Sur l'analyse factorielle des proximités (suite)

J.-P. Benzecri

Publi. Inst. Statist. Univ. Paris
14 (1965), 65-80

SUR L'ANALYSE FACTORIELLE DES PROXIMITÉS (suite) [1]

BENZECRI J.-P.

1. Notation.

Une mesure de probabilité sur l'axe réel sera notée $f_i(x)\,dx$, qu'elle ait ou non une densité ; la lettre f sera employée pour toute mesure, l'indice i sera une expression quelconque désignant une mesure particulière ; on posera

$$P_i(x) = \int_{-\infty}^{x} f_i(t)\,dt \quad ; \quad Q_i(x) = 1 - P_i(x),$$

et l'on notera $S_i(y)$ la fonction, inverse de P_i, définie pour $y \in (0,1)$. Si la mesure $f_i(x)\,dx$ comporte des masses ponctuelles (resp. laisse des intervalles de R vides de masse), la fonction croissante P_i (resp. S_i) a des discontinuités ; nous conviendrons qu'en un point de discontinuité, P_i (ou S_i) prend toutes les valeurs comprises entre la limite à gauche et la limite à droite ; une égalité, e. g.

$$a < P_i(x) < b,$$

s'interprètera : toute valeur prise par P_i en x est comprise entre a et b.

Soit $X \in R^N$: $X = (x_1, \ldots, x_i, \ldots, x_N)$; on notera $f_X(x)\,dx$ la mesure de probabilité définie par N masses ponctuelles, éga-

---

(1) Nous démontrons ici les théorèmes énoncés au § III de l'article de même titre publié dans cette revue (Cf. Benzécri (J.-P.) : Sur l'analyse factorielle des proximités, *Publ. Inst. Statist. Univ. Paris 13* (1964), 235,282).

les à $1/N$, placées aux points d'abcisse x_i ; la suite de ces abcisses, ordonnées dans l'ordre croissant, sera écrite

$$x^1 < \ldots < x^2 < \ldots < x^N, \text{ ou } X(1) < \ldots < X(r) < \ldots < X(N).$$

Soit $\nu_{p\sigma}(X)\,dX$: la distribution normale suivante sur E^p

$$\nu_{p\sigma}(X)\,dX = (2\pi)^{-p/2} \cdot \sigma^{-p} \cdot e^{-|x|^2/2\sigma^2} \cdot dX,$$

où dX est l'élément de volume, et $|X|$ la norme ordinaire.

Le nombre aléatoire réel positif $|X|$ a une distribution de probabilité, (qu'on peut appeler "loi de χ à p paramètres", vu qu'on appelle classiquement loi de χ^2, la distribution de $|X|^2$), dont nous noterons $f_{p\sigma}(x)$ la densité

si $x > 0$: $f_{p\sigma}(x) = A(p) \cdot \sigma^{-p} \cdot e^{-x^2/2\sigma^2} \cdot x^{p-1}$

si $x < 0$: $f_{p\sigma}(x) = 0$,

où $$A(p) = \frac{2^{(2-p)/2}}{\Gamma(p/2)}.$$

Etant donnée une mesure de probabilité $f_i(x)\,dx$, et $a \in R^+$ on définit une nouvelle mesure $f_{ia}(x)\,dx$ par la formule

$x \in (-a, a)$: $P_{ia}(x) = P_i(x)$

$x < -a$: $P_{ia}(x) = 0$

$x > a$: $P_{ia}(x) = 1$.

Autrement dit, $f_{ia}(x)\,dx$ se déduit de $f_i(x)\,dx$ par concentration aux points d'abcisse a (resp. $-a$) de la masse contenue dans l'intervalle (a, ∞) (resp. $(-\infty\ -a)$).

En vue de comparer la mesure $f_X(x)\,dx$ (associée à un système fini de nombres, X ;) à une loi $f_i(x)\,dx$ posons la

Définition 1 : La distance latérale $d(f_i, f_j)$ de deux mesures de probabilités $f_i(x)\,dx$ et $f_j(x)\,dx$ est

$$d(f_i, f_j) = \sup_{y \in (0,1)} |S_i(y) - S_j(y)| ;$$

On peut encore écrire

$$d(f_i, f_j) = \inf_{d \in D} d \ ; \text{ avec}$$

$$D = \{d | (x - x' > d) \Longrightarrow [(P_i(x) \geq P_j(x')) \wedge (P_j(x) \geq P_i(x'))]\}.$$

On vérifie que la distance latérale satisfait aux axiomes d'une distance. On a deux lemmes

Lemme 1 : *Soit* $X \in R^N$;

$$X = (x_1, \ldots, x_N) \ ; \{x_i\} = \{x^r\} \ ; \ x^1 < \ldots < x^r < \ldots < x^N.$$

On a quelle que soit la mesure $f_i(x)\,dx$

$$\forall r = 1, \ldots, N \ ; \forall y \in \left(\frac{r-1}{N}, \frac{r}{N}\right) : |x^r - S_i(y)| \leq d(f_\chi, f_i).$$

Lemme 2 : *Soit* f_i, f_j *deux mesures de probabilités sur* R ; $\{x_h\}$ *une progression arithmétique de raison inférieure ou égale à* d, *comprenant un nombre fini de points dont le plus à gauche est* -a, *le plus à droite* a. *Sous les hypothèses*

$$\forall x_h : P_j(x_h - d) \leq P_i(x_h) \leq P_j(x_h + d) \ ;$$

$$P_i(a + d) = 1 \ ; \ P_i(- a - d) = 0,$$

on a

$$d(f_i, f_{ja}) \leq 2d.$$

(Notons que si le support de f_i et de f_j est le demi-axe positif, il suffit de prendre les x_h en progression de 0 à a).

2. Pseudo-échantillons.

Définition 2 : Soit X un point aléatoire de R^N, de distribution de probabilité $\mu(X)\,dX$; α, un indice aléatoire entier prenant de façon équiprobable les valeurs de 1 à N ; on suppose que α et X sont des variables aléatoires indépendantes. La distribution de probabilité du nombre aléatoire réel : α-ème coordonnée de $X = x_\alpha$, sera appelée distribution déduite de μ, et notée f_μ. On dira que X est un pseudo-échantillon d'effectif N, de la distribution f_μ.
f_μ peut encore être défini ainsi : $P_\mu(x)$ est l'espérance mathématique de la grandeur aléatoire $P_\chi(x)$ (l'espérance mathématique du nombre divisé par N, des coordonnées de X inférieures à x).

Si les N coordonnées de X sont des grandeurs aléatoires, indépendantes ou non, ayant toutes la même distribution, cette distribution est f_μ ; si ces coordonnées sont indépendantes, X est un échantillon, d'effectif N, de f_μ.

Les problèmes que nous étudierons, rentrent sous l'énoncé général suivant.

Problème : Comparer la distribution aléatoire f_x, à la distribution f_μ. Supposons que avec une probabilité supérieure à $1 - \eta$, soit inférieure à ε de la distance latérale

$$d(f_{\mu a}, f_x),$$

$\left(\text{où } a = S_\mu \left(1 - \frac{1}{2N}\right).\right)$; on aura, d'après le Lemme 1, avec une probabilité supérieure à $1 - \eta$

$$\forall r = 1, \ldots, N : \left| x^r - S_\mu \left(\frac{2r - 1}{2N}\right)\right| < \varepsilon,$$

autrement dit, les coordonnées de X seront, presque surement, toutes connues à un ε près, d'après leur rang r.

Pour démontrer l'énoncé

$$\{\text{Prob. } (d(f_{\mu a}, f_x) > \varepsilon) < \eta\},$$

il suffit en notant (cf. lemme 2) $\{x_h\}$ une progression arithmétique de raison $< \varepsilon/2$, allant de $-a$ à $+a$, que l'on majore par η la somme des probabilités suivantes

$$\text{Prob. } \left(P_x \left(a + \frac{\varepsilon}{2}\right) \neq 1\right)$$

$$\text{Prob. } \left(P_x \left(-a - \frac{\varepsilon}{2}\right) \neq 0\right)$$

$$\sum_h \text{Prob. } \left(P_x \left(x_h + \frac{\varepsilon}{2}\right) < P_\mu(x_h)\right)$$

$$\sum_h \text{Prob. } \left(P_x \left(x_h - \frac{\varepsilon}{2}\right) > P_\mu(x_h)\right),$$

ces dernières sommes étant étendues à toute la progression arithmétique des x_h.

Dans la suite nous étudierons le problème posé dans ce n° d'abord pour un échantillon de loi de χ, puis pour un pseudo-échantillon d'une telle loi, construit à partir d'un échantillon d'une loi normale spatiale.

3. Etude d'un échantillon de loi de χ.

Proposition 1 : *Notons* $\mu(p, N)(X)\, dX$, *la distribution d'un échantillon d'effectif* N *de la distribution* f_{p1} ; *on a*

$$f_{\mu(p,N)} = f_{p1}.$$

Quels que soient p *entier positif, et* η, ε *réels positifs, on peut déterminer* $N(p ; \eta ; \varepsilon)$ *tel que, si* $N > N(p ; \eta ; \varepsilon)$ *on ait*

$$\{\text{Prob. } (d(f'_{p1a}, f_\chi) > \varepsilon) < \eta\}$$

où a, *comme ci-dessus, est défini par* $Q_{p1}(a) = 1/(2N)$.

Preuve : Il suffit d'établir que pour N suffisamment grand est inférieure à η la somme des probabilités

$$\text{Prob } (P_\chi(a + \varepsilon/2) \neq 1)$$

$$(2a : \varepsilon) \operatorname*{Max}_{x \in (0,a)} \text{Prob } (P_\chi(x + \varepsilon/2) < P_{p1}(x))$$

$$(2a : \varepsilon) \operatorname*{Max}_{x \in (0,a)} \text{Prob } (P_\chi(x - \varepsilon/2) > P_{p1}(x))$$

où $(2a : \varepsilon)$ est le quotient entier par excès de a par $\varepsilon/2$.

Omettant la majoration de la première de ces probabilités, nous démontrerons la proposition par le

Lemme 3 : *Quand* N *tend vers l'*∞, *tendent vers 0 les expressions*

$$\frac{a}{\varepsilon} \operatorname*{Max}_{x \in (0,a)} \text{Prob } (Q_\chi(x + \varepsilon) > Q_{p1}(x))$$

$$\frac{a}{\varepsilon} \operatorname*{Max}_{x \in (0,a)} \text{Prob } (Q_\chi(x - \varepsilon) < Q_{p1}(x).$$

dans la démonstration du lemme nous n'écrirons pas l'indice p1.

Pour x fixé, $Q_\chi(x)$ a une distribution binomiale, de moyenne $Q(x)$ et de variance

$$\sigma^2 = \frac{1}{N} P_x(x)\ Q_x(x) < \frac{1}{N} Q_x(x)$$

On a donc, d'après l'inégalité de Bienaymé

$$\text{Prob}\ (Q_x(x+\varepsilon) > Q(x)) < \frac{Q(x+\varepsilon)}{N(Q(x) - Q(x+\varepsilon))^2} = \frac{1}{N}\ \pi\ (x\ ;\ \varepsilon\ ;\ +)$$

$$\text{Prob}\ (Q_x(x+\varepsilon) < Q(x)) < \frac{Q(x-\varepsilon)}{N(Q(x-\varepsilon) - Q(x))^2} = \frac{1}{N}\ \pi\ (x\ ;\ \varepsilon\ ;\ -)$$

Pour majorer les π nous ferons usage du

Lemme 4 : *Quand* x *tend vers l'infini, tend vers* 1 *le rapport de* $Q(x)$ *à* $A(p) \cdot e^{-x^2/2} \cdot x^{p-2}$.

Avant de considérer la preuve du lemme 4, nous en déduirons le lemme 3, donc la proposition,

Il résulte immédiatement du lemme 4 que tendent vers l'infini avec x les deux nombres $\pi(x\ ;\ \varepsilon\ ;\ +)$ et $\pi(x\ ;\ \varepsilon\ ;\ -)$. Pour prouver le lemme 3 il suffit donc d'établir que quand N, donc a, tendent vers l'infini, tendent vers zéro les deux nombres

$$\frac{a}{N\varepsilon}\ \pi(a\ ;\ \varepsilon\ ;\ +) \quad \text{et} \quad \frac{a}{N\varepsilon}\ \pi(a\ ;\ \varepsilon\ ;\ -).$$

On a d'après le lemme 4, $\left(\text{en remplaçant } \frac{1}{N} \text{ par } 2\,Q(a)\right)$

$$\frac{a}{N\varepsilon}\ \pi(a\ ;\ \varepsilon\ ;\ +) \sim \frac{2a}{\varepsilon} Q(a)\ \frac{Q(a+\varepsilon)}{(Q(a))^2} \sim \frac{2a}{\varepsilon}\ e^{-a\varepsilon-(\varepsilon^2/2)}$$

$$\frac{a}{N\varepsilon}\ \pi(a\ ;\ \varepsilon\ ;\ -) \sim \frac{2a}{\varepsilon}\ Q(a)\ \frac{Q(a-\varepsilon)}{(Q(a-\varepsilon))^2} \sim \frac{2a}{\varepsilon}\ e^{-a\varepsilon+(\varepsilon^2/2)}$$

d'où les limites souhaitées et la preuve du lemme 3 donc de la proposition.

Pour prouver le lemme 4, on place la courbe représentative de $f_{p1}(x)$ entre une droite tangente et une exponentielle négative : pour x_o suffisamment grand, et $x > x_o$ on a les inégalités

$$f_{p1}(x)\left(1 + (x - x_o)\left(-x_o + \frac{p-1}{x_o}\right)\right) < f_{p1}(x)$$

$$f_{p1}(x) < f_{p1}(x_o)\ e^{-x_o(x-x_o)/2}.$$

Ceci prouve que $Q_{p1}(x_o)$ est de l'ordre de $\frac{1}{x_o} f_{p1}(x)$ et de plus qu'une fraction arbitrairement élevée donnée de la masse de la mesure $f_{p1}(x)\,dx$, au delà de x_o, est contenue dans un intervalle $(x_o, x_o + h)$ où h est de l'ordre de $1/x_o$; dans un tel intervalle, on peut, si x_o est grand, assimiler $f_{p1}(x)$ à

$$f_{p1}(x_o)\ .\ e^{-x_o(x-x_o)}$$

d'où, par intégration, le lemme 4.

4. Etude d'un pseudo-échantillon de loi de χ.

Soit l'élément aléatoire : $\mathfrak{X} = (X_1, \ldots, X_i, \ldots, X_n) \in E^{pn}$, échantillon d'effectif n de la loi normale ν_{p1} sur E^p. Chacune des distances

$$d_{ij} = |X_i - X_j|$$

est une variable aléatoire réelle ; qui a pour distribution $f_{p\sqrt{2}}(x)\,dx$. Le système $X(\mathfrak{X})$, noté en abrégé X, de ces $N = n(n-1)/2$ distances d_{ij} est un pseudo-échantillon de la distribution $f_{p\sqrt{2}}(x)\,dx$; nous noterons $\delta_n(X)\,dX$ la distribution de ce pseudo-échantillon ;

$$f_{\delta_n} = f_{p\sqrt{2}}$$

Avec ces notations, on peut énoncer la

Proposition 2 : *Quels que soient* p *entier positif, et* η, ε, *réels positifs, on peut déterminer* $n(p\,;\eta\,;\varepsilon)$ *tel que, si* $n > n(p\,;\eta\,;\varepsilon)$ *on ait, pour le pseudo-échantillon* X *de distribution* $\delta_n(X)\,dX$

$$\{\text{Prob}\,(d(f_{p\sqrt{2}a}, f_X) > \varepsilon) < \eta\},$$

où a, *comme ci-dessus, est défini par* $Q_{p\sqrt{2}}(a) = \frac{1}{2N} = \frac{1}{n(n-1)}$.

Preuve : Procédant comme pour la proposition 1, nous nous bornerons à démontrer le

Lemme 5 : *Quand* n *tend vers l'infini, tendent vers zéro les expressions*

$$\frac{a}{\varepsilon} \underset{x\in(0,a)}{\text{Max}} \text{Prob}\,(Q_x(x+\varepsilon) > Q_{p\sqrt{2}}(x)) = \frac{a}{\varepsilon} \underset{x\in(0,a)}{\text{Max}} \Pr(x\,;+)$$

$$\frac{a}{\varepsilon} \underset{x\in(0,a)}{\text{Max}} \text{Prob}\,(Q_x(x-\varepsilon) < Q_{p\sqrt{2}}(x)) = \frac{a}{\varepsilon} \underset{x\in(0,a)}{\text{Max}} \Pr(x\,;-)$$

Faute d'avoir trouvé comme pour la preuve du lemme 3, des majorations valables sur tout l'intervalle $(0, a)$, nous fragmenterons cet intervalle, ce qui complique la preuve.

a) *Etude du nombre aléatoire* $Q_x(x)$ (*pour* x *fixé*) ; ce nombre est positif, compris entre 0 et 1, et a pour moyenne

$$Q_{p\sqrt{2}}(x) = Q_{p1}(x/\sqrt{2}) \sim A(p)\, 2^{(2-p)/2}\, x^{p-2}\, e^{-x^2/4}.$$

Pour évaluer la variance de la distribution de $Q_x(x)$, nous exprimerons ce nombre aléatoire comme une moyenne de N nombres

$$Q_x(x) = \frac{1}{N} \sum_{(i,j)} Q_{x_{ij}}(x),$$

où chacun des N termes $Q_{x_{ij}}(x)$ est ainsi défini

$$Q_{x_{ij}}(x) = 0 \quad \text{si } d_{ij} = |X_i - X_j| < x,$$

$$Q_{x_{ij}}(x) = 1 \quad \text{si } d_{ij} = |X_i - X_k| > x.$$

Si i, j, k, l sont quatre entiers distincts, d_{ij} et d_{kl}, donc aussi $Q_{x_{ij}}(x)$ et $Q_{x_{kl}}(x)$, sont des grandeurs aléatoires indépendantes. La variance $\sigma^2(N\,;x)$ de $Q_x(x)$ est l'espérance mathématique de

$$\frac{1}{N^2}\left[\sum_{ij} (Q_{x_{ij}}(x) - Q_{p\sqrt{2}}(x))\right]^2 ;$$

Si l'on développe en N^2 termes, le carré de cette somme de N termes, on trouve qu'il n'y a que les N termes carrés, et $n(n-1)(n-2)$ termes rectangles, qui soient non-nuls ; on a

$$\sigma^2(N\,;x) = \frac{1}{N} Q_{p\sqrt{2}}(x)\, P_{p\sqrt{2}}(x) + \frac{n(x-1)(x-2)}{N^2} (Q^{(2)}_{p\sqrt{2}}(x) - (Q_{p\sqrt{2}}(x))^2),$$

où $Q^{(2)}_{p\sqrt{2}}(x)$ est la probabilité que l'on ait

$$Q_{x_{12}}(x) = Q_{x_{13}}(x) = 1 \text{ i.e } ; d_{12} > x \; ; d_{13} > x.$$

On peut majorer $Q^{(2)}_{p\sqrt{2}}(x)$ par $Q_{p\sqrt{2}}(x)$; on peut encore montrer que $Q^{(2)}_{p\sqrt{2}}(x)$ admet un équivalent produit d'un monôme en x par $e^{-x^2/3}$. Par de telles évaluations, nous n'avons pas su simplifier la preuve du lemme 5. Nous utilisons seulement l'inégalité

$$\sigma^2(N\ ;\ x) \leq \frac{1}{N} + \frac{n(n-1)(n-2)}{N^2} < \frac{4}{n}\ .$$

b) *Une majoration déduite de l'espérance mathématique de* $Q_Y(x)$.

C'est le

Lemme 6 : *Pour* n, *donc* a, *tendant vers l'infini, tend vers zéro le produit*

$$\frac{a}{\varepsilon} \operatorname*{Max}_{x\in\left(\frac{a}{10},a\right)} \text{Prob}\ (Q_x(x+\varepsilon) > Q_{p\sqrt{2}}(x)).$$

Preuve : $Q_x(x+\varepsilon)$ a pour espérance mathématique $Q_{p\sqrt{2}}(x+\varepsilon)$ d'où (Tchebichev, Bienaymé) la majoration

$$\Pr(x\ ;\ +) < \frac{Q_{p\sqrt{2}}(x+\varepsilon)}{Q_{p\sqrt{2}}(x)} \sim e^{-(2\varepsilon x+\varepsilon^2)/4}.$$

Le Lemme 6 en résulte facilement.

c) *Majorations déduites de la variance de* $Q_Y(x)$. C'est le

Lemme 7 : *Pour* n, *donc* a, *tendant vers l'infini tendent vers zéro les produits*

$$\frac{a}{\varepsilon} \operatorname*{Max}_{x\in\left(0,\frac{a}{10}\right)} \Pr(x\ ;\ +)$$

$$\frac{a}{\varepsilon} \operatorname*{Max}_{x\in\left(0,\frac{a}{10}\right)} \Pr(x\ ;\ -)$$

Preuve : On a les majorations

$$\Pr(x\ ;\ +) < \frac{\sigma^2(N\ ;\ x)}{[Q_{p\sqrt{2}}(x+\varepsilon) - Q_{p\sqrt{2}}(x)]^2} < \frac{4}{n[Q_{p\sqrt{2}}(x+\varepsilon) - Q_{p\sqrt{2}}(x)]^2}$$

On peut trouver x_1, tel que si $x > x_1$ on ait

$$[Q_{p\sqrt{2}}(x+\varepsilon) - Q_{p\sqrt{2}}(x)]^2 > B(p)\ e^{-x^2/4}$$

où B(P) est un nombre qui ne dépend que de p (se référer à l'expression asymptotique de $Q_{p\sqrt{2}}(x)$). On en déduit facilement que tend vers zéro, quand a tend vers l'infini, le produit

$$\frac{a}{\varepsilon} \max_{x \in (x_1, \frac{a}{10})} P_r(x\ ;\ +) < \frac{a}{\varepsilon^n} \frac{1}{B(p)} e^{a^2/400}$$

Car

$$\frac{1}{2N} = Q_{p\sqrt{2}}(a) \sim A\ 2^{(2-p)/2}\ a^{p-2}\ e^{-a^2/4}, \text{ et}$$

$$\frac{1}{n} \sim \frac{1}{\sqrt{2N}}.$$

Sur l'intervalle fixe $(0, x_1)$, $\frac{a}{\varepsilon} Pr(x\ ;\ +)$ tend vers zéro comme a/n. D'où, la partie du Lemme relative à $Pr(x\ ;\ +)$. L'on procèderait de même pour $Pr(x\ ;\ -)$.

d) *Retour à la Définition du pseudo-échantillon.* Après les Lemmes 6 et 7, il ne manque plus à la preuve du lemme 5, donc de la proposition 2 que le

Lemme 8 : *Quand* a *tend vers l'infini, tend vers zéro le produit*

$$\frac{a}{\varepsilon} \max_{x \in (\frac{a}{10}, a)} Pr(x\ ;\ -).$$

Pour simplifier les notations de la preuve, nous démontrerons l'énoncé suivant

Lemme 8' : *Quand* a, *(donc* n), *tend vers l'infini, tend vers zéro le produit*

$$\frac{a}{\varepsilon} \max_{x \in (\frac{a}{20}, a-\varepsilon)} \text{Prob}\ (Q_x(x - \varepsilon) < Q_{p\sqrt{2}}(x + \varepsilon)).$$

Pour prouver le Lemme, il faut donner une minoration presque sûre du nombre des distances d_{ij} supérieures à $x - \varepsilon$. Pour cela, revenons à l'échantillon $\mathfrak{X}$ d'effectif n, utilisé pour construire le pseudo-échantillon X d'effectif N. Pour minorer l'ensemble des $d_{ij} > x - \varepsilon$, nous ne considérerons que les points X_i situés dans R^p au-delà de la sphère centrée à l'origine de rayon $x/2$. Pour que deux tels points X_i, X_j soient à une distance supérieure à $x - \varepsilon$,

on a la condition suffisante suivante : Soit N_i (resp N_j) le point d'intersection OX_i (resp OX_j) avec la sphère de rayon $x/2$; N'_j le point diamétralement opposé à N_j ; pour que $|N_i N_j|$, donc $|X_i X_j| = d_{ij}$, soit supérieure à $x - \varepsilon$, il suffit que N_i et N'_j se trouvent sur une même calotte sphérique de rayon $\sqrt{x\,\varepsilon/2}$ de la sphère de rayon $x/2$. D'où l'utilité du

Lemme 9 : *Soit* $\Sigma(x)$ *la sphère de centre* O *et de diamètre* x *dans* E^p

$$\Sigma(x) = \{u \mid u \in E^p \ ; \ |u| = x/2\}.$$

On peut trouver une constante $K(p)$ *telle que, pour* $x/\varepsilon > 10$, *il existe sur* $\Sigma(x)$ *une famille* $\{C_\beta\}_{\beta \in B}$ *de calottes possédant les propriétés suivantes*

- *toute calotte* C_β *est de rayon* $\sqrt{x\,\varepsilon/2}$.

- *deux calottes distinctes sont d'intersection vide.*

- *si* C_β *appartient à la famille, il en est de même de sa symétrique par rapport à* O, *notée* $C_{\beta'}$.

- *l'effectif de la famille est compris entre les deux bornes*

$$K(p)\left(\frac{x}{\varepsilon}\right)^{(p-1)/2} \quad \text{et} \quad \frac{1}{K(p)}\left(\frac{x}{\varepsilon}\right)^{(p-1)/2}.$$

- *l'aire de la réunion des calottes est une fraction supérieure à* $K(p)$ *de l'aire de* $\Sigma(x)$.

Pour $p = 1$ ou 2, la preuve est claire : si $p = 1$, $\Sigma(x)$ se compose de deux points qui sont aussi les deux calottes de la famille ; si $p = 2$, $\Sigma(x)$ est un cercle, que l'on partage en un nombre pair d'arcs égaux. Pour étudier le cas général, on remarque que l'aire de $\Sigma(x)$ est de l'ordre de x^{p-1}, tandis que l'aire d'une des calottes est de l'ordre de $x\,\varepsilon^{(p-1)/2}$.

Soit sur la sphère $\Sigma(x)$ une famille $\{C_\beta\}_{\beta \in B}$ de calottes satisfaisant aux conditions du Lemme, notons G_β l'ensemble des points de E^p situés à une distance de O supérieure à $x/2$, sur une demi-droite issue de l'origine rencontrant $\Sigma(x)$ en un point de C_β. L'espérance mathématique du nombre des points de l'échantillon $\mathfrak{X}$ situés dans G_i est minorée par

$$n \,.\, (K(p))^2 \,.\, f_{p1}(x/2) \,.\, \left(\frac{\varepsilon}{x}\right)^{(p-1)/2} = n\ \varphi(x, \varepsilon).$$

Nous montrerons que pour $x \in (a/20,\ a - \varepsilon)$, l'on a presque sûrement, dans chaque $G_{\beta'}(n/2)\ \varphi(n, \varepsilon)$ points de l'échantillon $\mathfrak{X}$, et que la considération des couples de points dans G_β, $G_{\beta'}$ suffit à assurer qu'alors $Q_x(x - \varepsilon) < Q_{p\sqrt{2}}(x + \varepsilon)$.

Notons $e(\beta\ ; \mathfrak{X})$ l'effectif des points de $\mathfrak{X}$ situés dans G_β; on a le

Lemme 10 : *Quand* a *et* n *tendent vers l'infini, tend vers zéro le produit*

$$\frac{a}{\varepsilon} \max_{x \in (\frac{a}{20},\, a-\varepsilon)} \text{Prob}\left\{\exists \beta \in B : e(\beta, \mathfrak{X}) < \frac{n}{2}\ \varphi(x, \varepsilon)\right\}$$

Preuve du Lemme 10 : d'après l'inégalité de Bienaymé, on a pour un β donné

$$\text{Prob}\left\{e(\beta) < \frac{n}{2}\ \varphi(x, \varepsilon)\right\} < \frac{4}{n\varphi(x, \varepsilon)}\ , \text{ d'où}$$

$$\text{Prob}\left\{\exists_{\beta \in} B : e(\beta, \mathfrak{X}) < \frac{n}{2}\ \varphi(x, \varepsilon)\right\} < \frac{4}{K(\rho)\ n\varphi(x, \varepsilon)} \left(\frac{x}{\varepsilon}\right)^{(p-1)/2}.$$

Du membre de droite de cette inégalité, nous possédons une expression asymptotique, donc valable pour $x > \frac{a}{20}$, si a est assez grand. Cette expression asymptotique comporte des termes monomes en x et a, termes que nous n'écrirons pas, les termes exponentiels donnant la limite cherchée

$$\frac{1}{\varphi(x, \varepsilon)} \sim \text{monome}\ .\ e^{-x^2/8}\ ;\ \frac{1}{n} \sim \text{monome}\ .\ e^{-a^2/8}$$

d'où pour $1/n\ .\ \varphi(x, \varepsilon)$ si $x \leq a - \varepsilon$

$$\frac{1}{n\ .\ \varphi(x, \varepsilon)} < \text{monome}.\, e^{-(a^2 - x^2)/8} < e^{-a\varepsilon/8}.\ \text{monome}.$$

D'où le Lemme 10. La démonstration de la proposition 2 va s'achever par le

Lemme 11 : *pour* a *suffisamment grand on a l'implication*

$$\left\{(x \in (a/20,\ a - \varepsilon)) \wedge \left(\forall \beta \in B : e(\beta, \mathfrak{X}) > \frac{n}{2}\ \varphi(x, \varepsilon)\right)\right\}$$
$$\Longrightarrow \left\{Q_x(x - \varepsilon) > Q_{p\sqrt{2}}(x + \varepsilon)\right\}$$

Preuve : On peut minorer $N.Q_x(x - \varepsilon)$, nombre des $d_{ij} > x - \varepsilon$, par la somme des produits $e(\beta, \mathfrak{X}) . e(\beta', \mathfrak{X})$ des effectifs de $\mathfrak{X}$ dans les secteurs symétriques G_β, $G_{\beta'}$, d'où, (en minorant le nombre des couples G_β, $G_{\beta'}$, cf. Lemme 9) l'inégalité

$$N . Q_x(x - \varepsilon) > \frac{1}{2} . K(\rho) . \left(\frac{x}{\varepsilon}\right)^{(p-1)/2} \frac{n^2}{4} (\varphi(x, \varepsilon))^2.$$

Comme dans le Lemme 10, on écrira en utilisant les développements asymptotiques sans préciser les termes monomes

$$Q_x(x - \varepsilon) > \text{monome } e^{-x^2/8}$$

Or

$$Q_{p\sqrt{2}}(x + \varepsilon) \sim e^{-(x+\varepsilon)^2/8} . \text{ monome.}$$

D'où le Lemme 11, les Lemmes 8 et 5 et la proposition 2.

5. Preuve des théorèmes du § III.

Etant donné un point (S ou $\mathfrak{X}$ ...) de E^{pn}

$$(X_1, \ldots, X_i, \ldots, X_n) \; ; \; X_i \in E^p,$$

On note X(S), $X(\mathfrak{X})$ etc. le système des $N = n(n - 1)/2$ distances

$$|X_i - X_j| \in R^N.$$

On note G(S), $G(\mathfrak{X})$, le point de E^{pn} ainsi défini

si $\mathfrak{X} = (X_i)$; $X_G = \frac{1}{n} \sum_i X_i$; $G(\mathfrak{X}) = ((X_i - X_G))$.

Il est clair que l'on a

$$\forall \mathfrak{X} : G(G(\mathfrak{X})) = G(\mathfrak{X}); \forall \mathfrak{X} : X(G(\mathfrak{X})) = X(\mathfrak{X}),$$

et que l'équation

$$G(\mathfrak{X}) = \mathfrak{X}$$

définit un sous-espace L de dimension $(n - 1) p$ de E^{pn}.

Dans ce n°, S est l'élément aléatoire défini au § III n° 5b ; la distribution de S est invariante par rotation sur la sphère de $E^{p(n)}$ ainsi définie

$$G(S) = S \; ; \; |S|^2 = np$$

Comme au n° 4, $\mathfrak{X}$ est l'élément aléatoire normal sphérique de support E^{pn} ; $G(\mathfrak{X})$ est normal sphérique dans le sous-espace linéaire L ; en particulier,

$$|G(\mathfrak{X})|$$

suit une loi de χ à $(n-1)p$ paramètres ; et si on impose la condition : $|G(\mathfrak{X})| = \rho$, la distribution conditionnelle pour $G(\mathfrak{X})$ est homothétique, dans le rapport $\rho/\sqrt{np}$ de celle de l'élément aléatoire S.

Etant donné un $\mathfrak{X}$ (ou $S) \in R^{pn}$, d'ordonnance ,

$$\omega(i,j) = \omega(\mathfrak{X})\,(i,j)$$

nous noterons

$$D(\varepsilon \; ; \mathfrak{X})$$

la propriété

$$\forall i,j : \; |X_i - X_j| - S_{p\sqrt{2}}\left(\frac{2\omega(i,j) - 1}{2N}\right) < \varepsilon$$

On notera que $D(\varepsilon \; ; \mathfrak{X}) \Longleftrightarrow D(\varepsilon \; ; G(\mathfrak{X}))$.

Ceci dit, le théorème 2 résulte immédiatement de la proposition 2 et de la

Proposition 3 : *Quels que soient* η, ε, p *donnés, il existe* $n'(p \; ; \eta \; ; \varepsilon)$ *tel que, si* $n > n'$ (p, η, ε)

$$(\text{Prob}\,(D(\varepsilon \; ; \mathfrak{X})) > 1 - \eta) \Longrightarrow (\text{Prob}\,(D(2\varepsilon, S)) > 1 - 2\eta).$$

Preuve : n' sera choisie telle que soit satisfaite la condition, (où a est définie par $Q_{p\sqrt{2}}(a) = \frac{1}{2N}$)

$$\text{Prob}\left\{\left|2a\left(\frac{|G(\mathfrak{X})|}{\sqrt{np}} - 1\right)\right| > \varepsilon\right\} < \eta$$

cela est possible vu que $|G(\mathfrak{X})|$ est un χ à $(n-1)p$ paramètres.

On aura pour $n > n'$

$$\text{Prob}\left\{\left\{\left|2a\left(\frac{|G(\mathfrak{X})|}{\sqrt{np}} - 1\right)\right| < \varepsilon\right\} \wedge D(\varepsilon\,;\mathfrak{X})\right\} > 1 - 2\eta$$

Il existe donc un ρ satisfaisant aux deux conditions

$$2a\left(\frac{\rho}{\sqrt{np}} - 1\right) < \varepsilon \quad \text{et}$$

Prob $(D(\varepsilon\,;G(\mathfrak{X}))$, modulo la condition $|G(\mathfrak{X})| = \rho) > 1 - 2\eta$. (se souvenir ici de ce que $D(\varepsilon\,;G(\mathfrak{X})) \Longleftrightarrow D(\varepsilon\,;\mathfrak{X})$).

Or la distribution de $G(\mathfrak{X})$ modulo $|G(\mathfrak{X})| = \rho$ est homothétique de celle de S dans le rapport $\rho/\sqrt{np}$, d'où l'on déduit

$$\text{Prob}\left(D\left(\varepsilon\,;\frac{\rho}{\sqrt{np}}\,S\right)\right) > 1 - 2\eta$$

mais en vertu de la première condition imposée à ρ cela implique

$$\text{Prob}\,(D(2\varepsilon\,;S)) > 1 - 2 \qquad \text{c. q. f. d.}$$

On notera que notre démonstration du théorème 2 par les propositions 2 et 3, laisse dans l'ombre le fait, pourtant clair, que $X(S)$ est mieux connu par son ordonnance $\omega(S)$ que ne l'est $X(\mathfrak{X})$ par $\omega(\mathfrak{X})$, car, dans le second cas, l'incertitude (faible il est vrai pour n grand) sur les dimensions de la figure inconnue s'ajoute à l'incertitude sur sa forme.

Le Théorème 1, résulte facilement du Lemme 2 du § III, n° 4 et du

Lemme 12 : *Soit* $\{X_i\}$ *un échantillon d'effectif* N *de la loi normale*

$$\nu_{p1}(X)\,dX \quad \textit{sur} \quad E^p :$$

a) *Pour* ε *donné et* $N \longrightarrow \infty$, *tend vers zéro la probabilité qu'un seul point de l'échantillon se trouve à une distance de l'origine supérieure à* $a + \varepsilon$ *où* a *est défini par*

$$\frac{1}{2N} = Q_{p1}(a) \sim A(p)\, e^{-a^2/2}\, a^{p-2}\ ;$$

tandis que tend vers l'∞ la densité de l'échantillon à l'intérieur de la sphère de rayon $a - \varepsilon$ *; en particulier*

b) *Soit* α *un nombre réel donné ; notons* $C(\alpha)$ *l'ensemble des pavés cubiques de* E^p, *d'arête* α, *dont les sommets ont pour coordonnées des multiples entiers de* α. *Quand* $a \longrightarrow \infty$, *tend vers un la probabilité que tout cube de* $C(\alpha)$ *situé tout entier à une distance de l'origine moindre que* $(a - \varepsilon)$, *se trouve au moins un point de l'échantillon.*

Preuve : a) résulte de la proposition 1 ; démontrons b). Les cubes envisagés sont en nombre inférieur à $\left(\frac{2a}{\alpha}\right)$. Sur un de ces cubes, l'espérance mathématique du nombre des X_i présents est supérieure à

$$N \alpha^p \, e^{-(a-\varepsilon)^2/2} = Nq \; ;$$

d'où pour l'absence de tout X_i dans un cube donné une probabilité au plus égale à e^{-Nq}. Et pour l'absence de point de l'échantillon dans un au moins des cubes envisagés une probabilité majorée par

$$\left(\frac{2a}{\alpha}\right)^p \, e^{-Nq} = q'$$

N est à un monome près en a, de la forme $e^{a^2/2}$, donc q' tend vers zéro, quand a tend vers l'infini, comme

$$a^p \, e^{-(e^{a\varepsilon})}.$$

Ceci établit le lemme.

Notons enfin que pour p donné, n assez grand, les moments d'inertie principaux du nuage des $\{X_i\}$ dans E^p, sont compris presque sûrement entre an et An (où a, A sont des nombres réels donnés tels que $a < 1 < A$). Ceci permet de justifier par le théorème 1 du §II n° 5 l'application de l'analyse factorielle usuelle à la reconstruction d'une figure donnée par son ordonnance, (les valeurs approchées des distances étant celles fournies par une loi de χ, selon le théorème 2).

IMPRIMERIE LOUIS-JEAN
Ouvrages scientifiques
TYPO-OFFSET
GAP (Hautes-Alpes)
Dépôt légal n° 159
1965

INSTITUT INTERNATIONAL DE STATISTIQUE

L'Institut International de Statistique (IIS) est une association autonome ayant pour but de développer et perfectionner les méthodes statistiques et leur application dans les divers pays du monde.

REVUE DE

L'INSTITUT INTERNATIONAL DE STATISTIQUE

La REVUE DE L'IIS présente:

- des articles scientifiques
- des rapports et informations concernant:
  - les bureaux officiels de statistique
  - les instituts de recherches statistiques
  - les sociétes de statistique
- la Bibliographie Statistique Internationale
- des comptes rendus critiques d'ouvrages statistiques
- le calendrier des réunions internationales

Le prix d'abonnement au volume annuel (trois livraisons) de la REVUE DE L'IIS est de F 41 (port compris)

INSTITUT INTERNATIONAL DE STATISTIQUE
2 Oostduinlaan, La Haye, Pays-Bas.

ARTICLES

VOLUME 32 (1964)

INFORMATIONS STATISTIQUES

Rapports et informations sur l'organisation et les activités des

- bureaux officiels de statistique
- institut de recherches statistiques
- sociétés de statistique.

CALENDRIER DES REUNIONS INTERNATIONALES

ARTICLES

VOLUME 33 (1965)

Récents développements dans le plan d'expérience

Dugué, D. - L'équipe française de chercheurs sur le plan d'expériences

Barra, J.R. - Carrés latins et Eulériens (avec discussion)

Guérin, R. - Vue d'ensembles sur les plans en blocs incomplets équilibrés et partiellement équilibrés

Rényi, A. - On the foundation of information theory (avec discussion)

Freiberger, F., Grenander, U. - On the formation of statistical meteorology

Grenander, U. - Stochastic models in the theory of pursuit

Godambe, V.P. - A review of the contributions towards a unified theory of sampling from finite populations

Puri, M.L. - On the combinations of independent two sample tests of a general class

Tekse, K. - Describing the geographical distribution of the population

Erlander, S., Gustavsson, J. - Simultaneous confidence regions in normal regression analysis with an application to road accidents

Gumbel, E.J. - A short estimation of the parameters in Fréchet's distribution.

Medin, K. - Crop yield estimation and crop insurance in Sweden

Zaremba, S.K. - Kurtosis and determinate components in linear processes

BIBLIOGRAPHIE STATISTIQUE INTERNATIONALE

Une bibliographie selectionnée sur la théorie et les méthodes statistiques, et sur leurs applications dans tous les domaines. Chaque volume annuel de la Revue comprend environ 2000 titres de livres et de communications publiées dans le monde entier.

COMPTES RENDUS

QUELQUES AUTRES PUBLICATIONS DE L'IIS

Bibliographie Statistique Internationale:

Trois livraisons chaque année. (Tirés à part de la Revue de l'IIS imprimés au recto seulement)
Prix du volume annuel, port payé . . . . . F 17.50

Bulletin de l'Institut International de Statistique:

Comptes rendus des sessions biennales de l'IIS.
Publié tous les deux ans.
Prière de se renseigner sur les tomes disponibles et leur prix.

A History of the International Statistical Institute, 1885 - 1960, par J. W. Nixon. . . . . F 12.25

Veuillez envoyer vos commandes et demandes de renseignements à:

L'Institut International de Statistique
2 Oostduinlaan
LA HAYE, Pays-Bas

PUBLIE POUR L'IIS PAR OLIVER & BOYD, LTD

Statistical Theory and Method Abstracts

Quatre livraisons et une table des matières chaque année.
Prix du volume annuel . . . . . £ 7.10.-

Dictionary of Statistical Terms (2ème ed.) par M. G. Kendall et W. R. Buckland. . . . . 32 sh.

Glossary of terms in official statistics . English-French, French-English. J. W. Nixon, Réd.. . . . . 42 sh.

Bibliography of basic texts and monographs on statistical methods par W. R. Buckland et R. A. Fox. . . . . 35 sh.

Veuillez envoyer vos commandes à:
Messrs Oliver & Boyd, Ltd.
Tweeddale Court
EDIMBOURG, Ecosse