Mesure de concordance des opinions de deux individus d'un pannel

A two ways measure of opinions condorcance

Farid Beninel et Michel Grun-réhomme

Édition électronique

URL : <http://journals.openedition.org/msh/2822>

DOI : 10.4000/msh.2822

ISSN : 1950-6821

Éditeur

Centre d'analyse et de mathématique sociales de l'EHESS

Édition imprimée

Date de publication : 1 mars 2001

ISSN : 0987-6936

Référence électronique

Farid Beninel et Michel Grun-réhomme, « Mesure de concordance des opinions de deux individus d'un pannel », *Mathématiques et sciences humaines* [En ligne], 153 | Printemps 2001, mis en ligne le 10 février 2006, consulté le 23 juillet 2020. URL : <http://journals.openedition.org/msh/2822> ; DOI : <https://doi.org/10.4000/msh.2822>

MESURE DE CONCORDANCE DES OPINIONS DE DEUX INDIVIDUS D'UN PANEL

Farid BENINEL* et Michel GRUN-REHOMME†

RÉSUMÉ – *Étant donné un panel d'individus interrogés à dates régulières, nous nous intéressons à la mesure a posteriori de l'amplitude de la concordance des opinions exprimées par deux individus quelconques de ce panel. Les individus sont décrits par leurs réponses, aux différentes dates, à une même variable qualitative et la mesure de l'amplitude de la concordance d'opinions est réalisée au moyen d'un indice d'association à deux voies ou indice de concordance. L'indice considéré incrémente en cas d'opinions concordantes à une date donnée le score correspondant à cette concordance. L'élaboration des scores se basant sur l'idée qu'une concordance d'opinions est d'autant plus significative qu'elle se réalise sur une opinion rare. Dans le but de mettre en évidence les concordances significatives, nous associons à cet indice une variable permutacionnelle dont nous étudions la distribution ; il découle de cette modélisation des problèmes d'analyse combinatoire et des outils mathématiques pour les résoudre. Pour illustrer et valider, des simulations sont réalisées à partir des données issues d'une enquête de conjoncture réalisée par l'INSEE.*

MOTS CLÉS – Concordance d'opinions, Enquête de conjoncture, Indice d'association à deux voies, Statistique permutacionnelle, Trajectoire de réponses.

SUMMARY – A two ways measure of opinions concordance
A measure of association quantifies the degree of opinions concordance between two individuals among a panel. This measure deals with qualitative data whose items are possible opinions of the individuals; these possible opinions are the same for the different dates. In order to test the degree of concordance, we associate a permutational statistic and derive its probability distribution. Simulations are given to illustrate and validate this model.

KEYWORDS – Two ways association indices, Opinions concordance, Perception survey of the economic situation, Permutational statistics, Responses trajectories.

Article communiqué par B. Leclerc.

*Université de Poitiers, Département STID-IUT, 8 rue Archimède 79000 Niort, France, e-mail : Farid.Beninel@univ-poitiers.fr.

†Université Paris 2, Centre d'Économétrie, 92 rue d'Assas 75006 Paris, e-mail : grun@u-paris2.fr.

1 INTRODUCTION

Nous nous intéressons ici, pour une date donnée, à la mesure de l'amplitude de la concordance des opinions de deux individus d'un panel.

En France, de nombreuses enquêtes sont réalisées auprès des entreprises, principalement par l'INSEE et la Banque de France ; l'analyse des réponses permettant de cerner la perception de la conjoncture pour le court terme.

Souvent, les questions posées, à la faveur de ces enquêtes, sont qualitatives et concernent l'évolution temporelle de différents facteurs (ventes, prix, investissements, emploi...). Généralement les individus ont le choix entre trois réponses : *favorable (ou niveau supérieur à la moyenne)*, *défavorable (ou inférieur)*, *normal (ou stabilité)*.

Pour des raisons évidentes de confidentialité, seuls des traitements de nature à agréger ces données sont publiés, voir réalisés.

L'approche développée ici porte sur la comparaison par paires de trajectoires de réponses individuelles.

Le recours à des indices s'exprimant en fonction des termes du tableau de contingence associé est quelquefois envisagé ; précisons que pour ce genre de tableau de contingence l'unité décomptée est la paire de réponses.

Les travaux relatifs aux indices associés aux tableaux de contingence remontent à Condorcet, 1785 [8] ; nous trouvons dans [1], [11], [12], [18], [24] une bibliographie concernant ces indices.

Cependant, initialement conçus pour mesurer l'association entre deux variables qualitatives observées sur un ensemble d'individus, à partir du tableau de contingence, ces indices ne permettent pas une comparaison fine de deux trajectoires de réponses. Leur défaut réside dans le fait qu'ils pondèrent de la même façon les différentes dates d'observation.

Ainsi, l'indice de concordance que nous préconisons, affecte à une paire de réponses concordantes un poids d'autant plus élevé que cette réponse est rare dans le panel *i.e.* deux individus seront considérés en accord ou concordant pour le nombre de réponses similaires mais aussi pour l'importance de celles-ci.

L'indice de concordance étant défini, nous lui associons, pour mettre en évidence des concordances ou discordances observées significatives, une variable aléatoire d'univers associée, un univers de paires de permutations ; chaque permutation opérant sur les réponses d'un des deux individus.

L'étude de la distribution de cette variable permutationnelle est envisagée sous l'hypothèse nulle d'indépendance des réponses conditionnellement à la distribution des différents types d'avis (ou différentes modalités réponses) pour chacun des deux individus.

Plus précisément, sous H_0 , l'univers associé à cette variable (un univers de paires de permutations) est muni de la mesure uniforme. Ainsi, considérant la valeur de l'indice de concordance comme la valeur observée de la variable permutationnelle, nous en dégageons un niveau de signification.

Cette démarche remonte à Daniels [1944] et préside à un ensemble de travaux menés en analyse combinatoire et statistique. Parmi ces travaux nous retiendrons ceux de Mantel [1967] dans un contexte de régression, ceux de Le Calvé [1976] et Hubert [1983] s'inscrivant eux dans un contexte de construction de similarités significatives.

Cette approche permutationalle est déclinée de longue date par I. C. Lerman dans le cadre de la *classification basée sur la vraisemblance du lien* et dans des directions différentes ; ces directions sont dictées notamment par la nature des entités à comparer et la nature des données relatives à ces entités. C'est ainsi, par exemple, que Lerman [20] s'intéresse à la comparaison d'objets décrits par des variables de type quelconque tandis que Lerman [19] s'intéresse à la comparaison de classes (ou catégories d'objets).

Notre apport, dans le cadre de l'approche permutationalle, réside dans la prise en compte de la notion de temps ; la variable permutationalle associée à notre modèle d'indice de concordance consistant en une combinaison linéaire de variables de Bernoulli, nous développons des outils combinatoires pour en étudier la distribution.

L'usage de ces outils mathématiques pouvant être étendu à l'étude d'autres statistiques, telles les statistiques linéaires de rang (cf. [7], [26] pour la définition de ces statistiques).

Cet article est organisé comme suit :

Dans la section 2, nous présentons le modèle d'indice de concordance et la statistique permutationalle associée ; la section 3 porte sur l'étude des paramètres caractéristiques de la statistique permutationalle ; la section 4 est consacrée aux simulations.

Conventions et symboles

: cardinal

A_n^j : nombre d'arrangements de j éléments parmi n

$[[\]]$: intervalle d'entiers de $\mathbb{Z}$

$\binom{n}{k_1, \dots, k_m}$: désigne la m -partition

$\{(U_1, \dots, U_m) \in P([1, n])^m : \forall k, l \in [1, r] (k \neq l) \# U_l = k_l, \quad U_l \cap U_k = \emptyset\}$, ainsi que son cardinal.

$I_{[1, p]}$: ensemble des intervalles de $[1, p]$.

2 MODÉLISATION

Soient S un panel de n individus, $[1, p]$ l'intervalle d'entiers indexant les dates d'observation et $[-1, r]$ l'ensemble des réponses possibles.

Pour $k \in [0, r]$, S_k^t est le sous-ensemble d'individus du panel donnant la réponse k à la date t ; S_{-1}^t désignant l'ensemble des individus sans opinion ou absents du panel à cette même date.

La variable étudiée consiste en $X(S \times \llbracket 1, p \rrbracket \rightarrow \llbracket -1, r \rrbracket)$ définie par $\forall(\omega, t) \in S \times \llbracket 1, p \rrbracket$, $X(\omega, t) = k \iff \omega \in S_k^t$.

Posons $X(\omega) = \{X(\omega, t) : t \in \llbracket 1, p \rrbracket\}$ et $K(\omega) = \{t : \omega \in S \setminus S_{-1}^t\}$.

La paire $(X(\omega), K(\omega))$ constitue par définition la trajectoire de l'individu ω . Comme nous le verrons dans l'application, $K(\omega)$ est plus exactement un sous-intervalle de $\llbracket 1, p \rrbracket$; ceci signifie qu'un individu répond sur toutes les dates où il est présent dans le panel et qu'il ne peut intégrer et sortir de celui-ci qu'une fois.

2.1 L'INDICE DE CONCORDANCE

Considérons deux individus distincts $\omega_i, \omega_j \in S$ de trajectoires respectives $(X_i, K_i), (X_j, K_j)$. L'indice de concordance est une application $T(S \times S \rightarrow \mathbb{R})$; son élaboration concrète s'appuie sur l'axiomatique ci-après.

Symétrie: $T(\omega_i, \omega_j) = T(\omega_j, \omega_i)$

Finitude: $\exists m, M \in \mathbb{R} : M \geq T(\omega_i, \omega_j) \geq m$

Maximalité: $T(\omega_i, \omega_i) \geq T(\omega_j, \omega_i)$

2.1.1 Justification des axiomes

L'axiome de *symétrie* se base sur l'idée que l'opinion exprimée par un individu est inconnue des autres individus au moment où ils s'expriment pour la même date; *a priori*, le modèle de concordance est symétrique.

L'axiome de *finitude* permet de quantifier le plus grand désaccord et le plus grand accord possibles entre deux individus distincts.

L'axiome de *maximalité* traduit l'idée que deux individus distincts ne peuvent s'accorder plus qu'un individu avec lui-même.

2.1.2 Le modèle d'indice

Assimilant les individus de S aux trajectoires correspondantes, l'indice de concordance est aussi défini comme l'application $T(\llbracket -1, r \rrbracket^p \times I_{\llbracket 1, p \rrbracket}^2 \rightarrow \mathbb{R})$ associant aux individus $(X_i, K_i), (X_j, K_j)$ la valeur $T(X_i, X_j | K_i, K_j)$; nous convenons que dans le cas où $K_i = K_j$, $T(\cdot | K_i, K_j) = T(\cdot | K_i) = T(\cdot | K_j) = T(\cdot | K(i, j))$ avec $K(i, j) = K_i \cap K_j$.

Pour $l, m \in \llbracket 0, r \rrbracket$, $T_{l,m}(\cdot | K_i, K_j)$ désigne la contribution des occurrences (l, m) (i.e. le 1^{er} individu donne la modalité l comme réponse et le second la modalité m pour une même date) à la valeur de l'indice.

Le modèle d'indice est déterminé par le choix des parties de K_i, K_j que nous considérons, la définition des contributions $T_{l,m}$ et la façon d'agréger ces contributions. Nous nous intéressons au modèle donné par $T(\cdot | K_i, K_j) = \sum_{l,m} T_{l,m}(\cdot | K(i, j))$,

avec

$$T_{l,m}(X_i, X_j | K(i, j)) = \sum_{t \in K(i, j)} a_{l,m}(t) \delta(l, X_i(t)) \delta(m, X_j(t))$$

Ici $\delta(w, x)$ est le coefficient de Kronecker *i.e.* $\delta(w, x) = 1$ ou 0 selon que $w = x$ ou non) et $(a_{l,m}(t))_t$ la séquence de scores associée aux occurrences (l, m) pour les différentes dates.

Par la suite, nous nous restreignons à un choix de scores nuls pour les discordances (*i.e.* $\forall t \quad a_{l,m}(t) = \delta(l, m) a_{l,m}(t)$), correspondant au fait que la variable qualitative étudiée est considérée comme nominale.

Dans le cas où cette variable est considérée comme ordinale, le score $a_{l,m}$ est d'autant plus grand que l est proche de m . À titre illustratif, dans le cas de 3 modalités réponses, l'occurrence $(0, 1)$ ou $(1, 2)$ traduirait un désaccord moindre par rapport à $(0, 2)$; en conséquence, nous affectons aux discordances $(0, 1)$, $(1, 2)$ un score plus grand que celui affecté à $(0, 2)$.

2.2 LA VARIABLE PERMUTATIONNELLE

Les individus ω_i, ω_j étant fixés, nous considérons $T(\omega_i, \omega_j)$ comme la valeur observée parmi un ensemble $\chi(\omega_i, \omega_j)$ de valeurs possibles correspondant aux valeurs de l'indice calculé après *perturbation* des réponses pour chacun des deux individus.

Autrement dit, ces valeurs possibles sont celles d'une variable aléatoire T^* et les façons de les *perturber* définissent l'univers associé.

Plus précisément, les façons de perturber consistent en des paires de permutations qui, appliquées à la paire (X_i, X_j) , conduisent à des *paires de même type*.

G_p étant l'ensemble des permutations de $\llbracket 1, p \rrbracket$ et pour $A \subset \llbracket 1, p \rrbracket$, $G_{p,A} = \{\sigma \in G_p : \forall t \in \bar{A} \quad \sigma(t) = t\}$ est le sous-ensemble de permutations laissant invariant le complémentaire de A .

Les paires de permutations appliquées à (X_i, X_j) sont prises dans $G_{p,K(i,j)}^2$. Appliquées à un vecteur réponse, les permutations de $G_{p,K(i,j)}$ modifient les réponses données sur les dates de $K(i, j)$ et laissent invariantes les réponses données sur les dates du complémentaire $\bar{K}(i, j)$.

$G_{p,K(i,j)}$ est isomorphe à G_q où $q = \#K(i, j)$; ainsi $\#G_{p,K(i,j)} = q!$.

DÉFINITION 2.1 Les paires $(Y_i, Y_j), (Z_i, Z_j) \in \llbracket -1, r \rrbracket^{2p}$ sont du même type si et seulement si $\exists \sigma_1, \sigma_2 \in G_{p,K(i,j)} : \forall t \in K(i, j) \quad Y_i(t) = \sigma_1 \circ Z_i(t), Y_j(t) = \sigma_2 \circ Z_j(t)$.

Cette définition permet de construire la notion de type, à une partie $A \subset \llbracket 1, p \rrbracket$ près, correspondant aux seules dates où les réponses de ω_i, ω_j peuvent être modifiées; ici $A = K(i, j)$.

Ce choix se base sur les justifications suivantes :

Les opinions de deux individus sujets à dépendance mutuelle sont celles exprimées à des dates où les deux individus en question sont présents dans le panel ; d'où l'idée de ne modifier pour ω_i, ω_j que les opinions exprimées sur les dates de $K(i, j)$.

Le choix des permutations comme façon de modifier permet, étant donné l'individu, de maintenir à même niveau la fréquence des différents type d'avis.

Le type de paire de vecteurs réponses défini précédemment fournit une relation d'équivalence $\xi_{K(i,j)}$ sur $[-1, r]^{2p}$. Ainsi l'ensemble des valeurs possibles, parmi lesquels $T(\omega_i, \omega_j)$ est la valeur observée, consiste en la classe d'équivalence :

$$\chi(\omega_i, \omega_j) = \left\{ T(Y_i, Y_j | K_i, K_j) : (Y_i, Y_j) \xi_{K(i,j)}(X_i, X_j) \right\}$$

Notons T^* la variable permutationnelle associée à la paire d'individus (ω_i, ω_j) . Elle est définie par :

$$\forall (\sigma_1, \sigma_2) \in G_{p,K(i,j)}^2 \quad T^*(\sigma_1, \sigma_2) = T(\sigma_1 \circ X_i, \sigma_2 \circ X_j | K(i, j))$$

Définissant T_k^* la variable définie sur le même univers par $T_k^*(\sigma_1, \sigma_2) = T_{k,k}(\sigma_1 \circ X_i, \sigma_2 \circ X_j | K(i, j))$, nous avons $T^* = \sum_k T_k^*$.

À des fins de comparaison avec la valeur observée $T(X_i, X_j | K(i, j))$ (où pour calculer la p -value associée à cette valeur observée), nous nous intéressons dans ce qui suit à la distribution de la variable T^* , sous l'hypothèse H_0 d'indépendance des réponses des individus ω_i, ω_j conditionnellement aux fréquences des différents types d'avis.

Ainsi, étudier la distribution de T^* sous H_0 consiste à munir $G_{p,K(i,j)}^2$ de la mesure uniforme.

3 RÉSULTATS MATHÉMATIQUES

Pour $k \in [0, 2]$, désignons par $u_k(X_i)$ le nombre de fois sur la période $K(i, j)$ où l'individu ω_i donne la réponse k i.e. $u_k(X_i) = \sum_{t \in K(i,j)} \delta(k, X_i(t))$.

PROPOSITION 3.1 *Les paires $(Y_i, Y_j), (Z_i, Z_j) \in [-1, 2]^{2p}$ sont de même type si et seulement si $\forall k \in [0, r] \quad (u_k(Y_i), u_k(Y_j)) = (u_k(Z_i), u_k(Z_j))$.*

La preuve tient du fait que les couples de permutations appartenant à $G_{p,K(i,j)}^2$, appliquées à une paire de $[-1, r]^{2p}$, laissent invariante la distribution des fréquences des différents types de réponse.

En conséquence, la donnée des paires $\{(u_k(i), u_k(j))\}_k = \{(u_k(X_i), u_k(X_j))\}_k$ détermine de façon unique la distribution de T^* .

3.1 LA VARIABLE PERMUTATIONNELLE « NOMBRE DE CONCORDANCE »

Considérons pour $k \in [0, r]$ la variable permutationnelle $N_k (G_{p,K(i,j)}^2 \rightarrow \mathbb{N})$ définie par $N_k(\sigma_1, \sigma_2) = \sum_{t \in K(i,j)} \delta(k, \sigma_1 \circ X_i(t)) \delta(k, \sigma_2 \circ X_j(t))$.

N_k donne pour une permutation des réponses de ω_i et ω_j le nombre de réalisations de l'occurrence (k, k) sur la période de $K(i, j)$.

Remarque. Étant donnés $Z \in \llbracket -1, r \rrbracket^p$, $\sigma \in G_p$ et $t \in \llbracket 1, p \rrbracket$, nous écrirons indifféremment $Z(\sigma(t))$ ou $\sigma \circ Z(t)$ (i.e. $Z(\sigma(t)) = \sigma \circ Z(t)$).

PROPOSITION 3.2 *Pour tout $k \in \llbracket 0, r \rrbracket$, la distribution de N_k sous l'hypothèse H_0 est l'hypergéométrie $H[q, u_k(i), u_k(j)]$.*

Preuve. Rappelons que $q = \#K(i, j)$; les $u_k(i)$ réponses k de l'individu ω_i étant fixées relativement aux dates de $K(i, j)$ (il y a $\binom{q}{u_k(i)}$ possibilités pour ce faire), avoir l concordances (k, k) revient à avoir l succès à l'issue de $u_k(j)$ tirages équiprobables sans remise (il y a $\binom{u_k(i)}{l} \binom{q - u_k(i)}{u_k(j) - l}$ possibilités); le domaine des réalisations de N_k (ou valeurs de l) est $\llbracket m_k, M_k \rrbracket$ avec

$$m_k = \max(0; u_k(i) + u_k(j) - q)$$

et

$$M_k = \min(u_k(i), u_k(j)) \quad \square$$

3.2 DÉNOMBREMENT DES PERMUTATIONS INTÉRESSANTES

Dans ce qui suit, nous présentons un résultat permettant le dénombrement de permutations d'un certain type et utilisable lors du calcul des moments de T^* ; ce résultat est introduit pour la première fois dans [3] pour le cas de 2 modalités réponses et peut être étendu à un nombre quelconque de modalités.

Soient $Z \in \llbracket -1, 2 \rrbracket^p$, $I \subset \llbracket 1, p \rrbracket$ et $E = \{E_m \subset I : m \in \llbracket -1, r \rrbracket, \#E_m \leq u_m(Z)\}$. Posons $G_{p,I}(Z, E) = \{\sigma \in G_{p,I} : \forall k \in \llbracket -1, r \rrbracket, \forall t \in E_k Z(\sigma(t)) = k\}$; il s'agit de l'ensemble des permutations de $G_{p,I}$, qui appliquées à Z , mettent sur les dates de E_k des réponses k .

Convenons de ce que A_n^j désigne le nombre d'arrangements de j éléments parmi n .

PROPOSITION 3.3 $\#G_{p,I}(Z, E) = (\#I - \sum_{m=-1}^r \#E_m) \prod_{m=-1}^r A_{\#E_m}^{\#E_m}$.

Preuve. À chaque permutation σ mettant des réponses k sur E_k ($k \in \llbracket -1, r \rrbracket$) correspondent les ensembles $\sigma^{-1}(E_k)$.

Il y a $(\#I - \sum_{m=-1}^r \#E_m) \prod_{m=-1}^r \#E_m!$ permutations appartenant à $G_{p,I}$ qui pour tout k font correspondre à $\sigma^{-1}(E_k)$ respectivement E_k .

Il y a $\binom{U_m(Z)}{\#E_m}$ possibilités de choix de l'ensemble antécédent de E_m ; par suite

$$\#G_{p,I}(Z, E) = (\#I - \sum_{m=-1}^r \#E_m) \prod_{m=-1}^r (\#E_m! \binom{U_m(Z)}{\#E_m}). \quad \square$$

3.3 PARAMÈTRES CARACTÉRISTIQUES

Nous nous intéressons ici à la détermination de $E_0(T^*)$, $\text{Var}_0(T^*)$, respectivement l'espérance et de la variance de T^* sous l'hypothèse nulle ; il suffit (par l'exploitation de la relation $T^* = \sum_k T_k^*$) d'expliciter $E_0(T_k^*)$ et $E_0(T_k^* T_l^*)$ pour $k, l \in \llbracket 0, r \rrbracket$.

PROPOSITION 3.4 $\forall k \in \llbracket 0, r \rrbracket \quad E_0(T_k^*) = \frac{u_k(i) u_k(j)}{q^2} \sum_{t \in K(i,j)} a_k(t)$.

Preuve. Les permutations σ_1, σ_2 parcourant $G_{p,K(i,j)}$ de cardinal $q!$, nous avons :

$$E_0(T_k^*) = \frac{1}{(q!)^2} \sum_{\sigma_1, \sigma_2} \sum_{t \in K(i,j)} a_k(t) \delta(k, X_i(\sigma_1(t))) \delta(k, X_j(\sigma_2(t)))$$

En intervertissant les sommes, nous obtenons :

$$E_0(T_k^*) = \frac{1}{(q!)^2} \sum_{t \in K(i,j)} a_k(t) \sum_{\sigma_1} \delta(k, X_i(\sigma_1(t))) \sum_{\sigma_2} \delta(k, X_j(\sigma_2(t))) \quad (1)$$

$\sum_{\sigma_1} \delta(k, X_i(\sigma_1(t)))$ est le nombre de permutations de $G_{p,K(i,j)}$ qui appliquées aux réponses de ω_i mettent en position t , une réponse k .

En application de la proposition 3.3, nous avons :

$$\sum_{\sigma_1} \delta(k, X_i(\sigma_1(t))) = u_k(i)(q-1)! \quad (2)$$

De façon analogue, nous avons :

$$\sum_{\sigma_2} \delta(k, X_j(\sigma_2(t))) = u_k(j)(q-1)! \quad (3)$$

Utilisant les équations (2) et (3) et substituant dans (1), nous achevons la preuve. $\square$

PROPOSITION 3.5 $\forall k \in \llbracket 0, r \rrbracket$

$$E_0(T_k^{*2}) = \frac{u_k(i) u_k(j)}{q^2} \left[\sum_{t \in K(i,j)} a_k^2(t) + \frac{(u_k(i)-1)(u_k(j)-1)}{(q-1)^2} \sum_{t,s \in K(i,j), t \neq s} a_k(t) a_k(s) \right]$$

Preuve.

$$E_0(T_k^{*2}) = \frac{1}{(q!)^2} \sum_{\sigma_1, \sigma_2} \sum_{t,s} a_k(t) a_k(s) \delta(k, X_i(\sigma_1(t))) \delta(k, X_j(\sigma_2(t))) \delta(k, X_i(\sigma_1(s)))$$

$\delta(k, X_j(\sigma_2(s)))$ avec σ_1, σ_2 parcourant $G_{p,K(i,j)}$.

En intervertissant les sommes, nous obtenons :

$$E_0(T_k^{*2}) = \frac{1}{(q!)^2} \sum_{t,s} a_k(t) a_k(s) \sum_{\sigma_1} \delta(k, X_i(\sigma_1(t))) \delta(k, X_i(\sigma_1(s)))$$

$$\sum_{\sigma_2} \delta(k, X_j(\sigma_2(t))) \delta(k, X_j(\sigma_2(s)))$$

$\sum_{\sigma_1} \delta(k, X_i(\sigma_1(t))) \delta(k, X_i(\sigma_1(s)))$ est le nombre de permutations de $G_{p,K(i,j)}$ qui, appliquées aux réponses de ω_i , mettent une réponse k sur les positions t, s .

En distinguant le cas $t = s$ du cas $t \neq s$ et en application de la proposition 3.3, nous obtenons :

$$\sum_{\sigma_1} \delta(k, X_i(\sigma_1(t))) \delta(k, X_i(\sigma_1(s))) = \begin{cases} u_k(i)(q-1)! & \text{si } t = s \\ u_k(i)(u_k(i) - 1)(q-2)! & \text{sinon} \end{cases} \quad (4)$$

De la même façon :

$$\sum_{\sigma_2} \delta(k, X_j(\sigma_2(t))) \delta(k, X_j(\sigma_2(s))) = \begin{cases} u_k(j)(q-1)! & \text{si } t = s \\ u_k(j)(u_k(j) - 1)(q-2)! & \text{sinon} \end{cases} \quad (5)$$

nous obtenons le résultat annoncé en substituant dans l'expression de $E_0(T_k^{*2})$ par l'exploitation de (4) et (5). $\square$

PROPOSITION 3.6 $\forall k, l \in \llbracket 0, r \rrbracket : k \neq l$,

$$E_0(T_k^* T_l^*) = \frac{u_k(i)u_l(i)u_k(j)u_l(j)}{q^2} \left[\frac{1}{(q-1)^2} \sum_{t \neq s} a_k(t)a_k(s) \right]$$

Preuve. Posant $\delta(k, w, z) = \delta(k, w)\delta(k, z)$, nous avons :

$$E_0(T_k^* T_l^*) = \frac{1}{(q!)^2} \sum_{\sigma_1, \sigma_2} \sum_{t, s} a_k(t)a_l(s) \delta(k, X_i(\sigma_1(t)), X_j(\sigma_2(t))) \delta(l, X_i(\sigma_1(s)), X_j(\sigma_2(s))),$$

avec σ_1, σ_2 parcourant $G_{p,K(i,j)}$. En intervertissant les sommes, nous avons :

$$E_0(T_k^* T_l^*) = \frac{1}{(q!)^2} \sum_{t, s} a_k(t)a_l(s) \sum_{\sigma_1} \delta(k, X_i(\sigma_1(t))) \delta(l, X_i(\sigma_1(s))) \quad (6)$$

$$\sum_{\sigma_2} \delta(k, X_j(\sigma_2(t))) \delta(l, X_j(\sigma_2(s)))$$

$\sum_{\sigma_1} \delta(k, X_i(\sigma_1(t))) \delta(l, X_i(\sigma_1(s)))$ est le nombre de permutations de $G_{p,K(i,j)}$ qui appliquées aux réponses de ω_i mettent une réponse k sur la position t et une réponse l sur la position s . Ainsi par l'utilisation de la proposition 3.3, nous avons :

$$\sum_{\sigma_1} \delta(k, X_i(\sigma_1(t))) \delta(l, X_i(\sigma_1(s))) = \begin{cases} u_k(i)u_l(i)(q-2)! & \text{si } t \neq s \\ 0 & \text{sinon} \end{cases} \quad (7)$$

De la même façon, nous avons :

$$\sum_{\sigma_2} \delta(k, X_j(\sigma_2(t))) \delta(l, X_j(\sigma_2(s))) = \begin{cases} u_k(j)u_l(j)(q-2)! & \text{si } t \neq s \\ 0 & \text{sinon} \end{cases} \quad (8)$$

En exploitant (7) et (8) et en substituant dans (6), nous avons le résultat. $\square$

3.4 DISTRIBUTION DE PROBABILITÉ

Posons $W_k(t) = \delta(k, X_j(t))\delta(l, X_j(t))$ et notons $\Phi_{q,r}$ la fonction de répartition de T^* .

Posons, pour $r \in \mathbb{N}$, $A = \{a_k(t) \in \mathbb{R} : k \in \llbracket 0, r \rrbracket, t \in \llbracket 1, q \rrbracket\}$, $x \in \mathbb{R}$,

$$N_A(k_0, \dots, k_r | x) = \#\{(U_0, \dots, U_r) \in \binom{q}{k_0, \dots, k_r} : \sum_{k=0}^r \sum_{t \in U_k} a_k(t) \leq x\}$$

Parce qu'obéissant aux conditions (C_1) , (C_2) ci-après, les variables $\{W_k(t)\}_{t,k}$ seront dites de *Bernoulli mutuellement emboîtées* :

$$(C_1) \quad W_k(t) \sim B(p_{k,t}),$$

$$(C_2) \quad W_k(t)[W_l(t) - \delta(k, l)] = 0.$$

Écrivant la statistique T^* comme combinaison linéaire de variables *Bernoulli mutuellement emboîtées* (i.e.) $T^* = \sum_{k=0}^r \sum_{t=1}^q a_k(t) W_k(t)$, nous appréhendons plus facilement sa fonction de répartition sous H_0 . Celle-ci est donnée par :

$$\Phi_{q,r}(x / H_0) = \sum_{k_0, \dots, k_r} P_{H_0}(N_0 = k_0, \dots, N_r = k_r) \frac{N_A(k_0, \dots, k_r | x)}{q \binom{q}{k_0, \dots, k_r}}$$

Le calcul exact de $\Phi_{q,r}(x / H_0)$ se fait en utilisant la proposition 3.2 pour le calcul de $P_{H_0}(N_0 = k_0, \dots, N_r = k_r)$ et un algorithme récursif (du type *partition de nombres entiers* ou *comptage des solutions admissibles dans des problèmes de Knapsack* pour le calcul de $N_A(k_0, \dots, k_r | x)$).

La programmation de cet algorithme est réalisée pour $r = 0, 1$. Le programme est disponible auprès des auteurs ou sous www.mathlabo.univ-poitiers.fr (cf. [4] pour plus de détails sur cet algorithme). Cet algorithme permet les calculs, indépendamment des propriétés intrinsèques des scores A , pour des valeurs de q assez grandes ($q > 50$).

Les problèmes à venir, en matière d'étude de la distribution $\Phi_{q,r}$, consistent en :

pb1 : L'optimisation de l'algorithme de calcul de la distribution exacte i.e. augmenter r , augmenter q .

pb2 : L'étude du comportement asymptotique ($rq \rightarrow +\infty$) de $\Phi_{q,r}$ postulant des hypothèses acceptables quant aux scores A .

4 SIMULATIONS

Nous considérons les données obtenues à l'issue d'une enquête de conjoncture réalisée chaque mois et relative au commerce de détail. Plus précisément cette enquête a été réalisée par l'INSEE et concerne les grandes entreprises du commerce de détail dont le chiffre d'affaires pour l'année 1998 dépasse 2 millions FF (ou 304861 Euros).

Le questionnaire porte sur la situation des entreprises et nous nous intéressons plus particulièrement au volume de ventes.

La question qui nous intéresse est formulée de la façon suivante : *Le volume de ventes réalisé ce mois-ci, est-il pour cette période de l'année supérieur à la normale, normal ou inférieur à la normale ?*

Dans ce travail le nombre de modalités est égal à 3 ; celles-ci sont codées respectivement 0, 1, 2 (dans la saisie originelle elles sont codées respectivement 1, 3, 5 ; la non réponse quant à elle est codée 9).

Nous considérons 10 entreprises qui ont répondu toute l'année et appartenant à 3 secteurs d'activités différents ; à savoir les supermarchés (surface de vente supérieur à 2500 m²), l'habillement et les produits surgelés ; dans l'étude du niveau de concordance par paire d'individus, la donnée de 10 entreprises permet 45 simulations sur données réelles.

Ces 3 secteurs sont notés respectivement M, H, S. Ces 10 entreprises font partie de l'échantillon aléatoire de 50 entreprises extraites du panel ; ces 50 entreprises ont été observées tout au long de l'année. Le tableau ci-après contient les données obtenues.

Tableau 1 : Réponses individuelles

	Jan	Fév	Mar	Avr	Mai	Jun	Jul	Aou	Sep	Oct	Nov	Déc
1M	0	0	0	0	0	1	1	1	1	1	0	0
2M	2	2	2	0	0	0	0	2	1	2	2	1
3M	2	0	0	0	0	0	0	2	2	2	0	0
4M	0	2	0	0	2	2	0	2	0	2	2	1
5C	1	1	1	1	0	1	2	2	1	1	0	1
6C	2	2	2	2	2	2	2	0	0	1	1	2
7C	1	2	1	1	2	2	0	2	1	2	1	1
8DF	0	0	2	0	0	2	1	1	1	2	1	1
9DF	0	0	2	2	0	2	1	2	1	0	2	0
10DF	2	2	0	0	1	1	1	1	1	0	0	1

Tableau 2 : Distribution des réponses sur les 50 entreprises ;

n_{-1}^t est déduit de la relation $n_0^t + n_1^t + n_2^t + n_{-1}^t = 50$.

n_0^t	11	8	13	18	12	15	14	12	12	15	7	11
n_1^t	15	13	11	10	11	13	14	15	18	16	20	18
n_2^t	24	26	26	22	27	22	22	23	20	19	23	21

Les scores sont donnés par $a_k(t) = \frac{N+1}{n_k^t + 0.5n_1^t + 0.5}$ pour $k = 0, 2$ et $a_1(t) = \frac{N+1}{n_1^t + 0.5n_0^t + n_2^t + 0.5}$. Ce choix des scores repose sur les remarques suivantes :

- Le « +1 » au numérateur permet d'obtenir des scores toujours supérieurs à 1.
- Le « +0.5 » au dénominateur permet d'éviter un dénominateur nul.

Ainsi, dans l'ensemble des fonctions homographiques, nous avons une symétrie entre $a_0(t)$ et $a_2(t)$; plus précisément $\frac{1}{a_0(t)} + \frac{1}{a_2(t)} = 1$.

Examinons, à titre d'illustration, les deux situations ci-après.

Situation 1 : $n_0^t = 5$ $n_1^t = 5$ $n_2^t = 10$, nous obtenons alors $a_0(t) = \frac{21}{8}$.

Situation 2 : $n_0^t = 5$ $n_1^t = 10$ $n_2^t = 5$, Nous obtenons alors $a_0(t) = \frac{21}{10.5}$.

La concordance est plus grande dans le premier cas (le nombre d'avis défavorables concordants est plus élevé); la définition des scores tient compte de ces situations. Souvent, dans les enquêtes d'opinion ayant des propositions de réponses ordinales d'ordre 3, la réponse médiane correspond à une façon d'éviter de se prononcer, d'où l'idée d'attribuer un poids plus faible pour ces concordances d'opinion.

Nous donnons dans ce qui suit les tableaux de résultats relatifs aux données des tableaux 1 et 2.

Tableau 3 : Tableau de scores

	a_0	a_1	a_2
Jan	2.6842	1.5454	1.5937
Fév	3.4000	1.5937	1.4166
Mar	2.6842	1.6451	1.5937
Avr	2.1702	1.6721	1.8545
Mai	2.8333	1.6451	1.5454
Juin	2.3181	1.5937	1.7586
Juil	2.3720	1.5692	1.7288
Août	2.5500	1.5454	1.6451
Sep	2.3720	1.4782	1.7288
Oct	2.1702	1.5223	1.8545
Nov	2.9142	1.4366	1.5223
Déc	2.4878	1.4782	1.6721

Tableau 4 : Distribution individuelle des réponses.

	u_0	u_1	u_2
1M	7	5	0
2M	4	2	6
3M	8	0	4
4M	5	1	6
5C	2	8	2
6C	2	2	8
7C	1	6	5
8DF	4	5	3
9DF	5	2	5
10DF	4	6	2

Du tableau 3, il ressort que globalement, l'opinion générale sur la conjoncture est mauvaise en 1998 et nous ne constatons pas de retournement de tendance ; en effet $\min a_0(t) \geq \max[a_1(t), a_2(t)]$.

L'examen du tableau 4 montre que toutes les paires d'individus sont de *types* (cf. définition 1) distincts. Ainsi, la comparaison du niveau de concordance entre paires ne peut pas se faire grâce aux valeurs observées de celles-ci mais plutôt grâce aux *p-valeurs* associées.

L'annexe donne pour chaque paire la valeur de la concordance associée (T_{obs}^*), la moyenne, l'écart-type, l'amplitude, l'infimum et les *p-valeurs* P_N et P_{mc} associées à la valeur observée de la concordance.

P_N^* correspondant à la *p-valeur* calculée selon l'approximation gaussienne de la loi de T^* et P_{ex}^* la *p-valeur* correspondant à une approximation par *Monte Carlo* de la distribution exacte donnée au § 3.4.

Il ressort une détection des concordances significatives par les deux méthodes de calcul de *p-valeur*.

Remerciements : Les auteurs remercient B. Leclerc pour ses encouragements, I. C. Lerman pour ses différentes remarques et son aide à l'élaboration de l'historique portant sur la *statistique permutatonnelle* et les rapporteurs anonymes pour leurs suggestions et conseils.

BIBLIOGRAPHIE

- [1] ANDERBERG, M., *Cluster analysis for applications*, Academic Press, New York, 1973.
- [2] ARABIE, P., HUBERT, L. J., « Combinatorial data analysis », *Annual review of psychology* 43, 1992, p. 169-203.
- [3] BENINEL, F., BRETAGNOLLE, V., « Exact inference for association indices on capture-recapture data », 1999 [soumis].
- [4] BENINEL, F., CIUPERCA, G., « Exact distribution of statistics linear combination of Mutually nested Bernoulli variables », 2000 [soumis].
- [5] BENINEL, F., HUSSON, F., « Calculation for small sequences of the cumulative distribution function for particular family of discrete statistics », *Computational statistics*, 14 (2), 1999, p. 251-261.
- [6] BENINEL, F., GRUN-REHOMME, M., « Un indice de concordance pour l'étude comparative d'opinions sur la conjoncture », *Math. Inf. Sc. hum.*, 142, 1998, p. 17-25.
- [7] CAPÉRAÀ, P., VAN CUTSEM, B., *Méthodes et modèles en statistique non paramétrique*, Exposé fondamental, Presses de l'université Laval, Dunod, 1988.
- [8] CARITAT, M.-J. A., Marquis de Condorcet, *Essais sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*, Paris, reprint, Chelsea publ. 6, New York, 1974.
- [9] CARON, N., RAVALET, P., SAUTORY, O., *Estimation de la précision d'un solde d'opinions dans les enquêtes de conjoncture auprès des entreprises*, INSEE, Coll. Méthodes statistiques, 1996.
- [10] DANIELS, H. E., « The relation between measures of correlation in the universe of sample permutations », *Biometrika*, 33, 1944, p. 129-135.
- [11] GREEN, P. E., RAO, V. R., « A note on proximity measure and cluster analysis », *Journal of Marketing Research* 6, 1969.
- [12] GOODMAN, L. A., KRUSKAL, W. H., *Measures of association for cross classification*, Springer Verlag, Berlin, New York, 1979.
- [13] GRUN-REHOMME, M., LADIRAY, D., « Moyennes mobiles centrées et non centrées : Construction et comparaison », *Revue de Statistique appliquée.*, vol. XLII, 3, 1994, p. 33-61.
- [14] HUBERT, L. J., « Inference procedures for the evaluation and comparison of proximity matrices », *Numerical taxonomy*, Ed. J. Felsenstein, NATO ASI Series, Berlin, Springer Verlag, 1983.
- [15] HUBERT, L. J., *Assignment methods in combinatorial data analysis*, New York, Marcel Decker, 1987.
- [16] INSEE, *La rénovation des enquêtes de conjoncture*, Coll. Méthodes, 32, 1987.

- [17] LE CALVÉ, G., « Un indice de similarité pour des variables de type quelconque », *Statistique et analyse des données*, 01-02, 1976, p. 39-47.
- [18] LERMAN, I. C., « Étude de la notion de ressemblance », *Classification et analyse ordinale des données*, Dunod, 1981.
- [19] LERMAN, I. C., « Analyse classificatoire d'une correspondance multiple, typologie et régression », *Data Analysis and Informatics*, III, E. Diday (éd.), North Holland, 1984, p. 193-252.
- [20] LERMAN, I. C., « Construction d'un indice de similarité entre objets décrits par des variables d'un type quelconque. Application au problème de consensus en classification, *Rev. Stat. Appl.*, 35, 1987, p. 39-60.
- [21] LERMAN, I. C., « Conception et analyse de la forme limite d'une famille de coefficients statistiques d'association entre variables relationnelles I », *Math. Inf. Sci. hum.*, 118, 1992, p. 35-52.
- [22] LERMAN, I. C., « Conception et analyse de la forme limite d'une famille de coefficients statistiques d'association entre variables relationnelles II », *Math. Inf. Sci. hum.*, 119, 1992, p. 75-100.
- [23] MANTEL, N., « Detection of disease clustering and generalized regression approach », *Cancer Research*, 27 (2), 1967, p. 209-220.
- [24] MARCOTORCHINO, F., *Comparaison par paires et critères de contingence*, Étude F-071, Centre scientifique IBM Paris, 1984.
- [25] ROBERTS, G., EVANS, P. R., « A method for the detection of non random associations », *Behavioral Ecology and Sociobiology* 15, 1993, p. 349-354.
- [26] SIEGEL, S., *Non parametric statistics for the behavioral sciences*, McGraw-Hill, New York, Toronto, London, 1956.
- [27] SNIJDERS T. A. B. et al., « Distribution of some similarity coefficients for dyadic binary data in the case of associated attributes », *J. of Classification*, 17 (1), 1990, p. 5-31.

ANNEXE

Tableau de résultats associés à la valeur observée de la concordance

Paire	Valeur de la concordance associée T_{obs}^*	Moyenne	Écart-type	Amplitude	Infimum	<i>p-valeurs</i>	
						p_N^* (approx. gaussienne)	p_{mc}^* (approx. Monte Carlo)
1-2	6.479	7.32	4.390	15.149	0.000	0.573	0.405
1-3	16.489	12.04	1.824	12.894	6.660	0.008	0.002
1-4	7.538	8.17	3.742	16.188	0.000	0.567	0.515
1-5	10.339	8.21	5.634	13.003	1.437	0.355	0.296
1-6	1.522	4.31	3.301	9.632	0.000	0.799	0.877
1-7	1.478	4.75	4.513	11.526	0.000	0.764	0.791
1-8	15.680	9.27	6.360	19.857	0.000	0.159	0.022
1-9	14.453	8.82	4.796	17.782	0.000	0.121	0.041
1-10	13.955	9.92	6.875	19.857	0.000	0.291	0.330
2-3	12.932	10.20	2.632	19.028	0.000	0.152	0.098
2-4	3.001	9.54	3.152	23.974	0.000	0.980	0.973
2-5	5.957	5.46	3.383	16.147	0.000	0.445	0.521
2-6	4.604	8.88	2.657	20.042	2.993	0.946	0.974
2-7	8.828	6.57	2.720	15.616	0.000	0.204	0.250
2-8	11.724	7.23	3.916	20.414	0.000	0.128	0.117
2-9	8.987	8.97	3.475	23.848	0.000	0.500	0.583
2-10	8.137	6.66	4.047	18.385	0.000	0.359	0.301
3-4	9.089	11.92	2.698	19.542	2.170	0.850	0.704
3-5	7.393	4.54	2.024	12.857	0.000	0.079	0.021
3-6	1.594	7.87	2.222	13.511	0.000	0.997	0.958
3-7	5.872	4.49	1.905	10.597	0.000	0.235	0.345
3-8	6.858	8.54	2.457	16.984	0.000	0.648	0.705
3-9	10.366	11.37	2.643	19.542	2.170	0.751	0.776
3-10	9.361	7.99	2.334	15.540	0.000	0.281	0.182
4-5	3.123	4.85	2.984	14.450	0.000	0.719	0.735
4-6	7.090	9.05	2.551	15.618	2.939	0.774	0.815
4-7	12.070	6.00	2.479	13.970	0.000	0.001	0.006
4-8	9.946	7.44	3.512	18.557	0.000	0.239	0.152
4-9	7.610	9.78	3.307	24.987	0.000	0.742	0.785
4-10	7.749	6.44	3.572	17.135	0.000	0.356	0.285
5-6	3.899	5.16	2.662	13.232	2.915	0.680	0.691
5-7	9.464	8.05	2.942	16.653	0.000	0.316	0.256
5-8	5.790	7.75	4.850	16.585	1.437	0.656	0.721
5-9	5.957	5.61	3.641	16.147	0.000	0.504	0.412
5-10	7.435	8.51	4.875	16.653	2.915	0.587	0.596
6-7	6.157	7.51	2.345	15.616	0.000	0.719	0.741
6-8	4.789	6.34	3.090	16.489	1.417	0.691	0.724
6-9	5.042	8.20	2.807	17.113	1.417	0.870	0.902
6-10	3.010	5.50	3.136	16.147	0.000	0.785	0.755
7-8	8.066	6.83	3.355	15.195	0.000	0.364	0.321
7-9	6.360	6.09	2.851	15.616	0.000	0.464	0.390
7-10	4.373	6.92	3.449	16.653	0.000	0.770	0.805
8-9	17.355	7.67	4.202	20.414	0.000	0.010	0.003
8-10	8.241	6.45	7.473	23.540	0.000	0.405	0.305
9-10	3.648	6.20	5.571	18.385	0.000	0.677	0.709