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

F. BENINEL[*] *et M. GRUN-REHOMME*[**]

[*]Université de POITIERS. Département STID-IUT
[**]Université PARIS II. Centre d'Econométrie.

Résumé- ***Etant donné un panel d'individus interrogés à dates régulières, on s'intéresse à la mesure à 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 d'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, on associe à cet indice une variable permutationnelle dont on étudie 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, Phrases : ***Concordance d'opinions ; Enquête de conjoncture; Indice d'association ; Monte-Carlo(méthode de) ; p-value ; Statistique permutationnelle ; Trajectoire de réponses.***

AMS Subject Classification : ***Primary: 62D05, 62G10, 62G99, Secondary : 90A08, 90B50.***

1. Introduction

Dans ce travail, on s'intéresse à la mesure, pour une date donnée, 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 publiées, voire réalisées.

Dans ce travail, l'approche développée 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 associée est quelque fois envisagée; 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] ; on trouve dans [1], [11], [12], [18], [24] une bibliographie concernant ces indices, fort intéressante.

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

Ainsi l'indice de concordance que l'on préconise 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 celle-ci.

L'indice de concordance étant défini, pour mettre en évidence des concordances ou discordances observées significatives, on lui associe une variable aléatoire d'univers associé 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 *Ho*, l'univers associé à cette variable (un univers de paires de permutations) est muni de la mesure uniforme. Ainsi, considérant la valeur observée de la variable permutationnelle, on en dégage 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, on retiendra ceux de MANTEL (1967) dans un contexte de régression, ceux de LE CALVE (1976) et HUBERT (1983) s'inscrivant eux dans un contexte de construction de similarités significatives.

Cette approche permutationnelle 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 notamment dictées par la nature des entités à comparer et la nature des données relatives à ces entités. C'est ainsi, par exemple, que dans [20] LERMAN s'intéresse à la comparaison d'objets décrits par des variables de type quelconque tandis que dans [19] il s'intéresse à la comparaison de classes (ou catégories d'objets).

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

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

Cet article est organisé comme suit :

Dans la **section 2**, on présentera le modèle d'indice de concordance et la statistique permutationnelle associée.

La **section 3** portera sur l'étude de la distribution de probabilité de cette statistique permutationnelle sous ce qu'on considère comme hypothèse nulle.

Enfin, la **section 4** sera consacrée aux simulations.

2. Modélisation

2.1. Conventions

$\overline{A}$: complémentaire de A
: Cardinal
A_n^j : nombre d'arrangement

$$\begin{pmatrix} n \\ k_1,\ldots,k_s \end{pmatrix} =$$

$$\left\{(U_1,\ldots,U_s) \in [1,\, p]^s : \quad \forall k,l \in [1,\, r] \quad (k \neq l), \#U_l = k_l, \quad U_l \cap U_k = \varnothing\right\}$$

et désigne aussi le cardinal de cette *s*-partition.

$[\![\ \ ]\!]$: intervalle de Z .

$Int([\![1,\, p]\!])$: Ensemble des intervalle de $[[1,\, p]]$

Soient S un panel de n individus, $[\![1,\, p]\!]$ l'intervalle d'entiers indexant les dates 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 $X\ (S \times [\![1,\, p]\!] \to [\![1,\, p]\!])$ définie par $\forall(\omega,t) \in S \times [\![1,\, p]\!]$, $X(\omega,t) = k$ si et seulement si $\omega \in S_k^t$.

Posons $X(\omega)=\{X(\omega,t)\colon t \in [\![1,p]\!]\}$ et $K(\omega) = \left\{t :\ \omega \in S / S_{-1}^t\right\}$.

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

2.2. L'indice de concordance

Considérons deux individus distincts ω_i, ω_j de trajectoires respectives (X_i, K_i), (X_j, K_j). L'indice de concordance est une application T ($S \times S \rightarrow IR$) ; son élaboration concrète s'appuie sur l'axiomatique ci-après :

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

Finitude : $\exists m, M \in IR : \quad M \geq T(\omega_i, \omega_j) \geq m$,

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

2.2.1. 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 ; à 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 m*aximalité* traduit l'idée que deux individus distincts ne peuvent s'accorder plus qu'un individu et lui-même.

2.2.2. Le modèle d'indice

Identifiant les individus de S et les trajectoires correspondantes, l'indice de concordance est aussi défini comme l'application $T(\{[-1,r]_p \times Int([\![1,p]\!])\}^2 \rightarrow IR)$

associant aux individus (X_i, K_i) (X_j, K_j) la valeur $T(X_i, X_j / K_i, K_j)$ On convient de noter, dans le cas où $K_i = K_j$, $T(./K_i, K_j) = T(./K_i) = T(./K_{ji}) = T(./K(i,j))$ avec $K(i,j) \subseteq K_i \cap K_j$.

Pour $l, m \in [0, r]$, $T_{l,m}(./K_i, K_j)$ désigne la contribution des occurrences (l, m) (*i.e.* le premier 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 l'on considère, la définition des contributions $T_{l,m}$ et la façon d'agréger ces contributions.

Dans ce travail, on s'intéresse au modèle donné par

$$T(./K_i,K_j)=\sum_{l,m}T_{l,m}(./K_i,K_j),$$

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

Par la suite, on se restreint 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. A titre illustratif, dans le cas de trois modalités réponses, l'occurrence (0,1) ou (1, 2) traduirait un désaccord moindre par rapport à (0, 2) ; en conséquence, on affecte aux discordances (0, 1) , (1, 2) un score plus grand que celui affecté à (0, 2).

2.3. La variable permutationnelle

Les individus ω_i, ω_j étant fixés, on considère $T(\omega_i,\omega_j)$ 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.

Notons G_p l'ensemble des permutations de $[\![1, p]\!]$ et pour $A \subset [\![1, p]\!]$, posons $G_{p,A} = \left\{\sigma \in G_p : \ \forall t \in \overline{A} \ \ \sigma(t) = t\right\}$; il s'agit des permutations laissant invariant le complémentaire de A dans $[\![1, p]\!]$

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

Remarque 2.1. $G_{p,K(i,j)}$ est isomorphe à G_q où $q = \#K(i,j)$; ainsi $\#K(i,j) = q!$.

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

Ainsi défini, le type l'est au choix, d'une partie $A \subseteq [[1, p]]$ 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 sujet à dépendance mutuelle sont celles exprimées à des dates où les deux individus en question étaient 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 la fréquence des différents types d'avis.

Le type de paires 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 :

$$\forall(\sigma_1,\sigma_2) \in G^2_{p,K(i,j)} : \qquad T^*(\sigma_1,\sigma_2) = T(X_i \circ \sigma_1,\, X_j \circ \sigma_2 / K(i,j)).$$

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

A des fins de comparaison avec la valeur observée $T(X_i, X_j / K(i,j))$ (ou pour calculer la *p-value* associée à cette valeur observée), on s'intéresse 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 $G^2_{p,K(i,j)}$ 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.

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

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

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

3.1. La variable permutationnelle nombre de concordances

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

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

Proposition 3.1.1. Pour tout $k \in [\![0, r]\!]$ la distribution de N_k sous l'hypothèse H_0 est l'hypergéométrique $H(q,u_k(i),u_k(j))$.

Preuve : Les $u_k(i)$ réponses k de l'individu ω_i étant fixées relativement aux dates de $K(i,j)$, (il y a $\begin{pmatrix} \#K(i,j) \\ u_k(i) \end{pmatrix}$ possibilités pour ce faire), avoir l concordance (k,k) revient à avoir l succès à l'issue de $u_k(j)$ tirages équiprobables sans remise (il y a $\begin{pmatrix} u_k(i) \\ l \end{pmatrix}\begin{pmatrix} \#K(i,j)-u_k(i) \\ u_k(j)-l \end{pmatrix}$ possibilités) ; le domaine des réalisations de N_k est $[[m_k, M_k]]$ avec $M_k = \min(u_k(i),u_k(j))$ et $m_k = \max(0,q-u_k(i)-u_k(j))$. □

Dénombrement des permutations *intéressantes*

Dans ce qui suit, on présente 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 deux modalités réponses et peut être étendu à un nombre quelconque de modalités.

Soient $Z \in [[-1, 2]]^{2p}$, $I \subseteq [[1, p]]$ et $E = \{E_m \subset I :\ m \in [[-1, r]],\ \#E_m \le u_m(Z)\}$.

Posons $G_{p,I}(Z,E) = \{\sigma \in G_{p,I} :\ \sigma(t) \in E_k \Leftrightarrow Z \circ \sigma(t) = k\}$; 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.

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

Preuve : A chaque permutation σ mettant des réponses k sur E_k $k \in [[-1, r]]$, 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 $\begin{pmatrix} u_m(Z) \\ \#E_m \end{pmatrix}$ possibilités de choix de l'ensemble antécédent de E_m ; par suite

$\#G_{p,I}(Z,E) = (\#I - \sum_{m=-1}^{r} \#E_m) \prod_{m=-1}^{r} (\#E_m! \begin{pmatrix} u_m(Z) \\ \#E_m \end{pmatrix})$. □

3.3. Paramètres caractéristiques

On s'intéresse dans cette section à la détermination de $E_0(T^*)$, $Var_0(T^*)$ 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 $k,l \in [[0, r]]$

Proposition 3.4. $\forall k \in [[0, r]]$, $\quad E_0(T_k^*) = \dfrac{u_k(i)u_k(j)}{q^2} \sum_{t \in K(i,j)} a_k(t)$.

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

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

En intervertissant les sommes on obtient

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

$\sum_{\sigma_1} \delta(k, X_i \circ \sigma_1(t))$ 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, on a

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

De façon analogue, on a

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

Utilisant les équations (2) et (3) et substituant dans (1), on achève la preuve. □

Proposition 3.5. $\forall k \in [[0, r]]$

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

Preuve :

$E_0(T_k^{*2}) = \frac{1}{q!^2} \sum_{\sigma_1,\sigma_2} \sum_{t,s} a_k(t)a_k(s)\delta(k, X_i \circ \sigma_1(t))\delta(k, X_j \circ \sigma_2(t))\delta(k, X_i \circ \sigma_1(s))\delta(k, X_j \circ \sigma_2(s))$ avec σ_1, σ_2 parcourant $G_{p,K(i,j)}$.

En intervertissant les sommes, on obtient

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

$\sum_{\sigma_1} \delta(k, X_i \circ \sigma_1(t))\delta(k, X_i(\sigma_1(s))$ 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, on obtient

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

De la même façon

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

On obtient le résultat annoncé en substituant dans l'expression de $E_0(T_k^{*\,2})$ par l'exploitation (4) et (5). □

Proposition 3.6 . $\forall k,l \in [[0,r]]$, $k \neq l$,

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

Preuve : Posant $\delta(k,\omega,z)=\delta(k,\omega)\delta(k,z)$, on a

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

avec σ_1,σ_2 parcourant $G_{p,K(i,j)}$.

En intervertissant les sommes, on a

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

$\sum_{\sigma_1}\delta(k,X_i\circ\sigma_1(t))\delta(l,X_i(\sigma_1(s))$ 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, on a

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

De la même façon, on a

$$\sum_{\sigma_1}\delta(k,X_i\circ\sigma_1(t))\delta(l,X_i(\sigma_1(s))=\begin{cases}u_k(i)u_l(i)(q-2)! & \textit{si} \quad t\neq s\\ 0 & \textit{sinon}\end{cases} \tag{8}$$

En exploitant (7) et (8) et en substituant dans $E_0(T_l^*T_k^*)$, on a le résultat. □

3.4. Distribution de probabilité

Notons $\Phi_{q,r}$ la fonction de répartition de T^* et posons $W_k(t) = \delta(k, X_i(t))\delta(k, X_j(t))$,
tel que $W_k(t) \sim Bernoulli(p_{k,t})$ et $W_k(t)[W_l(t) - \delta(k,l)] = 0$.

Les variables $\{W_k(t)\}_{t,k}$ sont dites de *Bernoulli mutuellement emboîtées ou en abrégé B.m.e.*

Posons, pour $r \in IN$, $A = \{a_k(t) : k \in [0, r], t \in [1, q]\}$, $x \in IR$,

$$N_A(k_0,...,k_r / x) = \#\left\{(U_0,...,U_r) \in \binom{q}{k_0,...,k_1} : \sum_k \sum_{t \in U_k} a_k(t) \le x\right\}.$$

Ecrivant la statistique T^* comme combinaison de variables *B.m.e i.e.* $T^* = \sum_k \sum_t a_k(t) W_k(t)$, on appréhende plus facilement sa fonction de répartition sous H_0. Celle ci est donnée par

$$\Phi_{q,r}(x / H_0) = \sum_{k_0,...,k_r} P(N_0 = k_0,..., N_r = k_r) \frac{N_A(k_0,...,k_r / x)}{\binom{q}{k_0,...,k_1}},$$

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

Pour r = 0,1, l'algorithme est programmé et est disponible 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 réputées grandes $(q > 50)$.

Pour notre application $r = 2$, en l'absence d'un programme disponible actuellement pour ce cas, on usera d'approximations notamment par *Monte Carlo.*

Les problèmes à venir, en matière d'étude de la distribution $\Phi_{q,r}$, consistant 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 ($rq \to \infty$) de $\Phi_{q,r}$ postulant des hypothèses acceptables quant aux scores A.

4. Simulations

On considère 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 réalisé par l'INSEE et concerne les grandes entreprises du commerce de détail dont le chiffre d'affaires pour l'année 1998 dépasse 2 Millions FRF (ou 304861 Euros).

Le questionnaire porte sur la situation des entreprises et nous nous intéressons plus particulièrement en 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)

On considère 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 2.500 m²), l'habillement et les produits surgelés.

Ces 3 secteurs sont notés respectivement M, C, et DF. Ces 10 entreprises font partie de l'échantillon aléatoire de 50 entreprises, extrait du panel. Ces 50 entreprises ont été observées tout au long de l'année.

Dans l'étude du niveau de concordance par paire d'individus, calculé sur la base des douze mois, la donnée de 10 entreprises permet 45 simulations sur des données réelles.

Le tableau ci-après contient les données obtenues.

TABLEAU 1 : Réponses individuelles

	Jan	Fév	Mar	Avr	Mai	Jun	Jul	Aou	Sep	Oct	Nov	Dec
1 M	0	0	0	0	0	1	1	1	1	1	0	0
2 M	2	2	2	0	0	0	0	2	1	2	2	1
3 M	2	0	0	0	0	0	0	2	2	2	0	0
4 M	0	2	0	0	2	2	0	2	0	2	2	1
5 C	1	1	1	1	0	1	2	2	1	1	0	1
6 C	2	2	2	2	2	2	2	0	0	1	1	2
7 C	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^t_{-1} est déduit de la relation $n^t_0 + n^t_1 + n^t_2 + n^t_{-1} = 50$

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

Les scores sont donnés par $a_k(t) = \dfrac{N+1}{n^t_k + 0.5n^t_0 + 0.5}$ pour $k = 0, 2$ et

$$a_k(t) = \frac{N+1}{n^t_1 + 0.5(n^t_0 + n^t_2) + 0.5}.$$

Pour ces scores, le « +1 » au numérateur permet d'obtenir des scores toujours supérieurs à 1 et le « +0.5 » au dénominateur permet d'éviter un dénominateur nul.

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

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

Situation 1 : $n_0 = 5 \quad n_1 = 5 \quad n_2 = 10$, on obtient alors $a_0 = \dfrac{21}{8}$.

Situation 2 : $n_0 = 5 \quad n_1 = 10 \quad n_2 = 5$, on obtient alors $a_0 = \dfrac{21}{10.5}$.

La concordance est plus grande dans le premier cas (le nombre d'avis défavorables concordant est plus élevé) ; la définition des scores tient compte de ces situations.

Souvent dans les enquêtes d'opinions 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 faible pour ces concordances d'opinions.

TABLEAU 3 : Tableau de scores

	$a_0(t)$	$a_1(t)$	$a_2(t)$
Jan	2.6842	1.5454	1.5937
Feb	3.4000	1.5937	1.4166
Mar	2.6842	1.6451	1.5967
Avr	2.1702	1.6721	1.8545
May	2.8333	1.6451	1.5454
Jun	2.3181	1.5937	1.7586
Jul	2.3720	1.5692	1.7288
Aou	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
Dec	2.4878	1.4782	1.6721

TABLEAU 4 : Distribution individuelle des réponses

	u_0	u_1	u_2
1 M	7	5	0
2 M	4	2	6
3 M	8	0	4
4 M	5	1	6
5 C	2	8	2
6 C	2	2	8
7 C	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 on ne constate 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 *type (cf. définition 1)* distinct. Ainsi, la comparaison du niveau de concordance entre paires ne peut se faire grâce aux valeurs observées de celles ci, mais plutôt grâce aux *p-values* associées.

L'annexe 1 donne pour chaque paire, la concordance associée (T^*_{obs}), la moyenne (*Mean*), l'écart type (*Sd*), l'amplitude (*Range*), L'*inf* et les *p-values* P^*_N, P^*_{mc} associées à la valeur de la concordance.

P^*_N correspondant à la *p-value* calculée selon l'approximation Gaussienne de la loi de T^* et P^*_{mc} la *p-value* correspondant à une approximation par Monte-Carlo de la distribution exacte donnée dans §3.4.

Il ressort une détection simultanée des concordances significatives par les deux méthodes d'approximation.

5. Conclusion

La conception d'indice reste un domaine important, tant les données et les façons de comparer items ou individus sont diverses. L'originalité de notre approche réside dans les outils utilisées pour étudier statistiquement un indice.

Ces outils combinés avec des méthodes de simulations telles la méthode de Monte-Carlo permettent des résultats très fiables.

Des suites à ce travail sont envisagées ; du point de vue de la modélisation, celles ci consisteraient en des variations sur les données (*i.e.* extension de ce travail au cas ordinal...) et au passage des indices à 2 voies aux indices à *n* voies.

Du point de vue des outils mathématiques au service du modèle actuel, des variations peuvent être faites quant au lien de dépendance entre les variables *B.m.e* faisant ressortir davantage l'aspect temporel des données ; aussi, il s'agit d'appréhender la loi limite de la statistique permutationnelle postulant des hypothèses acceptables quant aux scores.

Références

1. Anderberg, M.,1973. Cluster analysis for applications. Academic press, New york.

2. Arabie, P. and Hubert, L.J., 1992. Combinatorial data analysis. *Annual review of psychology, 43, PP. 169-203.*

3. Beninel, F., Bretagnolle, V.,1999. Exact inference for association indices on capture-recapture data. Submitted.

4. Beninel, F., Ciuperca, G., 2000. Exact distribution of statistics linear combination of Mutually nested Bernoulli variables. *Proceedings of COMPSTAT 2000- Utrecht. North Holland. Short communications and posters. pp- 155 –156.*

5. Beninel,F., Husson,F., 1999. Calculation for small sequences of the cumulative distribution function for particular family of discrete statistics. *Computational statistics. Vol 14, n° 2, 251-261.*

6. Beninel,F.,Grun-rehomme,M., 1998. Un indice de concordance pour l'étude comparative d'opinions sur la conjoncture. *Math. Inf. Sc. Hum., Vol. 142, 17-25.*

7. Capéraà, P., Van Cutsem, B., 1988. Méthodes et modèles en statistique non paramétrique. Exposé fondamental. Presses de l'université Laval. *Dunod. 358 pages.*

8. Caritat, Marie-Jean Antoine, Marquis de Condorcet, 1785, 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.,1996. Estimation de la précision d'un solde d'opinions dans les enquêtes de conjoncture auprès des entreprises. *INSEE, Coll. Méthodes statistiques.*

10. Daniels, H.E, 1944. The relation between measures of correlation in the universe of sample permutations. *Biometrika, Vol.33, 129-135.*

11. Green,P.E., Rao,V.R., 1969. A note on proximity measure and cluster analysis. *Journal of Marketing Research.,6.*

12. Goodman,L.A., Kruskal,W.H., 1979. Measures of association for cross classification. *Springer Verlag, Berlin, New York.*

13. Grun-rehomme,M., Ladiray,D., 1994. Moyennes mobiles centrées et non centrées: Construction et comparaison. *Revue de Statistique appliquée., vol. XLII, 3,33-61.*

14. Hubert, L.J., 1983. Inference procedures for the evaluation and comparison of proximity matrices. *Numerical taxonomy, Ed. J. Felsenstein, NATO ASI series, Berlin, Springer Verlag.*

15. Hubert, L.J., 1987. Assignement methods in combinatorial data analysis. *New york, Marcel Decker.*

16. INSEE. La rénovation des enquêtes de conjoncture. *Coll. Méthodes., 32.*

17. Le Calvé, G. 1976. Un indice de similarité pour des variables de type quelconque. *Statistique et analyse des données, Vol. 01-02, PP. 39-47.*

18. Lerman, I.C.,1981. Etude de la notion de ressemblance. *Classification et analyse ordinale des données. Dunod.*

19. Lerman, I.C., 1984. Analyse classificatoire d'une correspondance multiple, typologie et régression. *Data Analysis and Informatics, III, E. Diday. North Holland, 193-252.*

20. Lerman, I.C., 1987. 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., n° 35, pp. 39-60.*

21. Lerman, I.C., 1992. Conception et analyse de la forme limite d'une famille de coefficients statistiques d'association entre variables relationnelles. *Math. Infor.& Sci. Hum., 30ème année, Paris, n°118, pp.35-52.*

22. Lerman, I.C., 1992. Conception et analyse de la forme limite d'une famille de coefficients statistiques d'association entre variables relationnelles. *Math. Infor.& Sci. Hum., 30ème année, Paris, n° 119, pp. 75-100.*

23. Mantel, N., 1967. Detection of disease clustering and generalized regression approach. *Cancer Research, Vol. 27, n° 2, pp. 209-220.*

24. Marcotorchino, F., 1984. Comparaison par paires et critères de contingence. *Etude F-071, Centre scientifique IBM Paris.*

25. Roberts, G., Evans, P.R, 1993. A method for the detection of non random associations. *Behavioral Ecology and Sociobiology. 15. 349-354.*

26. Siegel, S., 1956.Non parametric statistics for the behavioral sciences. *McGraw-Hill. New york, Toronto, London. 312 pages.*

27. Snijders T.A.B.*&al.* 1990. Distribution of some similarity coefficients for dyadic binary data in the case of associated attributes, *J. of Classification, 17, 1, pp.5-31.*

ANNEXE 1

Couple	T^*_{obs}	*Mean*	*Sd*	*Range*	*Inf*	P^*_N	P^*_{MC}
1-2	6.479	7.32	4.390	15.149	0.000	0.573	0.405
-3	16.489	12.04	1.824	12.894	6.660	0.008	0.002
-4	7.538	8.17	3.742	16.188	0.000	0.567	0.515
-5	10.339	8.21	5.634	13.003	1.437	0.355	0.296
-6	1.522	4.31	3.301	9.632	0.000	0.799	0.877
-7	1.478	4.75	4.513	11.526	0.000	0.764	0.791
-8	15.680	9.27	6.360	19.857	0.000	0.159	0.022
-9	14.453	8.82	4.796	17.782	0.000	0.121	0.041
-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
-4	3.001	9.54	3.152	23.974	0.000	0.980	0.973
-5	5.957	5.46	3.383	16.147	0.000	0.445	0.521
-6	4.604	8.88	2.657	20.042	2.993	0.946	0.974
-7	8.828	6.57	2.720	15.616	0.000	0.204	0.250
-8	11.724	7.23	3.916	20.414	0.000	0.128	0.117
-9	8.987	8.97	3.475	23.848	0.000	0.500	0.583
-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
-5	7.393	4.54	2.024	12.857	0.000	0.079	0.021
-6	1.594	7.87	2.222	13.511	0.000	0.997	0.958
-7	5.872	4.49	1.905	10.597	0.000	0.235	0.345
-8	6.858	8.54	2.457	16.984	0.000	0.648	0.705
-9	10.366	11.37	2.643	19.542	2.170	0.751	0.776
-10	9.362	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
-6	7.090	9.05	2.551	15.618	2.939	0.774	0.8015
-7	12.070	6.00	2.479	13.970	0.000	0.001	0.006
-8	9.946	7.44	3.512	18.557	0.000	0.239	0.152
-9	7.610	9.78	3.307	24.987	0.000	0.742	0.785
-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
-7	9.464	8.05	2.942	16.653	0.000	0.316	0.256
-8	5.790	7.75	4.850	16.585	1.437	0.656	0.721
-9	5.957	5.61	3.641	16.147	0.000	0.504	0.412
-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
-8	4.789	6.34	3.090	16.489	1.417	0.691	0.724
-9	5.042	8.20	2.807	17.113	1.417	0.870	0.902
-10	3.010	5.50	3.136	16.147	0.000	0.785	0.755
7-8	8.006	6.83	3.355	15.195	0.000	0.364	0.321
-9	6.360	6.09	2.851	15.616	0.000	0.464	0.390
-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
-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