REVUE DE STATISTIQUE APPLIQUÉE

A. PATRIS
N. CHAU
B. JAMBON
B. LEGRAS
F. KOHLER
J. MARTIN

Le coefficient de contingence de K. Pearson

Revue de statistique appliquée, tome 32, n° 2 (1984), p. 17-30
<http://www.numdam.org/item?id=RSA_1984__32_2_17_0>

© Société française de statistique, 1984, 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/

LE COEFFICIENT DE CONTINGENCE DE K. PEARSON

A. PATRIS(*), N. CHAU(*), B. JAMBON(**), B. LEGRAS(*), F. KOHLER(*), J. MARTIN(*)

(*) Section de Statistique et d'Informatique Médicale, Groupe INSERM U.115, Faculté de Médecine, Université Nancy I.

(**) O.R.S.T.O.M. détaché au Laboratoire d'Immunologie, Faculté de Médecine, Université Nancy I.

RESUME

Dans le cas de mesures qualitatives ou mixtes (quantitatif-qualititif) on examine des paramètres, qui, comme le coefficient de corrélation linéaire entre variables quantitatives, puissent servir d'indicateur de la force du lien ou permettent de faire des analyses multivariées de type analyse en composantes principales, coefficient de correlation partiel. . . Les auteurs veulent spécialement attirer l'intérêt du lecteur sur un coefficient introduit par K. PEARSON, modifié quelque peu par les auteurs qui en ont étudié le comportement par simulation et l'ont utilisé avec succès dans des analyses multidimensionnelles.

1. INTRODUCTION

Pour l'étude d'un problème dans lequel les variables mesurées sont quantitatives, le coefficient de corrélation linéaire est irremplaçable, surtout dans le cadre du modèle gaussien. Il sert de statistique pour effectuer les tests d'indépendance, comme échelle de la force du lien (échelle que chacun "voit" avec un peu d'expérience), et il intervient dans les notions de corrélations multiples, partielles, analyse en composantes principales, régression linéaire, analyse discriminante.

Cet article a pour but de présenter un coefficient introduit par K. PEARSON, un peu modifié par les auteurs, qui essaye de remplir le même rôle dans le cas où les variables mesurées sont qualitatives. L'utilisation de ce paramètre sera d'autant plus justifiée que les variables mesurées seront en fait sous-jacemment quantitatives. Ceci est un cas fréquent en médecine, domaine dans lequel cet outil a été utilisé. On y rencontre fréquemment des modalités du type "pas du tout", "un peu", "beaucoup". Dans ce cadre, le coefficient de PEARSON fournira une estimation du coefficient de corrélation linéaire entre les variables continues sous-jacentes. On aura ainsi une justification théorique pour l'utiliser comme échelle de la force du lien, pour construire des coefficients de corrélation partielle, ou dans les ACP lorsque, par exemple, le nombre total de modalités rend impossible une analyse des correspondances.

Nous allons commencer par un tour succinct de quelques paramètres répondant plus ou moins au problème posé. Nous étudierons ensuite plus en détail le coefficient proposé par PEARSON.

Mots Clés: **Variables qualitatives, Coefficient de corrélation, Coefficient de contingence de Pearson, Coefficient de Tschuprow, Phi-deux.**

2. SURVOL DE QUELQUES PARAMETRES

2.1. Le coefficient de corrélation des rangs

Si les caractères qualitatifs mesurés sont ordonnés, on peut utiliser le coefficient de corrélation des rangs. Ce qui revient à codet les modalités 1, 2, 3..... m. On obtient un coefficient de corrélation, défini aussi dans la population totale, avec lequel on peut mathématiquement utiliser les outils de l'analyse linéaire. Cependant, le codage peut apparaître arbitraire.

Il est bien sûr possible de coder autrement, en utilisant par exemple une analyse des correspondances. On peut cependant s'approcher d'un cercle vicieux, puisque le code obtenu sert à examiner des liens avec des variables qui ont permis de construire le code.

2.2. Le Phi-deux

Bien que peu utilisé pratiquement il est à la base de tous les paramètres qui vont suivre. Donnons en la définition : Soient deux variables aléatoires X et Y. Soit une mesure $\mu(x, y)$ sur l'image de (X, Y), qui ici sera soit la mesure de Lebesgue sur R^2 soit une mesure discrète si X et Y sont qualitatives. Soient f_X, f_Y, f les densités, supposées exister, de X, Y, (X, Y) par rapport à μ. La définition donnée par PEARSON [6] est :

$$\Phi^2 = \iint \frac{f^2(x,y)}{f_X(x)\, f_Y(y)}\, d\mu(x,y) - 1. \qquad (1)$$

$$= \iint \frac{(f(x;y) - f_X(x)\, f_y(y))^2}{f_X(x)\, f_Y(y)}\, d\mu(x,y)$$

si cette quantité existe. C'est en particulier le cas si (X, Y) suit une loi gaussien ou si X et Y ont un nombre fini de valeurs. Dans ce cas on a bien sûr :

$$\Phi^2 = \sum_i \sum_j \frac{[P(X = x_i; Y = y_i)]^2}{P(X = x_i)\, P(Y = y_i)} - 1$$

qui est estimée à l'airde d'un n-échantillon par la quantité :

$$\hat{\Phi}^2 = -1 + \sum_i \sum_j \frac{n_{ij}^2}{n_{i.}\, n_{.j}} \qquad (2)$$

Dans le cas gaussien, PEARSON [6] a montré que la relation entre le coefficient de corrélation ρ et Φ^2 était donnée par :

$$\rho^2 = \frac{\Phi^2}{\Phi^2 + 1} \; ; \quad \Phi^2 = \frac{\rho^2}{1 - \rho^2}$$

et que l'on pouvait estimer Φ^2 après découpage en classes en utilisant la relation (2) qui fournit alors un estimateur convergent vers Φ^2 si les classes deviennent

de plus en plus fines et si le nombre d'indivisus dans chacune des cases du tableau des effectifs croisés tend vers l'infini.

De manière générale, Φ^2 définit une échelle de lien variant de zéro à l'infini, et est égal à zéro si et seulement si X et Y sont indépendantes.

Dans le cas discret comme dans le cas gaussien on obtient donc un estimateur convergent asymptotiquement vers un paramètre défini dans la population totale. On pourrait en faire une échelle de la force du lien. Mais dans le cas continu, on ne voit pas l'intérêt de l'utiliser en remplacement du coeffieint de corrélation linéaire dont tout le monde à l'habitude.

2.3. Le Chi-deux

La statistique du Chi-deux que l'on peut écrire :

$$\chi^2 = n\hat{\Phi}^2$$

sert naturellement pour tester l'indépendance entre deux variables qualitatives. Elle peut aussi servir d'indicateur de la force du lien. Cependant c'est une échelle qui dépend de la taille de l'échantillon : s'il n'y a pas indépendance, le chi-deux tend vers l'infini en même temps que n. Un inconvénient analogue se produit si on utilise comme indice la fonction de répartition de la loi du chi-deux.

2.4. Les coefficients de Tschuprow et de Cramer

Ces deux coefficients ont été présentés par leurs auteurs comme mesure de la force du lien. Leur définition, fournie par exemple dans [5] est donné par :

$$\hat{C} = \sqrt{\frac{\hat{\Phi}^2}{m-1}}$$

$$\hat{T} = \sqrt{\frac{\hat{\Phi}^2}{\sqrt{(m_1-1)(m_2-1)}}}$$

où m_1 est le nombre de modalités de X, m_2 le nombre de modalités de Y et $m = \inf(m_1, m_2)$. Les coefficients sont très proches, et même égaux si $m_1 = m_2$. Dans le cas contraire $\hat{T}$ ne peut jamais atteindre la valeur 1, même si les modalités de X peuvent être reconstruites à partir de celles de Y, contrairement à $\hat{C}$. Ceci peut être un avantage ou un inconvénient selon que l'on prend ou non en compte le fait que si $m_1 \neq m_2$ les deux variables n'apportent pas la même quantité d'information.

Si X et Y ont un nombre fini de modalités déterminées à priori, $\hat{C}$ et $\hat{T}$ sont alors des estimateurs convergents de paramètres définis dans la population totale. Par contre, s'il y a des variables continues sous-jacentes et si le découpage en classes est d'autant plus fin que la taille de l'échantillon croît, on se heurte à une difficulté. En effet, $\hat{\Phi}^2$ va converger vers Φ^2 défini en (1) (s'il est fini), et le nombre de classes tendant vers l'infini, $\hat{C}$ et $\hat{T}$ vont converger vers zéro. On pourra d'ailleurs constater ce phénomène dans les simultations que l'on présentera plus loin. C'est donc un défaut important qui interdit des comparaisons de la force du lien quand les variables n'ont pas le même nombre de modalités.

Ces deux paramètres sont semble-t-il assez peu utilisés.(On pourra cependant voir une application médicale dans [1]). Mais le second vient de faire une timide réapparition grâce à l'analyse des données.

SAPORTA [9] a montré que dans le cas de variables discrètes, Φ^2 était le produit scalaire entre les opérateurs d'espérance conditionnelle (projecteurs orthogonaux sur les sous espaces engendrés par les indicatrices des modalités) le produit scalaire de deux opérateurs étant la trace du produit de composition de ces opérateurs. La norme d'un tel projecteur est égal au nombre de modalités moins 1 (on se place dans l'espace des variables de moyenne nulle). On en déduit un cosinus entre deux variables discrètes :

$$\cos(X, Y) = T^2 = \frac{\Phi^2}{\sqrt{(m_1 - 1)(m_2 - 1)}}$$

Ce paramètre fournit une véritable matrice de cosinus, et certains, dont l'auteur, s'en sont servi pour faire des ACP ou des analyses canoniques. Les inconvénients de ce paramètre sont les suivants :

- T^2 est bien plus petit que le coefficient de corrélation ρ. Son utilisation dans une ACP fournit alors des valeurs propres assez petites.
- Il a été montré par DAUDIN [3] que l'indépendance conditionnelle n'impliquait pas que le coefficient de corrélation partielle, obtenu à partir de T^2, soit nul. Et inversement.
- Comme pour T, mais plus rapidement encore, si on prend les classes de plus en plus fines, alors T^2 tend vers zéro (si n croit suffisamment). C'est pourquoi il était intéressant de rechercher d'autres coefficients.

3. LE COEFFICIENT DE PEARSON

L'idée de PEARSON repose sur le résultat que nous avons déjà enoncé : si l'on dispose d'un couple gaussien (X, Y), de coefficient de corrélation ρ, on a alors :

$$\rho^2 = \frac{\Phi^2}{\Phi^2 + 1}$$

Si on échantillonne et si l'on met en classes, alors

$$\hat{P} = \sqrt{\frac{\hat{\Phi}^2}{\hat{\Phi}^2 + 1}} \quad \text{(coefficient de contingence de PEARSON)} \qquad (3)$$

converge vers $|\rho|$ si les classes deviennent de plus en plus fines et si n croit suffisamment. Ce coefficient est exposé dans [2], [5], [6], [8].

Ce paramètre possède à priori de nombreux avantages. Si l'on suppose l'existence de variables continues gaussiennes sous-jacentes, on obtient quelque chose d'assimilable à un coefficient de corrélation linéaire. Compte tenu de la mise en classes, il suffit que les variables continues sous-jacentes puissent être rendues gaussiennes par une bijection. L'utilisation du coefficient de PEARSON peut donc être considérée comme un moyen de faire un changement de variable implicite.

On sera alors en droit d'utiliser la notion de corrélations partielles obtenues à partir de la matrice des $\hat{P}$. S'il n'y a pas de modèle gaussien le coefficient $\hat{P}$ fournit une échelle de la force du lien qui sera comparable à celle du coefficient de corrélation linéaire. Et dans la mesure où la matrice des coefficients n'aura pas trop de valeurs propres négatives, une A.C.P. aura toujours un sens.

Les qualités du coefficient $\hat{P}$ étant asymptotiques, il faut examiner son comportement réel. Pour cela les auteurs ont effectué quelques simulations et ont procédé à quelques analyses de problèmes réels en utilisant le coefficient de PEARSON.

4. SIMULATIONS ET COEFFICIENT DE PEARSON MODIFIE

Nous avons examiné le comportement de $\hat{P}$, $\hat{C}$ et $\hat{T}$ par simulations. Elles ont été effectuées pour des tailles d'échantillon $n = 75, 500, 2000$. Pour chaque valeur de n nous avons simulé des réalisations d'un échantillon d'un couple gaussien de coefficient

$$\rho = 0 \ ; \ 0.1 \ ; \ 0.2 \ ; \ \ldots \ ; \ 0.9 \ ; \ 0.95.$$

Le nombre de simulations a été de :

500 pour $n = 75$, 100 pour $n = 500$ et 40 pour $n = 2000$.

Pour le calcul de $\hat{\Phi}^2$, nous avons découpé chaque variable en 2, 3, 4, 6 et 8 classes de probabilités théoriques égales. Et nous n'avons calculé $\hat{\Phi}^2$ que pour un nombre de classes identique pour chaque variable. Cas pour lequel on a $T = C$. Chaque simulation nous a donc fourni 5 estimations de Φ^2. Dans les graphiques ne sont indiquées que celles pour lesquelles le nombre de classes est raisonnable relativement à la taille de l'échantillon. Les graphiques fournissent les moyennes simulations ainsi obtenues.

Pour chaque point de ces graphiques, si on note n le nombre de simulations effectuées pour estimer la moyenne de l'écart-type observé, alors la quantité $2s/\sqrt{n-1}$ est toujours inférieure à 0.015. Ceci donne une résolution graphique suffisante. Il ne faut cependant pas oublier que les moyennes obtenues dépendent peut-être du mode de découpage des classes, qui ici sont assez équilibrées.

L'analyse des résultats permet de constater que :

- Le coefficient de Tschuprow est assez dépendant du nombre de modalités et dans un sens gênant : $\hat{T}$ diminue et s'éloigne de ρ si le nombre de classes augmente. De plus on obtient une échelle de la force du lien assez compressée vers le bas.
- Le coefficient de PEARSON dépend lui aussi du nombre de modalités, mais varie dans un bon sens. L'important est de constater qu'il est presque une fonction affine de ρ et que l'estimation est d'autant meilleure que n et le nombre de classes croissent. Mais il est apparu nécessaire de l'améliorer, car les simulations montrent que ce paramètre a quelques inconvénients manifestes. (voir les graphes).

Soit $\nu = (m_1 - 1)(m_2 - 1)$ le nombre de degré de liberté.

Soit $m = \inf(m_1, m_2)$

D'après [5], vol. 2, on a asymptotiquement :

$$E(\hat{\Phi}^2) = \Phi^2 + \frac{\nu}{n}$$

et d'autre part la valeur maximum de $\hat{\Phi}^2$ est égale à $m - 1$. La transformation qui s'impose consiste alors à centrer $\hat{\Phi}^2$ pour améliorer l'estimateur au voisinage de $\rho = 0$, et à normer l'estimateur obtenu de façon à rendre son maximum égal à 1 quand $\rho = 1$.

Soit donc :

$$\hat{\Phi}'^2 \begin{cases} \Phi^2 - \frac{\nu}{n} \text{ si } \hat{\Phi}^2 > \frac{\nu}{n} \\ 0 \quad \text{sinon} \end{cases}$$

$$A = \frac{m - 1 - \nu/n}{m - \nu/n}$$

on pose alors :

$$\hat{P}' = \sqrt{\frac{\hat{\Phi}'^2}{A(\hat{\Phi}'^2 + 1)}}$$

et on a fourni les simulations de ce nouveau paramètre, noté coefficient de Pearson modifié. On constate une nette amélioration par rapport à P. Cependant, il reste encore une sous estimation systématique de ρ, qui diminue légèrement quand n croit. Ceci est dû au nombre fini de classes. On remplace en effet (1), l'intégrale d'un carré par (2), une somme des moyennes des carrés. Et l'estimation est d'autant plus mauvaise que la densité varie beaucoup à l'intérieur d'une case, ce qui est le cas quand ρ est proche de 1. On pourrait imaginer une procédure rectificatrice plus ou moins empirique. N'ayant pas trouvé de moyen théorique pour calculer cette rectification, les auteurs ont préféré laisser le paramètre tel quel.

Au niveau de sa moyenne, $\tilde{P}'$ dépend assez peu du nombre de classes, ce qui n'est pas le cas de sa variance. Les écarts-type observés dans les simulations sont fournis dans le tableau I. Pour $\rho = 0$, l'écart-type minimum est obtenu pour $m = 2$, et quand ρ croit, le minimum est obtenu pour des valeurs croissantes de m. On peut cependant constater que le choix de $m = 3$ pour $n = 75$, $n = 500$ et $m = 6$ pour $n = 2000$ fournit des écarts-types toujours assez proches du minimum. Ce sont donc des valeurs à conseiller. Comparons maintenant ces écarts-types à ceux du coefficient de corrélation linéaire R

$$\sigma_R^2 = \frac{1 - \rho^2}{n} \left(\frac{1 + 11\rho^2}{2n}\right) + 0\left(\frac{1}{n^3}\right)$$

dont quelques valeurs sont fournies dans le tableau II. On constate en premier lieu que le coefficient de PEARSON modifié a un écart-type supérieur à celui de R sauf pour ρ voisin de zéro, ce qui est dû à un effet de troncature. Le rapport entre les deux écarts-types est particulièrement important quand ρ dépasse0.9, mais diminue lorsque n croit. En particulier pour $n = 2000$ et $m = 6$, le rapport est presque toujours très proche de 1.

TABLEAU I

Ecarts-Types observés dans les simulations du Coefficient de PEARSON modifié, en fonction de la taille de l'échantillon n, du nombre de classes m et de la valeur de ρ.

n	m \ ρ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95
75	2	.103	.109	.149	.165	.170	.149	.130	.108	.087	.063	.046
75	3	.121	.118	.149	.157	.147	.124	.090	.076	.060	.045	.035
500	2	.034	.061	.066	.057	.059	.054	.044	.042	.031	.028	.020
500	3	.045	.064	.055	.047	.047	.039	.035	.026	.022	.019	.013
500	4	.045	.064	.058	.045	.034	.034	.025	.023	.019	.014	.012
500	6	.060	.075	.068	.056	.043	.035	.029	.023	.018	.011	.007
2000	2	.014	.032	.030	.032	.027	.029	.021	.017	.014	.011	.007
2000	3	.022	.029	.026	.025	.022	.022	.016	.015	.010	.010	.006
2000	4	0.24	.036	.027	.023	.022	.017	.016	.012	.009	.006	.005
2000	6	.028	.038	.025	.021	.018	.016	.014	.011	.008	.006	.005

TABLEAU II

Ecarts-types du coefficient de corrélation linéaire R en fonction de la taille de l'échantillon et de ρ.

n \ ρ	0	.1	.2	.3	.4	.5	.6	.7	.8	.9	.95
75	.115	.114	.111	.105	.098	.087	.075	.060	.043	.023	.012
500	.045	.044	.043	.041	.038	.034	0.29	.023	.016	.009	.004
2000	.022	.022	.021	.020	.019	.017	.014	.011	.008	.004	.002

Dans le cas où les variables qualitatives mesurées sont sous-jacemment continues, on peut mettre un signe sur $\hat{P}'$ pour se rapprocher le plus d'une véritable matrice de corrélation. Les auteurs se sont servis, dans ce but du coefficient de corrélation des rangs dont ils ont affecté le signe à $\hat{P}'$.

5. UTILISATION

Depuis environ 6 mois nous utilisons le coefficient de PEARSON modifié dans l'analyse de problèmes médicaux.

Son utilisation en tant qu'indicateur de la force du lien a donné satisfaction ainsi que le montrent les simulations dans le cas gaussien.

Dans le cadre purement qualitatif l'échelle obtenue n'est pas sans intérêt puisqu'il s'agit d'une normalisation de l'échelle du Φ^2.

L'utilisation de la matrice des coefficients de PEARSON modifié comme matrice de corrélation pose un petit problème pour les analyses multivariées : elle n'est pas généralement définie positive. Pour améliorer la situation, on aura bien évidemment intérêt à utiliser le coefficient avec son signe. L'utilisation de ce type de matrice sur une dizaine d'études, avec un nombre de variables allant de 20 à 70, a fourni des valeurs propres négatives représentant entre 1 % et 5 % de la trace de la matrice, ce qui est assez peu. On peut alors considérer que la représentation par une analyse en composantes principales ne pose aucun problème.

Le cas des analyses canoniques et des corrélations partielles est plus délicat.

L'obtention des corrélations partielles et l'analyse canonique nécessitent l'inversion de matrices des coefficients $\hat{P}'$ (pouvant s'interprêter comme associée à un opérateur de projection). Il est apparu qu'il était nécessaire de partir d'une matrice définie positive, sinon les petites valeurs propres négatives vont jouer un grand rôle dans la matrice inverse et vont provoquer des phénomènes inacceptables. Il faut donc modifier la matrice des coefficients $\hat{P}'$. Une première méthode consiste à ajouter sur la diagonale une quantité légèrement supérieure à la valeur absolue de la plus importante valeur propre négative. Pour obtenir des 1 sur la diagnoale, on renormalise ensuite en divisant chaque terme de la matrice par la valeur commune de la diagonale.

Une seconde méthode, utilisant la déflation, rend nulle les valeurs propres auparavant négatives. On rajoute une petite quantité sur la diagonale pour ne pas obtenir de matrice singulière et on renormalise la matrice pour obtenir une matrice de cosinus.

Les quelques essais effectués montrent que la seconde méthode transforme moins la matrice. La moyenne des valeurs absolues des différences entre les éléments de la matrice de départ et ceux de celle d'arrivée pouvant être par exemple de l'ordre de 0,02 dans la 2^{e} méthode et de 0,04 dans la première. Cependant la première méthode fournit une transformation identique pour tous les coefficients ce qui n'est pas le cas de la seconde.

Ceci a été utilisé dans l'étude de la dégradation du thymus chez des enfants morts en état de sous-nutrition à Dakar (voir [4], [10]).

Le thymus est un organe où se produit la maturation des lymphocytes T, cellules ayant un rôle important dans le système immunitaire. La maturation s'effectue en particulier par l'action d'une hormone tymique, le FTS. Cette hormone est produite uniquement dans deux éléments du thymus : les cellules épithéliales et les corps de Hassall. En état de sous-nutrition, le thymus subit une atrophie très importante, les effectifs de cellules "utiles" se réduisent, la quantité de FTS produite diminue et des remaniements tissulaires se produisent.

Un des problèmes abordés était de choisir entre les deux modèles (simplifiés) d'évolution suivants. Dans le modèle, on suppose que la sous-nutrition entraîne direcement une dégradation de la structure du thymus, ce qui entraîne a postériori la suppression de la production de FTS. Dans le modèle 2, la malnutrition entraîne directement l'arrêt de la production de FTS. L'activité du thymus est alors perturbée, et il s'ensuite une disparition des structures produisant le FTS. Ce deuxième modèle laisse supposer que l'injection de FTS pourrait produire un effet bénéfique chez les mal-nutris.

Dans la suite, nous noterons N1 le groupe de variables mesurant l'état nutritionnel et sensible à des perturbations immédiates (ex : le rapport Poids/Taille). N2 désignera les variables nutritionnelles mesurant des perturbations nécessairement longues (ex : le rapport Taille/Age). Le groupe FTS mesura la production de l'hormone FTS (quantité, nombre de sites de production). Le groupe structure mesurera le poids du thymus, le nombre de lymphocytes, de corps de Hassall, producteurs ou non de FTS, ainsi que la fibrose.

Pour des raisons pratiques, la majorité de ces variables n'a pu être mesurée qu'approximativement par un codage du type : peu, moyen, beaucoup. C'est pourquoi nous avons utilisé le coefficient de PEARSON modifié.

Pour choisir entre les deux modèles, nous avons utilisé la notion de covariance partielle que nous allons expliciter. Soient X et Y deux variables centrées réduites et soient $Z_1, \ldots, Z_p$ un ensemble de variables centrées engendrant un sous espace noté W. (On se place dans R^n, n = taille de l'échantillon, muni du produit scalaire usuel). Notons P_W le projecteur orthogonal sur W. On peut alors écrire :

$$X = X_1 + X_2 \quad \text{avec} \quad X_1 = P_W(X) \quad \text{et} \quad X_2 \perp W$$

$$Y = Y_1 + Y_1 \quad \text{avec} \quad Y_1 = P_W(Y) \quad \text{et} \quad Y_2 \perp W$$

Le coefficient de corrélation entre X et Y se décompose :

$$\rho(X, Y) = \text{Cov}(X_1, Y_1) + \text{Cov}(X_2, Y_2)$$

La covariance entre X_2 et Y_2, notée covariance partielle désigne alors la partie de la corrélation entre X et Y dûe à des phénomènes "indépendants" de $Z_1, \ldots, Z_p$. La quantité :

$$\rho(X_2, Y_2) = \frac{\text{Cov}(X_2, Y_2)}{(\text{Var}(X_2) \,.\, \text{Var}(Y_2))^{1/2}}$$

est la corrélation partielle entre X et Y pour $Z_1, \ldots, Z_p$ fixées.

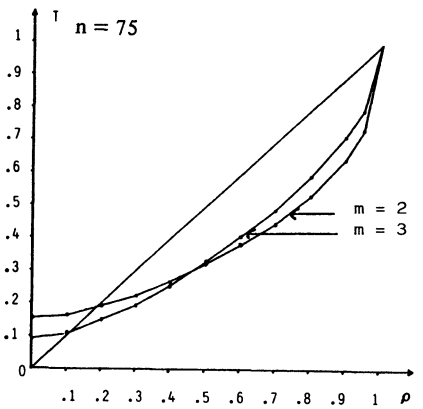
Nous avons préféré utiliser les covariances partielles Cov (X_2, Y_2) plutôt que les corrélations partielles puisque notre problème était de savoir si une partie importante de la liaison entre X et Y était ou non indépendante d'un groupe $Z_1, \ldots Z_p$. L'analyse des résultats semble en faveur du modèle 2. Les covariances partielles Nutrition-FTS pour structure fixée sont caractérisées par le fait suivant : les covariances partielles entre la nutrition et le nombre de cellules épithéliales ou de corps de Hassall producteurs de FTS sont toutes plus petites que 0.10. Au contraire les covariances partielles entre la nutrition et la quantité de FTS produite par cellule épithéliale ou par corps de Hassall sont plus élevées (de 0.10 à 0.26). Quant aux covariances partielles Nutrition/Structure pour FTS fixée, aucune ne dépasse 0.08. Pour obtenir de tels renseignements sur les lois conditionnelles avec les techniques usuelles de dénombrement, il aurait fallu étudier des croisements entre plus de deux variables simultanément. L'effectif de l'échantillon, n = 58, interdisait une telle démarche. Le modèle gaussien, qui paramétrise de façon simple les lois conditionnelles en permet l'estimation. Nous n'avons pas testé l'hypothèse selon laquelle nos données pouvaient se représenter à l'aide d'une loi gaussienne multidimensionnelle, l'échantillon étant faible. Mais les conclusions auxquelles on a abouti, ont le mérite d'être cohérentes et de correspondre à un processus etiopathogénique connu dans d'autres domaines (pancréas, foie).

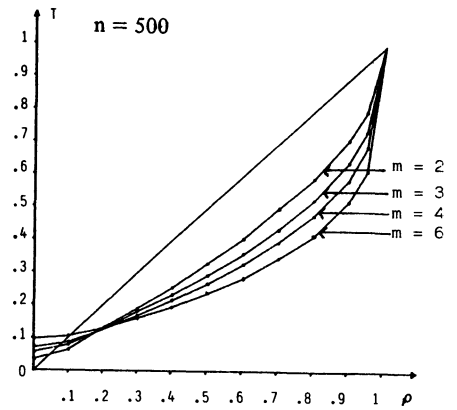
Un des problèmes important dans l'utilisation de cette technique c'est qu'il est impossible de fournir des intervalles de confiance sur les coefficients de corrélation partielle.

6. MELANGE QUANTITATIF-QUALITATIF

Bien souvent une étude comprend un mélange de variables quantitatives et qualitatives. Le calcul des coefficients de PEARSON modifiés nécessite la mise en classes des variables quantitatives et produit une perte d'information. Dans le cas de la liaison entre deux variables quantitatives, il est certainement préférable d'utiliser le coefficient de corrélation linéaire fourni par l'échantillon, plutôt que d'utiliser P' après avoir découpé en classes. On obtiendra alors une matrice de "corrélation" dont tous les éléments n'aurait pas été calculés par la même formule, mais qui restera néanmoins homogène.

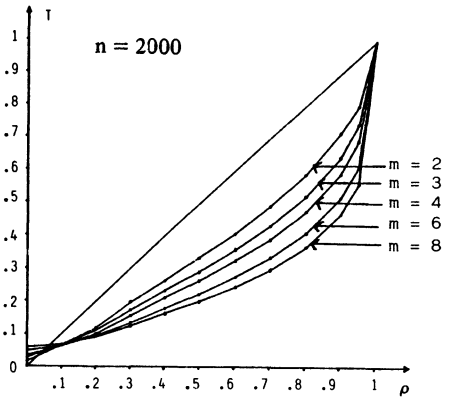
Quant au lien entre une variable quantitative et une variable qualitative (sous-jacemment continue), les auteurs n'ont pas trouvé de paramétres aussi précis que $\hat{P}'$: la mise en classes semble s'imposer. En effet, si on utilise la technique simple de l'analyse de la variance donnant une estimation de la variance résiduelle, celle-ci va fournir une estimation surestimée de la vairance conditionnelle de la variable mesurée en continu, et ceci à cause de la largeur des classes de l'autre variable. Il s'en suivra une sous-estimation du coefficient de corrélation. Et la rectification ne semble pas simple.



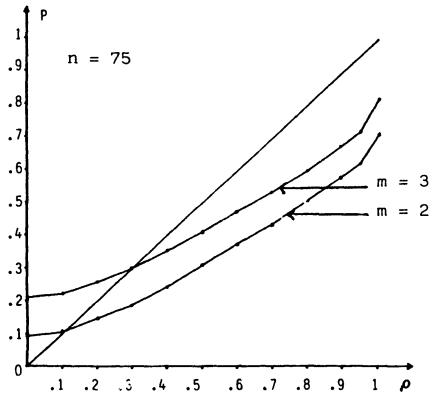


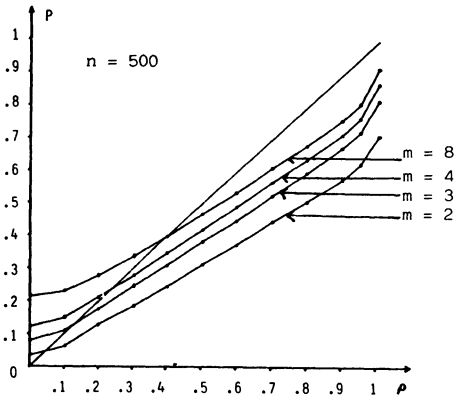
L'ordonnée à l'origine, de l'ordre de $\sqrt{\frac{m-1}{n}}$ est inférieure à celle obtenue pour $\hat{P}$. Cependant, quand ρ croit, $\hat{T}$ s'écarte d'autant plus de la diagonale que m augmente. A la limite $\hat{T}$ tend vers zéro. (Quand m et n $\longrightarrow \infty$)

Le maximum étant toujours égal à 1, on obtient une échelle qui ressemble assez peu à celle du coefficient de corrélation.



Moyennes des simulations du coefficient de TSCHUPROW

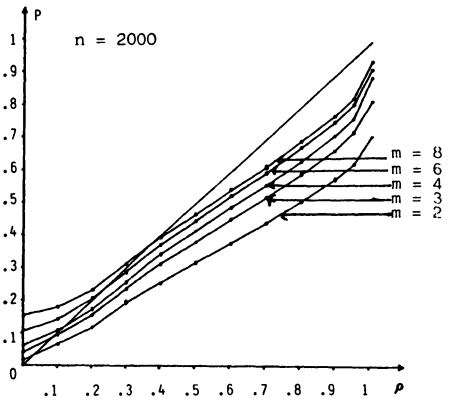




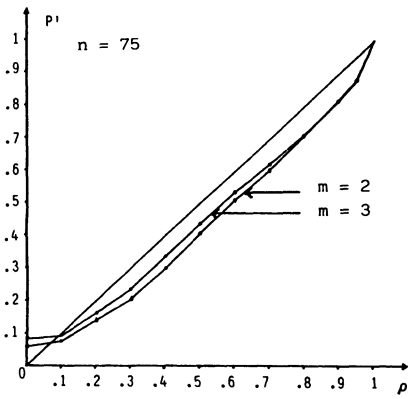
m désigne le nombre de classes construites sur X et Y pour obtenir $\hat{P}$.

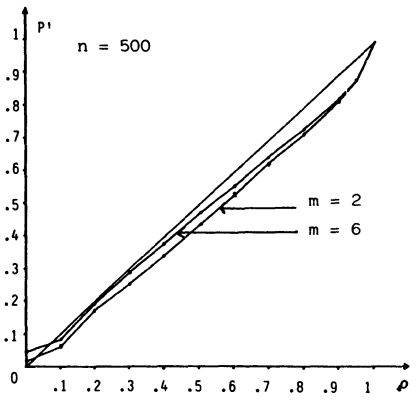
Les graphes obtenus forment des courbes parallèles, se rapprochant d'autant plus de la diagonale que m croit.

Le biais à l'origine est néanmoins important, de l'ordre de $\sqrt{\frac{m-1}{n}}$. Le maximum possible, atteint pour $\rho = 1$, est égal à $\sqrt{\frac{m-1}{m}}$.



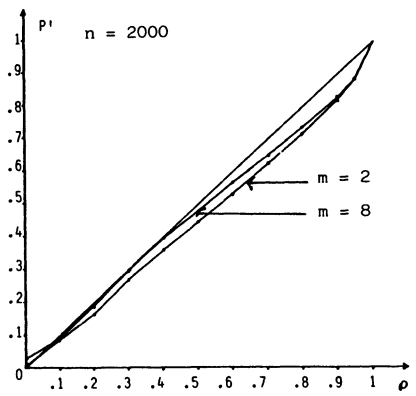
Moyennes des simulations du coefficient de PEARSON





Par soucis de clarté, seules les courbes pour les valeurs extrêmes de m sont fournies ; les autres sont comprises entre ces deux extrêmes et sont très rapprochées.

Les courbes sont assez proches de la diagonale. On s'en rapproche d'autant plus que n et m croissent et que ρ est petit.



Moyennes des simulations du coefficient de PEARSON modifié

REMERCIEMENTS

Nous remercions Mr. B. SIRANTOINE et Mme A. CROQUIN pour l'aide qu'ils nous ont apportée pour la réalisation de ce travail.

BIBLIOGRAPHIE

[1] COUDANE, FERY, SOMMELET, LACOSTE, LEDUC, GAUCHER. – Aseptic loosening of cemented total arthroplasties of the hip in relation to positioning of the prothesis : new utilization of the Tschuprow-Cramer statistical test, *Acta Orthop. Scand. 1981*, vol 52, n° 2 page 201-205.

[2] Dos GUPTAS. – Point biserial correlation coefficient and its generalization, *Psychometrika*, 25, 932-947, 1962.

[3] DAUDIN (1979). – Coefficient de Tschuprow partiel et indépendance conditionnelle, *Statistique et Analyse des données*, n° 3, p. 55-58.

[4] B. JAMBON, O. ZIEGLER, B. MAIRE, M.C. BENE, G. PARENT, A. PATRIS, J. DUHELLE. – *Tarissement de la secrétion de facteur thymique sérique F.T.S. et involution thymique chez les enfants sénégalais décédés en état de dénutrition protéino-énergétique,* Communication au 2e symposium sur les marqueurs de l'inflammation, Lyon, Juin 1983.

[5] KENDAL et STUART (1961). – *The Advanced Theory of Statistics*, vol. 2, London.

[6] K. PEARSON (1904). – On the theory of contingency and its relation to association and normal correlation, *Draper's Co. Memoirs, Biometric Series,* n° 1, London.

[7] Karl PEARSON'S. – *Early Statistical Papers,* Cambridge Univ. Press, London, 1948.

[8] Lothar SACHS. – *Applied Statistics*, Springer Verlag, 1982.

[9] G. SAPORTA (1976). – Quelques applications des opérateurs d'Escoufier au traitement des variables qualitatives, *Statistique et Analyse des données,* n° 1, p. 38-46.

[10] O. ZIEGLER. – *Les conséquences de la malnutrition protéino-énergétique sur la structure et le contenu du thymus en facteur thymique sérique (FTS),* Thèse de doctorat en Médecine, Univ. de Nancy I, 13 juin 1983.