REVUE DE STATISTIQUE APPLIQUÉE

REMERY
KAMINSKI
FERROUILLAT
QUEFFELEC

Test de décrochement des distributions

Revue de statistique appliquée, tome 23, n° 2 (1975), p. 19-27
<http://www.numdam.org/item?id=RSA_1975__23_2_19_0>

TEST DE DÉCROCHEMENT DES DISTRIBUTIONS (1)

MM. REMERY, KAMINSKI, FERROUILLAT et QUEFFELEC
Informatique, marketing, management

L'idée du test de décrochement des distributions provient d'une intuition originale de M. Remery.

MM. Kaminsky, Ferrouillat et Queffelec ont mis en commun les résultats de calculs menés parfois séparément selon des méthodes différentes.

Enfin le Professeur Thionet, qui connaît bien ce type de statistique, a apporté un certain nombre de remarques sur la validité de ce test.

INTRODUCTION

Il s'agit d'un test de signification de la disparité entre deux groupes d'observations. La statistique utilisée est le "décrochement". On dira qu'une observation du 2è échantillon "décroche à gauche" s'il n'y a pas d'observation du 1er échantillon inférieure à celle-ci. On voit, par exemple, sur la Figure 1 qu'il y a : 3 décrochements à gauche.

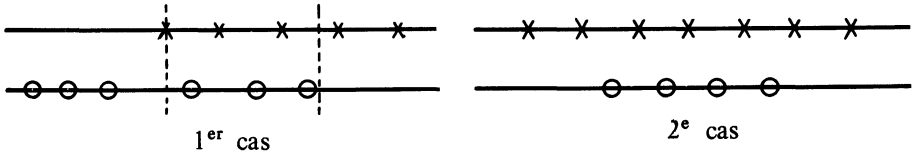


Figure 1

(1) Le test décrit dans cet article est lié à celui qu'avait proposé Tukey, pour un problème un peu moins général :

"J.W. TUKEY – A quick, compact, two sample test to Duckworth's specification. *Technometrics* **1** (1959) p. 31/48"

repris successivement par :

F.L. RAMSEY, *J.A.S.A.* **66** (Mars 1971) p. 149/151

W.J. WESTLAKE, *Technometrics* **13**, 4 (Nov. 1971) p. 901/903

D.F. BAUER, *J.A.S.A.* **67** (Sept. 1972) p. 687/690

En effet dans son article initial, Tukey envisageait la même statistique lorsque chaque échantillon "déborde" l'autre d'un seul côté. L'étude publiée ici concerne tous les cas, et aborde la question de l'efficacité du test.

(N.D.L.R.)

On définit de même les "décrochements à droite" et l'on constate que 4 cas de figure seulement sont possibles : les deux indiqués et les deux autres obtenus à partir de ceux-ci en faisant jouer un rôle symétrique aux deux échantillons.

Le test dont il est question a déjà été signalé par M. H. QUENOUILLE dans son livre "méthodes de calcul statistiques rapides" page 50 pour des échantillons de tailles élevées et égaux en nombre. Il le considère comme un test très rapide, indépendant de la forme des distributions mais d'efficacité faible.

• On trouvera dans cet article :

I – Exposé théorique

II – Efficacité

III – Utilisation et tables.

On peut bien sûr se reporter directement en III.

I – EXPOSE THEORIQUE DU TEST

On se donne deux échantillons :

– provenant d'une même population

– deux observations ne pouvant être identiques.

– de même taille N (à titre de simplification)

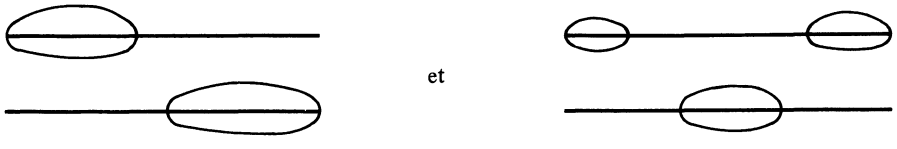
Dans ces conditions, on calcule successivement la probabilité d'avoir p décrochements : P(p) et la probabilité pour que le nombre n de décrochements soit inférieur ou égal à p : P ($n \leqslant p$).

1 – Calcul de P(p)

Evaluons d'abord la probabilité P(k , q) d'avoir k décrochements à gauche et q décrochements à droite dans les deux cas la figure ci-contre (voir Nota).

1er cas 2e cas

Nota : les 2 cas de figure comprennent les 2 cas particuliers suivants :



et

• dans le 1er cas, on a :

$$P_1(k, q) = \frac{N(N-1)^2\ (N-2) \ldots . (N-k)\ (N-2) \ldots . (N-q)}{(2N-1)\ (2N-2) \ldots . (2N-k-q-1)}$$

$$P_1(k, q) = \frac{N\Gamma^2\ (N)}{\Gamma\ (2N)}\ C^{N-k-1}_{2N-k-q-2}$$

• dans le 2è cas, on a :

$$P_2(k, q) = \frac{N\ (N-1)^2\ (N-2) \ldots . (N-k-q+1)}{(2N-1)\ (2N-2) \ldots . (2N-k-q-1)}$$

$$P_2(k, q) = \frac{N\Gamma^2\ (N)}{\Gamma\ (2N)}\ C^{N-2}_{2N-k-q-2}$$

La première formule est valable pour : $1 \leqslant k \leqslant N$ et $1 \leqslant q \leqslant N$ quant à la seconde pour : $1 \leqslant k \leqslant N$; $1 \leqslant q \leqslant N$; $2 \leqslant k + q \leqslant N$.

Nous pouvons maintenant calculer P (p) par la formule :

$$\text{pour } 2 \leqslant p \leqslant 2N-2\ ;\ P(p) = \sum_{\substack{k+q=p \\ k \geqslant 1 \\ q \geqslant 1}} P(k, q)$$

d'où :

– si $2 \leqslant p \leqslant N$

$$P(p) = \frac{N\Gamma^2\ (N)}{\Gamma\ (2N)} \sum_{k=1}^{p-1} C^{N-k-1}_{2N-p-2} + (p-1)\frac{N\Gamma^2\ (N)}{\Gamma(2N)}\ C^{N-2}_{2N-p-2}$$

– Si $N + 1 \leqslant p \leqslant 2N - 2$

$$P(p) = \frac{N\Gamma^2\ (N)}{\Gamma\ (2N)} \sum_{k=p-N+1}^{N-1} C^{N-k-1}_{2N-p-2} = \frac{N\Gamma^2\ (N)}{\Gamma\ (2N)}\ 2^{2N-p-2}$$

– si $p = 2N - 1$

$$P(p) = 0 \text{ (impossibilité évidente)}$$

– si $p = 2N$

$$P(p) = \frac{N\Gamma^2\ (N)}{\Gamma\ (2N)}$$

2 – Expression de P (p) pour certaines valeurs de p. ($2 \leqslant p \leqslant N$)

Il est intéressant de voir l'expression exacte de P (p) :

$$P(2) = P(3) = \frac{N(N-1)}{(2N-1)\ (2N-3)}$$

$$P(4) = \frac{N(N-1)\ (6N-17)}{4\ (2N-1)\ (2N-3)\ (2N-5)}$$

$$P(5) = \frac{N(N-1)(2N-7)}{2(2N-1)(2N-3)(2N-5)}$$

$$P(6) = \frac{N(N-1)(5N^2-40N+81)}{4(2N-1)(2N-3)(2N-5)(2N-7)}$$

$$P(7) = \frac{N(N-1)(3N^2-28N+68)}{4(2N-1)(2N-3)(2N-5)(2N-7)}$$

$$P(8) = \frac{N(N-1)(14N^3-217N^2+1155N-2094)}{16(2N-1)(2N-3)(2N-5)(2N-7)(2N-9)}$$

$$P(9) = \frac{N(N-1)(8N^3-140N^2+860N-1824)}{16(2N-1)(2N-3)(2N-5)(2N-7)(2N-9)}$$

$$P(10) = \frac{N(N-1)(9N^4-228N^3+2259N^2-10248N+17760)}{16(2N-1)(2N-3)(2N-5)(2N-7)(2N-9)(2N-11)}$$

3 – Expression de P (n ⩽ p) pour certaines valeurs de p.

A partir des expressions précédentes, on a :

$$P(n \leqslant 5) = \frac{N(N-1)(26N-71)}{4(2N-1)(2N-3)(2N-5)}$$

$$P(n \leqslant 6) = \frac{N(N-1)(57N^2-364N+578)}{4(2N-1)(2N-3)(2N-5)(2N-7)}$$

$$P(n \leqslant 7) = \frac{N(N-1)(30N^2-196N+323)}{2(2N-1)(2N-3)(2N-5)(2N-7)}$$

$$P(n \leqslant 8) = \frac{N(N-1)(494N^3-5513N^3+20435N-25350)}{16(2N-1)(2N-3)(2N-5)(2N-7)(2N-9)}$$

$$P(n \leqslant 9) = \frac{N(N-1)(502N^3-5653N^2+21295N-27174)}{16(2N-1)(2N-3)(2N-5)(2N-7)(2N-9)}$$

$$P(n \leqslant 10) = \frac{N(N-1)(1013N^4-17056N^3+107032N^2-298841N+316674)}{16(2N-1)(2N-3)(2N-5)(2N-7)(2N-9)(2N-11)}$$

C'est à partir de ces différentes expressions que sont constitués les tables I.

4 – Comportement de P (p) et P (n , p) pour les grandes valeurs de N

D'après les formules du paragraphe 1, on voit que pour N grand :

$$P(p) \simeq \frac{2(p-1)}{2^{p+1}} = \frac{p-1}{2^p}$$

En effet seul le cas $2 \leqslant p \leqslant N$ intervient, car les probabilités pour $p \geqslant N+1$ sont de l'ordre de $\frac{1}{\sqrt{N}}$.

Dans ce cas, le nombre de décrochements suit donc une loi binomiale négative, de moyenne et de variance 4.

$$P(n \leqslant p) \cong \sum_{k=2}^{p} \frac{k-1}{2^k} = \frac{p+1}{2^p}$$

5 – Etude de la moyenne et de la variance du nombre de décrochements

Soit X la variable statistique : "nombre de décrochements" ; on a :

$$E(X) = \sum_{p=2}^{2N-1} pP(p) + 2NP(2N)$$

Un calcul direct, simplifié par des considérations de symétrie, conduit à :

$$\boxed{E(X) = \frac{4N}{N+1}}$$

Le calcul de la variance conduit à la formule suivante :

$$\boxed{\mathrm{Var}(X) = 4\left\{\frac{N^3 - N^2 + 4N + 2}{(N+1)^2\,(N+2)} - \frac{1}{C_{2N}^{N}}\right\}}$$

II – EFFICACITE DU TEST DE DECROCHEMENT

Le tracé de la courbe d'efficacité a été déterminé par simulation, en considérant des couples d'échantillons provenant de variables normales réduites, dont la moyenne du 1er échantillon est décalée de la moyenne du second échantillon de la quantité Δm.

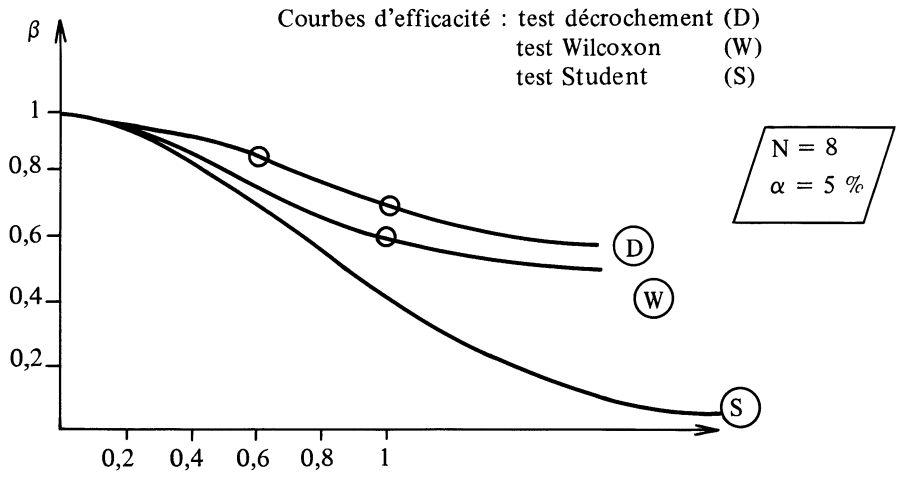
Nous avons, en parallèle, tracé la courbe d'efficacité du test de Wilcoxon. Le tableau suivant donne les résultats obtenus comparés au test de Student.

La quantité β représentant la probabilité d'accepter les deux échantillons comme provenant de la même population.

	Δm	0	0,2	0,6	1
β estimé	Test décrochement	0,88	0,98	0,82	0,68
	Test de Wilcoxon	0,92	0,96	0,72	0,64
	Test de Student	0,95	0,90	0,72	0,40

Risque 5 %
N = 8

Le calcul des estimations de β pour les deux premières lignes a été fait à partir de 50 tirages aléatoires d'échantillons normaux de taille 8. Le critère d'acception pour le test de décrochement étant X = 7.



Nous avons fait la même simulation pour des échantillons de taille N = 12 et obtenu des résultats analogues

III – TABLES ET UTILISATION DU TEST

On trouvera trois tables donnant :

Table I : la probabilité pour que le nombre de décrochements soit $\leqslant p$.

Table II : la probabilité pour que le nombre de décrochements soit $> p$.

Table III : pour différents niveaux de confiance, le nombre de décrochements nécessaires pour avoir un test significatif.

Le test de décrochement permet une comparaison globale de deux échantillons sans aucune hypothèse sur les natures des distributions. Sa simplicité et son efficacité même pour les petits échantillons en font dans de nombreux domaines, recherche ou contrôle de fabrication par exemple, un outil précieux.

c) Exemple et comparaison avec le test de X^2

1/ Soit la distribution d'observations suivantes provenant de 2 échantillons dont on veut tester l'appartenance à une même population : seules les observations qui "décrochent" ont été représentées dans le schéma ci-dessous.

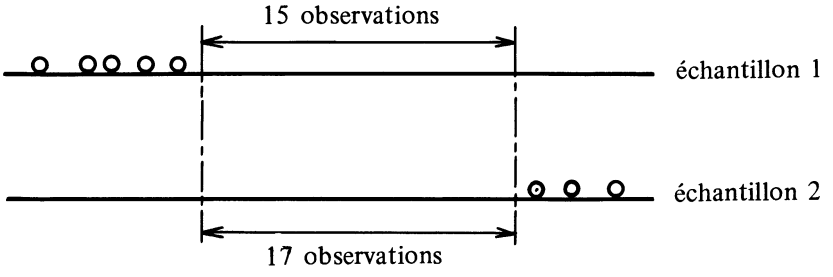


Table I

A l'intersection de la ligne N et de la colonne P, on lit la probabilité en % pour que le nombre de décrochements soit $\leqslant$ P.

	5	6	7	8	9	10	11	12	13	14	15	16
5	93,651	96,826	98,413	99,206	99,206	100	100	100	100	100	100	100
6	91,991	96,537	98,268	99,134	99,567	99,783	99,783	100	100	100	100	100
7	90,559	95,921	98,135	99,068	99,534	99,767	99,883	99,942	99,942	100	100	100
8	89,417	95,291	97,902	99,005	99,503	99,751	99,875	99,937	99,969	99,985	99,985	100
9	88,507	94,727	97,639	98,914	99,473	99,737	99,868	99,933	99,967	99,983	99,991	99,995
10	87,771	94,237	97,380	98,806	99,436	99,723	99,861	99,930	99,965	99,982	99,992	99,997
11	87,166	93,815	97,134	98,693	99,391	99,707	99,854	99,927	99,963	99,981	99,990	99,995
12	86,662	93,452	96,921	98,581	99,343	99,688	99,846	99,922	99,960	99,979	99,989	99,994
15	85,555	92,622	96,388	98,562	99,197	99,624	99,821	99,913	99,957	99,978	99,989	99,994
20	84,457	91,758	95,794	97,918	98,968	99,521	99,773	99,893	99,949	99,975	99,988	99,993
	81,250	89,062	93,75	96,484	98,047	98,925	99,415	99,688	99,829	99,908	99,951	99,974

Ainsi deux échantillons de taille 10 chacun, ont 97,380 % de chances d'avoir un nombre de décrochements $\leqslant$ 7.

Table II

On lit dans la case (P , N) la probabilité (en %) pour des échantillons de taille N, d'avoir un nombre de décrochements > P

N \ P	5	6	7	8	9	10	11
5	6,349	3,174	1,587	0,794	0,794	0	0
6	8,009	3,463	1,732	0,866	0,433	0,217	0,217
7	9,441	4,079	1,865	0,932	0,466	0,233	0,117
8	10,583	4,709	2,098	0,995	0,497	0,249	0,125
9	11,493	5,273	2,361	1,086	0,527	0,263	0,132
10	12,229	5,763	2,620	1,194	0,564	0,277	0,139
11	12,834	6,185	2,866	1,307	0,609	0,293	0,146
12	13,338	6,548	3,079	1,419	0,657	0,312	0,154
15	14,445	7,378	3,612	1,438	0,803	0,376	0,179
20	15,543	8,242	4,206	2,072	1,032	0,479	0,227
∞	18,750	10,938	6,250	3,516	1,953	1,075	0,585

Ainsi deux échantillons de taille 10 chacun, ont 2,620 % de chances d'avoir un nombre de décrochements > 7.

2/ Le test du Décrochement indique que l'on atteint le seuil de signification à 5 % (8 points à l'extérieur pour des échantillons de taille 20)

3) Le test χ^2 donne, en considérant 3 classes et le tableau de contingence suivant (2 classes extrêmes ayant 0 observations, dans l'un des échantillons, et une classe centrale).

	classe 1	classe 2	classe 3	Total
Echantillon n° 1	5	15	0	20
Echantillon n° 2	0	17	3	20
Total	5	32	3	40

$$\chi^2 = 2\left[\frac{(5-2,5)^2}{2,5} + \frac{(15-16)^2}{16} + \frac{(0-1,5)^2}{1,5}\right] = 8,13$$

TABLE III

Cette table donne le nombre minimum de décrochements nécessaires pour pouvoir conclure que les deux échantillons proviennent de populations différentes.

N	5 %	1 %	0,2 %	0,1 %
6	7	9	12	-
7	7	9	12	14
8	7	9	12	13
9	8	10	12	13
10	8	10	12	13
11	8	10	12	13
12	8	10	12	13
15	8	10	12	13
20	8	11	13	14
25	9	12	14	15
∞	9	12	14	15

4 – CONCLUSION

– Bien que le test du χ^2 soit peu valable vu le trop petit nombre d'observations des 4 classes extrêmes (2 fois 0 – 1 fois 3 – et 1 fois 5),

Son calcul donne :

– pour $(2 - 1)(3 - 1) = 2$ degrés de liberté.

– 5,99 comme seuil de signification à 5 % .

– Comme 8,13 (correspondant à 2 % environ) dépasse ce seuil, on conclut sensiblement comme pour le test du Décrochement (d'une manière un peu plus significative).