CALCUL DES PROBABILITÉS. — *Sur les rapports entre la quantité d'information au sens de Fisher et au sens de Wiener.* Note de M. **Marcel Paul Schutzenberger**, présentée par M. Émile Borel.

Limitons-nous au cas discret pour plus de simplicité et soit ξ une variable aléatoire pouvant prendre les états $A_i (i \in I)$ avec les probabilités respectives $Pr(A_i) = a_i$. Selon qu'il s'agit de statistique mathématique ou de théorie de l'information, la locution « *quantité d'information attachée à l'observation de* ξ » a deux sens bien différents. On donnera ici le principe d'une démonstration faisant apparaître ces deux expressions comme des cas particuliers de la solution générale d'un système de deux équations fonctionnelles qu'il est naturel d'imposer *a priori* à toute mesure de l'information (1).

Supposons que les a_i dépendent d'un paramètre inconnu θ qu'il s'agit d'estimer. D'après Fisher (2), la quantité d'information relative à θ attachée à l'observation de ξ est définie par :

$$H_F = \sum_{i \in I} \left(\frac{\partial a_i}{\partial \theta} \right)^2 \frac{1}{a_i} = - \sum_{i \in I} a_i \frac{\partial^2}{\partial \theta^2} \log a_i.$$

Sous certaines conditions très générales $1/H_F$ limite inférieurement la variance de toute valeur estimée $\hat{\theta}$ de θ.

D'après Wiener (3) et Shannon (4), la quantité d'information sur ξ est définie par :

$$H_W = - \sum_{i \in I} a_i \log_2 a_i,$$

H_W borne inférieurement le nombre moyen d'observations élémentaires nécessaires à la détermination de l'état ξ_0 pris par ξ dans une certaine épreuve (5).

(1) Une théorie entièrement différente (généralisant la probabilité au lieu de l'information) a été récemment développée par G. A. Barnard (*Roy. Stat. Soc.*, 31 janvier 1951). Cette théorie se place sur un plan différent de la nôtre et nous ne la discuterons pas ici.

(2) *Proc. Roy. Soc.*, A **146**, 1934, p. 1.

(3) *Cybernetics*, 1948, p. 47.

(4) *The Bell System Techn. Journ.*, **27**, 1948, p. 379.

(5) J. Ville et M. P. Schutzenberger, *Comptes rendus*, **232**, 1951, p. 206.

Soit maintenant $H(x)$ une mesure *quelconque* de la quantité d'information attachée à l'observation qui consiste à déterminer si l'état ξ_0 pris par ξ dans une certaine épreuve appartient ou non à l'ensemble $X(\Pr(\xi \in X) = x)$.

Soient X, Y et Z trois ensembles disjoints quelconques de probabilités respectives x, y et z partitionnant l'ensemble des $A_i (i \in I)(x + y + z = 1)$. Nous imposerons à H les trois conditions suivantes :

(1) *Continuité :* $H(x)$ est une fonctionnelle uniformément continue de x.

(2) *Symétrie :* $H(x) = H(1 - x)$.

(3) *Commutativité :*

$$H(x) + (1 - x)H(y/1 - x) = H(y) + (1 - y)H(x/1 - y) = H(x; y) = H(y; x).$$

(3) exprime qu'il est indifférent de déterminer d'abord si $\xi_0 \in X$ puis dans le cas contraire si $\xi_0 \in Y$ ou de déterminer d'abord si $\xi_0 \in Y$ puis dans le cas contraire si $\xi_0 \in X$. (2) et (3) impliquent que $H(x; y) = H(x; z) = H(y; z)$, et permettent d'associer à toute partition de l'ensemble des A_i en sous-ensembles disjoints une fonction symétrique de leurs probabilités qui est la quantité d'information attachée à la détermination de celui d'entre eux contenant ξ_0. Posons $K(y; z) = (1 - x)H(y/(1 - x))$. (3) devient :

$$K(x; y + z) + K(y; z) = K(y; z + x) + K(z; x) = K(z; x + y) + K(x; y)$$

dont la solution générale *continue* est : $K(x; y) = f(x) + f(y) - f(x + y)$. L'homogénéité de degré un de H impliquée par (3), montre que $f(x)$ doit être de la forme $x\mathrm{D}\log x$ où D est un *opérateur linéaire quelconque ;* d'où :

$$H(x) = x\mathrm{D}\log x + (1 - x)\mathrm{D}\log(1 - x).$$

On vérifie que pour H_F, $D = (\partial^2/\partial\theta^2)$ et pour H_W, $D = -1/\log 2$.

Généralisation. — A d'autres formes de D correspondent des généralisations utiles du concept d'information. Citons seulement :

a. Soient $r_{ji} (i \in I)$ des variables aléatoires indépendantes ne prenant que les valeurs 0 ou 1. Soit $\Pr(r_{ji} = 1) = p_i$. Les fonctions rationnelles $P_1, P_2, \ldots, P_\nu$ des p_i étant les probabilités des différents résultats d'une observation $\mathcal{O}$ sur les valeurs prises par les r_{ji} dans une épreuve, le nombre moyen de r_{ji} dont $\mathcal{O}$ permet d'affirmer qu'ils ont pris la valeur 1 est donné par

$$H_1 = \sum_{i=1} P_i D_3 D_2 D_1 \log P_i$$

où D_1 consiste à poser $p_1 = p_2 = \ldots = p_i = \varpi$; D_2, à diviser par $\log(1 - \varpi)$; D_3, à faire $\varpi = 1$. H_1 permet d'étudier le nombre moyen de r_{ji} ayant pris la

valeur 1 que l'on peut déterminer par une suite donnée d'observations indépendantes ou non.

b. Associons à tout A_i deux fréquences a_i et a'_i et considérons l'expression

$$H(a_i;;a'_i) = a_i \log \frac{a_i}{a'_i} + (1 - a_i) \log \frac{1 - a_i}{1 - a'_i} = H(1 - a_i;;1 - a'_i)$$

qui n'est nulle que pour $a_i = a'_i$ et qui peut être interprétée comme une pseudo-distance [puisque $H(a_i;;a'_i) \neq H(a'_i;;a_i)$] entre les deux fréquences. Manifestement cette expression est aussi une information généralisée satisfaisant à (1), (2) et (3). Si les a'_i sont des *fréquences théoriques* et les a_i les *fréquences correspondantes observées* dans une certaine suite de N épreuves indépendantes, cette expression est proportionnelle au logarithme de la vraisemblance des observations. On remarquera que l'expression $\Xi[a_i;;a'_i] = (a_i - a'_i)^2/a'_i(1 - a'_i)$ qui correspond à la mesure de concordance par χ^2 satisfait à une relation analogue mais distincte de (3) :

$$(3') \qquad \Xi[x;;x'] + (1 - x')\,\Xi\left[\frac{y}{1 - x};;\frac{y'}{1 - x'}\right]\frac{(y + z)^2}{(y' + z')^2} = \frac{\chi^2}{N},$$

où χ^2 est symétrique en les x, y et z.

(Extrait des *Comptes rendus des séances de l'Académie des Sciences*, t. 232, p. 925-927, séance du 5 mars 1951.)

GAUTHIER-VILLARS, IMPRIMEUR-LIBRAIRE DES COMPTES RENDUS DES SÉANCES DE L'ACADÉMIE DES SCIENCES.
13947-51 Paris. — Quai des Grands-Augustins, 55.