Statistics & Probability Letters 2 (1984) 19–21
North-Holland

A NOTE ON OPTIMUM GROUPING AND THE RELATIVE DISCRIMINATING POWER OF QUALITATIVE TO CONTINUOUS NORMAL VARIATES

Yung-tai HUNG and Anant M. KSHIRSAGAR

Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA

Received April 1983
Revised June 1983

Abstract: Ogawa (1951) considered the efficiency of estimation of the population mean from suitably chosen order statistics in large samples. Cox (1957) has considered the relative amount of information retained by grouping the normal curve. Cochran and Hopkins (1961) determined the discriminating power retained after partitioning normally distributed variates into qualitative ones in multivariate classification problems. And Connor (1972) discussed the asymptotic efficiencies of the test for the trend using m groups formed from a continuous variable. The same expression appears in all these investigations. This note throws some more light on the occurrence of the same expression in these seemingly unrelated problems.

Keywords: optimum grouping, efficiency of estimation using order statistics, multivariate classification with qualitative data.

1. Introduction

Consider a normal variate X with mean μ and variance σ^2, let the λ_i-quantile of the population be χ_i, $i = 1,\dots,m-1$, that is,

$$\lambda_i = \int_{-\infty}^{\chi_i} f(t)\,\mathrm{d}t, \quad 0 < \lambda_1 < \cdots < \lambda_{m-1} < 1, \tag{1.1}$$

where f denotes the density function of X. To estimate μ using the limiting distribution of suitably chosen order statistics, Ogawa (1951) derived the quantity

$$\varepsilon = \sum_{i=1}^{m} \frac{(\phi_i - \phi_{i-1})^2}{\Phi_i - \Phi_{i-1}} \tag{1.2}$$

where

$$\phi_0 = \Phi_0 = 0, \qquad \phi_i = \phi(U_i), \qquad \Phi_i = \Phi(U_i),$$
$$U_i = (\chi_i - \mu)/\sigma,$$

ϕ and Φ are the density and cumulative function respectively of an $N(0, 1)$ variate. ε is called the efficiency of estimation. Thus if σ^2 is known and m, the number of groups, is specified then an optimum spacing or maximum efficiency can be achieved by choosing the cutpoints χ_i, $i = 1,\dots, m-1$ so as to maximize (1.2).

On a somewhat different subject, however, in order to condense observations from X into a small number of groups while retaining as much information as possible, Cox (1957) proposed that group boundaries be chosen in such a way that the quantity

$$E\{(X - \xi_i)^2/\sigma^2\} \tag{1.3}$$

is minimized. Note that ξ_i denotes the mean of the ith group to which χ is assigned. (1.3) can also be written as

$$1 - \sum_{i=1}^{m} P_i(\xi_i - \mu)^2/\sigma^2 \tag{1.4}$$

where P_i is the probability of an observation falling into the ith group. The problem reduces to maximizing the second term of (1.4). The criterion was also proposed by Connor (1972) in a problem of maximizing the asymptotic efficiency of the test for the trend using optimal grouping. He also gave a list of other objectives relating to grouping which resulted in using this maximization criterion. We

0167-7152/84/$3.00 © 1984, Elsevier Science Publishers B.V. (North-Holland)

shall prove in this note that the quantity to be maximized here is mathematically identical to (1.2).

Yet on another different subject the same quantity appeared again in Cochran and Hopkins (1961). The standard discriminating procedure in the case of two K-variate normal populations Π_1 and Π_2 with different means but the same variance-covariance matrices gives the chance of misclassification $\Pr(\Pi_2 \mid \Pi_1)$ as

$$\alpha_1 = \Phi(-\tfrac{1}{2}\Delta_K) \tag{1.5}$$

where Δ_K is the Mahalanobis distance between the two populations. Suppose each continuous normal variable is now replaced by an m-category qualitative variable, and also suppose that all variables $X_1, \ldots, X_K$ are independent each with unit variance. Then let P_{si}, P'_{si} be the probability that an observation from Π_1 and Π_2 respectively falls into the ith group for the sth variate and let $R = \sum_{s=1}^{K} \log(P_{si}/P'_{si})$, Cochran and Hopkins (1961) found that using R as the classification criterion (that is, assigning an observation to Π_1 if $R \geqslant 0$, to Π_2 otherwise) the chance of misclassification, namely, $\Pr(R < 0 \mid \Pi_1)$ is asymptotically

$$\Phi\left(- \frac{E(R \mid \Pi_1)}{\sqrt{\operatorname{Var}(R \mid \Pi_1)}} \right) = \Phi(-\tfrac{1}{2}\Delta'_K), \tag{1.6}$$

say.

It turns out that minimizing (1.6) is the same as maximizing Δ'_K and Δ'^2_K/Δ^2_K is nothing but (1.2).

2. Proofs

We first show that the second term of (1.4) is in fact (1.2). Without loss of generality assume $\sigma^2 = 1$, $\mu = 0$ then

$$\sum_{i=1}^{m} P_i(\xi_i - \mu)^2/\sigma^2 = \sum_{i=1}^{m} P_i \xi_i^2. \tag{2.1}$$

Note that

$$P_i = \int_{U_{i-1}}^{U_i} \frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}t^2)\,\mathrm{d}t = \Phi_i - \Phi_{i-1} \tag{2.2}$$

where (u_i, u_{i-1}) are the boundaries of the ith group. Also,

$$\begin{aligned} \xi_i &= E(X \mid U_{i-1} < X < U_i) \\ &= \frac{1}{\Pr(U_{i-1} < X < U_i)} \int_{U_{i-1}}^{U_i} y \frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}y^2)\,\mathrm{d}y \\ &= (\phi_{i-1} - \phi_i)/(\Phi_i - \Phi_{i-1}). \end{aligned} \tag{2.3}$$

Putting ξ_i back into equation (2.1), the desired result follows.

We now show that Δ'^2_K/Δ^2_K is the same as (1.2). Assume that the mean of X_s $(s = 1, \ldots, K)$ is δ_s in Π_1 and 0 in Π_2, then assuming

$$\Phi\left(x - \frac{\delta}{2}\right) \doteq \Phi(x) - \frac{\delta}{2}\phi(x), \tag{2.4}$$

where $\doteq$ stands for approximate equality, neglecting higher order terms, we obtain,

$$\begin{aligned} P_{si} &= \int_{U_{i-1}+(1/2)\delta_s}^{U_i+(1/2)\delta_s} \frac{1}{\sqrt{2\pi}} \exp\left[-\tfrac{1}{2}(t-\delta_s)^2\right] \mathrm{d}t \\ &\doteq (\Phi_i - \Phi_{i-1}) + \tfrac{1}{2}\delta_s(\phi_{i-1} - \phi_i). \end{aligned} \tag{2.5}$$

Similarly,

$$P'_{si} \doteq (\Phi_i - \Phi_{i-1}) + \tfrac{1}{2}\delta_s(\phi_i - \phi_{i-1}). \tag{2.6}$$

so

$$\begin{aligned} \log\frac{P_{si}}{P'_{si}} &= \log\left[1 + \frac{\delta_s(\phi_{i-1} - \phi_i)}{\Phi_i - \Phi_{i-1} - (1/2)\delta_s(\Phi_{i-1} - \Phi_i)}\right] \\ &\doteq \frac{\delta_s(\phi_{i-1} - \phi_i)}{\Phi_i - \Phi_{i-1}}, \end{aligned} \tag{2.7}$$

ignoring higher order terms, and

$$\begin{aligned} E(R \mid \Pi_1) &= \sum_{s=1}^{K} \sum_{i=1}^{m} P_{si} \log(P_{si}/P'_{si}) \\ &\doteq \sum_{s=1}^{K} \tfrac{1}{2}\delta_s^2 \sum_{i=1}^{m} \frac{(\phi_{i-1} - \phi_i)^2}{\Phi_i - \Phi_{i-1}}, \end{aligned} \tag{2.8}$$

$$\operatorname{Var}(R \mid \Pi_1) \doteq \sum_{s=1}^{K} \delta_s^2 \sum_{i=1}^{m} \frac{(\phi_{i-1} - \phi_i)^2}{\Phi_i - \Phi_{i-1}} \tag{2.9}$$

(for derivations, see Kshirsagar (1972, p. 238)).

As a result,

$$\frac{E(R \mid \Pi_1)}{\sqrt{\operatorname{Var}(R \mid \Pi_1)}} \doteq \sqrt{\frac{1}{4} \sum_{s=1}^{K} \delta_s^2 \sum_{i=1}^{m} \frac{(\phi_{i-1} - \phi_i)^2}{\Phi_i - \Phi_{i-1}}}\,. \tag{2.10}$$

Putting (2.9) in (1.6), clearly,

$$\begin{aligned}\tfrac{1}{4}\Delta'^2_K &\doteq \tfrac{1}{4}\sum_{s=1}^{K}\delta_s^2\sum_{i=1}^{m}\frac{(\phi_{i-1}-\phi_i)^2}{\Phi_i-\Phi_{i-1}}\\ &=\tfrac{1}{4}\sum_{s=1}^{K}\delta_s^2 g(U),\end{aligned}\tag{2.11}$$

say.

Since $g(U)$ does not involve δ_s, the values of cutpoints that maximize $g(U)$ are the same for each s. Denote the maximum value of $g(U)$ by $g^*(U)$, the maximum of $\frac{1}{4}\Delta'^2_K$ will be $\frac{1}{4}\Delta^2_K g^*(U)$ where $\Delta^2_K = \Sigma^K_{s=1}\delta^2_s = K\delta^2_0$, say, is the square of the Mahalanobis distance between Π_1 and Π_2 based on the independent variables $X_1, \ldots, X_K$, all of which have unit variances. Note that assuming (2.4) and ignoring higher order terms in (2.7), Δ'^2_K is approximately equal to $\Sigma^K_{s=1}\delta^2_s g(U)$. Therefore Δ'^2_K/Δ^2_K is approximately equal to $g(U)$ which is (1.2). Hence the proof is completed.

References

Cochran, W.G. and C.E. Hopkins (1961), Some classification problems with multivariate qualitative data, *Biometrics* **17**, 10–32.

Connor, R.J. (1972), Grouping for testing trends in categorical data, *Journal of the American Statistical Association* **67**, 601–604.

Cox, D.R. (1957), Note on Grouping, *Journal of the American Statistical Association* **52**, 543–547.

Kshirsagar, A.M. (1972), *Multivariate Analysis*, Marcel Dekker, New York.

Ogawa, J. (1951), Contributions to the Theory of Systematic Statistics, I, *Osaka Mathematical Journal* **3**, 175–213.