ANNALES SCIENTIFIQUES
DE L'UNIVERSITÉ DE CLERMONT-FERRAND 2
Série Mathématiques

YU. V. LINNIK

Sur certaines questions de statistique analytique

Annales scientifiques de l'Université de Clermont-Ferrand 2, tome 8, série *Mathématiques*, n° 2 (1962), p. 53-61

<http://www.numdam.org/item?id=ASCFM_1962__8_2_53_0>

SUR CERTAINES QUESTIONS DE STATISTIQUE ANALYTIQUE

Yu. V. LINNIK
Professeur à l'Université de Léningrad (U.R.S.S.)

Cette communication concerne deux questions de la statistique analytique - la construction des zones similaires et la question des statistiques polynomiales.

1 - ENONCE DE LA QUESTION. PROPRIETES GENERALES DES ZONES SIMILAIRES -

La théorie des zones similaires a été fondée par J. Neyman et E. Pearson en 1933 dans le travail [1]. La revue des travaux dans ce domaine jusqu'en 1954 se trouve dans l'exposé bien connu de J. Neyman [2]. Le développement ultérieur de ce problème a été exposé par J. Neyman dans sa revue des problèmes de statistique au IV Congrès des Mathématiciens Soviétiques à Léningrad en juillet 1961.

Soit E_n l'espace Euclidien à n dimensions ; nous allons nous occuper de l'ensemble des mesures de probabilité P_ϑ où ϑ est un paramètre appartenant à un ensemble arbitraire $\bar{\vartheta}$. Toutes les mesures P_ϑ se rapportent à la même σ-algèbre dans E_n.

L'ensemble A de cette σ-algèbre se nomme zone similaire si la mesure $P_\vartheta(A)$ de cet ensemble a la même valeur pour tous les P_ϑ. Il va sans dire que seulement les zones similaires A pour lesquelles $P_\vartheta(A) \neq 0$ sont intéressantes. Les zones similaires ont une application dans la construction des tests éliminant les paramètres nuisibles. A une zone similaire A on peut faire correspondre une statistique $t = t(X)$ où $t = 1$ pour $X \in A$; $t = 0$ pour $X \notin A$; la distribution de t sera indépendante de ϑ. Inversement, si l'on a une statistique mesurable jouissant de cette propriété de l'indépendance, la zone $t < \xi$ sera évidemment une zone similaire quel que soit ξ. Une statistique t de cet espèce sera nommée dans l'exposé statistique zonale. Au lieu d'étudier les zones similaires, nous pouvons étudier les statistiques zonales t.

Si pour toute fonction continue $\varphi = \varphi(t)$ d'une statistique t, pour laquelle la moyenne $E\varphi(t) = U(\varphi, \vartheta)$ existe, cette dernière ne dépend pas de ϑ, la statistique t est zonale. Pour le voir, il suffit de prendre $\exp(itu)$ pour $\varphi(t)$, u étant un paramètre scalaire. Alors $E\varphi(t)$ sera la fonction caractéristique (f. c. dans ce qui suit) de la statistique t, d'où notre assertion.

Dans le cas où le paramètre ϑ prend ses valeurs dans l'intervalle numérique ouvert (a, b) et où l'ensemble des mesures P_ϑ remplit quelques conditions analytiques, on peut établir une relation entre les propriétés des statistiques zonales et la quantité d'information dans l'échantillon au sens de R. A. Fisher. Supposons que chaque mesure P_ϑ possède la densité de probabilité $L(X, \vartheta)$ (que nous supposons en outre continue par rapport à X et ϑ pour simplifier la discussion). Supposons que l'expression $\mathcal{J} = \frac{1}{L}\frac{\partial L}{\partial \vartheta}$ a un sens ; nous la nommerons l'informant. Il est bien connu que certaines conditions analytiques assez faibles étant remplies, nous avons : $E\mathcal{J}^2 = I$, (où I est la quantité d'information dans l'échantillon au sens de Fisher). La propriété de la statistique t d'être zonale est liée assez simplement au comportement de l'informant $\mathcal{J}$. Moyennant des conditions analytiques assurant la légitimité de quelques transformations simples, on démontre aisément le théorème suivant. (L'auteur a constaté après sa conférence que ce théorème avait été établi par J. Neyman en 1938 par une voie un peu différente).

THEOREME 1 - Pour que la statistique t soit zonale, il faut et il suffit que la régression de l'informant $\mathcal{J}$ par rapport à t soit nulle presque sûrement pour toutes les mesures P_ϑ :

$$E(\mathcal{J} \mid t) = 0 \qquad (1,1)$$

presque sûrement pour toutes les P_ϑ.

Soit t une statistique zonale ; nous avons à déduire l'égalité (1,1). La statistique t étant zonale, nous avons pour toutes les valeurs du paramètre u :

$$\int_{E_n} \dots \int e^{iut} L(X, \vartheta) dX$$

ne dépend pas de ϑ.

En admettant la légitimité de la différenciation par rapport à ϑ du noyau et l'existence de l'informant $\mathcal{J}$, nous obtenons :

$$\int_{E_n} \dots \int e^{iut} \frac{\partial L(X, \vartheta)}{\partial \vartheta} dX = E(e^{iut} \mathcal{J}) = 0 \qquad (1,2)$$

Moyennant des conditions analytiques assez faibles cela nous donne :

$$E(e^{iut} \mathcal{J}) = E(e^{iut} E(\mathcal{J} | t)) = 0 \qquad (1,3)$$

d'où l'on déduit aisément (1,1). Inversement, admettons (1,1) presque sûrement pour toutes les P_θ. Alors pour toutes valeurs de ϑ nous obtenons (1,3) et puis (1,2). Nous observons que la f. c. de t ne dépend pas de ϑ, donc t est une statistique zonale.

La condition (1,1) sera remplie sûrement si la statistique zonale t ne dépend pas de l'informant $\mathcal{J}$ parce que alors $E(\mathcal{J}|t) = E(\mathcal{J}) = 0$. De tels cas seront discutés plus loin lorsque nous parlerons de structures de J. Neyman (c. f. par exemple [3]) dans le cas d'existence de statistiques exhaustives de rang fini et dans le § 4.

Les statistiques zonales jouent un rôle pour ainsi dire opposé à celui des statistiques exhaustives ; elles ne comprennent aucune information sur le paramètre ϑ tandis que les statistiques exhaustives doivent comprendre toute cette information. Toute statistique t indépendante des statistiques exhaustives pour le paramètre ϑ doit être zonale : sà distribution conditionnelle, les statistiques exhaustives étant fixées, ne dépend pas de ϑ, et elle doit coïncider avec la distribution a priori. Dans cette communication nous discuterons certaines propriétés des statistiques zonales et certaines constructions qui généralisent les structures de J. Neyman.

2 - STATISTIQUES ZONALES LINEAIRES POUR LES ECHANTILLONS REPETES -

Dans ce paragraphe nous considérons le cas où les mesures P_θ sur E_n déterminent des vecteurs aléatoires de composantes indépendantes et identiquement distribuées (cas des échantillons répétés). Le paramètre ϑ appartient à un ensemble arbitraire. Nous considérons les statistiques zonales linéaires : $t = a_1 x_1 + \dots + a_n x_n$. Quelle que soit ξ, $P_\theta(t < \xi)$ ne dépend pas de ϑ. Il se trouve que la théorie des statistiques zonales linéaires est liée étroitement à la théorie des statistiques linéaires équidistribuées, développée dans les travaux de l'auteur [4] et [5]. En utilisant les méthodes élaborées dans ces travaux, nous pouvons déduire certains théorèmes sur les statistiques zonales linéaires. La statistique zonale t sera dite la plus simple si elle est de la forme $t = X_i - X_j$ (il va sans dire que seul le cas $i \neq j$ est intéressant). Si ϑ est un paramètre scalaire du type de translation (tel que, la distribution de X_i dépende de $X_i - \vartheta$ seulement), la différence $t = X_i - X_j$ sera évidemment une statistique zonale. Mais $t = X_i - X_j$ peut être statistique zonale dans certains autres cas. Soit par exemple, X_i des variables aléatoires de distributions indéfiniment divisibles pour toutes les valeurs de ϑ, supposons que la fonction spectrale dans les expressions correspondantes de P. Lévy pour les f. c. puisse être décomposée en parties "paire" et "impaire". La partie paire doit être indépendante du paramètre ϑ tandis que la partie impaire peut dépendre de ϑ. Il est aisé de voir que $t = X_i - X_j$ sera une statistique zonale. Soit $t = a_1 X_1 + \dots + a_n X_n$ statistique linéaire zonale où il y a des a_i qui ne s'annulent pas. Si tous les $a_i \neq 0$ ont les mêmes $|a_i|$, il est aisé de déduire en formant la f. c. que $X_i - X_j$ sera aussi une statistique zonale. Vue cette circonstance, nous admettrons que les nombres $b_j = |a_j|$ ne sont pas tous égaux à zéro où l'un à l'autre. Suivant la méthode des travaux [4], [5], formons "la fonction déterminante".

$$\sigma(z) = |a_1|^z + \dots + |a_n|^z = b_1^z + \dots + b_n^z \qquad (2,1)$$

Vues les conditions imposées aux a_i, les zéros réels et complexes de $\sigma(z)$ seront situés dans une bande verticale de largeur finie (cf. [4]). La borne supérieure exacte des abscisses des zéros de $\sigma(z)$ sera désignée par r.

THEOREME 2 - Soit $t = a_1X_1 + \ldots + a_nX_n$ une statistique linéaire zonale où les a_i ne sont pas tous égaux à zéro ou à l'un d'eux. (Donc, $r \neq \infty$). Supposons que pour toutes les valeurs de ϑ le moment d'ordre 2 m de X_i où $m = \frac{r}{2} + 1$ existe. Si la fonction caractéristique de X_i, $f(u, \vartheta) \neq 0$ pour toutes les valeurs réelles de u, la statistique linéaire $X_i - X_j$ sera aussi zonale.

Mme J. L. Romanovskaia a trouvé que ce théorème reste exact si l'on remplace toutes ces conditions par une seule : $f(u, \vartheta)$ est quasi-analytique dans un ensemble ouvert contenant $u = 0$, pour toutes les valeurs de ϑ.

Sans les conditions imposées sur l'existence des moments de X_i (la quasianalyticité de $f(u, \vartheta)$ aux environs de zéro est certainement une condition de ce type), on peut construire des exemples où le théorème 2 est inexact.

On peut déduire le théorème 2 en suivant les raisonnements du travail [4] (cf. [4], la déduction du théorème II, p. p. 208-224). Soit ϑ_0 une valeur quelconque du paramètre ϑ ; $(y_1, \ldots, y_n)$ un vecteur aléatoire aux composantes indépendantes distribuées comme les X_i ; $\varphi(u, \vartheta) = f(u, \vartheta) f(-u, \vartheta)$ le f. c. de $x_i - y_i$, vues les conditions du théorème 2, $\varphi(u, \vartheta) \neq 0$ pour toutes les valeurs de u, donc, $\varphi(u, \vartheta) > 0$. Formons : $\psi(u, \vartheta) = \text{Log}\, \varphi(u, \vartheta) - \text{Log}\, \varphi(u, \vartheta_0)$. Nous avons :

$$\sum_{j=o}^{n} \psi(b_j u, \vartheta) = 0 \qquad (2,2)$$

En outre, la fonction $\psi(u, \vartheta)$ est continue sur tout l'axe des u. La solution de l'équation (2,2) au moyen de la transformation de Laplace est effectuée et discutée dans [4], p. p. 208-234. La distinction du cas que nous considérons ici consiste en ce que $\psi(u, \vartheta)$ n'est pas le logarithme d'une f. c. mais la différence de deux logarithmes de cette espèce. En suivant les raisonnements du travail indiqué nous obtenons que l'existence du moment d'ordre 2 m entraîne l'égalité :

$$\psi(u, \vartheta) = \sum_{k=1}^{K} A_k(\vartheta)\, u^{2k} \qquad (2,3)$$

où les $A_k(\vartheta)$ sont des fonctions numériques de ϑ et K une constante. Comme $\sum_{j=1}^{K} b_j^{2k} > 0$ pour $k = 1, 2, \ldots, K$, (2,2) entraîne immédiatement la relation : $A_k(\vartheta) \equiv 0$ $(k = 1, \ldots, K)$ donc $\psi(u, \vartheta) \equiv 0$, d'où la conclusion que $X_i - X_j$ est une statistique zonale, C. Q. F. D.

3 - SUR UNE FAMILLE DE DISTRIBUTIONS ADMETTANT DES ZONES SIMILAIRES -

Dans ce qui suit nous étudierons une famille des mesures de probabilité P_θ données sur E_n et ayant une densité continue par rapport à la mesure de Lebesgue sur E_n. Le paramètre ϑ sera scalaire et appartenant à l'intervalle ouvert (a, b) (les nombres a, b peuvent être égaux à $-\infty$ ou $+\infty$).

Prenons un ensemble fini arbitraire $\{\vartheta_1, \ldots, \vartheta_s\}$ de valeurs du paramètre ϑ et considérons pour un ensemble borélien arbitraire A, le vecteur $\{P_{\theta_1}(A), P_{\theta_2}(A), \ldots, P_{\theta_s}(A)\}$. En vertu du théorème connu de A. A. Liapounoff [6], l'ensemble des valeurs de ce vecteur sera convexe. Comme cet ensemble contient les vecteurs $\{0, \ldots, 0\}$ et $\{1, \ldots, 1\}$, pour tout $\lambda \in (0,1)$ il existe un ensemble A tel que $P_{\theta_1}(A) = \ldots = P_{\theta_s}(A) = \lambda$; cet ensemble nous donne une zone similaire pour l'ensemble fini des mesures $\{P_{\theta_i}\}$, $i = 1, 2, \ldots, s$ (1). En considérant les mesures $P_{\theta i}$ sur l'ensemble E - A, nous pouvons faire une conclusion analogue en remplaçant λ par $\lambda P_{\theta_1}(E - A)$ et en répétant ce procédé, construire une statistique zonale à distribution discrète avec un nombre prescrit de points de croissance.

Ce raisonnement connu prouve l'existence de statistiques zonales non-triviales pour une classe assez étendue d'ensembles finis de mesures. Pour le cas des ensembles infinis de mesures $\{P_\theta\}$ on connait une méthode de construction de zones similaires (et statistiques zonales) s'il existe un système fini de statistiques exhaustives $\{X_1, \ldots, X_s\}$ tel que les surfaces de niveau : $X_1 = C_1$, $X_2 = C_2, \ldots, X_s = C_s$ permettent de former des systèmes locaux de coordonnées dans un sous-espace de E_n de dimension $< n$. Si en outre ce système de statistiques exhaustives $\{X_1, \ldots, X_s\}$ se trouve complet au sens restreint (cf. [3], p. p. 130-134), les zones similaires se construisent au moyen des structures de J. Neyman. (cf. le même livre), si cette dernière condition n'est pas remplie, elles existent tout de même et peuvent être construites.

---

(1) Cette application du théorème de A. A. Liapounov dans la théorie des zones similaires est due à J. Neyman (cf. [2]).

Dans ce paragraphe nous construirons une classe de familles $\{L(X,\vartheta)\}$ plus étendue que celle que permettent les structures de J. Neyman et discuterons ses propriétés.

Considérons un système de statistiques scalaires (non nécessairement exhaustives) $V_1, V_2, \ldots, V_r$ $(r < n)$ telles que les surfaces de niveau $V_1 = C_1, \ldots, V_r = C_r$ soient suffisamment régulières (par exemple, possédant des dérivées jusqu'à l'ordre 3) et permettant l'introduction de systèmes de coordonnées locales $\{\xi_1, \ldots, \xi_{n-r}\}$, de sorte que dans tout l'espace E_n nous avons le système de coordonnées $(V_1, V_2, \ldots, V_r, \xi_1, \ldots, \xi_{n-r})$. Les mesures P_ϑ engendrées par l'expression $L(X,\vartheta)dX$ doivent induire une distribution conditionnelle sur l'ensemble $V_1 = C_1, \ldots, V_r = C_r$, s'il n'est pas vide.

Supposons maintenant que pour un nombre quelconque $k \geqslant 0$, $L(X,\vartheta)$ satisfait pour tous les $\vartheta \in (a,b)$ et pour toutes les valeurs de X à l'équation différentielle :

$$\rho_0(V_1, V_2, \ldots, V_r, \vartheta)\frac{\partial^k L(X,\vartheta)}{\partial\vartheta^k} + \rho_1(V_1, \ldots, V_r, \vartheta)\frac{\partial^{k-1}}{\partial\vartheta^{k-1}}L(X,\vartheta) +$$

$$+ \ldots + \rho_k(V_1, \ldots, V_r, \vartheta)\,L(X,\vartheta) + \rho_{k+1}(V_1, \ldots, V_r, \vartheta) = 0 \qquad (3,1)$$

Les coefficients $\rho_i(V_1, \ldots, V_r, \vartheta)$ $(i = 0, 1, \ldots, k+1)$ sont supposés être suffisamment réguliers (par exemple trois fois différentiables par rapport à tous les arguments).

Les statistiques $V_1, \ldots, V_r$ étant fixées :

$$V_1 = V_{10}, \ldots, V_2 = V_{20} \qquad (3,2)$$

de telle façon que l'on puisse introduire les probabilités conditionnelles sur la surface (3,2), nous les désignons par $l(\xi_1, \ldots, \xi_{n-r}, V_1, \ldots, V_r, \vartheta)$. Nous avons :

$$L(X,\vartheta) = g(V_{10}, \ldots, V_{20}, \vartheta)\; l(\xi_1, \ldots, \xi_{n-r}, V_{10}, \ldots, V_{20}, \vartheta)\; J(V,\xi) \qquad (3,3)$$

où g est la densité des V_i et $\mathcal{J}(V,\xi)$ le jacobien de la transformation qui ne dépend pas de ϑ.

Posons dans l'équation (3,1) : $V_1 = V_{10}, \ldots, V_2 = V_{20}$ et exprimons $L(X,\vartheta)$ par la formule (3,3). Les $g(V_{10}, \ldots, V_{20}, \vartheta)$ sont des fonctions connues de ϑ. L'équation (3,1) se réduit à l'équation :

$$\pi_0(V_{i_0},\vartheta)\frac{\partial^k}{\partial\vartheta^k}\,\mathcal{J}.l(\xi, V_{i_0},\vartheta) + \pi_1(V_{i_0},\vartheta)\frac{\partial^{k-1}}{\partial\vartheta^{k-1}}\,\mathcal{J}l(\xi, V_{i_0},\vartheta) +$$

$$+ \ldots + \pi_k(V_i,\vartheta).\,\mathcal{J}l(\xi, V_{i_0},\vartheta) + \pi_{k+1}(V_{i_0},\vartheta) = 0 \qquad (3,4)$$

où nous désignons par $\pi_j(V_{i_0},\vartheta)$ les nouveaux coefficients et par $l(\xi, V_{i_0},\vartheta)$ la fonction de la formule (3,3) (la densité de probabilité). Nous admettons la légitimité de nos opérations.

Pour simplifier les notations, omettons l'argument V_{i_0} dans les notations ; posons :

$$\pi_j(V_{i_0},\vartheta) = \pi_j(\vartheta)\;;\; (j \leqslant k)\;;\; l(\xi, V_{i_0},\vartheta) = l(\xi,\vartheta).$$

Nous obtenons pour $l(\xi,\vartheta)$ l'équation :

$$\mathcal{J}\left[\pi_0(\vartheta)\,\frac{\partial^k l(\xi,\vartheta)}{\partial\vartheta^k} + \pi_1(\vartheta)\,\frac{\partial^{k-1} l(\xi,\vartheta)}{\partial\vartheta^{k-1}} + \ldots + \pi_k(\vartheta).\,l(\xi,\vartheta)\right] + \pi_{k+1}(\vartheta) = 0 \qquad (3,5)$$

Maintenant il faut obtenir, en partant de (3,5) une nouvelle équation pour $l(\xi,\vartheta)$ qui ne contienne pas ce terme en $l(\xi,\vartheta)$ et le terme indépendant de l. On peut l'obtenir formellement au moyen des opérations triviales suivantes : nous divisons (3,5) par $\pi_{k+1}(\vartheta)$, nous prenons la dérivée partielle par rapport à ϑ du résultat et après celà, nous divisons l'expression obtenue par le coefficient de $l(\vartheta)$: $\left(\frac{\pi_k(\vartheta)}{\pi_{k+1}(\vartheta)}\right)'$ après quoi nous différentions par rapport à ϑ encore une fois. On obtient de la sorte l'équation :

$$\rho_0(\vartheta)\,\frac{\partial^{k+2} l(\xi,\vartheta)}{\partial\vartheta^{k+2}} + \rho_1(\vartheta)\,\frac{\partial^{k+1} l(\xi,\vartheta)}{\partial\vartheta^{k+1}} + \ldots + \rho_{k+1}(\vartheta)\,\frac{\partial l(\xi,\vartheta)}{\partial\vartheta} = 0 \qquad (3,6)$$

Observons encore que sous certaines conditions analytiques assez faibles, l'équation (3,6) pour la densité conditionnelle $l(\xi, \vartheta)$ est équivalente à la possibilité d'exprimer $l(\xi, \vartheta)$ sous la forme :

$$l(\xi,\vartheta) = c_1(\xi)\, V_1(\vartheta) + \ldots + c_{K+2}(\xi)\, V_{K+2}(\vartheta), \tag{3,7}$$

où les $V_i(\vartheta)$ forment un système des fonctions suffisamment régulières et linéairement indépendantes.

Admettons maintenant que toutes les opérations qui nous ont guidés de l'équation (3,4) à l'équation (3,6) sont légitimes. Supposons encore que les fonctions $\rho_i(\vartheta)$ sont suffisamment régulières par rapport à leurs arguments (ϑ et V_{i_0}) et que $\rho_0(\vartheta)$ ne s'annule pas pour $\vartheta \in (a,b)$ et toutes les valeurs de V_{i_0}. Montrons maintenant comment on peut construire des zones similaires pour $L(X, \vartheta)$ dans ce cas.

Soit $\psi(\xi)$ une fonction continue de $\xi_1, \ldots, \xi_{n-r}$ telle que l'espérance mathématique conditionnelle.

$$\int\ldots\int \psi(\xi)\, l(\xi,\vartheta)\, d\xi = u_\psi(\vartheta) = u_\psi \tag{3,8}$$

existe. Supposons que l'opération (3,8) puisse être permutée avec l'opérateur différentiel dans la partie gauche de (3,6). Alors en posant $K + 2 = m$, nous obtenons :

$$\rho_0(\vartheta)\frac{d^m u_\psi}{d\vartheta^m} + \rho_0(\vartheta)\frac{d^{m-1}u_\psi}{d\vartheta^{m-1}} + \ldots + \rho_{m-1}(\vartheta)\frac{du_\psi}{d\vartheta} = 0 \tag{3,9}$$

En vertu du théorème de A.A. Liapounov mentionné au commencement de ce paragraphe, nous pouvons construire une zone similaire et une statistique zonale t pour les distributions $l(\xi, \vartheta_1), \ldots, l(\xi, \vartheta_s)$ pour tout système fini de nombres $\vartheta_1, \ldots, \vartheta_s$. Considérons les conditions qui permettent de déduire des équations (3,6) et (3,9) que cette statistique t sera zonale pour $\{l(\xi, \vartheta)\}$ pour toutes les valeurs $\vartheta \in (a,b)$.

Posons $\frac{du_\psi}{d\vartheta} = V_\psi$ et écrivons l'équation (3,9) sous la forme :

$$\rho_0(\vartheta)\frac{d^{m-1}\vartheta V_\psi}{d\vartheta^{m-1}} + \rho_1(\vartheta)\frac{d^{m-2}V_\psi}{d\vartheta^{m-2}} + \ldots + \rho_{m-1}(\vartheta)\, V_\psi = 0 \tag{3,10}$$

Considérons un système de solutions fondamentales de l'équation (3,10) : $V_1(\vartheta), \ldots, V_{m-1}(\vartheta)$ pour $\vartheta \in (a,b)$. En vertu de leur indépendance linéaire, le déterminant :

$$\begin{vmatrix} V_1(\vartheta_1) & \ldots & V_{m-1}(\vartheta_1) \\ \ldots & \ldots & \ldots \\ V_1(\vartheta_{m-1}) & \ldots & V_{m-1}(\vartheta_{m-1}) \end{vmatrix} = W(\vartheta_1, \ldots, \vartheta_{m-1}) \tag{3,11}$$

où les $\vartheta_i \in (a,b)$ sont des variables indépendantes, ne peut pas s'annuler identiquement. $W(\vartheta_1, \ldots, \vartheta_{m-1})$ étant continue, il ne s'annule pas dans un certain domaine $\vartheta_i \in [\vartheta_i^\circ - \varepsilon, \vartheta_i^\circ + \varepsilon]$ $(i = 1, 2, \ldots, m-1)$. Prenons maintenant $2m - 2$ nombres ϑ_i', ϑ_i'' appartenant aux intervalles $[\vartheta_i^\circ - \varepsilon, \vartheta_i^\circ + \varepsilon]$ et tels que $\vartheta_i' \neq \vartheta_i''$, $(i = 1, 2, \ldots, m-1)$ et construisons une statistique $t = t(\xi)$ qui est zonale pour $l(\xi, \vartheta)$, les paramètres égaux à $\vartheta_1', \vartheta_1'', \ldots, \vartheta_{m-1}', \vartheta_{m-1}''$. Si $\psi(t) = \psi(t(\xi))$ est une fonction continue pour laquelle l'expression (3,8) existe, nous avons (cf. § 1) :

$$u_\psi(\vartheta_i') = u_\psi(\vartheta_i'') \qquad (i = 1, 2, \ldots, m-1) \tag{3,12}$$

La fonction u_ψ, comme auparavant est supposée, suffisamment régulière. En vertu du théorème de Rolle, il existe des points $\vartheta_i \in [\vartheta_i^\circ - \varepsilon, \vartheta_i^\circ + \varepsilon]$ tels que :

$$u_\psi'(\vartheta_i) = V_\psi(\vartheta_i) = 0 \qquad (i = 1, 2, \ldots, m-1)$$

Maintenant nous avons, en substituant à $V_\psi(\vartheta)$ son expression par les solutions fondamentales :

$$V_\psi(\vartheta_i) = C_1^{(\psi)} V_1(\vartheta_i) + \ldots + C_{m-1}^{(\psi)} V_{m-1}(\vartheta_i) = 0 \tag{3,13}$$

$(i = 1, 2, \ldots, m-1)$. Nous avons vu que $W(\vartheta_1, \ldots, \vartheta_{m-1}) \neq 0$, donc $C_j^{(\psi)} = 0$ $(j = 1, 2, \ldots, m-1)$ et $V_\psi(\vartheta) = 0$ pour $\vartheta \in (a,b)$, donc $u_\psi(\vartheta) = u_\psi(\vartheta_1')$ pour $\vartheta \in (a,b)$.

Dans la formule (3,8) posons $\psi = \psi(\xi) = \exp itv$, v étant un paramètre scalaire. Dans ce cas $u_\psi = u_\psi(v, \vartheta)$ est la f. c. de t. Nous avons vu que u_ψ ne dépend pas de ϑ, de sorte que $t = t(\xi)$ sera une statistique zonale pour $\{l(\xi, \vartheta)\}$; $\vartheta \in (a,b)$. Ainsi, pour toute surface non-vide de la forme (3,2) nous pouvons construire des zones similaires. Par leur construction même, elles changeront continuement si l'on change les parties droites de (3,2) et nous pouvons "coller ensemble" ces zones comme on le fait dans la théorie des structures de J. Neyman (cf. [1]) ; alors en vertu du théorème des probabilités totales, nous obtiendrons une zone similaire pour la famille $L(X, \vartheta)$, $\vartheta \in (a,b)$; on peut construire comme cela les statistiques zonales.

4 - EXEMPLES. STATISTIQUES QUASI-EXHAUSTIVES -

1/ Si $\{V_1, \ldots, V_s\}$ forment un système de statistiques exhaustives pour le paramètre ϑ, nous avons, sous certaines conditions analytiques assez faibles, la décomposition connue :

$$L(X,\vartheta) = g(V_1, \ldots, V_s, \vartheta)\, h(\xi_1, \ldots, \xi_{n-s}, V_1, \ldots, V_s)\, \mathcal{J}(V, \xi) \tag{4,1}$$

où g est la densité de probabilité pour les statistiques, h la densité conditionnelle dans les coordonnées locales, indépendante de ϑ, $\mathcal{J}(V, \xi)$ le jacobien de la transformation également indépendant de ϑ. Ainsi, pour la densité conditionnelle $h(\xi, V)$ nous avons :

$$\frac{\partial h(\xi, V)}{\partial \vartheta} = 0 \tag{4,2}$$

Nous voyons que, sous certaines conditions analytiques, (4,2) est équivalent à la relation :

$$\frac{\partial}{\partial \vartheta}\left(\frac{1}{g} L(X, \vartheta)\right) = 0 \tag{4,3}$$

ce qui correspond à l'équation du type (3,1) pour l'ordre 1. De cette façon, l'existence de statistiques exhaustives de rang inférieur à la dimension de l'espace correspond à ce que l'équation du type (3,1) existe et est à l'ordre 1. Cela mène naturellement à l'introduction de la conception de statistiques quasi-exhaustives. Le système des statistiques $(V_1, \ldots, V_s)$ $(s < n)$ sera nommé un système de statistiques quasi-exhaustives pour le paramètre $\vartheta \in (a,b)$ si la densité des probabilités conditionnelles de l'échantillon $h(\xi_1, \ldots, \xi_{n-s}, V_1, \ldots, V_s) = h(\xi, V)$ obéit à l'équation différentielle de l'ordre $k \geqslant 1$:

$$\rho_o(\vartheta) \frac{\partial^k h(\xi, V)}{\partial \vartheta^k} + \rho_1(\vartheta) \frac{\partial^{k-1} h(\xi, V)}{\partial \vartheta^{k-1}} + \ldots + \rho_k(\vartheta) \frac{\partial h(\xi, V)}{\partial \vartheta} = 0 \tag{4,4}$$

où les $\rho_i(\vartheta)$ sont suffisamment réguliers et $\rho_o(\vartheta) \neq 0$ pour $\vartheta \in (a,b)$.

Il va sans dire que cette propriété est équivalente à l'expression de $h(\xi, V)$ sous la forme :

$$h(\xi, V) = C_1(\xi, V)\, u_1(\vartheta) + \ldots + C_k(\xi, V)\, u_k(\vartheta) \tag{4,5}$$

avec les $u_k(\vartheta)$ suffisamment réguliers et linéairement indépendants.

Tandis que la fixation des statistiques exhaustives élimine le paramètre de la distribution conditionnelle de l'échantillon, la fixation des statistiques quasi-exhaustives n'éliminera pas le paramètre ϑ, en général, mais rendra la dépendance de la distribution conditionnelle de l'échantillon du paramètre "peu essentielle".

De façon précise, si la distribution conditionnelle d'une statistique quelconque t, les statistiques quasi-exhaustives étant fixées, ne dépend pas de la valeur du paramètre ϑ pour un nombre fini de points convenablement choisis $\vartheta_1, \ldots, \vartheta_s$ elle ne dépendra pas de $\vartheta \in (a,b)$ et l'on peut généralement construire les zones, similaires au moyen de "collages".

On voit que l'équation (4,4) est de type plus restreint que l'équation (3,1).

2/ Cas où l'équation différentielle est du premier ordre.

Dans ce cas, l'équation (3,1) est de la forme :

$$\rho_o(V, \vartheta) \frac{\partial L(X, \vartheta)}{\partial \vartheta} + \rho_1(V, \vartheta)\, L(X, \vartheta) + \rho_2(V, \vartheta) = 0 \tag{4,6}$$

La solution explicite est bien connue. En posant :

$$\frac{\rho_1(V,\vartheta)}{\rho_0(V,\vartheta)} = P(V,\vartheta) \; ; \; \frac{\rho_2(V,\vartheta)}{\rho_0(V,\vartheta)} = Q(V,\vartheta),$$

la solution générale est de la forme :

$$L(X,\vartheta) = \exp - \int P(V,\vartheta)d\vartheta \; (C_1(X) - \int Q(V,\vartheta) \; (\exp \int P(V,\vartheta)d\vartheta)d\vartheta)$$

Dans le cas particulier :

$$\rho_2(V,\vartheta) \equiv Q(V,\vartheta) \equiv 0, \; C_1(X) = \exp C_2(X) \; ; \; P(V,\vartheta) =$$

$$= P(V) + h(\vartheta), \; \int h(\vartheta)\,d\vartheta = h_1(\vartheta), \; \int P(V,\vartheta)d\vartheta = \vartheta P(V) + h_1(\vartheta) \; ;$$

nous obtenons :

$$L(X,\vartheta) = \exp(C_2(X) + \vartheta P(V) + h_1(\vartheta))$$

où P(V) sera la seule statistique suffisante pour le paramètre ϑ dans le cas de la paramétrisation dite naturelle. Toutes les zones similaires seront dans ce cas (cf. [7]) les structures de J. Neyman. Dans le livre de M. Kendall ([8], pp. 283-285) au moyen de raisonnements assez compliqués sur les moments et sous des conditions très fortes, on discute un cas particulier de l'équation (4,6) lorsque les $\rho_i(V,\vartheta)$ ne dépendent pas des V.

3/ Echantillon répété.

Considérons une famille de densités unidimensionnelles de probabilités $\{\mu(X,\vartheta)\}$:

$$\mu(X,\vartheta) = \left(\exp \sum_{j=1}^{M} h_j(\vartheta)\, T_j(X)\right) \left(\sum_{i=o}^{N} \varphi_i(X)\, g_i(\vartheta)\right), \qquad (4,7)$$

où h_j, T_j, φ_i, g_i sont des fonctions suffisamment régulières ; $\vartheta \in (a,b)$; M, N des nombres entiers finis. On peut considérer cette famille comme celle que l'on obtient à partir de la famille initiale qui correspond à la valeur N = 0 au moyen de corrections du type Gram-Charlier. Pour un échantillon répété de volume n, nous obtenons :

$$L(X,\vartheta) = (\exp \Sigma h_j(\vartheta) V_j) \prod_{K=1}^{n} \left(\sum_{u=1}^{n} \varphi_i(X_K)\, g_i(\vartheta)\right) \qquad (4,8)$$

où $V_j = \sum_{K=1}^{n} T_j(X_K)$. Nous voyons que pour l'expression $\exp(-\Sigma h_j(\vartheta)V_j)\, L(X,\vartheta)$ on obtient une équation différentielle de l'ordre n(N + 1) dont les coefficients sont indépendants de X_i et dont les solutions sont du type : $\prod_{\mu=1}^{K} g_{i\mu}(\vartheta)$, c'est pourquoi $L(X,\vartheta)$ a le type (3,1) et nous pouvons construire des zones similaires comme auparavant.

5 - SUR LES STATISTIQUES POLYNOMIALES -

Dans le §1 nous avons remarqué que toute statistique indépendante des statistiques exhaustives est zonale ; en outre $E(\mathfrak{I}|t) = 0$ presque sûrement si $\mathfrak{I}$ est l'informant pour $L(X,\vartheta)$ et t une statistique zonale. Celà conduit à poser la question : une statistique zonale doit-elle être indépendante des statistiques exhaustives et de l'informant ? Dans ce cas la relation $E(\mathfrak{I}|t) = 0$ serait la conséquence directe de la relation connue : $E(\mathfrak{I}) = 0$ (on suppose que les conditions analytiques correspondantes sont remplies). Dans le cas de la construction des zones similaires et des statistiques zonales, au moyen des structures de J. Neyman, la statistique zonale sera évidemment indépendante des statistiques exhaustives. Si, en outre l'échantillon est répété et possède un système de statistiques exhaustives de rang connu nous avons (certaines conditions analytiques étant remplies) :

$$L(X,\vartheta) = \exp \sum_{j=1}^{M} h_j(\vartheta)V_j \,,$$

où h_j et V_j sont des fonctions du même type que celui de la formule (4,8). (On pose habituellement $V_1 \equiv h_1(\vartheta) \equiv 1$). Ici V_j sont les statistiques suffisantes. L'informant $\mathfrak{I}$ est de la forme :

$$\mathcal{J} = \sum_{j=1}^{M} h_j'(\vartheta)\, V_j$$

La statistique zonale construite au moyen des structures de J. Neyman sera indépendante des statistiques exhaustives, donc, aussi de $\mathcal{J}$; la condition $E(\mathcal{J}|t) = 0$ est évidente dans ce cas. S'il existe des statistiques quasi-exhaustives et si les statistiques zonales sont formées comme nous l'avons expliqué auparavant, elles seront indépendantes des statistiques quasi-exhaustives. Les propriétés indiquées ici donnent une raison nouvelle pour étudier le phénomène de l'indépendance des statistiques. Les phénomènes de cette sorte ont été étudiés pour le cas des statistiques polynomiales, quasipolynomiales et statistiques dites "tubes" par E. Lukacs [9], l'auteur [10], A. A. Zinger [11] et autres. (La littérature jusqu'en 1954 est indiquée dans la revue de E. Lukacs [9]). Mais dans le domaine même des statistiques polynomiales il y a plusieurs questions non-résolues. Nous ne savons pas jusqu'à présent classer les paires de statistiques polynomiales (à fortiori rationnelles), indépendantes pour l'échantillon répété normal, quoique plusieurs tests de la théorie des petits échantillons et de l'analyse de la variance soient basés sur des statistiques de cette espèce.

Quant aux statistiques polynomiales $P(X_1, \ldots, X_n)$; $Q(X_1, \ldots, X_n)$ pour un échantillon répété normal, $X_i \in N(0,1)$, on peut par exemple, formuler l'hypothèse suivante : si P et Q sont indépendantes, il existe une transformation orthogonale des coordonnées $X_1, \ldots, X_n$ qui transforme les statistiques P et Q en statistiques fonctions de deux systèmes d'arguments disjoints. La propriété inverse - que les statistiques de cette espèce sont indépendantes et restent indépendantes pour toutes les transformations orthogonales des coordonnées est triviale.

Cette hypothèse n'est ni prouvée ni refutée ; elle est exacte pour les polynômes de degré n'excédant pas 2.

Pour étudier les phénomènes de l'indépendance des statistiques polynomiales (qui peut être réduit au phénomène plus général des statistiques équidistribuées), on peut proposer une méthode valable pour le cas des mesures probabilistes quelconques sur E_n, possédant des moments de tous ordres.

Cette méthode est fondée sur l'application des éléments de la géométrie algébrique. Nous expliquerons son application au cas de l'hypothèse mentionnée plus haut sur l'indépendance des statistiques polynomiales d'un échantillon normal.

Soit $P(a,X)$ et $Q(a,X)$ deux statistiques de cette espèce, nous désignons par a et b les coefficients correspondants et par $X_1, X_2, \ldots, X_n$ les coordonnées de X. Une transformation orthogonale arbitraire portant sur X, transforme $P(a,X)$; $Q(a,X)$, en statistiques $P(a',X')$, $P(l',X')$, également indépendantes ; les X_i' peuvent être interprétés comme les coordonnées du même échantillon normal.

Nous aurons un ensemble dénombrable de relations :

$$EP^rQ^s = EP^rEQ^s \quad ; \quad r,s = 0,1,2,\ldots \tag{5,1}$$

qu'on peut écrire sous la forme :

$$\pi_{rs}(a,b) = 0 \quad ; \quad r,s = 0,1,2,\ldots \tag{5,2}$$

où les π_{rs} sont des polynômes en a, b. (Ils forment un idéal polynomial d'où l'on peut dégager une base finie, cf. [12]).

Pour vérifier cette hypothèse prenons un système fini quelconque de relations (5,2) (la remarque précédente montre que les systèmes infinis se réduisent aux systèmes finis). Nous obtenons un ensemble algébrique qui peut être décomposé en variétés sur le corps des nombres réels. Nous pouvons maintenant évaluer les dimensions de nos variétés et les comparer aux dimensions que l'on obtient en comptant les paramètres du groupe orthogonal (par exemple, I pour le volume de l'échantillon n = 2). Si les dimensions coïncident avec les dimensions calculées par la seconde voie, en appliquant les éléments de la théorie des variétés algébriques, nous prouvons notre hypothèse. L'évaluation des dimensions peut être effectuée en formant les ensembles algébriques tangents à : $\pi_{rs}(a,b) = 0$; $r,s = 1,2,\ldots,M$, c'est-à-dire en calculant les rangs des matrices qui consistent en éléments de la forme : $\frac{\partial \pi_{r,s}(a,b)}{\partial a}$, $\frac{\partial \pi_{rs}(a,b)}{\partial b}$ où a,b sont les coefficients ; r,s peuvent être choisis comme on veut (pour alléger le calcul). De cette façon on peut vérifier l'hypothèse pour tout volume n de l'échantillon fixé, et pour tout nombre m, maximum des degrés en X des polynômes $P(a,x)$; $Q(b,x)$, quoique les calculs se compliquent lorsque n et m croissent. A. A. Zinger a effectué les calculs correspondants pour les échantillons normaux de volume n = 2 et les polynômes de tous les degrés m_1, m_2 ; l'hypothèse se trouve exacte dans tous ces cas.

Cette méthode est valable pour étudier les statistiques polynomiales ; elle est inapplicable en général au cas des statistiques rationnelles. Mais l'hypothèse est inexacte pour ces dernières, comme on le voit en considérant l'exemple de deux statistiques indépendantes : $U + V$ et $\frac{U}{V}$ où $U = X_1^2 + \ldots + X_m^2$; $V = X_{m+1}^2 + \ldots + X_n^2$; $X_i \in N(0,1)$; les X_i forment un échantillon répété.

BIBLIOGRAPHIE

[1] J. NEYMAN, E. S. PEARSON - On the problem of most efficient test of statistical hypotheses. Philos. Trans. Roy. Soc. London, Ser. A (231) ; 1933 ; 289-337.

[2] J. NEYMAN - Current problems of mathematical statistics ; International Math. Congress, Amsterdam, 1954.

[3] E. LEHMANN - Testing statistical hypotheses, N. Y. J. Wiley, 1959.

[4] Yu. V. LINNIK - Formes linéaires et critères statistiques, I (en russe) Ukr. Matem. Journal, v. 5. N 2 ; 1953, 207-243.

[5] Yu. V. LINNIK - Formes linéaires et critères statistiques, II (en russe). Ukr. Matem. Journal, v. 5, N 3 ; 1953, 247-290.

[6] A. A. LIAPOUNOV - Sur les fonctions-vecteurs complètement additives. Izv. AN. SSSR, Sér. Mathem ; 4 ; 1940 ; 467-478.

[7] E. SVERDRUP - Similarity, Unbiasedness, Minimaxibility and Admissibility of Statisical Test Procedures. Skandinavisk Aktuerietidshrift, 1953 ; H 1-2 ; 64-86.

[8] M. KENDALL - The advanced theory of statistics. T. II, Ch. Griffin, London, 1955.

[9] E. LUKACS - Characterization of populations by properties of suitable statistics. Proc. III Berkeley Symposium ; T. II, 1956 ; 195-204.

[10] Yu. V. LINNIK - Sur les statistiques polynomiales en connexion avec la théorie des équations différentielles. Vestnik Leningradskogo Universiteta, N I ; 1956, 35-48.

[11] A. A. ZINGER - L'indépendance des statistiques quasipolynomiales et propriétés analytiques des distributions. Teoria Veroiatnostei i ee primen., v. 3, f. 3 ; 1958, 265-284.

[12] Yu. V. LINNIK - Polynomial statistics and polynomial ideals. Calcutta Math. Soc. Golden Jubilee Commen. Volume ; 1958-59, Part. I ; 95-98.