

# REVUE DE STATISTIQUE APPLIQUÉE

P. POTTIER

**Mesures de la liaison entre deux variables  
qualitatives : relation entre un coefficient de  
corrélation généralisé et le  $\chi^2$**

*Revue de statistique appliquée*, tome 42, n° 1 (1994), p. 41-61

[http://www.numdam.org/item?id=RSA\\_1994\\_\\_42\\_1\\_41\\_0](http://www.numdam.org/item?id=RSA_1994__42_1_41_0)

© Société française de statistique, 1994, tous droits réservés.

L'accès aux archives de la revue « *Revue de statistique appliquée* » (<http://www.sfds.asso.fr/publicat/rsa.htm>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme  
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

**MESURES DE LA LIAISON  
ENTRE DEUX VARIABLES QUALITATIVES :  
RELATION ENTRE UN COEFFICIENT  
DE CORRÉLATION GÉNÉRALISÉ ET LE  $\chi^2$ \***

P. Pottier

*Météo-France/SCEM/PREVI/COMPAS  
42 Avenue Coriolis, 31057 Toulouse Cedex*

**RÉSUMÉ**

Une comparaison théorique de deux critères d'interliaison statistique entre variables qualitatives est effectuée ici, permettant d'établir des résultats nouveaux. Bien que les deux critères considérés aient été construits à partir d'idées différentes, une définition de ces critères à partir d'une même matrice est présentée. Une relation analytique a été établie entre ces deux critères dans le cas particulier où l'une des deux variables ne possède que deux modalités. Dans le cas plus général où le nombre de modalités est quelconque, les deux critères étant exprimés à partir d'une même matrice et considérés comme des fonctions des éléments de cette matrice, on indique le degré de ces fonctions. L'étude théorique a été complétée par différentes simulations.

**Mots-Clés :** *Variables qualitatives, Interliaison statistique, Corrélation,  $\chi^2$ .*

**SUMMARY**

Different criteria can be used to measure the statistical relationship between qualitative variables. A theoretical comparison of two criteria is made, leading to new results. Although they have been made from different ideas, a definition of these criteria from a single matrix is presented. An analytic relation between these criteria is demonstrated in the case where one variable has two modalities. In the more general case where the number of modalities can be greater than 2, these criteria can be written as functions of the elements of the single matrix. The degree of this function is indicated. The theoretical study has been completed by different simulations.

**Key-Words :** *Qualitative variables, Statistical relationship, Correlation,  $\chi^2$ .*

---

\* Article proposé aux Publications de l'ISUP en décembre 1990 et réorienté sur la Revue de Statistique Appliquée en février 1993.

## Introduction

Dans cet article :

- on étudie différents critères de mesure de l'intensité de la liaison entre variables qualitatives  $A$  et  $B$  : le  $\chi^2$  de Pearson, le Tchuprov  $T$  et le coefficient de corrélation généralisé  $Q[A, B]$  proposé par G. Der Megreditchian;

- on présentera une définition du  $\chi^2$ , de  $Q[A, B]$  à partir d'une même matrice  $C$  (la matrice produit des matrices des probabilités conditionnelles de  $A$  par  $B$  et de  $B$  par  $A$ );

- le comportement du  $\chi^2$  et de  $Q[A, B]$  dans certains cas particuliers (indépendance, liaison fonctionnelle réciproque ou non) sera étudié à partir du comportement de la matrice  $C$  dans ces cas;

- on établira alors, dans le cas où l'un des deux caractères ne possède que 2 modalités, une relation directe entre  $\chi^2$  et  $Q[A, B]$ ;

- enfin, des simulations présenteront des résultats empiriques sur la distribution des critères  $\chi^2$  et  $Q[A, B]$ ; elles seront validées par la concordance, au niveau du  $\chi^2$ , entre ces résultats et les résultats théoriques bien connus.

## 1. Quelques critères de mesure de la liaison entre deux variables

### 1.1. Définition

Le coefficient de corrélation entre deux variables quantitatives  $x$  et  $y$  est défini par :

$$r(x, y) = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} \quad (1)$$

- où  $\text{cov}(x, y)$  est la covariance entre les variables aléatoires  $x$  et  $y$ ,
- $\sigma_x$  est l'écart-type de  $x$ ,
- $\sigma_y$  est l'écart-type de  $y$ .

### 1.2. Variables quantitatives

Quand on dispose de deux séries de  $N$  observations des variables aléatoires quantitatives  $\zeta$  et  $\eta$  :

- $(x_1, x_2, \dots, x_N)$  : série d'observations de  $\zeta$ ,
- $(y_1, y_2, \dots, y_N)$  : série d'observations de  $\eta$ ,

le coefficient de corrélation empirique est défini par :

$$\tilde{r}(\zeta, \eta) = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$

$$\text{Avec } \begin{cases} \bar{x} &= \frac{1}{N} \sum_{i=1}^N x_i \\ \bar{y} &= \frac{1}{N} \sum_{i=1}^N y_i \end{cases}$$

### 1.3. Variables qualitatives

Dans le cas où l'on cherche à quantifier l'intensité de la liaison entre deux variables qualitatives, cette formulation n'est plus utilisable.

On peut alors calculer les critères classiques du khi-deux  $\chi^2$  de Pearson (réf 1) et de Tchuprov  $T$  (réf 2), ou encore le coefficient de corrélation généralisé  $Q[A, B]$  proposé par Guy Der Mégréditchian (réf 3).

Nous allons ici nous intéresser à ces critères.

## 2. Notations, définitions et expressions des différents critères étudiés

### 2.1. Notations

On se place dans le cas où le caractère  $A$  possède  $s$  modalités :  $A_1, A_2, \dots, A_s$ ;  $B$  est un caractère à  $k$  modalités :  $B_1, B_2, \dots, B_k$ ,  $s$  et  $k$  sont des entiers vérifiant  $s \geq 2$  et  $k \geq 2$ .

On posera aussi  $\text{Mod } A = s$  et  $\text{Mod } B = k$ , pour rappeler que  $s$  est le nombre de modalités de  $A$  et  $k$  celui de  $B$ .

On suppose que l'on a noté  $A$  le caractère ayant le moins de modalités, c'est-à-dire  $s \leq k$ .

On dispose de  $N$  observations du couple  $(A, B)$ , résumées par la *table de contingence* ci-dessous : représentation des données sous forme d'une table à double entrée; à l'intersection d'une ligne  $i$  et d'une colonne  $j$ , on trouve le nombre de cas où l'on a observé à la fois la modalité  $i$  du caractère  $B$  et la modalité  $j$  du caractère  $A$ .

On note :

- $n_{ij}$  : le nombre d'observations simultanées de la modalité  $i$  de  $B$  et de la modalité  $j$  de  $A$ ,

$B$	$A$	$A_1$	$A_2$	..	$A_s$	Total
$B_1$		$n_{11}$	$n_{12}$	..	$n_{1s}$	$n_1$
$B_2$		$n_{21}$	$n_{22}$	..	$n_{2s}$	$n_2$
..		..	..	..	..	..
$B_i$		$n_{i1}$	$n_{i2}$	..	$n_{is}$	$n_i$
..		..	..	..	..	..
$B_k$		$n_{k1}$	$n_{k2}$	..	$n_{ks}$	$n_k$
Total		$m_1$	$m_2$	..	$m_s$	$N$

(2)

•  $m_i$  : le nombre d'observations de la modalité  $i$  de  $A$ , pour  $i$  variant de 1 à  $s$ , vérifiant :

$$m_i = n_{i1} + n_{i2} + \dots + n_{ji} + \dots + n_{ki}$$

•  $n_i$  : le nombre d'observations de la modalité  $i$  de  $B$ , pour  $i$  variant de 1 à  $k$ , avec :

$$n_i = n_{i1} + n_{i2} + \dots + n_{ij} + \dots + n_{is}$$

•  $N$  : le nombre d'observations du couple  $(A, B)$ , vérifiant :

$$N = \sum_{i=1}^s m_i = \sum_{i=1}^k n_i$$

•  $P(A/B)$  et  $P(B/A)$  les matrices de termes génériques respectifs :

–  $P(A_i/B_j)$  : probabilité d'avoir la modalité  $i$  de  $A$  si on a observé la modalité  $j$  de  $B$

–  $P(B_t/A_u)$  : probabilité d'avoir la modalité  $t$  de  $B$  si on a observé la modalité  $u$  de  $A$

**Expression 1** On définit  $C$  la matrice produit des matrices de probabilités conditionnelles de  $A$  par  $B$  et de  $B$  par  $A$ , par :

$$C = P(A/B).P(B/A)$$

$C$  est la matrice que l'on diagonalise quand on effectue l'analyse des correspondances de la table de contingence (2).

**2.2 Expression de  $C$  : matrice produit des matrices  
de probabilités conditionnelles de  $A$  par  $B$   
et de  $B$  par  $A$**

Avec les notations précisées en (2.1),

- $P(A/B)$  est de dimension  $s \times k$ ;
- $P(B/A)$  est de dimension  $k \times s$ ;
- $C$  est donc de dimension  $s \times s$ ,  $s$  étant la plus petite dimension entre  $s$  et  $k$ .

On estime les probabilités conditionnelles  $P(A_i/B_j)$  et  $P(B_t/A_u)$  par les fréquences empiriques calculées à partir de la table de contingence :

$$P(A_i/B_j) = \frac{n_{ji}}{n_j}$$

$$P(B_t/A_u) = \frac{n_{tu}}{m_u}$$

Le terme générique de la matrice produit  $C$  s'écrit alors, pour  $i = 1$  à  $s$  et  $l = 1$  à  $s$  :

$$c_{il} = \sum_{j=1}^k P(A_i/B_j)P(B_j/A_l) = \sum_{j=1}^k \frac{n_{ji}n_{jl}}{n_j m_l} \quad (3)$$

Compte tenu de la définition des  $c_{il}$  et des relations entre les  $n_{ij}$ ,  $m_i$  et  $n_i$ , on met en évidence les relations suivantes :

$$\begin{aligned} c_{il} &= \sum_{j=1}^k \frac{n_{ji}n_{jl}}{n_j m_l} \times \frac{m_l}{m_i} \times \frac{m_i}{m_l} = \frac{m_i}{m_l} \sum_{j=1}^k \frac{n_{jl}n_{ji}}{n_j m_i} = \frac{m_i}{m_l} c_{li} \\ \sum_{i=1}^s c_{il} &= \sum_{i=1}^s \sum_{j=1}^k \frac{n_{ji}n_{jl}}{n_j m_l} = \sum_{j=1}^k \frac{n_{jl}}{n_j m_l} \left( \sum_{i=1}^s n_{ji} \right) \\ &= \sum_{j=1}^k \frac{n_{jl}n_j}{n_j m_l} = \frac{1}{m_l} \sum_{j=1}^k \frac{n_{jl}n_j}{n_j} = 1. \end{aligned} \quad (4)$$

Nous allons maintenant définir les critères  $\chi^2$  et  $Q[A, B]$  et les exprimer en fonction de la matrice  $C$ .

**2.3 Expression du khi-deux  $\chi^2$**

Pour le calcul du khi-deux ( $\chi^2$ ) et du Tchuprov ( $T$ ), l'idée est de comparer la table de contingence observée à la table que l'on obtiendrait sous l'hypothèse d'indépendance des caractères  $A$  et  $B$ . Le  $\chi^2$  mesure alors une distance entre ces deux tables : s'il est nul, les caractères  $A$  et  $B$  sont indépendants. Cette distance étant

fonction à la fois du nombre  $s$  de modalités de  $A$ , du nombre  $k$  de modalités de  $B$  ainsi que de la taille  $N$  de l'échantillon, on définit le Tchuprov  $\mathcal{T}$ , indépendant de  $N$ ,  $s$  et  $k$  par :

$$\mathcal{T}^2 = \frac{\chi^2}{N\sqrt{(k-1)(s-1)}} \quad (5)$$

Calculons le khi-deux  $\chi^2$  : c'est la somme des contributions au  $\chi^2$  de chaque couple  $(A_i, B_j)$ , cette contribution pouvant être ainsi exprimée à partir des données de la table de contingence (2) :

- le nombre d'observations du couple  $(A_i, B_j)$  est  $n_{ji}$
- s'il y avait indépendance entre les caractères A et B, on observerait  $n'_{ji}$  fois le couple  $(A_i, B_j)$  :

$$n'_{ji} = \frac{n_j m_i}{N}$$

- la contribution du couple  $(A_i, B_j)$  au  $\chi^2$  est :

$$\frac{(n'_{ji} - n_{ji})^2}{n'_{ji}}$$

Le  $\chi^2$  se calcule donc par :

$$\chi^2 = \sum_{i=1}^s \sum_{j=1}^k \frac{(n'_{ji} - n_{ji})^2}{n'_{ji}}$$

Soit encore, en développant le résultat classique :

$$\frac{\chi^2}{N} = \sum_{i,j} \frac{n_{ji}^2}{n_j m_i} - 1,$$

il vient,

$$\frac{\chi^2}{N} = \sum_{i=1}^s \left( \sum_{j=1}^k \frac{n_{ji} n_{ji}}{n_j m_i} \right) - 1.$$

On peut enfin écrire, en utilisant la relation (3) :

## Expression 2

$$\frac{\chi^2}{N} = \sum_{i=1}^s c_{ii} - 1 = \text{Trace}(C) - 1, \quad (6)$$

et pour le Tchuprov (avec 5) :

### Expression 3

$$\mathcal{T} = \frac{\text{Trace}(C) - 1}{\sqrt{(k-1)(s-1)}}. \quad (7)$$

#### 2.4 Définition du coefficient de corrélation généralisé $\mathcal{Q}[A, B]$

Plaçons-nous tout d'abord dans le cas où  $A$  et  $B$  sont deux variables qualitatives ne possédant que deux modalités. On définit simplement le coefficient de corrélation  $\mathcal{Q}[A, B]$  entre les événements  $A$  et  $B$  comme le coefficient de corrélation ordinaire entre les variables indicatrices  $i(A)$  et  $i(B)$ ,  $i(A)$  étant définie par :

$$i(A) = \begin{cases} 1 & \text{si modalité } A \text{ observée} \\ 0 & \text{si modalité } \bar{A} \text{ (non-}A\text{) observée} \end{cases}$$

On peut alors écrire, avec la définition de l'équation (1) :

$$\mathcal{Q}[A, B] = r(i(A), i(B)) = \frac{\text{cov}(i(A), i(B))}{\sigma_{i(A)}\sigma_{i(B)}}$$

$$\text{avec } \begin{cases} E(i(A)) & = P(A) \\ V(i(A)) & = E(i(A)^2) - P(A)^2 = P(A)(1 - P(A)) \\ \text{Cov}(i(A), i(B)) & = E(i(A)i(B)) - P(A)P(B) = P(AB) - P(A)P(B) \end{cases}$$

on obtient :

$$\mathcal{Q}[A, B] = \frac{P(AB) - P(A)P(B)}{\sqrt{P(A)(1 - P(A))P(B)(1 - P(B))}}$$

que l'on peut écrire sous la forme (au signe près) :

$$\mathcal{Q}[A, B] = \sqrt{\text{Det}(P(A/B)P(B/A))} \quad (8)$$

En effet avec,

$$P(A/B) = \begin{pmatrix} P(A/B) & P(A/\bar{B}) \\ P(\bar{A}/B) & P(\bar{A}/\bar{B}) \end{pmatrix}$$

il vient,

$$\begin{aligned} \text{Det}P(A/B) &= P(A/B)P(\bar{A}/\bar{B}) - P(A/\bar{B})P(\bar{A}/B) \\ &= \frac{P(AB)}{P(B)} \frac{P(\bar{A}\bar{B})}{P(\bar{B})} - \frac{P(A\bar{B})}{P(\bar{B})} \frac{P(\bar{A}B)}{P(B)} \end{aligned}$$



$$\text{Soit encore, } \begin{cases} \text{Det}P(A/B) = \frac{P(AB)P(\bar{A}\bar{B}) - P(A\bar{B})P(\bar{A}B)}{P(B)P(\bar{B})} \\ \text{Det}P(B/A) = \frac{P(AB)P(\bar{A}\bar{B}) - P(A\bar{B})P(\bar{A}B)}{P(A)P(\bar{A})} \end{cases}$$

On obtient donc :

$$\text{Det}(P(A/B)P(B/A)) = \frac{[P(AB)P(\bar{A}\bar{B}) - P(A\bar{B})P(\bar{A}B)]^2}{P(A)P(\bar{A})P(B)P(\bar{B})} \quad (9)$$

D'autre part, on peut écrire :

$$P(AB) - P(A)P(B) = P(AB) - [P(AB) + P(A\bar{B})][P(AB) + P(\bar{A}B)]$$

Soit,

$$P(AB) - P(A)P(B) = P(AB)[1 - P(AB) - P(A\bar{B}) - P(\bar{A}B)] - P(A\bar{B})P(\bar{A}B)$$

Et enfin,

$$P(AB) - P(A)P(B) = P(AB)P(\bar{A}\bar{B}) - P(A\bar{B})P(\bar{A}B) \quad (10)$$

De (9) et (10) on déduit bien la relation (8).

Sous cette formulation (8), on peut aisément étendre la définition du coefficient de corrélation généralisé à des caractères  $A$  et  $B$  possédant plus de deux modalités. Le coefficient de corrélation généralisé  $Q[A, B]$  est toujours défini par la relation (8), quel que soit le nombre de modalités de  $A$  et  $B$ .

En utilisant la matrice  $C$  définie par la relation (3), on obtient l'expression suivante :

#### Expression 4

$$Q[A, B]^2 = \text{Det}(C) \quad (11)$$

### 3. Relations entre ces différents critères

#### 3.1 $\chi^2$ , $Q[A, B]$ et matrice $C$

La matrice produit des probabilités conditionnelles  $C$  est une matrice de taille  $s \times s$ . Sa trace est une fonction d'ordre 1 des  $c_{ij}$ , c'est-à-dire d'ordre 2 en  $n_{ij}$ . Son déterminant est une fonction d'ordre inférieur ou égal à  $s$  des  $c_{ij}$ , c'est-à-dire d'ordre  $2s$  en  $n_{ij}$ .

Nous avons vu (6, 7 et 11) que les critères  $\chi^2$ ,  $\mathcal{T}$  et  $\mathcal{Q}[A, B]$  pouvaient être définis à partir de cette matrice  $C$ .

*A priori,*

$$\left\{ \begin{array}{l} \chi^2 : \text{fonction de degrés} = 1 \text{ en } c_{ij} \text{ et } = 2 \text{ en } n_{ij} \\ \mathcal{Q}[A, B]^2 : \text{fonction de degrés} \leq s \text{ en } c_{ij} \text{ et } \leq 2s \text{ en } n_{ij} \\ \mathcal{T}^2 : \text{fonction de degrés} = 1 \text{ en } c_{ij} \text{ et } = 2 \text{ en } n_{ij} \end{array} \right.$$

Nous allons nous intéresser au comportement de  $C$  dans certains cas particuliers ce qui va nous permettre de retrouver ou de mettre en évidence des résultats sur les critères  $\chi^2$ ,  $\mathcal{T}$  et  $\mathcal{Q}[A, B]$ .

### 3.2 Cas particulier : indépendance

Dans ce paragraphe nous supposons que  $A$  et  $B$  sont indépendants.

On peut alors écrire :

$$\forall i = 1, s, \forall j = 1, k \quad P(A_i/B_j) = P(A_i) \text{ et } P(B_j/A_i) = P(B_j).$$

Dans l'expression de  $c_{ij}$ , il vient :

$$c_{ij} = \sum_{l=1}^k P(A_i/B_l)P(B_l/A_j) = \sum_{l=1}^k P(A_i)P(B_l) = P(A_i) \sum_{l=1}^k P(B_l) = P(A_i)$$

Toutes les colonnes de la matrice  $C$  sont donc égales et on obtient l'expression suivante de  $C$  dans le cas de l'indépendance :

$$C = \begin{pmatrix} P(A_1) & P(A_1) & \dots & P(A_1) \\ P(A_2) & P(A_2) & \dots & P(A_2) \\ \dots & \dots & \dots & \dots \\ P(A_s) & P(A_s) & \dots & P(A_s) \end{pmatrix}$$

$$\text{Avec cette expression de } C, \left\{ \begin{array}{l} \text{Trace}(C) = \sum_{i=1}^s P(A_i) = 1 \\ \text{Det}(C) = 0. \end{array} \right.$$

On obtient donc, dans le cas de l'indépendance :

#### Résultat 1 (Indépendance)

$$\left\{ \begin{array}{l} \mathcal{Q}[A, B] = 0 \\ \chi^2 = 0 \\ \mathcal{T} = 0 \end{array} \right.$$

Le  $\chi^2$  mesurant une distance entre la table de contingence observée et celle que l'on aurait si les variables étaient indépendantes, on retrouve naturellement  $\chi^2 = 0$  dans le cas particulier où  $A$  et  $B$  sont indépendants.

Pour  $Q[A, B]$ , *corrélation généralisée*, on a bien une corrélation nulle dans le cas de l'indépendance.

### 3.3 Cas particulier : liaison fonctionnelle réciproque

Nous considérons maintenant que  $A$  et  $B$  sont en liaison fonctionnelle réciproque. On supposera de plus que les modalités de  $A$  et les modalités de  $B$  ont été numérotées de façon à ce que  $P(A_i/B_j) = \delta_{ij}$ , pour  $1 \leq i \leq s$  et  $1 \leq j \leq s$ . Dans le cas de la liaison fonctionnelle réciproque, on a obligatoirement  $s = k$ .

On obtient alors pour  $1 \leq i, j \leq s$  :

$$c_{ij} = \sum_{l=1}^k P(A_i/B_l)P(B_l/A_j) = \delta_{ij}$$

La matrice  $C$  est donc la matrice identité d'ordre  $s$  et vérifie

$$\begin{cases} \text{Trace}(C) &= s \\ \text{Det}(C) &= 1. \end{cases}$$

On obtient donc, dans le cas de la liaison fonctionnelle réciproque :

**Résultat 2 (Liaison fonctionnelle réciproque; Mod  $A = \text{Mod } B = s$ )**

$$\begin{cases} Q[A, B]^2 &= 1 \\ \chi^2/N &= s - 1 \\ T^2 &= 1 \end{cases}$$

Le module de la *corrélation* généralisée  $Q[A, B]$  est bien égal à 1 dans le cas de la liaison fonctionnelle réciproque.

Pour le  $\chi^2$  et le module de  $T$ , on retrouve des valeurs connues de respectivement  $N[s - 1]$  et 1, dans le cas de la liaison fonctionnelle réciproque.

### 3.4 Cas particulier : liaison fonctionnelle non réciproque

Nous considérons maintenant que  $A$  et  $B$  sont en liaison fonctionnelle non réciproque. Dans ce cas, on a  $s < k$  (la connaissance de la modalité  $j$  de  $B$  permet de connaître exactement la modalité  $i$  de  $A$ ; la réciproque n'étant pas vérifiée,  $A$  ayant moins de modalités que  $B$ ).

On notera  $\mathcal{B}(i)$  l'ensemble des indices  $j$  des classes de la variable  $B$  pour lesquelles on a :  $A_i \cap B_j \neq \emptyset$ ;  $\mathcal{B}(i)$  est défini pour  $1 \leq i \leq s$ .

Dans le cas de la liaison fonctionnelle non réciproque,  $\{\mathcal{B}(i), i = 1, s\}$  réalise une partition non triviale de  $\{1, 2, \dots, k\}$ .

On obtient alors, avec les notations précisées en (2.1) :

- $n_{ji} = 0$  si  $j \notin \mathcal{B}(i)$ ;
- $n_j = \sum_{i=1}^s n_{ji} = n_{ji_j}$  avec  $i_j$  défini par  $j \in \mathcal{B}(i_j)$  existe et est unique;
- $m_i = \sum_{j=1}^k n_{ji} = \sum_{j \in \mathcal{B}(i)} n_{ji}$ .

On peut alors calculer :

$$P(A_i/B_l) = \frac{n_{li}}{n_l} = \begin{cases} 0 & \text{si } l \notin \mathcal{B}(i) \\ \frac{n_{li}}{n_{li}} = 1 & \text{si } l \in \mathcal{B}(i) \end{cases}$$

De même,

$$P(B_l/A_j) = \frac{n_{lj}}{m_j} = \begin{cases} 0 & \text{si } l \notin \mathcal{B}(j) \\ \frac{n_{lj}}{\sum_{t \in \mathcal{B}(j)} n_{tj}} & \text{si } l \in \mathcal{B}(j) \end{cases}$$

Le produit  $P(A_i/B_l)P(B_l/A_j)$  est donc nul si  $l$  n'appartient pas simultanément à  $\mathcal{B}(i)$  et  $\mathcal{B}(j)$ , ce qui n'est réalisable que si  $i = j$ . On obtient donc  $c_{ij} = 0$  si  $i \neq j$ .  $C$  est donc une matrice diagonale d'ordre  $s$ , de terme générique  $c_{jj}$  :

$$\begin{aligned} c_{jj} &= \sum_{l=1}^k P(A_j/B_l)P(B_l/A_j) = \sum_{l \in \mathcal{B}(j)} P(A_j/B_l)P(B_l/A_j) \\ &= \sum_{l \in \mathcal{B}(j)} 1 \cdot \frac{n_{lj}}{\sum_{t \in \mathcal{B}(j)} n_{tj}} = 1 \end{aligned}$$

La matrice  $C$  est donc la matrice identité d'ordre  $s$  et vérifie

$$\begin{cases} \text{Trace}(C) &= s \\ \text{Det}(C) &= 1. \end{cases}$$

On obtient donc, dans le cas de la liaison fonctionnelle non réciproque :

**Résultat 3 (Liaison fonctionnelle non réciproque; Mod  $A = s \prec$  Mod  $B = k$ )**

$$\begin{cases} Q[A, B]^2 &= 1 \\ \chi^2/N &= s - 1 \\ T^2 &= (s - 1)/(k - 1) \end{cases}$$

Le module de la *corrélation* généralisée  $Q[A, B]$  reste égal à 1 dans le cas de la liaison fonctionnelle non réciproque.

Pour le  $\chi^2$  et le carré de  $\mathcal{T}$ , on retrouve des valeurs connues de respectivement  $N[s - 1]$  et  $(s - 1)/(k - 1)$ , dans le cas de la liaison fonctionnelle non réciproque.

La liaison fonctionnelle réciproque n'est qu'un cas particulier de la liaison fonctionnelle non réciproque; notamment, le résultat 2 peut être obtenu à partir du résultat 3 en posant  $k = s$ ; la démonstration est alors la même, chaque  $\mathcal{B}(i)$  étant réduit à un singleton.

### 3.5 Cas où $A$ ne possède que 2 modalités

Dans ce paragraphe, on considère que  $A$  possède 2 modalités ( $s = 2$ ); les matrices à calculer sont donc de taille :

- $2 \times k$  pour la matrice  $P(A/B)$ ;
- $k \times 2$  pour la matrice  $P(B/A)$ ;
- carrée d'ordre 2 pour la matrice produit  $C = P(A/B)P(B/A)$ .

Le déterminant de  $C$ , matrice carrée d'ordre 2, se calcule simplement par :

$$\text{Det}C = c_{11}c_{22} - c_{12}c_{21}$$

On obtient donc les expressions suivantes de  $Q[A, B]^2$ , en utilisant la relation (4) du paragraphe 2.2 avec  $s = 2$  ( $c_{1l} + c_{2l} = 1$ ) :

$$Q[A, B]^2 = c_{11}c_{22} - c_{12}c_{21} = c_{11}c_{22} - (1 - c_{11})(1 - c_{22}) = c_{11} + c_{22} - 1$$

On reconnaît ici  $\text{Trace}(C) = c_{11} + c_{22}$  et finalement :

**Résultat 4** [ $A$  à 2 modalités :  $\text{Mod } A = s = 2, \text{Mod } B = k$ ]

$$Q[A, B]^2 = \text{Det}(C) = \text{Trace}(C) - 1 = \frac{\chi^2}{N} = \sqrt{k - 1} \cdot \mathcal{T}. \quad (12)$$

Ainsi, dans le cas où  $A$  possède 2 modalités et  $B$  un nombre quelconque  $k$  de modalités, une relation directe (égalité) a été établie entre  $Q[A, B]^2$  et  $\chi^2/N$ .

### 3.6 Cas où $A$ possède plus de 2 modalités

Considérons tout d'abord que  $A$  possède 3 modalités ( $s = 3$ ).

On doit donc calculer :

$$Q[A, B]^2 = \begin{vmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{vmatrix}$$

En remplaçant la première ligne par la somme des trois lignes du déterminant, en utilisant alors la relation (4) du paragraphe 2.2 et en soustrayant la première colonne aux deux autres, le calcul du déterminant devient :

$$Q[A, B]^2 = \begin{vmatrix} \sum_{i=1}^3 c_{i1} & \sum_{i=1}^3 c_{i2} & \sum_{i=1}^3 c_{i3} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{vmatrix} = \begin{vmatrix} 1 & 1 & 1 \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{vmatrix} = \begin{vmatrix} 1 & 0 & 0 \\ c_{21} & c_{22} - c_{21} & c_{23} - c_{21} \\ c_{31} & c_{32} - c_{31} & c_{33} - c_{31} \end{vmatrix}$$

$$Q[A, B]^2 = \begin{vmatrix} c_{22} - c_{21} & c_{23} - c_{21} \\ c_{32} - c_{31} & c_{33} - c_{31} \end{vmatrix}$$

Plus généralement, pour  $s$  quelconque, un calcul similaire montre que :

$$\begin{vmatrix} c_{11} & c_{12} & \dots & c_{1s} \\ c_{21} & c_{22} & \dots & c_{2s} \\ \dots & \dots & \dots & \dots \\ c_{s1} & c_{s2} & \dots & c_{ss} \end{vmatrix} = \begin{vmatrix} 1 & 1 & \dots & 1 \\ c_{21} & c_{22} & \dots & c_{2s} \\ \dots & \dots & \dots & \dots \\ c_{s1} & c_{s2} & \dots & c_{ss} \end{vmatrix} = \begin{vmatrix} c_{22} - c_{21} & \dots & c_{2s} - c_{21} \\ \dots & \dots & \dots \\ c_{s2} - c_{s1} & \dots & c_{ss} - c_{s1} \end{vmatrix}$$

On montre ainsi que le calcul de  $Det(C)$ , a priori fonction d'ordre  $s$  en  $c_{ij}$  se ramène à un déterminant d'ordre  $s - 1$ .  $Q[A, B]^2$  est une fonction de degrés  $s - 1$  en  $c_{ij}$ .

Le  $\chi^2$  est toujours d'ordre 1 en  $c_{ij}$ .

**Résultat 5** [Mod  $A = s \preceq$  Mod  $B = k$ ]

$$\left\{ \begin{array}{l} \chi^2 : \text{fonction de degrés} = 1 \text{ en } c_{ij} \text{ et } = 2 \text{ en } n_{ij} \\ Q[A, B]^2 : \text{fonction de degrés} \leq s - 1 \text{ en } c_{ij} \text{ et } \leq 2(s - 1) \text{ en } n_{ij} \\ T^2 : \text{fonction de degrés} = 1 \text{ en } c_{ij} \text{ et } = 2 \text{ en } n_{ij} \end{array} \right.$$

Dans le cas  $s = 2$ ,  $\chi^2$  et  $Q[A, B]^2$  sont des fonctions d'ordre 1 en  $c_{ij}$  et on a pu mettre en évidence une relation de proportionnalité entre ces deux critères.

Ces deux résultats nouveaux (Résultats 4 et 5) permettent de mieux appréhender le comportement des deux critères étudiés.

#### 4. Lois des critères $\chi^2$ et $Q[A, B]$

Pour le  $\chi^2$ , la loi est bien connue.

Pour une corrélation en général, la loi est bien connue dès lors que l'on se place dans le cadre gaussien. Pour la corrélation généralisée  $Q[A, B]$  introduite, dans le cas où les caractères  $A$  et  $B$  n'ont que 2 modalités, comme la corrélation ordinaire entre les variables indicatrices  $i(A)$  et  $i(B)$ , il n'est certes pas envisageable d'appliquer l'hypothèse gaussienne aux variables indicatrices.

Nous pouvons cependant étudier le comportement de  $Q[A, B]$  par simulation. Nous étudierons de même le comportement du  $\chi^2$ , ce qui permettra de valider les simulations réalisées.

##### 4.1 Loi du $\chi^2$

Sous l'hypothèse d'indépendance des caractères  $A$  ( $s$  modalités) et  $B$  ( $k$  modalités), le  $\chi^2$  suit une loi du khi-deux à  $\nu = (s - 1)(k - 1)$  degrés de liberté.

L'application du test du  $\chi^2$  impose :

$$\begin{cases} n_{ij} & \geq 5, \forall i, j \\ \frac{n_{ij}m_j}{N} & \geq 5, \forall i, j \\ N & \geq 50 \end{cases}$$

##### 4.2 Simulations de variables qualitatives

Les procédures de la librairie IMSL ont été utilisées pour simuler des échantillons de variables aléatoires à un nombre de modalités et des probabilités marginales données.

Le calcul des critères  $Q[A, B]$  et  $\chi^2$  a été effectué à partir des échantillons simulés.

Nous avons simulé  $N_s = 1000$  couples  $(A, B)$  de variables à respectivement 2 et 2, 2 et 3, 2 et 5, 3 et 3, 3 et 5 modalités, de probabilités marginales différentes pour chaque variable. Chaque échantillon simulé était de taille  $N = 1000$ .

On détermine aléatoirement les probabilités marginales de chacune de ces variables.

###### 4.2.1 Variables à 2 modalités

On tire  $p_1 \in U[0, 1]$  (uniforme sur  $[0, 1]$ ), et on pose  $p_2 = 1 - p_1$ ; on simule alors un échantillon de 1000 observations d'une variable  $A$  à 2 modalités de probabilités marginales  $\begin{cases} p(A_1) = p_1, \\ p(A_2) = p_2. \end{cases}$

On note  $f_1$  et  $f_2$  les fréquences empiriques calculées sur chaque échantillon simulé.

La comparaison entre les  $f_1$  et  $f_2$  calculés et les  $p_1$  et  $p_2$  demandés indique une bonne adéquation des simulations réalisées.

#### 4.2.2 Variables à 3 modalités

on tire  $p_1 \in U[0, 1]$ , on transforme  $p_1$  en  $\frac{2}{3}p_1$ ; de même pour  $p_2$ . Si  $p_1 + p_2 \geq 1$ , on recommence le tirage de  $p_1$  et  $p_2$ ; sinon, on pose  $p_3 = 1 - p_1 - p_2$ . On simule alors un échantillon de 1000 observations d'une variable  $A$  à 3 modalités de probabilités

$$\text{marginales } \begin{cases} p(A_1) = p_1, \\ p(A_2) = p_2, \\ p(A_3) = p_3. \end{cases}$$

Comme précédemment, les fréquences empiriques, calculées sur chaque échantillon simulé, sont «proches» des probabilités marginales demandées.

#### 4.2.3 Variables à 5 modalités

On procède comme dans les deux cas précédents, en tirant ici  $p_1, p_2, p_3, p_4 \in U[0, 1]$ , puis en considérant  $p_5 = 1 - 2/5(p_1 + p_2 + p_3 + p_4)$ . Si  $p_5$  ainsi défini vérifie  $p_5 > 0$ , on simule alors un échantillon de 1000 observations d'une variable  $A$  à 5 modalités de probabilités marginales  $(2/5p_1, 2/5p_2, 2/5p_3, 2/5p_4, 1 - 2/5(p_1 + p_2 + p_3 + p_4))$ .

### 4.3 Comparaison des valeurs du $\chi^2$ obtenues empiriquement à la distribution d'un $\chi^2$ théorique

Pour (Mod  $A$ , Mod  $B$ ) fixé, pour chacune des  $N_s$  simulations d'un échantillon de 1000 couples  $(A, B)$ , on calcule le  $\chi^2$ . A l'aide du logiciel STATGRAPHICS, on ajuste l'histogramme des  $N_s = 1000$  valeurs du  $\chi^2$  obtenues, par un  $\chi^2$  dont on détermine le nombre de degrés de liberté. On peut aussi réaliser l'ajustement du  $\chi^2$  aux valeurs obtenues pour les  $N_e$  ( $N_e \leq N_s$ ) couples  $(A, B)$  pour lesquels la table de contingence respecte la condition  $n_{ij} > 5$  (fichier extrait).

• Le tableau ci-dessous résume, pour chaque couple (Mod  $A$ , Mod  $B$ ) étudié, le nombre de degrés de liberté obtenus lors de l'ajustement des valeurs du  $\chi^2$  obtenues empiriquement par un  $\chi^2$ , à la fois sur le fichier complet ( $C$ ) de taille  $N_s$  et sur le fichier extrait ( $E$ ) de taille  $N_e$ ; ce nombre peut être comparé au résultat théorique attendu (ddl theor).

Mod $A$	Mod $B$	ddl theor	$N_s$	ddl ajust $C$	$N_e$	ddl ajust $E$
2	2	1	1000	1,023	918	1,008
2	3	2	1000	1,995	876	1,999
2	5	4	1000	3,858	726	3,728
3	3	4	1000	3,938	799	3,966
3	5	8	1000	8,052	660	8,030

• Nous présentons (figures 1 et 2) les histogrammes des valeurs du  $\chi^2$  obtenues avec respectivement (2,3) et (3,5) modalités sur les échantillons complets, comparés avec la courbe d'un  $\chi^2$  théorique à 2 puis 8 degrés de liberté.



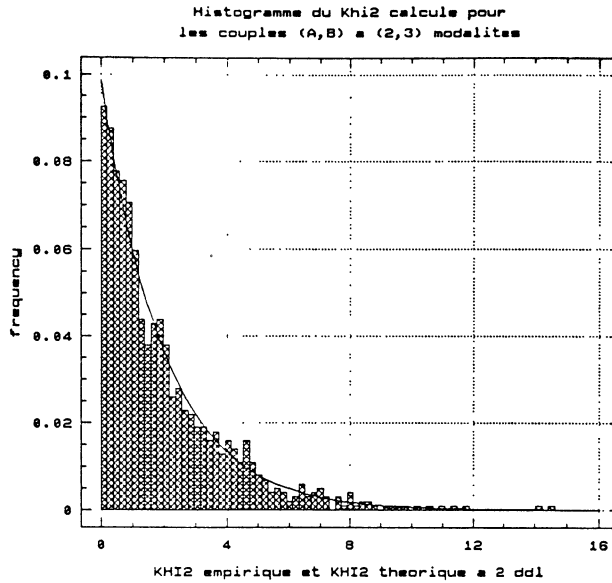


FIGURE 1

*Histogramme des valeurs du  $\chi^2$  calculées pour  $N_s = 1000$  échantillons de taille 1000 de couples (A, B) à 2 et 3 modalités, et comparaison à la loi du  $\chi^2$  à 2 ddl*

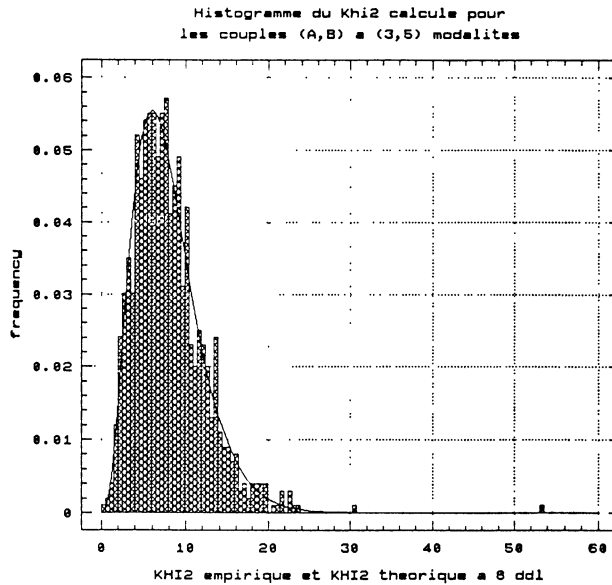


FIGURE 2

*Histogramme des valeurs du  $\chi^2$  calculées pour  $N_s = 1000$  échantillons de taille 1000 de couples (A, B) à 3 et 5 modalités, et comparaison à la loi du  $\chi^2$  à 8 ddl*

L'étude de l'ensemble des résultats montre que ceux-ci sont très peu différents que l'on considère les fichiers complets ou extraits et réalisent toujours une bonne approximation du résultat théorique attendu.

- De même, les fractiles empiriques obtenus sur les échantillons simulés estiment bien les fractiles théoriques attendus (par exemple figure 3 pour les couples  $(A, B)$  à respectivement 2 et 2 modalités).

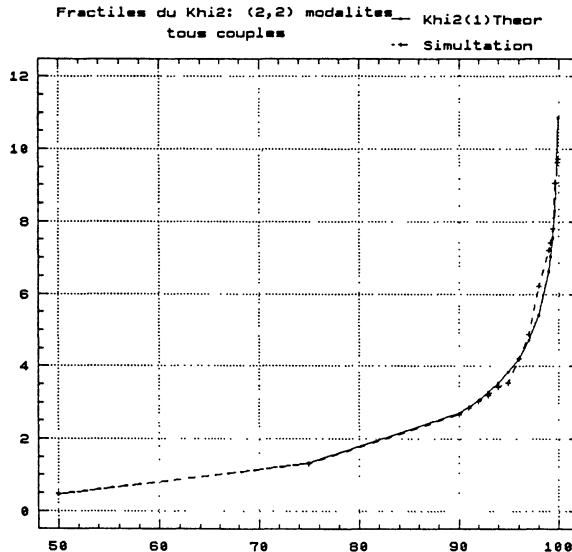


FIGURE 3

*Fractiles Empiriques ( $\chi^2$  calculé pour  $N_s = 1000$  échantillons de taille 1000 de couples  $(A, B)$  à 2 et 2 modalités) et Théoriques ( $\chi^2$  à 1 ddl); fractiles d'ordre supérieur à 0,5*

Les simulations ici réalisées sont validées par la concordance, pour le  $\chi^2$ , entre les résultats théoriques attendus et les résultats obtenus empiriquement; nous allons maintenant nous intéresser aux résultats de ces mêmes simulations pour le coefficient  $Q[A, B]$ .

#### 4.4 Distribution du coefficient $Q[A, B]$

Pour le coefficient de corrélation généralisé, nous ne disposons pas de résultats théoriques mais les simulations réalisées permettent d'évaluer empiriquement les fractiles de  $Q[A, B]$ . Les courbes ou les tables ainsi établies peuvent alors donner une indication sur le caractère significatif d'une valeur de  $Q[A, B]$  donnée.

- Les figures 4 et 5 présentent les histogrammes des valeurs de  $Q[A, B]$  obtenues pour une même valeur du couple (Mod  $A$ , Mod  $B$ ), 2 et 3 modalités, puis 3 et 5 modalités.

- La figure 6 présente les courbes des fractiles empiriques de  $Q[A, B]$  obtenus pour une même valeur du couple (Mod  $A$ , Mod  $B$ ).

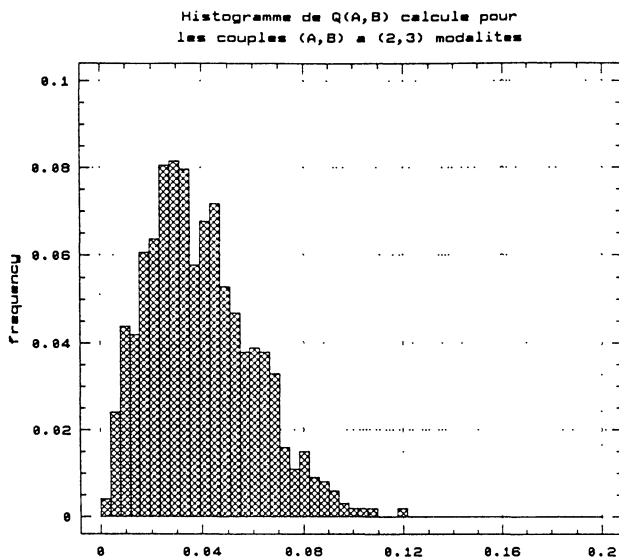


FIGURE 4  
Histogramme de  $Q[A, B]$  calculé pour  $N_s = 1000$  échantillons  
de taille 1000 de couples  $(A, B)$ , à 2 et 3 modalités

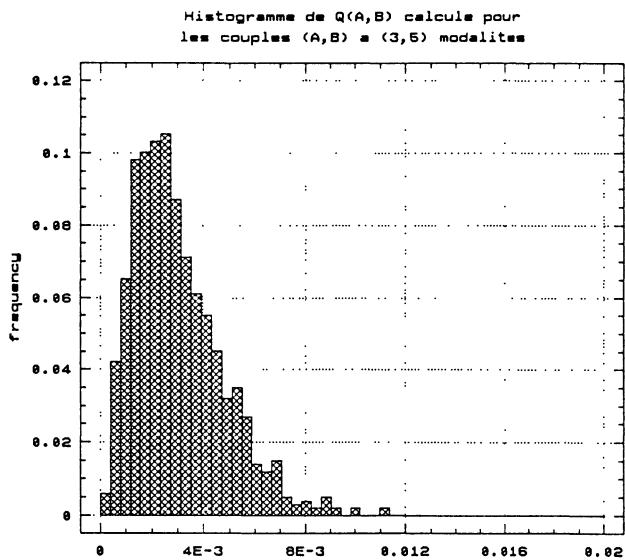


FIGURE 5  
Histogramme de  $Q[A, B]$  calculé pour  $N_s = 1000$  échantillons  
de taille 1000 de couples  $(A, B)$ , à 3 et 5 modalités

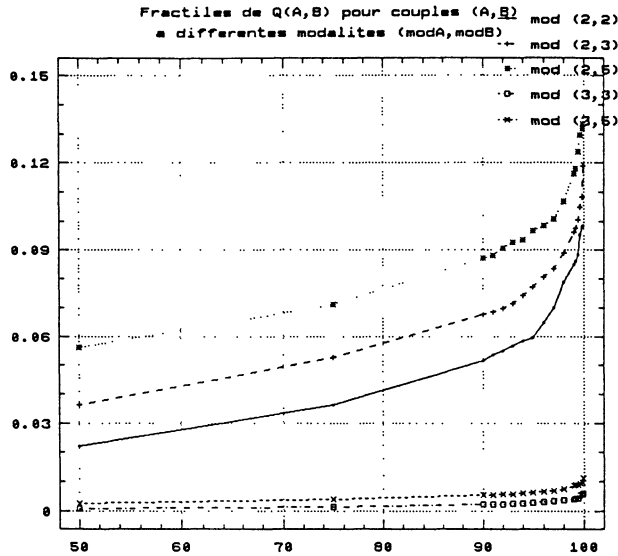


FIGURE 6

*Fractiles Empiriques des valeurs de  $Q[A, B]$ ; fractiles d'ordre supérieur à 0.5*

- Sur la figure 7, le coefficient  $Q[A, B]$  a été préalablement élevé à la puissance  $1/(s-1)$  pour pouvoir comparer les valeurs obtenues avec différents couples (Mod A, Mod B) : en effet, avec  $s = \inf(\text{Mod A}, \text{Mod B})$ , nous avons vu que  $Q[A, B]^2$  est de degrés  $2(s-1)$  en  $n_{ij}$ .

- Sur la figure 8, on détaille les mêmes courbes mais uniquement pour les fractiles d'ordre supérieur à 0.9, qui nous intéressent tout particulièrement pour tester la significativité d'une valeur de  $Q[A, B]$  donnée.

## 5. Conclusion

Nous nous sommes intéressés ici à deux critères d'interliaison statistique entre des variables qualitatives; le  $\chi^2$  de Pearson et un coefficient de corrélation généralisé, étendant la notion de coefficient de corrélation, défini classiquement entre des variables quantitatives, à des variables qualitatives.

Bien que ces deux critères aient été construits à partir d'idées différentes, distance (entre deux tables de contingence) pour le  $\chi^2$  et coefficient de corrélation ordinaire (entre les variables indicatrices des variables qualitatives considérées) pour le coefficient de corrélation généralisé  $Q[A, B]$ , nous avons pu présenter une définition de ces critères à partir d'une matrice commune : le  $\chi^2$  s'exprimant en fonction de la trace de cette matrice et le coefficient  $Q[A, B]$  en fonction de son déterminant.

Il est aussi intéressant de noter qu'une relation analytique a été établie entre ces deux critères dans le cas particulier où l'une des deux variables qualitatives ne possède que 2 modalités.

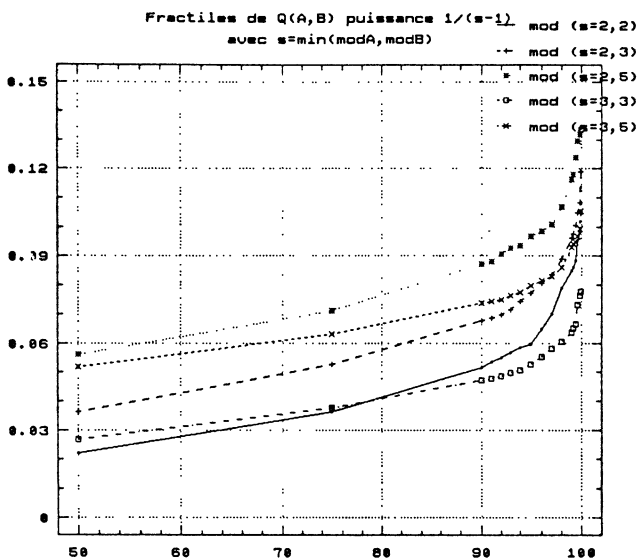


FIGURE 7

Fractiles Empiriques des valeurs de la racine d'ordre  $s - 1$  de  $Q[A, B]$ ;  
 $s = \min(\text{Mod}A, \text{Mod}B)$ ; fractiles d'ordre supérieur à 0.5

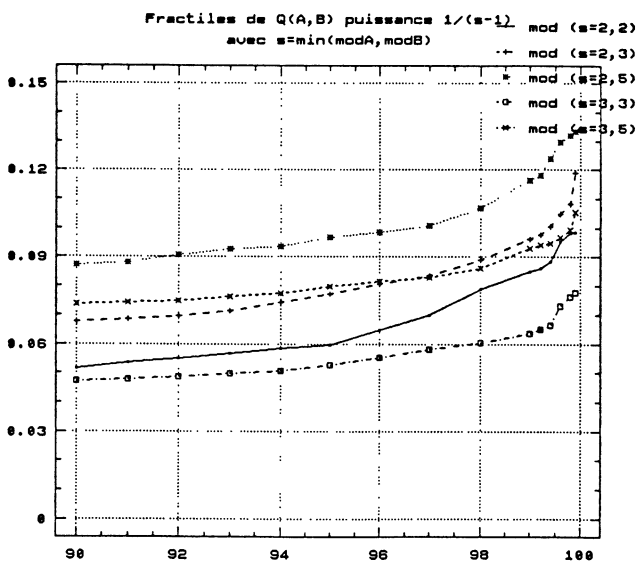


FIGURE 8

Fractiles Empiriques des valeurs de la racine d'ordre  $s - 1$  de  $Q[A, B]$ ;  
 $s = \min(\text{Mod}A, \text{Mod}B)$ ; fractiles d'ordre supérieur à 0.9

## 6. Références

- (1) CRAMER, 1945, «Mathematical Methods of Statistics», Wiley.
- (2) DER MEGREDITCHIAN, 1988, «Le traitement statistique des données multidimensionnelles», Tome 1, Ecole Nationale de la Météorologie, Paris.
- (3) VIALAR, 1978, «Calcul des probabilités et statistiques», Tome 3, Direction de la Météorologie, Paris.