A NOTE ON THE PROBABILITY OF THE CORRECT CLASSIFICATION WHEN THE DISTRIBUTIONS ARE NOT SPECIFIED

By Hirosi Hudimoto

(Received June 5, 1957)

1. Consider the composite population π of π_1 and π_2 in which a random member is assigned to the sub-population π_1 or π_2 with probability p or q $(=1-p)$, and let $F_1(x)$ and $F_2(x)$ be the distribution functions of π_1 and π_2, respectively. Now let x be the discriminant point between π_1 and π_2, that is, decide that an individual known to be a member of π belongs to π_1 or π_2 according as it is less than or equal to x, or not. When a random sample O_N of size N is drawn from π and m members of it belong to π_1 and the remainders to π_2, we employ $\hat{C}_N^{(1)}(x)=\frac{m}{N}c_m^{(1)}(x)+\frac{N-m}{N}[1-c_{N-m}^{(2)}(x)]$ as an estimate of the probability of the correct classification by x, $C(x)=pF_1(x)+q[1-F_2(x)]$, where $c_m^{(1)}(x)$ and $c_{N-m}^{(2)}(x)$ denote their empirical cumulative distribution functions. Concerning this we have given the following relation in the previous paper [3].

For a given positive number η and any x, we have

$$(1)\quad P_r\{|\hat{C}_N^{(1)}(x)-C(x)|>3\eta\}\leqq\frac{1}{5N^2\eta^4}\left[1+\left\{\frac{p^4}{(p-\eta)^2}+\frac{q^4}{(q-\eta)^2}\right\}\left\{1-\frac{1}{5N^2\eta^4}\right\}\right].$$

The following table will show the values of N that satisfies the relation $P_r\{|\hat{C}_N^{(1)}(x)-C(x)|>0.1\}\leqq 0.05$ for different values of p.

Table I

Table for the values of N satisfying the relation

$$P_r\{|\hat{C}_N^{(1)}(x)-C(x)|>0.1\}\leqq 0.05$$

P	1/2	1/3	1/4	1/5	1/6	1/7	1/8	1/9	1/10	1/20
N	2245	2298	2347	2385	2413	2435	2453	2467	2479	2541

Now, we intend to give another relation which will serve better because it is more precise than the previous inequality:

If $F_1(x)$ and $F_2(x)$ are assumed to be continuous, for given positive numbers α and η satisfying the relation $P_r\left\{\left|\frac{m}{N}-p\right|<\eta\right\}\geqq 1-\alpha$ and for any given positive numbers α_1 and α_2 with $.001\leqq\alpha_1\leqq.1$, $.001\leqq\alpha_2\leqq.1$, we have

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

PROOF. When we denote by A, A_i, B, B_1, B_2 and C the events $\left|\frac{m}{N}-p\right|\leqq\eta$, $m=i$, $p\left(\frac{k}{m}-F_1(x)\right)+q\left(F_2(x)-\frac{h}{N-m}\right)\leqq\eta_1+\eta_2$, $p\left(\frac{k}{m}-F_1(x)\right)\leqq\eta_1$, $q\left(F_2(x)-\frac{h}{N-m}\right)\leqq\eta_2$ and $\left|\frac{m}{N}-p\right|+p\left(\frac{k}{m}-F_1(x)\right)+q\left(F_2(x)-\frac{h}{N-m}\right)\leqq\eta+\eta_1+\eta_2$ where $\eta_1>0$, $\eta_2>0$, respectively, we have

$$(3)\qquad P_r\{C\}\geqq P_r\{C\cap A\}\geqq P_r\{B\cap A\}=P_r\{B\cap\sum_{N(p-\eta)\leqq i\leqq N(p+\eta)}A_i\}$$
$$=\sum_{N(p-\eta)\leqq i\leqq N(p+\eta)}P_r\{A_i\}P\{B|A_i\}$$

and

$$(4)\quad P_r\{B|A_i\}\geqq P_r\{B_1\cap B_2|A_i\}=P_r\{\overline{\overline{B}_1\cup\overline{B}_2}|A_i\}=1-P_r\{\overline{B}_1\cup\overline{B}_2|A_i\}\ ,$$

where $\overline{B}_1$ and $\overline{B}_2$ denote the complementary sets of B_1 and B_2, respectively.

On the other hand, if X is a random variable with the continuous distribution function $F(x)=P_r\{X\leqq x\}$, and if $\hat{c}_n(x)$ is the empirical distribution function of n observations from the population with $F(x)$, that is,

$$\hat{c}_n(x)=\begin{cases}0\ ; & x<x_1\\ \frac{k}{n}\ ; & x_k\leqq x<x_{k+1}\\ 1\ ; & x_n\leqq x\end{cases}$$

where $x_1<\cdots x_k<\cdots<x_n$, then, as is well known,

$$P_n(\epsilon)=P_r\{F(x)\leqq\min[\hat{c}_n(x)+\epsilon,\ 1]\}$$
$$=P_r\{F(x)\geqq\max[\hat{c}_n(x)-\epsilon,\ 0]\}\ \text{ for all } x,$$

is independent of the distribution $F(x)$, and the Smirnov's asymptotic expression for $P_n(\epsilon)$ is

$$P_n\left(\frac{z}{\sqrt{n}}\right)=1-e^{-2z^2}.$$

Besides, in [2], Z. W. Birnbaum has given the following explicit expression for $P_n(\epsilon)$:

$$P_n(\epsilon)=1-\epsilon\sum_{j=0}^{[n(1-\epsilon)]}\binom{n}{j}\left(1-\epsilon-\frac{j}{n}\right)^{n-j}\left(\epsilon+\frac{j}{n}\right)^{j-1}$$

for $0<\epsilon\leqq 1$, where $[n(1-\epsilon)]$=greatest integer of the numbers which are less than or equal to $n(1-\epsilon)$, and according to the tables in [2] the asymptotic values $\tilde{\epsilon}_{n,\alpha}=\frac{z}{\sqrt{n}}$ are greater than the exact values $\epsilon_{n,\alpha}$ for the probability level α with $.001\leqq\alpha\leqq.1$,

Therefore, let α_1 and α_2 are the probability levels such as

$$P_r\left\{F_1(x)\geqq\max.\left[\frac{k}{m}-\frac{\eta_1}{p},0\right]\middle|A_m\right\}=1-\alpha_1,$$

and

$$P_r\left\{F_2(x)\leqq\min\left[\frac{h}{m}+\frac{\eta_2}{q},1\right]\middle|A_m\right\}=1-\alpha_2$$

for each fixed m, where $.001\leqq\alpha_1\leqq.1$, $.001\leqq\alpha_2\leqq.1$. Then we have

$$\begin{aligned}&P_r\left\{p\left(\frac{k}{m}-F_1(x)\right)>p\sqrt{\frac{1}{2N(p-\eta)}\log\frac{1}{\alpha_1}}\middle|A_m\right\}\\ &\leqq P_r\left\{p\left(\frac{k}{m}-F_1(x)\right)>p\sqrt{\frac{1}{2m}\log\frac{1}{\alpha_1}}\middle|A_m\right\}\\ &\leqq P_r\left\{p\left(\frac{k}{m}-F_1(x)\right)>\eta_1\middle|A_m\right\}\leqq\alpha_1,\end{aligned}\tag{5}$$

and

$$\begin{aligned}&P_r\left\{q\left(F_2(x)-\frac{h}{N-m}\right)>q\sqrt{\frac{1}{2N(q-\eta)}\log\frac{1}{\alpha_2}}\middle|A_m\right\}\\ &\leqq P_r\left\{q\left(F_2(x)-\frac{h}{N-m}\right)>q\sqrt{\frac{1}{2(N-m)}\log\frac{1}{\alpha_2}}\middle|A_m\right\}\\ &\leqq P_r\left\{q\left(F_2(x)-\frac{h}{N-m}\right)>\eta_2\middle|A_m\right\}\leqq\alpha_2.\end{aligned}\tag{6}$$

Thus, if we apply $p\sqrt{\frac{1}{2N(p-\eta)}\log\frac{1}{\alpha_1}}$ and $q\sqrt{\frac{1}{2N(q-\eta)}\log\frac{1}{\alpha_2}}$ in-

stead of η_1 and η_2 into (3) and (4), respectively, we have the result to prove.

Concerning $\max_x C(x)$, as above *we have under the same conditions*

$$(7)\quad P_r\Big\{\max_x \hat{C}_N^{(1)}(x)-\Big(\eta+p\sqrt{\frac{1}{2N(p-\eta)}\log\frac{1}{\alpha_1}}+q\sqrt{\frac{1}{2N(q-\eta)}\log\frac{1}{\alpha_2}}\Big) \leqq \max_x C(x)\Big\}\geqq(1-\alpha)(1-\alpha_1-\alpha_2)\ .$$

Table II*

Table for the values of ϵ, such that $P_r\{\hat{C}_N^{(1)}(x)-\epsilon\leqq C(x)\}\geqq 0.95$,

N \ P	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{5}$
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

* Tables I and II were obtained by Kazuko Aihara of the Institute of Statistical Mathematics.

2. The result in Section 4 of the previous paper [3] is improved as follows.

If the condition of the optimal classification is satisfied, that is,

$$(8)\quad \begin{aligned} pf_1&\geqq qf_2 \quad \text{in } w_0\\ pf_1&<qf_2 \quad \text{in } R-w_0\end{aligned}$$

where R denotes the sample space, w_0 is the region $\{x; x\leqq x_0\}$, and f_1, f_2 are the density functions of π_1 and π_2, respectively, we have

$$(9)\quad \frac{1}{2}\leqq 1-\frac{1}{2}\rho(F_1,F_2)\leqq 1-\sqrt{pq}\rho(F_1,F_2) \leqq C(x_0)\leqq\frac{1}{2}[1+(1-4pq\rho^2(F_1,F_2))^{1/2}]$$

where $\rho(F_1,F_2)=\int_R\sqrt{f_1}\sqrt{f_2}dx$ is the affinity between F_1 and F_2 (see [1]), and $C(x_0)=\max_x C(x)$.

PROOF. From (8), we have

$$(10)\quad \begin{aligned} pf_1&\geqq\sqrt{pq}\sqrt{f_1f_2}\geqq qf_2 \quad \text{in } w_0\\ pf_1&<\sqrt{pq}\sqrt{f_1f_2}<qf_2 \quad \text{in } R-w_0\ .\end{aligned}$$

Therefore, we obtain the left hand side of the inequality (9) as follows:

$$
\begin{aligned}
(11)\qquad C(x_0) &= 1-p\int_{R-w_0} f_1 dx - q\int_{w_0} f_2 dx \\
&\geqq 1-\sqrt{pq}\,\rho(F_1, F_2) \\
&\geqq 1-\frac{1}{2}\rho(F_1, F_2) \quad \left(\text{because } \max \sqrt{pq}=\frac{1}{2}\right) \\
&\geqq \frac{1}{2} \qquad\qquad \text{(from } \rho(F_1, F_2)=1, \text{ when} \\
&\qquad\qquad F_1=F_2 \text{ and } 0\leqq\rho(F_1, F_2)\leqq 1)\,.
\end{aligned}
$$

Under the condition (8) we have

$$C(x_0)=\frac{1}{2}\left[1+\int_R |pf_1-qf_2|\,dx\right]$$

and

$$
\begin{aligned}
\int_R |pf_1-qf_2|\,dx &= \int_R |(\sqrt{pf_1}-\sqrt{qf_2})(\sqrt{pf_1}+\sqrt{qf_2})|\,dx \\
&\leqq \left[\int_R (\sqrt{pf_1}-\sqrt{qf_2})^2 dx \int_R (\sqrt{pf_1}+\sqrt{qf_2})^2 dx\right]^{1/2} \\
&\qquad\qquad \text{(by Schwarz's inequality)} \\
&= [1-4pq\,\rho^2(F_1, F_2)]^{1/2}\,.
\end{aligned}
$$

Therefore, we get

$$C(x_0)\leqq\frac{1}{2}[1+(1-4pq\,\rho^2(F_1, F_2))^{1/2}]\,.$$

Similarly we can treat the case of three groups having the sub-populations π_1, π_2 and π_3. That is, when

$$
\left.\begin{aligned} p_1f_1 &\geqq p_2f_2 \\ p_1f_1 &\geqq p_3f_3 \end{aligned}\right\} \text{ in } w_1
$$

$$
\left.\begin{aligned} p_2f_2 &\geqq p_1f_1 \\ p_2f_2 &\geqq p_3f_3 \end{aligned}\right\} \text{ in } w_2
$$

$$
\left.\begin{aligned} p_3f_3 &\geqq p_1f_1 \\ p_3f_3 &\geqq p_2f_2 \end{aligned}\right\} \text{ in } w_3
$$

where $w_1+w_2+w_3=R$, f_1, f_2 and f_3 denote the density functions of π_1, π_2 and π_3, then we have about the probability of the wrong classification, α, the following inequality

$$\alpha=p_1\int_{R-w_1} f_1 dx+p_2\int_{R-w_2} f_2 dx+p_3\int_{R-w_3} f_3 dx$$

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

where $\rho(F_i, F_j)$ denotes the affinity between π_i and π_j.

THE INSTITUTE OF STATISTICAL MATHEMATICS

REFERENCES

[1] Matusita, K., Decision rule, based on the distance, for the classification problem. *Ann. Inst. Stat. Math.*, Vol. VIII, No. 2, 1956.

[2] Birnbaum, Z. W. and Fred H. Tingey, One-sided confidence contours for probability distribution functions. *Ann. Math. Stat.*, Vol. 22, 1951.

[3] Hudimoto, H., On the distribution-free classification of an individual into one of two groups. *Ann. Inst. Stat. Math.*, Vol. VIII, No. 2, 1956.