Distance - discrimination et résumé exhaustif

B. P. Adhikari, D. D. Joshi

DISTANCE - DISCRIMINATION ET RÉSUMÉ EXHAUSTIF

par

B. P. ADHIKARI et D. D. JOSHI

INTRODUCTION

A. RÉSUMÉ HISTORIQUE

C'est avec Karl Pearson (voir [20]) qu'est commencée l'étude d'une distance entre deux univers statistiques. Pearson a proposé une distance entre deux ensembles de mesures anthropométriques, et il l'a appelée le "coefficient de similarité raciale". Soit

$$((x_{ij})) \qquad i=1, \ldots, p;\ j=1, \ldots, n_{1i}$$

et

$$((y_{ik})) \qquad i=1, \ldots, p;\ k=1, \ldots, n_{2i}$$

deux tels ensembles de mesures où les mesures du caractère i viennent de n_{1i} éléments dans le premier ensemble, et de n_{2i} éléments dans le deuxième. Dans les objets anthropologiques, les crânes anciens, par exemple, il y a souvent des cassures et des déformités, ce qui fait qu'on ne peut toujours mesurer toutes les grandeurs de tous ces objets d'un ensemble; c'est pourquoi les nombres n_{1i} $(i=1, \ldots, p)$ ainsi que les n_{2i} peuvent ne pas être tous égaux. Si x_i et y_i sont les valeurs moyennes du caractère i dans les deux ensembles et si s_i est son écart type quand les deux ensembles sont groupés, alors ledit coefficient (CRL = "coefficient of racial likeness") s'écrit

$$\mathrm{CRL} = \frac{1}{p} \sum_{i=1}^{p} \frac{n_{1i}\, n_{2i}}{n_{1i} + n_{2i}} \left(\frac{\overline{x}_i - \overline{y}_i}{s_i} \right)^2 ;$$

$$= \frac{1}{p} \frac{n\ n}{n_1 + n_2} \sum_{i=1}^{p} \left(\frac{\overline{x}_i - \overline{y}_i}{s_i} \right)^2$$

si $n_{1i} = n_1$ et $n_{2i} = n_2$ $(i=1, \ldots, p)$. Les deux faiblesses du CRL sont que, premièrement, il ne tient pas compte des corrélations entre les caractères, et que, deuxièmement, son espérance mathématique dépend fortement des nombres d'observations.

La question de choix entre deux hypothèses se trouve déjà traitée par Student quand il a défini son test de signification pour la moyenne d'une loi laplacienne. Le problème de tests, avec la considération explicite des hypothèses alternatives, a connu un grand développement dans les travaux de Neyman et E.S. Pearson [17, 18] mais l'insistance a toujours été mise sur ce qu'on appelle "l'hypothèse nulle". En 1936, Mahalanobis [11] a proposé

une mesure de distance entre deux lois laplaciennes à plusieurs variables et avec la même matrice de dispersion. Cette distance, qu'on désigne par D^2, est une sorte de généralisation de la fonction "t" de Student, et représente en même temps une amélioration du CRL de Karl Pearson. Soient $((x_{ij}))$ et $((y_{ik}))$ deux ensembles de mesures comme plus haut, $\overline{x_i}$, $\overline{y_i}$, les valeurs moyennes du caractère i dans les deux ensembles respectivement, et soit $((a_{ij}))$ la matrice $p \times p$ de dispersion quand les deux ensembles sont réunis. Soit $((a^{ij}))$ l'inverse de la matrice $((a_{ij}))$. Alors

$$D^2 = \sum_{i,j} a^{ij} (\overline{x_i} - \overline{y_i}) (\overline{x_j} - \overline{y_j})$$

Si μ_i, ν_i, α_{ij} $(i,j=1,\ldots,p)$ représentent les moyennes et les éléments de la matrice de dispersion (supposée commune) des deux univers laplaciens à p variables dont les échantillons ont été tirés, et si l'on pose

$$\Delta^2 = \sum_{i,j} \alpha^{ij} (\mu_i - \nu_i)(\mu_j - \nu_j)$$

comme la distance entre ces deux univers, alors on montre que l'espérance mathématique de D^2 est, à un facteur près, très peu différente de Δ^2. Ce facteur qui est une fonction de nombres d'observations, est donc incorporé dans la définition de D^2.

L'étude systématique de ce qu'on appelle "l'analyse discriminante" a été commencée par Fisher [5,6,7] en 1936. Il s'est posé le problème suivant : Soient à notre disposition les deux ensembles $((x_{ij}))$ et $((y_{ik}))$ d'observations, que l'on sait d'être deux échantillons de deux univers différents; si $(x_1, \ldots, x_p)$ sont les mesures des p caractères sur un nouvel élément, auquel des deux univers cet élément appartient-il?

Désignons par z_i le caractère i. Alors Fisher prend une fonction linéaire

$$z = \sum_{i=1}^{p} \lambda_i z_i$$

à coefficients λ_i arbitraires, et il choisit ces λ_i de telle manière que z soit une fonction qui différencie le mieux les deux univers. Les z_i sont les x_i dans le premier ensemble, et les y_i dans le second.

Soit W la variation interne de z dans les deux ensembles réunis, et B la variation entre les deux moyennes de z. Le rapport B/W représente d'une certaine manière la capacité de z de différencier les deux univers. On choisit les λ_i pour rendre B/W maximum. Dans ses études sur la discrimination, Fisher a montré que les travaux antérieurs (comme celui de Mahalanobis, par exemple) sont des cas particuliers de cette analyse discriminante.

S'inspirant des méthodes de Neyman et E.S. Pearson, Welch [21] a traité le problème de discrimination dans sa généralité de la manière suivante : Etant donné n observations $x_1, \ldots, x_n$, et deux densités de probabilité $f_1(x_1, \ldots, x_n)$ et $f_2(x_1, \ldots, x_n)$, le problème est de décomposer l'espace euclidien à n dimensions en deux régions disjointes, l'une comme région d'acceptation de f_1, et l'autre celle de f_2. Cette décomposition doit être faite suivant un critère donné d'optimalité. Welch a obtenu des méthodes de décomposition suivant deux critères différents : l'erreur commune minimum, et l'erreur totale minimum.

Rao [19] a généralisé cette analyse au cas de discrimination entre plus de deux densités de probabilité. Dans les méthodes de Welch comme dans

celles de Rao, on aboutit aux régions d'acceptation définies par des inégalités sur les rapports de vraisemblance. Or ces rapports constituent des résumés exhaustifs pour le choix entre diverses hypothèses, comme l'a montré Mourier [16].

La définition d'une distance entre deux lois de probabilité n'a pas toujours été donnée en liaison avec le problème de discrimination proprement dit. Le D^2 de Mahalanobis n'a trouvé sa justification au point de vue discrimination que dans les travaux postérieurs de Fisher. Depuis quelques années, avec Mourier, Rao et Matusita [12,13] pour les variables aléatoires réelles, et avec Kullback et Leibler [9] et Komogoroff pour les espaces abstraits, on voit l'accent mis sur cette liaison entre les distances et les erreurs de discrimination.

L'utilité d'une "distance" sur les ensembles de lois de probabilité est évidente dans les études de convergence de ces lois. Une première définition en vue de telles études est celle de Lévy [10] qui prend comme distance entre deux lois de répartition de probabilité la plus grande valeur de l'intervalle entre les deux courbes dans la direction de la droite $x + y = a$. La question de convergence sera hors de notre présente étude.

L'introduction des fonctions aléatoires et d'autres éléments aléatoires plus abstraits dans la théorie de la probabilité rend nécessaire la généralisation des méthodes de discrimination et des définitions de distance aux espaces abstraits sur lesquels sont définies des mesures de probabilité.

Dans la présente étude nous nous proposons donc d'examiner la plupart des distances et des "divergences" connues jusqu'à présent entre deux mesures de probabilité, pour essayer de voir leurs propriétés importantes et leur interrelations, s'il en existe. Nous nous plaçons dans le cas général d'un espace abstrait probabilisé $(X, \mathcal{S}, \mathfrak{M})$ où $\mathfrak{M}$ est une classe de mesures de probabilité définies sur un corps borélien $\mathcal{S}$ de sous-ensembles de X. La distance de Lévy ne se prêtant pas à une généralisation à un espace abstrait n'a pas été considérée.

B. PROPRIETES SOUHAITABLES D'UNE DISTANCE

Une distance définie sur l'ensemble $\mathfrak{M}$ de mesures de probabilité doit, pour être utile dans les applications, être liée à la facilité avec laquelle on peut distinguer deux lois l'une de l'autre, c'est-à-dire que la distance doit être en relation avec l'erreur de discrimination tel que plus la distance est grande, moins soit l'erreur (voir Mourier [16]). Comme l'erreur de discrimination dépend de la méthode de discrimination, alors le choix de la distance en dépendra aussi.

Il est souhaitable qu'une fonction, pour être une "distance", satisfasse aux propriétés d'une métrique, et qu'elle prenne toujours des valeurs finies. Mais d'après ce que nous venons de dire, la propriété d'être une métrique, à elle seule, n'est pas suffisante. Il faut que la distance soit une fonction décroissante de l'erreur, et qu'elle prenne sa valeur maximum pour deux mesures singulières ($\mu \perp \nu$).

Une transformation mesurable n'est qu'une extension du concept de fonction d'observations. Comme de telles fonctions ne contiennent pas plus d'informations que les observations elles-mêmes, elles ne peuvent pas diminuer l'erreur. Donc nous postulons qu'une distance sur l'espace des mesures ne doit pas augmenter sous les transformations mesurables à moins que cette transformation ne soit un résumé exhaustif.

Si l'on augmente le nombre des observations on doit être capable de distinguer entre deux mesures avec plus de précision. Ceci nous amène à un autre critère, à savoir : la distance doit augmenter dans les espaces produits de $(X, \mathcal{S})$ avec lui-même.

PRÉLIMINAIRES

1. Soit X un ensemble d'éléments x de nature quelconque. Nous appellerons x un point de l'espace X. Si $\mathcal{S}$ est une famille de sous-ensembles de X telle que

(i) si $E \in \mathcal{S}$, $F \in \mathcal{S}$, alors $E \cup F \in \mathcal{S}$ et $E \cap F \in \mathcal{S}$;

(ii) si $E \in \mathcal{S}$, alors $\overline{E} \in \mathcal{S}$, où $\overline{E}$ est l'ensemble complémentaire de E par rapport à X,

alors $\mathcal{S}$ est un corps de sous-ensembles de X. Il est évident que si $\mathcal{S}$ est un corps et si E_i $(i=1,\ldots,n)$ sont n ensembles appartenant à $\mathcal{S}$, alors

$$\bigcup_{i=1}^{n} E_i \in \mathcal{S}$$

Si $\mathcal{S}$ est telle que l'union de toute suite dénombrable $\{E_i\}$ d'ensembles de $\mathcal{S}$ appartient à $\mathcal{S}$, alors $\mathcal{S}$ est un corps borélien (ou σ-corps). Un espace X sur lequel est défini un corps borélien $\mathcal{S}$ de sous-ensembles de X sera appelé un espace mesurable, et nous le désignerons par le symbole $(X, \mathcal{S})$. On appelle ensemble mesurable tout sous-ensemble de X appartenant à $\mathcal{S}$.

2. Une fonction d'ensemble μ définie sur $\mathcal{S}$ est une mesure si elle est non-négative et complètement additive; μ est une mesure finie si $\mu(X) < \infty$; μ est une mesure σ-finie si l'on peut trouver une suite dénombrable ou finie $\{E_i\}$ d'ensembles mesurables telle que $\bigcup E_i = X$, et que $\mu(E_i) < \infty$ pour tout i. La mesure μ est une mesure de probabilité si $\mu(X) = 1$.

3. Soit $(X, \mathcal{S})$ et $(Y, \mathcal{T})$ deux espaces mesurables. On désigne par $X \otimes Y$ l'espace produit de X et de Y constitué par tout couple (x,y) où $x \in X$ et $y \in Y$. De même pour tout $E \in \mathcal{S}$ et $F \in \mathcal{T}$ $E \otimes F$ désigne l'ensemble de points (x,y) tels que $x \in E$ et $y \in F$. Le plus petit corps borélien sur $X \otimes Y$ qui contient tous les rectangles est désigné par $\mathcal{S} \otimes \mathcal{T}$. Si μ et ν sont des mesures définies sur $\mathcal{S}$ et $\mathcal{T}$ respectivement, on obtient une mesure $\mu \otimes \nu$ sur les ensembles de $\mathcal{S} \otimes \mathcal{T}$.

4. Soit μ, ν deux mesures telles que pour tout ensemble E pour lequel $\mu(E) = 0$, on ait $\nu(E) = 0$. Alors ν est dite absolument continue par rapport à μ, et on écrit

$$\nu \ll \mu$$

Si l'on a à la fois $\nu \ll \mu$ et $\mu \ll \nu$, alors on dit que μ et ν sont équivalentes, et on écrit $\mu \equiv \nu$.

Si pour μ et ν, il existe deux ensembles disjoints A et B tels que $A \cup B = X$, que pour tout ensemble mesurable E les produits $(A \cap E)$ et $(B \cap E)$ sont mesurables, et que

$$\mu(A \cap E) = \nu(B \cap E) = 0,$$

alors on dit que μ et ν sont singulières l'une par rapport à l'autre. Nous écrivons $\nu \perp \mu$.

5. Une fonction réelle f(x), de point x, est une fonction mesurable $(\mathcal{S})$ si pour tout nombre réel c, l'ensemble $\{x : f(x) \leq c\}$ appartient à $\mathcal{S}$.

On sait que

(Théorème de Radon-Nikodym) si $\nu \ll \mu$, alors il existe une fonction mesurable $(\mathcal{S})$, f(x), telle que $0 < f(x) < \infty$, et que pour tout $E \in \mathcal{S}$,

$$\nu(E) = \int_E f(x)\, d\mu(x)$$

Remarques : (i) La fonction f(x) est unique dans ce sens que s'il existe une autre fonction g(x), mesurable ($\mathcal{S}$), et telle que

$$\nu(E) = \int_E g(x)\, d\mu(x), \qquad E \in \mathcal{S}$$

alors

$$\mu\{x : f(x) \neq g(x)\} = 0$$

on écrit $d\nu(x) = f(x)\, d\mu(x)$ et aussi $f(x) = d\nu(x)/d\mu(x)$.

(ii) Toute relation suivie du symbole $[\mu]$ signifie que cette relation est vraie à tout point sauf aux points d'un ensemble de mesure-μ nulle.

6. Soit $\mathcal{M}$ une classe de mesures de probabilité définies sur $\mathcal{S}$. $\mathcal{M}$ est une classe de mesures dominée s'il existe une mesure λ (σ-finie) sur $\mathcal{S}$ telle que $\mu \ll \lambda$ pour toute $\mu \in \mathcal{M}$. La classe $\mathcal{M}$ est dite équivalente à une mesure λ ($\mathcal{M} \equiv \lambda$) si toutes les mesures dans $\mathcal{M}$ sont absolument continues par rapport à λ et que $\lambda(E) = 0$ pour tout ensemble $E \in \mathcal{S}$ pour lequel $\mu(E) = 0$ pour chaque $\mu \in \mathcal{M}$ simultanément.

Toute classe $\mathcal{M}$ dénombrable ou finie est dominée. Halmos et Savage [8] ont démontré que si $\mathcal{M}$ est dominée, il existe une mesure de probabilité λ telle que $\mathcal{M} \equiv \lambda$.

Dans ce qui suit nous supposerons que la famille $\mathcal{M}$ est toujours dominée et donc qu'il y aura toujours une mesure de probabilité (désignée toujours par λ) telle que $\mathcal{M} \equiv \lambda$.

7. Soit $(X, \mathcal{S})$ et $(Y, \mathcal{T})$ deux espaces mesurables dont les points sont désignés par x et y respectivement.

Une transformation $T(x) = y$ de $(X, \mathcal{S})$ sur $(Y, \mathcal{T})$ détermine une correspondance entre les points de X et ceux de Y de telle manière qu'à chaque point de X correspond un seul point de Y, et qu'à chaque point de Y correspond au moins un point de X.

Etant donné un ensemble $F \in \mathcal{T}$, l'ensemble de tous les points de X tels que $T(x) \in F$ est appelé l'image inverse de F, désigné par $T^{-1}(F)$. La transformation T est une transformation mesurable si pour tout $F \in \mathcal{T}$

$$T^{-1}(F) \in \mathcal{S}$$

L'ensemble de toutes les images inverses $T^{-1}(F)$, où $F \in \mathcal{T}$, constitue lui aussi un corps borélien $T^{-1}(\mathcal{T})$ de sous-ensembles de X. On a évidemment, $T^{-1}(\mathcal{T}) \subseteq \mathcal{S}$

8. Pour toute fonction g(y) on peut définir une fonction gT(x) sur X telle que $gT(x) = g(Tx)$. Si g(y) est mesurable ($\mathcal{T}$) alors gT(x) est mesurable ($T^{-1}(\mathcal{T})$). Pour toute mesure μ sur $\mathcal{S}$, on peut définir une mesure μT^{-1} sur $\mathcal{T}$ par la relation

$$\mu T^{-1}(F) = \mu(T^{-1} F), \quad F \in \mathcal{T}$$

Les lemmes suivantes sont dues à Halmos et Savage.

Lemme 1. Si f(x) est une fonction réelle sur X, alors une condition nécessaire et suffisante pour qu'il existe une fonction g(y) mesurable ($\mathcal{T}$) telle que $f(x) = gT(x)$ est que f(x) soit mesurable ($T^{-1}(\mathcal{T})$).

Lemme 2. Si g(y) est une fonction réelle sur Y, et μ une mesure sur $\mathcal{S}$, alors

$$\int_F g(y)\, d\mu\, T^{-1}(y) = \int_{T^{-1}(F} gT(x)\, d\mu(x), \qquad F \in \mathcal{T}$$

en ce sens que si l'une des deux intégrales existe, l'autre existe aussi et les deux sont égales.

9. Si μ est une mesure de probabilité définie sur $(X, \mathcal{S})$, alors toute fonction mesurable $(\mathcal{S})$, $f(x)$, est appelée une variable aléatoire. Soit ν l'intégrale indéfinie de $f(x)$ par rapport à la mesure μ, supposant que $f(x)$ est aussi intégrable. C'est-à-dire,

$$\nu(E) = \int_E f(x)\, d\mu(x), \qquad E \in \mathcal{S}$$

Alors ν est une mesure, et $\nu \ll \mu$.

Soit maintenant T une transformation mesurable de $(X, \mathcal{S})$ sur $(Y, \mathcal{T})$. On a

$$\nu T^{-1} \ll \mu T^{-1}$$

Il existe donc une fonction mesurable $(\mathcal{T})$, $g(y)$, telle que

$$\nu T^{-1}(F) = \int_F g(y)\, d\mu T^{-1}(y), \qquad F \in \mathcal{T}$$

On appelle $g(y)$ l'espérance mathématique conditionnelle de $f(x)$ par rapport à y, et on la désigne par $e_\mu(f/y)$. Si $f(x) = X_E(x)$ est l'indicatrice de l'ensemble $E \in \mathcal{S}$, alors $g(y)$ est la probabilité conditionnelle de E par rapport à y, désignée par $p_\mu(E/y)$.

10. Une transformation mesurable T est appelée un résumé exhaustif [3, 4, 8] pour la classe de mesures $\mathcal{M}$ si la probabilité conditionnelle $p_\mu(E/y)$ d'un ensemble $E \in \mathcal{S}$ par rapport à y est indépendante de μ; c'est à dire que si pour tout $E \in \mathcal{S}$ il existe une fonction $p(E/y)$ de y telle que

$$p_\mu(E/y) = p(E/y) \quad \left[\mu T^{-1}\right],$$

et cela pour toute mesure $\mu \in \mathcal{M}$. Alors nous avons le théorème suivant de Halmos et Sauvage [8].

Théorème 1. Une condition nécessaire et suffisante pour que la transformation T soit un résumé exhaustif pour une classe $\mathcal{M}$ de mesures (de probabilité) dominée est qu'il existe une mesure λ sur $\mathcal{S}$ telle que $\mathcal{M} \equiv \lambda$ et que $d\mu/d\lambda$ soit mesurable $(T^{-1}(\mathcal{T}))$ pour toute $\mu \in \mathcal{M}$.

Soit T un résumé exhaustif, et soit $f_\mu(x)$ la dérivée de Radon-Nikodym $\frac{d\mu}{d\lambda}$. Si l'on désigne par $g_\mu(y)$ l'espérance mathématique conditionnelle de $f_\mu(x)$ par rapport à y, alors Halmos et Savage ont démontré que

$$f_\mu(x) = g_\mu T(x) \quad [\lambda]$$

I. "L'AFFINITE" DE BHATTACHARYA ET LES DISTANCES Y ASSOCIÉES

Soit $\mu \in \mathcal{M}$, $\nu \in \mathcal{M}$ deux mesures sur $(X, \mathcal{S})$ et soit

$$d\mu = f_\mu(x) d\lambda_0 \qquad d\nu = f_\nu(x)\, d\lambda_0$$

Alors la fonction

$$\rho(\mu, \nu) = \int_X \sqrt{f_\mu(x)\, f_\nu(x)}\, d\lambda_0(x)$$

a été introduite par Bhattacharya [1] pour l'étude de la divergence entre deux lois de probabilité. Cette fonction mesure, d'une certaine manière, la proximité entre deux mesures de probabilité. Il est facile d'énumérer les propriétés suivantes de $\rho(\mu,\nu)$.

(i) $\rho(\mu,\nu) = \rho(\nu,\mu)$

(ii) $0 \leqslant \rho(\mu,\nu) \leqslant 1$

(iii) $\rho(\mu,\nu) = 1$ si et seulement si $\mu = \nu$

(iv) $\rho(\mu,\nu) = 0$ si et seulement si $\mu \perp \nu$

En vue de ces propriétés nous appelons $\rho(\mu,\nu)$, d'après Matusita [12] "l'affinité" entre μ et ν. Une fonction décroissante convenable de $\rho(\mu,\nu)$ peut se suggérer comme une distance entre μ et ν. Nous allons en étudier quelques-unes. Avant cela, nous allons établir quelques propriétés de $\rho(\mu,\nu)$ qui seront utilisées par la suite.

Lemme 3. Pour tout ensemble $E \in \mathcal{T}$,

$$\int_E \sqrt{f_\mu(x)\, f_\nu(x)}\, d\lambda_o(x) \leqslant \sqrt{\mu(E)\,\nu(E)}$$

On a l'inégalité de Holder :

$$\int_E |fg|\, d\lambda_o \leqslant \left(\int_E |f|^p\, d\lambda_o\right)^{1/p} \left(\int_E |g|^q\, d\lambda_o\right)^{1/q}$$

où f et g sont deux fonctions intégrables quelconques, $1/p + 1/q = 1$ et $E \in \mathcal{T}$.

Mettons $f = \sqrt{f_\mu(x)}$, $g = \sqrt{f_\nu(x)}$, $p = q = 2$. On obtient

$$\int_E \sqrt{f_\mu f_\nu}\, d\lambda_o \leqslant \left(\int_E f_\mu\, d\lambda_o\right)^{1/2} \left(\int_E f_\nu\, d\lambda_o\right)^{1/2} = \sqrt{\mu(E)\,\nu(E)}$$

Théorème 2. Soit T une transformation mesurable de $(X,\mathcal{T})$ sur $(Y,\mathcal{T})$, et soient $g_\mu(y)$ et $g_\nu(y)$ les fonctions définies par

$$d\mu T^{-1} = g_\mu(y)\, d\lambda_o T^{-1}$$
$$d\nu T^{-1} = g_\nu(y)\, d\lambda_o T^{-1}$$

soit

$$\rho_T(\mu,\nu) = \rho(\mu T^{-1}, \nu T^{-1}) = \int_Y \sqrt{g_\mu(y)\, g_\nu(y)}\, d\lambda_o T^{-1}(y)$$

Alors

$$\rho(\mu,\nu) \leqslant \rho_T(\mu,\nu)$$

On peut écrire

$$\rho_T(\mu,\nu) = \int_Y \sqrt{g_\mu(y)\, g_\nu(y)}\, d\lambda_o T^{-1}(y) = \int_Y \sqrt{g_\mu(y)/g_\nu(y)}\, d\nu T^{-1}(y)$$

Par la lemme 2,

$$\rho_T(\mu,\nu) = \int_X \sqrt{g_\mu T(x)/g_\nu T(x)}\, d\nu(x)$$

où $g_\mu T(x)$ et $g_\nu T(x)$ sont mesurables $(T^{-1}(\mathcal{T}))$. Mettons

$$g(x) = \frac{g_\mu T(x)}{g_\nu T(x)}$$

et soit $G_k^{(n)}$ l'ensemble

$$G_k^{(n)} = \left\{ x : \frac{k}{2^n} < \sqrt{g(x)} \leqslant \frac{k+1}{2^n} \right\} , \quad k = 0, 1, 2, \ldots$$

Alors

$$G_k^{(n)} \in T^{-1}(\mathcal{C}) , \quad \bigcup_k G_k^{(n)} = X$$

Soit

$$s_n = \sum_{k=0}^{\infty} \frac{k}{2^n} \nu\,(G_k^{(n)})$$

alors

$$\lim_{n\to\infty} s_n = \int_X \sqrt{g(x)}\, d\nu(x)$$

Or, pour $x \in G_k^{(n)}$, on a

$$\frac{k^2}{2^{2n}} < g(x) \leqslant \frac{(k+1)^2}{2^{2n}}$$

Donc

$$\frac{k^2}{2^{2n}}\, \nu\,(G_k^{(n)}) < \int_{G_k} g(x)\, d\nu(x) \leqslant \frac{(k+1)^2}{2^{2n}}\, \nu\,(G_k^{(n)})$$

C'est-à-dire

$$\frac{k^2}{2^{2n}} < \frac{\mu(G_k^{(n)})}{\nu(G_k^{(n)})} \leqslant \frac{(k+1)^2}{2^{2n}}$$

ou

$$\frac{k}{2^n} < \sqrt{\frac{\mu(G_k^{(n)})}{\nu(G_k^{(n)})}} \leqslant \frac{k+1}{2^n}$$

On a donc

$$\int_X \sqrt{g(x)}\, d\nu(x) = \lim_{n\to\infty} s_n = \lim_{n\to\infty} \sum_{k=0}^{\infty} \sqrt{\mu(G_k^{(n)})\, \nu\,(G_k^{(n)})}$$

Alors, par la lemme 3,

$$\rho(\mu,\nu) = \int_X \sqrt{f_\mu(x)\, f_\nu(x)}\, d\lambda_0(x) = \sum_{k=0}^{\infty} \int_{G_k^{(n)}} \sqrt{f_\mu(x)\, f_\nu(x)}\, d\lambda_0(x)$$

$$\leqslant \lim_{n\to\infty} \sum_{k=0}^{\infty} \sqrt{\mu\,(G_k^{(n)}) y(G_k^{(n)})} = \rho_T\,(\mu,\nu)$$

Halmos et Savage, dans leur démonstration du théorème cité plus haut, ont montré que si la transformation T est un résumé exhaustif pour $\mathcal{M}$, alors pour toute μ,

$$f_\mu(x) = g_\mu\, T(x) \quad [\lambda_0]$$

Donc, il est évident que l'affinité $\rho(\mu,\nu)$ reste invariable sous une transformation T qui est un résumé exhaustif.

Une autre propriété de $\rho(\mu,\nu)$ est la suivante :

Soit $(X \otimes X, \mathcal{J} \otimes \mathcal{J})$ le produit cartésien de $(X, \mathcal{J})$ avec lui-même, et soient $\mu \otimes \mu$, $\nu \otimes \nu$ les mesures produites correspondant à μ et à ν. Alors il est facile de voir que

$$\rho\left(\mu^{(2)}, \nu^{(2)}\right) = \rho\left(\mu \otimes \mu\,,\ \nu \otimes \nu\right) = \left[\rho\,(\mu,\nu)\right]^2$$

et que, de la même manière

$$\rho(\mu^{(n)}, \nu^{(n)}) = \left[\rho(\mu,\nu)\right]^n$$

Nous allons étudier maintenant les propriétés de trois fonctions de $\rho(\mu,\nu)$ que l'on peut proposer comme une distance entre μ et ν.

A)
$$d_H(\mu,\nu) = 1-\rho(\mu,\nu)$$
$$= \frac{1}{2}\int_X \left[\sqrt{f_\mu(x)} - \sqrt{f_\nu(x)}\right]^2 d\lambda_o(x)$$

Cette fonction avait été proposée par Kolmogoroff dans ses conférences à l'Institut Henri Poincaré en Novembre 1955. Cette fonction a des propriétés suivantes :

1) $d_H(\mu,\nu) = d_H(\nu,\mu)$
2) $o \leq d_H(\mu,\nu) \leq i$
3) $d_H(\mu,\nu) = 0$ si et seulement si $\mu = \nu$
4) $d_H(\mu,\nu) = 1$ si et seulement si $\mu \perp \nu$

Celles-ci sont des propriétés d'une "métrique" mais l'exemple suivant montre que $d_H(\mu,\nu)$ ne satisfait pas à l'inégalité triangulaire.

Exemple. Soit $f_1(x)$, $f_2(x)$ deux densités de probabilité laplaciennes $N(\mu_1,\sigma_1)$, $N(\mu_2,\sigma_2)$. Alors si $\mu_1 = \mu_2$,

$$\rho(f_1, f_2) = \sqrt{2\sigma_1\sigma_2/(\sigma_1^2+\sigma_2^2)}$$

Prenons trois densités de probabilités laplaciennes

$$N(\mu,\sigma),\ N(\mu, 4\sigma),\ N(\mu, 5\sigma)$$

Alors

$$\rho(f_1, f_2) = 0,68$$
$$\rho(f_2, f_3) = 0,98$$
$$\rho(f_1, f_3) = 0,63$$

C'est-à-dire

$$\rho(f_1, f_2) + \rho(f_2, f_3) > 1 + \rho(f_1, f_3)$$

ou

$$d_H(f_1, f_2) + d_H(f_2, f_3) < d_H(f_1, f_3)$$

5) D'après le théorème 2,

$$1-\rho(\mu,\nu) \geq 1-\rho_T(\mu,\nu)$$

C'est-à-dire que $d_H(\mu,\nu)$ n'augmente pas sous les transformations mesurables et reste invariable si la transformation T est un résumé exhaustif.

6)
$$1-\rho(\mu^{(n)}, \nu^{(n)}) = 1-\left[\rho(\mu,\nu)\right]^n \geq 1-\rho(\mu,\nu)$$

c'est-à-dire

$$d_H(\mu^{(n)}, \nu^{(n)}) = 1-\left[\rho(\mu,\nu)\right]^n \geq d_H(\mu,\nu)\ ;$$

si $\rho(\mu,\nu) < 1$, alors

$$\lim_{n\to\infty} d_H(\mu^{(n)}\ \nu^{(n)}) \to 1$$

7) Pour voir la relation entre $d_H(\mu,\nu)$ et l'erreur de discrimination, considérons la décomposition suivante de l'espace X.

$$R = \{x : f_\mu(x) > f_\nu(x)\}$$
$$\overline{R} = \{x : f_\mu(x) \leq f_\nu(x)\}$$

On sait que cette décomposition rend minimum l'erreur totale, c'est-à-dire rend minimum l'expression

$$\mu(\overline{E}) + \nu(E), \quad E \in \mathcal{S}$$

Or

$$\rho(\mu,\nu) = \int_X \sqrt{f_\mu(x)\, f_\nu(x)}\, d\lambda_0(x)$$
$$= \int_R \sqrt{\frac{f_\mu(x)}{f_\nu(x)}}\, d\nu(x) + \int_{\overline{R}} \sqrt{\frac{f_\nu(x)}{f_\mu(x)}}\, d\mu(x)$$
$$\geq \nu(R) + \mu(\overline{R}) = \varepsilon \text{ (erreur)}$$

C'est-à-dire

$$\varepsilon \leq 1 - d_H(\mu,\nu)$$

B) La deuxième fonction de $\rho(\mu,\nu)$ prise comme une distance est celle proposée par Matusita [12] et s'écrit comme

$$d_j(\mu,\nu) = \left(\int_X \left(\sqrt{f_\mu(x)} - \sqrt{f_\nu(x)}\right)^2 d\lambda_0(x)\right)^{\frac{1}{2}}$$
$$= \left[2\,(1 - \rho(\mu,\nu))\right]^{\frac{1}{2}}$$

On vérifie facilement les propriétés suivantes de $d_j(\mu,\nu)$:

1) $d_j(\mu,\nu) = d_j(\nu,\mu)$

2) $0 \leq d_j(\mu,\nu) \leq \sqrt{2}$

3) $d_j(\mu,\nu) = 0$ si et seulement si $\mu = \nu$

4) $d_j(\mu,\nu) = \sqrt{2}$ si et seulement si $\mu \perp \nu$

5) $d_j(\mu,\nu) + d_j(\nu,\xi) \geq d_j(\mu,\xi)$

cette dernière est une conséquence directe de l'inégalité de Minkowski.

6) $d_j(\mu,\nu)$ n'augmente pas sous les transformations mesurables, et reste invariable si la transformation T est un résumé exhaustif.

7) $$d_j(\mu^{(n)},\nu^{(n)}) = \sqrt{2}\left[1 - \rho(\mu^{(n)},\nu^{(n)})\right]^{\frac{1}{2}}$$
$$= \sqrt{2}\left[1 - \{\rho(\mu,\nu)\}^n\right]^{\frac{1}{2}} \geq \sqrt{2}\left[1 - \rho(\mu,\nu)\right]^{\frac{1}{2}} = d_j(\mu,\nu)$$

En plus, si $\rho(\mu,\nu) < 1$,

$$\lim_{n\to\infty} d_j(\mu^{(n)}, \nu^{(n)}) \to \sqrt{2}$$

8) Soit, comme plus haut, ϵ l'erreur totale de discrimination. Alors

$$\frac{1}{2} d_j^2(\mu,\nu) = 1 - \rho \leq 1 - \epsilon$$

ou

$$\epsilon \leq 1 - \frac{1}{2} d_j^2(\mu,\nu)$$

La relation entre $d_j(\mu,\nu)$ et des erreurs dans les tests des hypothèses a été étudiée par Matusita et les autres [12, 13, 14, 15].

C) La troisième définition de distance découlant de "l'affinité" est

$$d_C(\mu,\nu) = -\log \rho(\mu,\nu)$$

Une forme générale de cette fonction a été étudiée par Chernoff [2]. Il propose comme "divergence" entre μ et ν la fonction

$$-\log\left[\inf_{0<t<1}\int_X (f_\mu(x))^t\,(f_\nu(x))^{1-t}\,d\lambda_g(x)\right]$$

La distance $d_C(\mu,\nu)$ satisfait à toutes les propriétés que nous venons d'étudier pour les distances précédentes; on doit toutefois faire les remarques suivantes :

1) $d_C(\mu,\nu) = \infty$ pour $\mu \perp \nu$, et reste finie dans les autres cas.

2) $d_C(\mu,\nu)$ ne satisfait pas l'inégalité triangulaire,

3) $d_C(\mu^{(n)}, \nu^{(n)}) = n\, d_C(\mu,\nu)$

II. DISTANCE A L'ERREUR COMMUNE MINIMUM DE RAO

Soit $(E,\overline{E})$ une décomposition mesurable de X telle que $E \cup \overline{E} = X$ et que

$$\mu(\overline{E}) = \nu(E) = \alpha_E$$

Si, dans le problème de discrimination entre μ et ν, l'on prend E comme région d'acceptation de μ et $\overline{E}$ celle de ν, alors les deux erreurs de discrimation sont égales à α_E.

Soit

$$\alpha = \inf_{E \in \mathcal{J}} \alpha_E$$

Alors Rao [19] définit la distance entre μ et ν comme

$$d_R(\mu,\nu) = 1 - \alpha$$

On peut énumérer les propriétés suivantes de $d_R(\mu,\nu)$

1) $d_R(\mu,\nu) = d_R(\nu,\mu)$

2) $\frac{1}{2} \leq d_R(\mu,\nu) \leq 1$

3) $d_R(\mu,\nu) = \frac{1}{2}$ si et seulement si $\mu = \nu$

4) $d_R(\mu,\nu) = 1$ si et seulement si $\mu \perp \nu$

5) $d_R(\mu,\nu) + d_R(\nu,\xi) \geq d_R(\mu,\xi)$ (voir Rao [19])

6° La distance n'augmente pas sous les transformations mesurables. En effet, soit

$$\frac{d\mu}{d\lambda} = f_\mu(x) \qquad \frac{d\nu}{d\lambda} = f_\nu(x)$$

Si $(A,\overline{A})$ est une décomposition de l'espace X définie par

$$A = \left\{ x : \frac{f_\mu(x)}{f_\nu(x)} \geq k \right\}, \quad \overline{A} = \left\{ x : \frac{f_\mu(x)}{f_\nu(x)} < k \right\}$$

où k est choisi tel que

$$\mu(A) = \nu(\bar{A})$$

alors l'erreur commune minimum α est donnée par

$$\alpha = \mu(\bar{A}) = \nu(A)$$

(voir Welch [21]).

Soit T une transformation mesurable de $(X, \mathcal{J})$ sur $(Y, \mathcal{T})$, et soit $(A_T, \bar{A}_T)$ la décomposition de Y définie par

$$A_T = \left\{ y : \frac{g_\mu(y)}{g_\nu(y)} \geqslant k' \right\}, \quad \bar{A}_T = \left\{ y : \frac{g_\mu(y)}{g_\nu(y)} < k' \right\}$$

où

$$d\mu\, T^{-1} = g_\mu(y)\, d\lambda\, T^{-1},$$

$$d\nu\, T^{-1} = g_\nu(y)\, d\nu\, T^{-1},$$

et où k' est choisi tel que

$$\mu T^{-1}(A_T) = \nu T^{-1}(\bar{A}_T)$$

Alors l'erreur commune minimum α_T dans l'espace $(Y, \mathcal{T})$ est donnée par

$$\alpha_T = \mu\, T^{-1}(\bar{A}_T) = \nu\, T^{-1}(A_T)$$

Considérons maintenant la décomposition $(T^{-1}(A_T), T^{-1}(\bar{A}_T))$ de $(X, \mathcal{J})$. On a évidemment

$$\mu(T^{-1}(A_T)) = \nu(T^{-1}(A_T)) = \alpha_T,$$

c'est-à-dire que pour cette décomposition les deux erreurs de discrimination sont égales. Comme α est la valeur minimum de cette erreur commune, alors

$$\alpha \leqslant \alpha_T$$

De plus il est évident que la décomposition $(T^{-1}(A_T), T^{-1}(\bar{A}_T))$ de $(X, \mathcal{J})$ rend l'erreur commune minimum si l'ensemble A pour lequel $\inf_{E \in \mathcal{J}} \alpha_E$ est atteint est mesurable $T^{-1}(\mathcal{T})$. Dans le cas où T est un résumé exhaustif, les régions $A, \bar{A}$ sont elles aussi mesurables $T^{-1}(\mathcal{T})$, puisque les fonctions $f_\mu(x)$, $f_\nu(x)$ sont mesurables $(T^{-1}(\mathcal{T}))$. Donc dans ce cas

$$\alpha = \alpha_T$$

Ainsi nous voyons que la distance $d_R(\mu, \nu)$ n'augmente pas sous les transformations mesurables et reste invariable pour les résumés exhaustifs.

7) Désignons par $\alpha^{(2)}$ l'erreur commune minimum de discrimination entre les mesures $\mu \otimes \mu$ et $\nu \otimes \nu$ sur l'espace produit $(X \otimes X, \mathcal{J} \otimes \mathcal{J})$. Soit $\mathcal{R}$ la famille de tous les ensembles $G \in \mathcal{J} \otimes \mathcal{J}$ pour lesquels

$$(\mu \otimes \mu)(\bar{G}) = (\nu \otimes \nu)(G) = \alpha_G^{(2)}$$

Alors, par définition

$$\alpha^{(2)} = \inf_{G \in \mathcal{R}} \alpha_G^{(2)}$$

Soit $\mathcal{R}'$ la famille de tous les ensembles G' de la forme

$$E \otimes X$$

où $E \in \mathcal{J}$, et

$$\mu(\overline{E}) = \nu(E) = \alpha_E$$

On a

$$(\mu \otimes \mu)(\overline{G'}) = (\nu \otimes \nu)(\overline{G'}) = \alpha^{(2)}_{G'} = \alpha_E$$

Alors comme $\mathcal{R}' \subseteq \mathcal{R}$, nous avons

$$\alpha = \inf_{E \in \mathcal{J}} \alpha_E = \inf_{G' \in \mathcal{R}'} \alpha^{(2)}_{G'} \geqslant \inf_{G \in \mathcal{R}} \alpha^{(2)}_{G} = \alpha^{(2)}$$

Donc, si l'on désigne la distance entre $\mu \otimes \mu$ et $\nu \otimes \nu$ par $d_R(\mu^{(2)}, \nu^{(2)})$, on a

$$d_R(\mu^{(2)}, \nu^{(2)}) = 1 - \alpha^{(2)} \geqslant 1 - \alpha = d_R(\mu, \nu)$$

Ainsi on voit que la distance d_R augmente avec le nombre d'observations.

Dans le cas des lois laplaciennes à plusieurs variables $d_R(\mu,\nu)$ est liée à la distance D^2 de Mahalanobis par la relation

$$d_R(\mu,\nu) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{D}{2}} e^{-\frac{1}{2}y^2} \, dy$$

III. DISTANCES ASSOCIÉES A "L'INFORMATION" DE SHANNON

Kullback et Leibler [9] ont généralisé l'information de Shannon et ont proposé la forme suivante :

L'information de discrimination de μ contre ν est

$$I(\mu,\nu) = \int_X \left[\log \frac{f_\mu(x)}{f_\nu(x)}\right] f_\mu(x) \, d\lambda_0(x).$$

De même

$$I(\nu,\mu) = \int_X \left[\log \frac{f_\nu(x)}{f_\mu(x)}\right] f_\nu(x) \, d\lambda_0(x).$$

La "divergence" entre μ et ν est alors définie par Kullback et Leibler comme

$$J(\mu,\nu) = I(\mu,\nu) + I(\nu,\mu).$$

Une définition analogue a été proposée par Mourier [16] :

Soit :

$$\sum\nolimits^2_\mu = \int_X \left[\log \frac{f_\mu(x)}{f_\nu(x)}\right]^2 f_\mu(x) \, d\lambda_0(x) - \left[I(\mu,\nu)\right]^2$$

et

$$\sum\nolimits^2_\nu = \int_X \left[\log \frac{f_\nu(x)}{f_\mu(x)}\right]^2 f_\nu(x) \, d\lambda_0(x) - \left[I(\nu,\mu)\right]^2$$

Alors la distance $d_M(\mu,\nu)$ se définit par

$$d_M(\mu,\nu) = \frac{I(\mu,\nu) + I(\nu,\mu)}{\sum_\mu + \sum_\nu}$$

Nous allons énumérer les propriétés de ces deux distances successivement.

(a) La "divergence" $J(\mu,\nu)$ de Kullback et Leibler.

1) $J(\mu,\nu) = J(\nu,\mu)$.

2) $J(\mu,\nu) \geqslant o$, et peut atteindre la valeur infinie même pour les mesures non-singulières.

3) $J(\mu,\nu) = 0$ si et seulement si $\mu = \nu$.

4) L'inégalité triangulaire n'est pas toujours satisfaite par $J(\mu,\nu)$.

Exemple : Si $f_1(x)$ et $f_2(x)$ désignent les densités de probabilité de deux lois laplaciennes avec la même moyenne μ et les variances σ_1^2 et σ_2^2, alors

$$J(1,2) = \frac{1}{2}\,\frac{(\sigma_1^2 - \sigma_2^2)}{\sigma_1^2\,\sigma_2^2}$$

Si l'on prend trois lois

$$N(\mu, \sigma),\ N(\mu, 4\sigma),\ N(\mu, 5\sigma),$$

alors

$$J(1,2) = 225/32, \qquad J(2,3) = 81/800, \qquad J(1,3) = 576/50.$$

Donc

$$J(1,2) + J(2,3) < J(1,3).$$

5) $J(\mu,\nu)$ n'augmente pas sous les transformations mesurables, mais reste invariable si et seulement si T est un résumé exhaustif.

6) $J(\mu^{(n)}, \nu^{(n)}) = n\,J(\mu,\nu)$. Cette relation n'a évidemment de sens que si $J(\mu,\nu)$ est finie.

(b) La distance de Mourier.

1) $d_M(\mu,\nu) = d_M(\nu,\mu)$.

2) $d_M(\mu,\nu) \geqslant o$, et peut être indéterminée.

3) $d_M(\mu,\nu) = o$ si et seulement si $\mu = \nu$.

4) L'inégalité triangulaire n'est pas toujours satisfaite par $d_M(\mu,\nu)$.

Exemple. Soit

$$f_1(x) = N(\mu_1, \sigma_1),\quad f_2(x) = N(\mu_2, \sigma_2),\quad f_3(x) = N(\mu_3, \sigma_3).$$

Alors

$$I(1,2) + I(2,1) = \frac{1}{2\,\sigma_1^2\,\sigma_2^2}\left[(\sigma_1^2 - \sigma_2^2)^2 + (\mu_1 - \mu_2)^2(\sigma_1^2 + \sigma_2^2)\right]$$

$$\sum\nolimits_1^2 = \sigma_1^2\left[\frac{(\mu_1 - \mu_2)^2}{\sigma_2^4} + \frac{1}{2}\,\frac{(\sigma_1^2 - \sigma_2^2)^2}{\sigma_1^2\,\sigma_2^4}\right]$$

$$\sum\nolimits_2^2 = \sigma_2^2\left[\frac{(\mu_1 - \mu_2)^2}{\sigma_1^4} + \frac{1}{2}\,\frac{(\sigma_1^2 - \sigma_2^2)^2}{\sigma_1^4\,\sigma_2^2}\right]$$

Si l'on prend comme exemple les valeurs suivantes :

$$(\mu_1 = 1,\ \sigma_1^2 = 1),\ (\mu_2 = 10,\ \sigma_2^2 = 3),\ (\mu_3 = 21,\ \sigma_3^2 = 1),$$

on voit que

$$d_M(1,2) = \frac{164/3}{\sqrt{83/9} + \sqrt{245}} < 4,$$

$$d_M(2,3) = \frac{244/3}{\sqrt{365} + \sqrt{123/9}} < 4$$

et

$$d_M(1,3) = \frac{400}{20+20} = 10$$

Donc

$$d_M(1,2) + d_M(2,3) < d_M(1,3).$$

Nous n'avons pas pu trouver le comportement de $d_M(\mu,\nu)$ sous les transformations mesurables, mais il est facile de voir qu'elle reste invariable si la transformation T est un résumé exhaustif. La relation entre $d_M(\mu,\nu)$ et l'erreur asymptotique d'un type particulier a été examinée par Mourier.

IV. DISTANCE "EN VARIATION" DE KOLMOGOROFF

Etant donné un ensemble $\mathfrak{M}$ de mesures sur $(X,\mathcal{T})$, on sait que, pour deux mesures μ et ν dans $\mathfrak{M}$, la fonction

$$d_V(\mu,\nu) = \sup_{E\in\mathcal{T}} \left|\mu(E) - \nu(E)\right|$$

est une distance sur $\mathfrak{M}$. Si les mesures dans $\mathfrak{M}$ sont des mesures de probabilité avec $\mu(X) = 1$ pour toute $\mu \in \mathfrak{M}$, alors on peut écrire

$$d_V(\mu,\nu) = \sup_{E\in\mathcal{T}} (\mu(E) - \nu(E))$$

$$= \sup_{E\in\mathcal{T}} (\nu(E) - \mu(E))$$

$$= (\mu-\nu)^+ = (\mu-\nu)^- = \frac{1}{2}\left|\mu-\nu\right|$$

où ξ^+, ξ^- et $|\xi|$ sont les variations supérieure, inférieure et absolue de la mesure ξ. C'est pourquoi Kolmogoroff appelle $d_V(\mu,\nu)$ la distance en variation.

Si λ est une mesure de probabilité sur $(X,\mathcal{T})$ telle que $\mu,\nu \ll \lambda$ (on peut prendre λ comme, par exemple, $\frac{1}{2}(\mu+\nu)$), alors

$$d_V(\mu,\nu) = \mu\left\{x : f_\mu(x) > f_\nu(x)\right\} - \nu\left\{x : f_\mu(x) > f_\nu(x)\right\}$$

$$= \nu\left\{x : f_\mu(x) \leq f_\nu(x)\right\} - \mu\left\{x : f_\mu(x) \leq f_\nu(x)\right\}$$

$$= \frac{1}{2}\int \left|f_\mu(x) - f_\nu(x)\right| \, d\lambda(x),$$

où $f_\mu(x) = d\mu/d\lambda$, $f_\nu(x) = d\nu/d\lambda$.

Nous allons énumérer les propriétés suivantes de $d_V(\mu,\nu)$:

1) $0 \leq d_V(\mu,\nu) \leq 1$.

2) $d_V(\mu,\nu) = 0$ si et seulement si $\mu = \nu$.

3) $d_V(\mu,\nu) = 1$ si et seulement si $\mu \perp \nu$

4) Soit T une transformation mesurable de $(X,\mathcal{T})$ sur $(Y,\mathcal{C})$; alors

$$T^{-1}(\mathcal{C}) \subset \mathcal{T}$$

Donc

$$d_V(\mu T^{-1}, \nu T^{-1}) = \sup_{F \in \mathcal{T}} \left| \mu T^{-1}(F) - \nu T^{-1}(F) \right|$$

$$= \sup_{E \in T^{-1}(\mathcal{T})} \left| \mu(E) - \nu(E) \right|$$

$$\leqslant \sup_{E \in \mathcal{S}} \left| \mu(E) - \nu(E) \right|$$

$$= d_V(\mu, \nu)$$

C'est-à-dire que $d_V(\mu, \nu)$ n'augmente pas sous une transformation mesurable.

Soit, maintenant, T un résumé exhaustif. Alors si

$$g_\mu(y) = d\mu\, T^{-1} / d\lambda\, T^{-1}$$

et

$$g_\nu(y) = d\nu\, T^{-1} / d\lambda\, T^{-1}$$

on sait, d'après Halmos et Savage, que

$$f_\mu(x) = g_\mu T(x) \left[\lambda\right] \quad \text{et} \quad f_\nu(x) = g_\nu T(x) \left[\lambda\right]$$

Donc

$$d_V(\mu T^{-1}, \nu T^{-1}) = (\mu T^{-1} - \nu T^{-1}) \left\{ y : g_\mu(y) - g_\nu(y) > o \right\}$$

$$= (\mu - \nu) T^{-1} \left\{ y : g_\mu(y) - g_\nu(y) > o \right\}$$

$$= (\mu - \nu) \left\{ x : g_\mu T(x) - g_\nu T(x) > o \right\}$$

$$= (\mu - \nu) \left\{ x : f_\mu(x) - f_\nu(x) > o \right\}$$

$$= d_V(\mu, \nu).$$

C'est-à-dire que $d_V(\mu, \nu)$ reste invariable si la transformation T est un résumé exhaustif

5) Soit μ_1, ν_1 deux mesures sur $(X_1, \mathcal{T})$, chacune $\ll \lambda_1$, et μ_2, ν_2 deux mesures sur $(X_2, \mathcal{T}_2)$, chacune $\ll \lambda_2$.

Soit $d\mu_i = f_i(x_i)\, d\lambda_i$ et $d\nu_i = g_i(x_i)\, d\lambda_i$, $i = 1, 2$.

Alors

$$d_V(\mu_1 \otimes \mu_2, \nu_1 \otimes \nu_2) = \frac{1}{2} \int_{X_1 \otimes X_2} \left| f_1(x_1) f_2(x_2) - g_1(x_1) g_2(x_2) \right| d(\lambda_1 \otimes \lambda_2)(x_1, x_2)$$

$$= \frac{1}{2} \int_{X_1 \otimes X_2} \left| \left\{ f_1(x_1) - g_1(x_1) \right\} f_2(x_2) + \left\{ f_2(x_2) - g_2(x_2) \right\} g_1(x_1) \right| d(\lambda_1 \otimes \lambda_2)(x_1, x_2)$$

$$\leqslant \frac{1}{2} \int_{X_1} \left| f_1(x_1) - g_1(x_1) \right| d\lambda_1(x_1) + \frac{1}{2} \int_{X_2} \left| f_2(x_2) - g_2(x_2) \right| d\lambda_2(x_2)$$

$$= d_V(\mu_1, \nu_1) + d_V(\mu_2, \nu_2).$$

Si l'on prend le produit cartésien de $(X, \mathcal{T})$ avec lui-même, on obtient

$$d_v(\mu^{(2)}, \nu^{(2)}) \leq 2\, d_v(\mu,\nu)$$

De même

$$d_v(\mu^{(n)}, \nu^{(n)}) \leq n\, d_v(\mu,\nu).$$

Soit, maintenant, $\mathcal{R}$ la classe de tous les ensembles produits de la forme

$$E \otimes X$$

où $E \in \mathcal{J}$. Alors

$$\mathcal{R} \subset \mathcal{J} \otimes \mathcal{J}$$

Donc

$$\begin{aligned} d_v(\mu^{(2)}, \nu^{(2)}) = d_v(\mu\otimes\mu, \nu\otimes\nu) &= \sup_{E\in\mathcal{J}\otimes\mathcal{J}} \left|(\mu\otimes\mu)(E) - (\nu\otimes\nu)(E)\right| \\ &\geq \sup_{E\in\mathcal{R}} \left|(\mu\otimes\mu)(E) - (\nu\otimes\nu)(E)\right| \\ &= \sup_{E\in\mathcal{J}} \left|\mu(E) - \nu(E)\right| \\ &= d_v(\mu,\nu) \end{aligned}$$

Ainsi l'on voit que $d_v(\mu^{(n)}, \nu^{(n)})$ est une fonction positive, non-décroissante de n, qui est de plus toujours ≤ 1. Donc

$$\lim_{n\to\infty} d_v(\mu^{(n)}, \nu^{(n)}) \leq 1$$

6) Comme l'avait indiqué M. Kolmogoroff dans ses conférences, la distance $d_v(\mu,\nu)$ est associée à l'erreur totale minimum de discrimination. Soit, en effet, $\varepsilon(\mu,\nu)$ cette erreur. Alors

$$\begin{aligned} \varepsilon(\mu,\nu) &= \inf_{E\in\mathcal{J}} \left(\mu(\overline{E}) + \nu(E)\right) \\ &= \inf_{E\in\mathcal{J}} \left\{1 - (\mu(E) - \nu(E))\right\} \\ &= 1 - \sup_{E\in\mathcal{J}} \left(\mu(E) - \nu(E)\right) \\ &= 1 - d_v(\mu,\nu) \end{aligned}$$

7) Il subsiste des inégalités entre $d_v(\mu,\nu)$, $d_R(\mu,\nu)$ et "l'affinité" $\rho(\mu,\nu)$.

(i) Si α et ε sont l'erreur commune minimum de Welch et l'erreur totale minimum respectivement, alors $2\alpha \geq \varepsilon$. Donc

$$d_R(\mu,\nu) = 1 - \alpha \leq 1 - \varepsilon/2 = \frac{1}{2}\left(1 + d_v(\mu,\nu)\right).$$

(ii)

$$\begin{aligned} \rho(\mu,\nu) &= \int_X \sqrt{f_\mu(x)\, f_\nu(x)}\, d\lambda(x) = \int_{f_\mu > f_\nu} \sqrt{\frac{f_\mu(x)}{f_\nu(x)}}\, d\nu(x) + \int_{f_\mu \leq f_\nu} \sqrt{\frac{f_\nu(x)}{f_\mu(x)}}\, d\mu(x) \\ &\geq \nu\left\{x : f_\mu(x) > f_\nu(x)\right\} + \mu\left\{x : f_\mu(x) \leq f_\nu(x)\right\} \\ &= 1 + \mu\left\{x : f_\mu(x) \leq f_\nu(x)\right\} - \nu\left\{x : f_\mu(x) \leq f_\nu(x)\right\} \\ &= 1 - d_v(\mu,\nu) \end{aligned}$$

BIBLIOGRAPHIE

[1] BHATTACHARYA,A. : Bull. Cal. Math. Soc., 35 (1943), 99.
[2] CHERNOFF,H. : Ann. Math. Statist., 23 (1952), 493.
[3] DARMOIS,G. : C.R. Acad. Sci., 200 (1935), 1265.
[4] FISHER,R.A. : Phil. Trans. Roy. Soc. A, 222 (1921), 309.
[5] " : Ann. Eugenics, 7 (1936), 179.
[6] " : Ann. Eugenics, 8 (1938), 376.
[7] " : Ann. Eugenics, 9 (1939), 238.
[8] HALMOS, P. et SAVAGE, L: Ann. Math Statist., 20 (1949), 225.
[9] KULLBACK, S. et LEIBLER,R. : Ann. Math. Statist., 22 (1951), 79.
[10] LEVY,P. : Théorie de l'addition des variables aléatoires, Paris (1937), 47.
[11] MAHALANOBIS,P.C. : Proc. Nat. Inst. Sci. (India) 12 (1936), 49.
[12] MATUSITA,K. : Ann. Inst. Math. Statist. Tokyo, 2 (1950).
[13] MATUSITA,K. : ibid., 3 (1951), 17.
[14] MATUSITA,K. et AKAIKA,H. : ibid. 4 (1952), 11.
[15] MATUSITA,K., SUZUKI,Y. et HUDIMOTO,H. : ibid., 6 (1954), 133.
[16] MOURIER,E. : C.R. Acad. Sci., 223 (1946), 712.
[17] NEYMAN,J. et PEARSON,E.S. : Biometrika, 20A (1928), 175 et 263.
[18] " : Phil. Trans. Roy. Soc. A, 231 (1933a), 281.
[19] RAO,C.R. : Jour. Roy. Statis. Soc. B, 10 (1948), 159.
[20] TILDESLEY,M.L. : Biometrika, 13 (1921), 176.
[21] WELCH,B.L. : Biometrika, 31 (1939), 218.