

ON THE ONE-SAMPLE LOCATION HYPOTHESIS FOR MIXED BIVARIATE DATA

A. R. de Leon and K. C. Carrière

Department of Mathematical Sciences, University of Alberta
Edmonton, Alberta T6G2G1, Canada

Key Words : consistency; likelihood ratio test; location model; power function.

ABSTRACT

We study a hypothesis-testing problem involving the location model suggested by Olkin and Tate (1961). Specifically, we derive a likelihood ratio test of the associated location hypothesis as an alternative to the conventional method of carrying out separate tests for each of the parameters. A small sample Monte Carlo comparison indicates the general superiority of the former in terms of statistical power. We also comment briefly on the properties of the test.

1. INTRODUCTION

Data with mixtures of discrete and continuous variables frequently arise in practice. The analysis usually focuses on finding associations among the variables or on certain inferential problems about the parameters. A model used to study this type of data is the so-called location model. First introduced by Olkin and Tate (1961) and later extended by Little and Schluchter (1985) and Little and Rubin (1987), it has found numerous applications in multivariate, especially discriminant analysis (Krzanowski, 1993). Recent papers by Catalano and Ryan (1992), Fitzmaurice and Laird (1995), and Liu and Rubin

(1998) further extended the model by relaxing the homogeneity assumption and by allowing for covariates to be included in the model.

As the simplest such model, consider a discrete variable X which has a Bernoulli distribution, and a continuous variable Y whose conditional distribution for fixed X is normal. This model was studied in some detail by Tate (1954, 1955), who investigated the point biserial correlation as a measure of association between X and Y .

In this paper, we consider a bivariate random sample (X_i, Y_i) , $i=1, 2, \dots, n$, from the location model with parameters $\theta=(p, \mu_0, \mu_1)'$, and σ^2 . Here, X has a Bernoulli distribution with parameter p , and the conditional distributions of Y for $X=1$ and $X=0$ are assumed to be $N(\mu_1, \sigma^2)$ and $N(\mu_0, \sigma^2)$, respectively. The problem of interest is to test

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta \neq \theta_0,$$

for some specified $\theta_0=(p_0, \mu_{00}, \mu_{01})'$. The above null hypothesis is referred to in the literature as the one-sample location hypothesis, and much work has been done for the case with continuous data. Affifi and Elashoff (1969) tackled the two-sample mixed data problem and obtained two tests, one based on the Kullback-Liebler distance and another on the likelihood ratio approach. However, we are not aware of any work done in the one-sample case.

It is worth noting that the simple null hypothesis we consider here as well as elsewhere (for example, Affifi and Elashoff 1969) is of particular interest in such applications as quality control charting situations. There, the control limits are to be set up, simultaneously and more effectively charting both the discrete and continuous characteristics. In this context, the alternative hypothesis may correspond to a signal for the process being out of control. The absence of a signal in the multivariate chart precludes the presence of signals in the univariate charts.

In practice, the analytic strategy with mixed data has been to perform tests on the parameters separately. This approach entails the problem of multiple significance testing, to which the simplest solution is to adjust the level of each test to control the overall level. Such an approach may lose power quite substantially because the correlations between the variables are not utilized explicitly in constructing the test statistic (Pocock, Geller and Tsiatis, 1987). An alternative approach is to treat the problem in a multivariate setting to

come up with a single test based on all the variables. O'Brien (1984) and Pocock, Geller and Tsiatis (1987) studied one such global test statistic in the context of clinical trials. In this paper, we propose a test for mixed data using the likelihood ratio criterion.

The paper is organized as follows. We derive the likelihood ratio test and obtain the exact null and non-null distributions of the resulting test statistic in Section 2. The consistency and unbiasedness of the test are also briefly studied. The results of a power comparison of the proposed test against the commonly employed approach are presented in Section 3. We summarize our findings in Section 4.

2. THE LIKELIHOOD RATIO TEST

The likelihood under the location model is given by

$$L = p^{n_1}(1 - p)^{n-n_1}(2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{Q(\mu_0, \mu_1)}{2\sigma^2}\right\}, \tag{1}$$

where $Q(\mu_0, \mu_1) = \sum_{i \in A_0} (y_i - \mu_0)^2 + \sum_{i \in A_1} (y_i - \mu_1)^2$, $A_j = \{i | x_i = j\}$, $j = 0, 1$, and $n_1 = \sum_{i=1}^n x_i$.

It is well-known that the (unrestricted) maximum likelihood estimates of θ and σ^2 are

$$\hat{\theta} = (\hat{p}, \bar{y}_0, \bar{y}_1)', \quad \hat{\sigma}^2 = \frac{1}{n}Q(\bar{y}_0, \bar{y}_1), \tag{2}$$

where $\hat{p} = n_1/n$, $\bar{y}_j = \sum_{i \in A_j} y_i/n_j$, with $n_0 = n - n_1$. Note that $\hat{\sigma}^2$ is the usual (unadjusted) pooled variance estimate. Under H_0 , we only need to maximize the likelihood given in (1) with respect to σ^2 , and it is easy to see that the restricted maximum likelihood estimate in this case is

$$\hat{\sigma}_0^2 = \frac{1}{n}Q(\mu_{00}, \mu_{01}), \tag{3}$$

where μ_{00} and μ_{01} are the hypothesized values of μ_0 and μ_1 , respectively, under H_0 .

Let \hat{L} and \hat{L}_0 be the maximum likelihood estimates of (1) in the unrestricted and restricted (i.e., under H_0) cases, respectively. From (2) and (3), it is clear that \hat{L}/\hat{L}_0 yields the test statistic

$$\Lambda = A_{p_0}(n_1) \left(1 + \frac{2}{n-2}T^2\right), \tag{4}$$

where

$$T^2 = \left(\frac{n-2}{2}\right) \left\{ \frac{n_0(\bar{y}_0 - \mu_{00})^2 + n_1(\bar{y}_1 - \mu_{01})^2}{Q(\bar{y}_0, \bar{y}_1)} \right\},$$

and

$$A_{p_0}(n_1) = \left\{ \frac{\hat{p}^{n_1}(1-\hat{p})^{n-n_1}}{p_0^{n_1}(1-p_0)^{n-n_1}} \right\}^{2/n},$$

and we reject H_0 if Λ is too large. We see that the statistic given in (4) resembles the likelihood ratio statistic in the normal one-sample location problem (Bickel and Doksum, 1977, p. 212). Note as well that we require $n_1 \in [1, n-1]$ in order for all parameters to be estimable.

To derive the null distribution of (4), we note that $n_0(\bar{y}_0 - \mu_{00})^2/\sigma^2$, $n_1(\bar{y}_1 - \mu_{01})^2/\sigma^2$, and $Q(\bar{y}_0, \bar{y}_1)/\sigma^2$ are independent χ_1^2 , χ_1^2 and χ_{n-2}^2 variables, respectively, conditionally on n_1 . Because n_1 is binomially distributed with parameters (n, p_0) under H_0 , we have

$$\text{pr}(\Lambda > c | H_0) = \sum_{r=1}^{n-1} \binom{n}{r} \frac{p_0^r (1-p_0)^{n-r} \text{pr}(F_{2,n-2} > c_r)}{1-p_0^n - (1-p_0)^n}, \quad (5)$$

where $F_{u,v}$ denotes an F variable with df u and v and $c_r = (n-2)\{c/A_{p_0}(r) - 1\}/2$. Observe that T^2 and n_1 are independent under the null hypothesis. Critical values for Λ computed in Splus are listed in Table I for various values of n and p_0 at $\alpha=0.01$ and 0.05 . Note that Table I can also be used for $1-p_0=0.75$, 0.90 , and 0.95 , as the critical values are the same for these cases. For example, the critical value at $\alpha=0.01$ when $p_0=0.05$ with $n=30$ (i.e., 1.453066) is exactly the same as that when $p_0=0.95$.

The non-null distribution is obtained similarly except that the F variable now becomes a non-central F variable with df of 2 and $n-2$ and noncentrality parameter $\lambda = \{n_0(\mu_0 - \mu_{00})^2 + n_1(\mu_1 - \mu_{01})^2\}/\sigma^2$ (Johnson and Kotz, 1970). The power function of the test is then

$$\text{pr}(\Lambda > c_\alpha | \theta, \sigma^2) = \sum_{r=1}^{n-1} \binom{n}{r} \frac{p^r (1-p)^{n-r} \text{pr}\{F_{2,n-2}(\lambda_r) > c_{\alpha,r}\}}{1-p^n - (1-p)^n}, \quad (6)$$

where c_α is the $(1-\alpha)$ -quantile of the null distribution of Λ , $c_{\alpha,r} = (n-2) \times \{c_\alpha/A_{p_0}(r) - 1\}/2$, and $\lambda_r = \{(n-r)(\mu_0 - \mu_{00})^2 + r(\mu_1 - \mu_{01})^2\}/\sigma^2$. Probabilities and quantiles of non-central F distributions are readily available from standard statistical software for given degrees of freedom and non-centrality parameter.

Let $c_{\alpha,n}$ be the α -critical value of the test when the sample size is n .

TABLE I. Critical Values of the Null Distribution of Λ

p_0	α	n					
		10	15	20	25	30	50
0.05	0.01	3.798619	2.2596	1.791351	1.576424	1.453066	1.246445
	0.05	2.496786	1.741778	1.483547	1.360345	1.286645	1.161088
0.1	0.01	3.611616	2.211619	1.781555	1.577642	1.459268	1.255812
	0.05	2.381277	1.708373	1.477761	1.362857	1.294167	1.171379
0.25	0.01	3.626509	2.260955	1.825464	1.614376	1.489984	1.264061
	0.05	2.40079	1.752161	1.518428	1.397828	1.316231	1.175092
0.5	0.01	3.744688	2.327917	1.848998	1.622168	1.490199	1.263739
	0.05	2.504224	1.791348	1.527104	1.394488	1.315474	1.175038

Then, $c_{\alpha,n}$ satisfies

$$\text{pr}(\Lambda > c_{\alpha,n} | H_0) = \alpha.$$

Because $\chi_{n-2}^2 / (n-2) \rightarrow 1$ and $A_{p_0}(n_1) \rightarrow 1$ almost surely, we have $(n-2)(c_{\alpha,n} - 1) \rightarrow c_0$, where c_0 satisfies

$$\text{pr}(\chi_2^2 > c_0) = \alpha.$$

Hence, $c_{\alpha,n} \rightarrow 1$ as $n \rightarrow \infty$. By the strong law of large numbers, we get $T^2 \rightarrow \{(\mu_0 - \mu_{00})^2 + (\mu_1 - \mu_{01})^2\} / \sigma^2$ almost surely. Under the alternative hypothesis with $\mu_j \neq \mu_{0j}, j=0, 1$, we have $\{(\mu_0 - \mu_{00})^2 + (\mu_1 - \mu_{01})^2\} / \sigma^2 > 0$; thus the test is consistent in this case. However, the consistency fails to hold in the case where $p \neq p_0$ and $\mu_j = \mu_{0j}, j=0, 1$.

The unbiasedness of the test is easily established as well. By noting that

$$\text{pr}\{F_{2,n-2}(\lambda_r) > c_{\alpha,r}\} \geq \text{pr}(F_{2,n-2} > c_{\alpha,r}), \quad r = 1, \dots, n-1,$$

which is immediate from properties of the non-central F distribution (Bickel and Doksum, 1977, p. 303), and from properties of expectations (Casella and Berger, 1990, p. 56), we get

$$\text{pr}(\Lambda > c_\alpha | \theta) \geq \text{pr}(\Lambda > c_\alpha | \theta_0), \quad \theta \neq \theta_0,$$

which shows unbiasedness.

3. POWER COMPARISONS

In this section, we investigate the empirical power of the likelihood ratio test we derived against the separate test approach treating the parameters separately. The latter method entails carrying out tests of the following hypotheses simultaneously:

$$H_{01} : p = p_0 \quad \text{vs.} \quad H_{11} : p \neq p_0,$$

$$H_{02} : \mu_0 = \mu_{00} \quad \text{vs.} \quad H_{12} : \mu_0 \neq \mu_{00},$$

$$H_{03} : \mu_1 = \mu_{01} \quad \text{vs.} \quad H_{13} : \mu_1 \neq \mu_{01}.$$

The first pair above is tested using the exact binomial test while the latter two are tested using the standard one-sample t -test, using the pooled sample variance to estimate σ^2 . To control the overall level of the tests, a Bonferroni adjustment of each test's level is made by dividing the nominal level α by 3.

The relative power superiority of the likelihood ratio test compared to that of the separate test may be anticipated as the former utilizes the information about the dependency between the variables X and Y . Here we present the actual power values of the likelihood ratio test to confirm our conjecture as well as to show its relative merits over the separate test. As the power function for the separate test is not known, we compute values directly only for the likelihood ratio test using the power function given in (6), and use Monte Carlo simulation for the separate test. Samples of moderate sizes $n=15$ and 25 were generated from the location models with scale parameter $\sigma^2=25$ and location parameter $\theta=(p, \mu_0, \mu_1)'$ given by (a) $(0.35, 50, 25)'$, (b) $(0.35, 52.5, 22.5)'$, (c) $(0.4, 55, 22.5)'$, and (d) $(0.4, 55, 20)'$. In each case, the null parameter θ_0 was taken to be $(0.3, 50, 25)'$. To maximize the advantage of using the likelihood ratio test, the difference in the two mean values should be quite large and this influenced the choice of the parameters in our simulation. This is because, under the location model, X and Y become independent if the two means are equal. Conversely, the dependency becomes stronger when they are far

TABLE II. Power Comparisons of the Tests

Case	α	$n=15$		$n=25$	
		LRT ^a	ST ^b	LRT	ST
(0) ¹	0.05	0.05	0.0362	0.05	0.0429
	0.01	0.01	0.0093	0.01	0.0071
(a) ²	0.05	0.0626	0.0443	0.0656	0.0555
	0.01	0.0138	0.0089	0.0149	0.0096
(b) ³	0.05	0.3005	0.2358	0.5001	0.4258
	0.01	0.1142	0.0805	0.2484	0.1822
(c) ⁴	0.05	0.5974	0.4947	0.8499	0.7824
	0.01	0.3377	0.2384	0.6469	0.5243
(d) ⁵	0.05	0.8578	0.7533	0.9872	0.9664
	0.01	0.6199	0.4543	0.9306	0.8322

^aLikelihood ratio; ^bseparate test.; ¹ $\theta = (0.3, 50, 25)$ ';

² $\theta = (0.35, 50, 25)$ '; ³ $\theta = (0.35, 52.5, 22.5)$ ';

⁴ $\theta = (0.4, 55, 22.5)$ '; ⁵ $\theta = (0.4, 55, 20)$ '.

from each other, and it is precisely where we expect the likelihood ratio test to outperform the separate test. The performance of the likelihood ratio test will be equivalent to that of the separate test as the dependency between X and Y becomes negligible.

Table II presents the results of the power comparison. All samples were generated using Splus, with 10,000 repetitions in each case. The entries in Case (0) correspond to the situation when the null hypothesis is true, and hence give the levels (empirical in the case of the separate test) of the tests.

It is clear from Table II that the performance of the likelihood ratio test is superior to that of the separate test, as the power values are generally much higher for the former compared with those of the latter. This is true even in the case of a very slight departure from the true value as in (a) and (b). The comparison is most favorable to the likelihood ratio test when $\alpha=0.01$, and especially when $n=15$. This can be explained mainly by the fact that the separate test is a conservative method and becomes especially so for small

sample sizes. The entries in Case (0) indicate the actual level of the likelihood ratio test to be exactly at, and the separate test to be lower than the nominal level.

4. CONCLUDING REMARKS

This paper was concerned with the one-sample location hypothesis for mixed bivariate data. An ad-hoc approach that has been employed in practice entailing the carrying out of separate tests for each parameter of the location model was shown to be not very powerful. As an alternative, we have presented a test based on the likelihood ratio criterion that is simple and exact. It is simple in that (a) it provides a single test of the null hypothesis and thus avoids the problem of multiple testing; (b) the critical value can be easily computed; and (c) computing the statistic is straightforward, requiring no special software.

In terms of statistical power, the likelihood ratio test is also more appealing, in that it proved to be considerably more powerful than the conventional separate-test-for-each-parameter method in our simulations. In addition, it is unbiased as well as consistent, provided $\mu_j \neq \mu_{0j}$, $j=0, 1$. In conclusion, we recommend the likelihood ratio test, as it provides a practical and feasible test of the one-sample location hypothesis for mixed bivariate data.

ACKNOWLEDGEMENTS

A. R. de Leon was supported by a Studentship Award from the Alberta Heritage Foundation for Medical Research (AHFMR), AB, Canada. K. C. Carrière is a National Health Scholar with the National Health Research Development Program and a Heritage Senior Scholar with AHFMR. Additional support was provided by a grant from Natural Sciences and Engineering Research Council of Canada.

BIBLIOGRAPHY

- Affi, A. A., and Elashoff, R. M. (1969). Multivariate two sample tests with dichotomous and continuous variables. I. The location model. *Ann. Math. Statist.* **40**, 290-298.
- Bickel, P. J., and Doksum, K. A. (1977). *Mathematical Statistics: Basic Ideas and Selected Topics*. San Francisco: Holden Day.

- Casella, G., and Berger, R. (1990). *Statistical Inference*. Belmont: Duxbury.
- Catalano, P. J., and Ryan, L. M. (1992). Bivariate latent variable models for clustered discrete and continuous outcomes. *J. Am. Statist. Assoc.* **87**, 651-658.
- Fitzmaurice, G. M. and Laird, N. M. (1995). Regression models for a bivariate discrete and continuous outcome with clustering. *J. Am. Statist. Assoc.* **90**, 845-852.
- Johnson, N. L., and Kotz, S. (1970). *Continuous Univariate Distributions-2*. Boston: Houghton Mifflin.
- Krzanowski, W. J. (1993). The location model for mixtures of categorical and continuous variables. *J. Classification* **10**, 25-49.
- Little, R. J., and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley.
- Little, R. J., and Schluchter, M. D. (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika* **72**, 497-512.
- Liu, C., and Rubin, D. B. (1998). Ellipsoidally symmetric extensions of the general location model for mixed categorical and continuous data. *Biometrika* **85**, 673-688.
- O'Brien, P. C. (1984). Procedures for comparing samples with multiple endpoints. *Biometric* **40**, 1079-1087.
- Olkin, I., and 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., and Tsiatis, A. A. (1987). The analysis of multiple endpoints in clinical trials. *Biometrics* **43**, 487-498.
- Tate, R. F. (1955). Applications of correlation models for biserial data. *J. Am. Statist. Assoc.* **50**, 1078-1095.
- Tate, R. F. (1954). Correlation between a discrete and a continuous variable. *Ann. Math. Statist.* **25**, 603-607.

Received April, 2000; Revised May, 2000.