Journal of Statistical Computation and Simulation

ISSN: 0094-9655 (Print) 1563-5163 (Online) Journal homepage: http://www.tandfonline.com/loi/gscs20

On difference of means of correlated variates with incomplete data on both responses

Dinesh S. Bhoj

J. Statist. Comput. Simul., 1984, Vol. 19, pp. 275–289
0094-9655/84/1904-0275 $18.50/0

On Difference of Means of Correlated Variates with Incomplete Data on Both Responses

DINESH S. BHOJ
Rutgers University, Camden, NJ 08102, U.S.A.

(Received August 27, 1981)

A class of estimators is proposed for the difference of means when sampling from a bivariate normal distribution with means μ_1 and μ_2, variances σ_1^2 and σ_2^2, and correlation coefficient ρ, where some observations on either of the variables are missing. The new estimator from this class is used to obtain some test statistics for testing the equality of means. Some additional test statistics are also proposed and the empirical powers of all statistics are computed for different values of σ_1^2, σ_2^2 and ρ. These computations support the use of some of the new test statistics.

KEY WORDS: Combination of independent tests, empirical size and power, equality of means, Fisher's method, incomplete data, linear combination of independent variates, paired t test, unbiased estimator.

1. INTRODUCTION

Suppose that a sample of n independent pairs of observations $(x_1, y_1), \ldots, (x_n, y_n)$ has been drawn from the bivariate normal distribution with mean vector $\mu = (\mu_1, \mu_2)'$ and covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}.$$

In addition, n_1 observations $x_{n+1}, \ldots, x_{n+n_1}$ on x alone, and n_2 observations $y_{n+1}, \ldots, y_{n+n_2}$ on y alone are available. It is assumed that (x_i, y_i), x_{n+j}, y_{n+k} are mutually independent for $i=1, \ldots, n$, $j=1, \ldots, n_1$ and $k=1, \ldots, n_2$. We wish to use all of these data to draw inferences about the difference of means $\delta=\mu_1-\mu_2$. In the case of incomplete data on both responses there are no explicit expressions for the maximum likelihood estimators of the parameters. In the past, two different approaches have been commonly used in estimating these parameters. The first approach is to obtain the maximum likelihood estimates of μ and Σ implicitly (Rao, 1952, pp. 161–3) or to devise a simple iterative procedure for finding these estimates (Orchard and Woodbury, 1970). These estimates are not amenable to construct noniterative test statistics for testing hypotheses. The second approach is to get noniterative estimates of δ which can be used for testing hypothesis $\delta=0$. Lin and Stivers (1974) obtained the maximum likelihood estimate of δ under the assumption of known Σ and then replaced the elements of Σ by their maximum likelihood estimates based on complete n pairs of observations.

In constructing test statistics, three different procedures have been generally used. The first procedure, used by Lin and Stivers (1974) and Ekbohm (1976), is to start with a noniterative estimator, find an estimator of its variance and then form a statistic by taking a ratio of these quantities. The second procedure, devised by Woolson and Cole (1974), is to use the theory of the general linear hypothesis to derive the statistic under the assumption of equal variances and known ρ. This statistic was modified by Woolson, Leeper, Cole and Clarke (1976), by replacing ρ by its maximum likelihood estimator based on complete data. The third procedure is to obtain a statistic by combining two independent statistics which are based on complete and incomplete data, respectively. Bhoj (1978) proposed the test statistic which is a linear combination of the paired and unpaired t statistics. He used the expected squared lengths of the confidence intervals for δ to measure the increased precision of his statistic relative to the conventional paired t statistic. However, the performance of this statistic in terms of power remains to be investigated and compared with the known statistics.

In this paper we propose a noniterative estimate of δ and use it to derive two new test statistics for testing the hypothesis $\delta=0$. One more statistic, which is based on a combination of independent statistics, is proposed, and the statistic due to Bhoj (1978) is modified. The relative merits of the proposed and known test statistics are determined by Monte Carlo study.

The following notation is used in this paper:

$$n\bar{x}_1=\sum_{i=1}^{n} x_i,\quad n\bar{y}_1=\sum_{i=1}^{n} y_i,\quad n_1\bar{x}_2=\sum_{j=1}^{n_1} x_{n+j},\quad n_2\bar{y}_2=\sum_{k=1}^{n_2} y_{n+k}$$

$$(n+n_1)\bar{x}=\sum_{i=1}^{n+n_1} x_i,\quad (n+n_2)\bar{y}=\sum_{i=1}^{n+n_2} y_i.\quad a_{11}=\sum_{i=1}^{n}(x_i-\bar{x}_1)^2$$

$$a_{22}=\sum_{i=1}^{n}(y_i-\bar{y}_1)^2,\quad a_{12}=\sum_{i=1}^{n}(x_i-\bar{x}_1)(y_i-\bar{y}_1),\quad b_1=\sum_{j=1}^{n_1}(x_{n+j}-\bar{x}_2)^2,$$

$$b_2=\sum_{k=1}^{n_2}(y_{n+k}-\bar{y}_2)^2,\quad c_1=\sum_{i=1}^{n+n_1}(x_i-\bar{x})^2,\quad c_2=\sum_{i=1}^{n+n_2}(y_i-\bar{y})^2,$$

$$r=a_{12}/\sqrt{(a_{11}\, a_{22})},\qquad u=2a_{12}/(a_{11}+a_{22}).$$

2. ESTIMATION OF δ

We propose the following estimator of δ:

$$\hat{\delta}=\lambda_1\bar{x}_1-\lambda_2\bar{y}_1+(1-\lambda_1)\bar{x}_2-(1-\lambda_2)\bar{y}_2, \tag{2.1}$$

where λ_1 and λ_2 are uncorrelated with $\bar{x}_p$ and $\bar{y}_p$ $(p=1,2)$, respectively. $\hat{\delta}$ forms a class of unbiased estimators of δ. In this paper it is shown that some of the known estimators of δ belong to this class. Some new estimators from this class are considered and are used to construct new statistics for testing the hypotheses $\delta=0$. First we determine the values of λ_p so that the variance of $\hat{\delta}$ is minimized. Those values are given by

$$\begin{aligned}\lambda_1&=n(n+n_2+n_1\rho R)\{(n+n_1)(n+n_2)-n_1n_2\rho^2\}^{-1}\\ \lambda_2&=n(n+n_1+n_2\rho/R)\{(n+n_1)(n+n_2)-n_1n_2\rho^2\}^{-1},\end{aligned} \tag{2.2}$$

where $R=\sigma_2/\sigma_1$. It is interesting to note that these values of λ_p make $\hat{\delta}$ as the maximum likelihood estimator of δ when Σ is known (Lin and Stivers, 1974). For the case of unknown Σ they modified this estimator by estimating the elements of Σ by their maximum likelihood estimators based on complete observations. Another estimator which was also used by Ekbohm (1976) is

$$\hat{\mu}=\bar{x}-\bar{y}. \tag{2.3}$$

Note that $\hat{\mu}$ belongs to the class defined by (2.1) with $\rho=0$ in (2.2). For the case of unknown Σ and $R=1$ we propose the following estimator of δ:

$$\hat{\gamma}=\hat{\lambda}_1\bar{x}_1-\hat{\lambda}_2\bar{y}_1+(1-\hat{\lambda}_1)\bar{x}_2-(1-\hat{\lambda}_2)\bar{y}_2, \tag{2.4}$$

where

$$\begin{aligned}\hat{\lambda}_1&=n(n+n_2+n_1c)\{(n+n_1)(n+n_2)-n_1n_2c^2\}^{-1},\\ \hat{\lambda}_2&=n(n+n_1+n_2c)\{(n+n_1)(n+n_2)-n_1n_2c^2\}^{-1},\end{aligned} \tag{2.5}$$

and $c=0.5$ when $\rho>0$, and $c=-0.5$ when $\rho<0$. The numerical computations show that these values of c give the minimum average variance of $\hat{\gamma}$ over various values of ρ ($\rho=-1(0.1)1$). In these cases the variance of $\hat{\gamma}$ is smaller than that of $\hat{\mu}$ over a wide range of values of ρ except when ρ is near zero. Therefore one can expect that the test based on $\hat{\gamma}$ will be more powerful than those based on $\hat{\mu}$ for those values of ρ.

3. TEST STATISTICS

3.1 Test statistics based on $\hat{\delta}$, $\hat{\mu}$ and $\hat{\gamma}$

When $\sigma_1^2=\sigma_2^2=\sigma^2$, the variance of $\hat{\gamma}$ can be expressed as

$$\begin{aligned}\mathrm{Var}(\hat{\gamma})&=2\sigma^2(1-\rho)\hat{\lambda}_1\hat{\lambda}_2/n\\ &\quad+\sigma^2\{(\hat{\lambda}_1-\hat{\lambda}_2)^2/n+(1-\hat{\lambda}_1)^2/n_1+(1-\hat{\lambda}_2)^2/n_2\}.\end{aligned}$$

$2\sigma^2(1-\rho)$ in the first term and σ^2 in the second term can be estimated, as usual, from n complete pairs and n_1 and n_2 extra observations respectively. Thus the estimator of $\mathrm{Var}(\hat{y})$ is a linear combination of two independent χ^2 variates and can be taken as approximately distributed as χ^2. Our proposed statistic is

$$S_b = \{\hat{\lambda}_1\bar{x}_1 - \hat{\lambda}_2\bar{y}_1 + (1-\hat{\lambda}_1)\bar{x}_2 - (1-\hat{\lambda}_2)\bar{y}_2\}(d+e)^{-1/2},$$

where

$$d = (a_{11} + a_{22} - 2a_{12})\hat{\lambda}_1\hat{\lambda}_2\{n(n-1)\}^{-1},$$

$$e = (b_1 + b_2)\{(\hat{\lambda}_1 - \hat{\lambda}_2)^2/n + (1-\hat{\lambda}_1)^2/n_1 + (1-\hat{\lambda}_2)^2/n_2\}(n_1 + n_2 - 2)^{-1},$$

and $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are given by (2.5). When the hypothesis of equal means holds, S_b is approximately distributed as t with degrees of freedom given by

$$\nu_1 = [(d+e)^2/\{d^2/(n+1) + e^2/(n_1+n_2)\}] - 2.$$

The expression for the degrees of freedom is obtained by fitting first and second moments; see Searle (1971, p. 417). The statistic S_b is also computed when c in $\hat{\lambda}_1$ and $\hat{\lambda}_2$ is replaced by u and the resulting statistic is denoted by S_b'.

Our second statistic is

$$T_b = \frac{\hat{\lambda}_1\bar{x}_1 - \hat{\lambda}_2\bar{y}_1 + (1-\hat{\lambda}_1)\bar{x}_2 - (1-\hat{\lambda}_2)\bar{y}_2}{\sqrt{\left\{\frac{\hat{\lambda}_1^2}{n} + \frac{\hat{\lambda}_2^2}{n} - \frac{2r\hat{\lambda}_1\hat{\lambda}_2}{n} + \frac{(1-\hat{\lambda}_1)^2}{n_1} + \frac{(1-\hat{\lambda}_2)^2}{n_2}\right\}}\sqrt{\frac{S^*}{(n+n_1+n_2-2)}}},$$

where $S^* = b_1 + c_2$ for $n_1 \geqq n_2$ and $S^* = b_2 + c_1$ when $n_2 > n_1$, and $\hat{\lambda}_p (p=1,2)$ are given by (2.5). If r is replaced by ρ, the exact distribution of T_b would be Student's t with $n+n_1+n_2-2$ degrees of freedom. By following Lin and Stivers (1974), we claim that T_b is approximately t distributed with $n+n_1+n_2-4$ degrees of freedom. This claim is well supported by the simulation study to be discussed in Section 4.

Ekbohm (1976) compared most of the statistics available at the time and recommended four statistics which we shall use in our study. For the case of equal variances, he constructed his first statistic by dividing the estimator (2.1) by its estimated standard error, using the unbiased estimator σ^2 based on all observations, and replacing ρ by u. He denoted this statistic by Z_e, which is taken to be distributed as t with n degrees of freedom. His second statistic, S_e, was based on the estimator (2.3) and was derived by the method used in obtaining S_b. Lin and Stivers (1974) constructed the test statistic by dividing the estimator $\hat{\delta}$ by its estimated standard error, where σ_1^2, σ_2^2 and ρ were replaced by their maximum likelihood estimates based on n complete observations. This statistic, which we denote as Z_{1s}, is approximately distributed as t when the hypothesis of equal means holds. Another statistic proposed by these authors is similar to our proposed statistic T_b except that it is based on the estimator $\hat{\mu}$. We denote this statistic by T_{1s}. One of the statistics which was not included in Ekbohm's study is due to Woolson, Leeper, Cole and Clarke (1976). This statistic, to be denoted by S_w, was derived by using the theory of general linear hypothesis under the assumption of equal variances and known ρ and, then replacing ρ by u.

3.2 Test statistics based on combination of independent tests

When $\sigma_1^2 = \sigma_2^2$, we have two independent test statistics t_1 and t_2 for testing $\delta = 0$ where

$$t_1 = (\bar{x}_1 - \bar{y}_1)\sqrt{\{n(n-1)\}}/\sqrt{(a_{11} + a_{22} - 2a_{12})},$$

$$t_2 = (\bar{x}_2 - \bar{y}_2)/\sqrt{[\{(b_1 + b_2)/(n_1 + n_2 - 2)\}(1/n_1 + 1/n_2)]}.$$

The problem is to select a function of t_1 and t_2 to be used as the combined test statistic. Unfortunately, there is no uniformly most powerful procedure for combining the independent test statistics. However, Littel and Folks (1973) have shown that according to Bahadur relative efficiency, Fisher's method (Fisher, 1950, pp. 99–101) is asymptotically optimal among essentially all methods of combining independent tests. Hence, we suggest the statistic

$$L = -2\log_e(L_1 L_2),$$

where L_p is the observed level of significance obtained from $t_p (p=1,2)$. When the means are equal, L is distributed as χ^2 with four degrees of freedom.

For the case of equal variances, Bhoj (1978) proposed the statistic

$$T=\lambda t_1+(1-\lambda)t_2$$

where the value of $\lambda (0<\lambda<1)$ was determined by minimizing the average value of the ratio $E(L_T^2)/E(L_{t_1}^2)$ over the various values of $\rho>0$, where $E(L_T^2)$ and $E(L_{t_1}^2)$ are the expected square lengths of the confidence intervals on δ based on T and t_1, respectively. When the means are equal, T is the linear combination of two independent random variables which are distributed according to Student's t distribution with $f_1=n-1$ and $f_2=n_1+n_2-2$ degrees of freedom. The percentage points of T can be obtained by using approximations due to Patil (1965) and Ghosh (1975). Recently, Walker and Saw (1978) have shown that the exact percentage points of T can be computed by using only tables of t distribution if f_1 and f_2 are odd. However, this method is laborious and it has limited applications since it requires that f_1 and f_2 be odd.

Instead of approximating the distribution of T, one could transform each t_p into a new variable so that the distribution of the linear combination of these new variables has approximately a well known distribution. This can be achieved first by transforming t_p to a symmetric F_p via Cacoullos's (1965) result, and then to U_p $(p=1,2)$ by using Paulson's (1942) approximation; see Bhoj (1979). Then our statistic is

$$Z=\{\lambda U_1+(1-\lambda)U_2\}\{\lambda^2+(1-\lambda)^2\}^{-1/2} \tag{3.1}$$

where

$$U_p=\left(1-\frac{2}{9f_p}\right)(F_p^{1/3}-1)\left\{\frac{2}{9f_p}(F_p^{2/3}+1)\right\}^{-1/2}$$

and

$$F_p=1+(2t_p^2 f_p)+(2t_p/\sqrt{f_p})\sqrt{(1+t_p^2/f_p)}.$$

When the means are equal, Z is approximately normally distributed with zero mean and unit variance.

It remains to determine an adequate formula for λ. If $\sigma_1^2=\sigma_2^2=\sigma^2$, and σ^2 and ρ are assumed to be known, then t_1 and t_2 are to be replaced by two independent statistics which are distributed as unit normal. If these statistics are combined as in (3.1) then the value of λ which maximizes the power of the resulting statistic is given by

$$\lambda=[1+\sqrt{\{2n_1n_2(1-\rho)/n(n_1+n_2)\}}]^{-1}.$$

When ρ is unknown it may be replaced by r, and the resulting λ is used in the statistic Z. The values of λ, which were given by Bhoj (1978) for $\rho>0$, are close to the ones obtained from the above formula when ρ is replaced by 0.5.

4. A SIMULATION STUDY

The test based on t_1 is well-known, and some simulation results are available for the statistics Z_e, S_e, T_{1s}, Z_{1s} and S_w. However, further study was needed for the statistics S_b, S_b', T_b, Z and L proposed in Section 3. For this reason, using the Box-Muller (1958) technique, one thousand random samples were generated from a bivariate normal distribution with various values of μ, R^2, ρ, and combinations of n, n_1 and n_2. The hypothesis H_0: $\delta=0$ against H_1: $\delta>0$ was tested at the 5% level of significance and the resulting empirical levels and powers of the test statistics are compared.

First, we compare the statistics based on the estimator (2.1), S_w and L. The statistic S_w is conservative compared to Z_{1s} and Z_e, and this sometimes results in loss of power. In general, Z_{1s} has a larger size compared to the other statistics, and partly because of this it seems to give higher power. If we take into account the larger empirical sizes of Z_{1s}, then the statistics Z_e, Z_{1s} and S_w seem to have approximately the same power, at least when the variances are equal. The statistic L has a somewhat smaller power than these statistics when n_1 and n_2 are small. Indeed, in these cases L can be less powerful than t_1 when $\rho=0.9$. However, as is expected, when n, n_1 and n_2 are large, L becomes competitive with the statistics Z_e, Z_{1s} and S_w. L is particularly attractive because its exact distribution is known and the critical values are readily available.

The statistics S_e and T_{1s} based on the estimator (2.3), as expected, give higher powers when $|\rho|$ is small and they can be inferior to complete data test when $\rho>0.5$. The statistics S_b and T_b which we proposed in this paper perform very well over a wide range of values of ρ. They are preferred to Z_{1s} and Z_e for all negative values and up to moderately large positive values of ρ. As is expected, their powers are slightly smaller than those of S_e and T_{1s} when $|\rho|$ is very small. They can be less powerful than the complete data test when $\rho=0.9$ and $R=1$. The powers of S_b and T_b are approximately the same, but S_b is more robust against heteroscedasticity, at least when $n_1=n_2$. Therefore, we prefer S_b to T_b. The powers of S_b and S'_b are about the same for all ρ except when ρ is positive and large, in which case S'_b is superior to S_b. The empirical values of α, the size of the test, tend to be larger for S'_b for large negative values of ρ.

The statistics based on the estimator (2.3) are better than those based on the maximum likelihood estimators when $|\rho|\leqq 0.5$ and vice versa when $\rho>0.5$. The powers of the statistics based on the maximum likelihood estimator are not always higher than those based on the estimator (2.3) when the variables are highly negatively correlated. The statistics S_b and T_b based on the estimator (2.4) perform very well for all negative and up to moderately positive values of ρ. The powers of Z are closer to or slightly better than those based on the estimator (2.3) when $|\rho|<0.5$. Z is a good competitor to the statistics based on the maximum likelihood estimator when ρ is moderate or large. In addition, Z is better than these statistics for $\rho\leqq -0.5$.

The powers of many statistics under investigation are very close to one another. In order to facilitate the comparison of the powers of some of the new statistics with the known statistics, we have graphically displayed the power curves of only four statistics: t_1, S'_b, Z and the statistic which gave the maximum power from among the known statistics. In most cases, the dominating statistic from the known statistics is Z_{1s}. This is not surprising since Z_{1s} performs very well for $\rho\geqq 0.5$. It may be emphasized that empirical levels of Z_{1s} are also high. To avoid distracting the reader, the power results are not shown for the tests whose empirical sizes are beyond three standard deviations from the nominal level. This has happened for Z_{1s} in only one case among those plotted. The test statistics are parenthesized if the empirical α levels are more than two standard deviations from the nominal level. From Figures 1, 2, 3 and 4 it is observed that for

$R=1$ the powers of Z and S_b' are higher or almost the same when compared to the best statistic selected from the known statistics. For the heteroscedastic cases, the powers of Z and S_b' do not lag too far behind Z_{1s} when the effect of larger empirical values of α for Z_{1s} is taken into account. For some cases, Z, Z_{1s} and S_b' essentially have the same powers for the three chosen values of $\rho>0$. However, Z_{1s} is inferior to Z and S_b' when $\rho<0.5$. The statistics Z and S_b' are fairly robust for departures from the homoscedasticity assumption.

To summarize, we note that the three graphed test statistics are better than the complete data test. The three statistics are close competitors for large positive ρ. But when $\rho<0.5$, Z and S_b' are better than Z_{1s}. There is no single test statistic which is superior to all other statistics over the entire parameter space. We should choose the statistic which performs well from the significance level and power point of view over the wide range of values of ρ. Z retains the significance level close to the nominal level while S_b' seems to give slightly higher values of empirical α for large negative values of ρ. Hence, in conclusion, the use of Z is recommended at least for the homoscedastic or nearly homoscedastic cases, and when nothing is known about ρ. S_b' is recommended when it is known that the two variables are not highly negatively correlated, and the variances are not believed to differ much. When the variances are believed to be very unequal, Z_{1s} is slightly better than Z and S_b' for moderately large or large values of ρ, but its size may be too large for $n<10$ and in some cases even for large n.

Acknowledgement

The author would like to thank the associate editor and the referee for their helpful suggestions.

References

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

Bhoj, D. S. (1979). Testing equality of variances of correlated variates with incomplete data on both responses. *Biometrika* **66**, 681–683.

Box, G. E. P. and Muller, M. E. (1958). A note on the generation of random normal deviates. *Ann. Math. Statist.* **29**, 610–611.

Cacoullos, T. (1965). A relation between t and F distributions. *J. Am. Statist. Assoc.* **60**, 528–531.

Ekbohm, G. (1976). Comparing means in the paired case with incomplete data on both responses. *Biometrika* **63,** 299–304.

Fisher, R. A. (1950). *Statistical Methods for Research Workers,* London: Oliver and Boyd.

Ghosh, B. K. (1975). On the distribution of the difference of two t variables. *J. Am. Statist. Assoc.* **70,** 463–467.

Lin, P. E. and Stivers, L. E. (1974). On the difference of means with incomplete data. *Biometrika* **61,** 325–334.

Littel, R. C. and Folks, J. L. (1973). Asymptotic optimality of Fisher's method of combining independent tests II. *J. Am. Statist. Assoc.* **68,** 193–194.

Orchard, T. and Woodbury, M. A. (1970). A missing information principle: theory and applications. *Proc. 6th Berkeley Symp.* **1,** 697–715.

Patil, V. H. (1965). Approximation to the Behrens–Fisher distributions. *Biometrika* **52,** 267–271.

Paulson, E. (1942). An approximate normalization of the analysis of variance distribution. *Ann. Math. Statist.* **13,** 233–235.

Rao, C. R. (1952). *Advanced Statistical Methods in Biometric Research,* Wiley, New York.

Searle, S. R. (1971). *Linear Models,* Wiley, New York.

Walker, G. A. and Saw, J. G. (1978). The distribution of linear combinations of t-variables. *J. A. Statist. Assoc.* **73,** 87–88.

Woolson, R. F. and Cole, J. W. L. (1974). Comparing means of correlated variates with missing data. *Commun. Statist.* **3,** 941–948.

Woolson, R. F., Leeper, J. D., Cole, J. W. L. and Clarke, W. R. (1976). A Monte Carlo investigation of a statistic for a bivariate missing data problem. *Commun. Statist.—Theor. Meth.* **A5,** 681-688.

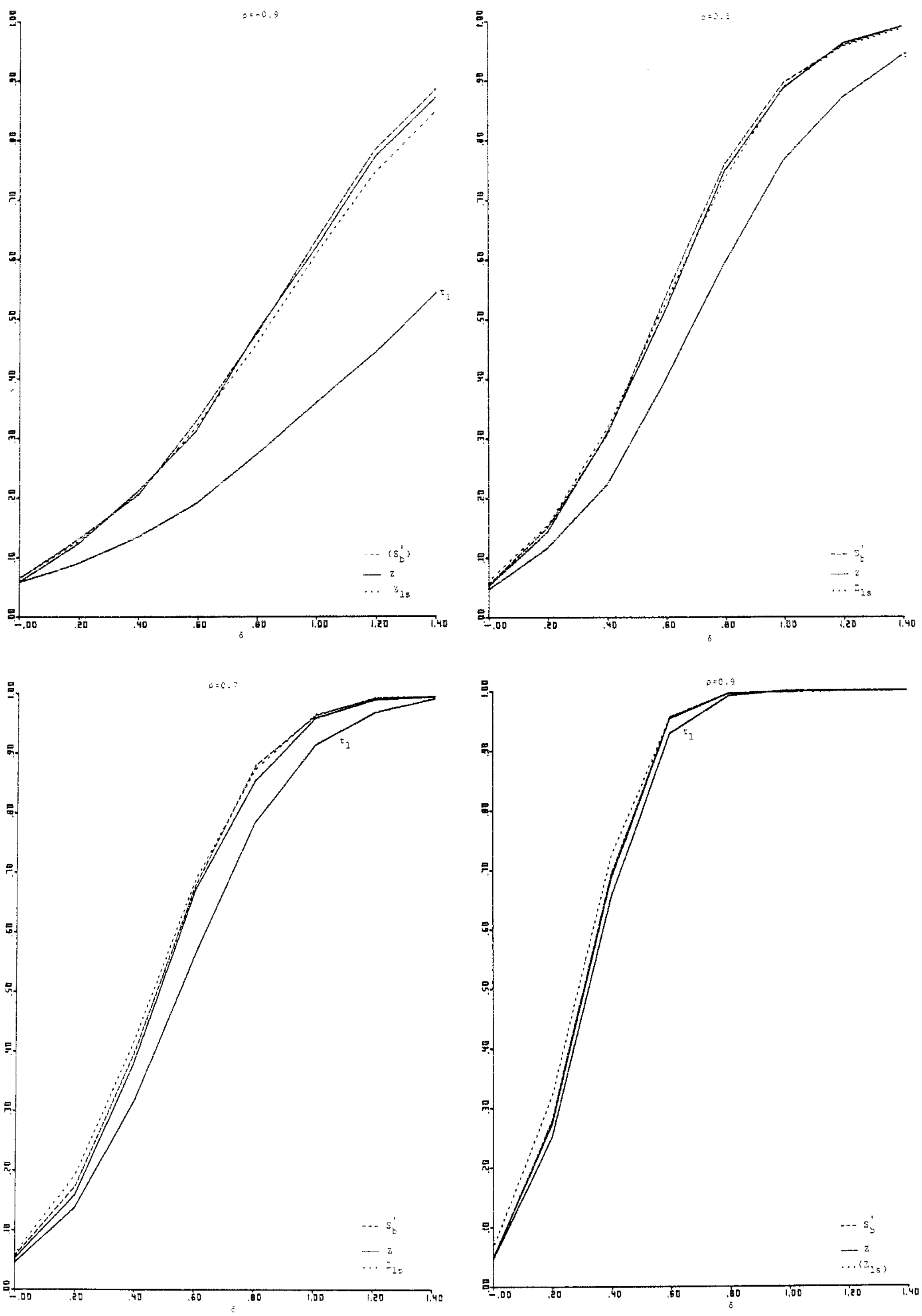

FIGURE 1 Estimasted power curves for $\sigma_1^2/\sigma_2^2=1$; $n=7$, $n_1=n_2=5$.

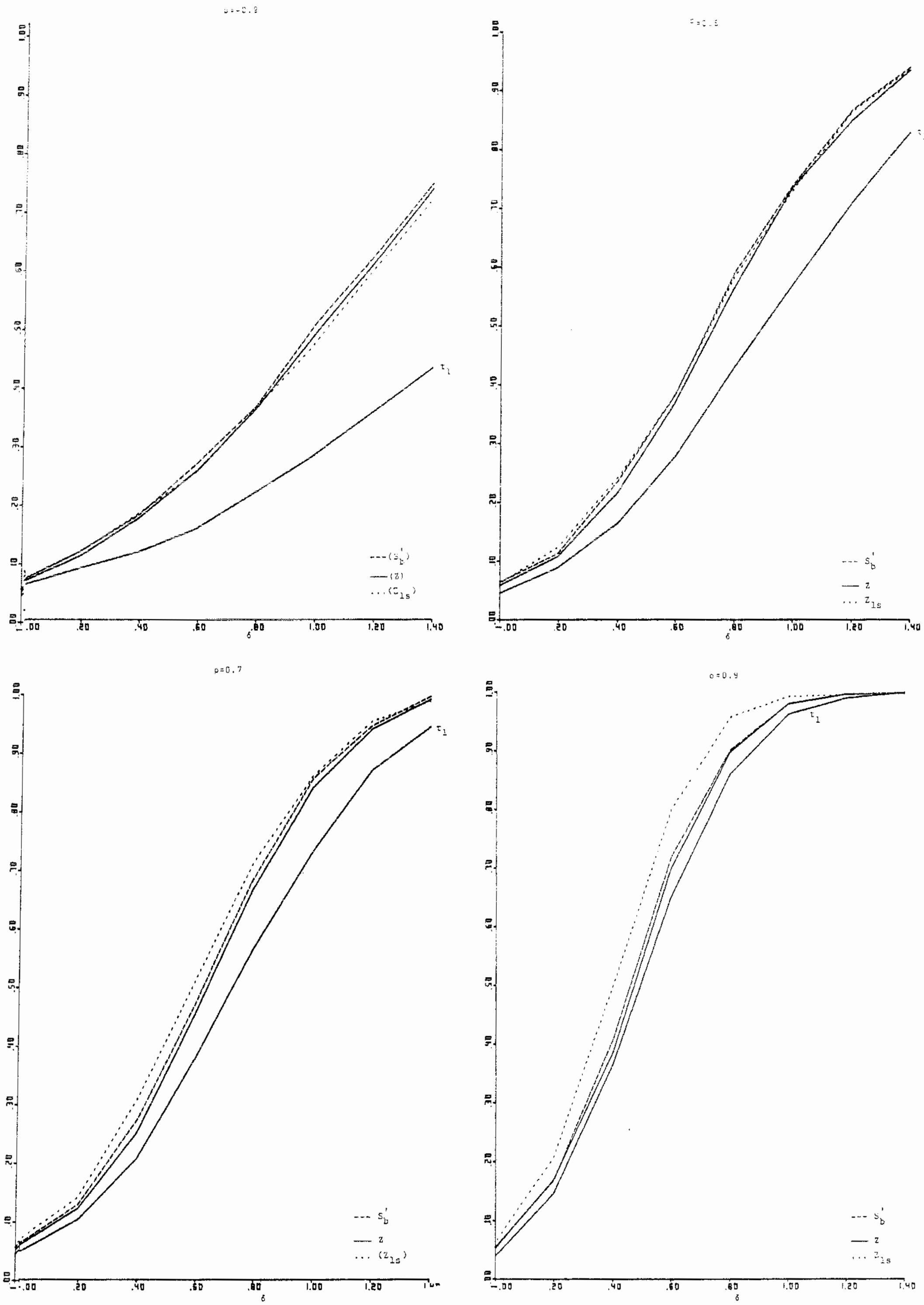

FIGURE 2 Estimated power curves for $\sigma_1^2/\sigma_2^2=2$; $n=7$, $n_1=n_2=5$.

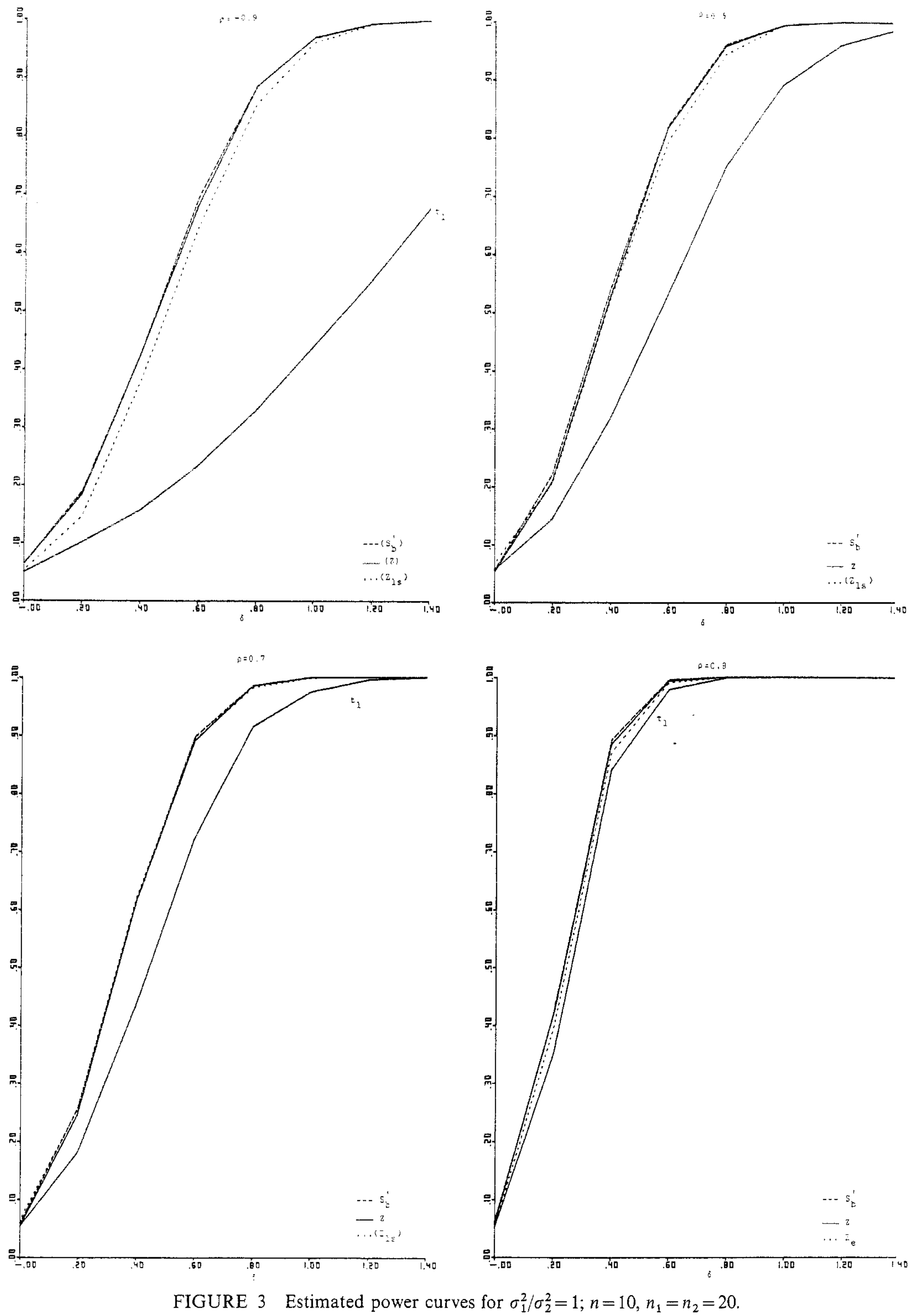

FIGURE 3 Estimated power curves for $\sigma_1^2/\sigma_2^2=1$; $n=10$, $n_1=n_2=20$.

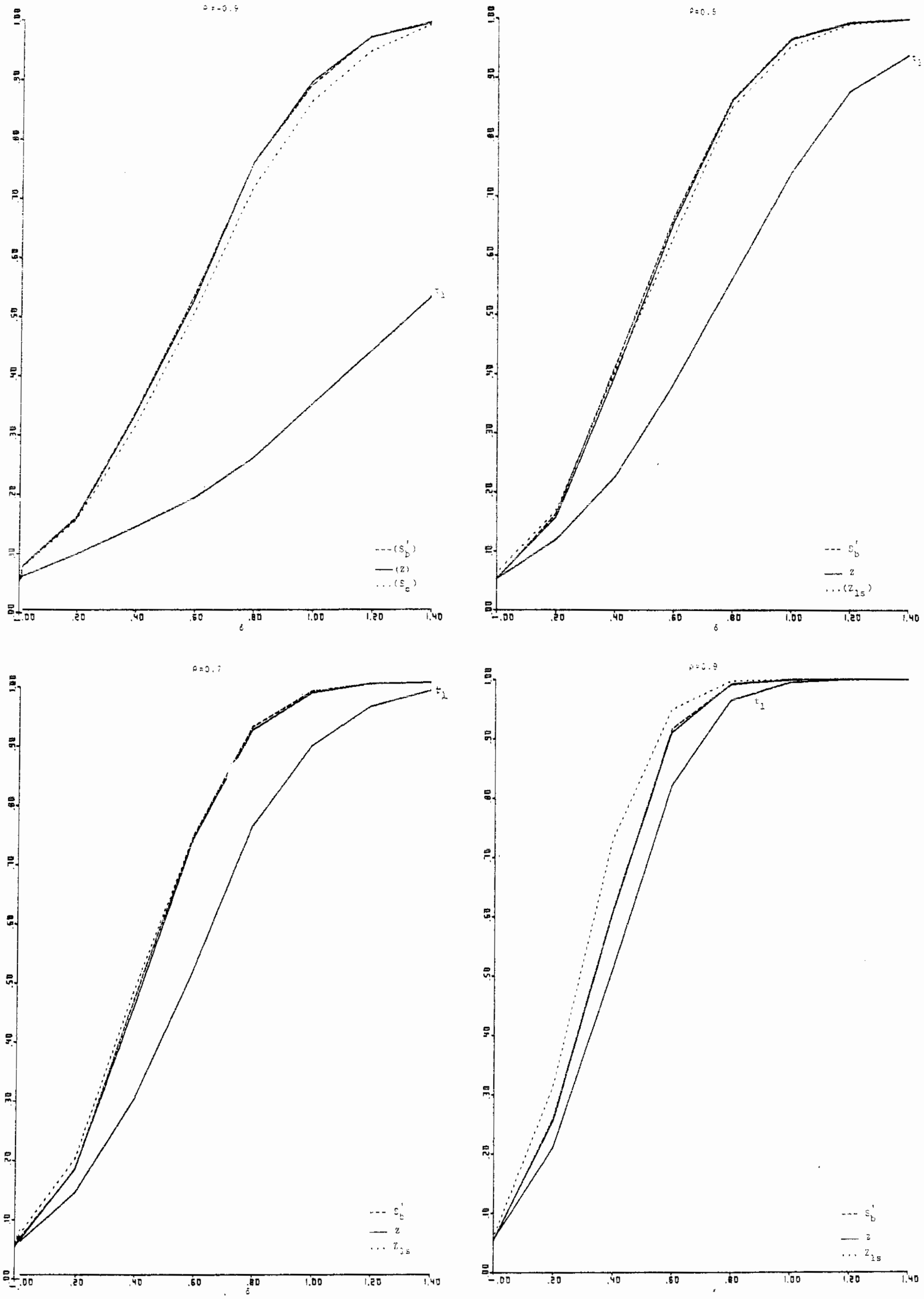

FIGURE 4 Estimated power curves for $\sigma_1^2/\sigma_2^2=2$; $n=10$, $n_1=n_2=20$.