分布函数間の一つの距離とその応用

松 下 嘉 米 男

(1953 年8月 受付)

A Distance between Distribution Functions and its Applications

KAMEO MATUSITA

The treatments of some classical problems, for example the problem of goodness of fit, are explained from the point of view of decision-making, using the distance between distribution functions given in [1].

Institute of Statistic1 Mathematics

§1. 序

統計数理に於いて，種々の問題を取扱ふに際して分布函数間に適切な距離を導入することが大変重要である．実際，適切な距離を導入したがために，問題が解決された例は幾多見受けられる．分布函数の間の距離でよく知られているものに次のようなものがある．即ち $F_1(x)$, $F_2(x)$ を夫々一次元の分布函数とするとき

$$D(F_1, F_2) = \sup_x |F_1(x) - F_2(x)|$$

をもつて $F_1(x)$ と $F_2(x)$ の距離とするものである．之は容易に考え付かれるものであり，又よく引合いに出されるものである．所謂 Lēvy の距離も之を modify したものである．この $D(F_1, F_2)$ を用いて，例えば適合度の問題或いは推定の問題が論ぜられる．即ち問題の確率変数 X に対する n 個の観測値 $x_1, \cdots, x_n$ を得たときに，之より経験的分布函数 $S_n(x)$ を作り，之と X の従う分布函数 $F(x)$ との距離 $D(F, S_n)$ を考え，之によつて問題を論じる．この $D(F, S_n)$ については，その漸近的分布が Kolmogorov により与えられているので，之を用いる根拠があるわけである．然るに，この $D(F, S_n)$ を用いても，推定の問題はさておき，適合度の検定の問題に於いては所謂第二種の過誤と云つたものの大きさがわからない．尚この適合度の問題に対しては χ^2 検定と云ふものがあるが，之についてもやはり第二種の過誤がわからないと云う欠陥がある．この χ^2 と云う量も理論的分布函数と経験的分布函数の間の隔りを表わす一つの量と考えられる．上記 $D(F, S_n)$ 或いは χ^2 のような量を基にしたのでは過誤全体を一まとめにして考えた "risk" を小さくすると云う立場に立つ決定函数を求めるわけには行かない．この様な決定函数を求めるためには如何にしても risk を評価出来る方法をとらなければならない．それには $D(F, S_n)$ 或いは χ^2 に代る適切な量を導入しなくてはならない．この他種々の所謂 non-parametric の問題に於いても分布函数の間に適切な距離を導入する事が望ましい．

以下に於いては，分布函数の間の一つの距離を定義し，それによつて各種の問題の取扱い方を述べようと思う．

§2. 分布函数間の距離と近似度

統計の諸問題に於いては問題となる分布函数は同一問題に於いては同時に discrete か或いは同時に連続である，或いはこれに reduce される．それで本節に於いても同時に discrete, 或いは連続な分布の間の距離を定義しそれと共に近似度 (affinity) なる量を導入する．

F_1, F_2 を今同時に discrete 或いは連続な分布とする．簡単のため F_1, F_2 は一次元の分布とする。そうすると F_1 及び F_2 による確率は適当な測度 m を用いて次の様に表わされる．

$$F(E)=\int_E p_1(x)dm,\quad F_2(E)=\int_E p_2(x)dm$$

$$p_1(x),\ p_2(x)\geqq 0$$

この時 F_1 と F_2 の距離として

$$||F_1-F_2||=\left(\int_R(\sqrt{p_1(x)}-\sqrt{p_2(x)})^2dm\right)^{\frac{1}{2}}$$

を考える．又 F_1, F_2 の近似度として

$$\rho(F_1,F_2)=\int_R(\sqrt{p_1(x)}\cdot\sqrt{p_2(x)}dm$$

なる量を考える．ここにRは分布の定義されている全空間である．この距離と近似度に関して次のようなことが成り立つ．

$$||F_1-F_2||^2=2(1-\rho)$$
$$0\leqq||F_1-F_2||^2\leqq 2$$
$$0\leqq\rho(F_1,F_2)\leqq 1$$

この ρ を用いると，問題になつている分布函数がわかつていて，そのいづれを採るべきかと云つた問題に対し，risk を予め与えた正数よりも小さくするような決定函数を与える事が出来る．このことに関しては，すでに他所に於いて述べた故，本稿に於いては，距離 $||\ ||$ を用いてそれ以外の典型的な問題の取扱い方について述べようと思う．

§3. 基本的関係とその応用

適合度の問題，二標本問題，或いは推定の問題を取扱うに際して，基本になる関係を次に述べる．連続分布は discrete な分布によつて如何程でも近似される故，以下考える分布は凡べて有限 discrete とする．

さて，確率変数 X の従う分布は $F:(E_1, p_1:\cdots:E_k p_k)$ とし（ここに $E_1,\cdots,E_k$ は事象，$p_1,\cdots,p_k$ はそれらの確率を表わす），n 回の観測中に $E_1,\cdots,E_k$ を夫々 $n_1,\cdots,n_k$ 回観測したとする．この時経験的分布 S_n を作り，これと F との距離

$$||F-S_n||=\left(\sum_{i=1}^{k}\left(\sqrt{p_i}-\sqrt{\frac{n_i}{n}}\right)^2\right)^{\frac{1}{2}}$$

を考えると，これに関して次の定理を得る．

定理．任意の正数 t に対し

$$P_r\left(||F-S_n||^2<\frac{k-1}{n}t\right)\geqq 1-\frac{1}{t}$$

が成立する．

この関係を用いると，例えば適合度の問題は次のやうに取扱われる．先づ問題は，確率変数Xが分布 $F_0:(E_1, p_{01}:\cdots:E_k, p_{0k})$に従うか，或いは $||F-F_0||\geqq\delta_0$ なる一つの分布 $F:(E_1, p_1:\cdots:E_k, p_k)$ に従うかを定めることにする．この時，任意の正数 ε に対し $t>\frac{1}{\varepsilon}$ のように t をとり，ついで $n\geqq 4\frac{k-1}{\delta_0^2}t$ のようにnをとる．そこで

$$||F_0-S_n||^2<\frac{k-1}{n}t$$

なるとき，Xは分布 F_0 に従い

$$||F_0 - S_n||^2 \geqq \frac{k-1}{n} t$$

なるとき，X は $||F-F_0||>\delta_0$ なる一つの分布 F に從つているとすると，之によつて定められる決定函数の risk は ε よりも小さくなる。ここで考える weight function は正しい決定をしたときに0で絶対値は常に1よりも小さいとする。

二標本問題もこれと同様に取扱われる。又推定の問題についても上の定理より直ちに未知の分布函数或いはパラメーターに対する confidence band 若しくは confidence interval を得ることが出来る。又，F に含まれるパラメーターを点推定する場合，$||F-S_n||^2$ を最小にするようにそのパラメーターを定める方法が考えられる。この推定方法を用いて，適合度問題に於いて問題の分布 F_0 が未知のパラメーターを含む場合を取扱うことが出来る。

（統計数理研究所）

参　考　文　献

〔1〕 Matusita, Kameo, "On the Theory of Statistical Decision Functions," *Annals of the Institute of Statistical Mathematics*, vol. III, 1951.

〔2〕 Matusita, Kameo and Hirotugu Akaike, "Note on the Decision Problem," *Annals of the Institute of Statistical Mathematics*, vol. IV, 1952.

〔3〕 松下嘉米男著，"統計数理の基礎理論" 増訂版 朝倉書店，1953.