ON THE DISTRIBUTION-FREE CLASSIFICATION OF AN INDIVIDUAL INTO ONE OF TWO GROUPS

By HIROSI HUDIMOTO

(Received Nov. 5, 1956)

1. Summary and introduction

In this note, we shall deal with two estimates $\hat{C}_N^{(1)}$ and $\hat{C}_N^{(2)}$ for the probability of the correct classification and with the evaluation of approach of the estimated discriminating point determined by $\hat{C}_N^{(1)}$ to the optimal classification point. In section 3, a certain numerical examples will be shown. In section 4, we shall state a relation between the correct probability of the optimal classification and the distance between the two distributions under consideration. $\hat{C}_N^{(2)}$ is evaluated using the concept of distance there. For this pourpose, Theorem V in [1] will give a usefull method.

2. Estimates of the probability of correct classification and evaluation of $\hat{C}_N^{(1)}$

We consider univariate populations π_1 and π_2 with probability density functions f_1 and f_2, respectively. Let us form the composite population π of π_1 and π_2 in which a random member is assigned to π_1 or π_2 with probabilities p or q. Then the distribution function of π can be denoted by pF_1+qF_2, where F_1 and F_2 denote the distribution functions of π_1 and π_2.

Now suppose a random sample of size N, O_N, is taken from π, and m members of it belong to π_1 and the remainders to π_2. When another sample value taken from π is classified into π_1 or π_2 according as its value is not greater than a preassigned value x or not, the probability of the correct classification is

$$C(x)=pF_1(x)+q[1-F_2(x)]\ . \tag{1}$$

Our problem is then to find an estimate of the value x_0 which maximizes $C(x)$ under the condition that $F_1(x)$, $F_2(x)$ are unknown. Denote the members of O_N belonging to π_1 by $u_1, \cdots, u_m$ and those belonging to π_2 by $v_1, \cdots, v_{N-m}$. We assume here that $u_1, \cdots, u_m$ and $v_1, \cdots, v_{N-m}$ are respectively ordered according to magnitude. Further, put

$$c_m^{(1)}(x)=\frac{k}{m} \qquad \text{for } u_k \leqq x < u_{k+1},$$

$$c_{N-m}^{(2)}(x)=\frac{h}{N-m} \qquad \text{for } v_h \leqq x < v_{h+1}$$

Then we have

$$(2) \qquad \hat{C}_N^{(1)}(x)=\frac{m}{N}c_m^{(1)}(x)+\frac{N-m}{N}[1-c_{N-m}^{(2)}(x)]=\frac{1}{N}[(N-m)+(k-h)]$$

as an estimate of $C(x)$ for x with $u_k \leqq x < u_{k+1}$ and $v_h \leqq x < v_{h+1}$.

On the other hand, when we divide the sample space (the real axis) into a finite number of intervals and denote by $\hat{f}_j^{(i)}$ $(i=1, 2)$ the frequency of u's or v's in the j-th interval, we have

$$(3) \qquad C_N^{(2)}=\frac{1}{2}+\frac{1}{2}\left[\sum_j \left|\frac{m}{N}\hat{f}_j^{(1)}-\frac{N-m}{N}\hat{f}_j^{(2)}\right|\right]$$

as an estimate of $C(x_0)$, provided that the condition of the optimal classification is satisfied, that is,

$$(4) \qquad \begin{aligned} pf_1 &\geqq qf_2 \qquad \text{in } w_0, \\ pf_1 &< qf_2 \qquad \text{in } R-w_0, \end{aligned}$$

where R denotes the sample space (see [3]). This is obtained as follows:

$$(5) \qquad C(x_0)=p\int_{w_0} f_1dx+q\int_{R-w_0} f_2dx=p\left(1-\int_{R-w_0} f_1dx\right)$$
$$+q\int_{R-w_0} f_2dx=p+\left[\int_{R-w_0}(qf_2-pf_1)dx\right]$$

or

$$(6) \qquad C(x_0)=1-p\int_{R-w_0} f_1dx-q\int_{w_0} f_2dx=(1-p)+p\int_{w_0} f_1dx-q\int_{w_0} f_2dx$$
$$=q+\left[\int_{w_0}(pf_1-qf_2)dx\right].$$

where w_0 is the region $\{x\,;\, x \leqq x_0\}$. Therefore,

$$(7) \qquad C(x_0)=\frac{1}{2}+\frac{1}{2}\left[\int_{w_0}(pf_1-qf_2)dx+\int_{R-w_0}(qf_2-pf_1)dx\right],$$

or

$$(8) \qquad C(x_0)=\frac{1}{2}+\frac{1}{2}\left[\int_R |pf_1-qf_2|dx\right].$$

Replacing the observed results for the probabilities and the density functions in (8) we obtain the formula (3). If we can take class intervals so that the histogram on them have $\hat{f}_j^{(i)}$ with $\frac{m}{N}\hat{f}_j^{(1)} \geqq \frac{N-m}{N}\hat{f}_j^{(2)}$ for $j \leqq k$ and $\frac{m}{N}\hat{f}_j^{(1)} < \frac{N-m}{N}\hat{f}_j^{(2)}$ for $j > k$, then $\hat{C}_N^{(2)}$ is nearly equal to $\max_x \hat{C}_N^{(1)}(x)$ and in this case classification point will be determined by the class mark of the k-th interval. Otherwise, the criterion is not given, but the evaluation of $C(x_0)$ can be carried out by $\hat{C}_N^{(2)}$ using the concept of distance.

As for formula (2) it will easily be seen that $E[\hat{C}_N^{(1)}(x)] = C(x)$ for every fixed x, and

$$
\begin{aligned}
(9)\quad |\hat{C}_N^{(1)}(x) - C(x)| &= \left| \frac{1}{N}[(N-m)+(k-h)] - [pF_1 + q(1-F_2)] \right| \\
&= \left| \left(\frac{k}{N} - pF_1\right) - \left(\frac{h}{N} - qF_2\right) - \left(\frac{m}{N} - p\right) \right| \\
&= \left| \frac{k}{m}\left(\frac{m}{N} - p\right) + p\left(\frac{k}{m} - F_1\right) - \frac{h}{N-m}\left(\frac{N-m}{N} - q\right) \right. \\
&\qquad \left. - q\left(\frac{h}{N-m} - F_2\right) - \left(\frac{m}{N} - p\right) \right| \\
&\leqq \left| \frac{m}{N} - p \right| + p\left| \frac{k}{m} - F_1 \right| + q\left| \frac{h}{N-m} - F_2 \right|
\end{aligned}
$$

Now, $\sum_{m=0}^{N} \binom{N}{m} p^m q^{N-m} \left(\frac{m}{N} - p\right)^4 = \frac{1}{N^2} pq\left\{3pq + \frac{1}{N}(1-6pq)\right\}$, therefore, $\sum_{m'} \binom{N}{m} p^m q^{N-m} \leqq \frac{1}{N^2\eta^4} pq\left\{3pq + \frac{1}{N}(1-6pq)\right\}$ where the summation $\sum_{m'}$ runs over m with $\left| \frac{m}{N} - p \right| > \eta$. Hence we get

$$
(10)\qquad P\left\{ \left| \frac{m}{N} - p \right| > \eta \right\} \leqq \frac{1}{N^2\eta^4} pq\left\{3pq + \frac{1}{N}(1-6pq)\right\} \leqq \frac{1}{5N^2\eta^4},
$$

similarly

$$
(11)\qquad P\left\{ p\left| \frac{k}{m} - F_1 \right| > \eta \right\} = P\left\{ \left| \frac{k}{m} - F_1 \right| > \frac{\eta}{p} \right\} \leqq \frac{p^4}{5m^2\eta^4}
$$

for every fixed m and x, and

$$
(12)\qquad P\left\{ q\left| \frac{h}{N-m} - F_2 \right| > \eta \right\} \leqq \frac{q^4}{5(N-m)^2\eta^4}
$$

for every fixed m and x.

On the other hand, if we denote by A, B, C, and A_i the events $\left|\frac{m}{N}-p\right|<\eta$, $p\left|\frac{k}{m}-F_1\right|+q\left|\frac{h}{N-m}-F_2\right|<2\eta$, $\left|\frac{m}{N}-p\right|+p\left|\frac{k}{m}-F_1\right|+q\left|\frac{h}{N-m}-F_2\right|<3\eta$ $m=i$, respectively, then we have

$$P\{C|A\}\geqq P\{B|A\}=\frac{P\{B\frown A\}}{P\{A\}}=\frac{P\{B\frown \sum_{N(p-\eta)<i<N(p+\eta)} A_i\}}{P\{A\}}$$
$$=\frac{\sum_{N(p-\eta)<i<N(p+\eta)} P\{B\frown A_i\}}{P\{A\}}=\frac{\sum_{N(p-\eta)<i<N(p+\eta)} P\{A_i\}P\{B|A_i\}}{P\{A\}}\ **$$

and by (11), (12),

$$P\{B|A_i\}\geqq 1-\frac{1}{5\eta^4}\left[\frac{p^4}{i^2}+\frac{q^4}{(N-i)^2}\right]$$
$$\geqq 1-\frac{1}{5\eta^4}\left[\frac{p^4}{N^2(p-\eta)^2}+\frac{q^4}{N^2(q-\eta)^2}\right]$$
$$=1-\frac{1}{5N^2\eta^4}\left[\frac{q^4}{(p-\eta)^2}+\frac{q^4}{(q-\eta)^2}\right].$$

where $P\{C|A\}$, etc$\cdots$denote conditional probabilities. Therefore, by (10)

$$\frac{P\{C\frown A\}}{1-\frac{1}{5N^2\eta^4}}\geqq\frac{P\{C\frown A\}}{P\{A\}}=P\{C|A\}$$
$$\geqq\frac{\sum_{N(p-\eta)<i<N(p-\eta)} P\{A_i\}}{P\{A\}}\left[1-\frac{1}{5N^2\eta^4}\left\{\frac{p^4}{(p-\eta)^2}+\frac{q^4}{(q-\eta)^2}\right\}\right]$$
$$=1-\frac{1}{5N^2\eta^4}\left[\frac{p^4}{(p-\eta)^2}+\frac{q^4}{(q-\eta)^2}\right]$$
$$P\{C\}\geqq P\{C\frown A\}\geqq\left(1-\frac{1}{5N^2\eta^4}\right)\left[1-\frac{1}{5N^2\eta^4}\left\{\frac{p^4}{(p-\eta)^2}+\frac{q^4}{(q-\eta)^2}\right\}\right].$$

Consequently,

$$(13)\qquad P\{|\hat{C}_N^{(1)}(x)-C(x)|>3\eta\}\leqq\frac{1}{5N^2\eta^4}\left[1+\left\{\frac{p^4}{(p-\eta)^2}+\frac{q^4}{(q-\eta)^2}\right\}\times\left\{1-\frac{1}{5N^2\eta^4}\right\}\right].$$

** I wish to express my thanks to Mr. K. Isii who kindly discussed with me these analytical derivations.

If we take $\eta=pq$, we get, therefore,

$$P\{|\hat{C}_N^{(1)}(x)-C(x)|>3pq\}\leqq\frac{3}{5N^2(pq)^4}-\frac{2}{25N^4(pq)^8}. \tag{13}$$

Now, restricting ourselves to the case where $pf_1\geqq qf_2$ for $x\leqq x_0$ and $pf_1<qf_2$ for $x>x_0$, that is, where only one maximum of $C(x)$ exists and $C(x)$ is monotone increasing in $x<x_0$ and monotone decreasing in $x>x_0$, we consider the sample point x^* which maximizes $\hat{C}_N^{(1)}(x)$ as an estimate of $C(x_0)$. In the following we shall give a relation between the estimate x^* and the value x_0.

Let c denote the maximum value of $C(x)$ in the outside of the open interval $|x-x_0|<\epsilon$ for fixed $\epsilon>0$. Obviously $c<C(x_0)$. Let δ be a small positive number such that $|x-x_0|<\delta\leqq\epsilon$ implies

$$C(x_0)-\frac{1}{3}(C(x_0)-c)<C(x)\leqq C(x_0) \tag{14}$$

Then, when $\frac{(C(x_0)-c)}{3}\geqq\eta$ and probabilistic inequality (13) are satisfied and when the interval $(x_0-\delta,\ x_0+\delta)$ includes certain members of O_N, in order that $|x^*-x_0|>\epsilon$, it is necessary that at least one of the values of $\hat{C}_N^{(1)}(x)$ at any sample points in the outside of the interval $(x_0-\epsilon,\ x_0+\epsilon)$ exceeds any of the values of $\hat{C}_N^{(1)}(x)$ within $(x_0-\delta,\ x_0+\delta)$. However, outside of $(x_0-\epsilon,\ x_0+\epsilon)$ we have

$$C(x)\leqq c<C(x_0)$$

and inside of $(x_0-\delta,\ x_0+\delta)$

$$C(x)>C(x_0)-\frac{1}{3}(C(x_0)-c).$$

If at each sample point the empirical distribution function $\hat{C}_N^{(1)}$ differs by at most $\frac{1}{3}(C(x_0)-c)$ from c, we have

$$|x^*-x_0|\leqq\epsilon.$$

Therefore, the probability that $|x^*-x_0|>\epsilon$ is at most equal to the probability that $|\hat{C}_N^{(1)}-C|\geqq3\eta$ for at least one sample point. Thus we get

$$P\{|x^*-x_0|>\epsilon\}\leqq\frac{1}{5N\eta^4}\left[1+\left\{\frac{p^4}{(p-\eta)^2}+\frac{q^4}{(q-\eta)^2}\right\}\left\{1-\frac{1}{5N^2\eta^4}\right\}\right]. \tag{15}$$

Therefore, when lower bounds for p and q are given beforehand, we can obtain a confidence interval for x_0.

However, in the above arguments, if the interval $(x_0-\delta,\ x_0+\delta)$ does not include certain members of O_N, δ must be extended as far as it contains a member of O_N lying within $(x_0-\delta,\ x_0+\delta)$, for it is meaningless to take 2δ less than the width of the interval in which $\hat{C}_N^{(1)}$ attains its maximum. As for ϵ and η, we are able to decide them so that $\epsilon \geqq \Delta$ and $\eta \leqq \int_{\Delta} |pf_1-qf_2|dx/3$, where $\Delta=|x_0-x|$, because $C(x_0)-C(x)=\int_{\Delta}|pf_1-qf_2|dx$ from the formula (7).

3. Numerical examples*

The following histograms show the frequencies of the observed values on the ultimate tensile strength (unit: kg/mm²) and the yield point (unit: kg/mm²) of the iron material (9 mmϕ: the diameter at a

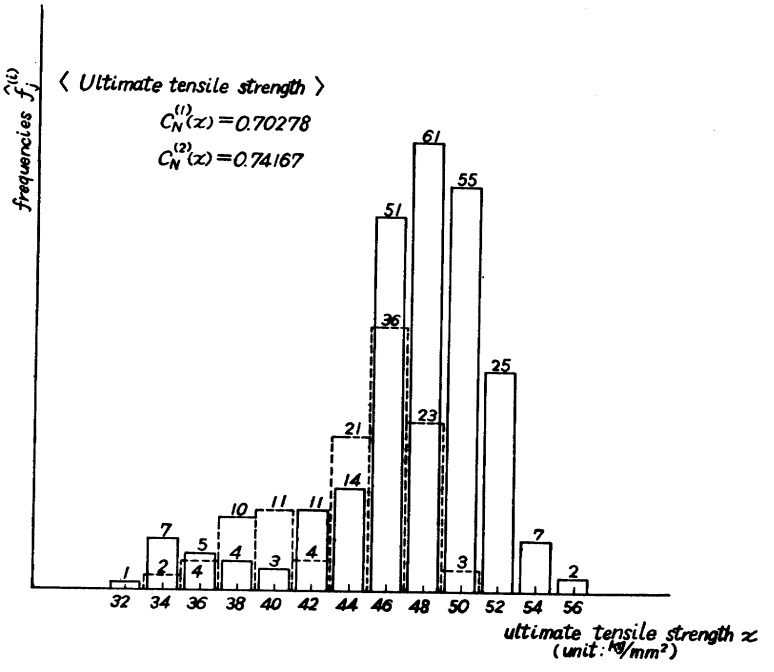

Fig. 1.

* These data were obtained by Mr. Takesaku Tutumi of the material test room in Nippon Telephone and Telegram Public Corporation, and numerical works were done by Kazuko Aihara and Eiko Ozaki of the Institute of Statistical Mathematics.

section) was made through the two different manufacturing processes. The dotted lines show the manufacturing process in which the products were rolled around the drum at the end of the production process and were cut at the normal temperature, and the real lines show the ordinary process.

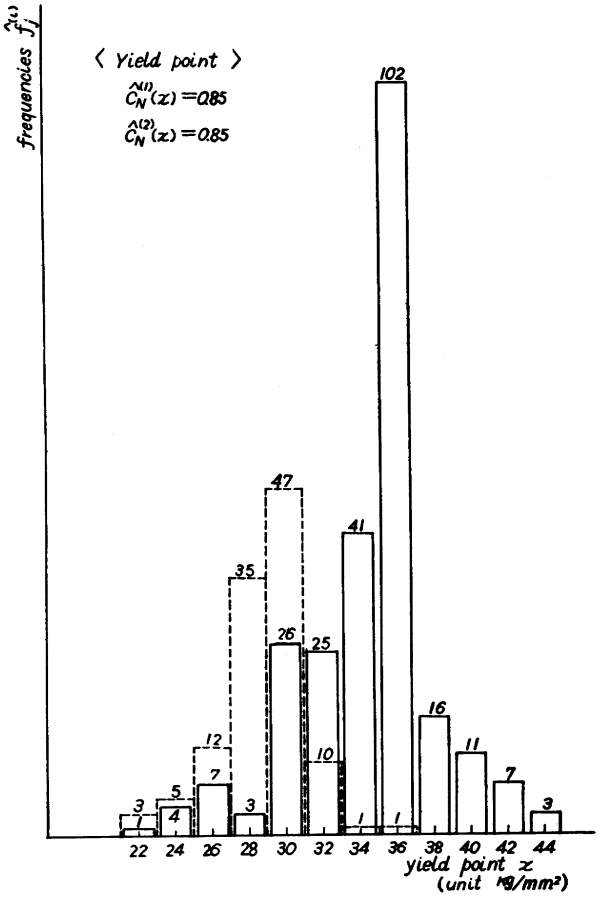

Fig. 2.

4. A relation between the correct probability $C(x_0)$ and the distance between F_1 and F_2

From the formula (7), it will easily be seen that the correct probability of the optimal classification is closely related to the concept of distance between F_1 and F_2.

In the case of $p=q=\frac{1}{2}$, from the optimal condition (4) we have

$(f_1-f_2)\geqq 0$ for $x\leqq x_0$, $(f_2-f_1)>0$ for $x>x_0$, and

$$C(x_0)=\frac{1}{2}+\frac{1}{4}\left[\int_R |f_1-f_2|dx\right]\geqq 1-\frac{1}{2}\rho(F_1, F_2)\ , \tag{16}$$

where $\rho(F_1, F_2)=\int_R \sqrt{f_1}\sqrt{f_2}dx$ is K. Matusita's affinity (see [1]). In case $p>q$, putting $p=\frac{1}{2}+\delta$, $q=\frac{1}{2}-\delta$, $q=p-2\delta$, from the optimal condition (4), we have $p(f_2-f_1)\leqq 2\delta f_2$ in w_0, and $p(f_2-f_1)>2\delta f_2$ in $R-w_0$, therefore,

$$C(x_0)=\frac{1}{2}+\frac{1}{2}\left[p\int_{w_0}(f_1-f_2)dx+2\delta\int_{w_0}f_2dx\right.$$
$$\left.+p\int_{R-w_0}(f_2-f_1)dx-2\delta\int_{R-w_0}f_2dx\right].$$

Now the negative parts of $p(f_1-f_2)$ are at most equal to $2\delta f_2$ in w_0, hence

$$C(x_0)\geqq\frac{1}{2}+\frac{1}{2}p\int_R|f_1-f_2|dx-\delta=q+\frac{1}{2}p\int_R|f_1-f_2|dx \tag{17}$$
$$\geqq q+p[1-\rho(F_1, F_2)]\ .$$

Similarly, in case $p<q$, we have

$$C(x_0)\geqq p+q[1-\rho(F_1, F_2)]\ . \tag{18}$$

Therefore, if p and q are given beforehand, $\hat{C}_N^{(2)}$ can be evaluated approximately using Theorem V in section 6 of [1].

THE INSTITUTE OF STATISTICAL MATHEMATICS

REFERENCES

[1] K. Matusita, Decision rules, based on the distance, for problems of fit, two samples, and estimation, *Ann. Math. Stat.*, Vol. 26 (1955).

[2] H. Aoyama, A note on the classification of observation data, *Ann. Inst. Stat. Math.*, Vol. II, No. 1 (1950).

[3] P. G. Hoel and R. P. Peterson, A solution to the problem of optimum classification, *Ann. Math. Stat.*, Vol. 20 (1949).

[4] A. Wald, On a statistical problem arising in the classification of an individual into one of two groups, *Ann. Math. Stat.*, Vol. 15 (1944).