REVUE DE STATISTIQUE APPLIQUÉE

G. DER MEGREDITCHIAN

Un test non paramétrique unilatéral de rupture d'homogénéité de « k » échantillons

Revue de statistique appliquée, tome 34, nº 1 (1986), p. 45-60

<http://www.numdam.org/item?id=RSA_1986__34_1_45_0>

UN TEST NON PARAMÉTRIQUE UNILATÉRAL DE RUPTURE D'HOMOGÉNÉITÉ DE « K » ÉCHANTILLONS

G. DER MEGREDITCHIAN
Météorologie Nationale

Dans de nombreuses disciplines scientifiques, et en météorologie en particulier, se pose avec acuité le problème de l'homogénéité des séries statistiques. Un échantillonnage étant effectué successivement au cours du temps, il est primordial souvent de pouvoir affirmer que les conditions de son obtention n'ont pas variées : les conditions climatiques sont-elles les mêmes pour une période donnée ou, au contraire, un fait nouveau est-il venu perturber les populations dont les échantillons sont extraits ? Peut-on affirmer, par exemple, que la loi de distribution de l'ozone a été modifiée par des facteurs exogènes tels que la pollution ambiante ? Les séries des échantillons annuels d'un élément météorologique (température, précipitations etc.) ont-elles évoluées sous l'influence d'une modification des conditions de leur mesure (déplacement du site, changement d'appareils etc.) ? Assiste-t-on à une modification du climat ?

L'outil statistique approprié pour répondre à ce genre de questions est la théorie des tests dits non paramétriques. Formulons maintenant le problème mathématique correspondant.

POSITION DU PROBLÈME

Le problème général de vérification de l'homogénéité de « K » échantillons se formule ainsi. Considérons K populations aléatoires définies par les K fonctions de répartition $F_j(x)$, $j = 1, k$.

On extrait « au hasard» de chaque population un échantillon de taille n_j :

$$x_1[j], \ldots, x_s[j], \ldots, x_{n_j}[j];\ j = 1, k.$$

On dispose ainsi de K échantillons indépendants extraits au hasard, de sorte, que l'on peut introduire les fonctions de répartition empiriques correspondantes :

$$\hat{F}_j[x] = \frac{1}{n_j} \sum_{s=1}^{n_j} \theta\,(x - x_s[j]),$$

où $\theta\,[z]$ désigne la fonction de Heaviside :

$$\theta[z] = 1\,, \quad \text{si} \quad z = 0 \quad \text{et} \quad \theta[z] = 0\,, \quad \text{si} \quad z < 0\,.$$

On cherche à vérifier l'hypothèse :

$$H_0 : F_1[x] = \ldots = F_j[x] = \ldots = F_k[x]\,, \qquad (1)$$

Mots-clés : **Homogénéité, Tests non paramétriques, Echantillon.**

contre l'hypothèse alternative.

$$H_1 : \exists\,(i,j) : F_i(x) \neq F_j(x)\,. \qquad (1')$$

Dans le cas classique de deux échantillons (K = 2) de nombreux tests ont été élaborés. Nous citerons simplement certains d'entre eux : le test des signes, le test de VAN DER VARDEN, le test de WALD-WOLFOWICZ, le test de KHI-DEUX, le test de WILCOXON, le test de Mann-WHITNEY, le test informationnel de KULLBACK et enfin le test de KOLMOGOROV-SMIRNOV. Le test de KOLMOGOROV-SIMIRNOV [18] occupe une place particulière dans la grande famille des tests non paramétriques par les multiples recherches qu'il a suscitées. SMIRNOV a étudié des statistiques basées sur l'écart entre fonction de répartition empiriques :

$$\mathscr{D}^{+}_{n_1,n_2} = \sup_{-\infty<x<+\infty} [\hat{F}_1(x) - \hat{F}_2(x)]; \; \mathscr{D}_{n_1,n_2} = \sup_{-\infty<x<+\infty} |\hat{F}_1(x) - \hat{F}_2(x)| \qquad (2)$$

pour lesquelles il a établi les relations asymptotiques :

$$P\left\{ \mathscr{D}^{+}_{n_1,n_2} < z\sqrt{\frac{n_1+n_2}{n_1 \cdot n_2}} \right\} \to 1 - e^{-2z^2}\,; \qquad (3)$$

$$P\left\{ \mathscr{D}_{n_1,n_2} < z\sqrt{\frac{n_1+n_2}{n_1 \cdot n_2}} \right\} \to K(z) = \sum_{s=-\infty}^{+\infty} (-1)^s\, e^{-2s^2x^2}\,. \qquad (3')$$

GNÉDENKO a introduit [10, 11] une méthode élégante pour étudier les probabilités exactes correspondantes. Il a défini un problème combinatoire donnant la solution exacte et pouvant être interprété soit en termes de cheminement aléatoire, soit en termes de schéma d'urne. Ce problème combinatoire peut être ramené en fait à un problème classique, le problème du scrutin de ANDRÉ 1, connu également sous le nom de lemme de BERTRAND 2.

GNÉDENKO a obtenu ainsi pour le cas des tailles égales des échantillons, $n_1 = n_2 = n$ les formules suivantes :

$$P\left\{ \mathscr{D}^{+}_{nn} < \frac{c}{n} \right\} = 1 - \frac{C^{n-c}_{2n}}{C^{n}_{2n}}\,, \qquad (4)$$

$$P\left\{ \mathscr{D}_{nn} < \frac{c}{n} \right\} = \sum_{k=-(n/c)}^{n/c} (-1)^k \frac{C^{n-kc}_{2n}}{C^{n}_{2n}}\,. \qquad (4')$$

Le cas des tailles non égales ($n_1 \neq n_2$) des échantillons fut étudié par KOROLIOUK [15, 16], STECK [19], CHANG LI CHIEN [20] et donne des formules plus compliquées.

Le cas général de plusieurs échantillons (k > 2) semble plus délicat à aborder. Nous rappellerons certaines statistiques.

Pour le cas de K = 3 échantillons OZOLS [17] a étudié la statistique unilatérale ($n_1 = n_2 = n_3 = n$) :

$$\mathscr{D}^{+}_{n} = \max\left\{ \sup_x [\hat{F}_2(x) - \hat{F}_1(x)],\, \sup_x [\hat{F}_3(x) - \hat{F}_2(x)] \right\}, \qquad (5)$$

et DAVID [3] la statistique bilatérale ($n_1 = n_2 = n_3 = n$) :

$$\mathscr{D}_{n} = \max\left\{ \sup_x [\hat{F}_2(x) - \hat{F}_1(x)],\, \sup_x [\hat{F}_3(x) - \hat{F}_2(x)],\, \sup_x [\hat{F}_3(x) - \hat{F}_1(x)] \right\}, \qquad (6)$$

Ils ont obtenu les lois exactes et asymptotiques de distribution de ces statistiques dans l'hypothèse H_0 d'homogénéité :

Statistique d'OZOLS [17] :

$$P\left\{ \mathscr{D}_n^+ < \frac{c}{n} \right\} =$$

$$\frac{1}{[n, n, n]} \{ [n, n, n] + [n + c, n + c, n - 2c] + [n - c, n, n + 2c] - [n - 2c, n, n + 2c] - 2 [n - c, n, n + c] \} .$$

$$\lim_{n \to \infty} P [\sqrt{n} \cdot \mathscr{D}_n^+ < z] = 1 - e^{-2z^2} - e^{-4z^2} + e^{-6z^2} \qquad (5')$$

On a introduit ici la notation $[\ell_1, \ell_2, \ell_3]$ pour désigner le coefficient polynomial :

$$[\ell_1, \ell_2, \ell_3] = \frac{[\ell_1 + \ell_2 + \ell_3] !}{\ell_1 ! \ell_2 ! \ell_3 !} . \qquad (7)$$

Statistique de DAVID [3] :

$$P \left[\mathscr{D}_n < \frac{c}{3n} \right] = 3 \sum_{i=1}^{n/c} \sum_{s=1}^{j} [A_{3s-i-1} - A_{3s-i}] ,$$

où

$$A_\ell = \frac{[n - ic, n - \ell c, n + (i - \ell) c]}{[n , n , n]} .$$

$$\lim P \{ \sqrt{n} \, \mathscr{D}_{3n} < z \} = 1 - 3 \sum_{i=1}^{\infty} \sum_{s=1}^{i} \{ e^{-z^2 b_{is}} - e^{-z^2 c_{is}} \} , \qquad (6')$$

où $b_{is} = i^2 + (3s - i - 1)(3s - 2i - 1)$ et $c_{is} = i^2 + (3s - i)(3s - 2i)$.

Dans le cas général ($k > 3$) les résultats sont peu nombreux. GIKHMAN [13] et KIEFER [14] ont obtenu la loi asymptotique de la statistique $\mathscr{D}_k^2$ du test bilatéral :

$$\mathscr{D}_k^2 = \sup_x \sum_{j=1}^{k} n_j \left[\hat{F}_j(x) - \frac{\sum_{j=1}^{k} n_j \hat{F}_j(x)}{N_k} \right]^2 ; \qquad (8)$$

$$\lim_{N_k \to \infty} P [\mathscr{D}_k^2 < x] = \frac{4}{\Gamma\left(\frac{k-1}{2}\right) . (2x)^{(k-1)/2}} \sum_{s=1}^{\infty} \frac{\mathscr{M}_s^{k-3} . e^{-\mathscr{M}_s^2 2x^2}}{J'_{(k-3)/2} (\mathscr{M}_s)} , \qquad (9)$$

où $\mathscr{M}_s$ désigne le s-ième zéro positif de la fonction de BESSEL $J_{(k-3)/2}(s)$ et $N_j = \sum_{i=1}^{j} n_i$.

Ce n'est que dans certains cas particuliers, que la formule (9) prend une forme relativement simple.

Ainsi pour $K = 2$ la formule (9) devient :

$$\lim_{N_2 \to \infty} P [\mathscr{D}_2^2 < x] = \frac{\sqrt{2\pi}}{x} \sum_{s=1}^{\infty} e^{\frac{-\pi^2 (2s-1)}{8x^2}} \qquad (9')$$

De même pour K = 4 nous avons encore une expression assez simple :

$$\lim_{N_4 \to \infty} P\,[\mathscr{D}_4^2 < x] = 1 + 2 \sum_{s=1}^{\infty} [1 - 4\,s^2x^2]\, e^{-2s^2x^2}. \qquad (9'')$$

FISZ et CHANG [8, 9] ont trouvé un système de combinaisons linéaires indépendantes des fonctions de répartition empiriques $F_i(x)$ qui leur a permis de remener le problème de la dimension K à la dimension 2. Pour expliciter cette affirmation considérons tout d'abord le cas de deux échantillons de tailles n_i et n_j générés par les fonctions de répartition $F_i(x)$ et $F_j(x)$. Introduisons les statistiques de SMIRNOV :

pour le test unilatéral :

$$\mathscr{D}_{ij}^+ = \sup\,[\hat{F}_i(x) - \hat{F}_j(x)]$$

pour le test bilatéral :

$$\mathscr{D}_{ij} = \sup\,|\hat{F}_i(x) - \hat{F}_j(x)|.$$

La distribution exacte de ces statistiques étudiée par GNÉDENKO et KOROLIOUK [10, 11, 15, 16] est bien connue et nous avons rappelé sa forme dans le cas particulier où $n_1 = n_2 = n$. Nous introduirons ici pour simplifier les notations suivantes pour leur fonction de répartition :

$$F_{\mathscr{D}_{ij}^+}[z] = \varphi_{n_i,n_j}^+[z] \;;\; F_{\mathscr{D}_{ij}}[z] = \varphi_{n_i,n_j}[z]\,. \qquad (10)$$

Les résultats asymptotiques correspondants s'expriment alors par les relations :

$$\lim_{\substack{n_i \to \infty \\ n_j \to \infty}} \varphi_{n_i,n_j}^+ \left[\frac{n_i + n_j}{n_i \cdot n_j}\, z\right] = 1 - e^{-2z^2}\;; \qquad (11)$$

$$\lim_{\substack{n_i \to \infty \\ n_j \to \infty}} \varphi_{n_i,n_j} \left[\frac{n_i + n_j}{n_i \cdot n_j}\, z\right] = K\,[z] = \sum_{k=-\infty}^{\infty} (-1)^k\, e^{-2k^2z^2}\,. \qquad (11')$$

Les statistiques retenues par FISZ et CHANG dans le cas général sont de la forme (on a posé ici $N_j = \sum_{i=1}^{j} n_i$) :

pour le test unilatéral :

$$\mathscr{D}_j^+ = \sup_x \left[\hat{F}_{j+1}(x) - \frac{1}{N_j} \sum_{i=1}^{j} n_i \hat{F}_i(x)\right]\;; \qquad (12)$$

pour le test bilatéral :

$$\mathscr{D}_j = \sup_x \left|\; \hat{F}_{j+1}(x) - \frac{1}{N_j} \sum_{i=1}^{j} n_i \hat{F}_i \;\right| \qquad (13)$$

Les statistiques $\mathscr{D}_1^+, \ldots, \mathscr{D}_j^+, \ldots, \mathscr{D}_{k-1}^+$ étant indépendantes [9], de même que les statistiques $\mathscr{D}_1^+, \ldots, \mathscr{D}_j^+, \ldots, \mathscr{D}_{k-1}^+$ on obtient immédiatement les relations :

$$P\,\{\forall\,(j = 1, k - 1) : \mathscr{D}_j^+ < z_j\} = \prod_{j=1}^{k-1} \varphi_{N_j,n_{j+1}}^+[z_j]\,. \qquad (14)$$

$$P\,\{\forall\,(j = 1, k - 1) : \mathscr{D}_j^+ < z_j\} = \prod_{j=1}^{k-1} \varphi_{N_j,n_{j+1}}[z_j]\,. \qquad (14')$$

En passant à la limite quand $N_{k \to \infty}$ on obtient des expressions commodes pour l'application pratique des tests :

$$\lim_{N_k \to \infty} P\left\{ \max_{1 \leqslant j < k} \mathscr{D}_j^+ < \frac{N_{j+1}}{N_j \cdot n_{j+1}} z \right\} = [1 - e^{-2z^2}]^{k-1} . \tag{15}$$

$$\lim_{N_k \to \infty} P\left\{ \max_{1 \leqslant j < k} \mathscr{D}_j < \frac{N_{j+1}}{N_j \cdot n_{j+1}} z \right\} = [K(z)]^{k-1} \tag{15'}$$

En pratique, il arrive souvent que l'on doive vérifier non pas l'homogénéité de tous les couples d'échantillons, mais plutôt la *rupture d'homogénéité* dans une série ordonnée d'échantillons. En particulier, les échantillons sont souvent prélevés successivement dans le temps et nous voulons vérifier si nous sommes toujours en présence de la même population. En d'autres termes, nous voulons vérifier l'hypothèse.

$$H_0 : F_1(x) = \ldots = F_i(x) = \ldots = F_k(x) , \tag{16}$$

contre l'hypothèse alternative

$$H_1 : \exists\, (j) \Rightarrow \hat{F}_{j+1}(x) \neq \hat{F}_j(x) . \tag{16'}$$

Nous proposons pour vérifier cette hypothèse un test basé sur les statistiques.

— Test unilatéral

$$G_k^+ = \max_i \sup_x [\hat{F}_{i+1}(x) - \hat{F}_i(x)] ; \tag{17}$$

— Test bilatéral

$$G_k = \max_i \sup_x | \hat{F}_{i+1}(x) - \hat{F}_i(x) | . \tag{17'}$$

Nous présentons ici les résultats obtenus pour le test unilatéral dans le cas d'égalité des tailles des échantillons $n_1 = \ldots = n_j = \ldots = n_k = n$, dont nous avons donné une première démonstration assez complexe en 1966 [4, 5, 6], mais pour lesquels une démonstration beaucoup plus simple a été obtenue maintenant.

PROBABILITÉS EXACTES

$$P\left\{ \forall\, (i > j) \sup_x [\hat{F}_i(x) - \hat{F}_j(x)] < \frac{\theta_i - \theta_j}{n} \right\} = \det \left(\frac{n\,!}{(n - \theta_i + \theta_j)\,!} \right) ;$$

$$P\left\{ \forall\, (i = 1, \ldots, k-1) \sup_x [\hat{F}_{i+1}(x) - \hat{F}_i(x)] < \frac{t_i}{n} \right\} =$$

$$= \det \left(\frac{n\,!}{(n - T_i + T_j)\,!} \right) ; \tag{18}$$

$$P\left\{ \max_i \sup_x [\hat{F}_{i+1}(x) - \hat{F}_i(x)] < \frac{t}{n} \right\} = \det \left(\frac{n\,!}{[n - (i-j)\,t]\,!} \right).$$

PROBABILITÉS ASYMPTOTIQUES

$$\lim_{n \to \infty} P \left\{ \forall\, (i > j) \sup_x [\hat{F}_i(x) - \hat{F}_j(x)] < (\theta_i - \theta_j) \sqrt{\frac{2}{n}} \right\} = \det \left(e^{-(\theta_i - \theta_j)^2} \right).$$

$$\lim_{n \to \infty} P \left\{ \forall\, (i = 1, \ldots, k-1) \sup_x [\hat{F}_{i+1}(x) - \hat{F}_i(x)] < t_i \sqrt{\frac{2}{n}} \right\} = e^{-2 \sum_{j=1}^{k} T_j^2} \times$$

$$\times \det \left(e^{2T_iT_j} \right). \quad (18')$$

$$\lim_{n \to \infty} P \left\{ \max_i \sup_x [\hat{F}_{i+1}(x) - \hat{F}_i(x)] < t \sqrt{\frac{2}{n}} \right\} = \prod_{r=1}^{k} [1 - e^{-2r^2t^2}]^{k-r}.$$

Nous avons adopté ici les conventions suivantes :

$$T_j = \sum_{s=1}^{j} t_s ;\ \forall\, (i > j) : \theta_i > \theta_j > 0 ;\ \theta_i, t_i, t$$

sont des entiers positifs.

Pour établir les formules exactes nous avons généralisé la voie traitée par GNÉDENKO [10, 11] dans le cas classique de 2 échantillons. Précisons tout d'abord le schéma combinatoire adéquat.

SCHÉMA COMBINATOIRE

Soit une urne comportant $L_k = \sum_{j=1}^{k} \ell_j$ boules, dont ℓ_j portent le numéro j.

Considérons les séries obtenues en tirant successivement toutes les boules (sans les remettre). Le nombre total de séries différentes est évidemment

$$(L_k)! / \prod_{j=1}^{k} \ell_j ! .$$

Appelons $v_i[s]$ le nombre de boules numérotées « i » contenues parmi les S premières boules de la série.

Soit :

$$A_{ij}[t_i, t_j, s]$$

l'événement défini par l'inégalité :

$$v_i[s] + t_i < v_j[s] + t_j .$$

Nous définissons encore deux événements

$$A_{ij}[t_i, t_j] = \prod_{s=1}^{L_k} A_{ij}[t_i, t_j, s].$$

$$A = \prod_{i>j} A_{ij}[t_i, t_j]. \quad (19)$$

Introduisons la fonction $f_k(\ell_1, \ldots, \ell_k / \theta_1, \ldots, \theta_k) \equiv f_k[\ell / \theta]$, à l'aide de la relation.

$$f_k\,[\ell_1,\ldots,\ell_k \,/\, \theta_1,\ldots,\theta_k] = P[A]\,.\,\frac{L_k\,!}{\prod_{j=1}^{k} \ell_j\,!}\,. \tag{20}$$

C'est par conséquent le nombre de séries pour lesquelles l'événement A est réalisé.

On vérifie alors aisément les relations suivantes qui établissent le lien entre le schéma combinatoire et la loi de distribution de la statistique du test non paramétrique :

$$P\left\{\forall\,(i > j)\ \sup_x\,[\hat{F}_i(x) - \hat{F}_j(x)] < \frac{\theta_i - \theta_j}{n}\right\} =$$
$$= f_k\,[n,\ldots,n \,/\, \theta_1,\ldots,\theta_k]\,\frac{(n\,!)^k}{(kn)\,!}$$

$$P\left\{\forall\,(i = 1,\ldots,k-1)\ \sup_x\,[\hat{F}_{i+1}(x) - \hat{F}_i(x)] < \frac{t_i}{n}\right\} =$$
$$= f_k\,[n,\ldots,n \,/\, T_1,\ldots,T_k]\,\frac{(n\,!)^k}{(kn)\,!} \tag{21}$$

$$P\left\{\max_i\ \sup_x\,[\hat{F}_{i+1}(x) - \hat{F}_i(x)] < \frac{c}{n}\right\} =$$
$$= f_k\,[n,\ldots,n \,/\, 0, c,\ldots,(k-1)\,c]\,\frac{(n\,!)^k}{(kn)\,!}$$

La démonstration est basée sur la généralisation de la méthode de Gnédenko (10) pour le cas de deux échantillons.

Etablissons maintenant le lemme suivant :

Lemme 1

La fonction $f_k(\ell_1,\ldots,\ell_k \,/\, \theta_1,\ldots,\theta_k) \equiv f_k\,(\ell/\theta)$ est caractérisée univoquement par les 4 propriétés suivantes :

1) $f_k[\ell/\theta] = \sum_{r=1}^{k} f_k[\ell^{(r)} \,/\, \theta^{(r)}]$,

où $\ell_i^{(r)} = \begin{cases} \ell_i \text{ si } i \neq r; \\ \ell_{r-1} \text{ si } i = r; \end{cases}$, $\theta_i^{(r)} = \begin{cases} \theta_i \text{ si } i \neq r; \\ \theta_{r+1} \text{ si } i = r. \end{cases}$

2) $f_k[\ell_1,\ldots,\ell_{i-1}, 0, \ell_{i+1},\ldots,\ell_k \,/\, \theta_1,\ldots,\theta_k] =$
$= f_{i-1}[\ell_1,\ldots,\ell_{i-1} \,/\, \theta_1,\ldots,\theta_{i-1}]\,.\,f_{k-i}[\ell_{i+1},\ldots,\ell_k \,/\, \theta_{i+1},\ldots,\theta_k]$. (22)

3) $f_k[\ell/\theta] = 0$ si $\exists\,(i,j)$ tel que $\theta_i = \theta_j$ et $i \neq j$.

4) $f_2[\ell_1, \ell_2 \,/\, \theta_1, \theta_2] = C_{L_2}^{\ell_1} - C_{L_2}^{\ell_1+\theta_1-\theta_2}$.

La démonstration de ces propriétés découle de la nature même du schéma combinatoire. L'univocité peut être aisément établie. Par conséquent; toute fonction des variables entières $\ell = \{\ell_1,\ldots,\ell_k\}, \theta = \{\theta_1,\ldots,\theta_k\}$ possédant ces 4 propriétés coincide avec la fonction $f_k[\ell/\theta]$.

Lemme 2

La fonction $f_k[\ell/\theta]$ est déterminée par la formule :

$$f_k[\ell/\theta] = L_k!\ \det\left(\frac{1}{[\ell_i + \theta_i - \theta_j]\,!}\right). \tag{23}$$

Démonstration

On remarque immédiatement que pour $k = 2$ on a l'identité :

$$f_2[\ell_1, \ell_2 / \theta_1\theta_2] = C_{L_2}^{\ell_1} - C_{L_2}^{\ell_1+\theta_1-\theta_2} = L_2!\ \begin{vmatrix} \dfrac{1}{\ell_1!} & \dfrac{1}{(\ell_1 + \theta_1 - \theta_2)\,!} \\ \dfrac{1}{(\ell_2 - \theta_1 + \theta_2)\,!} & \dfrac{1}{\ell_2!} \end{vmatrix} \tag{24}$$

Cette identité nous suggère l'introduction de la fonction

$$\psi_k(\ell/\theta) = L_k!\ \det\left(\frac{1}{(\ell_i + \theta_i - \theta_j)\,!}\right), \tag{24'}$$

pour laquelle nous avons déjà que

$$\psi_2(\ell/\theta) = f_2(\ell/\theta).$$

Nous démontrons ensuite, que la fonction $\psi_k(\ell/\theta)$ vérifie les propriétés caractéristiques 1, 2, 3, 4 figurant dans le lemme 1. Seule la démonstration de la propriété 1 est laborieuse et nécessite des développements assez longs, bien que ne comportant pas de difficulté majeure.

En définitive nous obtenons l'identité :

$$\psi_k(\ell/\theta) \equiv f_k(\ell/\theta). \tag{25}$$

Le lemme 2 permet d'établir immédiatement les lois de probabilité de la statistique G_k^+ sous les trois formes que nous avons rapportées.

Pour établir les lois asymptotiques, nous utiliserons la formule de Stirling pour démontrer la relation :

$$\frac{n!}{(n + t\sqrt{2n})\,!} = e^{-t^2+0(1/\sqrt{n})} \cdot e^{t\alpha_n}, \tag{26}$$

où α_n ne dépend pas de t.

On obtient alors compte tenu de (21), (23) et (26) :

$$P\left\{\forall\,(i > j)\ \sup_x\,[\hat{F}_i(x) - \hat{F}_j(x)] < (\theta_i - \theta_j\ \sqrt{2/n}\right\} =$$

$$= \det\left(\frac{n!}{[n + (\theta_i - \theta_j)\sqrt{2n}]\,!}\right) = \det\left(e^{-(\theta_i-\theta_j)^2+0(1\sqrt{n})} \cdot e^{(\theta_i-\theta_j)\alpha_n}\right) =$$

$$= \prod_{i=1}^{k} e^{\theta_i\alpha_n} \cdot \prod_{j=1}^{k} e^{-(\theta_j\alpha_n} \cdot \det\left(e^{-(\theta_i-\theta_j)^2+0(1/\sqrt{n})}\right) =$$

$$= \det\left(e^{-(\theta_i-\theta_j)^2+0(1/\sqrt{n})}\right).$$

En passant à la limite quand n → ∞ nous établissons la première formule asymptotique :

$$\lim_{n \to \infty} P\left\{\forall (i > j) \sup_x [\hat{F}_i(x) - \hat{F}_j(x)] < (\theta_i - \theta_j)\sqrt{\frac{2}{n}}\right\} = \det\left(e^{-(\theta_i - \theta_j)^2}\right).$$

La seconde en découle directement. Pour la troisième nous obtenons de (21) en posant $\theta_i = (i - 1)t$:

$$\det\left(e^{-(\theta_i - \theta_j)^2}\right) = e^{-2t^2 \sum_{j=1}^{k} (j-1)^2} \det\left(e^{2t^2(i-1)(j-1)}\right).$$

Mais $\det\left(e^{t^2(i-1)(j-1)}\right)$ n'est autre qu'un déterminant du type de VANDERMONDE, de sorte que nous obtenons en définitive la troisième formule asymptotique :

$$\lim_{n \to \infty} P\left\{\max_i \sup_x [\hat{F}_{i+1}(x) - \hat{F}_i(x)] < t\sqrt{\frac{2}{n}}\right\} = \prod_{r=1}^{k} [1 - e^{-2r^2t^2}]^{k-r}.$$

Quelques graphiques complémentaires illustrent la méthodologie proposée.

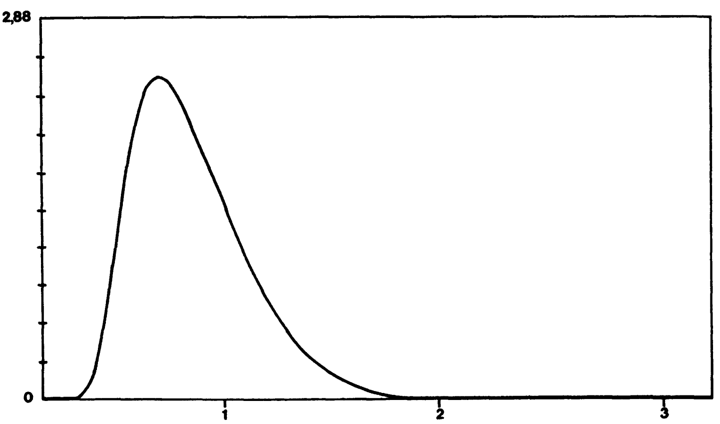

Figure 1. — Densité de probabilité de la loi de KOLMOGOROV.

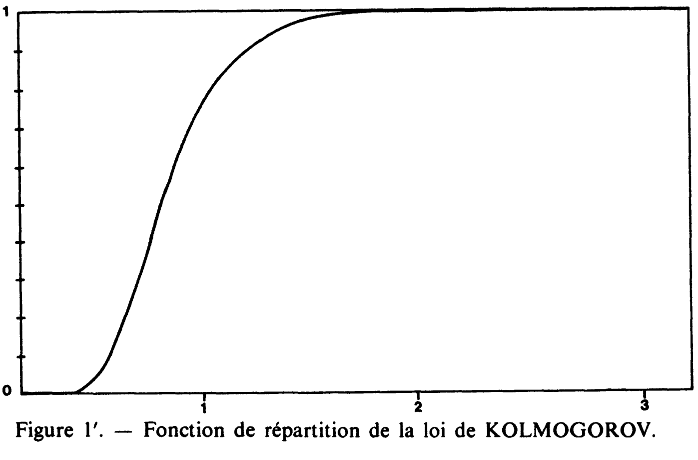

Figure 1'. — Fonction de répartition de la loi de KOLMOGOROV.

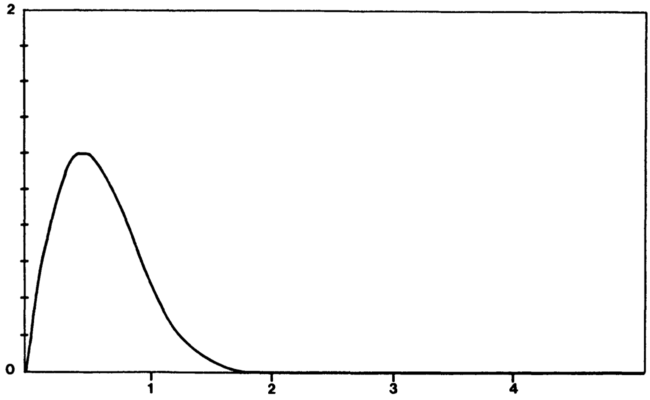

Figure 2. — Densité de probabilité de la loi de SMIRNOV.

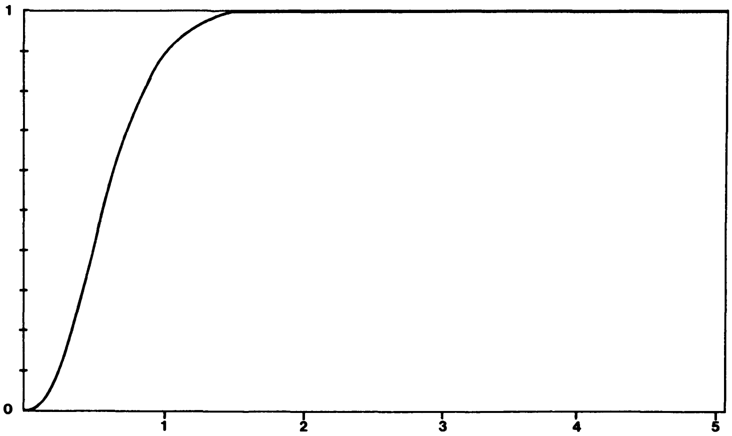

Figure 2'. — Fonction de répartition de la loi de SMIRNOV.

Ainsi on présente sur les figures 1 et 1' la densité de probabilité et la fonction de répartition de la loi de KOLMOGOROV, sur les figures 2 et 2' la densité de probabilité et la fonction de répartition de la loi de SMIRNOV, sur les figures 3 et 3' la forme asymptotique de la densité de probabilité et de la fonction de répartition de la statistique G_{20}^{+} correspondant à la vérification de l'homogénéité de K = 20 échantillons.

Pour une utilisation commode de ce test de rupture d'homogénéité on a présenté également sur la figure 4 les valeurs critiques de la distribution asymptotique de la statistique G_K^+ du test de G. DER MEGREDITCHIAN pour un nombre d'échantillons K variant de 2 à 100. On a souligné en particulier les courbes correspondant aux seuils $\alpha = 0{,}01$ et $\alpha = 0{,}05$. Les valeurs z_α sont définies par la relation :

$$\lim_{n \to \infty} P\{\sqrt{n/2}\, G_K^+ > z_\alpha\} = \alpha$$

Les résultats présentés ont permis l'élaboration d'un programme standardisé « HOMOGK » qui a été appliqué à la vérification de l'homogénéité de nombreux échantillons météorologiques.

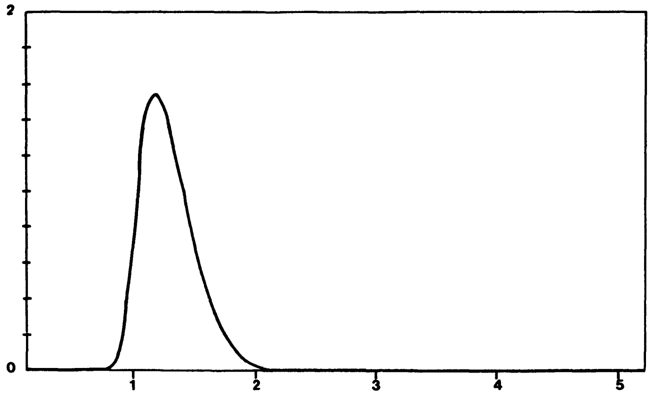

Figure 3. — Densité de probabilité de la statistique G_{20}^{+} du test de rupture d'homogénéité dans le cas de K = 20 échantillons.

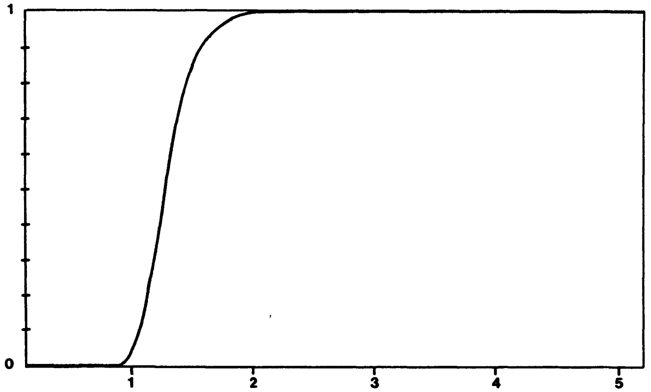

Figure 3'. — Fonction de répartition de la statistique G_{20}^{+} du test de rupture d'homogénéité pour le cas de K = 20 échantillons.

Le programme « HOMOGK » est basé sur le calcul des statistiques de 4 tests : FISZ et CHANG, GIKHMAN-KIEFER, DER MEGREDITCHIAN et la variante adéquate du test du KHI-DEUX. Nous avons prévu dans ce test une modification permettant dans le calcul des statistiques de tenir compte de la redondance des observations par l'introduction du nombre équivalent d'unités d'information indépendantes introduit dans [5, 6]. Par ailleurs le test du KHI-DEUX ne pouvant être appliqué que si les classes extrêmes comportent au moins 5 observations un regroupement des classes est automatiquement prévu si nécessaire.

Le programme HOMOGK a été largement appliqué à la vérification de l'homogénéité des années-échantillons d'observations des variables météorologiques. En effet si l'on considère une variable météorologique mesurée quotidiennement en une station donnée, sa variabilité annuelle impliquera une

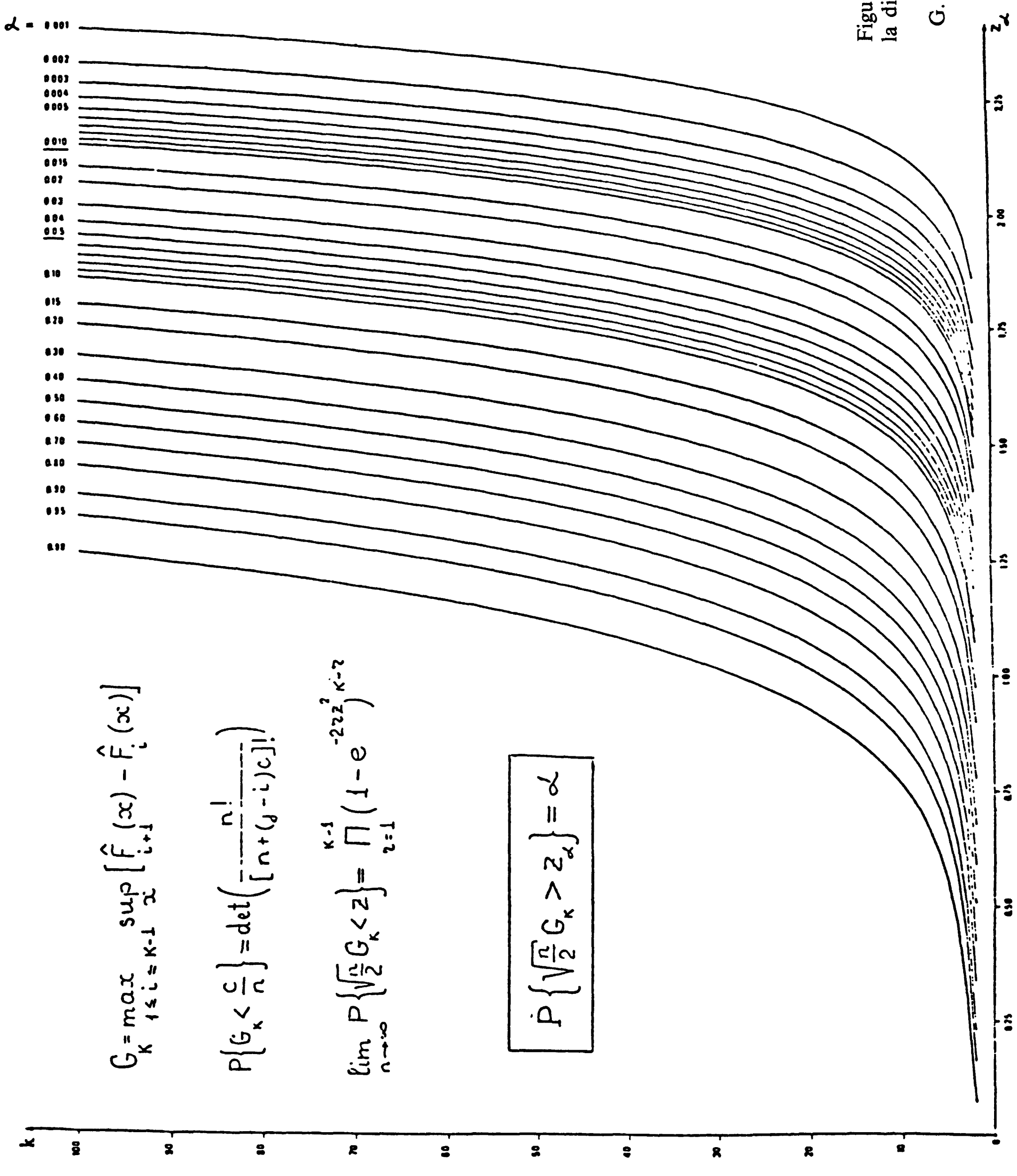

Figure 4. — Valeurs critiques Z_α de la distribution asymptotique de la statistique G_k^+ du test de G. DER MEGREDITCHIAN.

certaine distribution de ses valeurs, caractérisée par une fonction de répartition $F_a(x)$ pour une année « a » donnée. On peut alors se poser la question de la stabilité de cette distribution au cours de K années (a = 1, K). Autrement dit peut-on considérer homogènes les K-années-échantillons ? Comme nous n'avons aucune information de caractère théorique sur la forme concrète de la fonction de répartition théorique inconnue $F_q(x)$, nous sommes bien dans le cadre d'un test non paramètrique de vérification de l'homogénéité de « K » échantillons, de sorte que sans chercher à connaître la fonction $F_a(x)$, nous voulons simplement savoir si elle a évolué au cours des années. Nous sommes ainsi conduit à vérifier l'hypothèse

$$H_0 : F_1(x) = ... = F_a(x) = ... F_k(x) ,$$

contre l'hypothèse alternative

$$H_1 : \exists\ (i) \Rightarrow F_{i+1}(x) \neq F_i(x) .$$

Pour terminer nous donnerons un exemple d'application du programme à la vérification de l'homogénéité des années-échantillons d'ozone total, que nous avons conduits pour le fichier correspondant à 119 stations de mesure qui nous a été fourni par le Centre Mondial de Toronto dans le cadre de l'accord tripartite, U.S.A., U.K., France de surveillance de l'ozone.

En effet il est bien connu que l'ozone constitue pour l'organisme humain un bouclier protecteur contre les radiations néfastes. Or la communauté scientifique est aujourd'hui sensibilisée par les destruction éventulle de cet élément que provoquerait certaines causes exogènes telles la pollution ambiante, les vols stratosphériques, les bombes aérosols, le fréon etc. C'est en tout cas ce qu'il ressort de considérations théoriques basées sur des modèles physicochimiques. Il est donc extrêmement important de vérifier sur les fichiers disponibles le bien fondé de cette hypothèse pessimiste. En particulier nous illustrerons cette étude par l'application de HOMOGK aux 20 années d'observations de l'ozone total à la station Resolute.

Soulignons toutefois, que d'autres approches sont également possibles et ont également été appliquées à la résolution de ce problème complexe.

Le programme HOMOGK calcule tout d'abord les paramètres empiriques du fichier pour chacun des 20 échantillons années comme le montre le tableau 1. Il donne ensuite, pour un nombre de classes fixées des valeurs observées le nombre d'observations par classes, les fréquences relatives et les fréquences cumulées (présentées dans le tableau 2). Les données permettent alors le calcul des statistiques des quatre tests retenus que nous présentons sur le tableau 3.

Le programme sort également les résultats intermédiaires correspondant à la valeur des statistiques pour deux échantillons correspondant à deux années successives, ce qui peut être une information complémentaire intéressante.

En tout état de cause dans l'exemple de l'ozone total l'hypothèse de l'homogénéité des 20 échantillons n'est rejetée que pour 7 des 119 stations étudiées, de sorte que les données dont nous disposons ne nous permettent pas de conclure globalement à une modification significative de la structure du champ d'ozone total.

TABLEAU 1

Paramètres empiriques du fichier « Ozone total » calculés pour chaque année-échantillon Colonne 1 : Numéro de l'année; Colonne 2 : valeur moyenne; Colonne 3 : écart-type; Colonne 4 : 3$^{\text{ième}}$ moment centré; Colonne 5 : 4$^{\text{ième}}$ moment centré; Colonne 6 : coefficient d'asymétrie; Colonne 7 : coefficient d'aplatissement; Colonne 8 : entropie.

STATISTIQUES CALCULEES POUR CHAQUE ANNEE.

	MOYENNE	SIGMA	MIOU3	MIOU4	BETA1	BETA2	ENTROPIE
1	.575E+02	.648E+02	.417E+06	.860E+08	.154E+01	.189E+01	.232E+01
2	.214E+02	.213E+02	.115E+05	.658E+06	.119E+01	.202E+00	.108E+01
3	.833E+01	.139E+02	.735E+04	.358E+06	.272E+01	.650E+01	.337E+00
4	.724E+01	.478E+01	.133E+03	.195E+04	.122E+01	.742E+00	0.
5	.541E+02	.105E+03	.353E+07	.159E+10	.305E+01	.100E+02	.188E.01
6	.698E+02	.972E+02	.202E+07	.749E+09	.220E+01	.538E+01	.239E+01
7	.511E+02	.668E+02	.615E+06	.162E+09	.206E+01	.514E+01	.227E+01
8	.114E+02	.134E+02	.329E+04	.125E+06	.137E+01	.861E+00	.518E+00
9	.537E+02	.560E+02	.167E+06	.292E+08	.107E+01	-.198E-01	.197E+01
10	.837E+02	.101E+03	.229E+07	.958E+09	.221E+01	.613E+01	.271E+01
11	.597E+02	.928E+02	.190E+07	.695E+09	.237E+01	.637E+01	.215E+01
12	.236E+01	.301E+01	.396E+02	.316E+03	.145E+01	.855E+00	0.
13	.182E+02	.169E+02	.348E+04	.164E+06	.722E+00	-.930E+00	.381E+00
14	.480E+01	.681E+01	.650E+03	.131E+05	.206E+01	.310E+01	0.
15	.157E+02	.234E+02	.601E+05	.106E+03	.468E+01	.321E+02	.883E+00
16	.346E+02	.426E+02	.160E+06	.257E+08	.207E+01	.482E+01	.181E+01
17	.334E+02	.337E+02	.324E+05	.288E+07	.847E+00	-.770E+00	.149E+01
18	.446E+01	.456E+01	.121E+03	.167E+04	.128E+01	.862E+00	0.
19	.144E+02	.180E+02	.113E+05	.694E+06	.195E+01	.365E+01	.614E+00
20	.335E+02	.388E+02	.907E+05	.112E+08	.156E+01	.196E+01	.173E+01

TABLEAU 2

Fréquences cumulées (fonctions de répartition empiriques) des valeurs observées par classes des années-échantillons (K = 20, n = 20).

FREQUENCES CUMULEES

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	.50	.60	.74	.87	.88	.90	.94	.96	.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2	.74	.89	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
3	.94	.94	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
5	.65	.75	.81	.86	.89	.92	.92	.92	.94	.95	.97	.97	.97	.97	.97	.98	.98	.98	.98	.98
6	.56	.60	.70	.80	.84	.88	.94	.96	.98	.98	.98	.98	.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00
7	.52	.70	.77	.85	.90	.93	.94	.98	.98	.99	.99	.99	.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
8	.89	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
9	.52	.67	.67	.86	.90	.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
10	.40	.50	.64	.71	.79	.83	.86	.95	.95	.95	.98	.98	.98	.98	.98	.98	.98	.98	1.00	1.00
11	.61	.74	.76	.79	.84	.85	.92	.94	.95	.97	.98	.98	.98	.99	.99	.99	.99	1.00	1.00	1.00
12	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
13	.70	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
14	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
15	.80	.96	.99	.99	.99	.99	.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
16	.59	.79	.89	.94	.97	.98	.99	.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
17	.64	.73	.82	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
18	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
19	.89	.94	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
20	.60	.81	.89	.94	.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

TABLEAU 3

Valeurs observées des statistiques des tests non paramétriques d'homogénéité figurant dans le programme HOMOGK.

```
              TESTS D'HOMOGENEITE DES 20 ECHANTILLONS
              _______________________________________

TEST DE DER MEGREDITCHIAN

GK (+) =        .391E+00      GK =         .391E+00
SK (+) =        .194E+01      SK =         .194E+01

TEST DE FISZ ET CHANG

GK (+) =        .404E+00      GK =         .404E+00
SK (+) =        .223E+01      SK =         .223E+01

TEST DE GIKHMAN ET KIEFER

D2K =           .219E+02

TEST DU KHI-DEUX (DDL = 361)

S =             .359E-01      (AVEC REGROUPEMENT)
S =             .310E+03      (SANS REGROUPEMENT)
```

BIBLIOGRAPHIE

[1] ANDRÉ. — Solution directe du problème résolu par Bertrand. *C.R.A.S.*, Paris, Vol. 105, 1887.

[2] BERTRAND. — *Calcul des probabilités,* Paris, 1925.

[3] DAVID. — A three sample Kolmogorov-Smirnov test. *AMS,* Vol. 29, 1958.

[4] DER MEGREDITCHIAN. — Un critère unilatéral d'homogénéité de K échantillons. *Travaux du Centre Météorologique Mondial n° 9,* Moscou, 1966 (en russe).

[5] DER MEGREDITCHIAN. — *Thèse de doctorat,* Moscou 1969.

[6] DER MEGREDITCHIAN. — Sur la définition du nombre de stations indépendantes « équivalentes » à un système donné de stations corrélées (en russe). *Météorologie et Hydrologie n° 2* Moscou, 1963.

[7] FELLER. — *An introduction to Probalility Theory and its applications.* Wiley, 1957.

[8] FISZ. — A limit theorem for empirical distribution functions. *Bull. Polish. Acad. Scien. 5,* 1957.

[9] FISZ and CHANG. — Asymptoticaly independant linear functions of empirical distribution functions. *Science Record,* 1957.

[10] GNÉDENKO. — *Théorie des probabilités* (en russe) Moscou, 1961.

[11] GNÉDENKO. — Vérification de la constance des distributions des probabilités dans deux échantillons indépendants. *Math. Nachr., 12, 1-2, 1954.*

[12] GNÉDENKO et KOROLIOUK. — Sur l'écart maximal entre deux fonctions de répartition empiriques (en russe). *Comptes rendus de l'Académie des Sciences de l'URSS,* 80, 4, 1951.

[13] GIKHMAN. — Sur un test non paramétrique de vérification de l'homogénéité de « K » échantillons (en russe). *Théorie des probabilités et ses applications,* 1957, n° 3.

[14] KIEFER. — K samples analogues of the Kolmogorov-Smirnov and Cramer. Von Mises *tests AMS.* Vol. 30, n° 2, 1959.

[15] KOROLIOUK, YAROCHEVSKI. — Distribution de l'écart maximum entre deux fonctions de répartition empiriques, *Dopovidi Akad. Navk UKRAIN RSR,* 80, 4, 1951.

[16] KOROLIOUK. — Sur l'écart entre deux fonctions de répartition empiriques pour le cas de deux échantillons indépendants. *Bulletin de l'Académie des Sciences de l'U.R.S.S., série math.,* 19, 1, 1955.

[17] OZOLS. — Le théorème de Gnédenko-Koroliouk pour trois échantillons avec deux fonctions unilatérales (en letton). *Bulletin de l'Académie des Sciences de Lettonie,* 1956, Vol. 10.

[18] SMIRNOV. — Estimation de l'écart entre deux échantillons indépendants.*Bulletin de l'université de Moscou,* T 11, 2, 1939.

[19] STECK. The SMIRNOV two sample test as rank test, *AMS,* 1969, 10.

[20] CHANG LI CHIEN. Sur la distribution exacte de la statistique de SMIRNOV et son développement asymptotique. *Mathematica,* 4 (2), 1960.