إحصائيات التصنيف Z^* و Z

د. أحمد زوقو ميمون[1]

إذا كانت لدينا عينتان عشوائيتان من مجتمعين طبيعيين متعددين π_1 و π_2 وكانت لدينا مشاهدة X معرفة بانها تنتمي إلى احد المجتمعين π_1 أو π_2 فإننا نستطيع بإستخدام إحصائية التصنيف Z^* والتي هي عبارة عن إحصائية معدلة لإحصائية التصنيف Z تعيين ما إذا كانت المشاهدة X تنتمي إلى المجتمع π_1 أو المجتمع π_2

وعندما تحتوي العينتان على نفس الحجم فإن هذا البحث يستنتج توزيع Z^* ويشمل كذلك دراسة مقارنة لإحتمالات سوء التصنيف في حالة التصنيف الذي يعتمد على Z^* و Z .

(١) استاذ مشارك بقسم الإحصاء كلية الإقتصاد والتجارة/جامعة قاريونس .

sionality causes a considerable increase in D, the statistic $\overset{*}{Z}$ seems a better choice in reducing probability of misclassification.

REFERENCES

1. Cochran, W. G. and Bliss, C. I. (1948). Discriminant functions with covariance. Ann. Math. Statist. 19, 151–176.
2. Cramer, H. (1964). Mathematical Methods of Statistics. Princeton Univ. Press.
3. John, S. (1960). On some classification problems – 1. Sankhya 22, 301–308.
4. John, S. (1963). On classification by the statistics R and Z. Ann. Inst. Statist. Math. 14, 237–246.
5. Kudo, A. (1959). The classificatory problem viewed as a two-decision problem. Mem. Fac. Sci. Kyushu Univ. Ser. A. 13, 96–125.
6. Memon, A. Z. (1968). Z statistic in discriminant analysis. Ph.D. Dissertation, Iowa State University, Ames, Iowa.
7. Memon, A. Z. and Okamoto, M. (1970). The classification statistic $\overset{*}{W}$ in covariate discriminant analysis. Ann. Math. Statist. 41, 1491–99.
8. Memon, A. Z. and Okamoto, M. (1971). Asymptotic expansion of the distribution of the Z statistic in discriminant analysis. Journal of Multivariate Analysis. Vol. 1, 3, 294–307.
9. Memon, A. Z. (1976). Some results on $\overset{*}{Z}$ classification statistic. The Panjab University Journal of Mathematics, Vol. IX, 45–50.

with

$$A_1 = \tfrac{1}{2}D^{-2}\theta\{4\theta^3(D^2+2) + 4D\theta^2(D^2+2) + \theta[D^4 + 2D^2(3p+3q+5) + 8p] + 2D^3(p+q+1)\}, \qquad (2.8)$$

as the coefficient of $1/4N$ in $\Sigma_{i=1}^3 \overset{*}{L}_i(-\theta N/(N+1); D)$. The term B_1 is composed as

$$B_1 = B_{11} + B_{12} + B_{13}, \qquad (2.9)$$

these being coefficients of $1/8N^2$ in $\Sigma_{i=1}^3 \overset{*}{L}_i(-\theta N/(N+1); D), \tfrac{1}{2}[\Sigma_{i=1}^3 \overset{*}{L}_i(-\theta N/(N+1); D]^2$ and $\Sigma_{i\le j=1}^3 \overset{*}{Q}_{ij}(-\theta N/(N+1); D)$ respectively. Note that the terms $\overset{*}{L}_i$ and $\overset{*}{Q}_{ij}$ are defined in the paper (7). The coefficients in (2.9) are obtained as

$$B_{11} = -D^{-2}\theta^2[4\theta^2(3D^2+8) + 8D\theta(D^2+3) + D^4 + 2D^2(3p+3q+7) + 16p], \qquad (2.10)$$

$$B_{12} = \tfrac{1}{4}A_1^2, \qquad (2.11)$$

$$\begin{aligned} B_{13} = \tfrac{1}{6}D^{-4}\theta\{&8\theta^5(7D^4+24D^2+12) + 24\theta^4D(3D^4+10D^2+4) \\ &+ \theta^3[30D^6 + 3D^4(29p+29q+95) + 24D^2(10p+7q+17) + 48p] \\ &+ 4\theta^2D^3[D^4 + 3D^2(5p+5q+11) + 6(6p+5q+9)] \\ &+ 3\theta D^2[D^4(3p+3q+5) + 2D^2(6p^2+6q^2+12pq+19p+19q+15) \\ &+ 8(3p^2+4pq+4p)] + 6D^5(p+q+1)^2\}. \end{aligned} \qquad (2.12)$$

Now $\exp\left[-\dfrac{\theta D}{2(N+1)} + \dfrac{\theta^2N^2}{2(N+1)^2}\right]$ can be expanded with respect to N^{-1} as

$$(1 + A_2/N + B_2/2N^2 + \ldots)\, e^{\theta^2/2}, \qquad (2.13)$$

where

$$A_2 = -\theta D/2 - \theta^2, \qquad (2.14)$$

$$B_2 = \theta^4 + \theta^3 D + \theta^2(D^2/4 + 3) + \theta D. \qquad (2.15)$$

Using (2.6), (2.7) and (2.13) we have

$$\begin{aligned} \phi_1(\theta) &= (1 + A_1/4N + B_1/8N^2 + \ldots)(1 + A_2/N + B_2/2N^2 + \ldots)e^{\theta^2/2} \\ &= [1 + (A_1 + 4A_2)/4N + (B_1 + 4B_2 + 2A_1A_2)/8N^2 + \ldots]e^{\theta^2/2} \\ &= \left[1 + \frac{1}{4N}a_1(\theta) + \frac{1}{8N^2}a_2(\theta) + \ldots\right]e^{\theta^2/2}, \end{aligned} \qquad (2.16)$$

$a_1(\theta)$ and $a_2(\theta)$ are found to be as in (2.2) and (2.3). The use of Cramer's method (2, p. 225) of inverting a characteristic function of the above form completes the proof.

Since by interchanging $\bar{x}_1^*$ and $\bar{x}_2^*$ in (1.1) we obtain $-\overset{*}{Z}$, the following result is easily derived from above theorem.

Theorem 2

$$F(z|\pi_2) = 1 - [1 + a_1(d)/4N + a_2(d)/8N^2 + O_3]\Phi(-z). \qquad (2.17)$$

Corollary: The probability of misclassifying an observation into π_2 when it comes in fact from π_1 is given by

$$\{1 - [1 + a_1(d)/4N + a_2(d)/8N^2 + O_3]\Phi(z)\}_{z=D/2}. \qquad (2.18)$$

The other probability of misclassification has the same expression.

The above corollary follows from Theorem 1 and the procedure proposed in classifying an observation into its relevant population in the first section of this paper.

Remarks: Taking q = 0 in (2.1) the distribution of Z statistic is obtained as in (8).

Table 1 of (8) shows that the probability of misclassification decreases as the dimensionality p decreases or the Mahalanobis distance D^2 increases, when Z is used as a discriminant function. The terms $a_1(d)$ and $a_2(d)$ when compared with similar terms in (8) indicate a general increase in dimensionality. But as in using $\overset{*}{Z}$, D increases, it follows that the superiority of $\overset{*}{Z}$ over Z depends mainly on q and D. In situations where the introduction of a covariate of small dimenbeing the Mahalanobis distance between π_1 and π_2. Asymptotically, $(2D)^{-1}(\overset{*}{Z} + D^2)$ is a standardized normal variate. We shall obtain its distribution when $N_1 = N_2 = N$.

Theorem 1: If $D > 0$, an asymptotic expansion of the distribution of $(2D)^{-1}(\overset{*}{Z} + D^2)$ when $\binom{x}{y}$ comes from π_1 is given by

$$F(z|\pi_1) = (1 + a_1(d)/4N + a_2(d)/8N^2 + O_3)\,\Phi(z) \qquad (2.1)$$

where $d = d/dz$, $\Phi(z)$ is the cdf of N(0, 1), O_3 is the third order term with respect to N^{-1} and

$$a_1(d) = \tfrac{1}{2}D^{-2}\Big[4(D^2 + 2)d^4 + 4D(D^2 + 2)d^3 + \big[D^4 + 2D^2(3p + 3q + 1) + 8p\big]d^2 + 2D^3(p + q - 1)d\Big], \qquad (2.2)$$

$$\begin{aligned} a_2(d) = D^{-4}\Big[&(D^2 + 2)^2d^8 + 2D(D^2 + 2)^2d^7 + \tfrac{1}{6}[9D^6 + 2D^4(9p + 9q + 46) \\ &+ 12D^2(5p + 3q + 19) + 48(p + 2)]\,d^6 \\ &+ \tfrac{1}{2}D[D^6 + 2D^4(4p + 4q + 13) + 8D^2(3p + 2q + 10) + 16(p + 2)]d^5 \\ &+ \tfrac{1}{16}\{D^8 + 4D^6(7p + 7q + 17) + 4D^4[9(p + q)^2 + 76p + 72q + 135] \\ &+ 32D^2(3p^2 + 3pq + 21p + 14q + 18) + 64p(p + 2)\}d^4 \\ &+ \tfrac{1}{12}D^3\{D^4(3p + 3q + 5) + 18D^2[(p + q)^2 + 6p + 6q + 9] \\ &+ 24(p^2 + pq + 11p + 10q + 6)\}d^3 + \tfrac{1}{4}D^2\{D^4(p + q)^2 \\ &+ 4p + 4q + 7] + 4D^2]6(p + q)^2 + 13p + 13q + 13] + 16p(3p + 4q)\}d^2 \\ &+ D^5[(p + q)^2 + 2p + 2q + 5]d\Big]. \end{aligned} \qquad (2.3)$$

Proof: Memon and Okamoto (7) use the Fourier transform to obtain the distribution of $D^{-1}(\overset{*}{W} - \tfrac{1}{2}D^2)$; the limiting distribution of $\overset{*}{W}$ is $N(\tfrac{1}{2}D^2, D^2)$ as $\binom{x}{y} \in \pi_1$. Following the same approach the characteristic function of $(2D)^{-1}(\overset{*}{Z} + D^2)$ is

$$\phi_1(\theta) = \underline{E}\big[E\{\exp[-\theta(2D)^{-1}(\overset{*}{Z} + D^2)]|\pi_1\}\big], \qquad (2.4)$$

where $\theta = -it$ and $\underline{E}$ indicates expectation w.r.t. joint distribution of $(\bar{x}_1, \bar{x}_2, \bar{y}_1, \bar{y}_2, S)$. With $N_1 = N_2 = N$ as

$$\begin{aligned} \overset{*}{Z} &= \frac{N}{N + 1}[(\overset{*}{x} - \bar{x}_1^*)'\,\underline{S}^{-1}(\overset{*}{x} - \bar{x}_1^*) - (\overset{*}{x} - \bar{x}_2^*)'\,\underline{S}^{-1}(\overset{*}{x} - \bar{x}_2^*)] \\ &= \frac{N}{N + 1}[2\overset{*}{x}{}'\underline{S}^{-1}(\bar{x}_2^* - \bar{x}_1^*)' - (\bar{x}_1^* + \bar{x}_2^*)'\,\underline{S}^{-1}(\bar{x}_2^* - \bar{x}_1^*)] \\ &= \frac{-2N}{N + 1}\overset{*}{W}, \end{aligned} \qquad (2.5)$$

(2.4) can be written as

$$\underline{E}\left[E\left\{\exp\left[\left(\frac{N\theta}{N + 1}\right)\left(\frac{\overset{*}{W} - D^2/2}{D}\right) - \frac{\theta D}{N + 1}\right]\Bigg|\pi_1\right\}\right]$$

$$= e^{-\theta D/2(N+1)}\,\phi(-\theta N/(N + 1)) \qquad (2.6)$$

where ϕ is the same function as 4.1 of the paper (7). So, $\phi(-\theta N/N + 1)$ can be derived from there by changing θ to $-\theta N/N + 1$ in the expression for the asymptotic expansion of $\phi(\theta)$. We find that on simplifying,

$$\phi(-\theta N/(N + 1)) = (1 + A_1/4N + B_1/8N^2 + \dots)\, e^{\theta^2N^2/2(N+1)^2} \qquad (2.7)$$

Classification Statistics Z* and Z

AHMED ZOGO MEMON[1]

1. INTRODUCTION

Suppose we have an observation x from one of two p-variate populations $N(\mu_1, \Sigma_{11})$ and $N(\mu_2, \Sigma_{11})$ where the parameters μ_1, μ_2 and Σ_{11} are unspecified but Σ_{11} is positive definite. Given independent random samples from these populations, the problem of classification of x into its relevant population can be tackled by using the discriminant function Z proposed by Kudo (5) and John (3, 4) and studied by Memon (6), Memon and Okamoto (8). Sometimes there occur situations in texanomical problems when in addition to the knowledge of discriminators, information is available on a covariate y whose mean is known to be the same in both multivariate populations π_1 and π_2, that is, $\begin{pmatrix} x_i \\ y_i \end{pmatrix}$ has population π_i: $N\left[\begin{pmatrix} \mu_i \\ \nu \end{pmatrix}, \Sigma\right]$,

$i = 1, 2$, where $(x_i', y_i') = (x_{i1}, \ldots, x_{ip}, y_{i\,p+1}, \ldots, y_{i\,p+q})$ and the covariance matrix

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

is positive definite. Let

$$\begin{pmatrix} x_{i1} \\ y_{i1} \end{pmatrix}, \ldots, \begin{pmatrix} x_{iN_i} \\ y_{iN_i} \end{pmatrix} \quad i = 1, 2$$

be independent random samples from π_i. Although the covariate has no discriminating power by itself, Memon (6, 9) still proposes, like as in Cochran and Bliss (1), to utilize the additional information y in replacing x by $\overset{*}{x} = x - \hat{\beta}y$ in the statistic Z where $\hat{\beta}$ is a sample estimate of regression matrix β of x on y. According to this the modified criterion is

$$\overset{*}{Z} = \frac{N_1}{N_1 + 1} (\overset{*}{x} - \bar{x}_1^*)' \underset{=}{S}^{-1} (\overset{*}{x} - \bar{x}_1^*) - \frac{N_2}{N_2 + 1} (\overset{*}{x} - \bar{x}_2^*)' \underset{=}{S}^{-1} (\overset{*}{x} - \bar{x}_2^*), \qquad (1.1)$$

and the procedure of classification proposed by him is to assign $\binom{x}{y}$ to π_1 if $\overset{*}{Z} \leqq 0$ and to π_2 if $\overset{*}{Z} > 0$, where $\bar{x}_i^* = \bar{x}_i - \hat{\beta}\bar{y}_i$, $\hat{\beta} = S_{12} S_{22}^{-1}$, $\bar{x}_i$ and $\bar{y}_i$ denote the sample means, and $\underset{=}{S} = S_{11} - S_{12}S_{22}^{-1} S_{21}$ with

$$S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}$$

as the best unbiased estimator of Σ. Memon and Okamoto (7) study properties of the $\overset{*}{W}$ statistic and give asymptotic expansions of its distribution function and probabilities of misclassification that arise in using it. This paper follows the same approach in studying the $\overset{*}{Z}$ statistic when the sample from each population has the same size N, and thus extends the case $N_1 = N_2 = N$ of the paper by including the information on covariate in the discriminant function.

2. THE MAIN RESULT

We can easily see that as $N_1, N_2 \to \infty$, the limiting distribution of $\overset{*}{Z}$ is $N(-D^2, 4D^2)$ or $N(D^2, 4D^2)$ according as $\binom{x}{y} \in \pi_1$ or π_2, D^2

[1] Associate Professor, Statistics Department, Faculty of Economics and Commerce, University of Garyounis.