HOMENAJE AL PROFESOR SIXTO RIOS
TRABAJOS DE ESTADISTICA Y DE INVESTIGACION OPERATIVA
Vol. 36, Núm. 3, 1985, pp. 113 a 118

TEST D'INDEPENDANCE MULTIDIMENSIONNELLE

Daniel Dugue

Institut de Statistique de Paris. 24 Rue Jean Louis SINET.
92330 SCEAUX France.

Proff of a result generalizing a Hoeffding's paper about independence.

Key words: Non parametric test; Independence test.

AMS Classification (1980): Primary, 62G10.

Un contraste de independencia multidimensional

Demostración de un resultado generalizando el artículo de Hoeffding sobre independencia.

Palabras clave: Prueba no parámetrica; Prueba de independencia.

Clasificación AMS (1980) : Primaria, 62G10.

Qu'il me soit permis d'apporter en hommage au grand statisticien le Professeur Sixto Rios cette démonstration d'un théorème dont j'ai donné l'énoncé dans une note aux comptes rendus de l'Académie de Sciences de Paris mais qui était restée sans explications. C'est pour moi une ocasion de rappeler tour les liens qui unissent l'école statistique espagnole et l'école française. Souvent nos deux pays ont eu l'occasion d'échanger des professeurs, des chercheurs et des étudiants. Le professeur Sixto Rios y a beaucoup aidé et j'espère que cette tradition sera maintenue.

Notations. J'appèlerai A_k une matrice k x k dont tous les éléments a_{ij} sont égaux à 0 si $j > i$ et 1 si $j \leq i$, B_j la matrice $A_k\ {}^tA_k$ (ce qui donne $b_{ij} = \min(i, j)$), M_k la matrice k x k où $m_{ij} = 1$ et I_k la matrice unitè k x k.

Etant donné une variable aléatoire p dimensionnelle j'appèlerai $\dot{H}_n(x_1,...,x_p)$ la distribution empirique de n résultats, la distribution théorique étant supposée à composantes indépendantes, les lois marginales $F_i(x_i)$ (i=1,...,p) ètant continues.

$H_n^{(i)}(x_i)$ sera la distribution empirique marginale de la $i^{ème}$ composante.Appelons D_n l'expression :

$$H_n(x_1,...,x_p) + (-1)^1 \quad \Sigma^{(1)} H_n^{(i)}(x_i)\, H_n(x_1,..., x_i = +\infty ,..., x_p)$$

$$+ (-1)^2 \quad \Sigma^{(2)} H_n^{(i)} H_n^{(j)}(x_j)\, H_n(x_1,..., x_i = +\infty ,..., x_j = +\infty ,..., x_p)$$

$$+ .....$$

$$+ (-1)^p H_n^{(1)}(x_1) ... H_n^{(p)}(x_p)$$

$\Sigma^{(k)}$ est etendue aux C(p, k) termes obtenus en faisant dans H_n,k termes égaux à $+\infty$ et en multipliant cette valeur de H_n par le produit des distributions marginales correspondantes.

Voici les étapes de la démonstration qui doit conduire à la limite de la fonction caractéristique de la variable aléatoire.

$$G_n = n \int_{R^p} D_n^2(x_1,..., x_p)\, dH_n^{(1)}(x_1) ... dH_n^{(p)}(x_p)$$

et constitue un test d'indépendance des composantes.

1º) Les $H_n^{(i)}(x_i)$ tendent dans tous les sens du calcul des probabilités vers les $F_i(x_i)$ et par conséquent la limite de la fonction caractéristique de G_n sera la même que celle de :

$$n \int_{R^p} D_n^2(x_1,..., x_p)\, dF_1(x_1) ... dF_p(x_p)$$

2º) L'expression $\sqrt{n} \prod_i (F_i - H_n^{(i)})$, i=1,...,p tend uniformément vers 0 en $x_1,...,x_p$ car $\sqrt{n}\,[F_1(x_1) - H_n^{(1)}]$ tend en loi vers la loi normale et $\prod_i (F_i - H_n^{(i)})$; i= 2,...,p tend dans tous les sens du calcul des probabilités vers 0.

La limite de la fonction caractéristique de G_n sera donc la même que celle de:

$$n \int_{R^p} [\, D_n(x_1,..., x_p) - \prod_i [\, F_i(x_i) - H_n^{(i)}(x_i)\,]\,]^2\, dF_1(x_1)...dF_p(x_p)$$

$D_n - \prod_i (F_i - H_n^{(i)})$ s'écrit :

$$H_n(x_1,\dots x_p) - F_1(x_1) \dots F_p(x_p)$$

$$+(-1)^1 \Sigma^{(1)} H_n^{(i)}(x_i) \left[H_n(x_1,\dots, x_i=+\infty,\dots, x_p) - F_1(x_1) \dots F_{i-1}(x_{i-1})F_{i+1}(x_{i+1})\dots F_p(x_p)\right]$$

$$+\dots$$

le dernier terme $(-1)^p H_n^{(1)}(x_1)\dots H_n^{(p)}(x_p)$ disparaissant.

Les termes $\sqrt{n}\left[H_n(x_1,\dots, x_i=+\infty,\dots, x_p) - F_1(x_1)\dots F_{i-1}(x_{i-1}) F_{i+1}(x_{i+1})\dots F_p(x_p) \right]$ tendent en loi vers une limite. Les termes $H_n^{(i)}(x_i)$ tendent vers $F_i(x_i)$

Donc la limite de la fonction caractéristique de G_n est la même que celle de :

$$n \int_{R^p} C_n^2(x_1,\dots, x_p)\, dF_1(x_1) \dots dF_p(x_p)$$

avec

$$C_n(x_1,\dots, x_p) = H_n(x_1,\dots, x_p) - F_1(x_1) \dots F_p(x_p)$$

$$+ (-1)^1 \Sigma^{(1)} F_i(x_i) \left[H_n(x_1,\dots,x_i=+\infty,\dots,x_p) - F_1(x_1)\dots F_{i-1}(x_{i-1}) F_{i+1}(x_{i+1})\dots F_p(x_p) \right]$$

$$+(-1)^2 \Sigma^{(2)} F_i(x_i)\, F_j(x_j)\left[H_n(x_1,\dots,x_i=+\infty,\dots,x_j=+\infty,..,x_p) - F_1(x_1)\dots F_{i-1}(x_{i-1})F_{i+1}(x_{i+1})\dots\right.$$

$$\left. F_{j-1}(x_{j-1})F_{j+1}(x_{j+1})\dots F_p(x_p)\right] + \dots$$

3º) Etant donné les conditions de régularité de la fonction C_n on peut remplacer l'intégrale par une somme de Riemann dont la limite quand k (chaque dimension étant partagée en k segments) augmente indéfiniment sera la même que la limite de G_n.

Les F_i étant continues partageons chaque dimension en k segments dans lesquels la croissance des F_i sera 1/k . Dans chacun des k^p domaines appelons $\Lambda H_n(z_1,\dots, z_p)$ l'accroissement p-dimensionnel de H_n le point $z_1,\dots, z_p$ étant celui du domaine dont toutes les coordonnées sont les plus petites; appelons **C** l'ensemble de ces points. la limite en loi de l'intégrale

$$n \int_{R^p} C_n^2(x_1,\dots, x_p)\, dF(x_1) \dots dF(x_p)$$

sera donc la même que celle de la somme Riemann:

$$n \; \Sigma_C \; C_n^2 \; (z_1,..., z_p) \; (1/k^p)$$

4º) Les $\sqrt{n} \; [\; \Lambda \; H_n(z_1,..., z_p) - (1/k^p) \;]$ forment un ensemble de k^p variables aléatoires de valeurs moyennes nulles et dont la loi tend vers une loi normale à k^p dimensions dont la matrice de covariance est :

$$(1/k^p) \; I_{k^p} - (1/k^{2p}) \; M_{k^p} = \; \otimes_i \; (1/k) \; I_k^{(i)} - \; \otimes_i \; (1/k^2) \; M_k^{(i)}; \; i=1,...,p$$

quand n tend vers l'infini.

Si l'on appelle $X_{1,...,p}$ un vecteur à p indices dont les k^p coordonnées sont $\sqrt{n} \; [\; \Lambda \; H_n(z_1,..., z_p) - 1/k^p \;]$ rangées dans l'ordre lexicographique on voit que les différentes valeurs de :

$\sqrt{n} \; [\; H_n(z_1,..., z_p) - F_1(z_1) \; ... \; F_p(z_p) \;]$ sont les coordonnées du vecteur :

$$\otimes_i \; A_k^{(i)} \; X_{1,...,p} \; ; \; i=1,...,p$$

De même les valeurs de:

$\sqrt{n} \; [\; H_n \; (+ \infty, z_2,..., z_p) - F_2(z_2) \; ... \; F_p(z_p) \;]$ seront les coordonnées de

$$M_k^{(1)} \; \otimes_i \; A_k^{(i)} \; X_{1,...,p} \; ; \; i=2,...,p$$

et les valeurs de:

$$F_1(z_1) \; \sqrt{n} \; \; [\; H_n \; (+ \infty, z_2,..., z_p) - F_2(z_2) \; ... \; F_p(z_p) \;]$$

sont les coordonnées du vecteur :

$$[\; (1/k) \; A_k^{(1)} \; \otimes_i \; I_k^{(i)} \;] \; [\; M_k^{(1)} \; \otimes_i \; A_k^{(i)} \;] \; X_{1,...,p} = (1/k \;) \; A_k^{(1)} \; M_k^{(1)} \; \otimes_i A_k^{(i)} \; ; \; i=2,...,p$$

On voit aisément qu'il en est de même pour tous les termes de C_n. Pour obtenir la matrice donnant par multiplication de $X_{1,...,p}$ le terme dans lequel plusieurs coordonnées sont remplacées par $+ \infty$ il suffit de remplacer dans le produit $\otimes_i A_k^{(i)}$; . $i=1,...,p$ les $A_k^{(i)}$ correspondants par $(1/k \;) A_k^{(i)} \; M_k^{(i)}$; finalement on aura

$\sqrt{n}\ \{\ H_n(z_1,...,\ z_p)\ -\ F_1(z_1)...F_p(z_p)$

$+(-1)^1\ \Sigma^{(1)}\ F_i(z_i)\ [\ H_n(z_1,...,\ z_i=+\infty,...,\ z_p)\ -\ F_1(z_1)\ ...\ F_{i-1}(z_{i-1})\ F_{i+1}(z_{i+1})...F_p(z_p)\ ]$

$+(-1)^2\ \Sigma^{(2)}\ F_i(z_i)\ F_j(z_j)[\ H_n(z_1...,z_i=+\infty,...,z_j=+\infty,...,z_p)\ -\ F_1(z_1)...F_{i-1}(z_{i-1})F_{i+1}(z_{i+1})$

$...\ F_{j-1}(z_{j-1})\ F_{j+1}(z_{j+1})...\ F_p(z_p)] + ...\ \}$

$=\ \otimes_i\ [\ A_k^{(i)} - (1/k)\ A_k^{(i)}\ M_k^{(i)}\ ]\ X_{1,...,p} =\ \otimes_i A_k^{(i)}\ [\ I_k^{(i)} - (1/k)\ M_k^{(i)}\ ]\ X_{1,...,p}$

Et la forme quadratique $n\ \Sigma_C\ C_n^2(z_1,...,z_p)$ pourra s'écrire

${}^tX_{1,..,p}\ \{\ \otimes_i [\ I_k^{(i)} - (1/k)\ M_k^{(i)}\ ]\ \ {}^tA_k^{(i)}\ \ \otimes_i A_k^{(i)}\ [\ I_k^{(i)} - (1/k)\ M_k^{(i)}\ ]\ \}\ X_{1,...,p}\ ;\ i=1,...,p$

Le vecteur $X_{1,...,p}$ a une valeur moyenne nulle et une matrice de covariance égale à $\otimes_i\ (1/k)\ I_k^{(i)} - \otimes_i\ (1/k^2)\ M_k^{(i)}$; i=1,...,p . On aura donc suivant le résultat bien connu:

$$\lim_{n\to\infty} E\ [\ \exp iu\ n\ \Sigma_C\ C_n^2(z_1,...,z_p)\ 1/k^p] =$$

$$\det\ [\ \otimes_i\ I_k^{(i)} - (2\ iu\ /k^p)\ Q\ [\ \otimes_i\ (1/k)\ I_k^{(i)} - \otimes_i\ (1/k^2)\ M_k^{(i)}\ ]\ ]^{-1/2}\ ;\ i=1,...,p$$

en appelant Q la matrice figurant entre les deux accolades.

Comme $(\ (1/k)\ M_k)^2 = (1/k\ )\ M_k$ et comme det [I - RS] = det [I - SR] on voit aisément que :

$$\lim_{n\to\infty} E\ [\ \exp iu\ n\ \Sigma_C\ C_n^2(z_1,...,z_p)\ 1/k^p] =$$

$$\{\ \det\ [\ \otimes_i\ I_k^{(i)} - 2\ iu\ \otimes_i (1/k^{2p})\ A_k^{(i)}\ [\ I_k^{(i)} - (1/k)\ M_k^{(i)}\ ]\ \ {}^tA_k^{(i)}\ ]\ \}^{-1/2}$$

La matrice $(1/k^{2p})\ A_k\ {}^tA_k\ -\ (1/k^{3p})\ A_k M_k\ {}^tA_k$ s'écrit

$(\ (1/k)\ \min\ (i/k,\ j/k) - 1/k\ i/k\ j/k)\ )$

Ses valeurs propres tendent quand k augmente indéfiniment vers les valeurs propres de l'équation intégrale (classique dans l'étude du mouvement brownien) de noyau $\min(x,y) - xy$ $(0 \le x \le 1,\ 0 \le y \le 1)$. Ces valeurs propres sont $k^2\pi^2$ (k=1,2,...) Il en résulte que l'on a quand k augmente indéfiniment :

$$\lim_{n\to\infty} E\,[\exp iu\, G_n] = \prod_{k_1,\ldots,k_p\,=\,1}^{\infty} [\,1 - (2\,iu / k_1^2 \ldots k_p^2\,\pi^{2p})\,]^{-1/2}$$

Il reste et c'est un problème difficile et important en statistique à connaître la vitesse de convergence avec laquelle est atteinte la limite.

Sans aucune démonstration j'avais donné le théorème dans la note: Sur des tests d'indépendance "indépendants de la loi" (Comptes Rendus de l'Académie des Sciences Serie A t. 281 (22 décembre 1975)).

BIBLIOGRAPHIE

HOEFFDING, W. (1948). A non-parametric test of independence. *Ann. Math. Stat.* **19**, 546-557.