A c-SAMPLE NON-PARAMETRIC TEST FOR LOCATION IN A MIXED MODEL OF CONTINUOUS AND DISCRETE VARIABLES*

By Shashikala Sukhatme

New Delhi

1. Introduction

Let $Z_i = (Y_i, X_{1i}, X_{2i}, \cdots X_{ci})$ $i = 1, 2, \cdots, N$ be N independent observations from a $(c+1)$ variate distribution where for each i,

$X_{ji} = 0$ or 1, $\sum_{j=1}^{c} X_{ji} = 1$, $P\{X_{ji} = 1\} = p_j$, $P\{X_{ji} = 0\} = q_j = 1 - p_j$,

$\sum_{j=1}^{c} p_j = 1$, and $P\{Y \leqslant y \mid X_j = 1\} = F_j(y)$, $j = 1, 2, \cdots, c$. The distribution functions $F_1, \cdots, F_c$ are assumed to be absolutely continuous. In this paper we propose a median test for testing the hypothesis $H_0: F_1 = \cdots = F_c$. For this purpose, divide the observations $Y_1, \cdots, Y_N$ into c sets according as $X_{ji} = 1$, $j = 1, 2 \cdots, c$. Let $U_{ji}, \cdots, U_{jn_j}$ ($n_j > 0$ for each j, $\sum_{j=1}^{c} n_j = N$) denote those Y_j's for which the corresponding $X_{ji} = 1$. For given $n_1, \cdots, n_c$ the problem of testing the hypothesis H_0 reduces to testing the hypothesis that the c independent samples of U_{ji}'s ($i = 1, 2, \cdots, n_j$; $j = 1, 2, \cdots, c$) come from the same distribution. However, the problem under consideration differs from the usual c-sample problem in that the sample sizes $n_1, \cdots, n_c$ are random variables having a multinomial distribution with parameters $p_1 \cdots, p_c$.

We assume that F_j's differ only in location. Let $F_j(y) = F(y + \theta_j)$, $j = 1, 2, \cdots, c$ for some arbitrary choice of real numbers $\theta_1, \cdots, \theta_c$. Further we denote by H_N the hypothesis which specifies that $F_j(y) = F(y + \theta_j/\sqrt{N})$, $j = 1, 2, \cdots, c$ and for some pair (i, j) $\theta_i \neq \theta_j$.

Let $\tilde{W}$ denote the sample median of Y observations and m_j the number of U_{ji}'s ($i = 1, 2, \cdots, n_j$) that are less than $\tilde{W}$. Assume

* This research was performed while the author was a graduate student at Michigan State University, East Lansing, Michigan, and was sponsored in part by the Office of Ordnance Research.

$N = 2k + 1$. Clearly $\sum_{j=1}^{c} m_j = k$. The test statistic proposed for testing the hypothesis $H: F_1 = \cdots = F_c$ is then defined as

$$M = \sum_{j=1}^{c} \left(\frac{m_j - kp_j}{\sqrt{Np_j}} \right)^2 , \tag{1.1}$$

when $p_1, \cdots, p_c$ are known, and as

$$\hat{M} = \left(\frac{m_j - k\hat{p}_j}{\sqrt{N\hat{p}_j}} \right) , \tag{1.2}$$

where $\hat{p}_j = n_j/N$, when $p_1, \cdots, p_c$ are unknown. The test consists in rejecting the hypothesis if $M(\hat{M})$ is large.

In Section 2 we find the joint distribution of $m_1, \cdots, m_c$ and $\tilde{W}$ and in Section 3 the limiting distribution of M. In Section 4 the relative asymptotic efficiency of the median test based on M with respect to a corresponding parametric test based on multiple correlation coefficient is evaluated. Section 5 deals with the case when $p_1, \cdots, p_c$ are unknown and gives the asymptotic distribution of $\hat{M}$ under the hypothesis H_c, from which we conclude that the test based on $\hat{M}$ is asymptotically distribution-free.

2. Joint Distribution of $m_1, \cdots, m_c$ and $\tilde{W}$

Henceforth $f(.)$ denotes the probability density function of the random variables written in the parentheses.

LEMMA 2.1. The joint distribution of $m_1, \cdots, m_c$ and $\tilde{W}$ is given by

$$f(m_1, \cdots, m_c, \tilde{w})$$

$$= \frac{N!}{k! \prod_{j=1}^{c} m_j!} \prod_{j=1}^{c} [p_j F_j(\tilde{w})]^{m_j}$$

$$\times \left[1 - \sum_{j=1}^{c} p_j F_j(\tilde{w}) \right]^k \left[\sum_{j=1}^{c} p_j F_j'(\tilde{w}) \right], \tag{2.1}$$

where $m_1, \cdots, m_c$ is a partition of k, $\sum_{j=1}^{c} m_j = k$.

Proof.—Noting that the conditional probability density of $m_1, \cdots, m_c$ and $\tilde{W}$ for fixed values of $n_1, \cdots, n_c$ is given by

$$f(m_1, \cdots, m_c, \tilde{w} \mid n_1, \cdots, n_c)$$
$$= \left[\sum_{j=1}^{c} \frac{(n_j - m_j)}{[1 - F_j(\tilde{w})]} F_j'(\tilde{w}) \right]$$
$$\times \left[\prod_{j=1}^{c} \binom{n_j}{m_j} (F_j(\tilde{w}))^{m_j} ((1 - F_j(\tilde{w}))^{n_j - m_j} \right] \qquad (2.2)$$

and that $n_1, \cdots, n_c$ have the multinomial distribution $m(N; p_1, \cdots, p_c)$ given by

$$f(n_1, \cdots, n_c) = \frac{N!}{\prod_{j=1}^{c} n_j!} \prod_{j=1}^{c} p_j^{n_j}, \qquad (2.3)$$

we obtain by using

$$f(m_1, \cdots, m_c, \tilde{w})$$
$$= \sum_{n_1, \ldots n_c} f(m_1, \cdots, m_c\, \tilde{w} \mid n_1, \cdots, n_c) f(n_1, \cdots, n_c)$$

the required joint probability density given by (2.1).

Summing (2.1) over $m_1, \cdots, m_c$ we obtain the marginal distribution of $\tilde{W}$,

$$f(\tilde{w}) = \frac{N!}{k!\,k!} \left[\sum_{j=1}^{c} p_j F_j(\tilde{w}) \right]^k \left[1 - \sum_{j=1}^{c} p_j F_j(\tilde{w}) \right]^k$$
$$\times \left[\sum_{j=1}^{c} p_j F_j'(\tilde{w}) \right].$$

Under H_c: $F_1 = F_2 = \cdots = F_c$, integration over the domain $0 \leqslant F_j(\tilde{w}) \leqslant 1$ in (2 1) yields the distribution of $m_1, \cdots, m_c$ as

$$f(m_1, \cdots, m_c) = \frac{k!}{\prod_{j=1}^{c} m_j!} \prod_{j=1}^{c} p_j^{m_j}$$

which is multinomial distribution $m(k; p_1, \cdots, p_c)$.

Also note that $(m_1, \cdots, m_c)$ and $\tilde{W}$ are independent under the hypothesis H_c.

3. Asymptotic Distribution of M

We first prove the following lemma which gives the joint limiting distribution of $m_1, \cdots, m_c$ and $\tilde{W}$.

Lemma 3.1. Let

$$\nu_j = \frac{m_j - Np_j F_j(\xi)}{\sqrt{Np_j F_j(\xi)}}, \quad j = 1, 2 \cdots, c; \quad \eta = \sqrt{N}(\tilde{w} - \xi),$$

where ξ is such that

$$\sum_{j=1}^{c} p_j F_j(\xi) = \tfrac{1}{2}. \tag{3.1}$$

Assume that in some neighbourhood of ξ the density function $F_j'(y) = f_j(y)$ $(j = 1, 2, \cdots, c)$ has a continuous derivative. Then the asymptotic joint distribution of $\nu_1, \cdots, \nu_{c-1}$ and η is c-variate normal distribution with zero mean vector and covariance matrix Σ given by $\Sigma^{-1} = \Lambda = (\lambda ij)$ where

$$\lambda_{ii} = 1 + \frac{p_i F_i(\xi)}{p_c F_c(\xi)}, \quad i = 1, 2, \cdots, (c-1);$$

$$\lambda_{cc} = \sum_{i=1}^{c} \frac{p_i f_i^2(\xi)}{F_i(\xi)} + 2\left[\sum_{i=1}^{c} p_i f_i(\xi)\right]^2;$$

$$\lambda_{ij} = \frac{[p_i p_j F_i(\xi) F_j(\xi)]^{\frac{1}{2}}}{p_c F_c(\xi)}, \quad i \neq j = 1, 2, \cdots, (c-1)\cdot$$

$$\lambda_{ic} = f_i(\xi)\sqrt{\frac{p_i}{F_i(\xi)}} - \frac{f_c(\xi)}{F_c(\xi)}\sqrt{p_i F_i(\xi)}, \quad i=1, 2, \cdots, (c-1).$$

Proof.—Throughout this proof for convenience set $F_i = F_i(\xi)$ and $f_i = f_i(\xi)$. Using Taylor's expansion

$$F_j(\tilde{w}) = F_j\left(\xi + \frac{\eta}{\sqrt{N}}\right) = F_j + \frac{\eta}{\sqrt{N}} f_j + o\left(\frac{\eta^2}{N}\right),$$

$$j = 1, 2, \cdots, c;$$

$$1 - \sum_{j=1}^{c} p_j F_j(\tilde{w}) = \frac{1}{2} - \frac{\eta}{\sqrt{N}} \sum_{j=1}^{c} p_j f_j + o\left(\frac{\eta^2}{N}\right).$$

and substituting these in (2.1) we get

$$f(m_1, \cdots, m_c, \tilde{w})$$

$$= \left\{\frac{N(2k)!}{k!\,k!\,2^{2k}}\right\} \left\{\frac{k!}{\prod_{i=1}^{c} m_i!} \prod_{j=1}^{c} (2p_jF_j)^{m_j}\right\}$$

$$\times \left\{\left(\prod_{i=1}^{c}\left[1 + \frac{\eta}{\sqrt{N}}\frac{f_i}{F_i} + o\left(\frac{\eta^2}{N}\right)\right]^{m_i}\right)\right.$$

$$\left.\times \left(1 - \frac{-2\eta}{\sqrt{N}} \sum_{i=1}^{c} p_i f_i - o\left(\frac{\eta^2}{N}\right)\right)^k\right\}$$

$$\times \left\{\sum_{j=1}^{c} p_j f_j \left(\xi + \frac{\eta}{\sqrt{N}}\right)\right\}$$

$$= \{A_1\}\{A_2\}\{A_3\}\{A_4\}. \tag{3.2}$$

Note that ν_j's satisfy the relation

$$\sum_{j=1}^{c} \nu_j (p_jF_j)^{\frac{1}{2}} = 0 \tag{3·3}$$

Now consider the region S defined by

$$S = \{(\nu_1, \cdots, \nu_{c-1}, \eta) : a_1 \leqslant \nu_1 \leqslant b_1, a_2 \leqslant \nu_2 \leqslant b_2, \cdots,$$
$$a_c \leqslant \eta \leqslant b_c\}.$$

Using Stirling's approximation for $n!$

$$A_1 \sim \frac{N}{\sqrt{k\pi}}.$$

A_2 is independent of η and because of convergence of multinomial distribution to normal distribution, uniformly in S

$$A_2 \sim [(2\pi)^{(c-1)} (2p_cF_c) \prod_{j=1}^{c} (2kp_jF_j)]^{-\frac{1}{2}}$$

$$\times \exp. -\frac{1}{2}\left[\sum_{i=1}^{c-1} \nu_i^2 \left(1 + \frac{p_iF_i}{p_cF_c}\right) + \sum_{i \neq j=1}^{c-1} \nu_i \nu_j \frac{\sqrt{p_ip_jF_iF_j}}{p_cF_c}\right]$$

Using series expansion for $\log(1 + x)$, uniformly in S

$$\log A_3 = -\frac{\eta^2}{2}\left[\sum_{i=1}^{c}\frac{p_i f_i^2}{F_i} + 2\left(\sum_{i=1}^{c} p_i f_i\right)^2\right]$$

$$+ \sum_{i=1}^{c} \eta \nu_i f_i \left(\frac{p_i}{F_i}\right)^{\frac{1}{2}} + o(1).$$

Using continuity of f_i, we have, uniformly in S,

$$f(m_2, \cdots, m_c, \tilde{w})$$

$$\sim N\left(\sum_{j=1}^{c} p_j F_j\right)\left[k\pi (2\pi)^{(c-1)} (2p_c F_c) \prod_{j=1}^{c} (2kp_j F_j\right]^{-\frac{1}{2}}$$

$$\times \exp. -\frac{1}{2}\left[\sum_{i-1}^{c-1} \nu_i^2\left(1 + \frac{p_i F_i}{p_c F_c}\right)\right.$$

$$+ \eta^2\left\{\sum_{i=1}^{c-1}\frac{p_i f_i^2}{F_i} + 2\left(\sum_{i=1}^{c} p_i f_i\right)^2\right\}$$

$$+ \sum_{i \neq j=1}^{c-1} \nu_i \nu_j \frac{\sqrt{p_i p_j F_i F_j}}{p_c F_c}$$

$$\left. - 2\eta \sum_{i=1}^{c-1} \nu_i \left\{f_i \left(\frac{p_i}{F_i}\right)^{\frac{1}{2}} - \frac{f_c}{F_c} (p_i F_i)^{\frac{1}{2}}\right\}\right].$$

Now making the transformation $(m_1, \cdots, m_c, \tilde{W}) \rightarrow (\nu_1, \cdots, \nu_c, \eta)$ it is seen that

$$\lim_{N\to\infty} P\{a_1 \leqslant \nu_1 \leqslant b_1, \cdots, a_c \leqslant \eta \leqslant b_c\}$$

$$= \int_{a_1}^{b_1}\int_{a_2}^{b_2} \cdots \int_{a_c}^{b_c} f(\nu_1, \cdots, \nu_{c-1}, \eta)\, d\nu_1 d\nu_2 \cdots d\eta$$

where $f(\nu_1, \cdots, \nu_{c-1}, \eta)$ is the probability density function of normal distribution described in the present theorem.

The following lemma gives the asymptotic joint distribution of $\nu_1, \cdots, \nu_{c-1}$ and η under the hypothesis H_N which specifies that $F_j(y) = F(y + \theta_j/\sqrt{N})$, $j = 1, 2, \cdots, c$.

LEMMA 3.2. Under the hypothesis H_N the asymptotic joint distribution of $(\nu_1 \cdots, \nu_{c-1}, \eta)$ is the c-variate normal distribution given by

$$f(\nu_1, \cdots, \nu_{c-1}, \eta)$$

$$= \frac{f(\xi)}{\pi^{c/2} 2^{(c-2)/2} p_c^{1/2}} \times \exp. -\frac{1}{2} \left[4\eta^2 f^2(\xi) + \sum_{i=1}^{c-1} \nu_i^2 \left(1 + \frac{p_i}{p_c}\right) + \sum_{i \neq j=1}^{c-1} \nu_i \nu_j \frac{\sqrt{p_i p_j}}{p_c} \right].$$

Proof.—It is similar to that of Lemma 3.1.

Now we are in a position to obtain the limiting distribution of M defined by (1 1) under the hypothesis H_N.

THEOREM 3.1. Under the hypothesis H_N the asymptotic distribution of $2M$ is non-central χ^2 with $(c-1)$ degrees of freedom and non-centrality parameter

$$\lambda = 2\,[F'(\xi)]^2 \sum_{j=1}^{c} p_j (\theta_j - \bar{\theta})^2, \tag{3.4}$$

where

$$\bar{\theta} = \sum_{j=1}^{c} p_j \theta_j.$$

Proof.—Write

$$u_i = \frac{m_i - kp_i}{\sqrt{Np_i}} = \frac{[m_i - Np_i F_i(\xi)]\sqrt{F_i(\xi)}}{\sqrt{Np_i F_i(\xi)}} - \frac{kp_i - Np_i F_i(\xi)}{\sqrt{Np_i}},$$

$i = 1, 2, \cdots, c.$

Under the hypothesis H_N using Lemma 3.2 it follows that the asymptotic joint distribution of $(u_1, \cdots, u_c)$ is c-variate normal with means $\mu_i = \theta_i F'(\xi) \sqrt{p_i}$, and covariance matrix $\Sigma = (\sigma_{ij})$ of rank $(c-1)$ where $\sigma_{ii} = (1 - p_i)/2$, $i = 1, 2, \cdots, c$, $\sigma_{ij} = -\sqrt{p_i p_j}/2$, $i \neq j = 1, 2, \cdots, c$.

Hence noting that $\sum_{j=1}^{c} \sqrt{p_j}\, u_j = 0$ it follows that the limiting distribution of

$$2M = 2\left[\sum_{i=1}^{c-1} u_i^2\left(1+\frac{p_i}{p_c}\right) + \sum_{i\neq j=1}^{c} u_i u_j \frac{\sqrt{p_i p_j}}{p_c}\right]$$

is $\chi^2_{c-1}(\lambda)$, where λ is given by (3.4).

4. Asymptotic Efficiency

Let $F_j(y) = F(y+\theta_j)$, then H_c is true when $\theta_j = 0$. We now find the relative asymptotic efficiency of the c-sample median test with respect to the corresponding parametric test, when F_j ($j = 1, 2, \cdots, c$) is a normal distribution with mean μ_j and variance σ^2. The hypothesis H_c is true if and only if $\rho^2_{Y(X_1,\ldots,X_c)} = 0$, (Olkin and Tate[1]) where $\rho_{Y(X_1,\ldots,X_c)}$ is the multiple correlation coefficient between Y and X. Let R denote the sample multiple correlation coefficient between Y and X. If

$$\bar{U}_{..} = \frac{\left(\sum_{i,j} U_{ij}\right)}{N}, \quad \bar{U}_{j.} = \frac{\left(\sum_{i=1}^{n_j} U_{ji}\right)}{n_j},$$

then

$$T^2 = \frac{R^2}{1-R^2} = \frac{\sum_{j=1}^{c} n_j(\bar{U}_{j.} - \bar{U}_{..})^2}{\sum_{j,i}(U_{ji} - \bar{U}_{..})^2 - \sum_{j=1}^{c}(\bar{U}_{j.} - \bar{U}_{..})^2}$$

Also

$$\rho^2_{Y(X_1,\ldots,X_c)} = \frac{\sum_{j=1}^{c} \frac{[(\mu_j - \bar{\mu})^2 p_j]}{\sigma^2}}{1 + \sum_{j=1}^{c} \frac{[(\mu_j - \bar{\mu})^2 p_j]}{\sigma^2}},$$

where

$$\bar{\mu} = \sum_{j=1}^{c} p_j \mu_j .$$

Following Fisher[2] it is seen that under the hypothesis H_N the asymptotic distribution of $(N - c)\, T^2/(c - 1)$ is $\chi^2_{(c-1)}(\lambda')$ where the non-centrality parameter is given by

$$\lambda' = \sum_{j=1}^{c} \frac{p_j\,(\mu_j - \bar{\mu})^2}{\sigma^2}.$$

Also it is proved that the limiting distribution of $2M$ is $\chi^2_{c-1}(\lambda)$, where λ is given by

$$\lambda = 2\,[F'(\xi)]^2 \sum_{j=1}^{c} \frac{p_j\,(\mu_j - \bar{\mu})^2}{\sigma^2}.$$

Since the two test statistics are asymptotically distributed as a non-central χ^2 with the same number of degrees of freedom, following Andrews[3] and Hannan[4] it is seen that the asymptotic efficiency is given by the ratio of the two non-centrality parameters. Hence the asymptotic relative efficiency is found to be

$$e\,(M, R) = 2\sigma^2\,[F'(\xi)]^2 = \frac{1}{\pi}.$$

5. Case when $p_1, \cdots, p_c$ are Unknown

In this case we estimate p_j by $\hat{p}_j = n_j/N$, $j = 1, 2, \cdots, c$ and consider the test based on $\hat{M}$ defined by (1.2). It is interesting to note that the test of H_c based on $\hat{M}$ is asymptotically distribution-free, which is seen from Theorem 5.1.

THEOREM 5.1. Under the hypothesis H_c, $4\hat{M}$ is asymptotically distributed as a χ^2 variable with $(c - 1)$ degrees of freedom.

Proof.—Write

$$v_j = \frac{m_j - k\hat{p}_j}{\sqrt{N\hat{p}_j}} = \left(\frac{p_j}{\hat{p}_j}\right)^{\frac{1}{2}} \frac{(m_j - k\hat{p}_j)}{\sqrt{Np_j}} = \left(\frac{p_j}{\hat{p}_j}\right)^{\frac{1}{2}} w_j,$$

where

$$w_j = \frac{m_j - kp_j}{\sqrt{Np_j}} - \frac{k\,(\hat{p}_j - p_j)}{\sqrt{Np_j}}.$$

Let $v = (v_1, \cdots, v_c)$ and $w = (w_1, \cdots, w_c)$, then $v = wD$, where D is a diagonal matrix with $\sqrt{p_j}/\sqrt{\hat{p}_j}$ as its diagonal elements. Since $\operatorname{plim}_{N\to\infty} \hat{p}_j = p_j$ it follows that $\operatorname{plim}_{N\to\infty} (\sqrt{p_j}/\sqrt{\hat{p}_j}) = 1$ and hence the matrix

D converges in probability (element-wise) to identity matrix. An application of lemma of [5, Lemma 1] yields that the vectors v and w have the same limiting distribution. Also it can be proved that the asymptotic distribution of w is c-variate normal with zero mean vector and covariance matrix $\Sigma = (\sigma_{ij})$ of rank $(c-1)$ with $\sigma_{jj} = (1-p_j)/4$, $j = 1, 2, \cdots, c$ and $\sigma_{ij} = -\sqrt{p_i p_j}/4$, $i \neq j$, $1, 2, \cdots, c$. Noting that $\sum_{j=1}^{c} \sqrt{p_j}\, v_j$ converges in probability to zero as $N \to \infty$, the asymptotic distribution of $v_1, \cdots, v_{c-1}$ is given by

$$f(v_1, \cdots, v_{c-1})$$

$$= \frac{1}{(2\pi)^{(c-1)/2} \left(\frac{1}{4}\right)^{(c-1)/2} p_c^{1/2}}$$

$$\times \exp. -2\left[\sum_{i=1}^{c-1} v_i^2\left(1 + \frac{p_i}{p_c}\right) + \sum_{i\neq j=1}^{c-1} v_i v_j \frac{\sqrt{p_i p_j}}{p_c}\right].$$

Hence

$$4\hat{M} = 4\left[\sum_{i=1}^{c-1} v_i^2\left(1 + \frac{\hat{p}_i}{\hat{p}_c}\right) + \sum_{i\neq j=1}^{c-1} v_i v_j \frac{\sqrt{\hat{p}\,\hat{p}_j}}{\hat{p}_c}\right]$$

has the asymptotic distribution stated in the theorem.

6. Acknowledgement

It is a pleasure to record my sincere thanks to Professor Ingram Olkin for suggesting the problem and for his keen interest in its solution.

7. References

1. Olkin, I. and Tate, R. F. .. "Multivariate correlation models with mixed discrete and continuous variables," *Ann. of Math. Stat.*, 1961, 27, 446–465.

2. Fisher, R. A. .. "The general sampling distribution of the multiple correlation coefficient," *Proc. Roy. Soc. Lond.*, 1928, **7 A** (121), 654–73.

3. Fred C. Andrews .. "Asymptotic behaviour of some rank tests for analysis of variance," *Ann. of Math. Stat.*, 1954, **25**, 724–36.

4. Hannan, E. J. .. "The asymptotic power of certain tests based on multiple correlation coefficient," *Journal Royal Stat. Soc., London*, 1956, **18 B**, 227–33.

5. Chin Long Chiang .. "On regular best asymptotically normal estimates," *Ann. of Math. Stat.*, 1956, **27**, 336–51.