分類と或る二標本検定及びその関係に就いて

藤　本　熙

(1958 年 2 月受付)

The Problem of Classification and Two Sample Test

Hirosi HUDIMOTO

In this paper, the results of [22] on the distribution-free classification is rewritten and the applications is shown. We consider a certain two-sample test connected with these applications.

The Institute of Statistical Mathematics.

§1.* [22] に於て我々は次の様な問題を取扱つた

distribution function (dist. f.) F_1, F_2 をその sub-population π_1, π_2 の dist. f. としてもつ composite population π を構成するために, π の任意の random member x が π_1 のものである確率 p を仮定する. 即ち π は distribution function $F=pF_1+qF_2$ によつて特徴づけられる. 但し $q=1-p$. 今 π の member であることの判つている観測値が, 予め決めておいた値 x に等しいか, あるいは小さければ π_1 の, 然らざれは π_2 の member であると判定するとき, その組分けの正しい確率

(1) $$C(x) = pF_1(x) + q[1 - F_2(x)]$$

の estimate として, π より size N の random sample を得, その m 個が π_1, 残り $N-m$ 個が π_2 のものであつた時, p を $\frac{m}{N}$ で q を $\frac{N-m}{N}$ で, F_1, F_2 をその empirical distribution function $\hat{F}_1$, $\hat{F}_2$ でおきかえて,

(2) $$\hat{C}_N(x) = \frac{m}{N}\hat{F}_1(x) + \frac{N-m}{N}[1 - \hat{F}_2(x)]$$

を用いると, F_1, F_2 の連続性を仮定して次の様な確率不等式が成立つ. 但し観測値 $u_1<\cdots\cdots<u_m$, $v_1<\cdots\cdots<v_{N-m}$ に対して,

$$\hat{F}_1(x) = \frac{k}{m} \qquad u_k \leqq x < u_{k+1},$$

$$\hat{F}_2(x) = \frac{h}{N-m} \qquad v_h \leqq x < v_{h+1}$$

と定義する.

(3) $$\text{Prob.}\left\{ \hat{C}_N(x) - \left(\eta + p\sqrt{\frac{1}{2N(p-\eta)}\log\frac{1}{\alpha_1}} + q\sqrt{\frac{1}{2N(q-\eta)}\log\frac{1}{\alpha_2}}\right) \leqq C(x) \right\} \geqq (1-\alpha)(1-\alpha_1-\alpha_2)$$

が任意の $0.001 \leqq \alpha_1 \leqq 0.1$, $0.001 \leqq \alpha_2 \leqq 0.1$ 及び次の様に定めた α, η に対して成立つ. 但し

* 以下の計算値の凡ては相原和子によつて得られた.

α, η は binomial dist. により Prob.$\left\{\left|\frac{m}{N}-p\right|<\eta\right\}=1-\alpha$ なる様に定める. statistic (2) は David S. Stoller によつても議論されていることを, 氏からの書翰によつて知つた. しかしながら [9] では $\max \hat{C}_N(x)$ の consistency のみが取扱はれた.

以上の記述に於いて, π_1, π_2 の discriminant point としては, (2) に於ける $\hat{C}_N(x)$ を maximum にする点を採用しようとするのであるが, 若し $\max_x C(x)$ が一意的に決まるものであれば, (3) を採用すれば $\max \hat{C}_N(x)$ が $\max C(x)$ を過大評価することを防いでいる. 処で $\hat{C}_N(x)$ に就いては, fixed x に対しては, $E[\hat{C}_N(x)]=C(x)$ であるが, $\max \hat{C}_N(x)$ の mean value は nonnegatively biased $E[\max_x \hat{C}_N(x)] \geqq \max_x C(x)$ であることは注意を要する. 表 I には Prob.$\{\hat{C}_N(x)-\varepsilon \leqq C(x)\} \geqq 0.95$ となる様な N, ε の関係を種々なる p に就いて示した.

表 I Prob.$\{\hat{C}_N(x)-\varepsilon \leqq C(x)\} \geqq 0.95$ なる ε, N の関係

N \ p ; ε	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{5}$
400	0.208	0.202	0.194	0.187
900	0.136	0.132	0.127	0.122
1600	0.101	0.098	0.094	0.090
2500	0.081	0.078	0.075	0.072

さて此処で取扱つている分類の問題は, R. A. Fisher[13),14)] 以来比較的古くからの問題だと思うが, 例えば B. L. Welch[15)], 近頃でも A Wald[1)], R. v. Mises[2)], 林[3),4),5)], 松下[6)], C. F. Kossach[7)], P. G. Hoel and R. P. Peterson[8)], D. S. Stoller[9)], B. B. Day and M. M. Sandomire[10)], P. O. Johnson[11)], G. W. Brown[12)], 等, 斯う見て来るとその数は少くない.

処で

(4) $$pf_1 \geqq qf_2 \quad \text{in } W_0$$
$$pf_1 \leqq qf_2 \quad \text{in } R-W_0$$

但し f_1, f_2 は π_1, π_2 の density function, R は sample space, の条件のもとでは, 即ち (4) による組分けが正しい判別の確率を maximum にする[8)] 値 C は

(5) $$C=\frac{1}{2}\left[1+\int_R |pf_1-qf_2|dx\right]$$

であり, かつまた

(6) $$pf_1 \geqq \sqrt{pq}\sqrt{f_1 f_2} \geqq qf_2 \quad \text{in } W_0,$$
$$pf_1 < \sqrt{pq}\sqrt{f_1 f_2} < qf_2 \quad \text{in } R-W_0,$$

となるから, affinity[6),16)]

(7) $$\rho(F_1, F_2)=\int_R \sqrt{f_1 f_2}dx,$$
$$0 \leqq \rho(F_1, F_2) \leqq 1,$$

を用いると, 容易に次の関係が成立つ.

(8) $$\frac{1}{2} \leqq 1-\frac{1}{2}\rho(F_1, F_2) \leqq 1-\sqrt{pq}\,\rho(F_1, F_2)$$
$$\leqq C \leqq \frac{1}{2}\left[1+\left\{1-4pq\rho^2(F_1, F_2)\right\}^{\frac{1}{2}}\right].$$

かつ $\rho(F_1, F_2)$ は

(9)　$$2(1-\rho) = ||F_1 - F_2||^2 = \int_R (\sqrt{f_1} - \sqrt{f_2})^2 dx$$

に直して [16] の定理によつて評価出来る.

試みに Gaussian distribution $N(m_1, \sigma^2)$, $N(m_2, \sigma^2)$ をとると

(10)　$$1 - \frac{1}{2} e^{-\frac{|m_1-m_2|^2}{8\sigma^2}} < C$$

これを $\frac{1}{\sqrt{2\pi}} \int_{\frac{m_1-m_2}{2\sigma}}^{\infty} e^{-\frac{1}{2}t^2} dt$, 但し $m_1 \leqq m_2$, [3], [4], と比較して見ると次表の様になる.

表 II

$\frac{m_1-m_2}{\sigma}$	$1-\frac{1}{2}\rho$	$\frac{1}{\sqrt{2\pi}}\int_{\frac{m_1-m_2}{2\sigma}}^{\infty} e^{-\frac{t^2}{2}} dt$	$\frac{1}{2}\left[1+(1-\rho^2)^{\frac{1}{2}}\right]$
1	0.55876	0.69146	0.73517
2	0.69674	0.84135	0.89753
3	0.83768	0.93319	0.97292
4	0.93233	0.97725	0.99540
5	0.97803	0.99379	0.99952
6	0.99445	0.99865	0.99997

同様にして3つの group の場合には

(11)　$$p_1 f_1 \geqq p_2 f_2,\ p_1 f_1 \geqq p_3 f_3 \quad \text{in } W_1,$$
$$p_2 f_2 \geqq p_1 f_1,\ p_2 f_2 \geqq p_3 f_3 \quad \text{in } W_2,$$
$$p_3 f_3 \geqq p_1 f_1,\ p_3 f_3 \geqq p_2 f_2 \quad \text{in } W_3.$$

但し $W_1+W_2+W_3=R$——が充たされるならば, 判別の誤る確率は

(12)　$$\alpha \leqq \sqrt{p_1 p_2}\rho(F_1, F_2) + \sqrt{p_2 p_3}\rho(F_2, F_3) + \sqrt{p_3 p_1}\rho(F_3, F_1) - \min_{i,j} \sqrt{p_i p_j}\rho(F_1, F_2) \quad \text{for } i \neq j,\ i, j = 1, 2, 3.$$

更に §2 以下の記述の便を考えて, Gaussian distribution に就いて簡単な場合を書下して見ると,

i)　$$N(m_1, \sigma^2);\ f_1(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m_1)^2}{2\sigma^2}}$$
$$N(m_2, \sigma^2);\ f_2(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m_2)^2}{2\sigma^2}}$$

とすると,　$pf_1 \geqq qf_2,\quad W_0 = \{x;\ x \leqq x_0\}$,

から, よく知られている様に,

(13)　$$x_0 = \frac{1}{2}\left[(m_1 + m_2) + \frac{2\sigma^2}{m_2 - m_1} \log \frac{p}{q}\right].$$

また R. v. Mises の場合が, $x_0 = \frac{m_1+m_2}{2}$ の場合であることもよく知られている, [22].

ii)　$N(m_1, \sigma_1^2)$, $N(m_2, \sigma_2^2)$ の場合は $\frac{f_2}{f_1} = \frac{\sigma_1}{\sigma_2} e^{\frac{1}{2}\left[\left(\frac{x-m_1}{\sigma_1}\right)^2 - \left(\frac{x-m_2}{\sigma_2}\right)^2\right]} = \frac{p}{q}$ より, $m_2 > m_1$, $\sigma_2 > \sigma_1$ を仮定すれば,

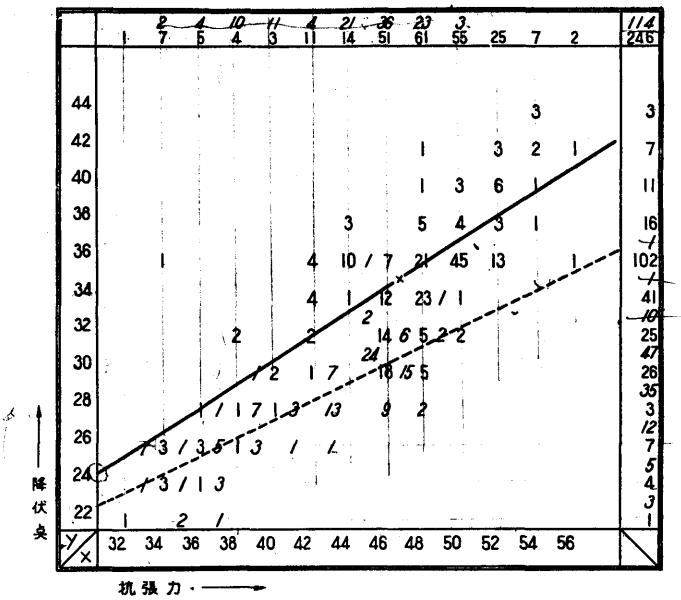
(14)　$$x_0 = \frac{\sigma_2^2 m_1 - \sigma_1^2 m_2}{\sigma_2^2 - \sigma_1^2} \pm \frac{\sigma_1\sigma_2(m_2 - m_1)}{\sigma_2^2 - \sigma_1^2} \sqrt{1 + \frac{2(\sigma_2^2 - \sigma_1^2)}{(m_2 - m_1)^2} \log \frac{p\sigma_2}{q\sigma_1}}$$

もしここで, $x_0 = \frac{\sigma_2^2 m_1 - \sigma_1^2 m_2}{\sigma_2^2 - \sigma_1^2} + \frac{\sigma_1\sigma_2(m_2 - m_1)}{\sigma_2^2 - \sigma_1^2}$ をとれば, R. v. Mises の場合となる.

§2. 図 I は某製鉄会社の製品に就いて

9 mmφ鉄筋の抗張力と降伏点を調べてその scatter diagram を示した．data は電信電話公社材料試験室の堤武作氏によつて得られたものである．図中 roman type の数字は通常工程の，italic type の数字は巻取工程による製品の個数を示している．巻取工程というのは，搬出の便を考えて

図 I



工程の終りで drum に巻取つたものを常温でもどして，切断したものである．試みにあとで必要となる数値のいくつかを示すと次の様である．

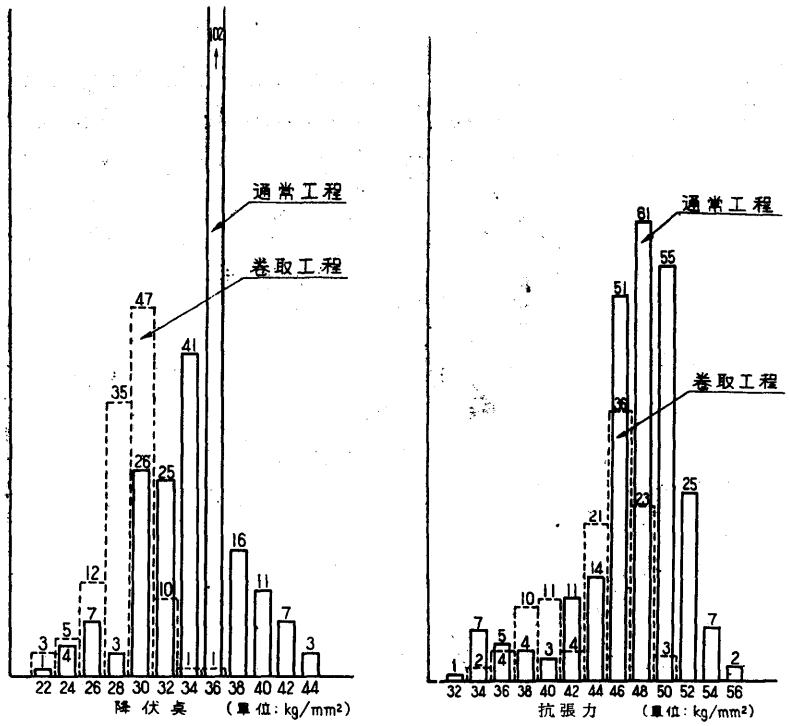
	〈巻取工程〉	〈通常工程〉
標本の数	m: 114,	n: 246,
抗張力（単位: kg/mm²）		
平均	$\bar{x}_1 = 44.158$,	$\bar{x}_2 = 47.211$,
分散	$s^2_{x_1} = 14.256$,	$s^2_{x_2} = 18.744$,
標準偏差	$s_{x_1} = 3.776$,	$s_x = 4.329$.
降伏点（単位: kg/mm²）		
平均	$\bar{y}_1 = 28.754$,	$\bar{y}_2 = 34.569$,
分散	$s^2_{y_1} = 5.431$,	$s^2_{y_2} = 13.595$,
標準偏差	$s_y = 2.330$,	$s_{y_2} = 3.687$.

図 II には抗張力 X, 降伏点 Y それぞれの histogram を示した．$\max_x \hat{C}_N(x)$, $\max_y \hat{C}_N(y)$ をそれぞれ求めて見ると，

$$\max_x \hat{C}_N(x) = 0.703, \qquad \max_y \hat{C}_N(y) = 0.850. \tag{15}$$

ここで，discriminant function を求める通常の手順に従つて，抗張力 X と降伏点 Y の linear function $Z = \lambda_1 X + \lambda_2 Y$ に於て，その平均差の2乗 $(\bar{Z}_1 - \bar{Z}_2)^2$ の分散に対する比を maximum にする様な λ_1, λ_2 を求めて

図 II



$$Z = -0.2739\,X + 0.7261\,Y$$

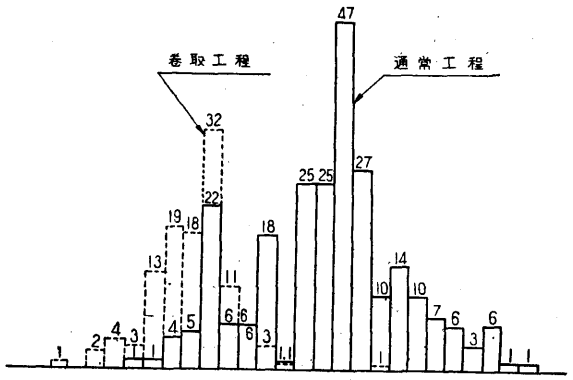
となつた．それより $\max \hat{C}_N(z)$ を求めると

(16)　$\max \hat{C}_N(z) = 0.86$

これから見ると，この判別に対する抗張力 X の側からの寄与の甚だ弱いことが予想される．従つてこの結果から降伏点 Y の側では，巻取のために降伏点が稍〻下降の傾向を示し製品のモロサを増加させると判断されよう．

処で試みに，sample の値をそのまま近似として用いて $\frac{1}{\sqrt{2\pi}}\int_{\frac{m_1-m_2}{\sigma_1+\sigma_2}}^{\infty} e^{-\frac{t^2}{2}}dt$ と比較して見ると[3),4),5)]，

図 III



(17)

$$\frac{1}{\sqrt{2\pi}}\int_{\frac{\bar{x}_1-\bar{x}_2}{s_{x_1}+s_{x_2}}}^{\infty} e^{-\frac{t^2}{2}}dt = 0.65,$$

$$\frac{1}{\sqrt{2\pi}}\int_{\frac{\bar{y}_1-\bar{y}_2}{s_{y_1}+s_{y_2}}}^{\infty} e^{-\frac{t^2}{2}}dt = 0.83$$

で何れも $\max \hat{C}_N(x)$, $\max \hat{C}_N(y)$ で求めた値が大きく出る．しかしながら，判別の確率が $\frac{1}{2}$ に近い処での差は当然大きくなる筈であるから，あながち大きく出過るとも言えない．後述する事柄から大まかな処推定誤差の範囲内と考えられるから，大して変る処はない．大体の目当を Z に就いて見ると次の様である．

(18) $$\frac{1}{\sqrt{2\pi}}\int_{\frac{\bar{x}_1-\bar{x}_2}{\sigma_{\bar{x}_1}+\sigma_{\bar{x}_2}}}^{\infty} e^{-\frac{t^2}{2}}dt = 0.84$$

次に check の意味をも含めて，以下の様に試みる．今 independent random variable X_1, X_2 の distribution function を $F_1(x)$, $F_2(x)$ として $X_1<X_2$ となる確率を求めて見ると，

(19) $$P = \text{Prob.}\{X_1 < X_2\} = \int_{-\infty}^{\infty} F_1(x)\, dF_2(x).$$

ここで $F_1(x)$ を empirical な distribution function $\hat{F}_1(x)$ でおきかえると，$F_2(x)$ が知られる場合の P の estimate として

(20) $$\hat{P} = \int_{-\infty}^{\infty} \hat{F}_1(x)\, dF_2(x)$$

を得る．$E(\hat{P})=P$. $F_2(x)$ が前述の数値，$\bar{x}_2=47.211$, $s_x=4.162$ (但し s_x は within の variance をとつた) の Gauss-分布と見做して調べて見ると，

(21) $$\hat{P}_x = \text{est. Prob.}\{X_1<X_2\} = 0.707$$

但し，est. Prob.{ } は Prob{ } の estimate を示す．同様に $\bar{y}_2=34.569$, $s_y=3.318$ を用いて，

(22) $$\hat{P}_y = \text{est.Prob.}\{Y_1 < Y_2\} = 0.887.$$

$\hat{P}$ に対する評価は，もし $F_2(x)$ の分布が正確に判つているならば，

(23) $$\text{Prob.}\{P < \hat{P} + \varepsilon\} > 1 - e^{-2m\varepsilon^2}, \quad \text{但し } m;\ X_1 \text{ の sample size}$$

上式はいわゆる N. Smirnov の近似評価であるが，sample size >50 であればその喰違いは 0.05 以下で，かつ近似に於ける ε は正確な値のそれより，同じ確率に対して常に大きく出ることが確められている，Z. Birmbaum and H. Tingey[17]，から実際的には正確なものである．(23) は Z. Birmbaum[18] に与えられている．

処で，もし F_1, F_2 共にその分布が正確に判らない場合

(24) $$\hat{\hat{P}} = \int_{-\infty}^{\infty} \hat{F}_1(x)\, d\hat{F}_2(x)$$

は，Mann-Whitney の U-statistic[20], U-statistic が Hoeffding の statistic とまぎらわしければ U-test の alternative form である，mn で割つたものに一致する．今 $\frac{1}{2}[F_1(x)+F_2(x)]$ の様な compound distribution function, を媒介して考えて見ると，

$$\alpha = \int_{-\infty}^{\infty} F_1(x)\frac{d(F_1(x)+F_2(x))}{2} + \int_{-\infty}^{\infty}\frac{[F_1(x)+F_2(x)]}{2}dF_2(x)$$

$$= 1 + \int_{-\infty}^{\infty}[F_1 - F_2]\, d\frac{F_1+F_2}{2}$$

処が $\int_{-\infty}^{\infty} F_1 dF_1 = \int_{-\infty}^{\infty} F_2 dF_2 = \frac{1}{2}$ であることから $\alpha = \frac{1}{2} + \int_{-\infty}^{\infty} F_1 dF_2 = \frac{1}{2} + P.$

よつて

(25) $$P - \frac{1}{2} = \int_{-\infty}^{\infty}[F_1(x) - F_2(x)]d\frac{F_1(x)+F_2(x)}{2}.$$

F_1, F_2 を $\hat{F}_1$, $\hat{F}_2$ でおきかえて estimate をつくると

$$\int_{-\infty}^{\infty}[\hat{F}_1(x) - \hat{F}_2(x)]d\frac{\hat{F}_1(x)+\hat{F}_2(x)}{2}$$

$$= \frac{1}{2}\left[\frac{1}{m}\sum_{i=1}^{m}\left(\frac{i}{m} - \frac{r_i - i}{n}\right) + \frac{1}{n}\sum_{j=1}^{m}\left(\frac{R_j - j}{m} - \frac{i}{n}\right)\right]$$

但し r_i. Rj は m 個の X_1, n 個の X_2 の combined sample に於ける i 番目の ordered X_1

のrank 及び j 番目の ordered X_2 の rank を示す.
従つて

$$(26)\qquad \int_{-\infty}^{\infty}[\hat{F}_1(x)-\hat{F}_2(x)]d\frac{\hat{F}_1(x)+\hat{F}_2(x)}{2}=\frac{1}{mn}\sum_{j=1}^{n}(R_j-j)-\frac{1}{2}+\frac{n-m}{4mn}$$
$$=\hat{\hat{P}}-\frac{1}{2}+\frac{n-m}{4mn}.$$

もし $F_1\equiv F_2$ のもとでの $\hat{\hat{P}}$ の fluctuation, 即ち $\hat{\hat{P}}=\frac{1}{2}$ の仮定を reject する目的のためには, $F_1\equiv F_2$, かつ $m/n\underset{n\to\infty}{\to}r$ ($\neq 0$, constant) を仮定して,

$$(27)\qquad \hat{\hat{P}}-\frac{1}{2}+\frac{n-m}{4mn}\leqq\sup(\hat{F}_1-\hat{F}_2)$$

より,

$$(28)\qquad \text{Prob.}\left\{\hat{\hat{P}}\leqq\frac{1}{2}+\varepsilon+\frac{|n-m|}{4mn}\right\}\geqq 1-e^{-2\frac{mn}{N}\varepsilon^2}$$

の程度に見積ればよかろう.

$$(29)\qquad \hat{\hat{P}}_x=\text{est. Prob.}\{X_1<X_2\}=0.671,$$
$$\hat{\hat{P}}_y=\text{est. Prob.}\{Y_1<Y_2\}=0.872.$$

(28) を用いて $\text{Prob.}\left\{\hat{\hat{P}}-\frac{1}{2}\leqq\varepsilon_\alpha\right\}\geqq 1-\alpha$ なる ε_α を求めると, $\varepsilon_\alpha=\sqrt{\frac{N}{2mn}\log\frac{1}{\alpha}}$ より, $\alpha=0.05$ で $\varepsilon_{0.05}\fallingdotseq\sqrt{\frac{N}{mn}}\cdot 1.22\fallingdotseq 0.14$. 従つて $\hat{\hat{P}}_x$ はかろうじて significance level 0.05 で reject 出来る程度で, 判別に対する弱さが判る. また (14) より得た結果から est. $P_r\{Z_1<Z_2\}$ を求めると,

$$(30)\qquad \hat{\hat{P}}_z=0.912$$

少し荒い評価ではあるが, $\sigma^2(\hat{\hat{P}})\leqq\frac{P(1-P)}{\min(m,n)}$ で Gauss 近似が効くと見て,

$$(31)\qquad \text{est.}\,\sigma(\hat{\hat{P}}_x)\leqq 0.044,$$
$$\text{est.}\,\sigma(\hat{\hat{P}}_y)\leqq 0.037,$$

で当つて見ると, $\hat{\hat{P}}_y$, $\hat{\hat{P}}_z$ に差があるといえる程度のものでもない.

(25) に於て $\frac{1}{2}(F_1(x)+F_2(x))$ を compound Gaussian distribution と見做して, §2のはじめに掲げた sample からの値を代用すると, Y に就いては

$$(32)\qquad \hat{P}_z=0.890$$

何れも大差のない値を示すが, Gauss で当嵌めたのでは, 少し尾の部分が低く出過るのかも知れない. 即ち一方乃至両方を Gauss と仮定した場合, 然も両方を仮定した場合の方が一方を仮定した場合より, わずかではあるが一層高くなるという理由でである. 勿論判然と左様言えると言うのではないが——.

§3. 今迄の記述に示した我々の目的は

group の数が2つの場合の classification と two-sample test の或るものと関連づけて見ようとしたのであるが, その目的のためには未だ舌たらずであるのをまぬがれぬかも知れない. 処でその様な理由に就いては, 幾つも掲げることが出来ようが, (2) を与えた様な考えに従つて discriminant point を決め判別の確率を求める場合, はじめに与えた distribution function $F=pF_1+qF_2$ が妥当する様なものに制限されねば正確には当嵌らないということである. 従つてその様な意味で適応性の広さから, min-max の方法に従う R. v. Mises の考えに優位を認めざる

を得ない．しかしながら，$\int_{-\infty}^{x_0} dF_1(x)=\int_{x_0}^{\infty} dF_2(x)$ によつて discriminant point x_0 を決める場合，F_1, F_2 は empirical な distribution に就いての話ではないから，$\hat{F}_1$, $\hat{F}_2$ にひき直した場合での sample size との関連が望まれる．これは次の様な素朴な手順を用いる場合の評価は簡単である．

今もし $F_1(x_0)+F_2(x_0)=1$ なる $x=x_0$ を以て discriminant point とすれば，そのときの正しい判別の確率は $\max\{F_1, F_2\}$ であるから，これらを empirical distribution function $\hat{F}_1$, $\hat{F}_2$ でおきかえて，$\hat{F}_1(x^*)+\hat{F}_2(x^*)=1$ を採用することにし，x_0 を x^* で estimate することにすれば，そのときの正しい判別の確率は $\max\{\hat{F}_1(x^*), \hat{F}_2(x^*)\}-\varepsilon$ の型で estimate 出来よう．従つてこの場合の ε の大きさを評価すればよい．処が，

$$\begin{aligned}(33)\qquad &\text{Prob.}\{(\hat{F}_1(x)-F_1(x))+(\hat{F}_2(x)-F_2(x))<\varepsilon\}\\ &\geqq \text{Prob.}\{\sup[\hat{F}_1(x)-F_1(x)]+\sup[\hat{F}_2(x)-F_2(x)]<\varepsilon\}\\ &\geqq (1-e^{-2m\varepsilon_1^2})(1-e^{-2(N-m)(\varepsilon-\varepsilon_1)^2}).\end{aligned}$$

従つてこれから ε の大いさを N に関係づけてはかることが出来る．

また (1) に於ける $C(x)$ の大いさは $|F_1-F_2|$ の大いさに関係するが，$P-\frac{1}{2}=\int_{-\infty}^{\infty}[F_1(x)-F_2(x)]d(pF_1(x)+qF_2(x))$ となるから当然 P の大いさに依存することになるであろう．処で F_1, F_2 が Gaussian distribution $N(m_1, \sigma_1^2)$, $N(m_2, \sigma_2^2)$ の場合に，念のためたしかめて見ると，実は，

$$P=\frac{1}{2\pi\sigma_1\sigma_2}\iint_{x_1-x_2<0} e^{-\left[\frac{(x_1-m_1)^2}{2\sigma_1^2}+\frac{(x_2-m_2)^2}{2\sigma_2^2}\right]}dx_1dx_2=\frac{1}{2\pi\sigma_1\sigma_2}\iint_{t_1-t_2<m_2-m_1} e^{-\left[\frac{t_1^2}{2\sigma_1^2}+\frac{t_2^2}{2\sigma_2^2}\right]}dt_1dt_2.$$

であるから，$t=t_1-t_2$, $s=\frac{\sigma_2}{\sigma_1}t_1+\frac{\sigma_1}{\sigma_2}t_2$ によつて，

$$\begin{aligned}(34)\qquad P&=\frac{1}{2\pi(\sigma_1^2+\sigma_2^2)}\int_{-\infty}^{m_2-m_1} e^{-\frac{t^2}{2(\sigma_1^2+\sigma_2^2)}}dt\int_{-\infty}^{\infty} e^{-\frac{s^2}{2(\sigma_1^2+\sigma_2^2)}}ds\\ &=\frac{1}{\sqrt{2\pi}\sqrt{\sigma_1^2+\sigma_2^2}}\int_{-\infty}^{m_2-m_1} e^{-\frac{t^2}{2(\sigma_1^2+\sigma_2^2)}}dt.\end{aligned}$$

これは Gauss の場合で，$\int_{-\infty}^{x_0} dF_1(x)=\int_{x_0}^{\infty} dF_2(x)$ で決めた discriminant point $x_0=\frac{m_1\sigma_2+m_2\sigma_1}{\sigma_1+\sigma_2}$ によつて求めた正しい判別の確率 $\frac{1}{\sqrt{2\pi}}\int_{\frac{m_1-m_2}{\sigma_1+\sigma_2}}^{\infty} e^{-\frac{t^2}{2}}dt$ にほとんど一致することが判る．従つて $Z=\sum\lambda_i X_i$ の型で分類を行おうとするとき，判別の力が強い factor として X_i をひろい出す場合の操作としては，χ^2 を用いるよりは $\hat{\hat{P}}=\frac{1}{2}$ か否かを基準にして $\hat{\hat{P}}$ を見てゆく方が直接的である様に見える $P=\frac{1}{2}$ の仮説のものでは (28) が成立つことは既に述べたが，この場合，勿論分散は H. B. Mann and D. R. Whitney[20] の結果から

$$(35)\qquad \sigma^2(\hat{\hat{P}})=\frac{N+1}{12mn},\quad N=m+n$$

で与えられる．少しばかり事情は違うが fit に対する χ^2 と $\sup|F_1-\hat{F}_1|$ の比較は F. J. Massey[21] に見える．

また必要であれば，(34) の分散は次の様にして求まる．

$$\begin{aligned}(36)\qquad \sigma^2(P)=\frac{1}{mn}\{&P+(m-1)\text{Prob.}\{X_{11}<X_2,\ X_{12}<X_2\}\\ &+(n-1)\text{Prob.}\{X_1<X_{21},\ X_1<X_{22}\}-(N-1)P^2\}\end{aligned}$$

但し X_{11}, X_{12} はそれぞれ dist. f. F_1 をもつ independent random variable, X_{21}, X_{22} はそれぞれ dist. F_2 をもつ independent random variable とする.

——であるが, Prob.$\{X_{11}<X_2\}$, Prob.$\{X_{12}<X_2\}$, Prob.$\{X_1<X_{21}\}$, Prob.$\{X_1<X_{22}\}$ は (34) で与えられるから, $X_{11}-X_2$, $X_{12}-X_2$ をそれぞれ, T_1, T_2 で, X_1-X_{21}, X_1-X_{22} をそれぞれ S_1, S_2 で示すと, T_1, T_2 及 S_1, S_2 の covariance は

(37)
$$C\{T_1, T_2\} = \sigma_2^2$$
$$C\{S_1, S_2\} = \sigma_1^2$$

従つてその相関係数は

(38)
$$\rho(T_1, T_2) = \frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2},$$
$$\rho(S_1, S_2) = \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}.$$

よつて,

(39)
$$\text{Prob.}\{X_{11} < X_2,\ X_{12} < X_2\}$$
$$= \frac{1}{2\pi(\sigma_1^2+\sigma_2^2)\sqrt{1-\left(\frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}\right)^2}} \int_{-\infty}^{m_2-m_1}\int_{-\infty}^{m_2-m_1} \exp.\left[-\frac{1}{2\left(1-\left(\frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}\right)^2\right)}\right.$$
$$\left.\times\left\{\frac{t_1^2}{\sigma_1^2+\sigma_2^2} - 2\frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}\cdot\frac{t_1t_2}{\sigma_1^2+\sigma_2^2} + \frac{t_2^2}{\sigma_1^2+\sigma_2^2}\right\}\right] dt_1\,dt_2$$
$$= \frac{1}{2\pi\sigma_1^2\left(1+2\frac{\sigma_2^2}{\sigma_1^2}\right)^{\frac{1}{2}}} \int_{-\infty}^{m_2-m_1}\int_{-\infty}^{m_2-m_1} \exp.\left[-\frac{\sigma_1^2+\sigma_2^2}{2\sigma_1^2(\sigma_1^2+2\sigma_2^2)}\right.$$
$$\left.\times\left\{t_1^2 - 2\frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}t_1t_2 + t_2^2\right\}\right] dt_1\,dt_2.$$

同様にして,

(40)
$$\text{Prob.}\{X_1 < X_{21},\ X_1 < X_{22}\}$$
$$= \frac{1}{2\pi\sigma_2^2\left(1+2\frac{\sigma_1^2}{\sigma_2^2}\right)^{\frac{1}{2}}} \int_{-\infty}^{m_2-m_1}\int_{-\infty}^{m_2-m_1} \exp.\left[-\frac{\sigma_1^2+\sigma_2^2}{2\sigma_2^2(\sigma_2^2+2\sigma_1^2)}\right.$$
$$\left.\times\left\{t_1^2 - 2\frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}t_1t_2 + t_2^2\right\}\right] dt_1\,dt_2.$$

これらを (36) に代入して得られる.

(統計数理研究所).

参 考 文 献

1) WALD, Abraham, On a statistical problem arising in the classification of an individual into one of two groups, *Ann. Math. Stat.*, Vol. XV (1944), pp. 145–162.
2) MISES, R. v., On the classification of observation data into distinct groups, *Ann. Math. Stat.*, Vol. XVI (1945), pp.
3) HAYASHI, C., On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematico-statistical point of view, *Ann. Inst. Stat. Math.*, Vol. 3 (1951), pp. 69–98.
4) 林 知己夫, 数量化理論とその応用例 (II), 統計数理研究所彙報, 第4巻第2号, 19–30.
5) 林 知己夫, 数量化理論とその応用例 (III), 統計数理研究所彙報, 第5巻第1号, 27–31.
6) MATUSITA, K., Decision rule, based on the distance, for the classification problem, *Ann. Inst. Stat. Math.*, Vol. VIII, (1956), pp. 67–77.

7) KOSSACK, Carl F., Some techniques for simple classification, *Proceedings Berkeley Symposium on Mathematical statistics and Probability*, (1949), pp. 345–352.

8) HOEL, P. G. and PETERSON, R. P., A solution to the problem of optimum classification, *Ann. Math. Stat.*, Vol. XX (1949), pp. 433–438.

9) STOLLER, David S., Univariate two-population disribution-free discrimination, *Jour. Amer. Stat. Ass.*, Vol. 49 (1954), pp. 770–777.

10) DAY, Besse B. and SANDOMIRE, Marion M., Use of the discriminant function for more than two groups, *Jour. Amer.* Stat. *Ass.*, Vol. 37 (1942), pp. 461–472.

11) JOHNSON, Palmer O., The quantification of qualitative data in discriminant analysis, *Jour. Amer. Stat. Ass.*, Vol. 45 (1950), pp. 65–76.

12) BROWN, George W., Discriminant function. *Ann. Math. Stat.*, Vol. XVIII (1947) pp. 514–528.

13) FISHER, R. A., Statistical methods for research workers.

14) FISHER, R. A., The use of multiple measurements in taxonomic problems, *Ann. Eugenics* 7 (1936), pp. 179–188.

15) WELCH, B. L., Note on discriminant functions, *Biometrika*, 31 (1939), pp. 218–220.

16) MATUSITA, K., Decision rules, based on the distance, for problems of fit, two samples, and estimation, *Ann. Math. Stat.*, Vol. XXVI, (1955), pp. 631–640.

17) BIRNBAUM, Z. and TINGEY, H., One-sided confidence contours for probability distribution functions, *Ann. Math. Stat.*, Vol. XXII, (1951), pp. 592–596.

18) BIRNBAUM, Z. W., On a use of the Mann-Whitney statistic, *Proceedings of the Third Berkeley Symposium*, Vol. I, (1956), pp. 13–17.

19) MARSHALL, A. W., A large-sample test of the hypothesis that one of two random variables is stochastically larger than the other, *Jour. Amer. Stat. Ass.*, Vol. 46 (1951), pp. 366–374.

20) MANN, H. B., and WHITNEY, D. R., On a test of whether one of two random variables is stochastically larger than the other, *Ann. Math. Stat.*, Vol. XVIII (1947), pp. 50–60.

21) MASSEY F. J., The Kolmogorov-Smirnov test for goodness of fit, *Jour. Amer. Stat. Ass.*, Vol. 46 (1951), pp. 68–78.

22) HUDIMOTO, H., A note on the probability of the correct classification when the distributions are not specified, *Ann. Inst. Stat. Math.*, Vol. IX (1957), pp. 31–36.

22) 青山博次郎, 松下嘉米男, 林 知己夫, 水野 坦, 社会現象の統計数理, 朝倉書店.