

## On combining independent tests in linear models

Scott M. Jordan, K. Krishnamoorthy\*

*Department of Statistics, University of Southwestern Louisiana, Lafayette, LA 70504-1006, USA*

Received April 1994

---

### Abstract

The problem of testing the common mean of several normal linear models when the variances are unknown and unequal is considered. A minor modification to the combined tests given in Zhou and Mathew (1993a) is proposed. The powers of the modified tests are numerically compared with their original versions and a traditional test. Comparison studies indicate that the modified tests have better power properties than their original versions. Practical situations where the modified tests are preferable to the traditional test are pointed out.

*Keywords:* Combined test; Fisher's test; Recovery of interblock information

---

### 1. Introduction

In the analysis of balanced incomplete block design (BIBD) with treatment effects fixed and block effects random, Cohen and Sackrowitz (1989) proposed an exact test that combines both intrablock and interblock F-tests. It is also shown by Monte Carlo methods that the combined test has excellent power properties compared to the F-test based on intrablock information alone. Their combined test may be more appropriate for BIBD since it is known that intrablock variance is smaller than interblock variance. However, extending their innovative idea, Zhou and Mathew (1993a) proposed combined tests for testing the common parameter involved in several independent normal linear models. The combined tests are essentially based on the weighted average of the  $p$ -values of independent F-tests available from different models, where the weights are inversely proportional to the sample variances. These results are further generalized by Mathew et al. (1993). For other related work in this area, we refer the readers to the references given in these three papers just cited.

In this note we consider this problem in the context of testing the common mean of several normal populations with unknown and unequal variances, although the idea will clearly work in other situations too (see Remark 3.2). For motivation and applications of this particular context, for example, see Cohen and Sackrowitz (1984). In the following section, we observe first that the weights used to combine independent

---

\* Corresponding author. Tel.: (+1-318) 482 6771.

tests considered in Cohen and Sackrowitz (1989), Zhou and Mathew (1993a) and Mathew et al. (1993) are independent of the sample sizes from different populations. When the sample sizes are unequal, weighing independent tests appropriately based on their sample sizes, one can improve their combined tests. In fact, we show by intuitive argument and numerical comparisons that the modified tests are better than their original versions and in general they are the appropriate ones to use.

## 2. Main result

Suppose that we have independent samples from  $k$  normal populations with the same mean  $\mu$  but different unknown variances  $\sigma_1^2, \dots, \sigma_k^2$ . Let  $\bar{x}_i$  and  $s_i^2$  respectively denote the mean and variance of the  $i$ th sample of size  $n_i$ ,  $i = 1, \dots, k$ . The problem of interest here is to test

$$H_0: \mu = 0 \quad (\text{vs}) \quad H_a: \mu \neq 0. \quad (2.1)$$

Let  $t_i^2 = n_i \bar{x}_i^2 / s_i^2$  and  $F_i(\cdot)$  denote the cumulative distribution of F-random variable with 1 and  $n_i - 1$  degrees of freedom,  $i = 1, \dots, k$ . Further, let

$$Z_i = -\ln(1 - F_i(t_i^2)) \quad \text{and} \quad W = \sum_{i=1}^k Z_i. \quad (2.2)$$

The traditional Fisher's test is based on  $W$ . Note that  $2Z_i$ 's are independently distributed as chi-squared random variables with 2 degrees of freedom. So for a given level of significance  $\alpha$ , Fisher's test rejects  $H_0$  if

$$2W > \chi_{2k}^2(\alpha), \quad (2.3)$$

where  $\chi_{2k}^2(\alpha)$  denotes the upper  $100(1 - \alpha)$  percentile point of a chi-squared distribution with  $2k$  degrees of freedom. Notice that Fisher's test gives equal weight to all the independent F-tests although the population variances are different. So it is logical to assign to each test a weight that is inversely proportional to the corresponding population variance. If the variances are unknown then their estimates can be used. Using this idea, Zhou and Mathew (1993a) proposed the test based on

$$Z = \sum_{i=1}^k \gamma_i Z_i, \quad (2.4)$$

where

$$\gamma_i = T_i^{-1} / \left( \sum_{i=1}^k T_i^{-1} \right) \quad \text{and} \quad T_i = ((n_i - 1)s_i^2 + n_i \bar{x}_i^2) / n_i \quad (2.5)$$

is an unbiased estimator of  $\sigma_i^2$  when  $H_0$  is true. They also showed that, for given  $\alpha$ , the test that rejects  $H_0$  when

$$\sum_{i=1}^k \frac{\gamma_i^{k-1} e^{-Z/\gamma_i}}{\prod_{j=1; j \neq i}^k (\gamma_i - \gamma_j)} \leq \alpha(1 + \eta), \quad (2.6)$$

where  $\eta = [\sum_{i < j} \bar{x}_i \bar{x}_j / (|\bar{x}_i| |\bar{x}_j|)] / [k(k-1)/2]$ , has exact size  $\alpha$ . An intuitive reasoning for using  $\eta$  is given in Cohen and Sackrowitz (1989) and Zhou and Mathew (1993a). To add to their arguments, we also observe from (2.6) and the definition of  $\eta$  that, for  $k = 2$ , this test accepts the null hypothesis whenever  $\bar{x}_1$  and  $\bar{x}_2$  have opposite signs no matter what is the size of the test and observed values of other statistics. This clearly agrees with the intuitive reasoning that when the sample means have different signs we expect the common mean  $\mu$  to be around zero.

In the above test (2.6), note that the weight  $\gamma_i$  associated with the  $i$ th independent test is inversely proportional to an estimate of  $\sigma_i^2$ ,  $i = 1, \dots, k$ . However, we recall that the power of the  $i$ th independent test is not only depending on  $\sigma_i^2$  but is also directly related to the sample size  $n_i$ . Therefore, instead of  $\gamma_i$ , we suggest using

$$\gamma_{i0} = n_i T_i^{-1} / \left( \sum_{i=1}^k n_i T_i^{-1} \right) \quad (2.7)$$

in (2.4) to get a better test. Thus, the combined test given in (2.6) can be modified accordingly. The modified test rejects  $H_0$  if

$$\sum_{i=1}^k \frac{\gamma_{i0}^{k-1} e^{-Z_0/\gamma_{i0}}}{\prod_{j=1; j \neq i}^k (\gamma_{i0} - \gamma_{j0})} \leq \alpha(1 + \eta), \quad (2.8)$$

where  $Z_0 = \sum_{i=1}^k \gamma_{i0} Z_i$ , and it also has exact size  $\alpha$ .

We next consider the case where one of the population variances is known to be smaller than the other as in the interblock analysis of a balanced incomplete block design. Suppose that  $k = 2$  and  $\sigma_1^2 < \sigma_2^2$ . Let  $U_i = (n_i - 1)s_i^2 + n_i \bar{x}_i^2$ ,  $i = 1, 2$ . In this case, Cohen and Sackrowitz (1989) suggested the test in (2.6) with

$$\gamma_1 = 1/[1 + \min(U_1/U_2, 1)] \quad \text{and} \quad \gamma_2 = 1 - \gamma_1 \quad (2.9)$$

so that  $\gamma_1 > \gamma_2$ . Instead, we suggest using

$$\gamma_1 = 1/[1 + \min(W_1/W_2, 1)] \quad \text{and} \quad \gamma_2 = 1 - \gamma_1, \quad (2.10)$$

where  $W_i = U_i/n_i^2$ ,  $i = 1, 2$ , in (2.6) if  $\sigma_1^2/n_1 < \sigma_2^2/n_2$ . In other words, we point out that the weights should reflect the inequality restriction of  $\sigma_i^2/n_i$ 's instead of  $\sigma_i^2$ 's. If the relationship between  $\sigma_1^2/n_1$  and  $\sigma_2^2/n_2$  is not known, then the test (2.8) can be used.

Next, in order to demonstrate the power improvement of the modified tests, in the following, we estimate the powers of all these combined tests using the Monte Carlo method.

### 3. Simulation results

In this section, simulated powers of the modified tests are compared with the simulated powers of their original versions. For simulation, the random variables are generated by IMSL subroutines RNNOR (for normal) and RNCHI (for chi-squared variable). Each simulation consists of 100 000 runs and, as we estimate the probabilities, the maximum standard error of an estimate is  $\sqrt{0.25/100\,000} = 0.0016$ . Hence, all the estimates are accurate at least to the second decimal place.

The simulated powers are presented in Table 1 for the following five tests:

1. Fisher's test in (2.3);
2. the test in (2.6) with  $\eta = 0$ ;
3. the modified test in (2.8) with  $\eta = 0$ ;
4. the test in (2.6) using  $\eta$ ;
5. the modified test (2.8) using  $\eta$ .

Note that when the sample sizes are equal, test 3 coincides with test 2, and test 5 coincides with test 4. For this reason, the comparisons are made only for unequal sample sizes.

Table 1 gives powers of the tests at  $\alpha = 0.05$  for sample sizes  $n_1 = 13$ ,  $n_2 = 5$  ( $n_1 = 5$ ,  $n_2 = 13$ ) and for various values of  $\sigma_2^2/\sigma_1^2$ . From the table we observe the following:

Table 1  
Simulated powers of the tests at  $\alpha = 0.05$

$\sigma_2^2/\sigma_1^2$	Test	$\mu$				
		0.2	0.4	0.6	0.8	1
1	1	0.10 (0.10)	0.25 (0.25)	0.52 (0.52)	0.79 (0.79)	0.94 (0.94)
	2	0.09 (0.09)	0.21 (0.21)	0.45 (0.45)	0.70 (0.70)	0.87 (0.87)
	3	0.10 (0.10)	0.25 (0.25)	0.51 (0.51)	0.76 (0.76)	0.92 (0.92)
	4	0.11 (0.11)	0.29 (0.29)	0.56 (0.56)	0.80 (0.80)	0.93 (0.93)
	5	0.12 (0.12)	0.32 (0.32)	0.61 (0.61)	0.84 (0.84)	0.95 (0.95)
2	1	0.09 (0.07)	0.22 (0.16)	0.46 (0.34)	0.71 (0.56)	0.90 (0.77)
	2	0.09 (0.07)	0.23 (0.14)	0.45 (0.28)	0.68 (0.46)	0.85 (0.67)
	3	0.10 (0.07)	0.26 (0.15)	0.50 (0.31)	0.75 (0.51)	0.91 (0.71)
	4	0.10 (0.09)	0.27 (0.20)	0.51 (0.39)	0.74 (0.60)	0.88 (0.79)
	5	0.11 (0.09)	0.29 (0.21)	0.54 (0.41)	0.77 (0.64)	0.91 (0.82)
3	1	0.09 (0.07)	0.21 (0.14)	0.43 (0.28)	0.68 (0.46)	0.87 (0.66)
	2	0.10 (0.07)	0.23 (0.12)	0.45 (0.23)	0.70 (0.38)	0.86 (0.57)
	3	0.10 (0.07)	0.26 (0.13)	0.49 (0.24)	0.75 (0.41)	0.91 (0.60)
	4	0.10 (0.08)	0.26 (0.17)	0.49 (0.32)	0.71 (0.52)	0.85 (0.71)
	5	0.10 (0.08)	0.27 (0.18)	0.51 (0.34)	0.73 (0.54)	0.87 (0.72)
4	1	0.09 (0.07)	0.21 (0.12)	0.42 (0.24)	0.67 (0.40)	0.85 (0.60)
	2	0.10 (0.06)	0.24 (0.11)	0.47 (0.21)	0.71 (0.34)	0.87 (0.51)
	3	0.10 (0.06)	0.26 (0.11)	0.50 (0.21)	0.75 (0.35)	0.91 (0.53)
	4	0.10 (0.08)	0.25 (0.16)	0.48 (0.29)	0.69 (0.47)	0.83 (0.66)
	5	0.10 (0.08)	0.26 (0.16)	0.49 (0.30)	0.70 (0.48)	0.84 (0.66)
5	1	0.09 (0.06)	0.21 (0.12)	0.41 (0.22)	0.66 (0.36)	0.85 (0.53)
	2	0.10 (0.06)	0.25 (0.11)	0.49 (0.19)	0.72 (0.32)	0.88 (0.48)
	3	0.10 (0.06)	0.26 (0.11)	0.51 (0.19)	0.75 (0.32)	0.91 (0.48)
	4	0.10 (0.07)	0.25 (0.15)	0.47 (0.27)	0.67 (0.44)	0.81 (0.62)
	5	0.10 (0.08)	0.25 (0.15)	0.48 (0.28)	0.68 (0.44)	0.82 (0.62)
10	1	0.09 (0.06)	0.20 (0.09)	0.40 (0.18)	0.64 (0.28)	0.82 (0.41)
	2	0.10 (0.06)	0.26 (0.10)	0.50 (0.18)	0.74 (0.29)	0.90 (0.42)
	3	0.10 (0.06)	0.27 (0.10)	0.51 (0.18)	0.75 (0.29)	0.91 (0.42)
	4	0.10 (0.07)	0.24 (0.13)	0.43 (0.23)	0.62 (0.37)	0.75 (0.53)
	5	0.10 (0.07)	0.24 (0.13)	0.44 (0.24)	0.62 (0.37)	0.75 (0.53)
20	1	0.08 (0.06)	0.20 (0.10)	0.40 (0.16)	0.62 (0.24)	0.81 (0.34)
	2	0.10 (0.07)	0.26 (0.11)	0.51 (0.18)	0.75 (0.28)	0.91 (0.40)
	3	0.10 (0.07)	0.26 (0.11)	0.51 (0.18)	0.76 (0.28)	0.91 (0.40)
	4	0.10 (0.07)	0.22 (0.12)	0.41 (0.21)	0.57 (0.33)	0.69 (0.46)
	5	0.10 (0.07)	0.23 (0.12)	0.41 (0.21)	0.57 (0.33)	0.69 (0.47)

$n_1 = 13, n_2 = 5$  ( $n_1 = 5, n_2 = 13$ ).

1. Test 3 has more power than test 2 and test 5 has more power than test 4 for all  $\sigma_2^2/\sigma_1^2$ . This indicates that using the modified tests instead of their original versions is certainly advantageous.
2. In comparisons of tests 3 and 5, we found that test 3 is preferable to test 5 for large values of  $\sigma_2^2/\sigma_1^2$  and vice versa for small values of  $\sigma_2^2/\sigma_1^2$ .
3. If  $\sigma_1^2$  is known to be smaller than  $\sigma_2^2$  and  $n_1 > n_2$ , then test 3 is certainly preferable to the Fisher's test.
4. Test 5 is highly recommended when  $\sigma_2^2/\sigma_1^2$  is not too large. This may be the situation in many applications. For example, when different instruments are used to measure like products to estimate the

Table 2  
Simulated powers of the tests at  $\alpha = 0.05$

$\sigma_2^2/\sigma_1^2$	Test	$\mu$				
		0.2	0.4	0.6	0.8	1
1	5	0.27 (0.27)	0.73 (0.73)	0.95 (0.95)	0.99(0.99)	0.99 (0.99)
	6	0.12 (0.24)	0.34 (0.70)	0.73 (0.94)	0.96 (0.99)	0.99 (0.99)
	7	0.27 (0.24)	0.73 (0.70)	0.95 (0.94)	0.99 (0.99)	0.99 (0.99)
2	5	0.18 (0.25)	0.52 (0.67)	0.84 (0.91)	0.97 (0.99)	0.99 (0.99)
	6	0.10 (0.22)	0.22 (0.63)	0.44 (0.90)	0.73 (0.99)	0.92 (0.99)
	7	0.18 (0.22)	0.52 (0.63)	0.84 (0.90)	0.97 (0.99)	0.99 (0.99)
3	5	0.14 (0.24)	0.40 (0.64)	0.72 (0.88)	0.92 (0.98)	0.98 (0.99)
	6	0.09 (0.21)	0.19 (0.60)	0.36 (0.88)	0.59 (0.98)	0.81 (0.99)
	7	0.14 (0.21)	0.41 (0.60)	0.72 (0.88)	0.92 (0.98)	0.99 (0.99)
4	5	0.12 (0.24)	0.33 (0.62)	0.63 (0.87)	0.86 (0.98)	0.96 (0.99)
	6	0.09 (0.20)	0.18 (0.58)	0.33 (0.86)	0.52 (0.98)	0.73 (0.99)
	7	0.12 (0.20)	0.34 (0.58)	0.64 (0.86)	0.86 (0.98)	0.96 (0.99)
5	5	0.11 (0.23)	0.29 (0.61)	0.56 (0.86)	0.80 (0.98)	0.93 (0.99)
	6	0.08 (0.20)	0.17 (0.56)	0.31 (0.85)	0.49 (0.98)	0.68 (0.99)
	7	0.11 (0.20)	0.30 (0.56)	0.57 (0.85)	0.81 (0.98)	0.94 (0.99)
7	5	0.10 (0.23)	0.24 (0.59)	0.46 (0.84)	0.70 (0.97)	0.87 (0.99)
	6	0.08 (0.19)	0.16 (0.55)	0.30 (0.84)	0.46 (0.97)	0.63 (0.99)
	7	0.10 (0.19)	0.25 (0.56)	0.47 (0.84)	0.71 (0.97)	0.88 (0.99)
10	5	0.09 (0.23)	0.20 (0.57)	0.38 (0.83)	0.60 (0.97)	0.78 (0.99)
	6	0.08 (0.19)	0.15 (0.54)	0.28 (0.82)	0.43 (0.97)	0.60 (0.99)
	7	0.09 (0.19)	0.21 (0.54)	0.40 (0.82)	0.61 (0.97)	0.79 (0.99)

$n_1 = 5, n_2 = 50$  ( $n_1 = 50, n_2 = 5$ ).

average quality or when different pharmaceutical companies estimate the effect of the same drug, we may expect that the underlying "variances" are not far away from each other.

We next consider the setup where it is assumed that  $\sigma_1^2$  is known to be smaller than  $\sigma_2^2$ . For this setup, we present in Table 2 the simulated powers of the following tests:

5. the modified test (2.8) using  $\eta$ ;

6. the test (2.6) using  $\eta$ ,  $\gamma_1 = 1/[1 + \min(U_1/U_2, 1)]$  and  $\gamma_2 = 1 - \gamma_1$  which reflect the inequality constraint between  $\sigma_1^2$  and  $\sigma_2^2$ ;

7. the test (2.6) using  $\eta$ ,  $\gamma_2 = 1/[1 + \min(W_2/W_1, 1)]$  and  $\gamma_1 = 1 - \gamma_2$  when  $\sigma_1^2/n_1 \geq \sigma_2^2/n_2$ ;  $\gamma_1 = 1/[1 + \min(W_1/W_2, 1)]$  and  $\gamma_2 = 1 - \gamma_1$  when  $\sigma_1^2/n_1 \leq \sigma_2^2/n_2$ .

We note that the tests 6 and 7 use the known relationship between the variances whereas test 5 does not. From Table 2 we observe the following:

1. It is surprising to note that test 5 has better power than tests 6 and 7 when  $n_1 > n_2$  although it does not use the knowledge of the relationships between the variances. Also, when  $n_1 < n_2$  test 5 is as good as test 7.

2. In comparisons of tests 6 and 7, when  $n_1 = 50$  and  $n_2 = 5$  (this implies that  $\sigma_1^2/n_1 < \sigma_2^2/n_2$ ), we found virtually no differences between the powers. However, there are situations (for example,  $n_1 = 13, n_2 = 5$  and  $n_1 = 20, n_2 = 10$ ) where we found the powers of test 7 are slightly more than those of test 6 (they are not reported here). On the other hand, when  $n_1 = 5$  and  $n_2 = 50$ , the performance of test 6 is inferior to the others

as long as  $\sigma_1^2/n_1 \geq \sigma_2^2/n_2$ . This is mainly because the  $\gamma_i$ 's used in test 6 reflect the inequality restriction between  $\sigma_i^2$ 's instead of between  $\sigma_i^2/n_i$ 's.

For small samples, we found that test 7 has slightly more power than test 5 when the relationship between  $\sigma_1^2/n_1$  and  $\sigma_2^2/n_2$  is known (not reported here). Thus, in general, there is an advantage in using the known information between the variances of  $\bar{x}_i$ 's. If the inequality constraint between the variances of  $\bar{x}_i$ 's is not known, then one can use test 5.

We also compared the powers of all these tests at different sample size configurations. They are not reported here since they all exhibited almost the same pattern as the powers in Tables 1 and 2.

**Remark 3.1.** In the interblock analysis of a BIBD, it is known that the intrablock variance is smaller than the interblock variance and the degrees of freedom corresponding to the former are larger than those of the latter. In other words,  $\sigma_1^2 < \sigma_2^2$  and  $\sigma_1^2/n_1 < \sigma_2^2/n_2$ . So, in this context only a little gain in power can be expected from using test 7 instead of test 6. This also explains Remark 2.1 of Mathew et al. (1993) to some extent.

**Remark 3.2.** We note that the idea of this article will clearly work in other situations as well. For example, the six combined tests proposed by Zhou and Mathew (1993b) for the multivariate case can be improved by incorporating sample sizes in the weights used to combine independent Hotelling  $T^2$  tests. Also, appropriate modification to the combined tests considered in Mathew et al. (1993) will yield better tests.

## References

- Cohen, A. and H.B. Sackrowitz (1984), Testing hypotheses about the common mean of normal distributions, *J. Statist. Plann. Inference* **9**, 207–227.
- Cohen, A. and H.B. Sackrowitz (1989), Exact tests that recover interblock information in balanced incomplete block designs, *J. Amer. Statist. Assoc.* **84**, 556–559.
- Mathew, T., B.K. Sinha and L. Zhou (1993), Some statistical procedures for combining independent tests, *J. Amer. Statist. Assoc.* **88**, 912–919.
- Zhou, L. and T. Mathew (1993a), Combining independent tests in linear models, *J. Amer. Statist. Assoc.* **88**, 650–655.
- Zhou, L. and T. Mathew (1993b), Combining independent tests in multivariate linear models, submitted for publication.