Ann. Inst. Statist. Math.
35 (1983), Part A, 63-75

SOME TESTS WITH UNBALANCED DATA FROM A BIVARIATE NORMAL POPULATION*

S. K. SARKAR, B. K. SINHA AND P. R. KRISHNAIAH

(Received Oct. 22, 1981; revised May 15, 1982)

Summary

Let (X, Y) follow a bivariate normal distribution, $N_2(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. In this paper, we have considered the problems of testing hypotheses $H_{01}: \mu_2=0$, $H_{02}: \mu_1=\mu_2$ and $H_{03}: \rho=0$ on the basis of an unbalanced data set-up consisting of n_1 paired observations, n_2 additional observations on X only and n_3 additional observations on Y only. Some new tests are proposed.

1. Introduction

Let us consider a situation where we want to infer about the characteristics of two variables, say X and Y. In many practical cases observing both X and Y may be expensive, time-consuming, or simply impossible because both may not be available simultaneously. On the other hand, we may have sample units providing information on X alone or Y alone. We would then observe both X and Y on some units and then X alone and/or Y alone on some other units. This kind of data set is often designated as incomplete or unbalanced.

Recently quite a bit of work has been done on some estimation and testing problems under an incomplete data set-up. The basic assumption is that X, Y follow a bivariate normal distribution $N_2(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. Suppose we have n_1 paired observations (X_i, Y_i), $i=1, 2, \cdots, n_1$, n_2 additional observations on X alone, X_i, $i=n_1+1, \cdots, n_1+n_2$, and n_3 additional observations on Y alone, Y_i, $i=n_1+n_2+1, \cdots, n_1+n_2+n_3$. In this paper we consider the following problems of testing hypotheses:

$$H_{01}: \mu_2=0\ , \quad H_{02}: \mu_1=\mu_2 \quad \text{and} \quad H_{03}: \rho=0$$

* This work was supported by the Air Force Flight Dynamics Laboratory under Grant AFOSR 77-3239 and the Air Force Office of Scientific Research under Contract F49629-82-K-001. Reproduction in whole or in part is permitted for any purpose of the United States Government.

under the cases (a) $n_1>0$, $n_2>0$, $n_3=0$ and (b) $n_1>0$, $n_2>0$, $n_3>0$.

The problem H_{01} under case (a) has been discussed in the papers by Khatri, Bhargava and Shah [5], Little [8], Tamhane [15] and most recently by Sarkar [13]. Usually the test is based on some studentized version of the maximum likelihood estimate (also known as the regression estimate) of μ_2. The test is not similar and the distribution of the resultant test statistic is very complicated. The test proposed by Sarkar is, of course, exact and similar and is based on an idea in Scheffé [14]. In Section 2 we put forward a few more solutions to this problem. Applying Wijsman's D-method (Wijsman [16]), a variety of similar critical regions are obtained in Subsection 2.1. When ρ is known, a locally optimum test of H_{01} can be obtained (Subsection 2.2).

The problem H_{02} under case (a) has been treated by Mehta and Gurland [9], [10], Morrison [11], Lin [7], Naik [12] and Little [8]. None of these tests is however exact similar. We propose an exact similar test in Section 3 by using a method similar to the one used by Scheffé [14] and Sarkar [13] for other problems.

For H_{03}, Eaton and Kariya [3] derive the locally most powerful invariant (LMPI) test under (a). For this problem, practical statisticians, however, often use a statistic which is based on the sample correlation coefficient obtained after predicting the missing Y-values from the observed values. In Section 4, we judge the merit of this test relative to a slight modification of the LMPI test and the usual correlation test which ignores the additional observations on X.

In Section 5, we deal with all of the above-mentioned problems under data set-up (b). We propose some similar tests for H_{01} and H_{02}. The second problem was considered by Bhoj [2]. The third problem has been considered by Eaton and Kariya [3], who derive the LMPI test by using Wijsman's representation theorem (Wijsman [17]). In Subsection 5.3, this is derived directly using an explicit expression of a maximal invariant statistic. A modification of this statistic which has a convenient distribution when the samples are moderately large is proposed.

2. Tests for H_{01} in case (a)

2.1. *Construction of a class of similar regions using D-method of Wijsman*

For the testing problem under consideration, we see that the underlying sufficient statistic, although belongs to a family of regular exponential densities, is not complete when H_{01} holds. Hence, we can use Wijsman's D-method (Wijsman [16]) for the construction of a class of similar regions.

Let $\boldsymbol{T}=(T_1,\cdots,T_m)$ denote a sufficient statistic for the parameter $\boldsymbol{\theta}$ in a given problem and let the density of $\boldsymbol{T}$ under the null hypothesis $H_0:\boldsymbol{\theta}\in\omega$ with respect to the m-dimensional Lebesgue measure be of the form

$$p_{\boldsymbol{\theta}}(\boldsymbol{t})=c(\boldsymbol{\theta})\exp\left(-\sum_{i=1}^{m}s_i(\boldsymbol{\theta})t_i\right)h(\boldsymbol{t})\,, \tag{2.1.1}$$

where $s_1,\cdots,s_m$ are real-valued functions of $\boldsymbol{\theta}$. Then $p_{\boldsymbol{\theta}}(\boldsymbol{t})$ is regular if $h(\boldsymbol{t})$ is bounded away from 0 on a closed m-dimensional cube C. We note that such a C can always be found in the problem under consideration. Under the incompleteness of the data, the fact that the sufficient statistic is not complete when H_0 holds is revealed by some parametric relations connecting the s's in its distribution under H_0. Let these relations be put into the form $P(s_1,\cdots,s_m)=0$, where P is a polynomial of degree d. Choose $G(\boldsymbol{t})$, a function of $\boldsymbol{t}$, such that $G(\boldsymbol{t})$ possesses all partial derivatives of dth order in the interior of C, vanishes outside C and has all partial derivatives of $(d-1)$ order continuous on the boundary of C. Then denoting by D the differential operator $D=P(\partial/\partial t_1,\cdots,\partial/\partial t_m)$, Wijsman suggested using $\phi(\boldsymbol{t})=\alpha+DG(\boldsymbol{t})/h(\boldsymbol{t})$ as a size α similar test of non-Neyman structure. Of course, here G has to be chosen suitably (subject to the restrictions mentioned above) to ensure that

$$-\alpha\leqq DG(\boldsymbol{t})/h(\boldsymbol{t})\leqq 1-\alpha \qquad \text{for all } \boldsymbol{t}\,. \tag{2.1.2}$$

Coming back to the problem of testing $H_{01}:\mu_2=0$, we note that the joint density of the sufficient statistics under H_{01} can be put in the exponential form (2.1.1) with

$$t_1=\sum_{i=1}^{n_1}x_i^2\,,\qquad t_2=\sum_{i=1}^{n_1}x_i\,,\qquad t_3=\sum_{i=1}^{n_1}y_i^2\,,\qquad t_4=\sum_{i=1}^{n_1}x_iy_i\,,$$

$$t_5=\sum_{i=1}^{n_1}y_i\,,\qquad t_6=\sum_{i=n_1+1}^{n_1+n_2}x_i^2\,,\qquad t_7=\sum_{i=n_1+1}^{n_1+n_2}x_i\,,$$

$$s_1=1/2\sigma_1^2(1-\rho^2)\,,\qquad s_2=-\mu_1/\sigma_1^2(1-\rho^2)\,,\qquad s_3=1/2\sigma_2^2(1-\rho^2)\,,$$

$$s_4=-\rho/\sigma_1\sigma_2(1-\rho^2)\,,\qquad s_5=\rho\mu_1/\sigma_1\sigma_2(1-\rho^2)\,,\qquad s_6=1/2\sigma_1^2\,,$$

$$s_7=-\mu_1/\sigma_1^2\,,$$

$$h(\boldsymbol{t})=(t_6-t_7^2/n_2)^{(n_2-3)/2}[(t_1-t_2^2/n_1)(t_3-t_5^2/n_1)-(t_4-t_2t_5/n_1)^2]^{(n_1-3)/2}\,,$$

$$C=\{\boldsymbol{t}:t_6\geqq t_7^2/n_2,\ t_1\geqq t_2^2/n_1,\ t_3\geqq t_5^2/n_1,\ (t_1-t_2^2/n_1)(t_3-t_5^2/n_1)\geqq(t_4-t_2t_5/n_1)^2\}\,.$$

Also, there exist three parametric relations of degree 2 in the problem given by $P_1(\boldsymbol{s})=s_1s_7-s_2s_6=0$, $P_2(\boldsymbol{s})=s_4s_7-2s_5s_6=0$, $P_3(\boldsymbol{s})=s_4^2+4s_3s_6-4s_1s_3=0$. It is possible to construct a similar region of non-Neyman structure using any one of these relations. Quite generally, we can choose

$G(\boldsymbol{t})$ as

$$(2.1.3)\qquad G(\boldsymbol{t})=\begin{cases} c(t_6-t_7^2/n_2)^{\nu_2}[(t_1-t_2^2/n_1)(t_3-t_5^2/n_1)-(t_4-t_2t_5/n_1)^2]^{\nu_1} & \\ \qquad\times\exp(-\alpha_1t_1-\alpha_3t_3-\alpha_6t_6)\,, & \text{for } \boldsymbol{t}\in C \\ 0\,, & \text{otherwise}\,, \end{cases}$$

where ν_1, ν_2, α_1, α_3, α_6 and c are constants to be suitably selected so that $G(\boldsymbol{t})$ possesses the desirable properties ((2.1.2) among others). The motivation behind the choice of $G(\boldsymbol{t})$ in the form (2.1.3) is that for α_1, α_3, $\alpha_6>0$, $G(\boldsymbol{t})$ will be bounded for all $\boldsymbol{t}\in C$; for large ν_1, $\nu_2>0$, $G(\boldsymbol{t})$ will be smooth inside C and also over the boundary of C and finally for sufficiently small c, $G(\boldsymbol{t})$ will satisfy (2.1.2).

Working with P_1, we note that

$$(2.1.4)\quad \begin{aligned} DG(\boldsymbol{t})=c[&2t_7\nu_2\{\alpha_1\Delta_1^{\nu_1}\Delta_2^{\nu_2-1}-\nu_1(t_3-t_5^2/n_1)\Delta_1^{\nu_1-1}\Delta_2^{\nu_2-1}\}/n_2 \\ &-\nu_1\{2t_5(t_4-t_2t_5/n_1)/n_1-2t_2(t_3-t_5^2/n_1)/n_1\} \\ &\times\{\nu_2\Delta_1^{\nu_1-1}\Delta_2^{\nu_2-1}-\alpha_6\Delta_1^{\nu_1-1}\Delta_2^{\nu_2}\}]\exp(-\alpha_1t_1-\alpha_3t_3-\alpha_6t_6)\,, \end{aligned}$$

where we have written

$$\Delta_1=(t_1-t_2^2/n_1)(t_3-t_5^2/n_1)-(t_4-t_2t_5/n_1)^2\,,\qquad \Delta_2=t_6-t_7^2/n_2\,.$$

Hence, choosing $\nu_1=(n_1-1)/2$ and $\nu_2=(n_2-1)/2$ (note that the differentiability properties of $G(\boldsymbol{t})$ are satisfied in this case), we get

$$(2.1.5)\qquad \frac{DG(\boldsymbol{t})}{h(\boldsymbol{t})}=\begin{cases} c[2t_7\nu_2\{\alpha_1\Delta_1-\nu_1(t_3-t_5^2/n_1)\}/n_2+\nu_1\{2t_2(t_3-t_5^2/n_1)/n_1 & \\ \qquad -2t_5(t_4-t_2t_5/n_1)/n_1\}\{\nu_2-\alpha_6\Delta_2\}] & \\ \qquad\times\exp(-\alpha_1t_1-\alpha_3t_3-\alpha_6t_6)\,, & \text{for } \boldsymbol{t}\in C \\ 0\,, & \text{otherwise}\,. \end{cases}$$

Finally, for any $\alpha_1, \alpha_3, \alpha_6>0$, we choose c appropriately so that $DG(\boldsymbol{t})/h(t)$ satisfies (2.1.2). It is clear that a variety of similar regions can be constructed using the other relations P_2, P_3 and combinations of P_1, P_2 and P_3.

Thus, although we point out the feasibility of Wijsman's D-method to construct a similar region of non-Neyman structure, this cannot be recommended for practical use because of two reasons. One is the obvious criticism of its always being a randomized test. A second point is that such a test is likely to be biased. Our calculations with $\phi(\boldsymbol{t})=\alpha+DG(\boldsymbol{t})/h(\boldsymbol{t})$ where $DG(\boldsymbol{t})/h(\boldsymbol{t})$ is defined by (2.1.5) show that for $\rho=0$ and $\mu_1=0$ the local power of $\phi(\boldsymbol{t})$ is less than α at least in one direction.

2.2. *A locally optimum test when ρ is known*

In this subsection we derive the LMPI test for H_{01} when ρ is known.

Without any loss of generality we can state the problem in canonical form as follows:

Given $\boldsymbol{Y}^{2\times 1}\sim N(\boldsymbol{\mu}, c\Sigma)$, $X\sim N(\mu_1, c_1\sigma_1^2)$, $\boldsymbol{V}^{(n_1-1)\times 2}\sim N(0, I_{n_1-1}\otimes\Sigma)$, $\boldsymbol{V}_1^{((n_2-1)\times 1}\sim N(0, \sigma_1^2 I_{n_2-1})$, $\mu'=(\mu_1, \mu_2)$, where $\boldsymbol{Y}, X, \boldsymbol{V}$ and $\boldsymbol{V}_1$ are mutually independent, $\otimes$ stands for Kronecker product, Σ denotes the bivariate variance-covariance matrix and c, c_1 are constants, we want to test H_{01}: $\mu_2=0$ against H_{11}: $\mu_2\neq 0$. Clearly

$$(2.2.1)\qquad \begin{aligned} &\boldsymbol{Y}'=(\bar{X}_1, \bar{Y}_1)=\sum_{i=1}^{n_1}(X_i, Y_i)/n_1\ , \qquad X=\bar{X}_2=\sum_{i=n_1+1}^{n_1+n_2} X_i/n_2\ , \\ &c=1/n_1\ , \qquad c_1=1/n_2\ , \\ &\boldsymbol{V}'\boldsymbol{V}=\begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22}\end{pmatrix}=\sum_{i=1}^{n_1}\begin{bmatrix} X_i-\bar{X}_1 \\ Y_i-\bar{Y}_1\end{bmatrix}\begin{bmatrix} X_i-\bar{X}_1 \\ Y_i-\bar{Y}_1\end{bmatrix}' \\ &\boldsymbol{V}_1'\boldsymbol{V}_1=S_{11}^*=\sum_{i=n_1+1}^{n_1+n_2}(X_i-\bar{X}_2)^2\ . \end{aligned}$$

To derive an LMPI test, note that the problem under consideration remains invariant under the transformation

$$(2.2.2)\qquad \boldsymbol{Y}\rightarrow A\boldsymbol{Y}+\boldsymbol{a}\ , \quad X\rightarrow A_{11}X+a_1\ , \quad \boldsymbol{V}\rightarrow \boldsymbol{V}A'\ , \quad \boldsymbol{V}_1\rightarrow \boldsymbol{V}_1A_{11}\ ,$$

where $A=\begin{pmatrix} A_{11} & 0 \\ 0 & A_{22}\end{pmatrix}$ is a 2×2 non-singular matrix and $\boldsymbol{a}=\begin{pmatrix} a_1 \\ 0\end{pmatrix}$ with $-\infty<a_1<\infty$. If G denotes the associated group, it follows clearly that $\nu(dA, da_1)=da_1((dA_{11}/|A_{11}|)(dA_{22}/|A_{22}|))$ is a left-invariant measure on G. Moreover, the differential $d\boldsymbol{Y}\,dX\,d\boldsymbol{V}\,d\boldsymbol{V}_1$ is relatively invariant with multiplier

$$\chi(A, \boldsymbol{a})=|A_{11}|^{n_1+n_2-2}|A_{22}|^{n_1-1}\ .$$

Hence, from Andersson [1] (for details see Eaton and Kariya [3]) we get that the ratio of the density of a maximal invariant statistic T under the alternative to null is given by

$$(2.2.3)\qquad r_\rho(T)=\frac{\int_G f((A, \boldsymbol{a})Z/\eta, \rho)\chi(A, \boldsymbol{a})\nu(dA, da_1)}{\int_G f((A, \boldsymbol{a})Z|\mu_2=0, \rho)\chi(A, \boldsymbol{a})\nu(dA, da_1)}\ ,$$

where $Z=(\boldsymbol{Y}, X, \boldsymbol{V}, \boldsymbol{V}_1)$, $(A, \boldsymbol{a})Z=(A\boldsymbol{Y}+\boldsymbol{a}, A_{11}X+a_1, \boldsymbol{V}A', \boldsymbol{V}A_{11})$, $\eta^2=\mu_2^2/\sigma_2^2$ and $f(\cdot)$ stands for the joint density of $\boldsymbol{Y}, X, \boldsymbol{V}$ and $\boldsymbol{V}_1$ under the parametrization $\mu_1=0$, $\mu_2=\eta$, $\sigma_1=\sigma_2=1$.

Evaluating $r_\rho(T)$ and using Ferguson [4], the LMPI test statistic T_1 can be written as

$$(2.2.4)\qquad T_1=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left[u_2\frac{\bar{y}_1}{\sqrt{A_2}}\,\frac{c^{-1}(c^{-1}+c_1^{-1})}{c^{-1}+c_1^{-1}(1-\rho^2)}-u_1\frac{\bar{x}_1-\bar{x}_2}{\sqrt{A_1}}\,\frac{c^{-1}c_1^{-1}\rho}{c^{-1}+c_1^{-1}(1-\rho^2)}\right]^2$$

$$\times(1-\rho^2)\exp\left(-\frac{1}{2}(u_1^2+u_2^2-2\rho u_1u_2A_3/\sqrt{A_1A_2})\right)$$
$$\times|u_1|^{n_1+n_2-3}|u_2|^{n_1-2}du_1du_2$$
$$\div\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2}(u_1^2+u_2^2-2\rho u_1u_2A_3/\sqrt{A_1A_2})\right)$$
$$\times|u_1|^{n_1+n_2-3}|u_2|^{n_1-2}du_1du_2\,,$$

where

$$\begin{aligned}&A_1=S_{11}+S_{11}^*(1-\rho^2)+c^{-1}c_1^{-1}(1-\rho^2)(\bar{x}_1-\bar{x}_2)^2/(c^{-1}+c_1^{-1}(1-\rho^2))\\ (2.2.5)\quad &A_2=S_{22}+c^{-1}(c^{-1}+c_1^{-1})(1-\rho^2)\bar{y}_1^2/(c^{-1}+c_1^{-1}(1-\rho^2))\\ &A_3=S_{12}+c^{-1}c_1^{-1}(1-\rho^2)\bar{y}_1(\bar{x}_1-\bar{x}_2)/(c^{-1}+c_1^{-1}(1-\rho^2))\,.\end{aligned}$$

Although the above integrals can be evaluated (Krishnaiah, Hagis and Steinberg [6]), the resulting expressions involve infinite sums unless both n_1+n_2-3 and n_1-2 are even integers. We consider below this special case, assuming $n_1-2=2r$ and $n_1+n_2-3=2s$. Writing $\bar{\rho}=\rho A_3/\sqrt{A_1A_2}$, it follows that T_1 has the expression

$$\begin{aligned}(2.2.6)\quad T_1=&\left\{\frac{(\bar{x}_1-\bar{x}_2)^2}{A_1}\left(\frac{c^{-1}c_1^{-1}\rho}{c^{-1}+c_1^{-1}(1-\rho^2)}\right)^2(1-\rho^2)\left[\sum_{j=0}^{r}\binom{2r}{j}\bar{\rho}^{2j}\frac{\Gamma(r-j+1/2)}{(1/2)^{r-j+1/2}}\right.\right.\\ &\left.\times\frac{\Gamma(s+j+1+1/2)}{((1-\bar{\rho}^2)/2)^{s+j+1+1/2}}\right]+\frac{\bar{y}_1^2}{A_2}\left(\frac{c^{-1}(c^{-1}+c_1^{-1})}{c^{-1}+c_1^{-1}(1-\rho^2)}\right)^2(1-\rho^2)\\ &\times\left[\sum_{j=0}^{r+1}\binom{2r+2}{2j}\bar{\rho}^{2j}\frac{\Gamma(r+1-j+1/2)}{(1/2)^{r-j+1+1/2}}\;\frac{\Gamma(s+j+1/2)}{((1-\bar{\rho}^2)/2)^{s+j+1/2}}\right]\\ &-2\rho\frac{\bar{y}_1(\bar{x}_1-\bar{x}_2)}{\sqrt{A_1A_2}}\;\frac{c_1^{-1}(c^{-1}+c_1^{-1})c^{-2}}{\{c^{-1}+c_1^{-1}(1-\rho^2)\}^2}(1-\rho^2)\left[\sum_{j=0}^{r}\binom{2r+1}{2r-2j}\right.\\ &\left.\left.\times\frac{\Gamma(r-j+1/2)}{(1/2)^{r-j+1/2}}\bar{\rho}^{2j+1}\frac{\Gamma(s+j+1+1/2)}{((1-\bar{\rho}^2)/2)^{s+j+3/2}}\right]\right\}\\ &\div\left[\sum_{j=0}^{r}\binom{2r}{2j}\bar{\rho}^{2j}\frac{\Gamma(r-j+1/2)}{(1/2)^{r-j+1/2}}\;\frac{\Gamma(s+j+1/2)}{((1-\bar{\rho}^2)/2)^{s+j+1/2}}\right]\,.\end{aligned}$$

The difficulty and hence the limitation in the use of the above test statistic, though locally optimum, is evident. We provide below upper $100\alpha\%$ points of the null distribution of T_1 for $\alpha=.1, .05, .025, .01$, and $(n_1, n_2)=(10, 10), (10, 15), (10, 20), (10, 25), (15, 15), (15, 20)$, $\rho=0.1(.1)0.9$. This table will help the practising statisticians to carry out the above test. To construct the table, 100 values of the statistic T_1 have been generated under the model $N(0, 0, 1, 1, \rho)$ (because of invariance) and the upper $100\alpha\%$ points have been recorded.

Table 2.1. Percentage points of T_1 for some values of n_1, n_2, ρ and α

$\rho \backslash \alpha$	.1	.05	.025	.010	$\rho \backslash \alpha$	.1	.05	.025	.010
	$n_1=10$, $n_2=10$					$n_1=10$, $n_2=15$			
.1	1.02700	1.93600	4.01174	7.90551	.1	0.91235	1.95222	3.54507	8.88342
.2	1.66065	3.42223	7.53221	19.96181	.2	1.77873	3.36245	7.22839	32.42370
.3	1.89769	3.18038	6.90620	15.67333	.3	1.78275	3.78613	7.63553	18.15987
.4	1.87179	3.38779	5.38235	15.52086	.4	1.93024	3.83829	7.20407	13.92252
.5	1.96051	3.24655	5.58057	10.55522	.5	1.71697	2.71018	6.04062	24.43401
.6	1.62903	2.60713	3.83758	7.66135	.6	1.80934	2.74452	3.79504	8.91943
.7	1.76184	2.43659	3.10068	7.59353	.7	1.55137	2.08636	2.89156	4.89674
.8	1.56079	2.12274	2.71521	3.97028	.8	1.42361	1.93422	2.27513	2.91562
.9	1.35136	1.79262	2.16924	2.74861	.9	1.31436	1.72435	2.04381	2.72910
	$n_1=10$, $n_2=20$					$n_1=10$, $n_2=25$			
.1	0.92333	1.56364	2.57478	5.58530	.1	0.93528	1.59186	2.63251	4.91956
.2	1.50072	3.09225	5.59342	22.79338	.2	1.48509	2.94104	4.22787	10.98097
.3	1.97239	3.74365	6.94326	29.35041	.3	1.94747	3.93975	10.44636	28.39296
.4	2.26504	4.39513	11.66648	30.50430	.4	1.98073	3.97174	7.45803	18.58330
.5	1.94876	3.37522	5.64184	11.60773	.5	1.98823	3.43432	4.99242	13.01333
.6	1.88433	2.71943	4.57048	8.02634	.6	1.68342	2.39340	3.54960	6.16087
.7	1.48671	2.03268	2.81866	3.48648	.7	1.46880	2.02406	2.73085	4.20848
.8	1.53299	2.04144	2.50352	2.99973	.8	1.46617	1.92087	2.32321	2.89250
.9	1.39636	1.92005	2.27580	2.93128	.9	1.41725	1.89358	2.22982	2.73891
	$n_1=15$, $n_2=15$					$n_1=15$, $n_2=20$			
.1	1.13451	2.41008	5.28146	13.78828	.1	1.43010	2.48732	4.83984	14.53893
.2	1.87126	3.64782	6.99666	20.00294	.2	1.84581	3.79643	7.47429	17.29544
.3	2.03030	3.80588	8.39736	22.38998	.3	2.15803	4.94392	10.18466	59.54926
.4	1.81273	2.91778	5.29046	13.05786	.4	1.86201	3.20983	5.99182	11.88610
.5	1.79297	2.92426	5.02641	9.03013	.5	1.81686	2.60587	3.74264	7.81223
.6	1.60150	2.20349	2.51541	3.79257	.6	1.54500	2.00971	2.72556	3.49335
.7	1.46881	1.97798	2.38057	3.10399	.7	1.49422	1.98493	2.72988	3.79701
.8	1.44568	1.88822	2.40460	2.98079	.8	1.38425	1.71862	2.12434	2.51977
.9	1.34815	1.77130	2.15666	2.45741	.9	1.27930	1.72909	2.06311	2.54026

3. Testing H_{02} in case (a)

Here our purpose is to propose a test for H_{02} which is exact and similar. For this we introduce Scheffé-type variables (Scheffé [14]) as in Şarkar [13]. Assuming $n_1 \leqq n_2$, we define

$$X_i^* = \lambda X_i + (1-\lambda) \sum_{j=1}^{n_2} c_{ij} X_{n_1+j}\,, \qquad i=1, 2, \cdots, n_1\,, \tag{3.1}$$

for some constant λ in such a way that (X_i^*, Y_i), $i=1, 2, \cdots, n_1$ are independently and identically distributed (i.i.d.) as bivariate normal with the mean vector (μ_1, μ_2) and with a minimum possible dispersion matrix. The purpose is achieved by taking $C=((c_{ij}))$: $n_1\times n_2$ a solution of $C1=1$ and $CC'=(n_1/n_2)I_{n_1}$ (see Sarkar [13]). The constant λ is chosen from a rough guess about the true regression of Y on X.

Now, (X_i^*, Y_i), $i=1, 2,\cdots, n_1$ are i.i.d. as $N_2(\mu_1, \mu_2, (\lambda^2+n_1(1-\lambda)^2/n_2)\cdot\sigma_1^2, \sigma_2^2, \rho\lambda(\lambda^2+n_1(1-\lambda)^2/n_2)^{-1/2}$. The test we propose for H_{02} is the usual paired t-test under this reduced set-up. Thus, when the alternative is H_1: $\mu_1\neq\mu_2$, we reject H_0 at the level of significance α if

$$(3.2)\qquad T_3=n_1(n_1-1)\bar{u}^2\Big/\sum_{i=1}^{n_1}(u_i-\bar{u})^2>F_{1,n_1-1;\alpha}\ ,$$

where $u_i=x_i^*-y_i$, $i=1, 2,\cdots, n_1$.

We see that under any alternative the distribution of T_3 is non-central F_{1,n_1-1} with the non-centrality parameter $\delta=n_1(\mu_1-\mu_2)^2/\{[\lambda^2+n_1(1-\lambda)^2/n_2]\sigma_1^2+\sigma_2^2-2\lambda\rho\sigma_1\sigma_2\}^{-1}$. The power function of this test is known to be monotonically increasing with δ. Hence, at a specified alternative the optimum choice of λ for which the denominator in δ is minimum is $\lambda_0=(n_1+n_2\beta)/(n_1+n_2)$, where $\beta=\rho\sigma_2/\sigma_1$. In practice, however, β is unknown. So λ_0 cannot be used. Comparing δ with the non-centrality parameter of the paired t-test which ignores the additional n_2 observations on X, we find that ours is better than the latter if

$$[\lambda^2+n_1(1-\lambda)^2/n_2]\sigma_1^2-2\lambda\sigma_1\sigma_2<\sigma_1^2-2\rho\sigma_1\sigma_2\ ,$$

whatever be $(\sigma_1^2, \sigma_2^2, \rho)$. The inequality can be expressed as follows:

$$1<\lambda<\{2\beta-(1-n_1/n_2)\}/(1+n_1/n_2)\qquad \text{if } \beta>1$$

and

$$\{2\beta-(1-n_1/n_2)\}/(1+n_1/n_2)<\lambda<1\qquad \text{if } \beta<1\,.$$

Hence, a rough knowledge about β in terms of its bounds will enable us to choose λ appropriately. Thus if ρ is known to be $\leqq 0$, we can choose any $\lambda\in[0, 1)$.

4. Testing H_{03} in case (a)

The locally most powerful invariant (LMPI) test for such a problem in the multivariate set-up was derived by Eaton and Kariya [3]. From their result, we see that the LMPI test for our problem is based on the statistic

$$(4.1)\qquad r_1^2=S_{12}^2/\tilde{S}_{11}S_{22}-S_{11}/(n_1-1)\tilde{S}_{11}$$

where $\tilde{S}_{11}=S_{11}+S_{11}^*+n_1n_2(\bar{X}_1-\bar{X}_2)^2/(n_1+n_2)$. For the sake of convenience, however, we may omit the second term in (4.1) which is small (in probability) for large sample sizes.

Another test using the predicted Y-values (missing) is obtained as follows. This is based on the square r_2^2 of the correlation coefficient r_2 obtained from the paired values

$$(X_i, Y_i)\,, \quad i=1, 2, \cdots, n_1\,, \qquad (X_i, Y_i^*)\,, \quad i=n_1+1, \cdots, n_1+n_2\,,$$

with $Y_i^*=\bar{Y}_1+b_{21}(X_i-\bar{X}_1)$ and $b_{21}=S_{12}/S_{11}$ is the sample regression coefficient of Y on X. It is easy to see that

$$r_2^2=b_{21}^2\tilde{S}_{11}/(S_{22.1}+b_{21}^2\tilde{S}_{11})\,, \tag{4.2}$$

where $S_{22.1}=S_{22}-b_{21}^2S_{11}$.

It is interesting to compare the test based on r_2^2 with those based on $r_1^{*2}=S_{12}^2/\tilde{S}_{11}S_{22}$, the approximate version of r_1^2, and r_3^2, the square of the correlation coefficient obtained by ignoring the additional observations. The distributions of r_1^{*2}, r_2^2 and r_3^2 under any alternative ρ^2 are the following.

Density of r_1^{*2}:

$$\sum_{i=0}^{\infty}\frac{\lambda^i}{i!}\sum_{j=0}^{i}(-1)^j\binom{i}{j}\frac{\Gamma((n_1-1)/2+i-j)\Gamma((n-1)/2+i)}{\Gamma((n_1-1)/2)\Gamma(1/2+i-j)\Gamma((n-2)/2)} \\ \times t^{1/2+i-j-1}(1-t)^{(n-2)/2-1}{}_2F_1(-j, n_2/2; (n_1-2)/2; 1-t)\,, \tag{4.3}$$

where $\lambda=\rho^2/(1-\rho^2)$, $n=n_1+n_2$, and

$${}_2F_1(a, b; c; x)=\sum_{j=0}^{\infty}\frac{\Gamma(a+j)\Gamma(b+j)\Gamma(c)}{\Gamma(a)\Gamma(b)\Gamma(c+j)}\frac{x^j}{j!}\,.$$

Density of r_2^2:

(4.4)

$$\sum_{i=0}^{\infty}\frac{\lambda^i}{i!}\sum_{j=0}^{i}(-1)^j\binom{i}{j} \\ \times\frac{\Gamma((n_1-1)/2+i)\Gamma((n_1-1)/2+i-j)\Gamma((n-1)/2+i)\Gamma(n_1/2+2i-j)}{\Gamma((n_1-1)/2)\Gamma(1/2+i-j)\Gamma((n_1-1)/2+i)\Gamma((n_1-2)/2)\Gamma(n/2+2i-j)} \\ \times t^{1/2+i-j+1}(1-t)^{(n_1-2)/2-1}{}_2F_1((n_1-1)/2+i-j, n_2/2; n/2+2i-j; t)\,.$$

Density of r_3^2:

$$\sum_{i=0}^{\infty}\frac{\lambda^i}{i!}\sum_{j=0}^{i}(-1)^j\binom{i}{j}\frac{1}{B(1/2+i-j, (n_1-2)/2)}t^{1/2+i-j-1}(1-t)^{(n_1-2)/2-1}\,. \tag{4.5}$$

From the above expressions of the densities, we can easily find out the slopes of the power functions of the above three tests at $\lambda=0$ (i.e., $\rho=0$). We also make the comparison of the power functions at some alternatives through simulation. The details appear in Tables 3 (a), (b).

Table 3(a). Slope of tests r_1^{*2}, r_2^2 r_3^2

	$\alpha=.05$			$\alpha=.025$			$\alpha=.01$		
	$n_1=10$	$n_1=15$	$n_1=20$	$n_1=10$	$n_1=15$	$n_1=20$	$n_1=10$	$n_1=15$	$n_1=20$
	$n_2=10$	$n_2=15$	$n_2=20$	$n_2=10$	$n_2=15$	$n_2=20$	$n_2=10$	$n_2=15$	$n_2=20$
r_1^{*2}	.6762	1.1221	1.5689	.4429	.7408	1.0397	.2293	.3885	.5487
r_2^2	.6329	1.1733	1.7309	.3469	.6734	1.0170	.1480	.3044	.4740
r_3^2	.8099	1.3825	1.9955	.4783	.8376	1.1989	.2241	.4056	.5896

Table 3(b). Simulated powers of tests r_1^{*2}, r_2^2, r_3^2 ($n_1=10$, $n_2=30$)

ρ	$\alpha=.05$			$\alpha=.025$			$\alpha=.010$		
	r_1^{*2}	r_2^2	r_3^2	r_1^{*2}	r_2^2	r_3^2	r_1^{*2}	r_2^2	r_3^2
0.0500	0.0517	0.522	0.0516	0.0253	0.0256	0.0257	0.0098	0.0116	0.0117
0.1000	0.0558	0.0590	0.0584	0.0296	0.0304	0.0293	0.0110	0.0136	0.0143
0.2000	0.0758	0.0868	0.0876	0.0391	0.0485	0.0477	0.0145	0.0244	0.0212
0.3000	0.1152	0.1314	0.1331	0.0590	0.0802	0.0807	0.0227	0.0443	0.0400
0.3500	0.1422	0.1649	0.1714	0.0769	0.1008	0.1025	0.0291	0.0564	0.0542
0.4000	0.1753	0.2008	0.2161	0.0992	0.1273	0.1347	0.0384	0.0725	0.0743
0.4500	0.2182	0.2459	0.2700	0.1266	0.1607	0.1764	0.0512	0.0925	0.0983
0.5000	0.2712	0.2964	0.3306	0.1639	0.1988	0.2267	0.0705	0.1180	9.1334

5. Testing H_{01}, H_{02} and H_{03} in case (b)

5.1. *Testing* H_{01}

For this testing problem, we propose a similar test. We assume that $n_1 \leqq n_2$. Then, using a method as in Sarkar [13] we can derive a similar test from the paired data and the additional data on X. Another similar test (Student's t-test) can be provided by using only the additional data on Y. These two similar tests are independent. Hence, by combining them suitably, we have a similar test utilizing all the available observations.

5.2. *Testing* H_{02}

We use the method as in Sarkar [13] to provide an exact similar test for this problem. We assume that $n_1 \leqq n_2$, $n_1 \leqq n_3$. Define new variables

$$
\begin{aligned}
X_i^* &= \lambda_1 X_i - (1-\lambda_1) \sum_{j=1}^{n_2} c_{ij}^{(1)} X_{n_1+j} \\
Y_i^* &= \lambda_2 Y_i - (1-\lambda_2) \sum_{k=1}^{n_3} c_{ik}^{(2)} Y_{n_1+n_2+k} ,
\end{aligned}
\tag{5.2.1}
$$

where the matrices $C_1=((c_{ij}^{(1)}))$ and $C_2=((c_{ik}^{(2)}))$ satisfy $C_1\mathbf{1}_{n_2}=\mathbf{1}_{n_1}$, $C_1C_1'=$

$(n_1/n_2)I_{n_1}$, $C_2\mathbf{1}_{n_3}=\mathbf{1}_{n_1}$, $C_2C_2'=(n_1/n_3)I_{n_1}$ and the constants λ_1 and λ_2 are to be specified later. Clearly, (X_i^*, Y_i^*), $i=1, 2, \cdots, n_1$ are i.i.d. as $N_2(\mu_1, \mu_2; [\lambda_1^2+(1-\lambda_1)^2 n_1/n_2]\sigma_1^2, [\lambda_2^2+(1-\lambda_2)^2 n_1/n_3]\sigma_2^2, \rho\lambda_1\lambda_2[\lambda_1^2+(1-\lambda_1)^2 n_1/n_2]^{-1/2}\cdot[\lambda_2^2+(1-\lambda_2)^2 n_1/n_3]^{-1/2})$. Hence, for testing $H_{02}: \mu_1=\mu_2$ against $H_{12}: \mu_1\neq\mu_2$, we propose the statistic

$$T_4=n_1(n_1-1)\bar{u}^2/S_u^2\ ; \tag{5.2.2}$$

where $u_i=X_i^*-Y_i^*$, $\bar{u}=\frac{1}{n_1}\sum_{i=1}^{n_1} u_i$ and $S_u^2=\sum_{i=1}^{n_1}(u_i-\bar{u})^2$. Under H_{02}, $T_4\sim F_{1,n_1-1}$, and under any alternative $T_4\sim F_{1,n_1-1}(\delta_1)$, where $\delta_1=n_1(\mu_1-\mu_2)^2\cdot\{[\lambda_1^2+(1-\lambda_1)^2 n_1/n_2]\sigma_1^2+[\lambda_2^2+(1-\lambda_2)^2 n_1/n_3]\sigma_2^2-2\lambda_1\lambda_2\rho\sigma_1\sigma_2\}^{-1}$. Comparing δ_1 with $\delta_2=n_1(\mu_1-\mu_2)^2/(\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2)$, the non-centrality parameter of the optimum test which ignores the additional data on both X and Y, we find that ours is better whenever

$$[1-\lambda_1^2-(1-\lambda_1)^2 n_1/n_2]+[1-\lambda_2^2-(1-\lambda_2)^2 n_1/n_3]\sigma_2^2/\sigma_1^2>2(1-\lambda_1\lambda_2)\beta\ . \tag{5.2.3}$$

So, a rough knowledge about the ratio of the variances and the regression coefficient can help us in the selection of appropriate λ_1 and λ_2. Thus, for example, if ρ is known to be negative, any λ_1, λ_2 such that $0\leqq\lambda_1, \lambda_2<1$ will be appropriate.

5.3. *Testing H_{03}*

As mentioned in the introduction, we provide a direct proof of a result of Eaton and Kariya [3]. Note that the problem of testing H_{03}: $\rho=0$ against $H_{13}: \rho\neq0$ remains invariant under the group of transformations

$$\begin{aligned} &X_i\to a+bX_i\ , \qquad X_{n_1+j}\to a+bX_{n_1+j}\ , \\ &Y_i\to c+dY_i\ , \qquad Y_{n_1+n_2+k}\to c+dY_{n_1+n_2+k}\ , \\ &\quad i=1, 2, \cdots, n_1,\ j=1, 2, \cdots, n_2,\ k=1, 2, \cdots, n_3, \\ &\quad -\infty<a, b, c, d<\infty,\ b\cdot d\neq0\ . \end{aligned} \tag{5.3.1}$$

The underlying minimal sufficient statistic is $(\bar{X}_1, \bar{Y}_1, S_{11}^{(1)}, S_{12}, S_{22}^{(1)}, \bar{X}_2, S_{11}^{(2)}, \bar{Y}_2, S_{22}^{(2)})$, where $\bar{X}_1$, $\bar{Y}_1$, S_{12}, and $\bar{X}_2$ were defined in (2.2.1), $S_{11}^{(1)}\equiv S_{11}$, $S_{11}^{(2)}\equiv S_{11}^*$, $S_{22}^{(1)}\equiv S_{22}$, $\bar{Y}_2=\frac{1}{n_3}\sum_{k=1}^{n_3} Y_{n_1+n_2+k}$ and $S_{22}^{(2)}=\sum_{k=1}^{n_3}(Y_{n_1+n_2+k}-\bar{Y}_2)^2$. Then, we have the following proposition whose proof is omitted.

PROPOSITION. A maximal invariant in the space of the minimal sufficient statistic under the group of transformations induced by (5.3.1) is

$$(t_1^2, t_2^2, t_3^2, u, v, w(t_1), w(t_2), w(t_3))\ , \tag{5.3.2}$$

where

$$t_1=\sqrt{n_1n_2/(n_1+n_2)}(\bar{X}_1-\bar{X}_2)/\sqrt{S_{11}^{(1)}}\,, \qquad t_2=\sqrt{n_1n_3/(n_1+n_3)}(\bar{Y}_1-\bar{Y}_2)/\sqrt{S_{22}^{(1)}}\,,$$

$$t_3=S_{12}/\sqrt{S_{11}S_{22}}\,, \qquad u=S_{11}^{(2)}/S_{11}^{(1)}\,, \qquad v=S_{22}^{(2)}/S_{22}^{(1)}\,,$$

and, for any t, $w(t)=+1$ or -1 according as $t>0$ or $\leqq 0$.

The density $f_\theta(\cdot)$ of $(t_1^2, t_2^2, t_3^2, u, v, w(t_1), w(t_2), w(t_3))$ at (x, y, z, u, v, w) is

$$\begin{aligned}
(5.3.3) \qquad f_\theta(\cdot)=&\text{const.}\ (xyz)^{-1/2}(1-z)^{(n_1-4)/2}u^{(n_1-1)/2}v^{(n_2-1)/2}\\
&\times\int_0^\infty\int_0^\infty e^{-1/2}\left\{\left(\frac{1}{1-\theta}+\frac{x}{1-c^2\theta}+u\right)S_{11}^{(1)}\right.\\
&\left.-\frac{1}{2}\left(\frac{1}{1-\theta}+\frac{y}{1-c^2\theta}+v\right)S_{22}^{(1)}\right\}\\
&\times\sum_{j=0}^{\infty}\frac{\theta^j}{(2j)!}\left(\frac{\sqrt{z}}{1-\theta}+\frac{cw\sqrt{x}\sqrt{y}}{1-c^2\theta}\right)^{2j}S_{11}^{(1)\,(n_1+n_2+2j-3)/2}\\
&\times S_{22}^{(1)\,(n_1+n_3+2j-3)/2}dS_{11}^{(1)}dS_{22}^{(1)}
\end{aligned}$$

where $\theta=\rho^2$, $c=\sqrt{n_2n_3/(n_1+n_2)(n_1+n_3)}$. From the above density, the LMPI test is obtained as that based on the statistic

$$\begin{aligned}
(5.3.4) \qquad &\left[S_{12}+\frac{n_1n_2n_3}{(n_1+n_2)(n_1+n_3)}(\bar{X}_1-\bar{X}_2)(\bar{Y}_1-\bar{Y}_2)\right]^2\Big/S_{11}^*S_{22}^*\\
&-\left\{S_{11}^{(1)}+\frac{n_1n_2^2n_3(\bar{X}_1-\bar{X}_2)^2}{(n_1+n_2)^2(n_1+n_3)}\right\}\Big/(n_1+n_3-1)S_{11}^*\\
&-\left\{S_{22}^{(1)}+\frac{n_1n_2n_3^2(\bar{Y}_1-\bar{Y}_2)^2}{(n_1+n_2)(n_1+n_3)^2}\right\}\Big/(n_1+n_2-1)S_{22}^*
\end{aligned}$$

where $S_{11}^*=S_{11}^{(1)}+S_{11}^{(2)}+n_1n_2(\bar{X}_1-\bar{X}_2)^2/(n_1+n_2)$, $S_{22}^*=S_{22}^{(1)}+S_{22}^{(2)}+n_1n_3(\bar{Y}_1-\bar{Y}_2)^2/(n_1+n_3)$.

The distribution of the statistic in (5.3.4) is extremely hard to obtain in small samples. In large samples, when $n_1/n_2\rightarrow\eta_1$ and $n_1/n_3\rightarrow\eta_2$, for some $0\leqq\eta_1$, $\eta_2<\infty$, an approximate version of the LMPI test is based on the statistic $r^{*2}=S_{12}^2/S_{11}S_{22}$. The density of r^{*2}, under H_{03}, is obtained as the following:

(5.3.5)

$$\begin{aligned}
&\frac{\Gamma((n_1+n_2-1)/2)\Gamma((n_1+n_3-1)/2)}{\Gamma(1/2)\Gamma((n_1-1)/2)\Gamma((n_1+n_2+n_3-2)/2)}(r^{*2})^{-1/2}(1-r^{*2})^{(n_1+n_2+n_3-4)/2}\\
&\qquad\times{}_2F_1(n_2/2, n_3/2; (n_1+n_2+n_3-2)/2; 1-r^{*2})\ .
\end{aligned}$$

Acknowledgement

The authors wish to thank Hina Vaidya for her valuable help in

computing the tables.

UNIVERSITY OF PITTSBURGH

REFERENCES

[1] Andersson, S. A. (1978). Invariant measures, *Tech. Report* No. 129, Stanford University.

[2] Bhoj, D. S. (1978). Testing equality of means of correlated variates with missing observations on both responses, *Biometrika*, **65**, 225-228.

[3] Eaton, M. L. and Kariya, T. (1980). Testing problems with additional or missing data, *Tech. Report* No. 370, University of Minnessota.

[4] Ferguson, T. S. (1967). *Mathematical Statistics: Decision Theoretic Approach*, Academic Press, Inc., New York.

[5] Khatri, C. G., Bhargava, R. P. and Shah, K. R. (1974). Distribution of regression estimate in double sampling, *Sankhyā*, C, **36**, 3-22.

[6] Krishnaiah, P. R., Hagis, P., Jr. and Steinberg, L. (1963). A note on the bivariate chi distribution, *SIAM Rev.*, **5**, 140-144.

[7] Lin, P. E. (1973). Procedures for testing the differences of means with incomplete data, *J. Amer. Statist. Ass.*, **68**, 699-703.

[8] Little, R. J. A. (1976). Inference about means from incomplete multivariate data, *Biometrika*, **63**, 593-604.

[9] Mehta, J. S. and Gurland, J. (1969). Testing equality of means in the presence of correlation, *Biometrika*, **56**, 119-126.

[10] Mehta, J. S. and Gurland, J. (1973). A test of equality of means in the presence of correlation and missing values, *Biometrika*, **60**, 211-213.

[11] Morrison, D. F. (1973). A test of equality of means of correlated variates with missing data on one response, *Biometrika*, **60**, 101-105.

[12] Naik, U. D. (1975). On testing equality of means of correlated variables with incomplete data, *Biometrika*, **62**, 615-622.

[13] Sarkar, S. K. (1979). A test for mean with additional observations, *Calcutta Statist. Ass. Bull.*, **28**, 47-56.

[14] Scheffé, H. (1943). On solutions of the Behrens-Fisher problem based on the t-distribution, *Ann. Math. Statist.*, **14**, 35-44.

[15] Tamhane, A. C. (1978). Inference based on regression estimator in double sampling, *Biometrika*, **65**, 419-427.

[16] Wijsman, R. A. (1958). Incomplete sufficient statistics and similar tests, *Ann. Math. Statist.*, **29**, 1028-1045.

[17] Wijsman, R. A. (1967). Cross-sections of orbits and their applications, *Proc. Fifth Berkeley Symp. Math. Statist. Prob.*, **1**, 389-400.