R. DE MISES

Sur les fonctions statistiques

Bulletin de la S. M. F., tome 67 (1939), p. 177-184 (supplément)

<http://www.numdam.org/item?id=BSMF_1939__67__S177_0>

Réun. intern. Math. (1937, Paris)
Bull. Soc. math. France,
Suppl. 1939, p. 177 à 184.

SUR LES FONCTIONS STATISTIQUES

Par M. R. de MISES.

Le théorème dit de Laplace-Tchebychef concernant l'addition de variables aléatoires est assez connu. Sous des conditions très larges on peut démontrer que, si $x_1, x_2, \ldots, x_n$ sont n variables soumises à des lois de probabilité indépendantes, la loi de probabilité de la somme $x_1 + x_2 + \ldots + x_n$ tend vers la fonction de Gauss, n augmentant infiniment. Évidemment au lieu de la somme on peut parler d'une fonction linéaire quelconque des $x_1, x_2, \ldots, x_n$, notamment de leur moyenne arithmétique. Mais il paraît être resté inaperçu jusqu'ici que le même fait s'étend à une classe de fonctions bien plus générale. Toute fonction qui, ainsi que la moyenne arithmétique, ne dépend que de la répartition des valeurs des $x_1, x_2, \ldots, x_n$ jouit, sous certaines conditions, de la propriété citée. Presque toute les fonctions de variables aléatoires que l'on étudie généralement en statistique mathématique, par exemple le coefficient de Lexis, le coefficient de corrélation, etc., appartiennent à cette classe de fonctions. C'est pourquoi je leur ai donné le nom de *fonctions statistiques*. Pour établir les notions et les théorèmes précis dans ce domaine il faut s'appuyer sur les définitions fondamentales, dues au génie de M. Volterra, définitions qui établissent la notion de *fonction de ligne*, de sa dérivée, etc. Je me permettrai dans ce qui suit, d'exposer quelques-uns de mes résultats dont une partie fut antérieurement publiée (1).

(1) Voir les références à la fin de cet article.

1. Fonction statistique. — Nous allons nous servir de deux notions apparentées, mais qu'il faut bien distinguer l'une de l'autre. J'appelle *répartition* des n variables réelles $x_1, x_2, \ldots, x_n$ la fonction $S(x)$ définie de sorte que $nS(x)$ donne pour tout x le nombre de ceux parmi les $x_1, x_2, \ldots, x_n$ dont la valeur ne dépasse pas x. Il est évident que $S(x)$ est une fonction monotone, non décroissante, représentée par une ligne-escalier qui monte de 0 à 1 par des marches dont les hauteurs sont des multiples entiers de $\frac{1}{n}$.

De l'autre côté toute fonction $V(x)$ monotone et non décroissante qui monte de 0 à 1 de façon quelconque sera dénommée une *distribution*. Toute répartition de n variables est ainsi une distribution particulière. Dans tout problème de probabilité (dans tout collectif) à une seule dimension l'ensemble des probabilités est donné par une telle distribution, $V(x)$ désignant la probabilité pour que la valeur de la variable aléatoire (du caractère distinctif du collectif) ne dépasse pas x.

Soit maintenant E un ensemble de distributions qui comprend au moins un certain nombre de répartitions. Un tel ensemble auquel appartiennent $V_1(x)$ et $V_2(x)$ sera dit *convexe*, si $V_1 + t(V_2 - V_1)$ lui appartient aussi, pour toute fraction proprement dite t. Attachons à tout élément $V(x)$ de E un nombre réel f; nous appellerons f une *fonction statistique définie sur* E et nous la désignerons par $f\{V(x)\}$. Si la *variable indépendante* $V(x)$ devient égale à la répartition $S(x)$ des variables $x_1, x_2, \ldots, x_n$, f ne dépend que de ces n variables et nous dirons aussi que $f\{S(x)$ est une fonction statistique des $x_1, x_2, \ldots, x_n$.

L'exemple le plus simple d'une fonction statistique s'écrit sous la forme d'une intégrale de Stieltjes comme il suit

$$f\{V(x)\} = \int x\, dV(x). \tag{1}$$

Quand $V(x)$ est la distribution des probabilités dans un collectif, f est *l'espérance mathématique de* x; si l'on prend pour $V(x)$ la répartition de n variables $x_1, x_2, \ldots, x_n$, alors f est leur moyenne arithmétique. En remplaçant x sous le signe dans (1) par une fonction quelconque $\varphi(x)$ on reçoit toujours une fonction statistique *linéaire*, définie pour toutes les $V(x)$ pour lesquelles l'intégrale existe. Soient

$$A = \int \alpha(x)\, dV(x), \quad B = \int \beta(x)\, dV(x), \quad \ldots, \tag{2}$$

de telles intégrales de Stieltjes, toute fonction $F(A, B, \ldots)$ sera une fonction statistique, en général non linéaire. Dans cet ordre d'idées entrent par exemple, les écarts d'ordre n rapportés à la valeur moyenne a, les coefficients de Lexis et de Gini et presque toutes les autres caractéristiques dont on se sert en statistique mathématique. Citons encore comme exemple d'une fonction statistique d'autre nature l'intégrale double

$$f\{V(x)\} = \iint \psi(x, y)\, dV(x)\, dV(y). \tag{3}$$

On voit bien que les fonctions statistiques constituent un cas particulier des *fonctions de ligne* introduites par M. Volterra.

2. Dérivées et formule de Taylor. — Il y aura lieu d'introduire la notion de dérivée et de formule Taylorienne pour une fonction statistique. Soient $V_1(x)$ une distribution fixe et $V(x)$ une distribution variable, toutes les deux appartenant à l'ensemble convexe E sur lequel la fonction f est définie. Nous dirons que f est dérivable *au point* $V_1(x)$, s'il existe une fonction f' qui dépend de $V_1(x)$ et d'une variable y, mais pas de $V(x)$, telle que

$$\frac{d}{dt} f\{V_1(x) + t(V - V_1)(x)\}_{t=0} = \int f'\{V_1(x), y\}\, d(V - V_1)(y). \tag{4}$$

Cette fonction $f'\{V_1(x), y\}$ sera appelée la dérivée de $f\{V(x)\}$ au *point* $V_1(x)$.

Pour toute fonction linéaire $f = \int \alpha(x)\, dV(x)$ la dérivée est indépendante de $V_1(x)$ et égale à $\alpha(y)$. Dans le cas $f = F(A, B, \ldots)$ où A, B, ... sont définies par (2), on trouve

$$f'\{V(x), y\} = \frac{\partial F}{\partial A}\alpha(y) + \frac{\partial F}{\partial B}\beta(y) + \ldots. \tag{5}$$

Pour l'écart d'ordre m on a

$$\begin{cases} f = M_m = \int (x - a)^m\, dV, \qquad a = \int x\, dV, \\ f' = (y - a)^m - m M_{m-1} y. \end{cases} \tag{6}$$

La fonction non linéaire (3) admet la dérivée

$$f'\{V(x), y\} = \int [\psi(x, y) + \psi(y, x)]\, dV(x). \tag{7}$$

D'une façon analogue on définit une dérivée de deuxième ordre au point $V_1(x)$ qui dépendra en général de $V_1(x)$ et de deux variables ordinaires y, z. Ainsi on peut établir une formule Taylorienne pour la fonction statistique $f\{V(x)\}$ qui est supposée deux fois dérivable sur tout le *segment* menant de $V_1(x)$ à $V(x)$

$$(8)\quad \begin{aligned} f\{V(x)\} - f\{V_1(x)\} &= \int f'\{V_1(x), y\}\, d(V - V_1)(y) \\ &+ \frac{1}{2}\int\int f''\{V_2(x), y, z\}\, d(V - V_1)(y)\, d(V - V_1)(z). \end{aligned}$$

Ici $V_2(x)$ est une distribution appartenant au *segment* en question, c'est-à-dire que

$$(9)\qquad V_2(x) = V_1(x) + \vartheta(V - V_1)(x), \qquad 0 \leqq \vartheta \leqq 1.$$

3. **Le problème.** — Abordons maintenant le problème du calcul des probabilités dont il s'agit dans le cas du premier théorème fondamental. On a une suite infinie de variables $x_1, x_2, x_3, \ldots$ dont chacune est assujettie à une certaine distribution (loi des probabilités) $V'_1(x)$, $V'_2(x)$, $V'_3(x)$, ... En d'autres termes, la probabilité pour que la $\nu^{\text{ième}}$ variable ne dépasse pas la valeur x est égale à $V'_\nu(x)$. Une expérience effectuée sur les n premiers de ces collectifs fournira un certain système de valeurs $x_1, x_2, \ldots, x_n$ et nous désignerons par $S_n(x)$ la répartition de ces n valeurs. Soit maintenant f une fonction statistique quelconque, par exemple (pour fixer les idées) l'écart d'ordre m

$$(10)\qquad f\{S_n(X)\} = \int (x - a)^m\, dS(x) = \frac{1}{n}\sum_{\nu=1}^{n} (x_\nu - a)^m,$$

où a désigne la moyenne arithmétique $x_1, x_2, \ldots, x_n$. On demande :

Quelle est la probabilité $P'_n(X)$ *pour que la valeur* $f\{S_n(x)\}$ *ne dépasse pas* X *et notamment quelle est l'allure asymptotique de la distribution* $P'_n(X)$?

Notre déduction sera basée sur *le lemme* suivant : soient A_n et B_n, pour tout entier n, deux fonctions dépendant de certaines variables aléatoires, de sorte qu'il existe des distributions $P_n(X)$ pour A_n et $Q_n(X)$ pour B_n [c'est-à-dire que $P_n(X)$ est la probabilité pour que A_n ne dépasse pas X, etc.]; alors si $Q_n(X)$ pour n infini s'approche d'une distribution $\Phi(X)$ à dérivée bornée et que, en même temps, l'espérance mathématique de $|A_n - B_n|$ tend vers zéro, $\Phi(X)$ sera également la limite de $P_n(X)$. La démonstration en est tout à fait élémentaire.

4. Démonstration. — Nous portons dans la formule Taylorienne (8) $S_n(x)$ à la place de $V(x)$ et la moyenne arithmétique des distributions V'_ν données

$$(11) \qquad V_n(x) = \frac{1}{n}[V'_1(x) + V'_2(x) + \ldots + V'_n(x)]$$

à la place de $V_1(x)$. En multipliant les deux membres par une constante H_n dont la valeur sera précisée tout de suite nous posons

$$(12) \qquad \begin{cases} A_n = H_n[f\{S_n(x)\} - f\{V_n(x)\}], \\ B_n = H_n \int f'\{V_n(x), y\}\, dT_n(y), \end{cases}$$

où T signifie la différence $S_n - V_n$. Ainsi l'équation (8) fournit

$$(13) \qquad A_n - B_n = \frac{1}{2} H_n \iint f''\{V(x), y, z\}\, dT_n(y)\, dT_n(z),$$

où $V(x)$ désigne une certaine distribution appartenant au segment de $S_n(x)$ à $V_n(x)$. Pour pouvoir appliquer le lemme cité ci-dessus il faut prouver que la distribution de B_n tend vers la Gaussienne et que l'espérance mathématique de la valeur absolue de (13) tend vers zéro.

Quant à la première de ces propositions il suffit de remarquer que B_n est une fonction statistique *linéaire* de $S_n(x)$. En effet, pour une distribution $V_n(x)$ fixe, l'expression $f'\{V_n(x), y\}$ est simplement une fonction de la variable y et, sauf le facteur, H_n/n, la première partie de

$$(14) \qquad B_n = \frac{H_n}{n} \sum_{\nu=1}^{n} f'\{V_n(x), x_\nu\} - \frac{H_n}{n} \sum_{\nu=1}^{n} \int f'\{V_n(x), y\}\, dV'_\nu(y)$$

est la *somme* des valeurs de cette fonction pour $y = x_1, x_2, \ldots, x_n$, tandis que la deuxième partie n'est qu'une constante. Donc nous n'aurons qu'à demander que les distributions $V'_\nu(y)$ considérées comme distributions des variables transformées $f'\{V_n(x), y\}$ satisfassent aux conditions pour que le théorème de limite classique soit valable. Par exemple, si l'on pose pour $\nu = 1, 2, 3, \ldots$

$$
(15) \qquad \left\{
\begin{aligned}
a_\nu &= \int f'\{V_n(x), y\}\, dV'_\nu(y),\\
r_\nu^2 &= \int [f'\{V_n(x), y\} - a_\nu]^2\, dV'_\nu(y)\\
C_\nu &= \int [f'\{V_n(x), y\} - a_\nu]^{2+\varepsilon}\, dV'_\nu(y),\\
&\qquad (s_n^2 = r_1^2 + r_2^2 + \ldots + r_n^2),
\end{aligned}
\right.
$$

il suffira que, pour un certain ε positif, les C_ν soient bornés et que l'expression $n^{-\frac{2}{2+\varepsilon}} s_n^2$ tende vers l'infini avec n. Mais naturellement on peut substituer pour (15) un autre système de conditions suffisantes, conformément aux théories connues. En tout cas on aura pour la distribution Q_n de B_n, si H_n/n est choisi égal à $\frac{1}{\sqrt{2}s_n}$,

$$
(16) \qquad \lim_{n \to \infty} Q_n(X) = \frac{1}{\sqrt{\pi}} \int e^{-u^2}\, du.
$$

5. **Fin de la démonstration. Énoncé.** — Tout dépend maintenant d'une appréciation de l'espérance mathématique de l'expression (13). Nous nous appuyerons sur le fait suivant qui, en substance, se déduit de l'inégalité bien connue de Tchebychef. Soit $\psi(x)$ une fonction non négative et

$$
(17) \qquad J = \int \psi(x) [S_n(x) - V_n(x)]^2\, dx,
$$

on aura pour l'espérance mathématique $E\{J\}$

$$
(18) \qquad E\{J\} \leqq \frac{1}{n} \int \psi(x)\, V_n(x) [1 - V_n(x)]\, dx.
$$

Pour appliquer cette formule nous restreignons les fonctions f admises à celles dont la deuxième dérivée $f''\{V(x), y, z\}$ dans un *point* quelconque $V(x)$ de l'ensemble convexe défini par les $V_n(x)$ satisfait la condition qui suit : il existe une fonction non négative $\psi(x)$ telle que pour $T_n(x) = S_n(x) - V_n(x)$ l'inégalité

$$
(19) \qquad \left| \iint f''\{V(x), y, z\}\, dT_n(y)\, dT_n(z) \right| \leqq \int \psi(x) T_n^2(x)\, dx,
$$

entraîne la relation

$$\lim_{n \to \infty} \frac{1}{s_n} \int \psi(x) V_n(x) [1 - V_n(x)]\, dx = 0. \tag{20}$$

Étant donné que H_n est égal à $\frac{n}{s_n\sqrt{2}}$ on voit que sous cette condition l'espérance mathématique de $|A_n - B_n|$ reste au-dessous d'une borne qui tend vers zéro si n augmente infiniment. Ainsi le résultat est établi que

$$\lim_{n \to \infty} P_n(X) = \frac{1}{\sqrt{\pi}} \int^{X} e^{-u^2}\, du, \tag{21}$$

si $P_n(X)$ signifie la probabilité de l'inégalité $f\{S_n(x)\} \leqq \frac{X}{H_n}$.

Résumons les conditions auxquelles $f\{S_n(x)\}$ doit répondre pour que (21) soit valable :

1° La fonction statistique $f\{S_n(x)\}$ est deux fois dérivable;

2° La première dérivée $f'\{S_n(x), y\}$ remplit les conditions qui découlent du théorème classique pour les fonctions linéaires;

3° La deuxième dérivée $f''\{S_n(x), y, z\}$ remplit la condition exprimée par (19) et (20).

Ajoutons que pour les fonctions statistiques de la forme $f = F(A, B, \ldots)$ où A, B, ... sont des simples intégrales de Stieltjes (2), la troisième condition prend une forme plus maniable qui permet de décider facilement dans quel cas elle est remplie. Citons enfin la fonction statistique

$$\omega^2\{V(x)\} = \int \lambda(x) [S_n(x) - V_n(x)]^2\, dx$$

comme exemple d'une fonction dont la première dérivée s'annule identiquement, de sorte que notre théorème n'y est pas applicable.

RÉFÉRENCES.

a. *Ergänzungssätze zu den Gesetzen der grossen Zahlen.* Dans mon livre : (*Wahrscheinlichkeitsrechnung und ihre Anwendungen*, Leipzig und Wien, Deuticke 1931, p. 102-197). Ici les lois des grands nombres pour les fonctions statistiques (sans introduire ce nom) sont déduites pour le cas de distributions discontinues.

b. Deux nouveaux théorèmes de limite dans le calcul des probabilités (*Revue de la Faculté des Sciences de l'Université d'Istanbul*, t. I, 1935, fasc. 1, p. 61-80). Les deux théorèmes de limite pour les fonctions statistiques sont signalés sans démonstrations.

c. Die Gesetze der grossen Zahl für statistische Funktionen. [*Monatshefte für Mathematik und Physik*, Bd. 43, (Wirtinger-Festband), 1936, p. 105-128.] Démonstrations complètes des lois des grands nombres (y compris la loi forte des grands nombres) pour les fonctions statistiques.

d. Les lois des probabilités pour les fonctions statistiques (*Annales de l'Institut Henri Poincaré*, vol. 6, 1936, p. 185-212). Démonstration complète du premier des théorèmes de limite.

R. DE MISES
(Istanbul).