CAHIERS DU BURO

I. C. LERMAN

Étude distributionnelle de statistiques de proximité entre structures finies de même type ; application à la classification automatique

Cahiers du Bureau universitaire de recherche opérationnelle. Série Recherche, tome 19 (1973), p. 3-53

<http://www.numdam.org/item?id=BURO_1973__19__3_0>

I – INTRODUCTION *

La question du choix d'une mesure de proximité entre structures algébriques de même type s'est posée à nous, de façon cruciale, dans le cadre du problème de la synthèse automatique de l'information contenue dans un tableau de données, à partir d'une hiérarchie de classifications des variables mises en jeu ; classification basée sur les ressemblances entre les comportements de la population étudiée vis à vis de chaque couple de variables. Cette population définit un ensemble E d'objets ou de sujets.

Nous distinguons principalement cinq types de variables dans les Sciences Humaines ; ces types se différencient par la structure algébrique qu'ils déterminent sur l'ensemble E :

a) La variable "attribut de description", *indicatrice d'une* partie de E ; celle formée de tous les éléments de E qui possèdent l'attribut.

b) La variable "caractère descriptif" *indicatrice d'une partition sur E ;* le caractère présente un ensemble fini de modalités sur lequel n'existe aucune structure, une même classe de la partition est formée du sous ensemble des objets de E possédant une certaine modalité du caractère.

Certes un attribut de description a définit bien une partition de E en deux classes $\{E_a , E_a^c\}$ où E_a (resp. E_a^c) est l'ensemble des éléments de E où a est présent (resp. absent) ; mais ce n'est pas pour autant que l'"attribut de description" pourra être considéré comme un "caractère à deux modalités" : en effet, les deux classes E_a et E_a^c sont à priori inégalement considérées par le spécialiste car c'est la présence de l'attribut qui est jugée significative et non son absence.

c) La variable "caractère aux modalités totalement ordonnées" *indicatrice d'un préordre total sur E ;* le caractère présente ici un ensemble fini de modalités sur lequel est donnée une structure d'ordre total. Une même classe du préordre est formée de l'ensemble des objets de E possédant

---

* Ce travail a fait l'objet d'un rapport du Centre de Mathématiques Appliquées et de Calcul (Maison des Sciences de l'Homme), Avril 1972

une certaine modalité du caractère. Rappelons que la donnée d'un préordre total est équivalente à la donnée d'une partition et d'un ordre total sur l'ensemble des classes de cette dernière.

d) La variable "rang" *indicatrice d'un ordre total sur E.* Cette variable est algébriquement un cas particulier de la précédente ; celui où chaque classe du préordre contient exactement un élément.

e) La variable "mesure" *indicatrice d'une mesure positive sur E.* Cette variable affecte à chaque élément de E un nombre réel positif.

Soit un tableau de données qui croise E avec un ensemble A d'éléments descriptifs ; nous supposons que A est formé de variables d'un même type algébrique. Rappelons que notre problème se situe dans le cadre de la recherche d'une chaîne de partitions permettant par l'étude des ressemblances entre variables d'organiser A en classes et sous-classes et de dégager ainsi les "dimensions" sous-jacentes au comportement de la population étudiée qui constitue E. Le point de départ de cette recherche est la définition d'un indice de proximité sur A attachant à chaque couple de variables un nombre sensé mesurer leur ressemblance. Relativement à un couple de variables définissant un couple de partitions sur E (type c)), la statistique du χ^2, attachée au tableau de contingence de croisement des deux partitions, sert en général à éprouver l'hypothèse H_o d'indépendance entre les deux variables où la distribution du χ^2 a une tendance asymptotique connue. Notre optique sera différente de celle adoptée dans la théorie des tests ; si χ_0^2 est la valeur du χ^2 associée au couple de partitions, χ_0^2 ou mieux "probabilité calculée dans l'hypothèse H_0 d'avoir $\chi^2 < \chi_0^2$" sera considérée comme une mesure du degré de dépendance entre les deux variables. H_0 a ainsi joué le rôle d'une hypothèse de référence pour l'établissement de la statistique de dépendance. On peut remarquer que cette dernière est aussi une statistique de proximité entre les deux partitions, parce que les diverses classes d'une même partition jouent un rôle symétrique les unes par rapport aux autres. Cependant l'approximation de la distribution de cette statistique dans l'hypothèse H_0 est sensible aux cases faiblement chargées du tableau de contingence de croisement des deux partitions ; d'autre part, la forme asymptotique de la distribution dépend de $(k-1)(h-1)$ degrès de liberté où k est le nombre de classes de l'une des deux partitions et h de l'autre, de sorte que dans la comparaison de toutes les paires de variables de A on aura à se référer à un grand nombre de distributions du χ^2 (dans les expériences courantes le cardinal de l'ensemble des paires de A est de l'ordre de 10^4) ; enfin, on ne voit pas clairement comment l'hypothèse H_0 tient compte de ce que les objets quelle manipule sont des partitions. Dans la comparaison d'un couple de variables de type d) définissant un couple (o, o')

d'ordres totaux sur E, M.G. Kendall considère une hypothèse N d'absence de liaison entre les deux variables qui tient précisément compte du type de structure qu'induisent les variables sur E ; en effet il se place dans $E \times E$ et étudie la distribution de la statistique

$$\sum_{(x,y)\in E\times E} a(x\,,\,y)\,b(x\,,\,y)$$

où $\quad a(x\,,\,y)$ (resp. $b(x\,,\,y)$) $= \begin{cases} 1 & \text{si } x < y \quad \text{pour } o \text{ (resp. } o'\text{)} \\ -1 & \text{si } x > y \quad \text{" } \quad o \text{ (resp. } o'\text{)} \end{cases}$

lorsque o' parcourt l'ensemble, muni d'une mesure de probabilité uniforme, de tous les ordres totaux qu'on peut définir sur E, (cf. [3]). C'est cette forme que nous donnerons à l'hypothèse N d'absence de liaison pour définir de façon "statistiquement pertinente" la proximité entre deux variables établissant le même type de structure sur E. Nous retrouvons de la sorte des statistiques connues dans les cas a), d) et e) ; il s'agit du coefficient d'association de K. Pearson pour a), du τ de M.G. Kendall dans le cas d) et d'une statistique dont la distribution, dans l'hypothèse N, est l'objet d'un théorème important dû à A. Wald et J. Wolfowitz (1944) (cf. [10]). Mais, nous trouvons des statistiques nouvelles dans les cas b) et c) où le tableau de contingence est le support de l'information ; pour b) une marge du tableau indique une partition aux classes étiquetées et pour c), un préordre total. Dans le cas b) nous nous plaçons dans l'ensemble F des paires d'objets distincts de E (i.e. des parties à deux éléments de E) et dans le cas c) dans l'ensemble $E \times E$, pour établir la statistique de proximité. L'hypothèse N, que nous exprimerons plus tard de façon plus précise, fixe dans le cas b) l'une des deux partitions et fait varier l'autre dans l'ensemble de toutes les partitions pour lesquelles la suite des cardinaux de la marge associée du tableau de contingence reste la même ; un théorème de dualité (cf. § 2.0) permet d'établir que *cette distribution ne dépend pas de celle de deux partitions fixée.* Le même théorème de dualité permet d'établir un résultat analogue pour la comparaison de deux préordres totaux où nous montrerons notamment que la statistique que propose M.G. Kendall pour étendre son τ est biaisée.

Quel que soit le type algébrique du couple de variables envisagé nous serons amenés à nous référer à la loi normale centrée et réduite pour la distribution dans l'hypothèse N de la statistique de proximité entre les deux variables.

La notion de proximité entre deux variables sera étendue à celle entre deux classes de variables de même type ; cette extension se fera par deux voies, les deux sont basées sur la distribution dans l'hypothèse N d'une

statistique de proximité entre les deux classes qui tient compte de leurs cardinaux ; la première statistique est la plus grande proximité entre deux éléments appartenant respectivement aux deux classes ; et la seconde, la somme des proximités attachées à l'ensemble des couples dont les deux composantes appartiennent respectivement aux deux classes.

L'algorithme classique qui à chaque pas réunit les deux classes les plus voisines produit un arbre détaillé de classifications de moins en moins fines "respectant" les ressemblances de l'ensemble D à classifier qui peut aussi bien être l'ensemble A des variables descriptives que l'ensemble E des objets. Pour étudier la cohérence des classes de la partition formée à un niveau donné, on introduit après R.N. Shepard (1962) et J.P. Benzecri (1964-65) (cf. [8] et [1]) l'"ordonnance" ω sur D qui est l'ordre total sur l'ensemble F des paires de D pour lequel, une paire p précède une paire p' si les deux composantes de p sont moins proches que ceux de p' au sens de la similarité établie sur D. La donnée d'une partition π sur D étant équivalente à celle d'un préordre total sur F en deux classes $R(\pi)$ (resp. $S(\pi)$) où $R(\pi)$) (resp. $S(\pi)$) est l'ensemble des paires réunies (resp. séparées) par la partition ; on se trouve ramenés à la comparaison de deux structures de même type sur D : préordres totaux sur l'ensemble F des parties à deux éléments de D. La base de la mesure de proximité sera card $(gr(\omega) \cap S(\pi) \times R(\pi))$ où $gr(\omega)$ est le graphe de l'ordre total dans $F \times F$. L'hypothèse N peut, soit faire varier l'ordre ω dans l'ensemble de tous les ordres totaux sur F en laissant fixée la partition π ; soit faire varier la partition π dans l'ensemble de toutes les partitions de même type sur D (i.e. dont la suite des cardinaux des classes est fixée), en laissant l'ordre ω sur F fixé. Nous obtenons asymptotiquement, dans des conditions assez générales, la même distribution qui est normale de moyenne $r.s/2$ et de variance $r.s(f + 1)/12$ où r = card $(R(\pi))$, s = card $(S(\pi))$ et $r + s = f$, où f = card (F).

La condensation de l'arbre des classifications à ses niveaux ou nœuds les plus significatifs est encore basée sur la distribution dans l'hypothèse N d'une statistique de même forme que card $(gr(\omega) \cap S \times R)$ mais conçue à partir de l'ensemble des paires laissées séparées à un niveau donné de l'arbre par rapport à celles qu'on vient de réunir à ce niveau.

II – INDICE DE PROXIMITÉ ENTRE VARIABLES DE MÊME TYPE

La partie la plus importante traitée ici concerne la comparaison d'un couple de partitions ou de préordres totaux définis respectivement par un couple de variables de type b) ou c), (cf. § I). L'étude comparative d'un couple de préordres totaux étant tout à fait parallèle à celle d'un couple de partitions ; nous grouperons ces deux études au paragraphe 2. Le paragraphe suivant, où on étudie la proximité entre deux attributs descriptifs définissant un couple de parties de E, peut être considéré comme d'introduction.

1. COUPLE DE VARIABLES INDICATEUR D'UN COUPLE DE PARTIES DE E.

Soit (E_a, E_b) le couple de parties indiqué par le couple (a, b) d'attributs descriptifs. Pour étudier la position relative de l'une des parties par rapport à l'autre, on introduit les cardinaux suivants : $s = \text{card}\,(E_a \cap E_b)$, $u = \text{card}\,(E_a \cap E_b^c)$ où $E_b^c = (E - E_b)$, $v = \text{card}\,(E_a^c \cap E_b)$ et $t = \text{card}\,(E_a^c \cap E_b^c)$. Les deux attributs de description étant supposés tels que leur présence simultanée chez un même objet de E peut être significative de leur ressemblance alors que leur absence commune "n'indique rien" ; la base de l'indice de proximité qui s'impose est s. Mais une telle statistique est manifestement biaisée ; en effet deux attributs fréquents (resp. rares) entraînent une valeur grande (resp. petite) de s, indépendamment de la position relative de E_a et de E_b. Pour ne retenir dans la statistique s que ce qui "peut être significatif", introduisons l'hypothèse N d'absence de liaison où à $s = \text{card}\,(E_a \cap E_b)$, on associe chacune des deux variables aléatoires

$$S_a = \text{card}\,(E_a \cap Y) \qquad \text{et} \qquad S_b = \text{card}\,(X \cap E_b)$$

où X (resp. Y) varie dans l'ensemble des parties de E à a_a (resp. n_b) éléments, muni d'une mesure de probabilité uniformément répartie.

Proposition

Les distributions de S_a et de S_b sont identiques

En effet ; $$P_r^N\{S_a = k\} = \frac{\binom{n_a}{k}\binom{n - n_a}{n_b - k}}{\binom{n}{n_b}} = \frac{\binom{n_b}{k}\binom{n - n_b}{n_a - k}}{\binom{n}{n_a}} = P_r^N\{S_b = k\}.$$

où P_r^N désigne la probabilité dans l'hypothèse N et où $\binom{m}{l}$ indique un coefficient binomial.

La loi de probabilité commune est de type hypergéométrique de moyenne $\mu = n_a n_b/n$ et de variance $\sigma^2 = n_a(n - n_a)n_b(n - n_b)/n^3$. La "bonne" mesure de proximité entre a et b, qui neutralise les effets statistiques dûs à la grandeur relative de n_a et de n_b, est obtenue en centrant et en réduisant s par référence à la loi commune de S_a et de S_b ; soit

$$Q(a, b) = \frac{s - \mu}{\sigma} \tag{1}$$

cette expression peut se mettre sous la forme

$$Q(a, b) = \sqrt{n}\ (st - uv)/\sqrt{(s + u)\ (s + v)\ (t + u)\ (t + v)}$$

qui n'est autre que le coefficient de K. Pearson dont le carré est le χ^2 attaché au tableau de contingence $\begin{array}{c|c} s & v \\ \hline u & t \end{array}$

Les rapport n_a/n et n_b/n étant fixés, la distribution commune de S_a et de S_b tend vers celle de la loi normale, pour $n \to \infty$. Pour n fixé, l'approximation normale est d'autant meilleure que $\min\ (n_a/n, n_b/n, 1 - (n_a/n), 1 - (n_b/n))$ est grand. De sorte que si q est la valeur observée de (1), on peut se référer à l'échelle de mesure définie par la loi normale et adopter comme valeur de la proximité entre a et b.

$$P(a, b) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{q} e^{-t2/2}\ dt \tag{2}$$

Nous allons à présent chercher à retrouver ces résultats connus d'ailleurs à propos d'autres types de variables.

2. COMPARAISON D'UN COUPLE DE PARTITIONS OU DE PREORDRES TOTAUX

2.0. Théorème de dualité

Nous désignons par $\mathcal{R}(m\,;m_1,\ldots,m_g)$ l'ensemble des partitions que l'on peut obtenir en associant à chaque partie e de E de cardinal m, l'ensemble de ses partitions $(e_1, e_2, \ldots, e_g)$ en classes étiquetées ; la i-ème e_i ayant pour cardinal m_i ; $m = m_1 + \ldots + m_g$. La donnée d'un élément de $\mathcal{R}(m\,;m_1,\ldots,m_g)$ correspond à celle d'une partie e de E de cardinal m et d'une application surjective φ de e dans $\{1, 2, \ldots, g\}$ pour laquelle $\varphi^{-1}(i) = e_i$ est de cardinal m_i, $i = 1, 2, \ldots, g$.

Le cardinal de $\mathcal{R}(m\,;m_1,\ldots,m_g)$ est

$$\binom{n}{m} \times \frac{m!}{m_1!\, m_2! \ldots m_g!} = \frac{n(n-1)\ldots(n-m+1)}{m_1!\, m_2! \ldots m_g!}$$

En effet $\binom{n}{m}$ est le nombre de parties de E à m éléments et $m!/m_1!\, m_2! \ldots m_g!$ est le nombre de partitions d'un ensemble fixé à m éléments, en classes étiquetées de cardinaux respectifs $m_1, m_2, \ldots, m_g$.

$\mathcal{R}(n\,;n_1,n_2,\ldots,n_k)$ indiquera l'ensemble des partitions $(E_1, E_2, \ldots, E_k)$ de E en classes étiquetées de cardinaux respectifs $n_1, n_2, \ldots, n_k$. Le cardinal de $\mathcal{R}(n\,;n_1,\ldots,n_k)$ est $n!/n_1!\, n_2! \ldots n_k!$.

Posons $J = \{1, 2, \ldots, g\}$ et $I = \{1, 2, \ldots, k\}$ où on suppose $g < k$; τ définira une application injective de J dans I pour laquelle $m_j \leqslant n_{\tau(j)}$ pour tout j de J.

Posons encore $u = (m_1, m_2, \ldots, m_g)$ et $t = (n_1, n_2, \ldots, n_k)$; et soient $\pi_0(u)$ et $\pi_0(t)$ deux partitions fixées appartenant respectivement à $\mathcal{R}(m\,;u)$ et $\mathcal{R}(n\,;t)$, on notera $\pi_0(u) = (e_1^0, e_2^0, \ldots, e_g^0)$ et $\pi_0(t) = (E_1^0, E_2^0, \ldots, E_k^0)$

Théorème

La proportion de partitions $\pi(u) = (e_1, e_2, \ldots, e_g)$ de $\mathcal{R}(m\,;u)$ pour lesquelles $e_j \subset E^0_{\tau(j)}$ pour tout j, est égale à la proportion de partitions $\pi(t) = (E_1, E_2, \ldots, E_k)$ de $\mathcal{R}(n\,;t)$ pour lesquelles $e_j^0 \subset E_{\tau(j)}$.

La première des deux proportions vaut

$$\frac{\prod_{j\in J} \binom{n_{\tau(j)}}{m_j}}{\dfrac{n(n-1)\ldots(n-m+1)}{m_1!\, m_2! \ldots m_g!}} = \frac{\prod_{j\in J} n_{\tau(j)}\,(n_{\tau(j)}-1)\ldots(n_{\tau(j)}-m_j+1)}{n(n-1)\ldots(n-m+1)} \quad (1)$$

En effet $\binom{n_{\tau(j)}}{m_j}$ est le nombre de parties de $E^0_{\tau(j)}$ à m_j éléments. Donc $\prod_{j\in J}\binom{n_{\tau(j)}}{m_j}$ est le nombre de g-uples de parties $(e_1, \ldots, e_g)$ tels que $\operatorname{card}(e_j) = m_j$ et $e_j \subset E^0_{\tau(j)}$ pour tout j de J.

D'autre part le nombre de partitions de $\mathcal{P}(n\,;t)$ pour les quelles $e^0_j \subset E_{\tau(j)}$ pour tout j de J est égal à

$$\frac{(n-m)!}{\prod_{j\in J}(n_{\tau(j)}-m_j)!\prod_{i\in(\tau(J))^c} n_i!}$$

où $(\tau(J))^c$ est le complémentaire dans I de l'image de J par τ. La proportion désignée dans le second membre du théorème est donc égal à

$$\frac{(n-m)!}{\prod_{j\in J}(n_{\tau(j)}-m_j)!}\Big/\frac{n!}{\prod_{j\in J} n_{\tau(j)}!}\,,$$

laquelle est égale à la proportion (1) ci-dessous calculée.

2.1. Distributions duales attachées à couple de partitions

Soit F l'ensemble des paires d'éléments distints de E :

$$F = \{\{x\,,\,y\}/\,x\in E\,,\,y\in E\,,\,x\neq y\}\,.$$

La donnée d'une partition π sur E est équivalente à la donnée d'une partie $R(\pi)$ de F telle que

$$\{x\,,\,y\}\in R(\pi) \quad \text{et} \quad \{y\,,\,z\}\in R(\pi) \Rightarrow \{x\,,\,z\}\in R(\pi)\,,$$

pour tout $x\,,\,y$ et z de E.

$R(\pi)$ est l'ensemble des paires dont les deux composantes sont dans une même classe de π ; la partie complémentaire $S(\pi)$ de $R(\pi)$ est formée des paires dont les deux composantes sont dans deux classes distinctes.

$$\operatorname{card}(R(\pi)) + \operatorname{card}(S(\pi)) = \operatorname{card}(F) = n(n-1)/2$$

Si $t = (n_1\ n_2, \ldots, n_k)$ est le type de la partition, les cardinaux de $R(\pi)$ et de $S(\pi)$ sont donnés par

$$\operatorname{card}(R(\pi)) = \sum_{1\leqslant i\leqslant k} n_i(n_i-1)/2\,,\ \operatorname{card}(S(\pi)) = \sum_{i<j} n_i\,n_j$$

Si $R(\pi_0)$ (resp. $R(\pi'_0)$) est l'ensemble des paires réunies par la partition $\pi_0(t)$ de $\mathcal{P}(n\,;t)$ (resp. $\pi'_0(t')$ de $\mathcal{P}(n\,;t')$), $\underline{\text{card}(R(\pi_0)\cap R(\pi'_0))}$ constituera la base de l'indice de proximité à établir entre les deux partitions ; il s'agit en effet du nombre de paires d'objets de E qui sont réunies par chacune des deux partitions π_0 et π'_0. On a

$$\text{card}(R(\pi_0)\cap R(\pi'_0)) = \sum_{(i,j)} n_{ij}(n_{ij}-1)/2$$

où n_{ij} est le cardinal de la classe $(i\,,j)$ de la partition, croisement des deux partitions π_0 et π'_0 .

A $\text{card}(R(\pi_0)\cap R(\pi'_0)$ associons deux variables aléatoires duales ; $\text{card}(R(\pi)\cap R(\pi'_0))$ et $\text{card}(R(\pi_0)\cap R(\pi'))$ où π (resp. π') est un élément aléatoire de $\mathcal{P}(n\,;t)$ (resp. $\mathcal{P}(n\,;t')$), muni d'une mesure de probabilité uniforme. On peut préciser la forme de t et de t' ; $t=(n_1,n_2,\ldots,n_k)$ et $t'=(n'_1,n'_2,\ldots,n'_h)$.

Le but principal de ce paragraphe est d'établir à l'aide du théorème précédent que la distribution de $\text{card}(R(\pi)\cap R(\pi'_0))$ *est la même* que celle de $\text{card}(R(\pi_0)\cap R(\pi'))$; on précisera d'autre part la *moyenne* et la *variance* de cette distribution commune. La statistique $\text{card}(R(\pi_0)\cap R(\pi'_0))$ centrée réduite, définira une "bonne" mesure de proximité entre les deux partitions.

Nous allons à présent travailler dans F^q où q est un exposant entier fixé. Si $(p_1,p_2,\ldots,p_q)$ est un q-uple de paires et a une auto-bijection de E, en posant $a(\{x\,,y\})=\{a(x)\,,a(y)\}$ et $a(p_1,\ldots,p_q)=(a(p_1),\ldots,a(p_q))$, a transforme $(p_1,p_2,\ldots,p_q)$ en un élément de F^q de même configuration c que $(p_1,p_2,\ldots,p_q)$; c'est-à-dire, tel que l'objet $a(x)$ remplace x dans les paires (composantes du q-uple) où était présent ce dernier. L'ensemble $G_q^{(c)}$ de tous les q-uples ayant la même configuration c, est l'ensemble des valeurs $\{(a(p_1),a(p_2),\ldots,a(p_q))\}$ où a décrit les n ! auto-bijections de E. On peut noter G_q^0 l'ensemble des q-uples de paires dont deux quelconques sont sans composante commune. Nous n'aurons pas besoin dans la suite de préciser, sauf pour $q\leqslant 2$, les différentes configurations c et les cardinaux des ensembles $G_q^{(c)}$.

Soient π_0 une partition fixée dans $\mathcal{P}(n\,;t)$ et $(p_1^0,p_2^0,\ldots,p_q^0)$ un q-uple fixé dans $G_q^{(c)}$; désignons par φ_0 (resp. φ) la fonction indicatrice de la partie $R(\pi_0)$ (resp. $R(\pi)$) de F que définit la relation d'équivalence associée à la partition π_0 (resp. associée à une partition π, élément courant de $\mathcal{P}(n\,;t)$)

2.1.1. Théorème

La proportion de q-uples $(p_1,p_2,\ldots,p_q)$ dans $G_q^{(c)}$ pour lesquels $\varphi_0(p_1)\varphi_0(p_2)\ldots\varphi_0(p_q)=1$, est égale à la proportion de partitions de $\mathcal{P}(n\,;t)$ pour lesquelles $\varphi(p_1^0)\varphi(p_2^0)\ldots\varphi(p_q^0)=1$.

Remplaçons le langage des proportions par celui, équivalent mais plus souple des probabilités où $G_q^{(c)}$ et $\mathcal{R}(n\,;t)$ sont respectivement munis d'une mesure de probabilité uniformément répartie, noté μ sur G_q^c et ν sur $\mathcal{R}(n\,;t)$. Dans ces conditions nous avons à établir que

$$\mu\{\varphi_0(p_1)\varphi_0(p_2)\ldots\varphi_0(p_q)=1\}=\nu\{\varphi(p_1^0)\varphi(p_2^0)\ldots\varphi(p_q^0)=1\}$$

Pour cela nous décomposerons chacun des évènements Q et Q^* respectivement sous les signes de μ et de ν en sous évènements disjoints qui se correspondent mutuellement

$$Q=\sum_{\alpha\in A}Q_\alpha\,,\quad Q^*=\sum_{\alpha\in A}Q_\alpha^*$$

où A est fini et nous montrerons que $\mu(Q_\alpha)=\nu(Q_\alpha^*)$.

Soit $\{p_1,p_2,\ldots,p_q\}$ l'ensemble des paires associé à un élément $(p_1,p_2,\ldots p_q)$ donné de $G_q^{(c)}$. La saturation de cet ensemble de paires (i.e. l'adjonction à cet ensemble de toute paire $\{x\,,\,y\}$ dont chacune des deux composante est présente dans au moins une paire p_i, $1\leqslant i\leqslant q$) définit une partition $(e_1,e_2,\ldots,e_g)$ sur un sous-ensemble e de E de cardinal m. Le type $u=(m_1,m_2,\ldots,m_g)$ est indépendant de l'élément $(p_1,p_2,\ldots,p_q)$ choisi dans $G_q^{(c)}$; d'autre part l'évènement Q, dont la probabilité μ ne dépend que de u, peut être décomposé en un sous évènements Q_α dont chacun est de la forme suivante ; $e_j'\subset E^0_{\tau(j)}$ pour tout j de $\{1\,,2\,,\ldots,h\}$, (1) ; où $(e_1',e_2',\ldots,e_h')$ est une partition de e moins fine que $(e_1,e_2,\ldots,e_g)$ (i.e. toute classe e_i' est une réunion de r_i classes e_j, $r_i\geqslant 1$) ; τ étant une injection de $\{1\,,2\,,\ldots,h\}$ dans $\{1\,,2\,,\ldots,k\}$. On notera $u'=(m_1',m_2',\ldots,m_h')$ le type de $(e_1',e_2',\ldots,e_h')$.

A l'évènement Q_α précèdent associons l'évènement Q_α^* :

$$e_j'^0\subset E_{\tau(j)}\qquad\text{pour tout } j \text{ de}\qquad\{1\,,2\,,\ldots,h\}\,,\tag{1}$$

où $(e_1'^0,e_2'^0,\ldots,e_h'^0)$ se déduit de $(e_1^0,e_2^0,\ldots,e_h^0)$ de la même façon que $(e_1',e_2',\ldots,e_h')$ se déduisait de $(e_1,e_2,\ldots,e_g)$; τ étant la même injection que ci-dessus.

La décomposition de Q^* en $\{Q_\alpha^*/\alpha\in A\}$ est duale de la décomposition de Q en $\{Q_\alpha/\alpha\in A\}$. Il nous reste à montrer que

$$\mu(Q_\alpha)=\nu(Q_\alpha^*)$$

$\mu(Q_\alpha)$ est la proportion de q-uples de $G_q^{(c)}$ dont la partition $(e_1',e_2',\ldots,e_h')$ associée remplit la condition (1) : $e_j'\subset E^0_{\tau(j)}$, pour tout j. Cette proportion est égale à la proportion de partitions de $\mathcal{R}(m\,;u')$ pour lesquelles on a la

condition (1) ; en effet, le nombre de q-uples de $G_q^{(c)}$ qui déterminent une meme partition $(e'_1, e'_2, \ldots, e'_h)$ est indépendant de la partition choisie dans $\mathcal{R}(m\,;u')$.

Parallèlement $\nu(Q_\alpha^*)$ est la proportion de partitions de $\mathcal{R}(n\,;t)$ pour lesquelles on a la condition (1^*), $e_j'^0 \subset E_{\tau(j)}$, pour tout j. Le théorème précédent (§ 2.0) établit l'égalité des deux proportions.

2.1.2. Expression du moment d'ordre r de la distribution de card$(R(\pi) \cap R(\pi'_0))$ ou de celle de card$(R(\pi_0) \cap R(\pi'))$

Rappelons que φ désigne la fonction indicatrice de la partie $R(\pi)$ de F associée à une partition variable π dans $\mathcal{R}(n\,;t)$ et soit ψ_0 la fonction indicatrice de la partie $R(\pi'_0)$ de F associée à une partition fixée de $\mathcal{R}(n\,;t')$.

$$\text{card}\,(R(\pi) \cap R(\pi'_0)) = \sum_{p \in F} \cdot \varphi(p)\psi_0(p) \tag{1}$$

$\mathcal{R}(n\,;t)$ étant muni d'une mesure uniforme de probabilité, le moment d'ordre r de la statistique (1) ci-dessus se met sous la forme

$$\mathcal{E}\{(\Sigma\, c(r\,;r_1, r_2, \ldots, r_q)\,(\varphi(p_{i_1})^{r_1} \ldots \varphi(p_{i_q})^{r_q}\,(\psi_0(p_{i_1})^{r_1} \ldots \psi_0(p_{i_q})^{r_q}\}\} \tag{2}$$

où le signe $\mathcal{E}$ désigne l'espérance mathématique ; la sommation pour $(r_1, r_2, \ldots, r_q)$ fixé, est étendue à toutes les permutations $(p_{i_1}, p_{i_2}, \ldots, p_{i_q})$ pouvant être obtenues à partir de chacune des parties à q éléments de F.

$$c(r\,;r_1, r_2, \ldots, r_q) = \frac{r!}{l_1!\, l_2! \ldots l_h!\, r_1!\, r_2! \ldots r_q!}$$

où h est le nombre d'entiers r_j distincts ; chacun d'entre eux se répétant respectivement $l_1, l_2, \ldots, l_h$ fois.

On a évidemment

$$\varphi(p_{i_1})^{r_1}\varphi(p_{i_2})^{r_2} \ldots \varphi(p_{i_q})^{r_q} = \varphi(p_{i_1})\varphi(p_{i_2}) \ldots \varphi(p_{i_q})$$

$$\psi_0(p_{i_1})^{r_1}\psi_0(p_{i_2})^{r_2} \ldots \psi_0(p_{i_q})^{r_q} = \psi_0(p_{i_1})\psi_0(p_{i_2}) \ldots \psi_0(p_{i_q})$$

Dans ces conditions l'expression (2) se met sous la forme

$$\Sigma\, c(r\,;r_1, r_2, \ldots, r_q)\,\mathcal{E}\{(\varphi(p_{i_1}) \ldots \varphi(p_{i_q}))\,(\psi_0(p_{i_1}) \ldots \psi_0(p_{i_q}))\} \tag{3}$$

De façon analogue, le moment d'ordre r de la statistique card$(R(\pi_0) \cap R(\pi'))$ se met sous la forme

$$\Sigma\, c(r\,;r_1,\ldots,r_q)\,(\varphi_0(p_{i_1})\ldots\varphi_0(p_{i_q}))\;\mathcal{E}(\psi(p_{i_1})\ldots\psi(p_{i_q})) \qquad (3')$$

où φ_0 (resp. ψ) est la fonction indicatrice de la partie $R(\pi_0)$ (resp. $R(\pi')$) de F, associée à une partition fixée de $\mathcal{R}(n\,;t)$ (resp. à une partition variable de $\mathcal{R}(n\,;t')$)

Pour une décomposition $(r_1, r_2, \ldots, r_q)$ fixée de r, considérons les parties correspondantes des sommes (3) et (3') ; soit

$$\Sigma\;\mathcal{E}(\varphi(p_{i_1})\ldots\varphi(p_{i_q})\,\psi_0(p_{i_1})\ldots\psi_0(p_{i_q}) \qquad (4)$$

$$\Sigma\,(\varphi_0(p_{i_1})\ldots\varphi_0(p_{i_q}))\quad(\psi(p_{i_1})\ldots\psi(p_{i_q}) \qquad (4')$$

Décomposons simultanément les sommes (4) et (4') selon les différents ensembles $G_q^{(c)}$ considérés ci-dessus. En vertu du théorème précédent (§ 2.1.1.), on a

$$\sum_{G_q^{(c)}}\mathcal{E}(\varphi(p_{i_1})\ldots\varphi(p_{i_q}))\,\psi_0(p_{i_1})\ldots\psi_0(p_{i_q}) =$$

$$\sum_{G_q^{(c)}}\varphi_0(p_{i_1})\ldots\varphi_0(p_{i_q})\;\mathcal{E}(\psi(p_{i_1})\ldots\psi(p_{i_q}))$$

Par conséquent, le moment d'ordre r de la distribution de card$(R(\pi)\cap R(\pi_0))$ est le même que celui de la distribution de card$(R(\pi_0)\cap R(\pi')$; d'où le théorème

Théorème.

$\mathcal{R}(n\,;t)$ et $\mathcal{R}(n\,;t')$ étant munis d'une mesure de probabilité uniforme, la distribution de card$(R(\pi)\cap R(\pi_0'))$ est la même que celle de card$(R(\pi_0)\cap R(\pi'))$.

2.1.3. Tendance centrale et Dispersion de la distribution de card $(R(\pi)\cap R(\pi_0'))$.

Nous commencerons par établir deux lemmes (1 et 2) qui précisent le théorème 2.1.1. pour $q = 1$ et $q = 2$ où nous expliciterons les différentes formes de $G_q^{(c)}$ et où nous déterminerons la valeur commune de la proportion que suppose l'énoncé du théorème. Nous pourrons alors calculer la *moyenne* et la *variance* de la distribution de card$(R(\pi)\cap R(\pi_0'))$.

2.1.3.1. Lemme 1.

Soient $p_0 = \{x_0, y_0\}$ une paire fixée dans F et π_0 une partition fixée dans $\mathfrak{P}(n\,;t)$. La proportion de paires de F dont les deux composantes sont réunies par la partition π_0, est égale à la proportion de partitions de $\mathfrak{P}(n\,;t)$ qui réunissent les deux composantes de p_0 ; la valeur commune de cette proportion est $\sum_i n_i(n_i - 1)/n(n-1)$; (on rappelle que $t = (n_1, n_2, \ldots, n_i, \ldots, n_k)$).

La valeur de la première des deux proportions indiquées apparaît clairement ; en effet, le nombre de paires réunies par π_0 est $\sum_i n_i(n_i - 1)/2$ et le cardinal de F est $n(n-1)/2$. La deuxième proportion peut se mettre sous la forme

$$\sum_i \left\{ \frac{(n-2)!}{n_1! \ldots n_{(i-1)}!\,(n_i - 2)!\,n_{(i+1)}! \ldots n_k!} \Big/ \frac{n!}{n_1!\,n_2! \ldots n_k!} \right\} =$$

$$\sum_i n_i(n_i - 1)/n(n-1)$$

où un terme de la somme représente la proportion de partitions dans $\mathfrak{P}(n\,;t)$ pour lesquelles x_0 et y_0 sont réunis dans la i-ème classe.

2.1.3.2. Lemme 2

Désignons ici par G (resp. H) l'ensemble des couples de paires (p, p') de la forme

$$(\{x, y\}, \{x, z\}) \quad (\text{resp. } (\{x, y\}, \{z, v\}))\ ;$$

c'est-à-dire, avec (resp. sans) composante commune ; on a

$$\text{card}(G) = n(n-1)(n-2) \quad \text{et} \quad \text{card}(H) = n(n-1)(n-2)(n-3)/4.$$

Soit π_0 une partition fixée dans $\mathfrak{P}(n\,;t)$.

(g) Si (p_0, p_0') est un couple de paires fixé dans G ; la proportion d'éléments (p, p') de G tels que p et p' soient formées de composantes réunies par π_0, est égale à la proportion de partitions de $\mathfrak{P}(n\,;t)$ pour lesquelles les deux composantes de p_0 et de p_0' sont réunies. La valeur commune de cette proportion est

$$\sum_{1 \leqslant i \leqslant k} n_i(n_i - 1)(n_i - 2)/n(n-1)(n-2) \qquad (1)$$

(h) Si (p_0, p'_0) est un couple de paires fixé dans H ; la proportion d'éléments (p, p') de H tels que p et p' soient formées de composantes réunies par π_0, est égale à la proportion de partitions de $\mathcal{R}(n ; t)$ pour lesquelles les deux composantes de p_0 et de p'_0 sont réunies. La valeur commune de cette proportion est

$$\Big\{ \sum_i n_i(n_i - 1)(n_i - 2)(n_i - 3) + \sum_i n_i(n_i - 1) \sum_{j \neq i} n_j(n_j - 1) \Big\} \Big/ n(n-1)(n-2)(n-3) \tag{2}$$

Démonstration de la partie (g)

Déterminons le nombre de couples de paires de la forme $(\{x, y\}, \{x, z\})$ pour lesquels x, y et z sont dans la i-ème classe de cardinal n_i de la partition π_0. A chaque partie $\{a, b, c\}$ à trois éléments de E correspond 6 couples différents de paires de G ; soit $(\{a, b\}, \{a, c\})$, $(\{a, b\}, \{b, c\})$; $(\{b, c\}, \{a, b\})$, $(\{b, c\}, \{a, c\})$; $(\{a, c\}, \{a, b\})$., $(\{a, c\}, \{b, c\})$. Le nombre de parties à 3 éléments de la i-ème classe étant $\binom{n_i}{3} = n_i(n_i - 1)(n_i - 2)/6$, $n_i(n_i - 1)(n_i - 2)$ est le nombre cherché. Il en résulte la valeur annoncée de la proportion définie dans la première partie de l'énoncée (g).

Posons $(p_0, p'_0) = (\{x_0, y_0\}, \{x_0, z_0\})$. La proportion de partitions de $\mathcal{R}(n ; t)$, pour lesquelles x_0, y_0 et z_0 se trouvent réunis dans la i-ème classe de cardinal

$$n_i \geqslant 3, \qquad \text{est} \qquad n_i(n_i - 1)(n_i - 2)/n(n-1)(n-2)$$

qui s'obtient de façon énumérative. D'où la valeur (1) ci-dessus de la proportion exprimée dans la seconde partie de l'énoncé (g) où on suppose $n_i \geqslant 3$ pour tout i.

Démonstration de la partie (h)

Le nombre de couples de paires de la forme $(\{x, y\}, \{z, v\})$, pour lesquels $\{x, y\}$ est incluse dans la i-ème classe de cardinal n_i et $\{z, v\}$, dans la j-ème classe de cardinal n_j de la partition π_0, est égal à

$$\binom{n_i}{2} \times \binom{n_j}{2} \qquad \text{pour} \qquad i \neq j$$

et

$$\binom{n_i}{2} \times \binom{n_i - 2}{2} \qquad \text{pour} \qquad j = i$$

où on supposera $n_i \geqslant 4$ pour tout i.

Il en résulte la formule (2) ci-dessus, concernant la première partie de l'énoncé (h), par la considération notamment du croisement de chaque paire $\{x, y\}$ d'une même classe avec l'ensemble des parties $\{z, v\}$ à deux éléments,disjointes de $\{x, y\}$ et pour lesquelles z et v sont dans une même classe.

Posons $(p_0, p_0') = (\{x_0, y_0\}, \{z_0, v_0\})$. La proportion de partitions dans $\mathfrak{P}(n; t)$ pour lesquelles la paire $\{x_0, y_0\}$ est contenue dans la i-ème classe de cardinal n_i et la paire $\{z_0, v_0\}$ dans la j-ème classe de cardinal n_j, est égale à

$$n_i(n_i - 1)\, n_j(n_j - 1)/n(n-1)\,(n-2)\,(n-3) \qquad \text{si} \quad j \neq i$$

et

$$n_i(n_i - 1)\,(n_i - 2)\,(n_i - 3)/n\,(n-1)\,(n-2)\,(n-3) \qquad \text{si} \quad j = i$$

D'où la valeur (2) ci-dessus de la proportion définie dans la seconde partie de l'énoncé (h) où on suppose $n_i \geqslant 4$ pour tout i.

Les deux distributions duales dont il est question dans l'énoncé suivant sont celles de $\operatorname{card}(R(\pi) \cap R(\pi_0'))$ et de $\operatorname{card}(R(\pi_0) \cap R(\pi'))$ envisagées au paragraphe 2.2. précédent.

Théorème

La *moyenne* commune et la *variance* commune, des deux identiques distributions duales, sont respectivement

$$\lambda\mu \qquad \text{et} \qquad \lambda\mu + \rho\sigma + (\theta\zeta - \lambda^2\mu^2) \simeq \lambda\mu + \rho\,\sigma$$

où

$$\lambda = \sum_{i=1} n_i(n_i - 1)/\sqrt{2n(n-1)}\,,$$

$$\rho = \sum_i n_i(n_i - 1)\,(n_i - 2)/\sqrt{n(n-1)\,(n-2)}$$

$$\theta = \frac{\left\{ \sum_i n_i^2(n_i - 1)^2 - 4 \sum_i n_i(n_i - 1)\,(n_i - 2) - 2 \sum_i n_i(n_i - 1) \right\}}{2\sqrt{n(n-1)\,(n-2)\,(n-3)}}$$

Les expressions de μ, σ et ζ ont respectivement la même forme λ, ρ et θ ; les n_i de $t = (n_1, n_2, \ldots, n_k)$ étant remplacés par les n_i' de $t' = (n_1', n_2', \ldots, n_h')$

En reprenant l'expression (1) du paragraphe 2.2, on a

$$\mathcal{E}(\text{card}(R(\pi) \cap R(\pi'_0))) = \sum_{p \in F} \psi_0(p)\mathcal{E}(\varphi(p)) = \mathcal{E}(\varphi(p)) \sum_{p \in F} \psi_0(p) \quad (1)$$

le Lemme 1 permet d'établir le premier résultat annoncé concernant la valeur de la moyenne.

Considérons pour l'expression suivante de la variance

$$\mathcal{E}\left\{\left(\sum_{p \in F} \varphi(p)\psi_0(p)\right)^2\right\} - \lambda^2\mu^2 \ , \quad (2)$$

le développement $\left(\sum_{p \in F} \varphi(p)\psi_0(p)\right)^2$; soit

$$\sum_{p \in F} \varphi(p)\psi_0(p) + \sum_{(p,p') \in J} \varphi(p)\psi_0(p)\varphi(p')\psi_0(p')$$

où $\qquad J = \{(p\ ,\ p')/p \neq p'\} \qquad$ et

où on a tenu compte de la relation $(\varphi(p)\psi_0(p))^2 = \varphi(p)\psi_0(p)$. L'expression (2) devient

$$\lambda\mu - \lambda^2\mu^2 + \sum_{J} \psi_0(p)\psi_0(p')\mathcal{E}(\varphi(p)\varphi(p')) \quad (3)$$

En partitionnant J en les deux classes G et H définies ci-dessus, ($J = G \cup H$) ; et en partageant la somme en deux parties : la première sur G et la seconde sur H ; on obtient le résultat annoncé à partir du Lemme 2.

Dans ces conditions l'indice de proximité que nous adoptons entre π_0 et π'_0 est

$$S(\pi_0, \pi'_0) = \frac{\text{card}(R(\pi_0) \cap R(\pi'_0)) - \lambda\mu}{\sqrt{\lambda\mu + \varphi\sigma}} \quad (I)$$

Posons $r = \text{card}(R(\pi_0))$ et $r' = \text{card}(R(\pi'_0))$ Soit T (resp. T') un élément variable de l'ensemble des parties de F à r(resp. r') éléments, muni d'une mesure de probabilité uniformément répartie. La distribution de card $(T \cap R(\pi'_0))$ est la même que celle de card$(R(\pi_0) \cap T')$; elle est hypergéométrique de moyenne $\lambda\mu$ et de variance $\lambda\mu(1 - r/f)\ (1 - r'/f)$ où $f = \text{card}(F)$; cette distribution est donc de *même moyenne* que celle envisagée ci-dessus mais de variance nettement plus petite ; son approximation par la loi normale est en général excellente.

L'étude comparative d'un couple de préordres totaux sera tout à fait parallèle à celle concernant un couple de partitions.

2.2. Distributions duales attachées à un couple de préordres totaux.

Soit E_2 l'ensemble des couples d'éléments distincts de E ; soit $(E \times E - \delta)$ où δ est la diagonale $\{(x, x)/x \in E\}$. La donnée d'un préordre total est équivalente à celle d'une partition et d'un ordre total sur l'ensemble des classes ; la *composition* d'un préordre total ω sur E est la suite $v = (n_1, n_2, \ldots, n_k)$ des cardinaux de ses différentes classes rangées selon l'ordre quotient. ω sera représenté par la partie R_ω de E_2 ; $R_\omega = \{(x, y)/x < y$ et non $y < x$ pour $\omega\}$. On a

$$\text{card}(R_\omega) = \sum_{i<j} n_i n_j = n(n-1)/2 - \sum_{i=1}^{k} n_i(n_i - 1)/2 .$$

Désignons par $\Omega(n ; v)$ (resp. $\Omega(n ; w)$) l'ensemble des préordres totaux sur E de composition $v = (n_1, \ldots, n_k)$ (resp. $w = (n'_1, \ldots, n'_k)$). Soit $(\omega_0, \bar{\omega}_0)$ un couple de préordres totaux de $\Omega(n ; v) \times \Omega(n ; w)$ et soit $R(\omega_0)$ (resp. $R(\bar{\omega}_0)$) l'ensemble de E_2 qui représente ω_0 (resp. $\bar{\omega}_0$) ;

$\text{card}(R(\omega_0) \cap R(\bar{\omega}_0)) = \text{card}\{(x, y)/x < y$ et non $y < x$ pour ω_0 et $\bar{\omega}_0\}$

constituera la *base* de la mesure de proximité entre les deux préordres totaux ; on a

$$\text{card}(R(\omega_0) \cap R(\bar{\omega}_0)) = \sum_{\substack{1 \leqslant i \leqslant (k-1) \\ 1 \leqslant j \leqslant (h-1)}} n_{ij} \sum_{p>i} \sum_{p>j} n_{pq}$$

où n_{ij} est le cardinal de l'intersection de la i-ème classe de ω_0 de la j-ème de $\bar{\omega}_0$.

Comme pour le cas des partitions ; à $\text{card}(R(\omega_0) \cap R(\bar{\omega}_0))$ associons deux variables aléatoires duales ; $\text{card}(R(\omega) \cap R(\bar{\omega}_0))$ et $\text{card}(R(\omega_0) \cap R(\bar{\omega}))$ où $R(\omega)$ (resp. $R(\bar{\omega})$) est la partie de E_2 associée à un élément courant de $\Omega(n ; v)$ (resp. $\Omega(n ; w)$) muni d'une mesure de probabilité uniforme. Posons pour abréger $R_0 = R(\omega_0)$ et $R'_0 = R(\bar{\omega}_0)$. Nous nous proposons d'établir ici, à l'aide du théorème qui constitue le second paragraphe que la distribution de $\text{card}(R(\omega) \cap R'_0)$ est la même que celle de $\text{card}(R_0 \cap R(\omega))$; on précisera d'autre part, la moyenne et la variance de cette distribution commune. La statistique $\text{card}(R_0 \cap R'_0)$ centrée et réduite définira une "bonne" mesure de proximité entre les deux préordres totaux.

Nous allons à présent travailler dans E_2^q où q est un exposant entier fixé. Si $(d_1, d_2, \ldots, d_q)$ est un q-uple de couples et a une auto-bijection de E ; prolongeons a à E_2^q en posant $a(d_1, \ldots, d_q) = (a(d_1), \ldots, a(d_q))$ où on pose $a(x, y) = (a(x), a(y))$. $(a(d_1), a(d_2), \ldots, a(d_q))$ a même configuration c que $(d_1, d_2, \ldots, d_q)$; c'est-à-dire, est tel que l'objet $a(x)$ remplace x dans les couples (composantes du q-uples) où était présent ce dernier. L'ensemble $G_q^{(c)}$ de tous les q-uples ayant la même configuration c, est l'ensemble des valeurs $\{(a(d_1), \ldots, a(d_q))\}$ où a décrit les n ! auto-bijections de E. On peut noter G_q^0 l'ensemble des q-uples de couples dont deux quelconques sont sans composante commune. Nous n'aurons pas besoin dans la suite de préciser, sauf pour $q \leqslant 2$, les différentes configurations c et les cardinaux associés des ensembles $G_q^{(c)}$.

Soient ω_0 un préordre total fixé dans $\Omega(n\,;v)$ et $(d_1^0, \ldots, d_q^0)$ un q-uple fixé dans $G_q^{(c)}$. Désignons par φ_0 (resp. φ) la fonction indicatrice de la partie $R(\omega_0)$ (resp. $R(\omega)$) de E_2 définie ci-dessus.

2.2.1. Théorème

La proportion de q-uples $(d_1, d_2, \ldots, d_q$ dans $G_q^{(c)}$ pour lesquels $\varphi_0(d_1)\varphi_0(d_2)\ldots\varphi_0(d_q) = 1$, est égale à la proportion dans $\Omega(n\,;v)$ de préordres totaux pour lesquels $\varphi(d_1^0)\varphi(d_2^0)\ldots\varphi(d_q^0) = 1$.

Il y a lieu d'établir que

$$\mu\{\varphi_0(d_1)\varphi_0(d_2)\ldots\varphi_0(d_q) = 1\} = \nu\{\varphi(d_1^0)\varphi(d_2^0)\ldots\varphi(d_q^0) = 1\} \qquad (1)$$

où le premier membre est la probabilité pour un q-uple $(d_1, d_2, \ldots, d_q)$ pris au hasard dans $G_q^{(c)}$ muni d'une mesure de probabilité uniforme, d'être tel que $\varphi_0(d_1)\varphi_0(d_2)\ldots\varphi_0(d_q) = 1$; le second membre est la probabilité pour un préordre total ω pris au hasard dans $\Omega(n\,;v)$ muni d'une probabilité uniforme, d'être tel que $\varphi(d_1^0)\varphi(d_2^0)\ldots\varphi(d_q^0) = 1$.

On commencera par remarquer l'isomorphisme entre $\Omega(n\,;v)$ et l'ensemble des partitions en classes étiquetées $\mathfrak{P}(n\,;v)$ mais où l'étiquette de la classe est son rang pour l'ordre quotient.

Comme dans le cas de la comparaison de deux partitions, nous allons décomposer chacun des évènements Q et Q^*, respectivement sous les signes μ et ν de la relation (1) à démontrer, en sous évènements disjoints qui se correspondent mutuellement

$$Q = \sum_{\beta \in B} Q_\beta \,,\; Q^* = \sum_{\beta \in B} Q_\beta^*$$

et nous établirons que $\mu(Q_\beta) = \nu(Q_\beta^*)$

Soit $(d_1, d_2, \ldots, d_q)$ un q-uple quelconque de $G_q^{(c)}$ et soit $\{x_1, x_2, \ldots, x_l\}$ le sous-ensemble des éléments de E dont chacun intervient au moins une fois comme composante de l'un des couples d_j, $1 \leqslant j \leqslant q$. Rappelons que si $\{x_1, x_2, \ldots, x_l\}$ et $\{x'_1, x'_2, \ldots, x'_l\}$ sont respectivement associés à deux q-uples distincts de $G_q^{(c)}$; il existe un bijection σ du premier ensemble dans l'autre telle que si x_k occupe la i-ème composante ($i = 1$ ou 2) de d_j, $\sigma(x_h)$ occupe la même composante de d'_j ou d'_j est la j-ème composante du second q-uple ; pour tout j, $1 \leqslant j \leqslant q$.

Considérons l'affectation des divers éléments de $e = \{x_1, \ldots . x_l\}$, associé à un q-uple aléatoire de $G_q^{(c)}$, dans les diverses classes de la partition $(E_1^\circ, E_1^\circ, \ldots, E_k^\circ)$ en classes étiquetées que définit le préordre total ω_0. Cette affectation définit une partition $(e_1, e_2, \ldots, e_k)$ pour laquelle on a

$$e_j \subset E^\circ_{\tau(j)} \qquad \text{pour tout} \qquad j,\ 1 \leqslant j \leqslant h \tag{2}$$

où τ est une injection de $\{1, 2, \ldots, h\}$ dans $\{1, 2, \ldots k\}$.

L'affectation se décompose en la constitution γ de la partition $(e_1, e_2, \ldots, e_h)$ de type $u = (l_1, l_2, \ldots, l_h)$, suivie de l'injection (2). Un évènement Q_β correspond à une affectation pour laquelle

$$\varphi_0(d_1)\varphi_0(d_2)\ldots\varphi_0(d_q) = 1$$

Dualement, constituons à partir de $(d_1^\circ, d_2^\circ, \ldots, d_q^\circ)$ la partition $(e_1^\circ, e_2^\circ, \ldots, e_h^\circ)$ de la même façon γ que l'a été $(e_1, e_2, \ldots, e_h)$ à partir de $(d_1, d_2, \ldots, d_q)$: "γ affecte la i-ème composante ($i = 1$ ou 2) du j-ème couple d_j dans la classe e_m". Un évènement Q_β^* dual de Q_β s'exprime par

$$e_j^\circ \subset E_{\tau(j)} \qquad \text{pour tout} \qquad j,\ 1 \leqslant j \leqslant h\ ; \tag{2*}$$

où on a $\varphi(d_1^\circ)\varphi(d_2^\circ)\ldots\varphi(d_q^\circ) = 1$, l'application τ étant la même que ci-dessus.

$\mu(Q_\beta)$ est la proportion de q-uples de $G_q^{(c)}$ dont la partition $(e_1, e_2, \ldots, e_h)$ γ-associée remplit la condition (2). Cette proportion est égale à celle de partitions dans $\mathcal{R}(l\ ;u)$ (cf. § 2.0) pour lesquelles on a (2) ; en effet le nombre de q-uples de $G_q^{(c)}$ qui déterminent par l'association γ une même partition $(e_1, e_2, \ldots, e_h)$ est indépendant de la partition choisie dans $\mathcal{R}(l\ ;u)$. Parallèlement $\nu(Q_\beta^*)$ est, en raison de l'isomorphisme entre $\Omega(n\ ;v)$ et $\mathcal{R}(n\ ;v)$, la proportion de partitions de $\mathcal{R}(n\ ;v)$ pour lesquelles on a (1*). Le théorème de dualité du paragraphe 2.0. établit précisément l'égalité de ces deux proportions.

En même temps que φ_0 et φ, on introduit ψ_0 (resp. ψ), fonction indicatrice de la partie $\mathcal{R}(\omega_0)$ (resp. $\mathcal{R}(\omega)$) de E_2 représentant un préordre total fixé (resp. courant) dans $\Omega(n\ ;\omega)$ (cf. début du paragraphe).

Par un argument analoque à celui du paragraphe 2.1.2, on établit le théorème suivant.

2.2.2 Théorème

$\Omega(n\,;v)$ et $\Omega(n\,;w)$ étant munis d'une mesure de probabilité uniformément répartie, la distribution de $\mathrm{card}(R(\omega)\cap R(\bar{\omega}_0))$ est la même que celle de $\mathrm{card}(R(\omega_0)\cap R(\bar{\omega}))$; ω et $\bar{\omega}$ parcourant respectivement $\Omega(n\,;v)$ et $\Omega(n\,;w)$.

2.2.3. Tendance centrale
et Dispersion de la distribution de card $(R(\omega)\cap R(\bar{\omega}_0))$

Nous commencerons par établir deux lemmes (1 et 2) qui précisent le théorème 2.2.1 pour $q = 1$ et $q = 2$ où nous expliciterons les différentes formes de $G_q^{(c)}$ et où sera calculée la valeur commune de la proportion que suppose l'énoncé du théorème. La détermination de la moyenne et de la variance de la distribution de $\mathrm{card}(R(\omega)\cap R(\bar{\omega}_0))$ s'en déduira.

2.2.3.1 Lemme 1

Soient $d_0 = (x_0, y_0)$ un couple fixé dans E_2 et ω_0 un préordre total fixé dans $\Omega(n\,;v)$. La proportion de couples $(x\,,\,y)$ de E_2 pour lesquels on a strictement $x < y$ pour ω_0, est égale à la proportion de préordres totaux de $\Omega(n\,;v)$ pour lesquels on a strictement $x_0 < y_0$; la valeur commune de cette proportion est $\sum_{i<j} n_i n_j / n(n-1)$ où on rappelle que $(n_1, n_2, \ldots, n_k)$ définit la composition v.

La valeur de la première des deux proportions indiquées apparaît clairement ; en effet, on a

$$\mathrm{card}\,(R(\omega_0)) = \sum_{i<j} n_i n_j \qquad \text{et} \qquad \mathrm{card}\,(E_2) = n(n-1)\,.$$

La deuxième proportion se met sous la forme

$$\sum_{i<j} \frac{(n-2)\,!}{n_1\,!\,n_2\,!\ldots(n_i-1)\,!\ldots(n_j-1)\,!\ldots n_k\,!} \Bigg/ \frac{n\,!}{n_1\,!\,n_2\,!\ldots n_i\,!\ldots n_j\,!\ldots n_k\,!} = \frac{\sum_{i<j} n_i n_j}{n(n-1)}$$

où un terme de la somme du premier membre représente le nombre de préordres totaux de composition v pour lesquels x_0 appartient à la i-ème classe et y_0 à la j-ème classe avec $i < j$.

L'introduction du Lemme 2 nécessite quelques définitions. Posons

$$E_4 = (E_2 \times E_2 - \Delta)$$

où Δ est la diagonale de $E_2 \times E_2$. On a

$$\text{card}\,(E_4) = n(n-1)\{n(n-1)-1\}.$$

E_4 est un ensemble de couples dont chaque composante est un couple d'éléments de E ; un objet de E_4 se met sous la forme $((x, y), (x', y'))$ où $x \neq y, x' \neq y'$ et $(x, y) \neq (x', y')$. Un élément (c, d) de E_4 peut prendre 6 formes distinctes, les deux extrêmes sont obtenues selon qu'aucune des deux composantes de c ne se répète dans d, ou que les deux se répètent. Les quatre autres formes sont obtenues lorsque l'une seulement des deux composantes de c se répète dans d, dans sa position ou non. Les six formes différentes sont

(0) $((x, y), (z, t))$

(1) $((x, y), (x, t))$; (1′) $((x, y), (z, x))$

(2) $((x, y), (z, y))$; (2′) $((x, y), (y, t))$

(3) $((x, y), (y, x))$

où des lettres différentes représentent des objets différents.

Soit $\{H, G_1, G_1', G_2, G_2', I\}$ la partition de E_4 où H est formé d'éléments de la forme (1) ; G_1', de la forme (1′) ; G_2, de la forme (2) ; G_2', de la forme (2′) et I, d'éléments de la forme (3).

On a

$$\text{card}\,(H) = n(n-1)(n-2)(n-3)$$

$$\text{card}\,(G_1) = \text{card}\,(G_1') = \text{card}\,(G_2) = \text{card}\,(G_2') = n(n-1)(n-2)$$

$$\text{card}\,(I) = n(n-1)$$

2.2.3.2. Lemme 2

Soit ω_0 un préordre total fixé dans $\Omega(n ; v)$ et soit $(c_0, d_0) = ((x_0, y_0), (x_0', y_0'))$ un élément fixé de E_4. Quelle que soit la forme de (c_0, d_0), la proportion de couples $((x, y), (x', y'))$ dans l'ensemble des couples de même forme que (c_0, d_0) pour lesquels on a $x < y$ et $x' < y'$ pour ω_0, est égale à la proportion de préordres totaux dans $\Omega(n ; v)$ pour lesquels $x_0 < y_0$ et $x_0' < y_0'$. La valeur commune de cette proportion est

$$\sum_{i<j} n_i n_j \sum_{i'<j'} m_{i'} m_{j'} \Big/ n(n-1)(n-2)(n-3) \quad \text{si} \quad (c_0, d_0) \in H$$

où

$m_{i'} = n_{i'}$ (resp. $(n_{i'} - 1)$) si $i' \neq i$ et $i \neq j$ (resp. $i' = i$ ou $i' = j$)

$m_{j'} = n_{j'}$ (resp. $(n_{j'} - 1)$) si $j' \neq i$ et $j' \neq j$ (resp. $j' = i$ *ou* $j' = j$)

$$\sum_{i<j} n_i n_j \sum_{h>i} m_h \Big/ n(n-1)(n-2) = \sum_{i<j} n_i n_j n_i^f \Big/ n(n-1)(n-2)$$

si $(c_0, d_0) \in G_1$

où $m_h = n_h$ (resp. $(n_h - 1)$) si $h \neq j$ (resp. $h = j$)

et où $n_i^f = \sum_{h>i} n_h - 1.$

$$\sum_{i<j} n_i n_j \sum_{h<j} m_h \Big/ n(n-1)(n-2) = \sum_{i<j} n_i n_j n_j^c \Big/ n(n-1)(n-2)$$

si $(c_0, d_0) \in G_1'$

où $m_h = n_h$ (resp. $(n_h - 1)$) si $h \neq i$ (resp. $h = i$)

et où $n_j^c = \sum_{h<j} n_h - 1.$

$$\sum_{i<j} n_i n_j n_i^c \Big/ n(n-1)(n-2) \quad \text{si} \quad (c_0, d_0) \in G_2$$

$$\sum_{i<j} n_i n_j n_j^f \Big/ n(n-1)(n-2) \quad \text{si} \quad (c_0, d_0) \in G_2'$$

$$0 \quad \text{si} \quad (c_0, d_0) \in I$$

Le nombre de couples de couples d'éléments de E de la forme (0), pour lesquels x, y, z et t appartiennent respectivement à la i-ème, j-ème, i'-ème et j'-ème classes de cardinaux respectifs n_i, n_j, $n_{i'}$ et $n_{j'}$, est égal à $n_i n_j m_{i'} m_{j'}$ où $m_{i'} = n_{i'}$ (resp. $(n_{i'} - 1)$) si $i' \neq i$ et $i' \neq j$ (resp. si $i' = i$ ou si $i' = j$) où on

suppose $i \neq j$ et $i' \neq j'$. En tenant compte de card (H), on obtient la première proportion annoncée de l'ensemble des couples $((x, y), (x', y'))$ dans H pour lesquels on a $x < y$ et $x' < y'$ pour ω_0.

Dualement, le nombre de préordres totaux de $\Omega(n ; v)$ pour lesquels x_0, y_0, z_0 et t_0 occupent respectivement la i-ème j-ème, i'-ème et j' classe, où $i \neq j$ et $i' \neq j'$, est égal à $(n-4)\,! \Big/ \prod_{i \leqslant h \leqslant k} l_h\,!$

où

$l_h = n_h$ si h est différent de chacun des indices

$$i, j, i' \quad \text{et} \quad j'$$

$l_i = (n_i - 1)$ (resp. $(n_i - 2)$) si $i' \neq i$ et $j' \neq i$ (resp. $i' = i$ ou $j' = i$)

$l_j = (n_j - 1)$ (resp. $(n_j - 2)$) si $i' \neq j$ et $j' \neq j$ (resp. $i' = i'$ ou $j' = j$)

de même

$l_{i'} = (n_{i'} - 1)$ (resp. $(n_{i'} - 2)$ si $i \neq i'$ et $j \neq i'$ (resp. $i = i'$ ou $j = i'$)

$l_{j'} = (n_{j'} - 1)$ (resp. $(n_{j'} - 2)$ si $i \neq j'$ et $j \neq j'$ (resp. $i = j'$ ou $j = j'$)

Le cardinal de $\Omega(n ; v)$ étant

$$n\,! \Big/ \prod_{1 \leqslant h \leqslant k} n_h\,!\,,$$

on obtient la même proportion que ci-dessus pour l'ensemble des préordres totaux dans $\Omega(n ; v)$ pour lesquels $x_0 < y_0$ et $z_0 < t_0$. Le calcul des autres proportions signalées dans l'énoncé du lemme est analogue à ce dernier ; nous allons cependant l'établir encore une fois dans le cas où (c_0, d_0) est de la forme $((x_0, y_0), (x_0, t_0))$.

Le nombre de couples de couples d'éléments de E de la forme (1) (i.e. $((x, y), ((x, t)))$, pour lesquels x, y et t appartiennent respectivement à la i-ème, j-ème et h-ème classes du préordre ω_0, est égal à $n_i n_j m_h$ où $i \neq j$, $i \neq h$ et où $m_h = n_h$ (resp. $(n_h - 1)$) si $h \neq j$ (resp. si $h = j$). On en déduit, en tenant compte de card(G_1), la proportion annoncée de l'ensemble des éléments $((x, y), (x', y'))$ dans G_1 pour lesquels on a $x < y$ et $x' < y'$ pour ω_0.

Dualement, le nombre de préordres totaux de $\Omega(n ; v)$ pour lesquels x_0, y_0 et t_0 occupent respectivement la i-ème, j-ème et h-ème classe, où $i \neq j$ et où $h > i$, est égal à

$$(n-3)\,!\ \Big/ \prod_{1 \leqslant p \leqslant k} l_p\,!$$

où

$l_p = n_p$ si p est différent de chacun des indices i, j et h

$l_i = (n_i - 1)$

$l_j = (n_j - 1)$ (resp. $(n_j - 2)$ si $h \neq j$ (resp. si $h = j$)

$l_h = (n_h - 1)$ (resp. $(n_h - 2)$) si $j \neq h$ (resp. si $j = h$).

En tenant compte de card$(\Omega(n\,;v))$, on obtient la même proportion que la dernière pour l'ensemble des préordres totaux dans $\Omega(n\,;v)$ pour lesquels $x_0 < y_0$ et $x_0 < t_0$.

Théorème

La *moyenne* commune et la *variance* commune des deux identiques distributions duales (celles de card $(R(\omega) \cap R(\overline{\omega}_0))$ et de card$(R(\omega_0) \cap R(\overline{\omega}))$), sont respectivement

$$\lambda\mu \quad \text{et} \quad \lambda\mu + \rho_{ff}\sigma_{ff} + \rho_{cc}\,\sigma_{cc} + 2\,\rho_{cf}\sigma_{cf} + (\Lambda M - \lambda^2\mu^2)$$

où

$$\lambda = \sum_{i<j} n_i n_j \Big/ \sqrt{n(n-1)}\ ,\quad \rho_{ff} = \sum_i n_i (n_i^f)^2 \Big/ \sqrt{n(n-1)\,(n-2)}\,.$$

$$\rho_{cc} = \sum_i n_i (n_i^c)^2 \Big/ \sqrt{n(n-1)\,(n-2)}\,,$$

$$\rho_{cf} = \sum_i n_i n_i^f n_i^c \Big/ \sqrt{n(n-1)\,(n-2)}$$

$$\Lambda = \sum_{i<j} n_i n_j \Big(\sum_{i'<j'} n_{i'} n_{j'} + n_i + n_j - 2n + 1 \Big) \Big/ \sqrt{n(n-1)\,(n-2)\,(n-3)}$$

Les expressions de μ, σ_{ff}, σ_{cc}, σ_{cf} et M ont respectivement la même forme que celles de ρ, ρ_{ff}, ρ_{cc}, ρ_{cf} et Λ ; les n_i étant remplacés par les n_i', $1 \leqslant i \leqslant h$, où $(n_1', n_2', \ldots, n_h') = w$

Remarquons que la variance est très sensiblement égale à

$$\lambda\mu + \rho_{ff}\sigma_{ff} + \rho_{cc}\sigma_{cc} + 2\,\rho_{cf}\sigma_{cf}.$$

En introduisant la fonction indicatrice φ(resp. ψ_0) de la partie $R(\omega)$ (resp. $R(\bar{\omega}_0)$) de E_2, on a

$$\mathcal{E}(\text{card}(R(\omega) \cap R(\bar{\omega}_0)) = \sum_{d \in E_2} \psi_0(d)\ \mathcal{E}(\varphi(d)) = \mathcal{E}(\varphi(d) \sum_{d \in E_2} \psi_0(d)$$

Lemme 1 ci-dessus permet d'établir le résultat concernant la valeur de la moyenne.

En adoptant, pour le calcul de la variance, le même shéma que celui considéré lors de la comparaison d'un couple de partitions ; on aura à déterminer

$$\sum_{(d,d') \in E_4} \psi_0(d)\psi_0(d')\ \mathcal{E}(\varphi(d)\varphi(d'));$$

En partitionnant E_4 selon les classes H, G_1, G_1', G_2 G_2' et I

$$(E_4 = H \cup G_1 \cup G_1' \cup G_2 \cup G_2' \cup I)\ ;$$

et en décomposant la somme en six parties, respectivement sur H, G_1, G_1', G_2, G_2' et I ; on obtient le résultat annoncé à partir du Lemme 2.

Dans ces conditions l'indice de proximité que nous adoptons entre ω_0 et $\bar{\omega}_0$ est

$$S(\omega_0, \bar{\omega}_0) = \frac{\text{card}(R(\omega_0) \cap R(\bar{\omega}_0)) - \lambda\mu}{\sqrt{\lambda\mu + \rho_{ff}\sigma_{ff} + \rho_{cc}\sigma_{cc} + 2\rho_{cf}\sigma_{cf}}} \qquad \text{(II)}$$

Posons $r = \text{card}(R(\omega_0))$ et $r' = \text{card}(R(\bar{\omega}_0))$. Soit T (resp. T') un élément variable de l'ensemble des parties de E_2 à r (resp. r') éléments, muni d'une mesure de probabilité uniformément répartie. La distribution de $\text{card}(T \cap R(\bar{\omega}_0))$ est la même que celle de $\text{card}(R(\omega_0) \cap T')$; elle est hypergéométrique de moyenne $\lambda\mu$ et de variance $\lambda\mu(1 - r/g)\ (1 - r'/g)$ où $g = \text{card}(E_2)$; cette distribution est donc de *même moyenne* que celle envisagée ci-dessus mais de variance notablement plus petite. L'approximation par la loi normale de cette distribution est en général excellente.

2.2.4. Comparaison de la statistique $S(\omega, \bar{\omega})$ avec celle de M.G. Kendall.

Dans son ouvrage "Rank Corrélation Methods" (Chapitre 3), M.G. Kendall propose d'étendre sa statistique τ, établie pour la comparaison de deux ordres totaux et dont nous reparlerons, au cas de deux préordres totaux et ce, en retenant l'algorithme de calcul lui ayant servi à déterminer τ. Par rapport à nos notations, la statistique se met sous la forme

$$T_{(\omega,\bar{\omega})} = \frac{\text{card}\,(R_\omega \cap R_{\bar{\omega}}) - \frac{1}{2}\sum_{(i,j)} n_{ij} \sum_{p>i} \sum_{q\neq j} n_{pq}}{\sqrt{\frac{n(n-1)}{4}\,\lambda\mu}}$$

Nous allons montrer, en nous appuyant sur un exemple, que cette statistique est *biaisée*. Posons

$$\Gamma = \frac{1}{2}\sum_{(i,j)} n_{ij} \sum_{\substack{p>i \\ q\neq j}} n_{pq} = \frac{1}{2}\sum_{(i,j)} n_{ij} \sum_{p>i} (n_p - n_{pj})$$

où $\quad n_p = \sum_j n_{pj}$

$$= \frac{1}{2}\left\{ \sum_{i<p} n_i n_p - \sum_{(i,j)} n_{ij} \sum_{p>i} n_{pj} \right\}$$

La moyenne de $\text{card}\,(R_\omega \cap R_{\bar{\omega}})$, par rapport à l'hypothèse N, étant

$$\lambda\mu = \frac{1}{n(n-1)} \sum_{i<j} n_i n_j \sum_{p<q} n'_p n'_q$$

La différence

$$\Gamma - \lambda\mu = \left(\sum_{i<j} n_i n_j\right)\left(\frac{1}{2} - \frac{1}{n(n-1)} \sum_{p<q} n'_p n'_q\right) - \frac{1}{2}\sum_{(i,j)} n_{ij} \sum_{p>i} n_{pj}$$

Compte tenu des relations de la forme

$$\sum_{i<j} n_i n_j = n(n-1)/2 - \sum_i n_i (n_i - 1)/2\,,$$

On a

$$4(\Gamma - \lambda\mu) = n(n-1)\,(1 - \sum_i n_i(n_i - 1)/\,n(n-1))$$

$$\left(\sum_p n'_p(n'_p - 1)/n(n-1)\right) - 2\sum_{(i,j)} n_{ij} \sum_{p>i} n_{pj}$$

Considérons l'exemple où le préordre ω comprend deux classes C_1 et C_2 telles que $\text{card}(C_1) = 4n/5$ et $\text{card}(C_2) = n/5$, le préordre $\bar{\omega}$ quatre classes sur lesquelles les éléments de C_1 (resp. C_2) se trouvent également répartis ; alors le tableau de contingence à la forme suivante où i est l'indice de ligne et j de colonne.

$n/5$	$n/5$	$n/5$	$n/5$	$n/5$	$4n/5$
$n/20$	$n/20$	$n/20$	$n/20$	$n/20$	$n/5$
$n/4$	$n/4$	$n/4$	$n/4$	$n/4$	n

$$\Gamma - \lambda\mu = \frac{1}{4}\left\{n(n-1) - \frac{4n}{5}\left(\frac{4}{5}n - 1\right) - \frac{n}{5}\left(\frac{n}{5} - 1\right)\right\}$$

$$\left\{4 \cdot \frac{n}{4}\left(\frac{n}{4} - 1\right)/n(n-1)\right\} - \frac{1}{2}\left(4 \times \frac{n}{5} \cdot \frac{n}{20}\right)$$

$$= -\frac{3}{50} \cdot \frac{n^2}{(n-1)} \quad ; \text{ soit en valeur absolue, un biais de } 6n/100.$$

On peut remarquer aussi que le dénominateur de la statistique T ne tient pas compte de la variance dans l'hypothèse N de la distribution de

$$\text{card}\,(R_\omega \cap R_{\bar{\omega}}) .$$

3. COMPARAISON D'UN COUPLE D'ORDRES TOTAUX SUR E

Il s'agit d'un cas particulier de l'étude précédente ; comme ci-dessus on se place dans E_2 pour représenter un ordre total o par son graphe

$$R(o) = \{(x, y)/(x, y) \in E_2 , x < y \quad \text{pour} \quad o\}$$

où $\text{card}\,(R(o)) = n(n-1)/2$.

Relativement à deux ordres totaux o_0 et o_0' sur E, la base de la mesure de leur proximité sera naturellement $\text{card}(R(o_0) \cap R(o_0'))$ qui se met sous la forme

$$\sum_{i=1}^{(n-1)} g(i)$$

où i dèsigne l'élément de rang i pour o_0 et $g(i)$ le nombre d'objets de E d'indice plus grand que i et situés à droite de i pour o'.

Soit O l'ensemble des ordres totaux sur E ; $\text{card}(O) = n!$. Par la même technique que ci-dessus on détermine simplement la *moyenne* et la *variance* de la distribution de $\text{card}(R(o_0) \cap R(o'))$, lorsque o' parcourt O muni d'une mesure uniforme ; ce sont respectivement

$$n(n-1)/4 \qquad \text{et} \qquad n(n-1)(2n+5)/72$$

La statistique

$$T(o_0, o'_0) = \frac{\text{card}(R(o_0) \cap R(o'_0) - n(n-1)/4}{\sqrt{n(n-1)(2n+5)/72}}$$

n'est autre que celle τ proposée par M.G. Kendall qui établit que la distribution commune de $\text{card}(R(o_0) \cap R(o'))$ et de $\text{card}(R(o_0) \cap R(o'))$ lorsque o (resp. o') décrit O muni d'une mesure de probabilité uniforme, est asymptotiquement, pour $n \to \infty$, normal

4. COUPLE DE VARIABLES INDICATEUR D'UN COUPLE DE MESURES SUR E

Les deux variables définissent respectivement les distributions suivantes $\xi = (x_1, x_2, \ldots, x_i, \ldots, x_m)$ et $\eta = (y_1, y_2, \ldots, y_i, \ldots, y_n)$ où x_i (resp. y_i) est la charge affectée au i-ème objet de E par la première (resp. la seconde) variable. La base de l'établissement du coefficient de proximité est naturellement

$$\sum_{i \leqslant i \leqslant n} x_i y_i \tag{1}$$

dont il y a lieu d'examiner la distribution dans l'hypothèse N que nous allons préciser. Pour cette hypothèse les distributions de chacune des deux variables sont fixées ; mais la position relative de l'une des distributions $(x_i / 1 \leqslant i \leqslant n)$ par rapport à l'autre $(y_i / 1 \leqslant i \leqslant n)$ est inconnue à priori. Il y a donc lieu d'étudier la loi de la statistique

$$\sum_{1 \leqslant i \leqslant n} x_{\sigma(i)} y_i \tag{1''},$$

la *moyenne* commune et la *variance* commune étant respectivement

$$\frac{1}{n} \sum_{1 \leqslant i \leqslant n} x_i \sum_{1 \leqslant i \leqslant n} y_i \qquad \text{et} \qquad \frac{1}{(n-1)} \sum_{1 \leqslant i \leqslant n} (x_i - \mu_x)^2 \sum (y_i - \mu_y)^2$$

où μ_x (resp. μ_y) est la moyenne des x_i (resp. des y_i) ;

$$\mu_x = \frac{1}{n} \sum_{1 \leqslant i \leqslant n} x_i \, , \; \mu_y = \frac{1}{n} \sum_{1 \leqslant i \leqslant n} y_i \; ; \text{ en désignant par}$$

$$\sigma_x^2 = \frac{1}{n} \sum_{1 \leqslant i \leqslant n} (x_i - \mu_x)^2 \qquad \text{et par} \qquad \sigma_y^2 = \frac{1}{n} \sum_{1 \leqslant i \leqslant n} (y_i - \mu_y)^2 \; ,$$

l'indice de proximité adopté entre les deux distributions ξ et η prend la forme suivante

$$M(\xi \, , \eta) = \Big(\sum_{1 \leqslant i \leqslant n} x_i y_i - \eta \mu_x \mu_y \Big) \Big/ \sqrt{\frac{n^2}{(n-1)} \sigma_x^2 \sigma_y^2} \qquad (2)$$

L'étude du caractère asymptotique de la distribution de la statistique (2) dans l'hypothèse N fait l'objet du théorème de A. Wald et J. Wolfowitz (cf. [10]) dont nous allons rappeler la forme précisée par Noether.

Théorème

Soit (ξ_n) (resp. (η_n)) une suite de suites de nombres réels où

$$\xi_n = (x_1^n, x_2^n, \ldots, x_n^n) \text{ (resp. } \eta_n = (y_1^n, y_2^n, \ldots, y_n^n))$$

est une suite de longueur n. Au couple (ξ_n, η_n) associons la variable aléatoire

$$X_n = \sum_i x_{\sigma(i)}^n y_i^n \; ;$$

Si (ξ_n) (resp. (η_n)) remplit la condition (V) (resp. (W)) ci-dessous précisée ; alors

$$(X_n - \mathscr{E}(X_n))/\sigma(X_n) \, ,$$

où $\mathscr{E}(X_n)$ et $\sigma^2(X_n)$ sont la moyenne et la variance de la variable aléatoire X_n, suit asymptotiquement une loi normale centrée réduite.

Conditions (V) et (W).

Relativement à une suite de suites de nombres réels

$$(\alpha_n) \qquad \text{où} \qquad \alpha_n = (a_1^n, a_2^n, \ldots, a_n^n)$$

est une suite de longueur n ; la condition (V) exprime que pour tout entier r fixé, $r \geqslant 3$,

$$\sum_{1 \leq i \leq n} (a_i^n - \bar{a}_n)^r \Big/ \Big(\sum_{1 \leq i \leq n} (a_i^n - \bar{a}_n)^2 \Big)^{r/2} = o(1)$$

et la condition (W)

$$\frac{1}{n} \sum_{1 \leq i \leq n} (a_i^n - \bar{a}_n)^r \Big/ \Big(\frac{1}{n} \sum_{1 \leq i \leq n} (a_i^n - \bar{a}_n)^2 \Big)^{r/2} = O(1)$$

où $\bar{a}_n = \frac{1}{n} \sum_{1 \leq i \leq n} a_i^n$, o et O correspondent aux notations de Landau.

La condition (W) a été introduite par Wald et Wolfowitz et celle, moins restrictive, (V) par Noether.

4.1. – Dans le cas d'un couple de variables de même type algébrique a), d) ou e) (cf. § I), des résultats théoriques montrent qu'on peut se référer à l'échelle définie par la loi normale centrée réduite pour juger de la valeur de l'indice de proximité entre les deux variables. Nous allons nous permettre, pour deux raisons, de faire également référence à la loi normale pour juger de la grandeur de chacun des deux coefficients de proximité $S(\pi, \pi')$ et $S(\omega, \bar{\omega})$ etablis respectivement pour un couple de variables de type b) ou c) (cf. formule (I) (resp. (II)) § 2.1.3 (resp. § 2.2.3)) ; la première raison est que le principe de l'établissement de la statistique de proximité est le même dans ces cas que dans ceux d) ou e) ; et la seconde est fournie par le commentaire ayant suivi chacune des deux formule (I) et (II) donnant l'expression de $S(\pi, \pi')$ et de $S(\omega, \bar{\omega})$. Finalement quel que soit le type algébrique du couple (ζ, θ) de variables ; si $u_0 = \cup(\zeta_0, \theta_0)$ est la valeur de la statistique de proximité établie sur (ζ_0, θ_0)

$$P(\rho_0, \theta_0) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u_0} e^{-x^2/2}\, dx \qquad (1)$$

définit une mesure de la "ressemblance" entre les deux variables où la notion de ressemblance est clairement remplacée par la notion de *vraisemblance* par rapport à l'hypothèse N. Nous ferons ci-dessous référence à ces deux échelles de mesure de la proximité entre deux variables de même type algébrique ; la première sera notée $\cup$ et la seconde P.

III – PROXIMITÉ ENTRE CLASSES DE VARIABLES DE MÊME TYPE

Relativement à un tableau de données qui croise un ensemble A de variables descriptives et un ensemble E d'objets ou de sujets, il y a lieu d'étendre la notion de proximité entre deux variables à celle entre deux classes de variables car de la sorte on établira une hiérarchie ascendante de classifications où à chaque pas on réunit les deux classes les plus voisines.

Soient B et C deux parties disjointes de A définissant deux classes de variables de cardinaux respectifs l et m. Considérons l'ensemble des valeurs de la mesure de proximité sur l'ensemble des couples (β, γ) où β décrit B et γ, C. Soit

$$\{\cup(\beta, \gamma)/(\beta, \gamma) \in B \times C\} \tag{1}$$

ou bien

$$\{P(\beta, \gamma)/(\beta, \gamma) \in B \times C\} \tag{2}$$

avec des notations que nous venons, ci-dessus, de préciser. Deux notions de proximité entre classes se présentent de façon, naturelle ; *la première* est basée sur *la somme des proximités* $\cup(\beta, \gamma)$ et la *seconde* sur *la plus grande proximité* observée entre un élément de B et un élément de C. Nous allons comme précédemment faire référence à l'hypothèse N d'absence de liaison pour établir chacune des deux statistiques de proximité.

Dans l'hypothèse N, l'ensemble suivant des valeurs de $\cup(\zeta, \theta)$;

$$\{\cup(\beta, \gamma)/\gamma \in C\}$$

est un échantillon de m points indépendants d'une variable aléatoire normale entrée réduite $\mathfrak{N}(0, 1)$; en notant

$$\cup(\beta, C) = \sum_{\gamma \in C} \cup(\beta, \gamma)$$

$\frac{1}{\sqrt{m}} \cup(\beta, C)$ est la réalisation d'une v.a. $\mathfrak{N}(0, 1)$.

Dans ces conditions l'ensemble des valeurs

$$\left\{ \frac{1}{\sqrt{m}} \cup(\beta, C) / \beta \in B \right\}$$

est, dans l'hypothèse N, un échantillon de l points indépendants d'une v.a. $\mathfrak{N}(0, 1)$; d'où

$$\frac{1}{\sqrt{l}} \sum_{\beta \in B} \frac{1}{\sqrt{m}} \cup(\beta, C)$$

est, dans l'hypothèse N, une réalisation d'une va. $\mathfrak{N}(0,1)$. Il en résulte que la première statistique de proximité à adopter entre les deux classes B et C, est

$$\cup(B, C) = \frac{1}{\sqrt{lm}} \sum_{(\beta,\gamma) \in C \times C} \cup(\beta, \gamma) \qquad (1')$$

Compte tenu de l'échelle de référence définie dans l'hypothèse d'absence de liaison par la loi $\mathfrak{N}(0,1)$, cette statistique permet de comparer sans biais les proximités $\cup(B, C)$ et $\cup(B', C')$ attachées à deux couples de classes.

La seconde mesure de proximité entre classes se conçoit à partir de

$$P'(B, C) = \max_{(\beta,\gamma) \in B \times C} P(\beta, \gamma).$$

qu'on peut écrire, $P'(B, C) = \max_{\beta \in B} P'(\beta, C)$ où $P'(\beta, C) = \max_{\gamma \in C} P(\beta, \gamma)$;

Or l'ensemble des valeurs $\{P(\beta, \gamma) / \gamma \in C\}$ constitue dans l'hypothèse N, un échantillon de m points indépendants d'une variable aléatoire uniformément répartie entre 0 et 1 ; d'où

$$P_r^N \{P'(\beta, C) < t\} = t^m \qquad \text{avec} \qquad 0 < t < 1 .$$

D'autre part, l'ensemble des valeurs $\{P'(\beta, C) / \beta \in B\}$ constitue dans l'hypothèse N, un échantillon de l points indépendants d'une variable aléatoire dont la fonction de répartition vient d'être établie ; par conséquent

$$P_r^N \{P'(B, C) < t\} = (t^m)^l = t^{lm} \ ;$$

et si t_0 est la valeur observée de $P'(B\ ,\ C)$, on retiendra comme mesure de la proximité entre les deux classes

$$P(B\ ,\ C) = t_0^{lm} \qquad (2')$$

Selon le type de données dont on dispose il peut être préférable d'utiliser soit (1'), soit (2') comme coefficient de proximité entre classes de variables d'un même type. Les premières expériences montrent que (2') produit des résultats très fins lorsque A est formé de variables de type a) ou c) (cf. § I) ; une même réalité sous-jacente à un tableau de données peut différemment se manifester selon qu'on utilise (1') ou (2').

IV – DISTRIBUTION D'UN CRITÈRE DE CLASSIFICATION SUR L'ENSEMBLE DES PARTITIONS DE TYPE FIXÉ

Un critère de classification est un indice de proximité entre une partition et une information relative à la ressemblance entre éléments de l'ensemble D à classifier, qui peut être l'ensemble A des variables descriptives ou celui E des objets décrits. Cette information se présente en général comme un indice de proximité sur D affectant à chaque paire $p = \{x , y\}$ d'éléments de D un nombre réel $\mathcal{S}(p)$ sensé refléter la ressemblance entre x et y. Nous n'avons pas discuté ci-dessus de l'établissement d'une mesure de similarité lorsque $D = E$; cependant la situation est tout à fait symétrique lorsque A est formé d'attributs descriptifs (variables de type a), cf. § I) ; d'autre part, des considérations géométriques et métriques permettent de concevoir l'indice de proximité dans le cas où A est formé de variables numériques ; enfin, si A est formé de variables ordinales (type c) ou d), § I), on se référera à la fonction de répartition observée de ces variables pour établir l'indice de similarité $\mathcal{S}(p)$. Posons

$$F = \{\{x , y\}/x \in D , y \in D , x \neq y\} \qquad \text{où} \qquad f = \operatorname{card}(F) = \frac{n(n-1)}{2}$$

avec $n = \operatorname{card}(D)$.

Si on suppose, ce qui est assez général, que l'application qui à chaque $p = \{x , y\}$ de F associe sa mesure $\mathcal{S}(p)$, est injective ; l'ordre ω sur F défini par

$$p < q \iff \mathcal{S}(p) < \mathcal{S}(q)$$

est total. Il s'agit de l'"ordonnance" sur D. Une telle information, relative à la ressemblance entre éléments de D, a des propriétés intéressantes de stabilité par rapport au choix de la mesure de similarité $\mathcal{S}$. (cf. [4], Chap. 1) ; d'autre part, sa donnée ramène le problème de la définition d'un critère de classification à la comparaison de deux structure de même type : préordres totaux

sur l'ensemble F ; en effet, la donnée d'une partition π sur D est équivalente à celle d'un préordre total à deux classes (S, R) sur F où S(resp. R) est l'ensemble des paires séparées (resp. réunies) par π ; pour faire image, S(resp. R) est l'ensemble des paires dont les deux composantes sont considérées éloignées (resp. proches) du point de vue de la partition. Le critère de classification qui sert à juger de l'adéquation d'une partition aux données permettra de mesurer la cohésion des classes formées à un niveau donné d'un arbre de classifications.

$\mathfrak{P}(n\,;\,t)$ désignera ici l'ensemble des partitions sur D de type

$$t = (n_1, n_2, \ldots, n_k)\,.$$

Posons $\mathcal{B}(r)$ l'ensemble des relations binaires symétriques b sur D pour lesquelles la parties $\mathcal{R}(b)$ de F, formée des paires $\{x, y\}$ pour lesquelles on a $b(x, y)$, est de cardinal r ; il y a une correspondance bijective entre $\mathcal{B}(r)$ et l'ensemble des parties de F à r éléments. En fixant $r = \Sigma\ n_i(n_i - 1)/2$ et en notant $\mathcal{R}(n\,;\,t)$ l'ensemble des relations d'équivalence associé à $\mathfrak{P}(n\,;\,t)$, on a $\mathcal{R}(n\,;\,t) \subset \mathcal{B}(r)$. Ω indiquera l'ensemble des ordres totaux pouvant être définis sur F, lequel est de cardinal $f!$.

La base de la construction de l'indice de proximité entre la partition π et l'ordre ω sera

$$\text{card}\,(gr(\omega) \cap S(\pi) \times R(\pi)) \tag{1}$$

où $gr(\omega)$ désigne le graphe dans $F \times F$ de ω, soit

$$\{(p, q)/(p, q) \in F \times F \quad \text{et} \quad p < q \quad \text{pour} \quad \omega\}.$$

Nous commencerons par préciser la forme limite des deux identiques distributions duales que sont celles de $\text{card}\,(gr(\omega) \cap S(b_0) \times R(b_0))$ et de $\text{card}\,(gr(\omega_0) \cap S(b) \times R(b))$ où b_0 (resp. ω_0) est un élément fixé de $\mathcal{B}(r)$ (resp. Ω) et où ω (resp. b) est un élément aléatoire de Ω (resp. $\mathcal{B}(r)$) muni d'une probabilité uniforme ; cette forme limite de la distribution commune est donnée par *la loi normale.* Nous comparerons ensuite la distribution de $\text{card}\,(gr(\omega_0) \cap S(\pi) \times R(\pi))$, où π décrit $\mathfrak{P}(n\,;\,t)$ muni d'une probabilité uniforme, à celle de $\text{card}\,(gr(\omega_0) \cap S(b) \times R(b))$ où on suppose $\mathcal{R}(n\,;\,t) \subset \mathcal{B}(r)$; les deux distributions ont *la même moyenne* et la forme asymptotique de la première distributions est, sous des conditions assez générales, la même que la seconde, qui est par conséquent *normale.*

1. DISTRIBUTIONS DUALES ATTACHEES AU COUPLE (ω_0 , b_0) de $\Omega \times \mathcal{B}(r)$.

1.1. Expression de $\text{card}(gr(\omega) \cap S \times R)$

R et S étant deux parties complémentaires de F,

$$\text{card}(gr(\omega) \cap S \times R) = \sum_{p \in R} \text{card}(gr(\omega) \cap S \times \{p\}) \qquad (1) ;$$

en associant à chaque p de F son rang pour ω :

$$k(p) = \text{card}\{p'/p' \in F \quad \text{et} \quad p' \leqslant p \quad \text{pour} \quad \omega\} ;$$

et en désignant par $k(p_i)$ le rang de la i-ème paire de R rencontrée en parcourant F de gauche à droite selon ω, on a

$$\text{card}(gr(\omega) \cap S \times \{p\}) = (k(p_i) - 1) - (i - 1) = k(p_i) - i ;$$

ainsi le second membre de (1) se met sous la forme

$$\sum_{1 \leqslant i \leqslant r} k(p_i) - r(r + 1)/2 \qquad \text{où} \qquad r = \text{card}(R)$$

et en introduisant la fonction indicatrice $((\epsilon(p)), p \in F)$ de R, la dernière expression devient

$$\text{card}(gr(\omega) \cap S \times R) = \sum_{p \in F} \epsilon(p)\, k(p) - r(r + 1)/2 \qquad (2).$$

1.2 Etude des deux distributions duales

La formule (2) montre qu'on peut ramener l'étude de chacune des deux distributions ; celle de

$\text{card}(gr(\omega) \cap S(b_0) \times R(b_0))$ et celle de $\text{card}(gr(\omega_0) \cap S(b) \times R(b))$,

à l'examen de la distribution de

$$\sum_{p \in F} \epsilon(p)\, k(p)$$

qui apparaît dans chacun des deux cas comme étant celle de la somme de r entiers parmi $1, 2, \ldots, f$.

A la suite des entiers $(1, 2, \ldots, f)$, attachons les caractéristiques suivantes

a) La moyenne $\mu(k) = \frac{1}{f}\frac{f(f+1)}{2} = \frac{(f+1)}{2}$

b) La variance $\mu_2(k) = \frac{1}{f}\sum_{1 \leqslant i \leqslant f} i^2 - \left(\frac{f+1}{2}\right)^2 = (f-1)(f+1)/2$

c) Le coefficient de Wald et Wolfowitz $W_h = \mu_h(k)/(\mu_2(k))^{h/2}$

$$W_h = \frac{1}{f}\sum_{1 \leqslant i \leqslant f}\left(i - \frac{f+1}{2}\right)^h \Big/ \left\{\frac{1}{f}\sum_{1 \leqslant i \leqslant f}\left(i - \frac{f+1}{2}\right)^2\right\}^{h/2} ;$$

un calcul simple montre que

$$|W_h| \leqslant 3^{h/2}$$

ce qui assure la condition (W) du théorème (§ II.4). D'autre part, considérons la condition (v) de Noether du même théorème relativement à la suite $(\epsilon(p)/p \in F)$; elle se met sous la forme

$$\sum_{p \in F}\left(\epsilon(p) - \frac{r}{f}\right)^h \Big/ \left\{ f\left(\frac{r}{f}\left(1 - \frac{r}{f}\right)\right\}^{h/2} \quad \text{où} \quad h \geqslant 3 \quad \text{et}$$

où $r = \text{card}(R)$, cette expression se met sous la forme

$$(-1)^h\left(\frac{r}{f}\right)^{h/2} s^{-(h-2)/2} + \left(\frac{s}{f}\right)^{h/2} r^{-(h-2)/2}$$

où $s = \text{card}(S)$.

Cette dernière quantité tend vers zéro pour $n = \text{card}(D)$ tendant vers l'infini pourvu que le rapport r/f ne tende ni vers zéro, ni vers un. Dans ces conditions, l'application du théorème de Wald et Wolfowitz donne

Théorème

La distribution de $\sum_{p \in F} \epsilon(p)k(p)$ qui est de *moyenne* $r(f+1)/2$ et de *variance* $rs(f+1)/12$ est asymptotiquement normale.

Il en résulte que la distribution de

$$\frac{\text{card}\,(gr(\omega_0) \cap S(b) \times R(b) - r.\,s/2}{\sqrt{r.\,s(f+1)/12}} \qquad (1),$$

où b est un élément aléatoire de $\mathcal{B}(r)$ muni d'une probabilité uniforme, est asymptotiquement normale centrée et réduite.

En remplaçant $k(p)$ par

$$k'(p) = \frac{k(p) - \mu(k)}{\sqrt{\mu_2(k)}} = \frac{k(p) - \dfrac{f+1}{2}}{\sqrt{(f^2-1)/12}},$$

(1) se met sous la forme

$$\frac{1}{\sqrt{r.\,s/(f-1)}} \sum_{p \in F} \epsilon(p) k'(p) \qquad (1')$$

1.3. Comparaison de la distribution de card $(gr(\omega_0) \cap S(\pi) \times R(\pi))$ à celle de card $(gr(\omega_0) \cap S(b) \times R(b))$

Compte tenu de la relation (2) du paragraphe (1.1) ci-dessus, il y a lieu de comparer la distribution de

$$C(\pi) = \sum_{p \in F} \epsilon(p)\, k(p), \qquad \text{à celle de} \qquad C(b) = \sum_{p \in F} \beta(p)\, k(p)\ ; \qquad (1)$$

où π(resp. b) est un élément aléatoire de $\mathcal{R}(n\,;t)$ (resp. $\mathcal{B}(r)$) muni d'une mesure de probabilité uniforme et où $(\epsilon(p)/p \in F)$ (resp. $(\beta(p)/p \in F)$) est la fonction indicatrice de la partie de F que définit la relation d'équivalence (resp. binaire) associée à π (resp. b)

Théorème 1.

Si l'ensemble $\mathcal{R}(n\,;t)$ des relations d'équivalence associé à $\mathcal{R}(n\,;t)$ est inclus dans $\mathcal{B}(r)$; la moyenne de $C(\pi)$ est égale à celle de $C(b)$; la valeur commune de la moyenne étant $r(f+1)/2$.

La moyenne de $C(b)$ a déjà été calculée ci-dessus ; celle de $C(\pi)$ se met sous la forme

$$\frac{1}{\text{card}\,(\mathcal{R}(n\,;t))} \sum_{\pi} \sum_{p} \epsilon(p)\, k(p),$$

où la première somme est étendue à F et la seconde à $\mathcal{P}(n\,;t)$. En inversant les deux signes sommes on obtient

$$\sum_{p} \left\{ \frac{1}{\operatorname{card}(\mathcal{P}(n\,;t))} \sum_{\pi} \epsilon(p) \right\} k(p).$$

La quantité entre accolades est la proportion de partitions dans $\mathcal{P}(n\,;t)$ pour lesquelles les deux composantes de p sont réunies ; cette proportion est égale à

$$\sum_{1 \leqslant i \leqslant k} n_i(n_i - 1)/n(n-1) = r/f \qquad \text{(cf. Lemme 1. §.II. 2.1.3)}$$

d'où le résultat.

Nous allons à présent comparer les moments d'ordre l de chacune des deux statistiques $C(\pi)$ et $C(b)$ où nous supposerons vérifiée la condition énoncée par le théorème ci-dessus.

1.3.1 Expression du moment d'ordre l de C(b) et de C(π)

$$\mathcal{E}\left(\sum_{p} \beta(p)\, k(p) \right)^l = \frac{1}{\operatorname{card}(\mathcal{B}(r))} \sum_{b} \left(\sum_{p} \beta(p)\, k(p) \right)^l \tag{1}$$

En développant, on obtient après inversion des deux signes sommes et en tenant compte de la relation $\beta(p) = 0$ ou 1.

$$\mathcal{E}(C(b))^l = \sum \mu(p_{i_1}, p_{i_2}, \ldots, p_{i_m})\, c(l\,; l_1, \ldots, l_m)\, k(p_{i_1})^{l_1} \ldots k(p_{i_m})^{l_m} \tag{2}$$

où on a

$l = l_1 + l_2 + \ldots + l_m$; la sommation est étendue à toutes les permutations $(p_{i_1}, p_{i_2}, \ldots p_{i_m})$ pouvant être obtenues à partir de chacune des parties à m éléments de $F = \{p_1, p_2, \ldots, p_f\}$; $\mu(p_{i_1}, \ldots, p_{i_m})$ est la proportion dans $\mathcal{B}(r)$ de relations binaires symétriques pour lesquelles on a $\beta(p_{i_1})\beta(p_{i_2}) \ldots \beta(p_{i_m}) = 1$;

$$c(l\,; l_1, l_2, \ldots, l_m) = \frac{l\,!}{e_1\,!\, e_2\,! \ldots e_h\,!\, l_1\,!\, l_2\,! \ldots l_m\,!} \tag{3}$$

où h est le nombre de l_j distincts, chacun d'entre eux se répétant $e_1, e_2, \ldots, e_h$ fois.

L'expression de $\mathfrak{S}(C(\pi))^l$ est obtenue à partir de celle (2) de $\mathfrak{S}(C(b))^l$ en remplaçant $\mu(p_{i_1}, p_{i_2}, \ldots p_{i_m})$ par $\nu(p_{i_1}, p_{i_2}, \ldots, p_{i_m})$ qui est la proportion de partitions dans $\mathfrak{P}(n\,;t)$ pour lesquelles les paires $p_{i_1}, p_{i_2}, \ldots, p_{i_m}$ sont formées de composantes réunies, soit $\epsilon(p_{i_1}) \ldots \epsilon(p_{i_m}) = 1$.

En constatant que la proportion dans $\mathfrak{B}(r)$ de relations binaires symétriques pour lesquelles on a $b(x, y)$, où $p = \{x, y\}$ est une paire donnée, est égale à

$$\binom{f-1}{r-1}\Big/\binom{f}{r} = r/f\,;$$

on se rend compte que $\mu(p_{i_1}, p_{i_2}, \ldots, p_{i_m})$ est, pour n grand, très sensiblement égal à

$$(r/f)^m \simeq \Big(\sum_{1 \leqslant i \leqslant k} \pi_i^2\Big)^m \qquad \text{où} \qquad \pi_i = n_i/n \qquad \text{avec} \qquad t = (n_1, n_2, \ldots, n_k) \tag{4}$$

Si $(u_1, u_2, \ldots, u_g)$ est le type de la partition γ définie par la saturation de l'ensemble des paires $\{p_{i_1}, p_{i_2}, \ldots, p_{i_m}\}$, on a

$$\nu(p_{i_1}, p_{i_2}, \ldots, p_{i_m}) \simeq$$

$$(\pi_1^{u_1} + \pi_2^{u_1} + \ldots + \pi_k^{u_1}) \ldots (\pi_1^{ug} + \pi_2^{ug} + \ldots + \pi_k^{ug}) = \prod_{j=1}^{g} \sum_{i=1}^{k} \pi_i^{uj} \tag{5}.$$

En effet, chaque produit, terme du développement de (5), est associé à une application φ de $\{1, 2, \ldots, g\}$ dans $\{1, 2, \ldots, k\}$; celui qui est associé à l'application φ pour laquelle l'image de j est i, définit la proportion de partitions de type $t = (n_1, \ldots, n_k)$ pour lesquelles les u_j objets de la j-ème classe de γ se trouvent réunis dans la classe i de cardinal n_i.

Le cardinal de l'ensemble des vecteurs $(p_1, p_2, \ldots, p_m)$ de F^m tels que

$$p_i \neq p_j \qquad \text{pour} \qquad i \neq j, \qquad \text{est} \qquad f(f-1) \ldots (f-m+1) \tag{6}\,;$$

en considérant la partition de cet ensemble en les différents sous-ensembles $G_m^{(c)}$ (cf. § II. 2.1), on a l'invariance de $\nu(p_{i_1}, p_{i_2}, \ldots, p_{i_m})$ (cf. formule (5) ci-dessus) pour $(p_{i_1}, p_{i_2}, \ldots, p_{i_m})$ parcourant un même ensemble $G_m^{(c)}$. Désignons par H_m l'ensemble des m-uples de paires dont deux quelconques sont sans composante commune, le cardinal de H_m est

$$\binom{n}{2}\binom{n-2}{2} \cdots \binom{n-2m-2}{2} \tag{7}$$

Le rapport de ce cardinal sur le précédent ((7)/(6)) tend rapidement vers 1 pour n tendant vers l'infini. On a de plus asymptotiquement

$$\text{card}\,(G_m^{(c)})/\text{card}\,(H_m) = O(\alpha/n^r)$$

où r est un entier supérieur à 1 ; r dépend de la configuration c et est directement lié au type $(u_1, u_2, \ldots, u_g)$ de la partition définie par la saturation de l'ensemble des paires d'un m-uples de $G_m^{(c)}$.

Décomposons chacune des deux sommes telles que (2) définissant respectivement $\mathcal{E}(C(b))^l$ et $\mathcal{E}(C(\pi))^l$ en deux parties $\Sigma^{(1)}$ et $\Sigma^{(2)}$ où $\Sigma^{(1)}$ est étendue à H_m et où $\Sigma^{(2)}$ est étendue à l'ensemble des vecteurs $(p_1, p_2, \ldots, p_m)$ pour lesquels il existe au moins deux paires distinctes p_i et p_j ayant une composante commune ; c'est à dire, étendue à la réunion des différents ensembles $G_m^{(c)}$ qui sont disjoints deux à deux. Le nombre de configurations (c) distinctes ne dépend que de m.

Les formules (4) et (5) nous montrent que les parts $\Sigma^{(1)}$ des expressions respectives de $\mathcal{E}(C(b))^r$ et de $\mathcal{E}(C(\pi))^r$ sont sensiblement égales (i.e. convergent rapidement vers la même forme limite). Nous venons de voir par ailleurs que la mesure du support de $\Sigma^{(2)}$ tend à être négligeable par rapport à celle de $\Sigma^{(1)}$. Pour que dans chacune des expression telles que (2) définissant respectivement $\mathcal{E}(C(b))^r$ et $\mathcal{E}(C(\pi))^r$, la valeur de la part de $\Sigma^{(2)}$ tende à être négligeable par rapport à celle de $\Sigma^{(1)}$, il suffit que la charge positive de la forme $(k(p_{i_1})^{l_1} \ldots k(p_{i_m})^{l_m})$ ne soit pas particulièrement forte sur les divers ensembles $G_m^{(c)}$ jusqu'à compenser la faiblesse de leur cardinal. De façon plus précise, soit sur F^m la mesure σ puissance m-ème de la mesure positive $\{k(p)/p \in F\}$;

$$\sigma(p_1, p_2, \ldots, p_m) = k(p_1)\, k(p_2) \ldots k(p_m)\,.$$

Considérons la partie de la somme (2) obtenue pour $(l_1, \ldots l_m)$ fixé ; si pour les différentes configurations (c), $\sigma(G_m^{(c)})/\sigma(H_m)$ tend vers zéro pour n tendant vers l'infini ; la partie de la somme étendue aux divers ensembles $G_m^{(c)}$ tend à être négligeable par rapport à celle étendue à H_m. D'où le théorème suivant

Théorème

Si pour tout m fixé, $m \geqslant 2$ et pour toute configuration (c), $\sigma(G_m^{(c)})/\sigma(H_m)$ tend vers zéro pour n tendant vers l'infini ; alors les moments de la distribution de $c(\pi) = \sum_p \epsilon(p)\, k(p)$ dans $\mathcal{E}(n\,;\,t)$ tendent vers eux de la distribution de $C(b) = \sum_p \beta(p)\, k(p)$ dans $\mathcal{E}(r)$, pour n tendant vers l'infini.

Dans ces conditions, la distribution de $\left(\Sigma\epsilon(p)\ k(p) - \frac{r(f+1)}{2}\right)\Big/\sqrt{\frac{rs(f+1)}{12}}$ dans $\mathcal{R}(n\,;t)$ muni d'une probabilité uniforme est asymptotiquement normale centrée réduite.

1.3.2 Extension des résultats précédents

Nous avons vu que dans l'ensemble $\mathcal{R}(n\,;t)$ des partitions de type fixé, le critère $C(\pi)$ s'exprime, à la constante $-\ r(r+1)/2$ additive près comme la somme des rangs des paires réunies par la partition π ; soit

$$\sum_{p\in R(\pi)} k(p) \qquad \text{où} \qquad R(\pi) \tag{1}$$

est l'ensemble des paires réunies par π.

Dans ces conditions, il semble naturel de considérer directement le critère

$$\sum_{p\in R(\pi)} \mathfrak{S}(p) \tag{2}$$

où $\{\mathfrak{S}(p)/p\in F\}$ est la mesure sur F définie par la similarité sur D.

Les résultats concernant la distribution dans $\mathcal{R}(n\,;t)$ de $\sum_{p\in F} \epsilon(p)\ k(p)$ se transposent à celle de $\sum_{p\in F} \epsilon(p)\ \mathfrak{S}(p)$; tout se passe comme si ; dans le passage de (2) à (1), la mesure attachée à F et reflétant les ressemblances entre éléments de D, était définie par la fonction de répartition de la distribution de $\mathfrak{S}(p)$ sur F. Soient α et λ la moyenne et l'écart type de cette distribution c'est-à-dire ,

$$\alpha = \frac{1}{f}\sum_{p} \mathfrak{S}(p) \qquad \text{et} \qquad \lambda^2 = \frac{1}{f}\sum_{p} (\mathfrak{S}(p) - \alpha)^2 \ .$$

En posant $c(p) = (\mathfrak{S}(p) - \alpha)/\lambda$, *la statistique de proximité* mesurant l'adéquation d'une partition π, correspondante à la formule (1′) qui termine le paragraphe 1.2 ci-dessus, est la suivante

$$\frac{1}{\sqrt{r\,.\,s/(f-1)}} \sum_{p\in F} \epsilon(p)\ c(p) \tag{2′}$$

Dans le cas où la mesure de similarité sur D définissant $\{\mathfrak{s}(p)/p \in F\}$ est statistiquement pertinente, le critère (2′) tient plus étroitement compte de l'information initiale que ne lefait (1′) et peut par conséquent être avantageusement utilisé. Cependant des raisons importantes justifient l'utilisation de (1′) ; la première est fournie par les résultats des travaux de R.N. Shepard et de J.P. Benzecri (cf. [8] et [1]) où on établit le résultat suivant ; Soient $C = \{i_1, i_2, \ldots, i_n\}$ et $C' = \{i'_1, i'_2, \ldots, i'_n\}$ deux configurations géométriques dans un même espace $\mathbf{R}^p$ et soient ω et ω' les ordonnances associées à C et à C' par la distance euclidienne. Si ω' se déduit de ω en remplaçant i_j par $\tau(i_j)$ pour une bijection τ de C sur C' ; alors on peut admettre que C' se déduit de C par un déplacement, une homothétie et une petite déformation. Si π et π' sont des partitions de C et de C' respectivement, telles que les classes de π' se déduisent des classes de π par la bijection τ ; on peut souhaiter d'un critère de classification qu'il juge également π et π'. C'est ce que fait clairement le premier critère par la comparaison de deux structures de même type : préordres totaux sur F. D'autre part ce critère tient compte des propriétés intéressantes de stabilité de ω auquelles nous avons déjà fait allusion, (cf. [4] Chap. 1). Mais, il reste certainement à comparer plus profondément d'un point de vue expérimental et théorique les deux critère.

Nous allons terminer le paragraphe 1.3 par une comparaison plus précise de la *variance* dans $\mathcal{R}(n\,;t)$ de $\sum_p \epsilon(p)c(p)$ et de celle dans $\mathcal{B}(r)$ de $\sum_p \beta(p)c(p)$.

1.3.3. Examen de la variance

Un couple de paire $(p\,,\,p')$ peut prendre l'une des trois formes suivantes : $(\{x\,,\,y\},\{x\,,\,y\})$, $(\{x\,,\,y\},\{x\,,\,z\})$ est $(\{x\,,\,y\},\{z\,,\,t\})$ où des lettres différentes désignent des objets différents. L'ensemble des couples de la première forme est la diagonale de F^2. Désignons par G l'ensemble des couples de la deuxième forme où les deux paires ont une composante commune,

$$\text{card}\,(G) = n(n-1)\,(n-2)\,.$$

Soit enfin H l'ensemble des couples de la troisième forme où les deux paires p et p' n'ont pas de composante commune,

$$\text{card}\,(H) \quad \text{vaut} \quad n(n-1)\,(n-2)\,(n-3)/4\,.$$

L'expression pour $l = 2$ du moment d'ordre l (cf. § 1.3.1) permet d'écrire les relations :

$$\mathcal{V}_\epsilon = \Big(\sum_{1\leqslant i\leqslant k} \pi_i^2\Big)\Big(\sum_{p\in F} c(p)^2\Big) + \Big(\sum_{1\leqslant i\leqslant k} \pi_i^3\Big)\Big(\sum_{G} c(p)\,c(p')\Big) + \Big(\sum_{1\leqslant i\leqslant k} \pi_i^2\Big)^2\Big(\sum_{H} c(p)\,c(p')\Big)$$

$$\mathcal{V}_\beta = \Big(\sum_{1\leqslant i\leqslant k} \pi_i^2\Big)\Big(\sum_{p\in F} c(p)^2\Big) + \Big(\sum_{1\leqslant i\leqslant k} \pi_i^2\Big)^2\Big(\sum_{G} c(p)\,c(p')\Big) + \Big(\sum_{1\leqslant i\leqslant k} \pi_i^2\Big)^2\Big(\sum_{H} c(p)\,c(p')\Big)$$

où $\mathcal{V}_\epsilon$ (resp. $\mathcal{V}_\beta$) est la variance dans $\mathcal{R}(n\,;t)$ (resp. $\mathcal{B}(r)$) de $\sum_p \epsilon(p)\,c(p)$ $\Big($resp. de $\sum_p \beta(p)\,c(p)\Big)$.

La différence entre $\mathcal{V}_\epsilon$ et $\mathcal{V}_\beta$ est

$$\mathcal{V}_\epsilon - \mathcal{V}_\beta = \Big\{\sum_i \pi_i^3 - \Big(\sum_i \pi_i^2\Big)^2\Big\}\Big\{\sum_G c(p)\,c(p')\Big\} \tag{1}$$

Propriété 1.

Les deux variances $\mathcal{V}_\epsilon$ et $\mathcal{V}_\beta$ sont égales si les différentes composantes n_i du type $t = (n_1, n_2, \ldots, n_k)$ sont égales

En effet dans ce cas on a $\pi_i = 1/k$ pour tout $i = 1, \ldots, k$; et

$$\sum_{1\leqslant i\leqslant k} \pi_i^3 = \Big(\sum_{1\leqslant i\leqslant k} \pi_i^2\Big)^2 = \frac{1}{k^2} \quad \text{d'où} \quad \mathcal{V}_\epsilon - \mathcal{V}_\beta = 0$$

Comme dans [6], considérons la formule d'analyse de la variance des proximités $c(x\,,y)$;

$$\frac{1}{n(n-1)} \sum_{\{(x,y)/x\neq y\}} c(x\,,y)^2 = \frac{1}{n}\sum_x \frac{1}{(n-1)} \sum_{\{y/y\neq x\}} \{c(x\,,y) - \bar{c}_x\}^2 + \frac{1}{n}\sum_x \bar{c}_x^2 \tag{2}$$

où $\bar{c}_x = \dfrac{1}{(n-1)} \sum_{\{y/y \neq x\}} c(x, y)$: moyenne des proximités à x

Le premier membre de (2) est égal à 1.

Si $\sum_i \pi_i^3$ est distinct de $\left(\sum_i \pi_i^2\right)^2$, la différence $(\mathcal{V}_\epsilon - \mathcal{V}_\beta)$ est fonction de $\sum_G c(p)\, c(p')$; notons alors la formule

Propriété 2.

$$\frac{1}{(n-1)}\left(1 + \frac{1}{n(n-1)} \sum_G c(p)\, c(p')\right) = \frac{1}{n} \sum_x \bar{c}_x^2 \tag{3}$$

En effet, on se rend compte que dans le développement de l'expression $\sum \left\{ \sum_{\{y/y \neq x\}} c(x, y) \right\}^2$; pour tout p, $(c(p))^2$ apparaît exactement deux fois et pour tout couple de paires (p, p') à composante commune, $c(p)\, c(p')$ apparaît exactement une fois ; par conséquent cette expression se met sous la forme

$$n(n-1) + \sum_G c(p)\, c(q) \qquad \text{car} \qquad \sum_p (c(p))^2 = n(n-1)/2$$

d'où la formule de l'énoncé.

Notons encore l'identité

$$\sum_G c(p)\, c(p') + \sum_H c(p)\, c(p') = \sum_{(F^2 - \Delta)} c(p)\, c(p') = -\, n(n-1)/2$$

où Δ est la diagonale de $F \times F$; en effet

$$\sum_{(F^2 - \Delta)} c(p)\, c(p') = \sum_{p \in F} c(p) \left\{ \sum_{p' \in F} c(p') - c(p) \right\} = -\sum_{p \in F} (c(p))^2$$

Il en résulte la relation

$$1 - \frac{1}{2(n-1)} + \frac{1}{n(n-1)^2} \sum_H c(p)\, c(p')$$

$$= \frac{1}{n} \sum_x \frac{1}{(n-1)} \sum_{\{y/y \neq x\}} \{c(x, y) - c_x\}^2 \tag{4}$$

Il est intéressant de remarquer que nous mesurions le caractère neutre d'un élément x, par rapport à une visée classificatoire, par la petitesse de la quantité

$$\frac{1}{(n-1)} \sum_{\{y/y \neq x\}} \{\mathcal{S}(x, y) - \overline{\mathcal{S}}_x\}^2 \tag{5}$$

où

$\mathcal{S}(x, y)$ est la mesure de similarité définie sur D ; cette quantité est précisément égale à $\frac{\lambda^2}{(n-1)} \sum_{\{y/y \neq x\}} \{\mathcal{S}(x, y) - \overline{\mathcal{S}}_x\}^2$ où λ^2 est la variance de la distribution $\{\mathcal{S}(p)/p \in F\}$.

La formule (4) montre que la somme des dispersions (5) dépend uniquement de $\sum_H c(p)\, c(p')$. La formule (3) peut être écrite sous la forme

$$\frac{1}{n(n-1)} \sum_G c(p)\, c(p') = \frac{1}{n} \sum_x \left\{ \frac{1}{\sqrt{n-1}} \sum_{\{y/y \neq x\}} c(x, y) \right\}^2 - 1 .$$

En désignant par $\Sigma_{(h)}\, c(p)$ une somme portée sur h valeurs de $\{c(p)/p \in F\}$, la distribution de

$$\frac{1}{\sqrt{n-1}} \sum_{(n-1)} c(p) \tag{6}$$

sur l'ensemble de toutes les parties à $(n-1)$ éléments de F, est de moyenne nulle et de variance 1 puisque la distribution $\{c(p)/p \in F\}$ est de moyenne nulle et de variance 1. L'échantillon suivant des valeurs de la statistique (6)

$$\left\{ \frac{1}{\sqrt{n-1}} \sum_{\{y/y \neq x\}} c(x, y)/x \in D \right\},$$

peut être considéré, dans l'hypothèse N d'absence de structure comme un échantillon de n valeurs indépendantes de la statistique (6). Par conséquent, dans l'hypothèse N, la variance empirique de cet échantillon converge presque sûrement vers 1, pour n tendant vers l'infini ; soit

$$P_r^N \left\{ \lim_{n \to \infty} \frac{1}{n} \sum_x \left\{ \frac{1}{\sqrt{n-1}} \sum_{\{y/y \neq x\}} c(x, y) \right\}^2 = 1 \right\} = 1 ,$$

d'où

Propriété 3.

La variance $\mathcal{V}_\epsilon$ est, dans l'hypothèse N, asymptotiquement presque sûrement égale à la variance $\mathcal{V}_\beta$.

Ce qui caractérise les cas réels c'est que la structure de la représentation des données s'écarte sensiblement de ce que peut être une réalisation de l'hypothèse N et la condition de l'énoncé ci-dessus est trop restrictive ; en effet, elle suppose qu'asymptotiquement, presque sûrement

$$\sigma(G)/\sigma(H) = \text{card}(G)/\text{card}(H) = 4/(n-3)$$

où

$$\sigma(G) = \sum_G \mathcal{S}(p)\,\mathcal{S}(p') \qquad \text{et} \qquad \sigma(H) = \sum_H \mathcal{S}(p)\,\mathcal{S}(p')$$

en supposant positive la mesure sur $F, \{\mathcal{S}(p)/p \in F\}$. Or nous avons vu précédemment qu'il suffit seulement que $\sigma(G)/\sigma(H)$ tende vers zéro pour n tendant vers l'infini (i.e. $\sigma(G)/\sigma(H) = \sigma(n)$) pour que $\mathcal{V}_\epsilon$ et $\mathcal{V}_\beta$ admettent la même forme limite. Une telle condition admet la convergence de la représentation des données vers la "classificabilité" (cf. [4] Chap. 4) pourvu que la tendance ne soit pas très forte ; il s'agit donc d'une hypothèse compatible avec la nature des données qui se présentent dans les Sciences Humaines.

2. NŒUDS SIGNIFICATIFS D'UN ARBRE DE CLASSIFICATIONS

Soit K l'ensemble des paires restant séparées à un niveau k de l'arbre des classifications et I l'ensemble des paires de K qu'on s'apprête à réunir en agrégant deux classes. Pour juger de la signification d'une telle agrégation commençons par considérer $\text{card}\{gr(\omega_K) \cap I \times J\}$ où ω_K est la restriction de ω à K et ou J est la partie complémentaire dans K de I, soit l'ensemble des paires laissées séparées au niveau $(k+1)$.

Des considérations analogues à celles du paragraphe 1.2 montrent que la distribution de la statistique ci-dessus envisagée, lorsque ω_K décrit uniformément l'ensemble de tous les ordres totaux sur K ; respectivement, lorsque I décrit de façon uniforme l'ensemble de toutes les parties de K de cardinal $i = \text{card}(I)$, est approximativement normale de *moyenne* $i \times j/2$ et de *variance* $i \times j(i+j+1)/12$, où $j = \text{card}(J)$. Par conséquent, la détection des nœuds les plus pertinents de l'arbre résulte de l'examen de la suite des valeurs de

$$\text{card}\{gr(\omega_K) \cap I \times J\} - i \times j/2\} \Big/ \sqrt{i \times j(i + j + 1)/12}$$

sur la suite des niveaux de l'arbre.

Il s'est avéré expérimentalement que la valeur de cette dernière statistique croît lorsqu'une classe en cours de formation se confirme et décroît devant l'arrêt de constitution d'une classe ayant quelque consistance au profit de celle de l'embryon d'une autre classe.

BIBLIOGRAPHIE

[1] J.P. BENZECRI. – "Analyse factorielle des proximités" I et II, *Publ. de l'Inst. de Stat. de l'Univ. de Paris,* XIII et XIV, 1964-65.

[2] J.P. BENZECRI. – "Théorie de l'Information et Classification d'après un tableau de contingence" cours I.S.U.P. 1968-69 ; Univ. Paris VI.

[3] M.G. KENDALL. – "Rank Correlation Methods", Charles Griffin, fourth edition, 1970.

[4] I.C. LERMAN. – "Les bases de la Classification automatique", Gauthier-Villars, "collection programmation", Paris, 1970.

[5] I.C. LERMAN. – "Analyse des Données préalable à une classification automatique", *Rev. Math. & Sc. Hum.* n° 32, 1970.

[6] I.C. LERMAN. – "Analyse du phénomène de la "sériation" ", *Rev. Math. & Sc. Hum.* n° 38.

[7] S. REGNIER. – "Sur quelques aspects mathématiques des problèmes de la classification automatique" *I.C.C. Bull.* vol. 4 (1965)

[8] R.N. SHEPARD. – "The analysis of proximities : Multidimensional scaling with an unkown distance function", *Psychometrika,* vol. 27, (1962).

[9] W.F. de LA VEGA. – "Techniques de classification automatique utilisant un indice de ressemblance", *Rev. Française de Sociologie,* déc. 1967.

[10] A. WALD & J. WOLFOWITZ. – "Statistical tests based on permutations of the observations", *Ann. Math. Stat.,* Vol 15, (1944). Ce travail développé peut être consulté dans l'ouvrage de D.A.S. Fraser., "Nonparametic Methods in Statistics, John Wiley, Third edition (1963) pp. 235-250.