REVUE DE STATISTIQUE APPLIQUÉE

E. DEROBERT

Un élément de comparaison des tests de Kramer, Friedman, et de la différence des extrêmes. Une étude de la puissance dans un cas particulier

Revue de statistique appliquée, tome 36, nº 2 (1988), p. 53-59

<http://www.numdam.org/item?id=RSA_1988__36_2_53_0>

Rev. Statistique Appliquée, 1988, **XXXVI** (2), 53-60

UN ÉLÉMENT DE COMPARAISON DES TESTS DE KRAMER, FRIEDMAN, ET DE LA DIFFÉRENCE DES EXTRÊMES
UNE ÉTUDE DE LA PUISSANCE DANS UN CAS PARTICULIER

E. DEROBERT

(Laboratoire SOREDAB, La Boissière-Ecole, 78120 Rambouillet)

RÉSUMÉ

Nous avons effectué une simulation afin d'apporter un élément de comparaison de la puissance de trois tests s'appuyant sur des méthodes de sommes de rangs, utilisées notamment (mais pas seulement) en dégustation et en analyse sensorielle.

Ces trois tests sont :

— le test de KRAMER (utilisé avec les tables correctes calculées par THOMPSON et WILLKE (1963))
— le test de FRIEDMAN
— le test de la différence des extrêmes.

Nous avons choisi d'étudier le cas de 5 juges et de 4 produits, qui avantageait (ou plutôt ne désavantageait pas) le test de KRAMER au niveau des risques de première espèce.

Malgré cette précaution, le test de KRAMER ne s'avère puissant que lorsqu'il s'agit de séparer un seul produit de tous les autres. Le test de la différence des extrêmes se comporte mieux; il améliore KRAMER en gardant l'avantage de la simplicité d'utilisation. Le test de FRIEDMAN (le plus ancien), qui utilise le plus d'informations, donne, dans l'ensemble, les meilleurs résultats.

Mots clés : *Test de Friedman, Tests de rangs, Puissance*

Plan

Table II : Valeurs critiques des tests de KRAMER, de FRIEDMAN et de la différence des extrêmes aux seuils de 5 % et 1 % (cas de 5 juges et 4 produits).

Table III : La puissance des trois tests (résultats de la simulation).

1. Rappel sur les tests de KRAMER, de FRIEDMAN, et de la différence des extrêmes

Décider si certains produits sont préférés à d'autres, implique l'usage de tests statistiques adaptés. Une procédure classique consiste à faire déguster P produits par J juges et à demander aux J juges de classer par ordre de préférence les P produits. On effectue alors, pour chaque produit, la somme des rangs que lui ont attribués les J juges. Les tests de KRAMER, de FRIEDMAN et de la différence des extrêmes sont basés sur ces sommes de rangs.

L'hypothèse nulle de ces tests est : « tous les produits sont équivalents » : c'est-à-dire que l'on considère, sous cette hypothèse nulle, que les P ! classements possibles parmi lesquels choisissent de façon indépendante les J juges, sont équiprobables. L'hypothèse alternative est : « il existe au moins un produit dont la qualité est différente de celle des autres » ([1]).

1.1. Le test de KRAMER

La décision par le test de KRAMER s'effectue en considérant seulement la somme des rangs la plus élevée et la somme des rangs la moins élevée. On rejette l'hypothèse nulle si la première somme est plus grande qu'une certaine valeur ou si la seconde est plus petite qu'une autre valeur. Si l'on appelle, $R_1, R_2, \ldots, R_p$, les sommes des rangs de chaque produit, les deux valeurs limites sont fixées à partir de la distribution conjointe ([2]) de $\max_{i=1,P} (R_i)$ et de $\min_{i=1,P} (R_i)$ sous l'hypothèse nulle.

Une autre façon de formuler le test de KRAMER est de considérer

$$K = \min \left(\min_{i=1,P} (R_i) - J,\ PJ - \max_{i=1,P} (R_i) \right)$$

Les valeurs critiques pour K ont été calculées par THOMPSON et WILLKE (1963) à 1, 3 et 5 % pour toutes les valeurs de J et de P inférieures ou égales à 15. Ces auteurs proposent aussi des résultats asymptotiques.

---

(1) Nous ne traiterons pas ici du test de KRAMER dans le cas où l'un des produits est prédéterminé.

(2) C'est en oubliant cette notion de distribution conjointe que KRAMER a publié des tables unilatérales au lieu de tables bilatérales (KRAMER 1960).

1.2. Le test de FRIEDMAN

La statistique utilisée est ici :

$$F = \frac{12}{JP(P+1)} \sum_{i=1}^{P} (R_i - \overline{R})^2 = \frac{12}{JP(P+1)} \left(\sum_{i=1}^{P} R_i^2 \right) - 3J(P+1)$$

avec

$$\overline{R} = \frac{1}{P} \sum_{i=1}^{P} R_i = \frac{J(P+1)}{2}$$

OWEN (1962) a calculé la distribution exacte de F pour $J = 3$, $P = 2(1)\,15$ et $J = 4$, $P = 2(1)\,8$. Les tables d'OWEN ont été contestées par QUADE (1984) qui en a proposé d'autres (1972) dans un rapport auquel nous n'avons pu accéder. En tout état de cause, la prudence s'impose.

On trouve aussi les valeurs critiques de ce test — à 1 et 5 % — dans HOLLANDER et WOLFE (1973) et dans la norme AFNOR NF V 09-018 (voir Table I). Il y a rejet de l'hypothèse nulle lorsque F est supérieur ou égal à ces valeurs critiques. Asymptotiquement la statistique F suit une loi du χ^2 à $P - 1$ degrés de liberté que l'on utilise dans les cas non tabulés. Cette approximation aboutit à des tests conservateurs (FRIEDMAN 1940). Notons qu'il existe une correction de F qui améliore la convergence asymptotique — voir BRADLEY (1968).

1.3. Le test de la différence des extrêmes

Comme le test de KRAMER, ce test s'appuie exclusivement sur $\max_{i=1,P} R_i$ et $\min_{i=1,P} R_i$. On prend en compte la différence $W = \max_{i=1,P} R_i - \min_{i=1,P} R_i$

$$(\text{ou encore } W = \max_{i,j \in \{1,\dots,P\}} |R_i - R_j|).$$

Cette statistique a été tabulée par Mc DONALD et THOMPSON (1967) où elle figure sous le nom de « Wilcoxon multiple comparaison method ». Des résultats asymptotiques figurent dans WILCOXON ET WILCOX (1964). Bien entendu le rejet de l'hypothèse nulle a lieu quand W dépasse les valeurs critiques de la statistique.

2. Comparaison des trois tests

Nous avons voulu estimer la puissance de ces trois tests dans un cas particulier. Nous avons choisi le cas $P = 4$, $J = 5$ qui présentait l'avantage que, pour le test de KRAMER, les risques réels de 1ère espèce sont peu éloignés des seuils classiques de 5 % et 1 %. C'est en effet un défaut structurel du test de KRAMER — que nous n'avons pas voulu prendre en compte pour notre simulation — que le nombre de valeurs discrètes qu'il peut prendre est plus réduit, ce qui implique qu'en général on serre de moins près des probabilités fixées à l'avance et que, par conséquent, on perd automatiquement en puissance en se voyant obligé de travailler avec un risque de 1ère espèce plus petit que celui auquel

TABLE I

Valeurs critiques du test de FRIEDMAN
(seuils de 0,05 et de 0,01)

Nombre de sujets J	Nombre d'échantillons (ou de produits) P					
	3	4	5	3	4	5
	Seuil de signification $\alpha = 0{,}05$			Seuil de signification $\alpha = 0{,}01$		
2	—	6,00	7,60	—	—	8,00
3	6,00	7,00*	8,53	—	8,20*	10,13
4	6,50	7,50*	8,80	8,00	9,30*	11,00
5	6,40	7,80	8,96	8,40	9,96	11,52
6	6,33*	7,60	9,49**	9,00	10,20	13,28**
7	6,00*	7,62	9,49**	8,85	10,37	13,28**
8	6,33	7,65	9,49**	9,00	10,35	13,28**
9	6,22	7,81**	9,49**	8,66	11,34**	13,28**
10	6,20	7,81**	9,49**	8,60*	11,34**	13,28**
11	6,54	7,81**	9,49**	8,90*	11,34**	13,28**
12	6,16	7,81**	9,49**	8,66*	11,34**	13,28**
13	6,00	7,81**	9,49**	8,76*	11,34**	13,28**
14	6,14	7,81**	9,49**	9,00	11,34**	13,28**
15	6,40	7,81**	9,49**	8,93	11,34**	13,28**

Notes 1 : on pourrait tabuler le test de FRIEDMAN dans le cas P = 2. Cependant, dans ce cas, il suffit d'appliquer la loi binomiale (ou son approximation normale) sur le nombre de fois où l'un des deux échantillons est préféré à l'autre. Ceci ramène en fait aux comparaisons par paires qui font l'objet de la norme NF V 09-012 (essai bilatéral).

2 : la quantité F ne peut prendre que des valeurs discontinues, cette discontinuité étant très accusée pour les petites valeurs de (J, P). Il en résulte qu'on ne peut pas obtenir des valeurs critiques correspondant exactement aux seuils 5 % et 1 %; les valeurs marquées d'un astéristique (*) correspondent à des seuils très légèrement supérieurs à 5 % et 1 %; les valeurs non marquées correspondent à des seuils réels inférieurs à 5 % et 1 %.

3 : les valeurs marquées d'un double astéristique (**) sont les valeurs critiques obtenues à l'aide de l'approximation par la loi de χ^2.

(extrait de la norme NF V 09-018 — janvier 1987, reproduit avec l'autorisation de l'AFNOR)

TABLE II

Valeurs critiques des tests de KRAMER (K), de FRIEDMAN (F) et de la différence des extrêmes (W) aux seuils de 5 % et 1 %
(cas de 5 juges et 4 produits)

	5 %		1 %	
	Valeurs critiques	Calcul du α	Valeurs critiques	Calcul du α
KRAMER → K	1	$0.045 \leqslant \alpha \leqslant 0.047$	0	$\alpha \leqslant 0.008$
FRIEDMAN → F	7.80	0.044	9.96	0.009
Différence des extrêmes → W	11	$\alpha \leqslant 0.037$	12	$\alpha \leqslant 0.013$*

* La borne supérieure calculée par Mc DONALD et THOMPSON est plus grande que le 0.01 théorique. Mais, à regarder le mode de calcul de cette borne, on peut penser avec les auteurs qu'elle est large et que la valeur réelle se situe aux alentours de 0.01.

on veut se référer [3]. La table II présente les valeurs critiques des trois statistiques ainsi que les risques de 1ère espèce correspondants.

2.1. La procédure de simulation

2.1.1. Les alternatives à l'hypothèse nulle

Par convention, nous utiliserons, pour l'écriture des hypothèses, la méthode classique qui consiste à souligner d'un trait les produits identiques. C'est ainsi que pour 4 produits, l'hypothèse nulle H_0 s'écrit : $\underline{1\ 2\ 3\ 4}$.

Si on supprime les cas symétriques, on peut opposer, à H_0, cinq hypothèses alternatives, de celle qui suppose qu'un seul produit est différent des 3 autres, à celle qui suppose que tous les produits sont différents :

H_1 : $\underline{1\ 2\ 3}\ \ \underline{4}$
H_2 : $\underline{1\ 2}\ \ \underline{3\ 4}$
H_3 : $\underline{1\ 2}\ \ \underline{3}\ \ \underline{4}$
H_4 : $\underline{1}\ \ \underline{2\ 3}\ \ \underline{4}$
H_5 : $\underline{1}\ \ \underline{2}\ \ \underline{3}\ \ \underline{4}$

(3) On peut théoriquement randomiser la décision d'acceptation ou de rejet de l'hypothèse nulle, mais, dans la pratique, cette solution est rarement adoptée.

2.1.2. Le comportement des cinq juges

Pour chacune des cinq hypothèses alternatives à l'hypothèse nulle il s'agit de simuler le comportement des juges.

Nous avons supposé le cas de figure suivant :

— Quatre des cinq juges dégustent convenablement, c'est-à-dire que leurs votes sont compatibles avec l'hypothèse testée. Par exemple pour H_4, ils votent (de façon équiprobable) 1-2-3-4 ou 1-3-2-4.
— Le cinquième juge classe les quatre produits n'importe comment. Son vote est considéré comme un tirage aléatoire (et équiprobable) parmi les 24 (4 !) classements possibles.

Nous simulons, par ce biais, un « bruit » qui va nous permettre, de façon tout-à-fait informelle, d'évaluer le comportement de nos trois tests.

2.1.3. Le calcul de la puissance

Nous avons considéré une à une les cinq hypothèses présentées ci-dessus, et réalisé à chaque fois le calcul exact de la puissance des trois tests. Cependant, l'hypothèse réellement testée à chaque fois est, elle, toujours la même. L'hypothèse H_0 étant « les produits sont identiques », l'hypothèse testée est son complémentaire $\overline{H}_0$, c'est-à-dire « il existe au moins un produit différent des autres ». Il est clair que nous avons laissé de côté les éventuelles procédures de sélection et de classification qui peuvent suivre la réalisation d'un tel test.

Quelle que soit l'hypothèse, la puissance est donc toujours la probabilité pour que le test décèle la non-identité des quatre produits. Les résultats figurent Table III.

TABLE III

La puissance des trois tests
(résultats de la simulation)

Hypothèse simulée	Risque de $1^{ère}$ espèce	KRAMER	FRIEDMAN	Différence des extrêmes
H_1 ($\underline{1\ 2\ 3}\ \ \underline{4}$)	5 %	**51,9 %**	45,0 %	43,1 %
	1 %	**25,6 %**	16,5 %	22,9 %
H_2 ($\underline{1\ 2}\ \ \underline{3\ 4}$)	5 %	32,8 %	**69,7 %**	47,0 %
	1 %	6,1 %	**35,0 %**	23,4 %
H_3 ($\underline{1\ 2}\ \ \underline{3}\ \ \underline{4}$)	5 %	56,3 %	**75,5 %**	66,7 %
	1 %	27,1 %	45,3 %	**46,9 %**
H_4 ($\underline{1}\ \ \underline{2\ 3}\ \ \underline{4}$)	5 %	66,7 %	73,4 %	**75,0 %**
	1 %	41,7 %	**50,5 %**	50,0 %
H_5 ($\underline{1}\ \ \underline{2}\ \ \underline{3}\ \ \underline{4}$)	5 %	66,7 %	**79,2 %**	75,0 %
	1 %	41,7 %	**54,2 %**	50,0 %

Pour chacune des hypothèses et chacun des niveaux de risque, le meilleur résultat est en caractères gras.

2.2. Les résultats de la simulation

Globalement, c'est le test de FRIEDMAN qui se comporte le mieux. Le test de KRAMER se comporte très bien dans le cas où un seul produit est différent de tous les autres (H_1). C'est même, dans ce cas, le meilleur (sans que la différence avec les autres tests soit énorme). Il se comporte à peu près honorablement dans les cas H_3, H_4, H_5 (encore que dans le cas H_3, il y ait un écart de puissance de presque 20 % avec le test de FRIEDMAN). En revanche, il s'écroule totalement dans le cas H_2 où il s'agit de distinguer 2 groupes de 2 produits. Le test de la différence des extrêmes est en général intermédiaire entre FRIEDMAN et KRAMER. Il utilise la même information que le test de KRAMER (le maximum et le minimum) mais l'exploite mieux, au point qu'il fait jeu égal avec le test de FRIEDMAN dans les cas H_3, H_4, H_5. Il s'avère tout de même un peu faible dans le cas H_2, même s'il s'en tire beaucoup mieux que KRAMER.

Références bibliographiques

[1] AFNOR (1987). — *Analyse sensorielle — Méthodologie — Essai de classement par rangs.* Norme NF V 09-018.

[2] J.V. BRADLEY (1968). — *Distribution free statistical tests.* Prentice-Hall, Englewood Cliffs, New Jersey.

[3] M. FRIEDMAN (1940). — *A comparaison of alternative tests of significance for the problem of m rankings.* Annals of mathematical statistics 11, 86-92.

[4] M. HOLLANDER, D.A. WOLFE (1973). — *Nonparametric statistical methods.* John Wiley, New-York.

[5] D.N. JOANES (1985). — *On a rank sum test due to KRAMER.* Journal of food science 50, 1 442-1 444.

[6] M.G. KENDALL (1962). — *Rank correlation methods* (3rd édit.) Griffin, London.

[7] A. KRAMER (1960). — *A rapid method for determining significance of differences from rank sums.* Food technology 14, 576-581.

[8] B.J. Mc DONALD, W.A. THOMPSON (1967). — *Rank sum multiple comparisons in one and two-way classification.* Biometrika 54, 487-497.

[9] D.B. OWEN (1962). — *Handbook of statistical tables.* Addison-Wesley, Reading, Massachussets.

[10] D. QUADE (1972). — *Average internal rank correlation.* Report SW 16/72, Mathematical Center, Amsterdam.

[11] D. QUADE (1984). — *Non parametrics methods in two-way layouts* (pages 185-228 de *Handbooks of statistic vol. 4* de KRISHNAIAH P.R. — SEN P.K. Elsevier Science Publishers, Amsterdam).

[12] W.A. THOMPSON, T.A. WILLKE (1963). — *On an extreme sum test for outliers.* Biometrika 50, 375-383.

[13] F. WILCOXON, R.A. WILCOX (1964). — *Some rapid approximate statistical procedures.* Lederle Laboratories, Pearl River, New-York.