㊹ 観測値の組分けについて

青　山　博次郎

§ 1.　緒　言　R.V. Mises[1] は先に観測値の組分けについて次のような問題を考えた。即ち一つの集団があつて，この各々に試行を施すとその結果一定の数 x（実数）を示し，又集団に属する各々のものはすべて n 個の class の中の一つに属しているものと仮定する。このn個の class は一定の x に対して n 個の確率密度 $f_1(x)$, $f_2(x)$, ……, $f_n(x)$ によつて特色づけられているとする。今観測の結果 x を得た時，これはどの class に属しているかを決定したいというのがその問題である。

この応用として林知己夫氏の仮釈放についての興味ある論文がある。[2] こゝでは Mises の立場と若干異なる立場から同様な結果が得られることを示そう。

§ 2.　class が二つの場合

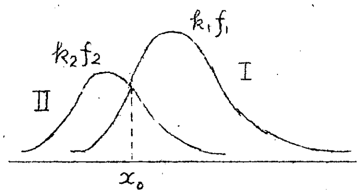
我々は一定の無限母集団を考え，この中には二つの密度函数 $f_1(x)$, $f_2(x)$ にて特色づけられる classes が混合しているものとする。

即ち

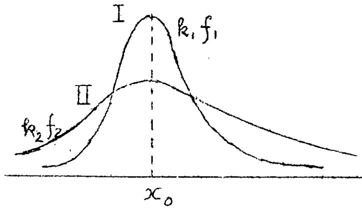
$$\int_{-\infty}^{\infty} f_1(x)\,dx = 1, \qquad \int_{-\infty}^{\infty} f_2(x)\,dx = 1 \tag{1}$$

$$\int_{-\infty}^{\infty} \{k_1 f_1(x) + k_2 f_2(x)\}\,dx = 1 \tag{2}$$

こゝに k_1, k_2 は $k_1+k_2=1$ を満足する定数であって我々には予め分っているものとは限らない。このとき一つの観測値 x を得たとして，一定数 x_0 をとり，$x>x_0$ ならば f_1 の class に，$x<x_0$ ならば f_2 の class に属するものと判断することにすると，信頼度はいかなる x_0 をえらぶときに最大となるかが問題になる。$f_1(x)$, $f_2(x)$ は共に unimodal とすると，図の如き場合が考えられよう。



第一図



第二図

信頼度 P を正しい判断をした率で表わすことにすれば

$$P=\int_{x_0}^{\infty}k_1f_1(x)dx+\int_{-\infty}^{x_0}k_2f_2(x)dx \tag{3}$$

とおくことができる。

先づ f_1, f_2 共に正規分布でその平均はそれぞれ m_1, m_2 標準偏差はそれぞれ σ_1, σ_2 とし，k_1, k_2 が既知の場合を考えてみよう。信頼度 P を最大とするためには $\dfrac{\partial P}{\partial x_0}=0$ より x_0 は

$$\frac{k_1}{\sigma_1}e^{-\frac{(x_0-m_1)^2}{2\sigma_1^2}}=\frac{k_2}{\sigma_2}e^{-\frac{(x_0-m_2)^2}{2\sigma_2^2}}$$

を満足することが分る。即ち

$$(\sigma_2^2-\sigma_1^2)x_0^2-2x_0(m_1\sigma_2^2-m_2\sigma_1^2)+(m_1^2\sigma_2^2-m_2^2\sigma_1^2-2\sigma_1^2\sigma_2^2\log\frac{k_1\sigma_2}{k_2\sigma_1})=0 \tag{4}$$

もし $m_1>m_2$ ならば，この二根のうち $x_0>m_2$ なるものをえらぶ。特に $\sigma_1=\sigma_2=\sigma$ のときは

$$x_0 = \frac{1}{2}\left\{ (m_1+m_2) - \frac{2\sigma^2}{m_1-m_2} \log \frac{k_1}{k_2} \right\} \tag{5}$$

これらの何れの場合に於ても，信頼度は（３）から計算できる。

以上は k_1，k_2 が既知の場合であるが，もし未知の場合ならば (3) を最大にする如き x，k_1 を定めることにする。 即ち

$$P = \frac{k_2}{\sqrt{2\pi}\,\sigma_2}\int_{-\infty}^{x} e^{-\frac{(x-m_2)^2}{2\sigma_2^2}}dx + \frac{k_1}{\sqrt{2\pi}\,\sigma_1}\int_{x}^{\infty} e^{-\frac{(x-m_1)^2}{2\sigma_1^2}}dx \tag{6}$$

但し　$k_1 + k_2 = 1$

$\dfrac{\partial P}{\partial k_1} = 0$ より

$$\frac{1}{\sigma_1}\int_{x_0}^{\infty} e^{-\frac{(x-m_1)^2}{2\sigma_1^2}}dx = \frac{1}{\sigma_2}\int_{-\infty}^{x_0} e^{-\frac{(x-m_2)^2}{2\sigma_2^2}}dx \tag{7}$$

このとき $\dfrac{\partial P}{\partial x} = 0$ より

$$\frac{f_1(x_0)}{f_2(x_0)} = \frac{k_2}{k_1} \tag{8}$$

が得られる。

一般の密度函数については（７）の代りに

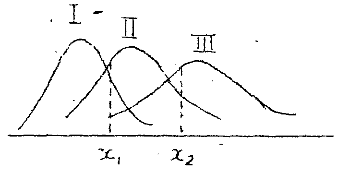
$$\int_{x_0}^{\infty} f_1(x)\,dx = \int_{-\infty}^{x_0} f_2(x)\,dx \tag{9}$$

が成立する。

§ 3. class が n 個の場合

n 個の classes の 密度函数をそれぞれ $f_1(x)$, $f_2(x)$, $\cdots$, $f_n(x)$ とし，これらがすべて k_1，k_2，$\cdots\cdots$，k_n の割合で母集団中に含まれているものとする。

前と同様にして信頼度を



第 三 図

$$P = k_1 \int_{-\infty}^{x_1} f_1(x)dx + k_2 \int_{x_1}^{x_2} f_2(x)dx + \cdots\cdots + k_n \int_{x_{n-1}}^{\infty} f_n(x)dx \qquad (10)$$

但し $$k_1 + k_2 + \cdots\cdots + k_n = 1 \qquad (11)$$

で表わすとき，この P を最大にする如く k_1，k_2，…… k_{n-1}，x_1，…… x_{n-1} を求めれば，これらの分点によつて

$$x < x_1 \ , \ x_1 < x < x_2 \ , \ x_2 < x < x_3, \cdots\cdots, x_{n-1} < x$$

に属する x をそれぞれ第1，第2，……，第n組に属するものと考えればよい。

$$\frac{\partial P}{\partial k_i} = 0, \quad \frac{\partial P}{\partial x_i} = 0 \qquad (i = 1, 2, \cdots\cdots, n-1)$$ を計算して

$$\int_{-\infty}^{x_1} f_1(x)dx = \int_{x_1}^{x_2} f_2(x)dx = \cdots\cdots = \int_{x_{n-1}}^{\infty} f_n(x)dx \qquad (12)$$

を満足する如く x_1，x_2，……，x_{n-1} を定めればよいことが分る。このとき

$$\frac{k_1}{f_2(x_1)} = \frac{k_2}{f_1(x_1)}, \quad \frac{k_2}{f_3(x_2)} = \frac{k_3}{f_2(x_2)}, \cdots\cdots, \frac{k_{n-1}}{f_n(x_{n-1})} = \frac{k_n}{f_{n-1}(x_{n-1})} \qquad (13)$$

が成立する。 即ち分点 x_1，x_2，……，x_{n-1} に於て

$$\frac{f_\nu(x)}{f_{\nu+1}(x)} = \frac{k_{\nu+1}}{k_\nu} \qquad (\nu = 1, 2, \cdots\cdots, n-1) \qquad (14)$$

が成立する。

もし k_1, k_2, ------, k_n が既知のときは（14）を満足する如く x_1, x_2, ------, x_{n-1} をえらべば，その信頼度は（10）によって計算される。

§4　多変数の場合

n 個の classes が密度函数 $f_1(x_1, x_2, \cdots, x_m)$, $f_2(x_1, x_2, \cdots, x_m)$, ----, $f_n(x_1, x_2, \cdots, x_m)$ をもって特長づけられているときも同様にして，信頼度は

$$P = k_1 \int_{R_1} f_1 \, dR_1 + k_2 \int_{R_2} f_2 \, dR_2 + \cdots + k_n \int_{R_n} f_n \, dR_n \tag{15}$$

但し　　$k_1 + k_2 + \cdots + k_n = 1$

で與えられる。こゝに R_ν は m 次元のある領域を示し，$R_1 + R_2 + \cdots + R_n$ は全 m 次元空間と一致するものとする。

こゝで $\dfrac{\partial P}{\partial k_\nu} = 0$ より

$$\int_{R_1} f_1 \, dR_1 = \int_{R_2} f_2 \, dR_2 = \cdots = \int_{R_n} f_n \, dR_n \tag{16}$$

なることは直ちに分る。また R_ν を種々に変へるとき（15）の最大値を求めるには相隣る領域 R_ν, $R_{\nu+1}$ にそれぞれ微小な増分及び減分 ΔR を與えるとき $\Delta P = 0$ なるべきことより

$$\frac{f_\nu(x_1, x_2, \cdots, x_m)}{f_{\nu+1}(x_1, x_2, \cdots, x_m)} = \frac{k_{\nu+1}}{k_\nu} \tag{17}$$

が境界上で成立することが示される。

（註）1) R.V. Mises: On the classification of observation data into distinct groups, Annals of Math. Statis. Vol. XVI No. 1, 1945

2) 林 知己夫：、Parole Predictionに於ける統計的方法の一応用について，再犯調査の基礎．ケースワーク研究会