8	9	9	9	10	10	10	11	11
15	4	5	6	6	7	7	8	8	9	9	10	10	11	11	11	12
16	4	5	6	6	7	8	8	9	9	10	10	11	11	11	12	12
17	4	5	6	7	7	8	9	9	10	10	11	11	11	12	12	13
18	5	5	6	7	8	8	9	9	10	10	11	11	12	12	13	13
19	5	6	6	7	8	8	9	10	10	11	11	12	12	13	13	13
20	5	6	6	7	8	9	9	10	10	11	12	12	13	13	13	14

Tables extraites de Sved et Eisenhart : « *Tables for testing randomness of grouping in a sequence of alternatives* » — Ann. Math. Stat. — Vol. 14 — P. 66 (Valeurs arrondies à l'unité).

TABLE X

Valeurs critiques $R_{0,975}$ du nombre total de suites homogènes

$$P [R \geq R_{0,975}] \neq 0,025:$$

$n \backslash m$	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
5	10	10	11	11												
6	10	11	12	12												
7	11	12	13	13	14	14	14	14	15	16						
8	11	12	13	14	14	15	15	16	16	16	17					
9			14	14	15	16	16	16	17	17	18	18	18	18	18	18
10			14	15	16	16	17	17	18	18	18	19	19	19	20	20
11		13	14	15	16	17	17	18	19	19	19	20	20	20	21	21
12		13	14	16	16	17	18	19	19	20	20	21	21	21	22	22
13		13	15	16	17	18	19	20	20	21	21	22	22	23	23	23
14		13	15	16	17	18	19	20	20	21	22	22	23	23	24	24
15			15	16	18	18	19	20	21	22	22	23	23	24	24	25
16				17	18	19	20	21	21	22	23	23	24	25	25	25
17				17	18	19	20	21	22	23	23	24	25	25	26	26
18				17	18	19	20	21	22	23	24	25	25	26	26	27
19				17	18	20	21	22	23	23	24	25	26	26	27	27
20				17	18	20	21	22	23	24	25	25	26	27	27	28

Tables extraites de Sved et Eisenhart (*loc. cit.*)
(Valeurs arrondies à l'unité,)

On y voit, par exemple, que pour

$$n = 8$$

$$m = 7$$

on a :

$$P (R \leq 4) \neq 0,025$$

$$P (R \geq 13) \neq 0,025$$

Au seuil considéré ($P = 0,05$), si le nombre des suites homogènes dépasse 4 et est inférieur à 13, il n'y a donc pas lieu de considérer le rangement comme différent d'un classement aléatoire : par exemple, il n'y aura pas lieu de s'étonner d'une telle succession de parties de jeu de pile (P) ou face (F).

Applications : suites de valeurs inférieures ou supérieures à la médiane.

L'étude du cas particulier $m = n = N/2$ peut être interprétée de la façon suivante :

Dans une séquence de N observations numériques, soient

(A) celles dont la valeur est inférieure à la médiane,

(B) celles dont la valeur est supérieure à la médiane,

leur effectif commun étant alors :

$$m = n = \frac{N}{2} \text{ (N étant supposé pair).}$$

Si on désigne encore par R le nombre total de suites homogènes des (a) et des (b), la distribution de R est caractérisée par :

$$\text{Moyenne } \frac{N+2}{2}$$

$$\text{Variance } \frac{N}{4} \left(1 - \frac{1}{N-1} \right) \approx \frac{N-1}{4}$$

Pour les petites valeurs de $n = \frac{N}{2}$, Sved et Eisenhart ont calculé la table suivante :

TABLE XI

Table des valeurs critiques du nombre total R des suites au-dessus et au-dessous de la médiane (1)

$m = n = \frac{N}{2}$	Valeurs critiques inférieures		Valeurs critiques supérieures	
	$R_{0,05}$	$R_{0,1}$	$R'_{0,05}$	$R'_{0,01}$
5	2	2	8	10
6	3	2	10	11
7	3	3	11	13
8	4	4	12	14
9	5	4	13	15
10	6	5	15	16
11	7	6	16	17
12	7	6	17	19
13	8	7	18	20
14	9	8	19	21
15	10	9	20	22
16	11	10	22	23
18	12	11	24	26
20	14	13	26	28
22	16	14	28	31
24	18	16	31	33
26	20	18	33	36
28	22	19	36	38
30	24	21	37	40

$P [R \leq R_{0,05}] = 0,05$
 $P [R \geq R'_{0,01}] = 0,01$

Pour $n = 30$ $N = 60$
 l'emploi de la loi normale donnerait

$$\bar{R} = \frac{62}{2} = 31 \quad V(R) = 15 \left(1 - \frac{1}{59} \right) = 14,746$$

$$\sigma = 3,84$$

d'où les valeurs

$$R_{0,05} = 31 - 1,64 \times 3,84 = 24$$

$$R_{0,05} = 31 + 1,64 \times 3,84 = 36$$

$$R'_{0,01} = 31 + 2,33 \times 3,84 = 40$$

$$R'_{0,01} = 31 - 2,33 \times 3,84 = 21,$$

valeurs voisines de celles données par la table.

(1) Extrait des tables de Sved et Einsehart (valeurs arrondies à l'unité).

Pour $n \geq 30$, on peut considérer cette distribution comme approximativement normale.

Dans les mêmes conditions, les distributions de

R_a nombre total des suites de (a)

R_b nombre total des suites de (b)

sont approximativement normales avec

$$\text{Moyenne } \frac{N + 2}{4}$$

$$\text{Variance } \frac{N + 2}{16} \left(1 - \frac{1}{N - 1} \right) \approx \frac{N + 1}{16}$$

Exemple : Un exemple de test basé sur la longueur des suites homogènes sera donné plus loin (P. 34).

II

TESTS BASÉS SUR LA LONGUEUR DES SUITES HOMOGÈNES

Si on désigne par R_{ak} (ou R_{bk}) le nombre de suites homogènes de (a) [ou de (b)] de longueur égale ou supérieure à k , la distribution de R_{ak} a été étudiée par Olmstead qui a donné des tables de

$$P [R_{ak} \geq 1]$$

en fonction de k pour $n = 10, 20$ et 40 (cf. [11]).

Pour $K \geq 8$ et N grand, on a sensiblement

$$E [R_{ak}] = V [R_{ak}] = \frac{N}{2^{k+1}}$$

de sorte que la distribution de R_{ak} est approximativement une distribution de Poisson :

$$P [R_{ak} = u] = e^{-\lambda} \frac{\lambda^u}{u!}$$

$$\text{avec } \lambda = \frac{N}{2^{k+1}}$$

de sorte que

$$P [R_{ak} \geq 1] = 1 - P [R_{ak} = 0] = 1 - e^{-\lambda}$$

Les résultats ci-dessus peuvent être appliqués à des échantillons provenant d'une population d'une variable continue en classant comme :

- (a) toute valeur inférieure à la médiane,
- (b) toute valeur supérieure à la médiane.

-Si N , nombre total de valeurs, est impair, $N = 2n + 1$, la médiane qui est la $(n + 1)^{\text{e}}$ observation sera laissée de sorte qu'il y aura toujours n valeurs (a) et n valeurs (b).

L'hypothèse d'une situation statistiquement contrôlée, ne manifestant pas d'anomalies significatives en delà et en deçà du niveau médian peut donc être testée à un double point de vue :

a) à l'aide du nombre total de suites (un nombre trop petit ou trop grand étant significatif);

b) à l'aide des longueurs des suites, l'arrivée de très longues suites étant significative.

Exemple (1) :

Deux solutions de concentration différente d'une même substance ont été placées l'une au-dessus de l'autre dans un mélangeur, la solution la plus concentrée ayant été versée la première. Après mélange, un robinet, placé à la partie inférieure du mélangeur, a permis de recueillir à intervalles réguliers 58 éprouvettes successives dont on a mesuré la concentration.

Les chiffres successifs sont donnés dans le tableau ci-après :

38	33	29	16	44	21	16	17	19	1
22	28	22	14	7	13	21	15	34	23
15	19	32	24	14	13	22	8	30	11
15	24	26	14	11	25	17	10	19	5
6	16	7	10	1	5	2	8	14	14
15	16	13	11	9	11	19	21		

Un mélange insuffisamment réalisé pourra être testé par la présence d'une concentration trop élevée, en général, pour les premières observations, trop basse pour les dernières.

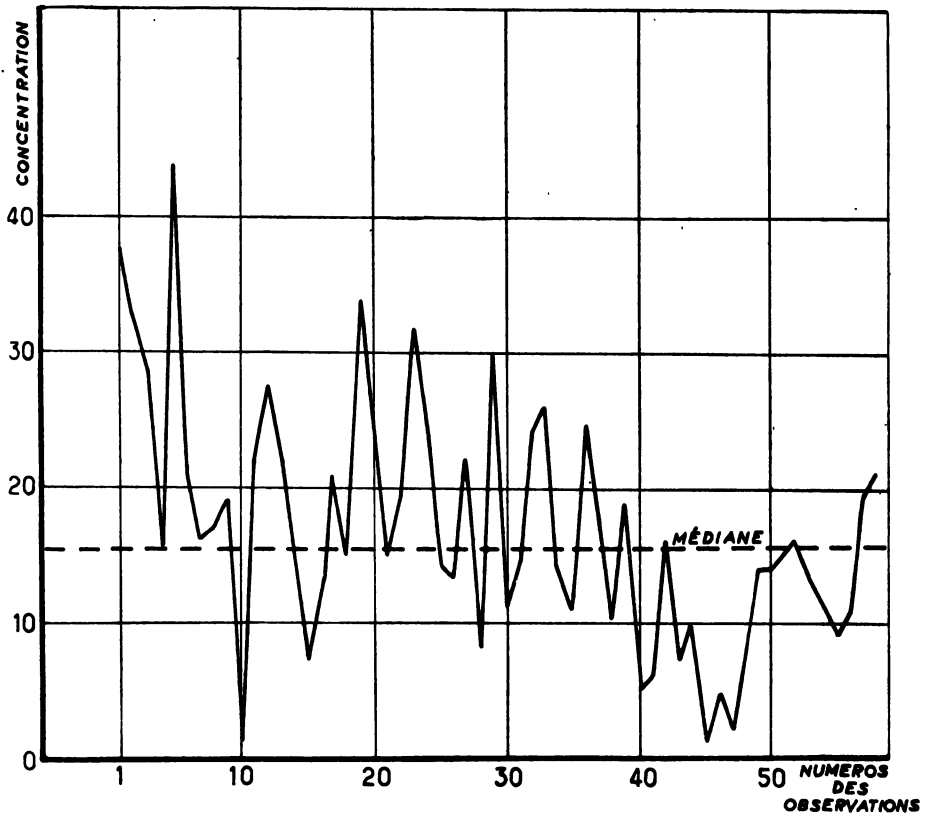
La médiane étant comprise entre 15 et 16, l'évolution de la concentration des éprouvettes successives est mise en évidence par le graphique ci-après qui semble montrer une certaine tendance décroissante.

La distribution des suites, au-dessus ou au dessous de la médiane, selon leur longueur, est donnée dans le tableau ci-après (cf. graphique I) :

DISTRIBUTION DES SUITES

Longueur de la suite l	Nombre observé			Nombre moyen théorique		
	en dessous r_{ai}	en dessus r_{bi}	Total r_i	en dessous ou en dessus $r_{ai} = r_{bi}$	Total \bar{r}_i	nombres cumulés $\bar{R}_k - \bar{R}_{k-1}$
1	5	6	11	7,6	15,2	15,0
2	4	4	8	8,8	7,6	7,4
3	1	2	3	1,9	3,8	3,6
4	1	0	1	0,9	1,8	1,7
> 5	1	1	2	0,8	1,6	0,8
TOTAL. . .	12	13	25	15,0	30,0	

(1) Tiré de Hald [13].



Graphique I. — Observations successives du degré concentration.

1 — Le nombre total des suites est $R = 25$, inférieur à sa valeur moyenne $\bar{R} = 30$.

On a :

$$V(R) = \frac{58 - 1}{4} = 14,25$$

$$\sigma_R = 3,8$$

$$t = \frac{25 + 1/2 - 30}{3,8} = -1,18 \quad (1)$$

Cet écart n'est pas significatif.

2 — Mais, d'autre part, on constate l'existence de deux suites de longueur $k = 9$, une de chaque côté de la médiane.

D'après ce qui précède, on a

$$P[R_{a,0} \geq 1] = 1 - e^{-\lambda}$$

$$\hat{\lambda} = \frac{N}{2^k + 1} = \frac{58}{2^{10}} = 0,0566$$

$$P[R_{a,0} \geq 1] = 0,055$$

(1) $+1/2$, correction de continuité pour tenir compte de ce que R est une variable discrète.

On aurait de même

$$P (R_{n_0} \geq 2) = 1 - e^{-\lambda} - \lambda e^{-\lambda} = 0,002$$

L'existence de deux suites de longueur 9, l'une de valeurs supérieures à la médiane au début des observations, l'autre de valeurs inférieures à la médiane vers la fin des observations, permet donc de conclure qu'il est vraisemblable que le mélange a été mal effectué, une part importante de la solution la plus concentrée étant restée dans la partie inférieure du mélangeur.

Remarques.

1 — La validité des conclusions ci-dessus est basée sur l'application du test envisagé aux nombres successifs du tableau ci-dessus : elle dépend donc partiellement de la précision des observations chiffrées.

Cette observation d'ordre général — valable dans toute étude statistique portant sur une documentation numérique représentant une réalité complexe que l'on veut juger — est particulièrement importante dans le cas présent.

Le simple fait, par exemple, de remplacer la 3^e mesure 16 par 15 supprimerait la première suite de 9 observations supérieures à la médiane et la remplacerait par trois suites de 3, 1 et 5 observations (alors que, par exemple, la moyenne générale des 58 observations, où la pente moyenne descendante du nuage de points seraient très peu modifiées).

L'emploi du test global (nombre de suites ou suites de longueur excessive) doit être interprété avec beaucoup d'esprit critique en y associant tous autres éléments d'information (par exemple, dans le cas actuel, la pente moyenne du nuage dont la signification peut être aussi testée [12]).

2 — Le tableau précédent donne pour (A) et (B) les nombres observés des suites des diverses longueurs et permet, au moins qualitativement, la comparaison avec leurs valeurs moyennes, dans le cas de distribution aléatoire.

Si on désigne par :

r_{ai} (ou r_{bi}), le nombre des suites de (a) (ou de b), de longueur i

r_i le nombre total des suites de (a) et (b), longueur i

$R_{ak} = \sum_{i=k}^N r_{ai}$, le nombre des suites de (a), de longueur égale ou supérieure à k

$R_{bk} = \sum_{i=k}^N r_{bi}$

R_k le nombre total des suites de (a) et (b) de longueur égale ou supérieure à k ,

l'analyse combinatoire permet de calculer les valeurs moyennes de ces diverses quantités (1).

Les valeurs moyennes de ces quantités sont données par les formules suivantes :

$$M (r_{ai}) = M (r_{bi}) = n \frac{(n+1) n \dots (n-i+1)}{N (N-1) \dots (N-i)} \simeq \frac{N}{2^{i+1}}$$

$$M (R_{ak}) = M (R_{bk}) = \frac{(n+1) n \dots (n-k+1)}{N (N-1) \dots (N-k+1)} \simeq \frac{N}{2^{k+1}}$$

(1) Cf. Hald' [13]

Les valeurs successives de $M(R_{ak})$ et $M(R_{bk})$ peuvent être calculées par récurrence à partir de :

$$M(R_{a_{k+1}}) = \frac{n-k}{N-K} M(R_{a_k})$$

$$M(R_{a1}) = \frac{n+1}{2} = \frac{N+2}{4}$$

$$M(r_{a_i}) = M(R_{a_i}) - M(R_{a_{i+1}})$$

On a encore :

$$M(R_k) = 2 M(R_{ak}) = \frac{N}{2^k}$$

$$M(R) = n + 1 = \frac{N+2}{2}$$

* * *

TROISIÈME PARTIE

TESTS BASÉS SUR LES SIGNES DES DIFFÉRENCES SUCCESSIVES

I

TEST BASÉ SUR LE NOMBRE DES DIFFÉRENCES NÉGATIVES (OU POSITIVES)

Considérons la série de n observations successives (série chronologique)

$$x_1, \dots, x_n$$

qui donne lieu à la séquence de $n - 1$ différences premières

$$x_2 - x_1 \dots x_n - x_{n-1}$$

Soit S le nombre des différences négatives de cette séquence. Dans l'hypothèse où l'échantillon est un échantillon aléatoire provenant d'une distribution continue, la distribution de S est indépendante de la distribution des x_i .

L'étude de S peut servir à tester le caractère aléatoire des variations de x . La distribution de S a été étudiée par G. H. Moore et W. A. Wallis [15], puis par H. B. Mann [16].

La valeur moyenne de S est :

$$\bar{S} = \frac{n-1}{2}$$

et sa variance

$$V(S) = \frac{n+1}{12}$$

De plus, lorsque n augmente, la distribution de S tend vers la distribution normale, l'approximation étant pratiquement suffisante dès que $n > 12$.

Moore et Wallis ont, pour $n = 2 (1) \dots 12$, calculé les probabilités

$$P (S \geq \bar{S})$$

On trouvera ci-après un extrait de cette table.

Remarque.

Dans l'application de la loi normale aux tests ci-dessus étudiés, relatifs à la distribution de variables discrètes, il serait théoriquement nécessaire, pour de petites valeurs des grandeurs étudiées, de tenir compte d'une correction de continuité.

Par exemple, l'intervalle de confiance à 0,95 pour une grandeur G de moyenne \bar{G} . sera pris égal à $G = \bar{G} - 1,96 \sigma_{\bar{G}} - 1/2$, $G_2 = \bar{G} + 1,96 \sigma_{\bar{G}} + 1/2$ avec

$$P (G \leq G_1) = 0,025$$

$$P (G \geq G_2) = 0,025$$

G_1 et G_2 devant pratiquement être ramenés à des valeurs entières, cette correction sera souvent sans intérêt.

Test des signes des différences successives de Moore et Wallis

$$P (n, S \leq S_p) \text{ pour } S_p \leq \frac{n-1}{2}$$

S_p	2	3	4	5	6	7	8	9	10	11	12
0	0,500	0,167	0,042	0,008	0,001						
1		0,500	0,250	0,081	0,024	0,006	0,001				
2			0,500	0,250	0,081	0,024	0,006	0,001			
3				0,500	0,250	0,081	0,024	0,006	0,001		
4					0,500	0,250	0,081	0,024	0,006	0,001	
5						0,500	0,250	0,081	0,024	0,006	0,001

D'après Moore et Wallis [14].

En raison de la symétrie, pour

$$S_p \geq \frac{n-1}{2}, \text{ on a } P (n, S \geq S_p) = P (n, S \leq n-1-S_p)$$

Pour $n > 12$, la distribution de

$$\frac{\left| S - \frac{n-1}{2} \right| - 1/2}{\sqrt{\frac{n+1}{12}}} = \frac{\sqrt{3} (| 2S - n + 1 | - 1)}{\sqrt{n+1}}$$

est approximativement la distribution normale réduite, la correction $- 1/2$ étant la correction de continuité pour emploi de la distribution normale pour une variable discrète (0, 1, ...).

II

TESTS BASÉS SUR LE NOMBRE ET LA LONGUEUR DES SUITES DE DIFFÉRENCES DE MÊME SIGNE

Dans une série d'observations successives

$$x_1, \dots, x_n$$

considérons la séquence des signes + ou - des $n - 1$ différences

$$x_{j+1} - x_j$$

correspondant à des suites de valeurs croissantes ou décroissantes et dont la longueur i est caractérisée par le nombre de signe + (ou -) de chaque suite homogène.

Désignons par

r_i le nombre de suites de longueur i

R_k le nombre de suites de longueur égale ou supérieure à k

$$R_k = \sum_{i=k}^{n-1} r_i$$

R le nombre total des suites

Les propriétés de ces suites ont été étudiées par Olmstead [11], qui a donné la table ci-après de l'espérance mathématique du nombre de suites croissantes ou décroissantes de longueur égale ou supérieure à k dans les rangements aléatoires de n nombres différents :

k	$M(R_k)$ (1)
1	$\frac{1}{3} (2n - 1)$
2	$\frac{1}{12} (3n - 5)$
3	$\frac{1}{60} (4n - 11)$
4	$\frac{1}{360} (5n - 19)$
5	$\frac{1}{2520} (6n - 29)$
6	$\frac{1}{20160} (7n - 41)$
7	$\frac{1}{181440} (8n - 55)$

(1) La formule générale est :

$$M(R_k) = \frac{2}{(k+2)!} [n(k+1) - (k^2 + k - 1)]$$

Par exemple, pour $k = 4$, on voit que $M(R_4) = 1$ pour $360 = 5n - 19$

Soit : $n = 76$

c'est-à-dire que dans les classements aléatoires de 76 nombres différents, on peut *en moyenne* compter obtenir une seule suite croissante ou décroissante de longueur 4 (cinq observations consécutives croissantes ou décroissantes).

On a, en particulier, pour le nombre total des suites :

$$\begin{aligned}M(R) &= 1/3 (2n - 1) \\V(R) &= 1/90 (16n - 29),\end{aligned}$$

la distribution de R étant approximativement normale pour $n > 20$

Pour $k \geq 5$, on a sensiblement

$$V(R_k) = M(R_k)$$

ce qui autorise pratiquement à envisager la possibilité d'une loi de Poisson pour les longues suites ($k \geq 5$).

On a alors

$$P(R_k = u) = e^{-\lambda} \frac{\lambda^u}{u!}$$

avec

$$\lambda = M(R_k)$$

En particulier

$$P(R_k \geq 1) = 1 - e^{-\lambda}$$

donnera la probabilité de trouver au moins une suite de longueur k dans un échantillon de n observations successives.

Ainsi par exemple :

$$P(R_6 \geq 1) < 0,05 \text{ pour } n < 153$$

L'utilisation simultanée du nombre des suites et de leur longueur pourra permettre de tester l'existence des tendances systématiques dans la variation des valeurs successives de x : le trop petit nombre des suites ou l'existence de suites trop longues étant significative de l'existence de telles tendances (systématiques ou cycliques).

BIBLIOGRAPHIE

- [1] WALD and WOLFOVITZ. *On a test whether samples are from the same population*, Ann. of. Math. Stat., vol. 11, 1940.
- [2] F. WILCOXON. *Individual Comparisons by ranking methods*. *Biometrics*, vol. 1, N° 6, 1945.
- [3] F. WILCOXON. *Probability tables for individual comparisons by ranking methods*. *Biometrics*, vol. 3, N° 3, Sept. 1947.
- [4] MANN and WHITNEY. *On a test of whether one of two random variables is stochastically larger than the other (order test U)*, The Ann. of. Math. Stat., vol. 18, N° 1, mars 1947.
- [5] KRUSKAL and WALLIS. *Use of ranks in one criterion variance analysis*, Journal of the American Statistical Association, vol 47, N° 260, décembre 1952.
- [6] COLIN WHITE. *The use of ranks in a test of significance for comparing two treatments*, *Biometrics*, vol. 8, N° 1, mars 1952.
- [7] KENDALL. *Advanced theory of statistics* (Tome II, p. 123), Griffin and C°. London.
- [8] MOOD. *The distribution theory of runs* Ann. Math. Stat., vol. 11, 1940, p. 367-392.
- [9] MOSTELLER. *Notes on an application of runs to quality control charts*, Ann. Math. Stat., vol 12, 1941, p. 228-232.

- [10] WOLFOWITZ. *Asymptotic distributions of runs up and down*, Ann. Math. Stat., vol. 15, 1944, p. 163-172.
- [11] OLMSTEAD. *Distribution of sample arrangements for runs up and down* (Tables, Ann. Math. Stat., vol. 17, 1947, p. 24-33.
- [12] MORICE et CHARTIER. *Méthode statistique*, 2 vol. Institut national de la Statistique. Paris, 1955.
- [13] HALD. *Statistical theory with engineering applications*, J. Wiley, 1952.
- [14] MOORE and WALLIS. *Time series significance tests based on signs of differences*. Journal of the Amer. Stat. Ass., vol. 38, juin 1943, p. 153-165.
- [15] SVED and EISENHART. *Tables for testing randomness of grouping in a sequence of alternatives*, Ann. of Math. Stat., vol. 14, 1943, p. 66-87.
- [16] MANN. *On a test of randomness based on signs of differences*, The Ann. of Math. Stat., vol., 16, 1945, p. 200-204.

E. MORICE,

Inspecteur général à l'I.N.S.E.E

*
*
*