

ANOVA extensions for mixed discrete and continuous data[☆]

A.R. de Leon*, Y. Zhu

Department of Mathematics & Statistics, University of Calgary, Calgary, AB, Canada T2N 1N4

Received 21 May 2006; received in revised form 21 July 2007; accepted 23 July 2007

Available online 7 August 2007

Abstract

This paper is concerned with ANOVA-like tests in the context of mixed discrete and continuous data. The likelihood ratio approach is used to obtain a location test in the mixed data setting after specifying a general location model for the joint distribution of the mixed discrete and continuous variables. The approach allows the problem to be treated from a multivariate perspective to simultaneously test both the discrete and continuous parameters of the model, thus avoiding the problem of multiple significance testing. Moreover, associations among variables are accounted for, resulting in improved power performance of the test. Unlike existing distance-based alternatives which rely on asymptotic theory, the likelihood ratio test is exact. In addition, it can be viewed as an extension to the mixed data setting of the classical multivariate ANOVA. We compare its performance against those of currently available tests via Monte Carlo simulations. Two real-data examples are presented to illustrate the methodology.

© 2007 Elsevier B.V. All rights reserved.

Keywords: General location model; Likelihood ratio test; Maximum likelihood; Multinomial distribution; Multivariate normal distribution; Power of the test; U-distribution

1. Introduction

Analysis of variance (ANOVA) is a popular and widely used statistical technique in many comparative studies. Whether comparing several treatments for depression in a clinical trial or discriminating between patients with favorable and unfavorable treatment prognoses, the ANOVA model is basic to a wide variety of statistical applications and is the appropriate procedure for testing location equality for several populations. However, in many medical and health studies, for example, treatment comparisons for some disease or disorder are carried out in terms of several outcome measures that include correlated continuous and discrete characteristics of the patients, in which case conventional ANOVA methods are not readily applicable.

Krzanowski (1980, 1975) describes data obtained in a study of psychosocial influences in breast cancer patients, conducted in King's College Hospital, London. They consist of measurements taken from two groups of women, one with malignant and another with benign breast tumors. There were a number of continuous and discrete variables which relate to psychosocial as well as physiological observations. One interest in the study is to see whether the two groups are different based on the observed outcomes. Mardia et al. (1979, p. 294) present data from university students on the number of GCE A-levels taken and the students' average grades (see also Morales et al., 1998;

[☆] This research is supported by the Natural Sciences and Engineering Research Council of Canada.

* Corresponding author. Tel.: +1 403 220 6782; fax: +1 403 282 5150.

E-mail address: adeleon@math.ucalgary.ca (A.R. de Leon).

Table 1
Empirical level and power of competing tests in the two-sample case with $C = 1$ and $S = 2$ at 5% level, based on 10,000 Monte Carlo samples

| Source of difference | Case | Test | $N_1 = N_2$ | | | $N_2 = 2N_1$ |
|------------------------------------------------|-------------------|-------------|-------------|----------|-----------|--------------|
| | | | $N = 40$ | $N = 80$ | $N = 120$ | $N = 120$ |
| No difference | (0) | LRT | 0.0513 | 0.0505 | 0.0497 | 0.0506 |
| | | Mahalanobis | 0.0698 | 0.0599 | 0.0547 | 0.0574 |
| | | Matusita | 0.0718 | 0.0611 | 0.0511 | 0.0576 |
| | | Bonferroni | 0.0475 | 0.0504 | 0.0471 | 0.0498 |
| Difference with respect only to \mathbf{x} | (a ₁) | LRT | 0.0741 | 0.1011 | 0.1301 | 0.1223 |
| | | Mahalanobis | 0.1054 | 0.1198 | 0.1436 | 0.1308 |
| | | Matusita | 0.1053 | 0.1194 | 0.1438 | 0.1311 |
| | | Bonferroni | 0.0654 | 0.1005 | 0.1261 | 0.1168 |
| | (a ₂) | LRT | 0.3433 | 0.6807 | 0.8511 | 0.8130 |
| | | Mahalanobis | 0.4200 | 0.7045 | 0.8591 | 0.8151 |
| | | Matusita | 0.4102 | 0.6988 | 0.8569 | 0.8127 |
| | | Bonferroni | 0.2619 | 0.6812 | 0.8686 | 0.8412 |
| Difference with respect only to Y | (b ₁) | LRT | 0.2719 | 0.5266 | 0.7272 | 0.6689 |
| | | Mahalanobis | 0.3177 | 0.5543 | 0.7418 | 0.6908 |
| | | Matusita | 0.3239 | 0.5577 | 0.7429 | 0.6911 |
| | | Bonferroni | 0.2666 | 0.5506 | 0.7479 | 0.6937 |
| | (b ₂) | LRT | 0.5051 | 0.8475 | 0.9645 | 0.9422 |
| | | Mahalanobis | 0.5633 | 0.8546 | 0.9657 | 0.9455 |
| | | Matusita | 0.5762 | 0.8589 | 0.9664 | 0.9481 |
| | | Bonferroni | 0.4248 | 0.7672 | 0.9279 | 0.9011 |
| Differences with respect to \mathbf{x} & Y | (c ₁) | LRT | 0.5325 | 0.8705 | 0.9740 | 0.9546 |
| | | Mahalanobis | 0.5820 | 0.8845 | 0.9779 | 0.9560 |
| | | Matusita | 0.5963 | 0.8893 | 0.9797 | 0.9568 |
| | | Bonferroni | 0.4355 | 0.7792 | 0.9383 | 0.9077 |
| | (c ₂) | LRT | 0.7459 | 0.9766 | 0.9990 | 0.9968 |
| | | Mahalanobis | 0.7961 | 0.9788 | 0.9985 | 0.9972 |
| | | Matusita | 0.7984 | 0.9790 | 0.9986 | 0.9972 |
| | | Bonferroni | 0.5489 | 0.9211 | 0.9902 | 0.9854 |

Krzanowski, 1983). The average A-level grade obtained is a continuous variable and the number of A-levels taken is a categorical variable. The students were also grouped according to their final degree classification into seven groups presented in Table 1 of Krzanowski (1983). One of the study objectives is to separate the degree classifications using the academic achievement data.

In these applications, one may proceed by testing hypotheses concerning the location parameters of the mixed data separately by applying conventional methods for discrete and continuous variables. Alternatively, one can consider the use of global tests to compare the parameters of several such populations (Pocock et al., 1987). Global tests based on an appropriate multivariate model for mixed data combine information from all the variables by fully exploiting the multivariate nature of the data thus resulting in increased power for the tests (de Leon, 2007; de Leon and Carrière, 2000). Unfortunately, standard multivariate approaches do not directly apply, and suitable methods have not been widely studied.

Despite the recent interest in mixed data analysis, only a few papers have considered similar problems. Afifi and Elashoff (1969) are the first to address ANOVA-like testing in mixed data situations. They consider the problem in the two-sample case, for which they derived likelihood ratio and information-theoretic tests. More recent work by Morales et al. (1998) introduces a general class of dissimilarity or entropy-type measures to obtain asymptotic tests for multisample location hypotheses involving mixed continuous and discrete data. Similar distance-based large-sample tests are provided by Bar-Hen and Daudin (1998, 1995) and Nakanishi (2003) for the two-sample case.

We revisit in this paper the problem of testing for differences among several groups with correlated mixed data. We adopt the general location model (GLOM) (Olkin and Tate, 1961) for the joint distribution of the mixed data and derive

a likelihood ratio test (LRT) for comparing the locations of mixed-variate populations. The proposed exact LRT, given in Section 3, can be viewed as a generalization to the mixed data setting of the classical ANOVA for continuous data (Mardia et al., 1979, Chapter 12). We investigate the performance of the test vis-à-vis other competing approaches for the two- and multisample cases via Monte Carlo simulations. The simulation results comparing the level and power of the tests are reported in Section 4. The methodology is illustrated by real-data examples and the relative merits of the proposed test are discussed in Section 5. Section 6 concludes the paper with a brief discussion.

2. Preliminaries: GLOM

Suppose $\mathbf{x}^\top = (X_1, \dots, X_S)$ and $\mathbf{y}^\top = (Y_1, \dots, Y_C)$ are vectors of binary and continuous variables, respectively, such that X_s is either 1 or 0 and $\sum_s X_s = 1$. The vectors \mathbf{x} and \mathbf{y} are said to be jointly distributed according to the GLOM if and only if (i) \mathbf{x} has probability distribution $p(\mathbf{x}) = \prod_s \pi_s^{x_s}$ and (ii) given $X_s = 1$, \mathbf{y} has the C -dimensional multivariate normal distribution with mean $\boldsymbol{\mu}_s^\top = (\mu_{1s}, \dots, \mu_{Cs})$ and covariance matrix $\boldsymbol{\Sigma}$, denoted $N_C(\boldsymbol{\mu}_s, \boldsymbol{\Sigma})$. The model is denoted by $GLOM(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\pi}^\top = (\pi_1, \dots, \pi_S)$ is the vector of state probabilities and $\boldsymbol{\mu}^\top = (\boldsymbol{\mu}_1^\top, \dots, \boldsymbol{\mu}_S^\top)$ is the $CS \times 1$ vector of state means. Here, $\pi_s = P(X_s = 1) > 0$ and $\sum_s \pi_s = 1$. A detailed discussion of this model is found in Schafer (1997, Chapter 9).

Assume independent samples from G $GLOM(\boldsymbol{\pi}_g, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ populations, with location parameters $\boldsymbol{\theta}_g^\top = (\boldsymbol{\pi}_g^\top, \boldsymbol{\mu}_g^\top)$, $g = 1, \dots, G$, and common covariance matrix $\boldsymbol{\Sigma}$, are available. This populations may arise as classifications, for example, of patients based on natural characteristics (e.g., sex, age) or as treatment groups (e.g., patients treated with different dosages of various drugs). The likelihood function is given by $\mathcal{L} = \prod_{g,s} \pi_{gs}^{n_{gs}} |2\pi\boldsymbol{\Sigma}|^{-N/2} \exp\{-\text{tr}[\boldsymbol{\Sigma}^{-1} \sum_{g,s} n_{gs} (\mathbf{S}_{gs} + \mathbf{d}_{gs}\mathbf{d}_{gs}^\top)/2]\}$, where $\mathbf{d}_{gs} = \bar{\mathbf{y}}_{gs} - \boldsymbol{\mu}_{gs}$, \mathbf{S}_{gs} and $\bar{\mathbf{y}}_{gs}$ are the sample mean and sample covariance matrix (uncorrected for bias), respectively, of the continuous data belonging to state s in population g , $N = \sum_g N_g = \sum_s n_{gs}$, $N_g = \sum_s n_{gs}$, $n_{gs} = \sum_g n_{gs}$ with n_{gs} the number of observations belonging to state s in population g , and $\text{tr}(\cdot)$ denotes the trace of a matrix. Maximum likelihood estimates (MLEs) of $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_G$ and $\boldsymbol{\Sigma}$ follow from standard results on multinomial and multivariate normal distributions, and are given by $\hat{\boldsymbol{\theta}}_g^\top = (\hat{\boldsymbol{\pi}}_g^\top, \bar{\mathbf{y}}_g^\top)$ and $N\hat{\boldsymbol{\Sigma}} = \sum_g N_g \mathbf{S}_{g\cdot}$, where for population $g = 1, \dots, G$, $N_g \hat{\boldsymbol{\pi}}_g^\top = (n_{g1}, \dots, n_{gS})$, $\bar{\mathbf{y}}_g^\top = (\bar{y}_{g1}, \dots, \bar{y}_{gS})$, and $N_g \mathbf{S}_{g\cdot} = \sum_s n_{gs} \mathbf{S}_{gs}$. The matrix $\mathbf{W} = N\hat{\boldsymbol{\Sigma}}$ is analogous to the within-group “sum of squares and products” (SSP) matrix in multivariate ANOVA (Mardia et al., 1979, p. 334).

The condition $n_{gs} > 0$ for all g, s , (i.e., each state has at least one observation) is necessary so that all unknown parameters will be estimable and should work well when the total sample size N is appreciably larger than the total number of states S . This results in a truncated multinomial distribution for $(n_{11}, \dots, n_{GS})^\top$, conditional on $\mathbf{n}^\top = (n_{\cdot 1}, \dots, n_{\cdot S})$. When this is not the case, a few states may be collapsed to reduce the number of parameters. Alternatively, linear restrictions may be imposed on the model as in Schafer (1997, p. 341), and the parameters in \mathcal{L} are then expressed in terms of the restricted parameters.

3. Mixed-data ANOVA: LRT

We extend the classical one-way multivariate ANOVA problem to mixed binary and continuous populations by testing $H: \boldsymbol{\theta}_1 = \dots = \boldsymbol{\theta}_G$ against K : at least one inequality. Under the hypothesis H of complete homogeneity, the G samples can be treated as constituting one sample from $GLOM(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$. The restricted MLEs are then obtained the usual way as $N\tilde{\boldsymbol{\pi}} = \sum_g N_g \hat{\boldsymbol{\pi}}_g$, $\tilde{\boldsymbol{\mu}}^\top = (\bar{y}_{\cdot 1}, \dots, \bar{y}_{\cdot S})$, and $N\tilde{\boldsymbol{\Sigma}} = \mathbf{W} + \mathbf{B}$, with $\mathbf{B} = \sum_{g,s} n_{gs} (\bar{\mathbf{y}}_{gs} - \bar{\mathbf{y}}_{\cdot s})(\bar{\mathbf{y}}_{gs} - \bar{\mathbf{y}}_{\cdot s})^\top$ the between-groups SSP matrix (Mardia et al., 1979, p. 334), where $n_{\cdot s} \bar{\mathbf{y}}_{\cdot s} = \sum_g n_{gs} \bar{\mathbf{y}}_{gs}$. The likelihood ratio statistic is $\lambda^{2/N} = b(n_{11}, \dots, n_{GS}) |\mathbf{W}| / |\mathbf{W} + \mathbf{B}|$, where $b^{N/2}(n_{11}, \dots, n_{GS}) = \prod_{g,s} N_g^{n_{gs}} n_{\cdot s}^{n_{gs}} / (N^{N_g} n_{gs}^{n_{gs}})$. Critical values and p -values are calculated using the following exact null distribution of $\lambda^{2/N}$, given \mathbf{n} :

$$P_H(\lambda^{2/N} \leq \ell | \mathbf{n}) = \sum_{n_{11}, \dots, n_{GS}} \frac{\omega(n_{11}, \dots, n_{GS})}{M} P\left(U \leq \frac{\ell}{b(n_{11}, \dots, n_{GS})}\right), \quad (1)$$

where $U \sim U(C; S(G-1), N-SG)$, the U -distribution of dimension C with degrees of freedom $S(G-1)$ and $N-SG$ (Mardia et al., 1979, p. 81), $M = \sum_{n_{11}, \dots, n_{GS}} \omega(n_{11}, \dots, n_{GS})$, and $\omega(n_{11}, \dots, n_{GS}) = \prod_s n_{\cdot s}! \prod_g (N_g! / \prod_s n_{gs}!)$. The summations in (1) are taken over all $\{n_{11}, \dots, n_{GS}\}$ such that $n_{gs} > 0$, for all g, s . Details of the derivations are found in Appendix A.

Note that $\lambda^{2/N}$ generalizes the LRT statistic in the classical one-way multivariate ANOVA (i.e., the case $S = 1$; see, e.g., Mardia et al., 1979, p. 138) to the case of mixed discrete and continuous data. It likewise extends the mixed-data one-sample LRT of de Leon (2007) to the multisample case. The LRT in the two-sample case appears in Afifi and Elashoff (1969) and is an extension of the LRT for testing equality of means of two multivariate normal distributions, or equivalently, the two-sample Hotelling T^2 test, to the case of mixed multinomial and continuous data.

While the LRT statistic is derived from the unconditional likelihood function, it is assessed conditionally on the total counts across samples for the categorical states. The conditioning is sensible because it eliminates the effect of the nuisance parameter π . It is also computationally convenient because it reduces the burden of getting critical values and p -values to summing tail probabilities for the U-distribution over the set of counts for the two-way *state* \times *sample* table with fixed marginal totals, instead of summing over all possible counts with fixed sample sizes. An S-plus program based on fast Fourier transform (Bilodeau and Brenner, 1999) to evaluate exact probabilities of U can be used to calculate critical values and p -values.

For the special case $C = 1$ and $S = 2$, the LRT statistic becomes $\lambda^{2/N} = b(n_1, \dots, n_G) \{1 + \sum_g [n_g (\bar{Y}_{g1} - \bar{Y}_{.1})^2 / (N\hat{\sigma}^2) + (N_g - n_g) (\bar{Y}_{g2} - \bar{Y}_{.2})^2 / (N\hat{\sigma}^2)]\}^{-1}$, where $b^{N/2}(n_1, \dots, n_G) = \hat{\pi}^n (1 - \hat{\pi})^{N-n} / \prod_g \hat{\pi}_g^{n_g} (1 - \hat{\pi}_g)^{N_g - n_g}$, $N\hat{\sigma}^2 = \sum_g [n_g S_{g1}^2 + (N_g - n_g) S_{g2}^2]$, $N\hat{\pi} = n$, $N_g \hat{\pi}_g = n_g$, \bar{Y}_{gs} and S_{gs}^2 are the respective sample state mean and variance of observations belonging to state s from population g , $n = \sum_g n_g$, $N = \sum_g N_g$, with n_g and $N_g - n_g$ the numbers of observations belonging to states 1 and 2, respectively, from population $g = 1, \dots, G$.

From properties of U-distributions (e.g., Bilodeau and Brenner, 1999, p. 181), the exact null distribution (1) of $\lambda^{2/N}$, given n , simplifies to $P_H(\lambda^{2/N} \leq \ell | n) = M^{-1} \sum_{n_1, \dots, n_G} \prod_g \binom{N_g}{n_g} P(2(G-1)F_{2(G-1), N-2G} / (N-2G) \geq b(n_1, \dots, n_G) / (\ell - 1))$, where $M = \sum_{n_1, \dots, n_G} \prod_g \binom{N_g}{n_g}$, with the summations taken over all $\{n_1, \dots, n_G\}$ such that $1 \leq n_g \leq N_g - 1$, for all g . The power $\beta(\theta_1, \dots, \theta_G | n)$ of the test at $\theta_1, \dots, \theta_G$, conditional on n , is obtained similarly and involves the multivariate extended hypergeometric distribution (Harkness, 1965).

For the two-sample case ($G = 2$), the power function $\beta(\theta_1, \theta_2 | n)$ of the test at $\theta_1 \neq \theta_2$, conditional on n , becomes

$$\beta(\theta_1, \theta_2 | n) = \sum_{n_1} \frac{\rho^{n_1} \binom{N_1}{n_1} \binom{N-N_1}{n-n_1}}{M_\rho} P\left(\frac{2}{N-4} F_{2, N-4}^{\delta(n_1, n_2)} \geq \frac{b(n_1, n_2)}{\ell_\alpha} - 1\right), \tag{2}$$

where $M_\rho = \sum_{n_1} \rho^{n_1} \binom{N_1}{n_1} \binom{N-N_1}{n-n_1}$, $F_{u,v}^\delta$ is the noncentral F random variable with degrees of freedom (u, v) and noncentrality parameter δ , $\delta(n_1, n_2) = [n_1 n_2 (\mu_{11} - \mu_{21})^2 / n + (N_1 - n_1)(N_2 - n_2) (\mu_{12} - \mu_{22})^2 / (N - n)] / \sigma^2$, ℓ_α is the α -critical value, $\rho = \pi_1 (1 - \pi_2) / [\pi_2 (1 - \pi_1)]$, with the summations taken over the set $\{\max(1, n - N_2 + 1), \dots, \min(n - 1, N_1 - 1)\}$. The LRT in this case is an extension of the LRT for testing equality of means of two normal distributions, or equivalently, the two-sample t -test, to the case of mixed binary and continuous data. It also extends the one-sample mixed-data LRT of de Leon and Carrière (2000) to the two-sample case.

4. Size and power investigation

We report in this section the results of Monte Carlo simulations comparing the performance of the LRT and other tests.

4.1. Other tests

4.1.1. Distance-based tests

Morales et al. (1998) propose using a class of f -dissimilarity measures for mixed data to obtain a test. Using the negative of the Matusita affinity (Matusita, 1967), the estimated f -dissimilarity becomes $\hat{D}_f = -\sum_s \prod_g \hat{\pi}_{gs}^{1/G} \exp\{(\sum_g \bar{y}_{gs})^\top \hat{\Sigma}^{-1} (\sum_g \bar{y}_{gs}) / (2G^2) - \sum_g \bar{y}_{gs}^\top \hat{\Sigma}^{-1} \bar{y}_{gs} / (2G)\}$, and the test statistic $2N(\hat{D}_f + 1)$ is asymptotically distributed as a mixture of $\chi_{S(C+1)-1}^2$ random variables, under H , where χ_{df}^2 is the chi-square distribution with df degrees of freedom. With equal sample sizes, $2N(\hat{D}_f + 1)$ is $\chi_{(G-1)[S(C+1)-1]}^2$ asymptotically, under H .

If $G = 2$, the above test is equivalent to that proposed by Bar-Hen and Daudin (1998) based on the estimated Matusita distance for mixed data (Krzanowski, 1983) given by $\widehat{D}_{\text{Mat}} = 2 - 2 \sum_s \sqrt{\widehat{\pi}_{1s} \widehat{\pi}_{2s}} \exp(-\widehat{A}_s^2/8)$, where $\widehat{A}_s^2 = (N - 2S)(\bar{\mathbf{y}}_{1s} - \bar{\mathbf{y}}_{2s})^\top \mathbf{W}^{-1}(\bar{\mathbf{y}}_{1s} - \bar{\mathbf{y}}_{2s})$ is the squared Mahalanobis distance in state $s = 1, \dots, S$. The test statistic is $4N_1N_2\widehat{D}_{\text{Mat}}/(N_1 + N_2)$, which follows $\chi_{S(C+1)-1}^2$ asymptotically, under H. Observe that $\widehat{D}_{\text{Mat}} = 2(1 + \widehat{D}_f)$ for $G = 2$ and equal samplesizes.

Another alternative in the two-sample case is provided by Nakanishi (2003) and Bar-Hen and Daudin (1995). The test is based on the estimated generalized Mahalanobis distance for mixed data given by $\widehat{D}_{\text{Mah}} = \sum_s (\widehat{\pi}_{1s} - \widehat{\pi}_{2s}) \log(\widehat{\pi}_{1s}/\widehat{\pi}_{2s}) + \frac{1}{2} \sum_s (\widehat{\pi}_{1s} + \widehat{\pi}_{2s}) \widehat{A}_s^2$. An asymptotic test is obtained by referring $N_1N_2\widehat{D}_{\text{Mah}}/(N_1 + N_2)$ to $\chi_{S(C+1)-1}^2$.

4.1.2. Bonferroni-corrected multiple tests

We also consider the separate-tests approach, which treats the parameters separately. The hypotheses $H_1 : \boldsymbol{\pi}_1 = \dots = \boldsymbol{\pi}_G$ and $H_{s+1} : \boldsymbol{\mu}_{1s} = \dots = \boldsymbol{\mu}_{G_s}$, $s = 1, \dots, S$, are tested simultaneously against the usual alternatives of at least one inequality. Hypothesis H_1 is tested using an asymptotic χ^2 -test (exact in the case $S = 2$) while the one-way multivariate ANOVA test is used to test each H_{s+1} , with $\boldsymbol{\Sigma}$ estimated by the pooled sample variance. A Bonferroni adjustment of each test's level is made by dividing the nominal level α by $S + 1$ to control the overall level of the tests.

In the two-sample case, the standard two-sample Hotelling T^2 test (or the two-sample t -test for $C = 1$) is used to test each H_{s+1} , with $\boldsymbol{\Sigma}$ estimated by the pooled sample variance.

4.2. Results

To assess the performance in finite samples of the exact LRT vis-à-vis the asymptotic tests described in the previous sections, we conducted a series of Monte Carlo simulations as follows. To evaluate the level and power of the LRT, the following four scenarios are considered: (0) no differences between populations; (a) there is difference between populations only with respect to binary vector \mathbf{x} ; (b) there is difference between populations only with respect to continuous vector \mathbf{y} ; and (c) populations are different with respect to both variable types. Note that (0) corresponds to the null case where H is true.

For the two-sample case ($G = 2$), we generate data from the GLOM with $C = 1$ and $S = 2$. This model corresponds to the case of a continuous variable Y_g and a binary vector $\mathbf{x}_g^\top = (X_{g1}, X_{g2})$, $g = 1, 2$. The location parameter is then $\boldsymbol{\theta}_g^\top = (\pi_g, \boldsymbol{\mu}_g^\top)$, where $\boldsymbol{\mu}_g^\top = (\mu_{g1}, \mu_{g2})$ with μ_{gs} the mean of Y for state $s = 1, 2$, in population g . Data are simulated from this GLOM with sample sizes $N = 40, 80$, and 120 , in the balanced case ($N_1 = N_2$), and $N_1 = 40, N_2 = 80$, in the unbalanced case, with 10,000 repeats. A check is performed at each repeat to ensure that each state has at least one observation (i.e., $n_{gs} > 0$ for all g, s).

For (0), we take $\pi_1 = \pi_2 = 0.5$, $\mu_{11} = \mu_{21} = 0$, $\mu_{12} = \mu_{22} = 5$, and $\sigma^2 = 1$. Note that, under the GLOM, \mathbf{x} and Y become uncorrelated (in fact, independent) if the two state means are equal. Conversely, the dependence becomes stronger when they are far from each other.

For (a), we consider cases (a₁) $\pi_2 = 0.4$ and (a₂) $\pi_2 = 0.8$. For (b), we have (b₁) $\boldsymbol{\mu}_2^\top = (0.8, 5)$ and (b₂) $\boldsymbol{\mu}_2^\top = (0.8, 5.8)$. For (c), we have (c₁) $\pi_2 = 0.4$ and $\boldsymbol{\mu}_2^\top = (0.8, 5.8)$; and (c₂) $\pi_2 = 0.8$ and $\boldsymbol{\mu}_2^\top = (0.8, 5.8)$.

We also consider the two-sample case with $C = 2$ and $S = 3$. This model corresponds to the case of a continuous vector $\mathbf{y}_g^\top = (Y_{g1}, Y_{g2})$ and a binary vector $\mathbf{x}_g^\top = (X_{g1}, X_{g2}, X_{g3})$, $g = 1, 2$. The location parameter is then $\boldsymbol{\theta}_g^\top = (\boldsymbol{\pi}_g^\top, \boldsymbol{\mu}_g^\top)$, where $\boldsymbol{\pi}_g^\top = (\pi_{g1}, \pi_{g2}, \pi_{g3})$ and $\boldsymbol{\mu}_g^\top = (\boldsymbol{\mu}_{g1}^\top, \boldsymbol{\mu}_{g2}^\top, \boldsymbol{\mu}_{g3}^\top)$, with $\boldsymbol{\mu}_{gs}$ the mean vector of \mathbf{y}_g for state $s = 1, 2, 3$, in population g . Data are simulated from this GLOM with sample sizes $N = 60, 120$, and 180 , in the balanced case ($N_1 = N_2$), and $N_1 = 60, N_2 = 120$, in the unbalanced case, with 10,000 repeats. The same check as above is performed at each repeat to ensure that each state has at least one observation (i.e., $n_{gs} > 0$ for all g, s).

For the null case (0), we take $\boldsymbol{\pi}_1^\top = \boldsymbol{\pi}_2^\top = (0.3, 0.4, 0.3)$, $\boldsymbol{\mu}_{11}^\top = \boldsymbol{\mu}_{21}^\top = (0, 2)$, $\boldsymbol{\mu}_{12}^\top = \boldsymbol{\mu}_{22}^\top = (5, 7)$, $\boldsymbol{\mu}_{13}^\top = \boldsymbol{\mu}_{23}^\top = (10, 12)$, $\text{var}(Y_{g1}) = \text{var}(Y_{g2}) = 1$, and $\text{cov}(Y_{g1}, Y_{g2}) = 0.5$. For (a), we consider cases (a₁) $\boldsymbol{\pi}_2^\top = (0.2, 0.5, 0.3)$, and (a₂) $\boldsymbol{\pi}_2^\top = (0.2, 0.6, 0.2)$. For (b), we have (b₁) $\boldsymbol{\mu}_{21}^\top = (0.7, 2.7)$ and $\boldsymbol{\mu}_{22}^\top = (5.7, 7.7)$; and (b₂) $\boldsymbol{\mu}_{21}^\top = (0.7, 2.7)$, $\boldsymbol{\mu}_{22}^\top = (5.7, 7.7)$, and $\boldsymbol{\mu}_{23}^\top = (10.7, 12.7)$. For (c), we have (c₁) $\boldsymbol{\pi}_2^\top = (0.2, 0.5, 0.3)$, $\boldsymbol{\mu}_{21}^\top = (0.7, 2.7)$, $\boldsymbol{\mu}_{22}^\top = (5.7, 7.7)$, and $\boldsymbol{\mu}_{23}^\top = (10.7, 12.7)$; and (c₂) $\boldsymbol{\pi}_2^\top = (0.2, 0.6, 0.2)$, $\boldsymbol{\mu}_{21}^\top = (0.7, 2.7)$, $\boldsymbol{\mu}_{22}^\top = (5.7, 7.7)$, and $\boldsymbol{\mu}_{23}^\top = (10.7, 12.7)$.

Finally, we consider the balanced three-sample case ($G = 3$) with $C = 1$ and $S = 2$. Data are simulated from this GLOM with sample sizes $N = 60, 120, 180$, and 240 , with $N_1 = N_2 = N_3$, and 10,000 repeats. The same check as above is performed at each repeat to ensure that each state has at least one observation (i.e., $n_{gs} > 0$ for all g, s).

Table 2
Empirical level and power of competing tests in the two-sample case with $C = 2$ and $S = 3$ at 5% level, based on 10,000 Monte Carlo samples

| Source of difference | Case | Test | $N_1 = N_2$ | | | $N_2 = 2N_1$ |
|---------------------------------------------------------|-------------------|-------------|-------------|-----------|-----------|--------------|
| | | | $N = 60$ | $N = 120$ | $N = 180$ | $N = 180$ |
| No difference | (0) | LRT | 0.0483 | 0.0494 | 0.0504 | 0.0492 |
| | | Mahalanobis | 0.0866 | 0.0656 | 0.0602 | 0.0620 |
| | | Matusita | 0.0814 | 0.0643 | 0.0593 | 0.0599 |
| | | Bonferroni | 0.0510 | 0.0495 | 0.0494 | 0.0495 |
| Difference with respect only to \mathbf{x} | (a ₁) | LRT | 0.0805 | 0.1185 | 0.1685 | 0.1581 |
| | | Mahalanobis | 0.1402 | 0.1574 | 0.1979 | 0.1796 |
| | | Matusita | 0.1251 | 0.1532 | 0.1934 | 0.1742 |
| | | Bonferroni | 0.0838 | 0.1257 | 0.1805 | 0.1624 |
| | (a ₂) | LRT | 0.1375 | 0.2822 | 0.4208 | 0.3633 |
| | | Mahalanobis | 0.2319 | 0.3435 | 0.4662 | 0.3990 |
| | | Matusita | 0.2118 | 0.3344 | 0.4580 | 0.3922 |
| | | Bonferroni | 0.1541 | 0.3316 | 0.4949 | 0.4256 |
| Difference with respect only to \mathbf{y} | (b ₁) | LRT | 0.3696 | 0.7176 | 0.9171 | 0.8766 |
| | | Mahalanobis | 0.4681 | 0.7594 | 0.9285 | 0.8931 |
| | | Matusita | 0.4641 | 0.7599 | 0.9292 | 0.8923 |
| | | Bonferroni | 0.3254 | 0.6507 | 0.8766 | 0.8327 |
| | (b ₂) | LRT | 0.5053 | 0.8878 | 0.9867 | 0.9670 |
| | | Mahalanobis | 0.6002 | 0.9117 | 0.9892 | 0.9728 |
| | | Matusita | 0.6005 | 0.9125 | 0.9890 | 0.9726 |
| | | Bonferroni | 0.3995 | 0.7623 | 0.9409 | 0.9101 |
| Differences with respect to \mathbf{x} & \mathbf{y} | (c ₁) | LRT | 0.5533 | 0.9199 | 0.9927 | 0.9815 |
| | | Mahalanobis | 0.6563 | 0.9377 | 0.9934 | 0.9844 |
| | | Matusita | 0.6495 | 0.9382 | 0.9941 | 0.9844 |
| | | Bonferroni | 0.4188 | 0.7912 | 0.9482 | 0.9149 |
| | (c ₂) | LRT | 0.6216 | 0.9524 | 0.9979 | 0.9941 |
| | | Mahalanobis | 0.7186 | 0.9646 | 0.9983 | 0.9949 |
| | | Matusita | 0.7133 | 0.9647 | 0.9984 | 0.9949 |
| | | Bonferroni | 0.4695 | 0.8441 | 0.9691 | 0.9477 |

For the null case (0), we take $\pi_1 = \pi_2 = \pi_3 = 0.5$, $\mu_{11} = \mu_{21} = \mu_{31} = 0$, $\mu_{12} = \mu_{22} = \mu_{32} = 6$, and $\sigma^2 = 1$. For (a), we consider cases (a₁) $\pi_3 = 0.65$, and (a₂) $\pi_2 = 0.35$, $\pi_3 = 0.65$. For (b), we have (b₁) $\mu_2^\top = (-0.5, 5.5)$, and (b₂) $\mu_2^\top = (-0.5, 5.5)$, $\mu_3^\top = (0.5, 6.5)$. For (c), we have (c₁) $\pi_3 = 0.65$, and $\mu_2^\top = (-0.5, 5.5)$, $\mu_3^\top = (0.5, 6.5)$; and (c₂) $\pi_2 = 0.35$, $\pi_3 = 0.65$, and $\mu_2^\top = (-0.5, 5.5)$, $\mu_3^\top = (0.5, 6.5)$.

Tables 1 and 2 report the results for the two-sample case with $C = 1$, $S = 2$, and $C = 2$, $S = 3$, respectively. Table 3 displays those for the three-sample case with $C = 1$, $S = 2$. In general, the LRT and the Bonferroni-corrected tests appear to have empirical levels close to the nominal level. While the LRT consistently attained the nominal level (i.e., all the empirical levels are within the 95% approximate confidence limits based on the binomial distribution), the Bonferroni-adjusted tests produced satisfactory levels for the majority of the cases, but exhibited slight conservatism for the two-sample case with $C = 1$, $S = 2$, $N_1 = N_2 = 20$. The levels of the Bonferroni-corrected tests are also quite unstable and produce high variabilities. The distance-based tests, on the other hand, had empirical levels that uniformly exceeded the nominal level. There is a noticeable improvement, however, in the latter's empirical levels as the sample sizes increase.

The results in Tables 1–3 also demonstrate that the LRT is reasonably powerful in rejecting the null hypothesis under various alternatives. As expected, by increasing the sample sizes, a more powerful test can be produced. The power also improves when differences between populations exists for both discrete and continuous data; moreover, the power increases with the distance between the parameters of the populations. It should be mentioned that the power values displayed for LRT in Table 1 were calculated using (2); simulated empirical power values were very close to the theoretical values, as expected.

Table 3

Empirical level and power of competing tests in the balanced three-sample case with $C = 1$ and $S = 2$ at 5% level, based on 10,000 Monte Carlo samples

| Source of difference | Case | Test | $N_1 = N_2 = N_3$ | | | |
|------------------------------------------------|-------------------|------------|-------------------|-----------|-----------|-----------|
| | | | $N = 60$ | $N = 120$ | $N = 180$ | $N = 240$ |
| No difference | (0) | LRT | 0.0490 | 0.0492 | 0.0506 | 0.0490 |
| | | Matusita | 0.0781 | 0.0621 | 0.0599 | 0.0554 |
| | | Bonferroni | 0.0499 | 0.0470 | 0.0495 | 0.0493 |
| Difference with respect only to \mathbf{x} | (a ₁) | LRT | 0.1053 | 0.1611 | 0.2476 | 0.3211 |
| | | Matusita | 0.1564 | 0.1943 | 0.2774 | 0.3453 |
| | | Bonferroni | 0.1083 | 0.1638 | 0.2551 | 0.3414 |
| | (a ₂) | LRT | 0.2312 | 0.4710 | 0.7055 | 0.8315 |
| | | Matusita | 0.3155 | 0.5204 | 0.7301 | 0.8467 |
| | | Bonferroni | 0.2569 | 0.5252 | 0.7516 | 0.8789 |
| Difference with respect only to Y | (b ₁) | LRT | 0.2190 | 0.4322 | 0.6195 | 0.7855 |
| | | Matusita | 0.2808 | 0.4721 | 0.6436 | 0.8025 |
| | | Bonferroni | 0.1859 | 0.3753 | 0.5437 | 0.7129 |
| | (b ₂) | LRT | 0.5939 | 0.9296 | 0.9913 | 0.9996 |
| | | Matusita | 0.6710 | 0.9414 | 0.9923 | 0.9996 |
| | | Bonferroni | 0.5225 | 0.8803 | 0.9819 | 0.9977 |
| Differences with respect to \mathbf{x} & Y | (c ₁) | LRT | 0.6405 | 0.9587 | 0.9972 | 1.0000 |
| | | Matusita | 0.7154 | 0.9683 | 0.9973 | 1.0000 |
| | | Bonferroni | 0.5382 | 0.9011 | 0.9851 | 0.9987 |
| | (c ₂) | LRT | 0.7659 | 0.9855 | 0.9995 | 1.0000 |
| | | Matusita | 0.8246 | 0.9886 | 0.9996 | 1.0000 |
| | | Bonferroni | 0.6058 | 0.9351 | 0.9947 | 0.9995 |

The findings regarding the Bonferroni correction confirm those obtained earlier by Pocock et al. (1987). The Bonferroni-corrected tests generally performed best for those cases in which the difference between populations is only with respect to one parameter. For those alternatives in which the populations differ with respect to two or more parameters, the approach seriously lacked power. Moreover, the empirical levels associated with the Bonferroni correction may not be stable.

The distance-based tests, on the other hand, exhibited good power in all the cases considered. However, the simulation results also indicated inflated empirical levels for the distance-based tests. Note as well that the power values of the distance-based tests get closer to the nominal 5% with increasing sample sizes. This is not surprising since these are asymptotic tests.

From the limited comparison studies, we can conclude that the LRT performs relatively well, exhibiting reasonably high power while at the same time controlling the levels to be at the nominally stated 5%. The Bonferroni correction, while generally able to achieve the nominal level, yielded comparatively lower power than those of the LRT and distance-based tests. The latter, while exhibiting empirical levels that are above the nominal 5%, also attained relatively good power.

5. Examples

5.1. Breast cancer data

The data were obtained in a study investigating psychosocial influences on breast cancer patients in London (Krzanowski, 1980). They consist of mixed discrete and continuous measurements taken on each of $N = 137$ women with breast tumors, $N_1 = 78$ of these being benign, and $N_2 = 59$ being malignant. We consider the psychosocial observation *direction of hostility*, which is scored in the range 0–10 and treated as a continuous variable, and the three-state nominal variable *feelings*. We are primarily interested in determining whether the two patient groups are significantly

different with respect to the mixed variables by modeling their joint distribution via the GLOM. The GLOM is appropriate in this case, as it allows us to account for the correlations in the data. Normality checks for *direction of hostility* for each of the three *feeling* states suggest that normality is a reasonable assumption.

Repeated two-sample t -tests on the state means and an asymptotic χ^2 -goodness of fit test for the multinomial probabilities yielded a minimum p -value of 0.0113. At a Bonferroni-adjusted overall 5% level (i.e., each test has level 0.0125), this indicates rejection of the hypothesis of no difference between the two groups. Applying now the LRT, we obtained a test statistic value of 0.1235 and an exact p -value of 0.0619, leading to acceptance of H_0 . The test based on \widehat{D}_{Mah} yielded an approximate p -value of 0.04809, indicating a marginally significant result, while that based on \widehat{D}_{Mat} gave an approximate p -value of 0.0519, resulting in the acceptance of the null hypothesis. Overall, we conclude that borderline evidence exists for difference between the two patient groups.

5.2. Academic achievement data

The data were collected from $N = 382$ university students on the number of GCE A-levels taken and the students' average grades (Mardia et al., 1979, p. 294). The average A-level grade obtained is a continuous variable and the number of A-levels taken is a categorical variable with $S = 3$ states (i.e., two, three or four A-levels). The students were also grouped according to their final degree classification into seven groups, among which we consider the groups II(ii) (i.e., students with lower second class degrees), III (i.e., students with third class degrees), and P (i.e., students who obtained a 'Pass') in what follows. Let Y_g denote the average A-level grade and let $\mathbf{x}_g^T = (X_{g1}, X_{g2}, X_{g3})$ represent the number of A-levels taken, with $X_{g1} = 1$ if four A-levels taken and 0 otherwise, $X_{g2} = 1$ if three A-levels taken and 0 otherwise, and $X_{g3} = 1$ if two-levels taken and 0 otherwise.

Repeated ANOVA F -tests on the state means and an asymptotic χ^2 -goodness of fit test for the multinomial probabilities yielded a minimum p -value of 0.0432. At a Bonferroni-adjusted overall 5% level (i.e., each test has level 0.0125), this indicates acceptance of the hypothesis of no difference among the three groups. Applying now the LRT, we obtained an exact p -value of 0.1819, leading to acceptance of H_0 . Morales et al.'s (1998) distance-based test yielded an approximate p -value of 0.3362, resulting in the same conclusion. Overall, we conclude that no evidence exists for differences among the three degree classifications.

6. Discussion

This paper is concerned with ANOVA-like tests of location for mixed multivariate data distributed according to the GLOM. By modeling the joint distribution of the mixed variables by the GLOM, the resulting exact LRTs can be viewed as extensions of classical normal theory ANOVA. The likelihood ratio approach was employed to construct global tests of mixed data location hypotheses because it allows for a general non-ad hoc approach of simultaneously accounting for both the discrete (i.e., multinomial) and continuous variables in the data, thus avoiding the problem of multiple testing. The approach parallels that of Afifi and Elashoff (1969) and is an alternative to the distance-based tests proposed by Nakanishi (2003), Bar-Hen and Daudin (1995, 1998) and Morales et al. (1998). These tests, it should be noted, are all asymptotic, unlike the exact LRTs proposed in the paper.

Results of a Monte Carlo study show that the LRT provides a valid procedure for hypothesis testing over a range of parameter configurations. The LRT is able to attain the nominal level, and has reasonably high power in various cases. While the distance-based tests provided generally satisfactory results, the simulations demonstrate them to be quite liberal, especially for small sample sizes.

Bonferroni-corrected multiple tests work reasonably well in most cases considered in the simulation study. Simulations indicate slight conservatism for small sample sizes, but no apparent deterioration is observed as the number of tests increases. It appears that the correction works best when only one test is significant and is least effective in cases where several parameters are different. The instability in the levels of the Bonferroni-corrected tests also poses a disadvantage. Although the Bonferroni correction is simple, albeit tedious, to apply, it considers only the minimum p -value among several tests, which may lead to skewed results (Pocock et al., 1987).

The assumption of homogeneity needs to be relaxed to widen the applicability of the proposed test. This can be done in two ways (Krzanowski, 1983). Within-population homogeneity or across-population heterogeneity occurs if we assume only that $\Sigma_{g1} = \dots = \Sigma_{gS} = \Sigma_g$, for $g = 1, \dots, G$. On the other hand, there is within-state homogeneity or

across-state heterogeneity if we assume only that $\Sigma_{1s} = \dots = \Sigma_{Gs} = \Sigma_{\cdot s}$, for $s = 1, \dots, S$. The same complications that it engenders in the case of continuous data are anticipated in the mixed data case.

A reviewer has suggested a possible extension of the LRT to variable selection in discriminant analysis. Letting $\lambda_{(q)}$ be the likelihood ratio statistic based on $q = 1, \dots, S + C - 1$, variables, and $\lambda_{(q+1)}$ be the corresponding ratio when a variable is added, a straightforward likelihood-based approach to carry this out is to test for a significant change in $\lambda_{(q)}$ by looking at the ratio $\lambda_{(q)}/\lambda_{(q+1)}$ and selecting that variable that maximizes $\lambda_{(q)}/\lambda_{(q+1)}$, or equivalently, minimizes $\lambda_{(q+1)}$. For $G=2$ and $S=1$ (i.e., no discrete variables), this test is equivalent to Rao's (1973, p. 368) test. This approach, however, is expected to share the same shortcomings of its counterpart in the continuous case (see, e.g., Rencher and Larson, 1980). This requires further study and the work is in progress. We hope to report the results in a future paper.

In conclusion, the results obtained in the paper unify and extend to mixed data settings, traditional ANOVA testing. The proposed LRT is easily understood and simple to apply, exact, and can readily accommodate any number of discrete and continuous variables. We thus recommend the test for most practical applications.

Acknowledgment

The authors are grateful to W.J. Krzanowski for providing them with the breast cancer data. They thank the Editor, Prof. S.P. Azen, an Associate Editor, and two referees for helpful comments and suggestions on earlier versions of the paper.

Appendix A. Derivation of LRT and its null distribution

The LRT statistic easily follows from the MLEs under H and K. Following Mardia et al. (1979, p. 138), let $\mathbf{W}_s^\top = (\mathbf{W}_{1s}^\top, \dots, \mathbf{W}_{Gs}^\top)$, for $s=1, \dots, S$, where \mathbf{W}_{gs} represents the n_{gs} observations belonging to state s from population g , $g=1, \dots, G$. It can be shown that $\mathbf{W} = \sum_s \mathbf{W}_s^\top \mathbf{C}_1 \mathbf{W}_s$ and $\mathbf{B} = \sum_s \mathbf{W}_s^\top \mathbf{C}_2 \mathbf{W}_s$, where $\mathbf{C}_1 = \sum_g [\text{diag}(\mathbf{1}_g) - \mathbf{1}_g \mathbf{1}_g^\top / n_{gs}]$ and $\mathbf{C}_2 = \sum_g (\mathbf{1}_g \mathbf{1}_g^\top / n_{gs} - \mathbf{1} \mathbf{1}^\top / n_{\cdot s})$, with $\mathbf{1}_g$ denoting the $n_{\cdot s} \times 1$ vector with 1 in the positions corresponding to the g th sample and 0 elsewhere, and $\mathbf{1} = \sum_g \mathbf{1}_g$.

Now under H, \mathbf{W}_s is a sample from $N_C(\mu_s, \Sigma)$ whence, by Theorems 3.4.4 and 3.4.5 of Mardia et al. (1979, p. 68), $\mathbf{W}_s^\top \mathbf{C}_1 \mathbf{W}_s \sim W_C(\Sigma, n_{\cdot s} - G)$ and $\mathbf{W}_s^\top \mathbf{C}_2 \mathbf{W}_s \sim W_C(\Sigma, G - 1)$ are independent, where $W_C(\Sigma, m)$ is the C -dimensional Wishart distribution with parameters Σ and m . Since $\mathbf{W}_1, \dots, \mathbf{W}_S$ are independent given $\{n_{11}, \dots, n_{GS}\}$, it follows that $\mathbf{W} \sim W_C(\Sigma, N - SG)$ and $\mathbf{B} \sim W_C(\Sigma, S(G - 1))$ are independent. Hence, given $\{n_{11}, \dots, n_{GS}\}$, $|\mathbf{W}|/|\mathbf{W} + \mathbf{B}| \sim U(C; S(G - 1), N - SG)$ under H, provided $N \geq C + GS$.

Next, the conditional joint distribution under H, of $\mathbf{n}_1, \dots, \mathbf{n}_G$ given $\mathbf{n} = \sum_g \mathbf{n}_g$, where $\mathbf{n}_g = (n_{g1}, \dots, n_{gS})^\top$, is obtained by observing that \mathbf{n}_g is multinomial with parameters N_g and π_g and, under H, \mathbf{n} is also multinomial with parameters N and $\pi_1 = \dots = \pi_G = \pi$. Thus, under H,

$$p_H^*(\mathbf{n}_1, \dots, \mathbf{n}_G | \mathbf{n}) = \frac{\prod_g (N_g! / \prod_s n_{gs}!)}{N! / \prod_s n_{\cdot s}!}.$$

The expression in (1) now follows.

References

- Afifi, A.A., Elashoff, R.M., 1969. Multivariate two sample tests with dichotomous and continuous variables. I. The location model. *Ann. Math. Statist.* 40, 290–298.
- Bar-Hen, A., Daudin, J.J., 1995. Generalization of the Mahalanobis distance in the mixed case. *J. Multivariate Anal.* 53, 332–342.
- Bar-Hen, A., Daudin, J.J., 1998. Asymptotic distribution of Matusita's distance: application to the location model. *Biometrika* 85, 477–481.
- Bilodeau, M., Brenner, D., 1999. *Theory of Multivariate Statistics*. Wiley, New York.
- de Leon, A.R., 2007. One-sample likelihood ratio tests for mixed data. *Comm. Statist. Theory Methods* 36 (1), 129–141.
- de Leon, A.R., Carrière, K.C., 2000. On the one-sample location hypothesis for mixed bivariate data. *Comm. Statist. Theory Methods* 29 (11), 2573–2581.
- Harkness, W.L., 1965. Properties of the extended hypergeometric distribution. *Ann. Math. Statist.* 36, 938–945.
- Krzanowski, W.J., 1975. Discrimination and classification using both binary and continuous variables. *J. Amer. Statist. Assoc.* 70, 782–790.
- Krzanowski, W.J., 1980. Mixtures of continuous and categorical variables in discriminant analysis. *Biometrics* 36, 493–499.
- Krzanowski, W.J., 1983. Distance between populations using mixed continuous and categorical variables. *Biometrika* 70, 235–243.

- Mardia, K.V., Kent, J.T., Bibby, J.M., 1979. *Multivariate Analysis*. Academic Press, New York.
- Matusita, K., 1967. On the notion of affinity of several distributions and some of its applications. *Ann. Inst. Statist. Math.* 19, 181–192.
- Morales, D., Pardo, L., Zografos, K., 1998. Informational distances and related statistics in mixed continuous and categorical variables. *J. Statist. Plann. Inference* 75, 47–63.
- Nakanishi, H., 2003. Test of hypotheses for the distance between populations on the mixture of categorical and continuous variables. *J. Japanese Soc. Comput. Statist.* 16, 53–62.
- Olkin, I., Tate, R.F., 1961. Multivariate correlation models with mixed discrete and continuous variables. *Ann. Math. Statist.* 32, 448–465 (correction in 36, 343–344).
- Pocock, S.J., Geller, N.L., Tsiatis, A.A., 1987. The analysis of multiple endpoints in clinical trials. *Biometrics* 43, 487–498.
- Rao, C.R., 1973. *Linear Statistical Inference and its Applications*. second ed. Wiley, New York.
- Rencher, A.C., Larson, S.F., 1980. Bias in Wilks' Λ in stepwise discriminant analysis. *Technometrics* 22, 349–356.
- Schafer, J.L., 1997. *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London.