2022 ■ Vol. 18 ■ no. 1

Les distributions multinomiales, leur mesure par les tests khi2 et G, leur approximation par la loi χ^2 et l'analyse des tableaux de fréquences par le test G

Louis Laurencelle [a]

[a]Université du Québec à Trois-Rivières

Abstract ■ La variable multinomiale, une extension de la binomiale, s'exprime par un vecteur de fréquences associées à des catégories d'observations indépendantes, la probabilité d'une observation dans chaque catégorie étant à son tour représentée par un vecteur correspondant de probabilités. Nous revoyons la structure, la genèse et les applications statistiques principales des multinomiales. L'accent est mis sur le test G (ou test du logarithme du quotient de vraisemblances, ln QV) d'une multinomiale, sur ses propriétés, sa distribution et son approximation par la loi du χ^2, comparativement au test classique du khi2. Enfin, grâce à l'additivité et la propriété de décomposition additive du G, l'essai s'achève avec l'illustration de l'analyse de variation de tableaux de fréquences d'une et de deux (ou plusieurs) dimensions par le test G, incluant le fractionnement d'un effet principal, le calcul des effets simples et le test d'interaction. ■ The multinomial variable, an extension of the binomial, is embodied in a vector of frequencies associated with categories of independent observations, the probability of an observation in each category being in turn mapped in a corresponding vector of probabilities. We review the structure, genesis and main statistical applications of multinomials. The focus is on the G test (or log likelihood ratio test, LLR) on a multinomial, its properties, its distribution and its approximation by the chi-square probability distribution, with comparison with the classical chi-square test. Finally, thanks to the additivity and additive decomposition property of G, the essay concludes by exemplifying the analysis of variation in frequency tables of one and two (or more) dimensions by the G test, comprising the splitting of a main effect, the calculation of simple effects and an interaction test.

Keywords ■ Multinomials; Statistical properties; Frequency tables; G test, Chi-square test.

louis.laurencelle@gmail.com

10.20982/tqmp.18.1.p001

Acting Editor ■ Denis Cousineau (Université d'Ottawa)

Introduction

L'unité d'une loi multinomiale est une variable multiple discrète représentée par le multiplet d'entiers $\{x_1, x_2, \ldots, x_k\}$ tel que $x_1+x_2+\cdots+x_k = n, x_i \geq 0$ et $n > 0$. Chaque élément numérique x_i dénote généralement une fréquence, un nombre d'occurrences, un décompte, l'objet compté appartenant à la catégorie i indiquée.

Exemple 1. Un exemple simple illustre le contexte d'application de ces traitements. Soit une série statistique de n éléments, $e_1, e_2, \ldots, e_n$, chaque élément étant classé dans l'une de k catégories, telle la courte série ci-dessous :

e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8	e_9	e_{10}	e_{11}	e_{12}
A	C	C	A	D	C	B	D	B	C	A	C

Cette série présente potentiellement deux caractères d'intérêt statistique, soit celui de la répartition des $n = 12$ éléments parmi les $k = 4$ catégories identifiées, soit celui des groupements en suites des éléments selon leur sériation (du 1er au 12e) et des transitions d'une catégorie à l'autre. Le lecteur intéressé à ce second aspect trouvera réponse et documentation dans Lauren-

CELLE (2020). Quant aux fondements mathématiques de la distribution multinomiale, ses propriétés, son estimation et ses variantes, le lecteur est référé à l'ouvrage remarquable de JOHNSON, KOTZ et BALAKRISHNAN (1977). Nous nous intéressons ici au seul aspect de la répartition des éléments en catégories et aux statistiques sommatives qui en découlent.

La répartition des $n = 12$ éléments ci-dessus en $k = 4$ catégories est complètement exprimée dans le tableau suivant illustrant l'Exemple 1 :

Catégorie :	A	B	C	D
Fréquence :	3	2	5	2

tableau synthèse à propos duquel différents questionnements sont possibles. C'est cette répartition de fréquences catégorielles, 3, 2, 5 2 comme ci-dessus ou, plus généralement, $\{x_1, x_2, \ldots, x_k\}$, qui constitue la variable multinomiale.

Bien que d'autres traitements statistiques soient possibles, l'application courante faite à une variable multinomiale consiste à vérifier si les fréquences x_i se répartissent à peu près conformément ou non à un modèle de répartition donné, par exemple un modèle égalitaire, ou en rampe croissante, ou à forme gaussienne, le profil de modèle étant généralement fourni ou exprimable sous la forme d'un multiplet correspondant de probabilités, $\pi_1, \pi_2, \ldots, \pi_k$, $\pi_i > 0$, sommant à 1. Parmi quelques autres procédures de test (voir à ce sujet l'excellent exposé de KENDALL & STUART, 1979, chap. 30), deux méthodes cousines ressortent, connues sous le nom du test khi2 (attribué à K. Pearson) et du test G ou logarithme du quotient de vraisemblances, la distribution de ces deux statistiques tendant asymptotiquement vers la loi χ^2.

Brièvement, dans le cas d'une variable multinomiale simple, le test khi2, dorénavant noté X^2, consiste à mesurer une distance entre les k fréquences d'occurrence observées x_i et les k valeurs attendues correspondantes, $n \cdot \pi_i$, la statistique bien connue étant alors :

$$X^2 = \sum_{i=1}^{k} \frac{(x_i - n\pi_i)^2}{n\pi_i}. \tag{1}$$

Pour un modèle à probabilités égales, soit $\pi_i = n/k$, le calcul devient plus simple, soit :

$$X^2 = \frac{k \sum_{i=1}^{k} x_i^2}{n} - n, \tag{2}$$

avec pour ce dernier la valeur maximale $n(k-1)$.

Quant au test G (parfois noté G^2), il découle d'un quotient (QV) de vraisemblances (ou probabilités conditionnelles), l'une dépendant du modèle de probabilités testé, l'autre de la donnée multinomiale observée, cette dernière donnant la valeur de vraisemblance maximale associée aux données observées, soit :

$$\begin{aligned} QV &= \frac{{\pi_1}^{x_1} \times {\pi_2}^{x_2} \times \cdots \times {\pi_k}^{x_k}}{\left(\frac{x_1}{n}\right)^{x_1} \left(\frac{x_2}{n}\right)^{x_2} \cdots \left(\frac{x_k}{n}\right)^{x_k}} \\ &= n^n \times \left(\frac{\pi_1^{x_1} \times \pi_2^{x_2} \times \cdots \pi_k^{x_k}}{x_1^{x_1} \times x_2^{x_2} \times \cdots \times x_k^{x_k}} \right). \end{aligned} \tag{3}$$

Ce quotient varie depuis à peu près 0 (signalant une grande discordance entre modèle et données) jusqu'à 1 (indiquant une concordance parfaite entre modèle et données). Analogiquement avec la relation connue entre la loi χ_2^2 et la variable uniforme $u \sim U(0..1)$, soit $\chi_2^2 = -2 \cdot \ln u$, l'expression

$$G = -2 \cdot \ln QV = -2 \left[\sum_{i=1}^{k} x_i \ln \pi_i - \sum_{i=1}^{k} x_i \ln (x_i/n) \right] \tag{4}$$

se distribue asymptotiquement comme χ_{k-1}^2 pour n élevé (WILKS, 1935, 1938). Différentes expressions algébriques de ce calcul sont possibles, notamment :

$$G = 2 \left[\sum_{i=1}^{k} x_i (\ln x_i - \ln \pi_i) - n \ln n \right]. \tag{5}$$

Une formule plus commode pour un test à probabilités π_i égales serait :

$$G = 2 \left[\sum_{i=1}^{k} x_i \ln x_i - n \ln(n/k) \right], \tag{6}$$

avec pour valeur maximale $2n \ln k$. Noter enfin que les expressions de type '$0 \ln(0)$' sont de valeur 0. Dans le but louable de rapprocher la distribution du G de la fonction de répartition de la loi χ^2, WILLIAMS (1976) propose la correction suivante,

$$G_W = G/q_W, \text{ où } q_W = 1 + k^2/(6 \cdot n \cdot dl), \tag{7}$$

« dl » étant les degrés de liberté concernés. Un multiplet de k éléments présentant $k-1$ degrés de liberté, la correction à appliquer ici se ramène à $q_W = 1 + (k+1)/6n$, son impact s'évanouissant pour de grands rapports de $\sqrt{n}/k$.

Ces trois statistiques, X^2, G et G_W, se distribuent approximativement comme χ^2 avec $k - 1 - p$ degrés de liberté, p étant le nombre de paramètres du modèle de probabilités estimés à partir des k éléments.

Génération de la variable multinomiale

Principe : À partir des k-partitions complétées de n, les A_p arrangements de chaque partition [1] et les P_p permutations

1. Dans ce texte, le traitement d'une partition proprement définie étant anecdotique, nous prenons la liberté de noter occasionnellement par « partition » ce qui en fait est une k-partition, si le contexte textuel ne présente pas d'ambiguïté.

Encadré 1 ■ FONCTIONS AUXILIAIRES

```
function min(a,b :integer) :integer;
begin if a<b then min:=a else min:=b end;

function qsup(a,b:integer):integer;
var q:integer;
begin q:=a div b; if q*b<a then q:=q+1;
          qsup:=q
end;
```

Encadré 2 ■ PROCÉDURE D'ÉNUMÉRATION (Générer les k-partitions complétées)

```
procedure genkpart(j,k,kr,xant,reste :integer);
  { Le vecteur vec[1.. k] : integer reçoit la partition }
  { L'appel à la procédure est : genkpart(1,k,k,n,n) }
var x :integer;
begin
    if j<k then for x:=min(reste,xant) downto qsup(reste,kr) do
       begin vec[j]:=x; genkpart(j+1,k,kr-1,x,reste-x) end
    else begin vec[k]:=reste;
        TRAITER LE VECTEUR {cf. Algorithme 2 plus bas, pour exemple}
    end
end.
```

associées à chaque partition ou arrangement sont définies. Par « arrangement », nous désignons ici la série des valeurs de fréquence des k catégories d'éléments et, par « permutations », les répartitions possibles des n éléments donnés à travers les k catégories. Par exemple, la série 5 1 1, une 3-partition de $n = 7$, présente 3 arrangements de fréquences (511, 151, 115), chacun comportant ${}_nC_{f_1,f_2,\ldots,f_k} = {}_7C_{5,1,1} = 42$ permutations d'éléments.

Le poids combinatoire associé à chaque arrangement ou chaque partition détermine aussi les probabilités ponctuelles attachées aux valeurs de la statistique qui en est tirée, la distribution de probabilités reposant donc sur la combinatoire de n éléments répartis en k catégories, d'ordre k^n.

Dans le cas d'une multinomiale égalitaire (ou équiprobable), tous les arrangements et permutations associés à une partition ont un poids équivalent, égal à $Rp = A_p \times P_p$, sommant à k^n pour l'ensemble des partitions; le poids relatif de chaque partition, sa probabilité, est donc de Rp/k^n. Dans le cas d'une multinomiale à probabilités diverses, les A_p arrangements sont distincts et toutes les permutations associées à un arrangement ont un poids déterminé par la conjonction des probabilités de chaque catégorie d'élément, soit Rp_{jA_i}. Le poids global relatif de l'arrangement est de $Rp_{jA_i} \times P_p$ et le poids total des k-partitions somme à $\sum_p Rp_{jA_i} \times P_p = 1$.

Note sur le vocabulaire : Il faut distinguer ici les vocables 'partitions de l'entier n', 'k-partitions pleines' et 'k-partitions complétées'.

— Chaque partition de l'entier n (> 0) est un multiplet formé par une série d'entiers (>0) totalisant à n et énumérés en ordre non croissant. Par exemple, les partitions de $n = 5$ sont : 5, 41, 32, 311, 221, 2111, 11111, et celles de $n = 7$ sont : 7, 61, 52, 511, 43, 421, 4111 331, 322, 3211, 31111, 2221, 22111, 211111, 1111111. Le nombre de partitions de n croît très rapidement; la formule $10^{\sqrt{n+3}-2}$ (attribuée à R. F. Churchhouse par Good, Gover & Mitchell, 1970) en donne une bonne approximation.

— Chaque k-partition pleine[2] de n est un k-tuplet formé par une série de k entiers (>0) énumérés en ordre non croissant et dont le total donne n. Les k-partitions « pleines » désignent ici les partitions comportant k entiers (plus grands que 0). Pour $n = 5$, les 3-partitions pleines sont donc 331 et 221; pour $n = 7$, nous obtenons 511, 421, 331 et 322.

— Chaque k-partition complétée de n est un k-tuplet formé par une série de k entiers énumérés en ordre non croissant et totalisant à n, série éventuellement complétée par un ou plusieurs « 0 ». Les k-

2. Les qualificatifs « pleines » et « complétées » sont spécifiés ici pour aider à distinguer les deux espèces de partitions partielles.

partitions complétées correspondent donc aux t-partitions pleines telles que $t \leq k$. Pour $n = 5$, les 3-partitions complétées sont 500, 410, 320, 311, 221 ; pour $n = 7$, nous avons 700, 610, 520, 511, 430, 421, 331, 322. Les encadrés 1 et 2 décrivent un algorithme permettant de produire l'ensemble des k-partitions complétées d'ordre (k, n).

— Une k-partition comporte k fréquences qui peuvent donner lieu à A_p arrangements selon la diversité de ces fréquences, le nombre d'arrangements étant donné par ${}_kC_{\{\#f\}}$, ou $\#f$ dénote le décompte des fréquences différentes. Par exemple, la 6-partition de 20, {6 4 4 4 1 1}, a comme décompte 1 (6), 3 (444) et 2 (11), d'où ${}_kC_{\{\#f\}} = {}_6C_{1,3,2} = 60$ arrangements possibles (des fréquences à travers les k catégories). Noter que les arrangements ont tous un poids équivalent pour les multinomiales à probabilités égales, ce qui n'est pas le cas autrement. Quant aux P_p permutations d'éléments, elles ont toutes un poids égal dans une partition donnée (et dans ses arrangements), leur nombre étant évidemment ${}_nC_{f_i}$, ici ${}_{20}C_{6,4,4,4,1,1} \approx 2,444 \times 10^{11}$. Le poids total de cette partition est alors de $60 \times 2,444 \times 10^{11} \approx 1,467 \times 10^{13}$, relativement à une masse (ou réalisations possibles) de $k^n = 6^{20} \approx 3,656 \times 10^{15}$.

Exemple 2. Notre exemple A initial présentait $n = 12$ éléments répartis en $k = 4$ catégories, la répartition catégorielle constatée étant 3, 2, 5, 2. La k-partition correspondante 5 3 2 2 fait partie d'un ensemble de 34 partitions, dont les première et dernière sont 12 0 0 0 et 3 3 3 3, et elle chapeaute 12 arrangements, incluant 3 2 5 2. Le nombre de permutations d'éléments pour cette partition et cet arrangement est de ${}_{12}C_{3,2,5,2} = 166320$. Dans une multinomiale à probabilités égales, notre variable 3 2 5 2 s'unit donc aux 11 autres arrangements de la même k-partition et son poids consolidé est de $12 \times 166320 = 1995840$, par rapport au nombre total de réalisations possibles de cet ensemble, soit $k^n = 4^{12} = 16777216$, correspondant à une probabilité globale de $1995840/16777216 \approx 0,119$.

Le nombre de réalisations ou répartitions possibles de n éléments en k catégories correspond à la fonction k^n. Le tableau 1 illustre cette progression.

De la lecture de ce tableau, le lecteur conclura facilement que le traitement exact des distributions multinomiales à taille n élevée et comportant plus de 4 ou 5 catégories reste aujourd'hui un défi informatique considérable, et qu'il est alors excusable de se replier sur des approximations par simulation Monte Carlo.

On trouvera des exposés sur la théorie des partitions et leurs variantes dans les traités sur la combinatoire, p. ex. R. C. Bose et B. Manvel, *Introduction to combinatorial theory*, New York, Wiley 1984 ou R. A. Brualdy, *Introductory combinatorics*, New York, North-Holland, 1977.

La génération de la distribution de probabilités d'une multinomiale à éléments équiprobables est basée ici sur les k-partitions complétées, les k^n réalisations possibles d'une multinomiale $M(k, n)$ étant dérivées des arrangements et permutations associés à chaque k-partition.

Génération des k-partitions complétées [algorithmes en langage Pascal/Delphi] (voir Encadrés 1 et 2)

Le procédé appliqué ici est récursif, la récursion (mathématique ou informatique) servant communément à solutionner nombre de problèmes de combinatoire.

Calcul de la fréquence (R) des réalisations de chaque partition (voir Encadré 3)

Les n éléments (ou fréquences) $x_1 \ x_2 \ x_3 \ldots x_k$ de la k-partition (tels que $x_i \geq x_{i+1}$) génèrent $P_p = {}_nC_{\{x_i\}}$ permutations et, pour chaque élément x_i, son nombre d'exemplaires t_i est compté et leur permutation établie, fournissant $A_p = {}_kC\{t_i\}$ arrangements, la fréquence ou poids de probabilité de la partition dans la distribution multinomiale égalitaire étant $R_p = A_p \times P_p/k^n$.

Par exemple, la partition 4 2 1 1 pour $n = 8$ éléments en $k = 4$ catégories génère $P_p = {}_8C_{4,2,1,1} = 840$ permutations et $A_p = {}_4C_{1,1,2} = 12$ arrangements, son poids ou fréquence étant donc de $12 \times 840 = 10080$ parmi la population des $k^n = 4^8 = 65536$ réalisations des 15 différentes k partitions possibles, soit une probabilité de 0,1538. Le lecteur pourra véri fier que la k-partition 4 2 0 de $n = 6$ a un poids de $R = 90$ ($= 15 \times 6$) parmi les 7 3-partitions de population $3^6 = 729$ éléments, ou une probabilité de 0,1218. Ces nombres croissent « exponentiellement » avec n et k : ainsi, pour la 8-partition 5 4 3 3 2 1 1 1 de $n = 20$ parmi les 434 partitions possibles, nous calculons $R_p \approx 3,9422 \times 10^{16}$ relativement à une population de $8^{20} \approx 1,1529 \times 10^{18}$ réalisations, avec une probabilité de 0,0342.

En ce qui touche les tests X^2, G et G_W, comme d'autres tests de procédure semblable, la valeur du test s'obtient en faisant la somme d'une expression pour chacun des k éléments du multiplet. Or, dans le cas d'une multinomiale inégalitaire, chaque élément a une probabilité π_i spécifique, de sorte que les A_p arrangements associés à chaque partition ont chacun une valeur singulière : par exemple, la partition 5 3 2 associée au vecteur de probabilités { 0,3 0,3 0,4 } donnerait lieu à 6 arrangements (5 3 2, 5 2 3, 3 5 2, etc.), leurs couplages avec les probabilités de chaque élément ([5 :0,3, 3 :0,3. 2 :0,4], [5 :0,3, 2 :0,3. 3 :0,4], etc.) produisant des totaux (généralement) différents. Dans le cas d'une multinomiale égalitaire, tous les éléments du multiplet sont couplés avec une probabilité égale, de sorte que la valeur de la statistique calculée dépend uniquement de la partition. Bref, dans le cas égalitaire, le poids proba-

Tableau 1 ■ k-partition, arrangements et réalisations selon n

n	k	k-partitions	arrangements	réalisations (k^n)
4	3	4	15	81
5	3	5	21	243
10	3	14	66	59 049
25	3	65	351	8,473 $\times 10^{11}$
100	3	884	5 151	5,154 $\times 10^{47}$
100	4	8037	176 851	1,607 $\times 10^{60}$

bilitaire de chaque valeur distincte de la statistique (X^2, G ou G_W) est exactement proportionnel au produit du nombre de permutations avec le nombre d'arrangements qui s'y rattachent. Pour une multinomiale inégalitaire, le nombre possible de valeurs distinctes des statistiques calculées correspond au nombre d'arrangements, chaque arrangement ayant alors le poids permutationnel associé à la partition.

GÉNÉRATION D'UNE MULTINOMIALE PAR SIMULATION

Générer une variable multinomiale (égalitaire) de n éléments en k catégories équiprobables

Définitions

— $V[0..k-1]$, un vecteur de k éléments
— U, U', des v.a.i. uniformes de 0 à 1 ;
— $P[1..npart]$, $R[1..npart]$, des vecteurs où $npart$ = nombre de k-partitions ; le vecteur P contient soit la partition concernée, soit la statistique associée à celle-ci ;
— $T[0..Tmax]$, un vecteur de taille élevée ($Tmax \geq 1000$) comme tableau développé. La technique dite du « tableau développé » avec ses variantes est documentée dans LAURENCELLE (2001), p. 42-49. Voir aussi le splendide ouvrage sur les techniques de simulation statistique de Luc DEVROYE (1986).
— $Pmax$, numéro ($< Tmax$) de la cellule maximale occupée dans T ;
— $Res[0..npart]$, un vecteur de probabilités résiduelles (déchets du tableau développé) standardisées et cumulées.

Le traitement de toutes les multinomiales d'ordre (k, n), qu'elles soient à probabilités égales ou inégales, peut se baser sur l'algorithme proposé à l'encadré 2, en y joignant les calculs de la ou des statistiques choisies et ceux de leurs poids (ou probabilités). Ce calcul exact devenant rapidement prohibitif, en raison du nombre de « réalisations » multinomiales générées, soit k^n, on peut l'approximer par un échantillonnage aléatoire, de type Monte Carlo. Voici deux devis de programmation, la première procédure étant simple et relativement peu efficace, la seconde, basée sur la technique du tableau développé, plus complexe et très rapide.

Algorithme 1 (produire une réalisation aléatoire d'une multinomiale égalitaire d'ordre (k, n), élément par élément)

Produire une k-partition aléatoire d'ordre (k, n)

1. Mettre à zéro le vecteur multinomial V.
2. Itérer n fois < $j \leftarrow \lfloor U \times k \rfloor$ et $V[j] \leftarrow V[j] + 1$ > puis produire V, en y greffant la valeur de la statistique attenante).

Note : Le poids de chaque réalisation V produite est de 1 (parmi les k^n possibles), et sa probabilité égale à 1 / Taille de l'ensemble de multinomiales produites pour cette partition.

Algorithme 2 (produire une multinomiale aléatoire égalitaire d'ordre (k, n) une partition à la fois, incluant son poids caractéristique et optionnellement la valeur de la statistique attenante)

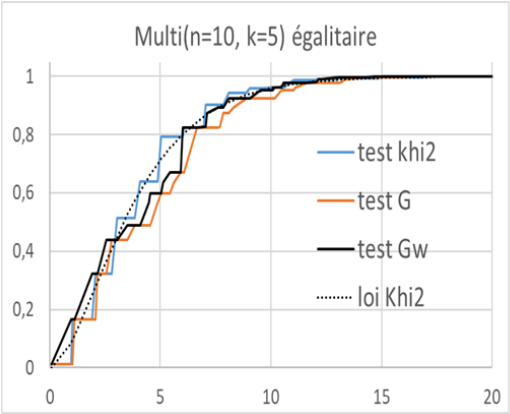
Préparer l'ensemble des k-partitions égalitaires d'ordre (k, n)

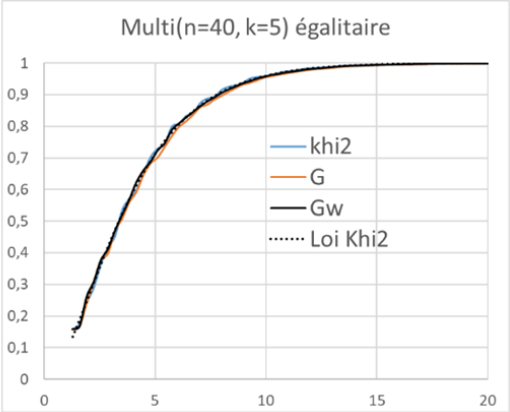
1. Établir un vecteur P des npart k-partitions de n (voir Encadré 2) et un autre R_P de leurs poids correspondants (en ajoutant optionnellement la valeur de la statistique associée à chaque partition).
2. Organiser les données jumelées poids : (vecteur/valeur) selon la technique du tableau développé, le numéro de chaque partition dans P étant inscrit au tableau développé T dans un nombre de cases à peu près proportionnel à son poids, soit ncases= $\lfloor R_P \times Tmax \rfloor$;
3. Récolter les déchets fractionnaires de l'allocation des poids R_P au tableau T, soit $R_P \times Tmax -$ ncases, en les plaçant dans le tableau résiduel Res. En fin de course, réorganiser le tableau Res en ordre décroissant des résidus, en y jumelant les numéros de la statistique associée).

Produire une k-partition aléatoire

1. Obtenir $j \leftarrow \lfloor Tmax \times U \rfloor$; si $j \leq Pmax$ alors $i \leftarrow T\lfloor j \rfloor$ sinon repérer avec i (de 1 à npart] dans Res à partir de U' ;
2. Produire la partition P[i] et son poids R[i] (incluant op-

FIGURE 1 ■ La forme des distributions et leurs moments





tionnellement la statistique associée)

Générer une multinomiale de n éléments en k catégories de probabilités inégales

L'algorithme 1, dans le cas de vecteurs à probabilités inégales, consiste ici, à l'instar de l'algorithme 2 pour probabilités égales, à établir un tableau développé basé non sur les partitions, mais sur les probabilités des éléments du vecteur multinomial, les vecteurs P, R et Res associés étant alors de taille k (plutôt que $npart$). Chaque pige U permet alors de produire un élément selon la probabilité de sa catégorie.

L'algorithme 2, du même niveau que celui appliqué ci-dessus à une multinomiale égalitaire, consiste à appliquer la même logique du tableau développé, cette fois en fonction des arrangements plutôt que des partitions : le poids probabiliste de chaque arrangement $f_1\ f_2 \ldots f_k$ est calculé comme $\pi_1^{f_1} \times \pi_2^{f_2} \times \cdots \times \pi_k^{f_k}$, puis multiplié par le nombre de permutations approprié. Les arrangements étant évidemment plus nombreux que les partitions, la masse de calculs requise est beaucoup plus élevée. Néanmoins, les procédures suggérées pour le cas de probabilités inégales devraient rester réalistes, c.-à-d. tant soit peu moins efficaces, que celles escomptées pour celui des probabilités égales.

Notes complémentaires

Valeurs maximales des tests X^2 et G

Tests calculés sous hypothèse uniforme ($\pi_j = 1/k$). La valeur maximale apparaît sous le multiplet $n\ 0\ 0\ 0 \ldots 0_k$. Pour le X^2, elle est de $n \cdot (k-1)$ et pour le G, de $2n \cdot [\ln(n) - \ln(n/k)] = 2n \ln(k)$.

Tests calculés sous hypothèse à probabilités inégales. La valeur maximale apparaît encore sous multiplet $n\ 0\ 0\ 0 \ldots 0_k$ couplé à $\pi_{min}\ \pi_2\ \pi_3 \ldots \pi_k$ (π_{min} étant associée à la fréquence n, quel que soit son rang dans le multiplet) et $\sum_{j=2}^{k} \pi_{\mathrm{j}} = 1 - \pi_{\mathrm{min}}$. Le maximum possible du calcul X^2 dépend alors de la valeur la plus petite de π, soit π_{min}, et il est de $n \cdot (1 - \pi_{min})/\pi_{min}$ et de $2\ n \cdot \ln\ \pi_{min}$ pour le G.

Étude de la distribution des tests khi2, G et G_W sous une loi multinomiale égalitaire

Dans les panneaux de la figure 1 construits à partir d'une segmentation arbitraire de l'axe des variables, les trois variables en lice (X^2, G, G_W) montrent des fonctions de répartition (FR) saltatoires, à côté de la FR continue de la loi χ^2 (tracée en pointillé), toutes étant découpées par intervalles égaux sur l'axe des abscisses. Tel que prévu, l'irrégularité des trois fonctions discrètes dépend de la granularité de la variable mesurée, c.-à-d. du nombre de valeurs distinctes réparties sur l'axe horizontal. La figure de gauche, pour une multinomiale $M(n = 10, k = 5)$, se décline selon 30 partitions (et arrangements multinomiaux) différentes, celle de droite pour $M(n = 40, k = 5)$, sur 1747 partitions.

Les histogrammes élaborés pour inscrire la FR de nos trois variables dans un classement commun ont pour effet de produire des fonctions graphiques « interpolées », la crête (en ordonnée) sommant chaque bosse de la fonction correspondant ainsi à une valeur de la variable (en abscisse) qui précède sa position sur l'axe qui lui correspond tout juste : c'est ce qu'illustrent les graphiques précédents.

À la figure 2, chaque variable est présentée cette fois séparément, échelonnée selon les valeurs discrètes de sa propre FR. Le modèle de la loi χ^2, on le voit, convient de mieux en mieux à la forme pourtant discontinue des répartitions des tests X^2 (comme l'avait démontré K. Pear-

son) et G (tel qu'indiqué par WILKS, 1935, 1938). Jusqu'à quel point cette correspondance est précise et, si différence il y a entre le X^2 et le G_W (ou le G), est-ce qu'on peut déterminer celui des deux qui serait le plus fiable, le plus proche de la loi χ^2 ? Les analyses suivantes nous éclaireront là-dessus.

Les moments des distributions constituent une façon de mesurer leurs similitudes et différences. Le tableau 2 rapporte ces calculs relativement aux statistiques[3] X^2 et G_W et la loi χ^2, soit l'espérance E, la variance σ^2, l'indice d'asymétrie γ_1 (= $+\sqrt{\beta_3}$) et l'indice d'aplatissement ou voussure γ_2 ($= \beta^4 - 3$), ce pour chacune de nos deux multinomiales.

Dans ce tableau, nous remarquons surtout que l'asymétrie (positive) et la voussure (pointue) de G_W pour $M(n = 10, k = 5)$ sont sensiblement moins fortes que celles de la loi χ^2 et du X^2, ces disparités disparaissant à peu près pour $M(n = 40, k = 5)$. Rappelons que chacune de ces deux distributions comparées (X^2 et G_W) est exacte et qu'il s'agit ici de vérifier à quel point la loi χ^2 leur convient pour en décrire le comportement et en déterminer les probabilités. Les données examinées jusqu'ici montrent qu'effectivement la loi χ^2 décrit globalement et raisonnablement bien l'allure et les propriétés distributionnelles des statistiques X^2, G et G_W, à quelques détails près. Examinons maintenant de plus près cette conformité.

L'écart entre les distributions

Les graphiques de la figure 1 montrent d'abord et permettent de deviner ensuite les écarts plus ou moins importants entre les FR présentées, soit entre les FR de chaque variable par rapport à celle de la loi χ^2. Ces écarts entre deux FR peuvent être individuellement évalués par la différence $d_{1,2}(x) = [P(x : f_1) - P(x : f_2)]$, la fonction f_1 étant ici la fréquence cumulative relative de la variable discontinue, la fonction f_2 étant la FR de la loi χ^2 correspondant à chaque valeur x de la variable discrète.[4] Le test de Kolmogorov-Smirnov (SOKAL & ROHLF, 1981) cible l'écart absolu maximal, soit $d_{max} = \sqrt{C} \times max|d_{1,2}(x)|$, où C est le nombre de classes de fréquences comparées : il permet de tester approximativement l'importance (la vraisemblance) de cet écart sous hypothèse nulle d'équivalence.

Le tableau 3 fournit des statistiques sur la comparaison des FR déjà documentées, statistiques toutes issues de la différence $d = d_{\text{Var vs Loi}}(x) = [P(x : \text{Variable calculée}) - P(x : \text{Loi } \chi^2)]$. Ces statistiques sont :

— $m(d)$: moyenne des différences $d = d_{\text{Var,Loi}}(x)$ à travers les x ;
— $\Sigma(d^2)$: somme des différences carrés ;
— $m(|d|)$: moyenne des différences absolues ;
— $max(|d|)$: valeur de différence absolue maximale.

Ces statistiques sont établies d'abord sur toute l'étendue de chaque variable, depuis sa valeur minimale jusqu'à son maximum, puis sur l'aile positive de l'étendue, soit celle débutant juste avant $P(x) \leq 0,90$ jusqu'à $P(x) = 1$, le second calcul voulant documenter la portion des FR dans laquelle se logent traditionnellement les valeurs critiques (P = 0,1, 0,025, 0,05, 0,005, 0,001) appliquées dans les tests d'hypothèses. Les tableaux 3 et 4 résument ces calculs.

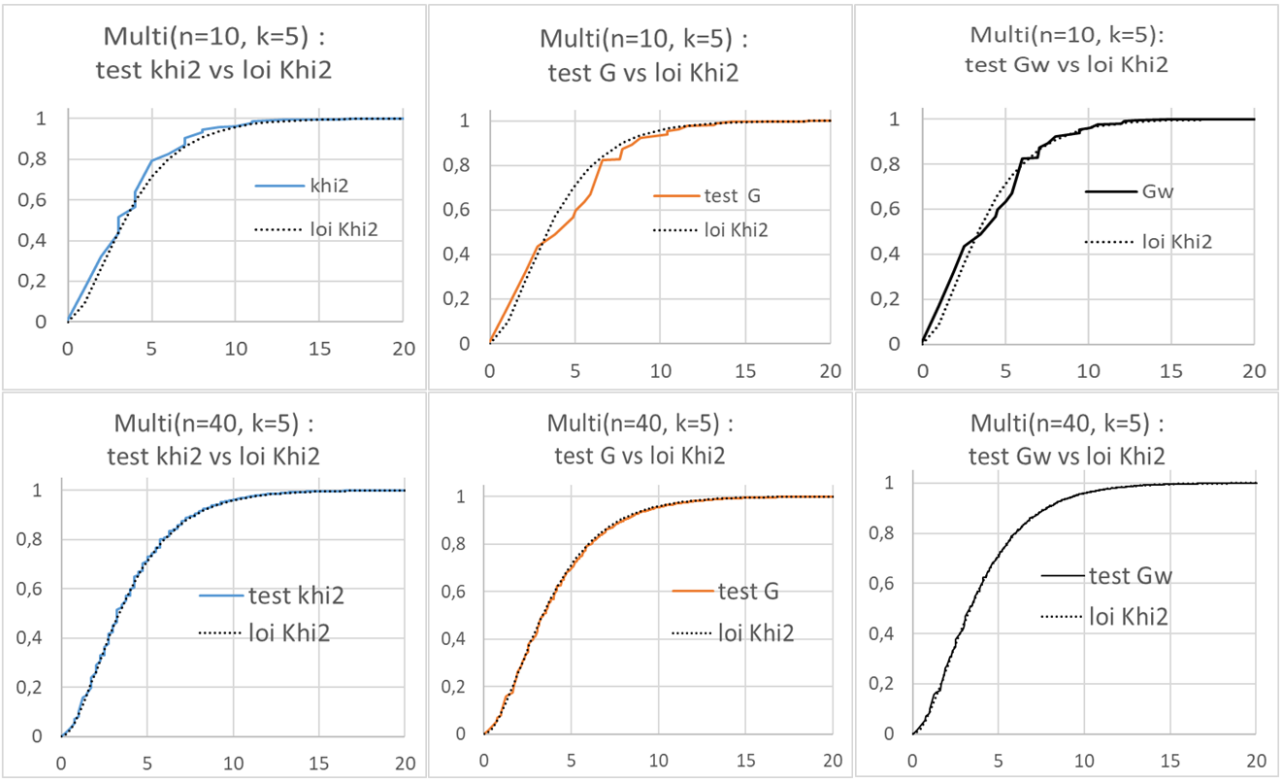
Le lecteur aura réalisé que l'amplitude de l'écart (d) entre distributions varie inversement avec le nombre (C) de strates ou valeurs que doit traverser chaque FR entre 0 à 1 ; aussi, n'est-il pas surprenant que les modèles à plus grande expansion, par exemple les $M(n = 40, k = 5)$ par rapport aux $M(n = 10, k = 5)$, présentent de plus petites valeurs d'écart : l'estimation à faire tient alors à la comparaison d'un test avec l'autre pour la même multinomiale. Quant aux valeurs calculées de la statistique $max(|d|)$, que considère le test Kolmogorov-Smirnov, elles n'atteignent pas les valeurs critiques appliquées (au seuils de 0,05 et 0,01), sauf dans le cas du X^2 au second tableau : pour ces données, on peut donc considérer que les variables X^2, G et G_W ne s'écartent pas dramatiquement de la loi du χ^2. Néanmoins, les statistiques descriptives de ces écarts permettent de jauger la proximité des distributions.[5] Primo, les indices associés au test G_W sont systématiquement meilleurs (plus bas) que ceux du G non corrigé, ce qui nous annonce une conclusion à venir en faveur de la version corrigée. Secundo, les indices de G_W sont tous meilleurs que ceux du X^2, notamment dans la zone de probabilité de 0,90 à 1, là où se jouent les tests de significativité statistique. Cela indique que, à tout le moins pour ces deux multinomiales, le test G_W peut remplacer, voire qu'il remplace avantageusement le test X^2. L'examen ponctuel et anec-

3. Nous omettons ici la statistique G non corrigée (au profit de G_W), leurs indices de forme γ_1 et γ_2 étant identiques et seules l'espérance et la variance de G étant légèrement plus élevées. Les données de la section suivante montreront à l'évidence que G_W frôle de plus près que G la répartition de la loi χ^2.

4. D'autres tests sur l'« égalité » des FR sont aussi disponibles, notamment en comparant d'une part les fréquences générées par la multinomiale considérée et ses partitions aux fréquences tirées de la loi χ^2 pour les intervalles de valeurs correspondants. Appliquerions-nous alors un calcul du khi2 interprété par le χ^2, ou par le G_W interprété par lui-même, ou bien... ? La présente étude étant essentiellement exploratoire et la preuve récursive étourdissante, nous nous en tiendrons ici à nos statistiques descriptives et au petit test (ponctuel) du Kolmogorov-Smirnov.

5. La grandeur absolue des statistiques d'écarts d, lesquelles représentent les différences entre deux probabilités cumulatives et sont donc bornées entre 0 et 1, est fortement influencée par la granularité des variables en jeu (examinée à la section suivante). Cette granularité dépend ici du nombre de partitions de la multinomiale concernée : ce nombre est de 30 pour $M(n = 10, k = 5)$ et de 1747 pour $M(n = 40, k = 5)$, d'où des indices de différences plus minces pour la seconde multinomiale.

FIGURE 2 ■ Fonctions de distribution (FR) comparées des statistiques X^2, G et G_W *vs* celle de la loi χ^2 pour deux multinomiales égalitaires



dotique d'autres multinomiales (p. ex. $M(10,3)$, $M(20,3)$, $M(20,5)$, $M(50,7)$) aboutit à la même conclusion.

La granularité, la précision et l'exactitude des valeurs critiques

Granularité, précision et exactitude pour un test de significativité sont des propriétés inter-reliées, la variable multinomiale donnant lieu à des valeurs discrètes dont la FR saute d'une valeur à l'autre, de façon discontinue, et certaines k-partitions distinctes produisent des résultats de calcul égaux, aussi appelés *répétats*, le plus souvent dans le cas du X^2 comme on verra bientôt. Une granularité distributionnelle grossière, c.-à-d. faible, donnera lieu à des sauts de probabilité plus importants, réduisant la précision du test et faussant ou brouillant éventuellement la conclusion de significativité. La granularité ne serait parfaite que dans le cas d'une variable continue, dans laquelle chaque configuration distincte de données donne lieu à une valeur distincte de la variable, la granularité n'étant alors limitée que par la précision (les décimales !) des valeurs obtenues.

La granularité optimale. On retrouve deux niveaux globaux de granularité dans les multinomiales. Pour les multinomiales égalitaires (à probabilités égales des k éléments), la granularité maximale correspondrait au nombre de k-

Tableau 2 ■ Moments des statistiques X^2, G et G_W *vs* ceux de la loi χ^2 pour deux multinomiales égalitaires

		$M(n=10,k=5)$		$M(n=40,k=5)$	
	loi χ^2	test X^2	test G_W	test X^2	test G_W
E	4,000	4,000	4,256	4,000	4,014
σ^2	8,000	7,200	8,094	7,800	8,132
γ_1	1,414	1,416	0,883	1,414	1,445
γ_2	3,000	3,472	0,759	3,143	3,198

Tableau 3 ■ Indices d'écart des statistiques X^2, G et G_W vs le χ^2 pour une multinomiale égalitaire $M(n = 10, k = 5)$*

	Toute l'étendue			L'étendue de $P = 0,9$ à $P = 1$		
	test X^2	test G	test G_W	test X^2	test G	test G_W
$m(d)$	0,0181	-0,0207	-0,0011	0,0098	0,0060	0,0033
$\Sigma(d^2)$	0,0299	0,0837	0,0458	0,0051	0,0024	0,0007
$m(\|d\|)$	0,0202	0,0314	0,0241	0,0099	0,0067	0,0043
$max(\|d\|)$	0,0817	0,1323	0,0836	0,0399	0,0258	0,0155

Note. * : Les valeurs critiques applicables pour le test Kolmogorov-Smirnov sont, respectivement, aux seuils de 5% et 1%, de 0,2417 et 0,2899 pour toute l'étendue (C = 30), ou de 0,3014 et 0,3612 pour l'aile droite évaluée (C = 19) (cf. Tableau 32 dans Rohlf et Sokal, 1981).

Tableau 4 ■ Indices d'écart des statistiques X^2, G et G_W vs le χ^2 pour une multinomiale égalitaire $M(n = 40, k = 5)$*

	Toute l'étendue			L'étendue de $P = 0,9$ à $P = 1$		
	test X^2	test G	test G_W	test X^2	test G	test G_W
$m(d)$	0,00058	-0,00121	0,00003	0,00058	-0,00121	0,00003
$\Sigma(d^2)$	0,01786	0,02148	0,00683	0,00203	0,00450	0,00026
$m(\|d\|)$	0,00100	0,00134	0,00051	0,00038	0,00063	0,00015
$max(\|d\|)$	0,03326**	0,02704	0,03149	0,00845	0,01091	0,00412

Note. * : Les valeurs critiques *approximatives* pour le test Kolmogorov-Smirnov sont, respectivement, aux seuils de 5% et 1%, de 0,0325 et 0,0390 pour toute l'étendue ($C = 1747$), et de 0,0338 et 0,0405 pour l'aile droite seulement ($P(x) \geq 0,90$, C = 1604 à 1615), cf. Rohlf et Sokal, 1981. ** $p < 0,05$.

partitions tandis que, pour les multinomiales inégalitaires (à probabilités inégales), elle correspond au nombre d'arrangements, substantiellement plus élevé puisque mettant en jeu un vecteur de probabilités inégales. Par exemple, pour une $M(n = 10, k = 5)$ vs une $M(n = 40, k = 5)$, nous comptons respectivement 30 vs 1747 k-partitions, et 1001 et 135751 arrangements. Estimant (grossièrement) la granularité (Gra) comme le fractionnement du domaine [0 .. 1] de la FR en petits intervalles de probabilité, soit $Gra \approx 1/$ (nombre de valeurs distinctes), les variables $M(n = 10, k = 5)$ et $M(n = 40, k = 5)$ égalitaires ont pour granularité optimale $1/30 \approx 0,0333$ et $1/1747 \approx 0,0006$, et les inégalitaires, $1/1001 \approx 0,0010$ et $1/135751 \approx 0,0^5737$.[6]. Il ressort de ces premières constatations (a) que les multinomiales à paramètres n et k plus élevés ont une plus forte granularité (correspondant à des « grains » de probabilité plus petits) et que les inégalitaires ont une granularité plus fine que les égalitaires.

La granularité réelle. Une mesure plus raffinée, et plus juste, de l'utilité d'une mesure est sa capacité discriminante, c.-à-d. le nombre de discriminations et de catégories distinctes qu'elle peut occasionner *dans un contexte d'incertitude maximale*, c.-à-d. pour des catégories à probabilités égales. Or, les « catégories » ou niveaux des FR de nos variables sont très loin d'être équiprobables : pour une $M(n = 20, k = 5)$ contrastez par exemple les vecteurs (20 0 0 0 0) et (4 4 4 4 4). Dans le cas d'éléments équiprobables, le premier vecteur n'a que 1 chance sur 5^{20} de se produire (donnant un $X^2 = 80$) contre $\approx 3,055 \times 10^{11}$ chances pour le second (donnant des $X^2 = 0$), le poids relatif du premier ne contribuant guère à la puissance de mesure de nos variables. Laurencelle (2014), éq. 19', propose un indice de capacité discriminante, de formule $k^* = 1/\sum_{-\infty}^{\infty} p_j^2$, estimant le nombre de catégories (ou intervalles) de valeurs équiprobables que la mesure peut présenter, les p_j étant les probabilités individuelles que les valeurs naturelles (vraisemblablement non équiprobables) contiennent. Le tableau 5 documente cet aspect.

Prenons l'exemple de la multinomiale égalitaire $M(n = 40, k = 5)$. Cette multinomiale génère 1747 vecteurs ou 5-partitions, depuis (20 0 0 0 0) jusqu'à (4 4 4 4 4). De ce nombre, le X^2 produit 369 valeurs distinctes, certaines partitions occasionnant donc un total de 1378 valeurs répétées de la statistique, alors que le G (ou G_W) en distingue 1724, sans répétitions : la perte d'information (sur les compositions différentes des vecteurs multinomiaux) du X^2 est donc considérable. Si on imagine maintenant des distributions qui seraient aplanies, dans lesquelles les données auraient été redistribuées en catégories de capacités égales, quel nombre de catégories obtiendrions-nous alors dans chaque cas. Ce nombre, la capacité discriminante k^* le fournit : pour notre exemple,

6. L'exposant interne dénote le nombre de « 0 » à intercaler dans la valeur inscrite, p. ex. $0,0^5737$ indique 0,00000737.

Tableau 5 ■ Granularité de diverses multinomiales égalitaires

Modèle	Nombre de partitions	Valeurs distinctes		Capacité discriminante ($k*$)		Granularité réelle ($1/k*$)	
		X^2	G, G_W	X^2	G, G_W	X^2	G, G_W
$M(10,5)$	30	22	29	7,424	9,745	0,135	0,103
$M(20,5)$	192	85	186	14,327	25,259	0,070	0,040
$M(30,5)$	674	195	665	21,376	47,478	0,047	0,021
$M(40,5)$	1747	369	1724	28,417	76,199	0,035	0,013
$M(50,5)$	3765	593	37226	35,442	110,914	0,028	0,009
$M(75,5)$	16019	1446	15954	52,997	222,960	0,019	0,004
$M(100,5)$	46262	2701	46102	70,554	370,030	0,014	0,003

Note. * Les valeurs critiques approximatives pour le test Kolmogorov-Smirnov sont, respectivement, aux seuils de 5% et 1%, de 0,0325 et 0,0390 pour toute l'étendue de x, et de 0,0338 et 0,0405 pour l'aile droite seulement ($P(x) \geq 0,90$) (cf. Rohlf et Sokal, 1981). ** $p < 0,05$

les 369 catégories inégales du X^2 équivalent à 28,417 catégories à probabilités égales et information maximale, aboutissant à une granularité de $\approx 0,035$; pour le G (ou G_W), les nombres passent de 1724 à 76,199, pour une granularité de 0,013.

La concordance des valeurs critiques. Pour des fins de tests de significativité, tests relatifs à l'hypothèse nulle : *'Les données se comportent comme reflétant des éléments catégoriels équiprobables'*, la disponibilité de valeurs critiques reste une considération importante. Les valeurs critiques issues de la loi χ^2 conviennent-elles dans ce contexte de données multinomiales égalitaires ?

Le lecteur aura déjà compris que, pour le traitement de toutes données discrètes, le recours à une loi continue, telle celle du χ^2, reste à valider, sans quoi il serait nécessaire de trouver une fonction appropriée à ces données ou bien de déterminer et publier de longues tables de valeurs critiques établies par calculs individuels. Le tableau 6 présente un échantillon indicatif de valeurs critiques, applicables aux seuils (unilatéraux droits) de significativité α de 0,10, 0,05 et 0,01. On peut y faire trois constatations provisoires :

1. Pour les modèles et seuils critiques explorés, la loi χ^2 navigue moyennement près des valeurs critiques réelles établies pour nos trois tests.
2. Le χ^2 apparaît systématiquement plus proche du test G_W que du G brut, la correction de (Williams, 1976) à cette fin s'avérant efficace.
3. Globalement, les valeurs du test G_W concordent davantage avec le χ^2 que celles du test X^2.

Le lecteur doit encore se rappeler que les valeurs critiques indiquées pour les tests X^2, G et G_W sont toutes exactes (sous réserve de l'arrondissement occasionnel des nombres) tandis que la loi χ^2 se propose comme une approximation. Cela étant et vu la commodité de recourir à la fonction χ^2 pour finaliser un test d'hypothèses statistiques, il est pertinent de vérifier auquel des trois tests examinés l'approximation sied le mieux.

Le tableau 7, établi selon le même contexte de modèles multinomiaux et de valeurs que le précédent, fait voir la zone de rejet (ou probabilité extrême) délimitée pour chaque cas par l'application de la valeur critique du modèle théorique du χ^2, la grandeur de zone attendue étant celle du seuil α. Le caractère approximatif du critère χ^2 ressort à nouveau, la probabilité obtenue se révélant rarement juste et plus souvent conservatrice (critère trop sévère) que libérale (critère trop bas). L'approximation s'améliore évidemment avec l'augmentation de la taille (n) de la multinomiale et de sa division (k). De nos trois tests comparés, les versions G et G_W apparaissent ici équivalentes et toutes deux sont plus souvent mieux approchées par la loi χ^2 que n'est le test X^2.

Étude de la distribution des tests khi2, G et G_W sous une loi multinomiale inégalitaire

Le traitement exact d'une multinomiale à probabilités inégales repose sur les mêmes principes que celui des multinomiales égalitaires : le déploiement des distributions prend son origine dans les k-partitions concrétisant la structure du vecteur multinomial, chacune de ces partitions se déployant en un ou en plusieurs arrangements, chaque arrangement chapeautant les permutations d'éléments semblables qui le composent. La différence ici tient à ce que, les k catégories d'éléments traités ayant des probabilités différentes, chaque arrangement d'une k-partition reçoit un poids de probabilité, c.-à-d. une portion de combinatoire, variable : les opérations de traitement des données s'en trouvent multipliées. Par exemple, une $M(10,5)$ donne lieu à 30 partitions et 1001 arrangements (avec $5^{10} = 9,766 \times 10^6$ réalisations); pour une $M(30,5)$, on a 674 partitions et 46376 arrangements (avec $5^{30} = 9,313 \times 10^{20}$ réalisations); pour une $M(50,5)$,

Tableau 6 ■ Valeurs critiques exactes des tests X^2, G et G_W versus les valeurs correspondantes du χ^2 (voir Note)

	$\alpha = 0,10$			$\alpha = 0,05$			$\alpha = 0,01$		
Modèles	X^2	G	G_W	X^2	G	G_W	X^2	G	G_W
$M(10,5)$	7	8,823	8,021	9	10,411	9,464	13	13,380	13,285
$M(30,5)$	$7,667^{3/8}$	8,147	7,885	$9,333^{2/10}$	9,928	9,607	$13^{2/7}$	14,368	13,905
$M(50,5)$	$7,6^{4/6}$	7,910	7,755	$9,4^{4/14}$	9,697	9,507	$13,2^{0/6}$	13,551	13,285
$M(100,5)$	$7,7^{13/20}$	7,864	7,780	$9,5^{2/20}$	9,576	9,482	$13,2^{13/24}$	13,428	13,295
χ^2_4		7,779			9,488			13,277	
$M(10,3)$	3,8	4,013	3,762	6,2	8,109	7,602	9,8	9,755	9,145
$M(30,3)$	$4,2^{1/2}$	4,587	4,487	$5,6^{1/2}$	6,313	6,175	$8,6^{1/2}$	9,316	9,113
$M(50,3)$	4,48	4,737	4,675	6,04	5,980	5,901	8,92	9,375	9,251
$M(100,3)$	4,58	4,622	4,592	6,02	6,051	6,011	9,14	9,243	9,181
χ^2_2		4,605			5,991			9,210	

Note. Le rapport de forme a/b placé en exposant de quelques valeurs du test X^2 indique et décrit un débordement de seuil. Par exemple, le rapport 3/8 pour le X^2 critique d'un $M(10,5)$ au seuil $\alpha = 0,10$ signifie que 3 configurations quelconques de données parmi les 8 fournissant la valeur 7,667 n'atteignent pas le seuil ou, autrement dit, que seules 5 configurations sur les 8 produisant un X^2 = 7,667 ont un probabilité $P \leq 0,10$. L'occurrence des débordements dans le calcul du X^2 dépend en partie de la conjonction des facteurs premiers des paramètres n et k, leur fréquence dépendant surtout de n, et nous avons constaté qu'ils n'adviennent que très rarement pour les tests G et G_W.

Tableau 7 ■ Zone extrême réelle délimitée dans la FR de chaque test par les valeurs critiques théoriques du χ^2

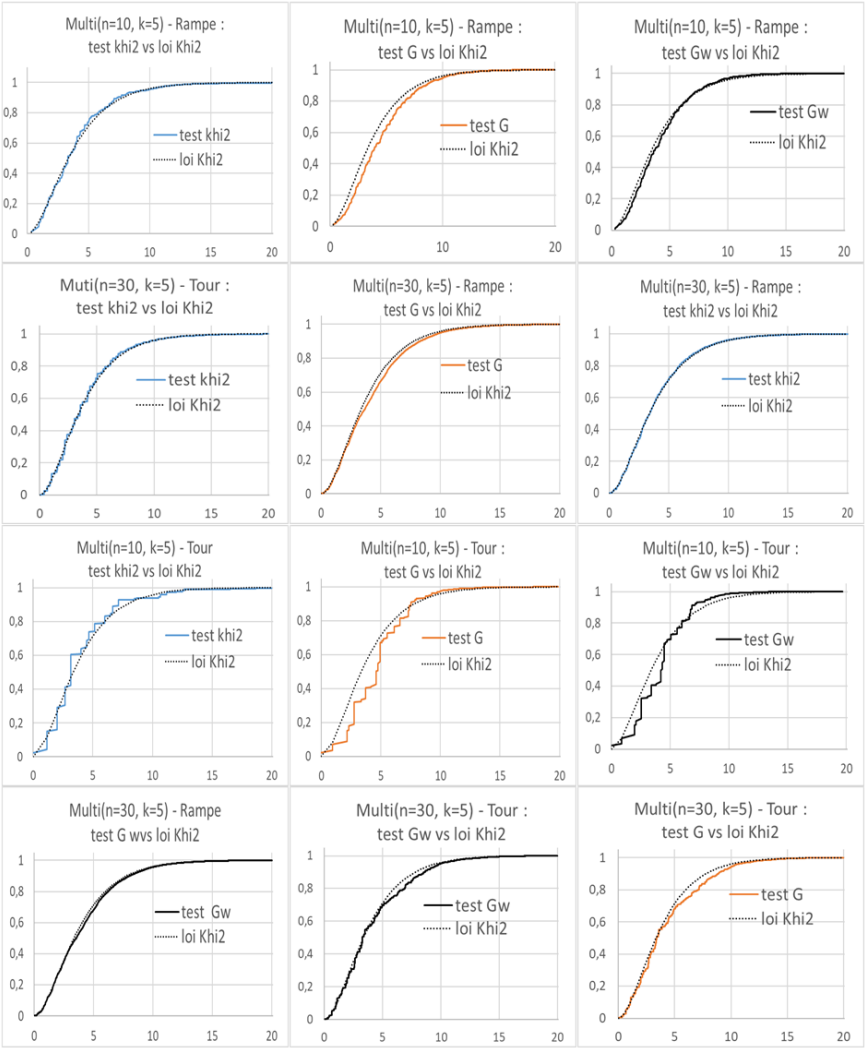
	$\alpha = 0,10$			$\alpha = 0,05$			$\alpha = 0,01$		
Modèles	X^2	G	G_W	X^2	G	G_W	X^2	G	G_W
$M(10,5)$	0,065	0,106	0,075	0,037	0,060	0,037	0,005	0,010	0,001
$M(30,5)$	0,084	0,118	0,101	0,039	0,055	0,049	0,009	0,014	0,012
$M(50,5)$	0.094	0,105	0,098	0,044	0,054	0,050	0,008	0,010	0,010
$M(100,5)$	0,097	0,103	0,100	0,045	0,052	0,050	0,009	0,011	0,010
$M(10,3)$	0,080	0,057	0,057	0,022	0,044	0,044	0,006	0,006	0,001
$M(30,3)$	0,080	0,086	0,086	0,033	0,051	0,051	0,009	0,010	0,007
$M(50,3)$	0,092	0,092	0,092	0,046	0,045	0,045	0,008	0,009	0,009
$M(100,3)$	0,092	0,096	0,092	0,051	0,051	0,049	0,010	0,010	0,009

on a 3765 partitions et 316251 arrangements (avec $5^{50} = 8,882 \times 10^{34}$ réalisations). Quelle que soit l'efficacité de calcul de nos ordinateurs, il reste des opérations qu'il est malcommode ou peu parcimonieux de transférer à la machine d'où, vu la lourdeur des fichiers de données, nous nous sommes contentés ici de quelques coups de sonde pour estimer la performance (exacte) des trois tests à l'étude.

Les graphiques de la figure 3 montrent les distributions calculées pour deux modèles multinomiaux à probabilités inégales, soit le modèle « Rampe », à probabilités proportionnelles aux nombres 1, 2, 3, 4 et 5 (où $\pi_j = x_j/15$) et le modèle « Tour », calqué sur les nombres 1, 1, 1, 1, 6 (où $\pi_j = x_j/10$), chaque ensemble de réalisations étant basé sur des échantillons de $n = 10$ et $n = 30$ éléments.

Le modèle « Rampe », qui correspond à un contexte de croissance régulière des fréquences, nous informe aussi sur le comportement qu'auraient les tests dans le cas, par exemple d'une vérification de la représentativité d'un échantillon de personnes par rapport à leur population d'origine (voir Laurencelle, 2012, pour une application avec le X^2), celle-ci étant campée en k (ici 5) catégories. Il appert ici que les tests X^2 et G_W s'en tirent tous deux très bien quant à leur approximation par la loi χ^2. Pour le modèle « Tour », un exemple d'inégalités assez importantes, la loi continue χ^2 peine à épouser la croissance saltatoire de la FR multinomiale, le test X^2 bénéficiant de la même approximation. Notons toutefois que, dans les quatre ensembles de cas présentés, l'aile droite des FR, là où se logent les zones critiques des tests, les approximations χ^2 semblent grossièrement comparables d'un test à l'autre.

FIGURE 3 ■ FR des statistiques X^2, G et G_W vs la loi χ^2 pour 4 multinomiales inégalitaires



L'écart entre les distributions

Les tableaux 8 à 11 donnent le résultat des calculs de l'écart entre la FR de chaque test et celle du χ^2 pour les quatre cas représentés plus haut.

Les indices non normés, tels $m(d)$, $\Sigma(d^2)$ et $m(|d|)$, apparaissent ici à l'avantage du test X^2 et occasionnellement du G, quoique les différences entre les tests soient souvent légères. Il reste notable que la plupart des sanctions du Kolmogorov-Smirnov, soit 13 sur 18, se montrent significatives, 6 sur 9 l'étant dans la portion critique (droite) des distributions, soit partout sauf pour les tests G et G_W dans le cas du modèle Rampe avec $n = 10$. Alors, dans les cas de tests à base multinomiale de probabilités inégales, serait-il requis de contrôler chaque fois le résultat en générant une distribution exacte ou Monte Carlo appropriée ? Le tableau 12 propose une réponse à cette question.

L'examen des enjambements de la zone critique des tests (Tableau 12) nous rassure plus ou moins sur la précision de l'approximation χ^2 même si, pour les tests X^2 et G_W, la probabilité rapportée par le test jouxte assez souvent le seuil α imposé. Le modèle Tour, formé sur une disposition très déséquilibrée des probabilités multinomiales, présente des écarts d'approximation légèrement plus importants que le modèle Rampe, ces deux modèles jurant avec ceux des modèles égalitaires examinés plus haut, l'approximation par la loi χ^2 y étant systématiquement meilleure.

Tableau 8 ■ Indices d'écart des statistiques X^2, G et G_W vs le χ^2 pour une multinomiale inégalitaire $M(n = 10, k = 5)$ de modèle « Rampe »*

	Toute l'étendue			L'étendue de $P = 0,9$ à $P = 1$		
	test X^2	test G	test G_W	test X^2	test G	test G_W
$m(d)$	0,028	-0,024	-0,0035	0,004	-0,0029	0,0042
$\Sigma(d^2)$	0,1254	1,4156	0,2974	0,0143	0,0319	0,0212
$m(\|d\|)$	0,0064	0,0208	0,0100	0,0023	0,0033	0,0042
$max(\|d\|)$	0,0413	0,1112‡	0,0669‡	0,1008‡	0,0263	0,0114

Note. * Les valeurs critiques applicables pour le test Kolmogorov-Smirnov sont, respectivement, aux seuils de 5% et 1%, de 0,0429 et 0,0515 pour toute l'étendue de x. Pour l'aile droite seulement ($P(x) \geq 0,90$), elles sont de 0,0505 et 0,0605 au X^2 et de 0,0509 et 0,0610 aux G et G_W. (cf. Rohlf et Sokal, 1981). ‡ $P < 0,01$.

Tableau 9 ■ Indices d'écart des statistiques X^2, G et G_W vs le χ^2 pour une multinomiale inégalitaire $M(n = 30, k = 5)$ de modèle « Rampe »*

	Toute l'étendue			L'étendue de $P = 0,9$ à $P = 1$		
	test X^2	test G	test G_W	test X^2	test G	test G_W
$m(d)$	-0,0015	-0,0023	-0,0015	-0,0001	-0,0010	-0,0004
$\Sigma(d^2)$	0,1744	4,0021	1,37093	0,0235	0,4027	0,0887
$m(\|d\|)$	0,0007	0,0030	0,0016	0,0003	0,0010	0,0004
$max(\|d\|)$	0,0245‡	0,0557‡	0,0380‡	0,0068†	0,0192‡	0,0098‡

Note. * Les valeurs critiques applicables pour le test Kolmogorov-Smirnov sont, respectivement, aux seuils de 5% et 1%, de 0,0063 et 0,0076 pour toute l'étendue de x. Pour l'aile droite seulement ($P(x) \geq 0,90$), elles sont de 0,0065 et 0,0078 (cf. Rohlf et Sokal, 1981). † $P < 0,05$, ‡ $P < 0,01$.

Pour le chercheur rigoureux et qui doit avoir recours à l'approximation par la loi χ^2, le prochain test qu'il devra faire sur ses données sera-t-il juste ?

Conclusions sur les distributions

Kendall et Stuart (1979, chap. 30, notamment p. 445 seq.) discute en profondeur l'application du test X^2 et du G (sous la dénomination LR ou « log ratio de vraisemblances », notamment pour le cas multinomial appliqué au test d'ajustement. Ces formes de calcul ont toutes deux la loi χ^2 comme distribution asymptotique sous $n \to \infty$. Pour les cas de tailles n plus courantes, la discontinuité et l'irrégularité relatives des probabilités multinomiales font que l'approximation χ^2 est imparfaite et mérite d'être considérée.

— *La granularité de la distribution.* La variable calculée, X^2, G ou G_W, saute d'une valeur à l'autre par intervalles irréguliers, le χ^2 estimant sa probabilité par bonds : la largeur d'intervalle et la hauteur du bond produisent l'imprécision de l'estimation rendue. Cette granularité dépend d'abord des paramètres n et k et elle augmente avec eux, ce qui explique la conformité asymptotique de distribution de la variable multinomiale avec le χ^2. Or, le calcul de la variable X^2

Tableau 10 ■ Indices d'écart des statistiques X^2, G et G_W vs le χ^2 pour une multinomiale inégalitaire $M(n = 10, k = 5)$ de modèle « Tour »*

	Toute l'étendue			L'étendue de $P = 0,9$ à $P = 1$		
	test X^2	test G	test G_W	test X^2	test G	test G_W
$m(d)$	0,0017	-0,0119	-0,0003	0,0033	0,0034	0,0091
$\Sigma(d^2)$	0,3396	2,1226	1,2507	0,0518	0,0326	0,2348
$m(\|d\|)$	0,0078	0,0178	0,0168	0,0033	0,0034	0,0091
$max(\|d\|)$	0,1341‡	0,2426‡	0,1914‡	0,0561‡	0,0234	0,0552†

Note. * Les valeurs critiques applicables pour le test Kolmogorov-Smirnov sont, respectivement, aux seuils de 5% et 1%, de 0,0429 et 0,0515 pour toute l'étendue de x. Pour l'aile droite seulement ($P(x) \geq 0,90$), elles sont de 0,0464 et 0,0557 au X^2 et de 0,0466 et 0,0559 aux G et G_W. † $P < 0,05$ ‡ $P < 0,01$.

Tableau 11 ■ Indices d'écart des statistiques X^2, G et G_W vs le χ^2 pour une multinomiale inégalitaire $M(n = 30, k = 5)$ de modèle « Tour »*

	Toute l'étendue			L'étendue de $P = 0,9$ à $P = 1$		
	test X^2	test G	test G_W	test X^2	test G	test G_W
$m(d)$	0,0001	-0,0027	0,0018	-0,0001	-0,0007	-0,0004
$\Sigma(d^2)$	0,4491	4,9789	2,7367	0,0397	0,0000	0,0000
$m(\|d\|)$	0,0008	0,0027	0,0019	0,0003	0,0007	0,0005
$max(\|d\|)$	0,0443‡	0,0726‡	0,0606‡	0,0123‡	0,0360‡	0,0360‡

Note. * Les valeurs critiques applicables pour le test Kolmogorov-Smirnov sont, respectivement, aux seuils de 5% et 1%, de 0,0063 et 0,0076 pour toute l'étendue de x. Pour l'aile droite seulement ($P(x) \geq 0,90$), elles sont de 0,0065 et 0,0077 (cf. Rohlf et Sokal, 1981). ‡ $P < 0,01$.

Tableau 12 ■ Zone extrême réelle délimitée dans la FR de chaque test par les valeurs critiques théoriques du χ^2

	$\alpha = 0,10$			$\alpha = 0,05$			$\alpha = 0,01$		
Modèles	X^2	G	G_W	X^2	G	G_W	X^2	G	G_W
$M(10,5)$ Rampe	0,082	0,127	0,098	0,052	0,065	0,039	0,011	0,009	0,005
$M(30,5)$ Rampe	0,093	0,122	0,110	0,049	0,064	0,055	0,011	0,013	0,011
$M(10,5)$ Tour	0,065	0,075	0,056	0,055	0,030	0,016	0,010	0,005	0,001
$M(30,5)$ Tour	0,085	0,148	0,129	0,051	0,073	0,065	0,013	0,015	0,013

étant plus grossier que celui des G et G_W (recourant à une mise au carré plutôt qu'un logarithme), le X^2 sépare moins bien les configurations multinomiales différentes et leurs probabilités. Les tests G et G_W, avec une granularité plus importante, bénéficient ainsi du même effet d'approximation que les paramètres n et k. Good et collègues. (1970), comparant les distributions lissées du G et du χ^2, arrivent à la même conclusion.

— *Multinomiales à probabilités égales ou inégales.* Kendall et Stuart (1979, p. 455-458), qui aborde la question de l'impact du classement des données d'une multinomiale sur la précision du test X^2, cite Wise (1963), lequel, se basant sur une étude des probabilités exactes de deux petits modèles, $M(n = 10, k = 4$, tous $\pi_i = 0,25)$ et $M(n = 8, k = 3, \pi_i = 0,25, 0,25, 0,5)$, conclut entre autres que les probabilités réelles du X^2 sont mieux respectées et l'approximation χ^2 meilleure lorsque les probabilités π_i et, particulièrement, les produits $n_i \cdot \pi_i$ (ou « fréquences théoriques ») sont égaux ou presque égaux. Toutefois, le regroupement de valeurs X^2 consécutives en classes hybrides afin d'égaliser leurs poids de probabilité (surtout dans l'aile droite de la distribution) risque d'amortir la sensibilité du test de significativité et de compromettre sa validité.
Dans le cas de notre étude, on a pu vérifier que les multinomiales inégalitaires produisent des distributions non pas moins régulières, mais moins bien modélisées par la loi χ^2. Sans être dramatique, cet écart de distribution cautionne, à notre avis, la recommandation suivante : emprunter la loi χ^2 pour tester les variables X^2 et G_W sous hypothèse d'équiprobabilité, et pour l'hypothèse d'altéroprobabilité, recourir plutôt à une table de valeurs critiques exactes ou à une preuve distributionnelle directe (exacte ou par simulations Monte Carlo) si $n \leq 30$, cette borne étant arbitraire.

— *En général.* Des données de l'ensemble de notre étude, étude exploratoire et par définition incomplète et provisoire, il ressort que le test X^2 classique garde généralement sa valeur et sa crédibilité dans le contexte considéré. Quant au test G, il nous apparaît globalement évident que la correction de Williams (1976) s'impose, le test G_W se révélant être un compétiteur valable, souvent plus précis par rapport à la loi χ^2 que n'est le test X^2, et systématiquement plus discriminant.

Décomposition additive des sources de variation pour les tests G et G_W

Par sa méthode de calcul, le test G (ou G_W) se prête facilement à une décomposition additive (et orthogonale) de ses composants (Shaffer, 1973; Sokal & Rohlf, 1981; Black & Laurencelle, 1987), ce qui rend possible une forme d'analyse de variation semblable à l'analyse de variance et permet, par exemple, de fractionner un effet global en parties ou bien de mesurer des effets principaux et leur interaction de premier ordre ou d'ordre supérieur. Quant au test X^2, la décomposition additive de sa valeur, appelée aussi partitionnement, est possible (Castellan,

1965; Bresnahan & Shapiro, 1966; Shaffer, 1973; Fagen & Mankovich, 1980; Beh, 2001), mais elle est complexe, le nombre de propositions faites à cet égard témoignant à lui seul qu'il s'agit de méthodes plutôt 'savantes', mutuellement peu compatibles et qu'on voit mal recommander au chercheur.[7]

Nous illustrons ci-après trois contextes d'application du test G avec ses techniques. Pour ce faire et pour alléger le texte et la lecture, nous utiliserons la forme initiale du G, dépourvue de la correction de Williams (1976) et redonnée à l'équation 8,

$$G_W = G/q_W, \text{ où } q_W = 1 + k^2/(6 \cdot n \cdot dl). \tag{8}$$

Le G_W possède toutes les propriétés du G original, la correction visant seulement à rapprocher sa distribution de celle de la loi χ^2.

Une note initiale permettra de bien contextualiser notre exposé. Soit un vecteur de fréquences quelconque $f_1 \quad f_2 \ldots f_k$ associé à un vecteur de catégories donné : l'intérêt porte ici sur la variation des données du premier vecteur. Or, contrairement aux données de l'analyse de variance, lesquelles sont des mesures, le vecteur étudié ici représente en fait une répartition de fréquences, chaque fréquence référant à des mesures ou des entités comparables ou identiques. Plutôt que de porter sur une question concernant le niveau moyen ou la variance d'une mesure d'un élément particulier dans un échantillon ou une population, on s'intéresse aux nombres d'éléments qui présentent l'une ou l'autre des catégories d'entités considérées dans l'étude. Cette différence de contexte n'interdira pas d'appliquer les principes d'analyse statistique utiles et bien connus pour les mesures, il faudra cependant y opérer certaines transpositions : c'est ce que nous nous proposons de faire dans les alinéas suivants.

Le partitionnement d'un vecteur.

Exemple 3. Soit le vecteur de fréquences 2 6 4 8 . Sous l'hypothèse d'égalité des probabilités, la formule (6) s'applique, soit $G = 2[\Sigma_{i=1}^{k} f_i \ln f_i - n \ln(n/k)]$. Cette formule, on l'a vu, exprime un quotient de vraisemblances, celle liée aux données observées versus celle découlant du modèle de probabilité (ici, une modèle égalitaire, sans variation), mais elle reflète aussi, comme la formule du X^2,([8]) la variation qu'il y a entre les éléments du vecteur analysé. Le vecteur représente $n = 2+6+4+8 = 20$ éléments classés en $k = 4$ catégories, d'où[9] :

$$\begin{aligned} G &= 2[2\ln 2 + 6\ln 6 + 4\ln 4 + 8\ln 8 - 20\ln(20/4)] \\ &\approx 4,2576 \text{ (}dl\text{ = 3), ;} \end{aligned}$$

comme pour la variance (et les calculs du X^2), les degrés de liberté (dl) associés au G sont le nombre de fréquences libres de varier sous les contraintes imposées. Dans le cas d'un test d'égalité de k composants, seul le total n est (ordinairement) fixé d'avance, d'où $dl = k - 1$, soit ici, 3.

La valeur du G ayant 3 degrés de liberté, elle peut être décomposée, c.-à-d. que l'analyse peut être appliquée sur 2 voire 3 sous-vecteurs mutuellement orthogonaux. Présentons le vecteur original complet 2 6 4 8 sous la forme 2 | 6 | 4 | 8 , nous considérons en premier exemple les 3 sous-vecteurs 2 | 6 4 8 , 6 | 4 8 et 4 | 8 , de 1 degré de liberté chacun. La notation $a \mid bcd$, par exemple, indique la présence de deux groupes d'éléments homogènes, c.-à-d. que les valeurs à traiter dans chaque groupe sont égalisées. Le vecteur $a \mid bcd$ équivaut ainsi au vecteur $a \mid (b+c+d)/3, (b+c+d)/3, (b+c+d)/3$. Le partitionnement 2 | 6 4 8 devient donc 2 6 6 6 , 6 | 4 8 devient 6 6 6 et 4 | 8 demeure 4 8. Ces 3 vecteurs partitionnés sont dits orthogonaux (c.-à-d. indépendants) en ce que les sommes des produits des coefficients du calcul des paires de sous-vecteurs sont toutes nulles, comme le lecteur peut le vérifier d'après le tableau ci-dessous :[10]

Fréquences	2	6	4	8
{2 | 6 4 8 }	3	−1	−1	−1
{6 | 4 8}	0	2	−1	−1
{4 | 8}	0	0	1	−1

Par exemple, pour le premier sous-vecteur 2 | 6 4 8 , le premier groupe, unique, est « 2 », et le second, triple, est « 6 4 8 » devenant « 18/3 18/3 18/3 », d'où :

7. L'obstacle principal à une décomposition additive simple du test X^2 vient de sa formule de calcul, soit l'ingrédient $(f_{oi} - f_{ti})^2/f_{ti}$ qui comporte une division par un dénominateur (f_{ti}) variable, lequel introduit un effet de non-linéarité dans le calcul additif global; tel n'est pas le cas pour le test G et son calcul.

8. En effet, pour un test du X^2 sous hypothèse égalitaire, celui-ci est une simple reformulation de la variance, selon l'équation $X^2 = s^2k(n-1)/n$.

9. Pour le lecteur curieux, cette valeur de G, dotée de $k - 1 = 3$ degrés de liberté, a selon la loi χ^2 une probabilité extrême (p_{ext}) de 0,235, qui aboutirait à un non-rejet de l'hypothèse d'égalité des catégories. Le G_W correspondant serait de 4,0873 ($p_{ext} = 0,252$) et le test X^2 est de 4,0 ($p_{ext} = 0,261$) : nous revenons plus bas à ce X^2. L'objectif de cette section étant d'expliquer les procédés de décomposition du G, nous ne reviendrons pas ici sur l'estimation de sa significativité.

10. Deux vecteurs sont dits mutuellement orthogonaux si leur produit scalaire (i.e. la somme des produits de leurs éléments correspondants) est nul. Les coefficients présentés au tableau en exemple satisfont cette condition.

$$G\{2 \mid 6\ 4\ 8\} = 2 \cdot [2\ln 2 + 18/3 \ln 18/3 + 18/3 \ln 18/3 + 18/3 \ln 18/3 - 20 \ln(20/4)$$
$$= 2 \cdot [2\ln 2 + 18 \ln(18/3) - 20 \ln(20/4)] \approx 2,8984.$$

Nous obtenons de même :

$$G\{6|4\ 8\} = 2 \cdot [6 \ln 6 + 12 \ln(12/2) - 18 \ln(18/3) = 0$$

et

$$G\{4 \mid 8\} = 2 \cdot [4 \ln 4 + 8 \ln 8 - 12 \ln(12/2)] \approx 1,3592,$$

la somme des G des 3 sous-vecteurs orthogonaux étant $\approx 4,2576$ ($= 2,8984 + 0 + 1,3592$), égale comme prévu au $G = 4,2576$ global. Le vecteur initial, basé sur $k = 4$ groupes et possédant $k-1 = 3$ degrés de liberté, a ainsi été parfaitement décomposé en 3 sous-vecteurs de 2 groupes dotés chacun de 1 degré de liberté.

Sous hypothèse d'égalité des probabilités, une formule générale de calcul pour tout vecteur ou sous-vecteur constitué de c groupes serait :

$$G\{g_1, g_2, \ldots | \sum_j k_j\} = 2 \cdot [t_1 \ln(t_1/k_1) + t_2 \ln(t_2/k_2) + \cdots + t_c \ln(t_c/k_c) - \sum_j t_j \ln(\sum_j t_j / \sum_j k_j)], \tag{9}$$

où t_j et k_j dénotent respectivement la somme et le nombre des éléments du groupe j.

La décomposition du X^2 global. Le X^2 global de notre vecteur 2 | 6 | 4 | 8 est, par notre formule (2), de $3 \sum_{j=1}^{k} f_j^2/20 = 4$. Le calcul des X^2 des trois sous-vecteurs examinés plus haut donne X^2(2 | 6 4 8) = X^2 (2 6 6 6) $\approx 2,4$, X^2(6 | 4 8) = X^2(6 6 6) = 0 et X^2(4 | 8) = $X^2(4\ 8) \approx 1,333$, la somme des trois composants (orthogonaux) étant 3,733, différente du G global de 4. Une autre décomposition orthogonale donnerait encore un autre total, par exemple celle des sous-vecteurs 4 4 | 6 6 , 4 4 et 6 6, le total des 3 valeurs G étant $\approx 4,133$. L'usager peut s'accommoder de ces « inexactitudes », toutefois un calcul à la fois orthogonal et additif reste préférable, est plus élégant et vérifiable, et il est promis par le test, le G, lequel a autant sinon plus de mérites distributionnels que le X^2 global.

Les exercices suivants portant sur deux décompositions distinctes d'un vecteur permettront au lecteur de mieux s'approprier la méthode de calcul.

Exercice 1. Le vecteur 2 2 3 3 4 4 5 5 comporte $k = 8$ sommant à $n = 28$. Soit les 4 sous-vecteurs 2 | 2 , 3 | 3 , 4 | 4 , 5 | 5 et le sous-vecteur 2 2 | 3 3 | 4 4 | 5 5 . Montrer que ces sous-vecteurs sont mutuellement orthogonaux et que la somme de leurs valeurs G égale le G du vecteur original.

Réponse à l'exercice 1. Le G global de 2,9433 (7 degrés de liberté) se décompose en 4 G de valeur nulle (1 degré de liberté chacun) associés aux 4 sous-vecteurs à groupes égaux, en plus du $G = 2.9433$ (3 degrés de liberté) associé au sous-vecteur comparant *entre eux* les 4 précédents.

Exercice 2. Soit le même vecteur initial qu'à l'exercice 1 et les 3 sous-vecteurs 2 | 2 | 3 , 3 | 4 | 4 | 5 | 5 et 2 2 3 | 3 4 4 5 5 . Effectuer les mêmes opérations que ci-dessus,

Réponse à l'exercice 2. Du même G global de 2,9433 (7 degrés de liberté), on détache les $G\{2|2|3\} \approx 0,2747$ (2 degrés de liberté), $G\{3|4|4|5|5\} \approx 0,6876$ (4 degrés de liberté) et $G\{2\ 2\ 3 \mid 3\ 4\ 4\ 5\ 5\} \approx 1,9810$ (1 degré de liberté), ces $k - 1 = 3$ G partiels sommant au G global.

L'analyse d'un tableau factoriel simple

Exemple 4. Le tableau 13 présente sous l'hypothèse (a) une matrice 3×4 de fréquences $f_{i,j}$ observées, flanquée à droite, en italique, des totaux de rangée, et en bas des totaux de colonnes, la matrice à droite sous (b) donnant les valeurs $f_{i,j} \ln f_{i,j}$ correspondantes. Les 12 éléments du tableau varient des valeurs 3 à 12, comme varient aussi les totaux de rangées et de colonnes. Le test G permet de quantifier ces variations (qui à proprement parler ne sont pas des variances), ce qui se fait encore par notre simple formule (6). Pour la clarté de l'exposé qui suivra, il est utile d'expliciter toutes les phases du calcul. La formule (6), ici $G = 2 \cdot \left[\sum_{i=1}^{k} f_i \ln f_i - n \ln(n/k)\right]$, suppose que l'on calcule d'abord les produits $f_{i,j} \ln f_{i,j}$ constituant la partie b du tableau 13.

Tableau 13 ■ Données de fréquences sous deux dimensions, avec calculs de base

(a)	$f_{i,j}$			Total	(b)	$f_{i,j} \ln f_{i,j}$	
	5	8	7	*20*	8,047	16,636	13,621
	9	6	6	*21*	19,775	10,751	10,751
	3	3	9	*15*	3,296	3,296	19,775
	12	8	7	*27*	28,819	16,636	13,621
Total	*29*	*25*	*29*	*83*			

Le G global. La somme des valeurs du tableau 1b est $\approx 166,023$, d'où $G = 2 \cdot [166,023 - 83 \ln(83/12)] \approx 11,013$: il s'agit ici du G global, mesurant la variation globale présente dans le tableau, sans considération de sa structure en rangées et colonnes.

Le G des rangées et des colonnes. La variation du tableau 13a attribuable aux rangées du tableau 1a est reflétée par les totaux de rangées, et la formule (6) s'y applique encore, soit $G = 2 \cdot [20 \ln 20 + 21 \ln 21 + 15 \ln 15 + 27 \ln 27 - 83 \ln(83/4)]$, le total de 83 étant ici divisé par le nombre de groupes (ou rangées) considérés. Nous obtenons $G_{\text{rangées}} \approx 3,513$. Le lecteur vérifiera que, pour les colonnes, nous avons $G_{\text{colonnes}} \approx 0,392$, une mesure de variation plus petite que pour les rangées, comme il se doit ici.

Le G résiduel ou d'interaction. La mesure de variation globale étant $G_{global} \approx 11,013$, celles des rangées et colonnes en ayant extrait une quantité de 3,513 et 0,392, soit 3,905, il en reste 11,013 - 3,905 = 7,107 à expliquer. La formule, évidente, est donc :

$$G_{\text{résiduel}} = G_{\text{global}} - G_{\text{rangées}} - G_{\text{colonnes}}, \qquad (10)$$

illustrant l'additivité parfaite de l'analyse par le test G.

Pour des fins d'évaluation statistique, ces quatre quantités s'appuient respectivement sur $R \cdot C - 1$, $R - 1$, $C - 1$ et $(R-1)(C-1)$ degrés de liberté (*dl*), soit, pour $R = 4$ et $C = 3$, $dl_{global} = 11$, $dl_{\text{rangées}} = 3$, $dl_{\text{colonnes}} = 2$ et $dl_{\text{résidu}} = 6$. Cette variation résiduelle, familière aux praticiens de l'analyse de variance, peut être attribuée à une interaction, c.-à-d. un effet combiné (et non additif) des rangées et colonnes, c.-à-d. des facteurs de catégorisation de données en rangées et en colonnes.

Le composant résiduel, ou d'interaction, obtenu par simple soustraction, peut aussi est calculé directement dans le cas du G. Ce résidu émane du fait qu'à chaque fréquence conjointe $f_{i,j}$ sont associés les totaux (r_i) de la rangée et de la colonne (c_j) concernés. Si les effets présumés des facteurs représentés par la rangée et la colonne étaient indépendants l'un de l'autre, la valeur $f_{i,j}$ serait proportionnelle à leurs influences séparées. Or, la probabilité d'aboutir dans la rangée i est estimée par sa proportion, $\pi_{i,.} = r_i/n$, n étant la somme de toutes les fréquences, tout comme la probabilité d'apparaître dans la colonne j est estimée par $\pi_{.,j} = c_j/n$, de sorte que, s'il n'y a pas d'interaction, la probabilité de tomber dans la cellule de coordonnées (i, j) est $\pi_{i,j} = \pi_{i,.} \cdot \pi_{.,j} = c_i \times r_j/n^2$. La valeur de fréquence attendue selon cette hypothèse et ce calcul, dite fréquence théorique, est $ft_{i,j} = n \times \pi_{i,j} = c_i \times r_j/n$, à confronter avec la fréquence observée $f_{i,j}$. Cette fréquence théorique est la même que celle appliquée dans le test X^2 classique sur un tableau de fréquences.

Le tableau 14 en partie (a) fournit ces fréquences théoriques. Par exemple, pour la cellule de coordonnées (1, 1) avec $f_{1,1} = 5$, nous avons $c_1 = 29$ et $r_1 = 20$, d'où $f_{t_{1,1}} = 29 \times 20/83 \approx 6,988$.

S'il y avait eu une (double) proportionnalité stricte dans les données originales du tableau 13a, c.-à-d. une indépendance statistique des rangées et des colonnes, les valeurs rapportées auraient été très proches voire égales à celles du tableau 14a. La mesure attribuable à une variation proportionnelle dans le tableau 13a s'obtiendrait, comme d'habitude, par l'application de la formule (6) sur les données $f_{i,j} \ln ft_{i,j}$, au tableau 14b, soit $G_{\text{proportionnel}} = 2 \cdot [9,721 + 13,366 + \cdots + 15,710 - 83 \ln(83/12)] = 2 \cdot (162,469 - 160,517) \approx 3,904$. Rappelant que le G global a été évalué à 11,012, sa différence d'avec la mesure proportionnelle, soit $11,012 - 3,904 \approx 7,108$, nous redonne la portion de variation non proportionnelle du tableau original. Le lecteur aura noté que cette seconde approche du calcul de la variation résiduelle, plus laborieuse qu'une simple soustraction, a tout de même l'avantage de mettre en relief les éléments du tableau qui sous-tendraient davantage la non-proportionnalité. Pour ce faire, il suffira de confronter les éléments des tableaux 13b à 14b et d'en juger les différences, le G de non-proportionnalité étant simplement la somme des différences individuelles $G_{i,j} = 2 \cdot [f_{i,j} \ln f_{i,j} - f_{i,j} \ln f_{t_{i,j}}]$.

Le test de sources de variation. Le G, on l'a vu plus haut, repose sur un calcul de probabilité et peut, de lui-même, servir à évaluer la vraisemblance d'un modèle probabiliste appliqué à un ensemble de données de fréquences, l'approximation de sa distribution par la loi χ^2 étant cependant raisonnablement bonne et surtout plus commode. Y a-t-il quelque chose de significatif dans les données de fréquences de notre tableau 13 ? Nous référant aux valeurs critiques de la loi χ^2 au seuil α unilatéral droit de 0,05,

Tableau 14 ■ Fréquences « théoriques » $ft_{i,j}$ et leurs expressions $ft_{i,j}$ relatives au tableau 13.

(a)	$ft_{i,j}$			(b)	$f_{i,j} \ln ft_{i,j}$		
	6,988	6,024	6,988		9,721	13,366	13,609
	7,337	6,325	7,337		17,937	11,067	11,959
	5,241	4,518	5,241		4,970	4,524	14,909
	9,434	8,133	9,434		26,932	16,767	15,710

nous obtenons $G_{\text{rangées}}(dl = 3) < 7,815$, $G_{\text{colonnes}}(dl = 2) < 5,991$, $G_{\text{résidu}}(dl = 6) < 12,592$, d'où rien de significatif ne ressort de cet exemple … pédagogique.

Le calcul des « effets simples ». En analyse de variance, on désigne par « effets simples » la portion de variance imputable à un facteur du plan (ou tableau) d'analyse, cela pour chaque niveau d'un autre facteur avec lequel il est croisé. Soit les facteurs Rangées et Colonnes dans un plan à deux facteurs ou dimensions. La variation Rangées reflète l'effet global, c.-à-d. moyen, des Rangées tel que mesurable pas la somme ou moyenne de chaque Rangée à travers toutes les Colonnes du tableau. Cependant, chaque Rangée individuelle, c.-à-d. la combinaison de facteurs expérimentaux qu'elle représente, peut donner lieu à un profil de variation qui lui est propre et que l'effet global des Rangées va dissimuler en partie en le moyennant à travers les Colonnes. Ce sont ces variations idiosyncrasiques de l'effet Rangée et de l'effet Colonne qu'on retrouve exprimées dans le terme résiduel qu'on vient de calculer.

Les données de fréquences de notre exemple 4 (Tableau 14) ont donc permis d'estimer les effets principaux des facteurs Rangée ($G = 3,513$) et Colonne ($G = 0,392$) ainsi que celui de leur interaction ($G = 7,107$). Dans les cas d'analyses se rapportant à un plan à deux ou à plusieurs facteurs croisés, il est aussi possible de tester l'effet d'un facteur *pour chaque niveau ou chaque combinaison de niveaux des autres facteurs* concernés : ces effets *décroisés* sont désignés sous l'appellation « effets simples ».

Effets simples des colonnes. Considérons d'abord les effets simples du facteur Colonnes. Au tableau 13a, les colonnes de la rangée 1 présentent les fréquences 5 8 7, la statistique G pour cette rangée est encore de forme $G = 2\left[\sum_j^k f_j \ln f_j - \sum_j^k f_j/k\right]$, k dénotant le nombre d'éléments considérés, soit $G(C : \text{L1}) = 2 \cdot [5 \ln 5 + 8 \ln 8 + 7 \ln 7 - 20 \ln(20/3)] \approx 0,734$, le premier effet simple du facteur Colonne. Les effets simples des Colonnes aux rangées 2 à 4 sont G(C :L2) = 0,824, G(C :L3) = 4,450 et G(C :L4) = 1,501, chaque statistique étant dotée de $k-1 = 2$ degrés de liberté. La somme des effets simples d'un facteur correspond à l'effet principal (ou global) du facteur, à quoi s'ajoute l'effet d'interaction qu'il peut avoir avec le facteur associé. La somme des 4 valeurs calculées, soit $0,723+0,824+4,450+1,501 \approx 7,498$ est donc égale au total de l'effet Colonnes (0,392) et de son interaction avec l'effet Rangées (7,107), soit 7,499 (aux arrondissements près). D'ailleurs, le lecteur aura remarqué que le total des degrés de liberté extraits par les effets simples des Colonnes, soit $dl = 4 \times 2 = 8$, est un remaniement des degrés de liberté de l'effet principal des Colonnes (2) et de l'interaction des Colonnes avec les Rangées (2×3), l'ensemble des variations considérées étant analysées sous des angles différents.

Effets simples des rangées. À titre d'exercice, le lecteur est invité à faire le calcul des effets simples des données des rangées du tableau 13a.

Le test X^2 en comparaison. Comparons une fois de plus les calculs G opérés précédemment aux calculs correspondants du X^2, en recourant encore une fois aux données du tableau 13a. Utilisant notre formule (2), nous obtenons, à tour de rôle, le $X^2_{\text{global}} = 4 \times 3 \left(\sum_{i=1}^4 \sum_{j=1}^3 f_{i,j}^2\right)/n$, où $n = \sum_{i=1}^4 \sum_{j=1}^3 f_{i,j}$, soit $12 \cdot (52+82+72+\cdots+272)/83 - 83 \approx 10,542$. Nous obtenons de même le $X^2_{\text{rangées}} \approx 3,506$ et le $X^2_{\text{colonnes}} \approx 0,386$. Par soustraction, nous avons un $X^2_{\text{résiduel}} \approx 6,650$. Par ailleurs, le calcul direct du X^2 résiduel ou d'interaction est en fait la procédure classique initialement proposée par K. Pearson (et amendée ensuite par R. A. Fisher), utilisant la formule (1) et les « fréquences théoriques » $ft_{i,j}$ déjà exposées et exploitées dans le calcul du test G. Aux données de fréquences originales $f_{i,j}$ présentées au tableau 13a correspondent les $ft_{i,j}$ inscrites au tableau 14a. L'application de la formule (1) à ces données, soit $\sum_{i=1}^4 \sum_{j=1}^3 (f_{i,j} - ft_{i,j})^2/ft_{i,j}$ fournit le X^2 d'interaction de Pearson, ici 7,343. La différence avec le 6,650 résiduel, différence due à la non-additivité stricte du X^2, n'est pas légère : le lecteur en jugera.

Mise en garde. Le G et le X^2 sont deux tests basés sur la différence entre les fréquences comparées, leurs valeurs dépendent entièrement et uniquement de la grandeur de cette différence, compte tenu du vecteur de probabilités qui accompagne ces catégories de fréquences. Ces deux tests, à sanction de significativité unilatérale droite, sont donc ici essentiellement bilatéraux. Par exemple, le test sur une proportion, selon les hypothèses $H_0 : E(x/n = \pi)$ contre $H_1 : E(x/n > \pi)$ ne peut être directement instrumenté par le G (ou le X^2), dont la distribution de probabilités évolue selon les valeurs calculées croissantes de G (ou X^2) plutôt que selon les valeurs croissantes de la variable

x, cette seconde option étant la bonne. Or, ces deux distributions ne coïncident que rarement. Les tests G et X^2 ne sont donc pas recommandés, notamment pour les tests à contre-hypothèse unilatérale.

Épilogue

Comparativement au test X^2, le test G présente une plus forte granularité, laquelle rend ses valeurs plus fidèles aux probabilités qu'elles représentent et contribue à augmenter sa précision. On peut analyser, c.-à-d. décortiquer additivement un G global en parties indépendantes, comme on le fait pour des variables continues par analyse de variance, ce que le X^2 ne permet de faire que gauchement, sinon après des manœuvres mathématiques complexes. Enfin, à défaut de disposer d'une loi générative des distributions du G ou d'un calcul informatique laborieux, l'approximation du G par la loi χ^2 s'y est montrée très acceptable, équivalant en général à celle rendue pour le X^2 et la surclassant souvent.

Remerciements

L'auteur tient à remercier Denis Cousineau, professeur et rédacteur de TQMP, pour son aide inappréciable à structurer et clarifier le texte soumis. Ses suggestions ont permis de transformer un amalgame hétéroclite de contenus en un article à peu près cohérent.

Références

BEH, E. J. (2001). Partitioning Pearson's chi-squared statistics for singly ordered two-way contingency tables. *Australian & New-Zeland Journal of Statistics*, *43*, 327.

BLACK, P. & LAURENCELLE, L. (1987). Le test G pour les tableaux de fréquences et sa décomposition orthogonale. *Lettres Statistiques*, *8*, 97-114.

BRESNAHAN, J. L. & SHAPIRO, M. M. (1966). A general equation and technique for the exact partitioning of chi-square contingency tables. *Psychological Bulletin*, *66*, 2525-262.

CASTELLAN, N. J., Jr. (1965). On the partitioning of contingency tables. *Psychological Bulletin*, *64*, 330-33.

COCHRAN, W. G. (1952). The chi-2 test of goodness of fit. *Annals of Mathematical Statistics*, *23*, 315-345.

CRESSIE, N. & READ, T. R. C. (1989). Pearson's X2 and the loglikelihood ratio statistic G2 : A comparative Review. *International Statistical Review / Revue Internationale de Statistique*, *57*, 19-43.

DEVROYE, L. (1986). *Non-uniform random variate generation*. New York : Springer-Verlag.

FAGEN, R. M. & MANKOVICH, N. J. (1980). Two-act transitions, partitioned contingency tables, and the 'significant cells' problem. *Animal Behavior¸*, *28*, 1017-1023.

GOOD, I. J., GOVER, T. N. & MITCHELL, G. J. (1970). Exact distributions for X2 and for the likelihood-ratio statistic for the equiprobable multinomial distribution. *Journal of the American Statistical Association*, *65*, 267.

HEALY, M. J. R. (1969). Exact tests of significance in contingency tables. *Technometrics*, *11*, 393-395.

HOEFFDING, W. (1965). Asymptotically optimal tests for multinomial distributions. *Annals of Mathematical Statistics*, *36*, 369-401.

JOHNSON, N. L., KOTZ, S. & BALAKRISHNAN, N. (1977). *Discrete multivariate distributions*. New York : Wiley.

KENDALL, M. & STUART, A. (1979). *The advance theory of statistics, vol. 2 : Inference and relationship (4e édition)*. New York : Macmillan.

LAURENCELLE, L. (1978). Permutations complètes par rotations récursives. *Lettres statistiques*, *4*, 1-11.

LAURENCELLE, L. (2001). *Hasard, nombres aléatoires et méthode Monte Carlo*. Québec : Presses de l'Université du Québec.

LAURENCELLE, L. (2012). La représentativité d'un échantillon et son test par le Khi-deux [Testing the representativeness of a sample]. *The Quantitative Methods for Psychology*, *8*, 173-181.

LAURENCELLE, L. (2014). The discriminating capacity of a measuring instrument : Revisiting Bloom (1942)'s theory and formula. *The Quantitative Methods for Psychology*, *10*, 5-12.

LAURENCELLE, L. (2020). Une étude statistique des séries de deux ou plusieurs types d'éléments. *The Quantitative Methods for Psychology*, *16*, 390-416.

READ, T. R. C. (1984). Small-sample comparisons for the power divergence goodness-of-fit statistics. *Journal of the American Statistical Association*, *79*, 929-935.

ROHLF, F. J. & SOKAL, R. R. (1981). *Statistical tables (2e édition)*. New York : Freeman.

SHAFFER, J. P. (1973). Testing specific hypotheses in contingency tables : Chi-square partitioning and other methods. *Psychological Reports*, *33*, 343-348.

SOKAL, R. R. & ROHLF, F. J. (1981). *Biometry : the principles and practice of statistics in biological research (3e edition)*. New York : Freeman.

WILKS, S. S. (1935). The likelihood test of independence in contingency tables. *Annals of Mathematical Statistics*, *6*, 190-196.

WILKS, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Mathematical Statistics*, *9*, 60-62.

WILLIAMS, D. A. (1976). Improved likelihood ratio tests for complete contingency tables. *Biometrika*, *63*, 33-37.

WISE, M. E. (1963). Multinomial probabilities and the ?2 and X2 distributions. *Biometrika*, *50*, 145-154.

WISE, M. E. (1964). A complete multinomial distribution compared with the X2 approximation and an improvement of fit. *Biometrika*, *51*, 277-281.

Citation

LAURENCELLE, L. (2022). Les distributions multinomiales, leur mesure par les tests khi2 et G, leur approximation par la loi khi-carré et l'analyse des tableaux de fréquences par le test G. *The Quantitative Methods for Psychology*, *18*(1), 1-20. doi :10.20982/tqmp.18.1.p001

Copyright © 2022, *LAURENCELLE*. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Received: 10/04/2021