The Korean Communications
in Statistics Vol. 8, No. 1, 2001
pp. 219-227

Non-Conservatism of Bonferroni-Adjusted Test

Gyeong Bae Jeon [1] and Sung Duck Lee [2]

Abstract

Another approach (multi-parameter measurement method) of interlaboratory studies of test methods is presented. When the unrestricted normal likelihood for the fixed latent variable model is unbounded, we propose a method of restricting the parameter space by formulating realistic alternative hypothesis under which the likelihood is bounded. A simulation study verifies the claim of conservatism of level of significance based on assumptions about central chi-square distributed test statistics and on Bonferroni approximations. We showed a randomization approach that furnished empirical significance levels would be better than a Bonferroni adjustment.

Keywords : Interlaboratory studies, Bonferroni-adjusted Test, Maximum likelihood. Empirical distribution, Latent variable model.

1. Introduction

There are two ways in which interlaboratory studies are conducted. The first is designed to monitor laboratories to see good agreement between the results obtained by different laboratories, by allowing each laboratory to compare its results with those obtained by other laboratories and take remedial action, if necessary. It is offen referred to as "proficiency testing." Here, the laboratorise themselves are of primary concern. The second is concerned not so much with the laboratories as with the method of measurement. Tests performed with presumably identical materials, in presumably identical circumstances usually do not yield identical results. This is attributed to unavoidable random error inherent in every test procedure; the factors that may influence the outcome of a test cannot all be completely controlled. Many different factors contribute to the variability in application of a test method, including the operator, equipment used, calibration of the equipment, and environment.

[1] Researcher, Bank of Korea, Seoul, Korea
E-mail : gjeon@channeli.net

[2] Professor, Dept. of Statistics, Chungbuk National University, Cheongju, Korea, 361-763
E-mail : sdlee@cbucc.chungbuk.ac.kr

Committee Ell of the American Society for Testing Materials(ASTM) has issued a software program with its standard E691 deals with the evaluation of methods of measurement in terms of reproducibility and repeatability. As a supplement to E691, J. Mandel proposed additional calculations using an unweighted average to estimate the materials's property and fitting a row-linear model to examine what mathematical model underlies the data. Proctor(1991) used nonlinear, generalized least squres to fit to sample covariances of the latent model(2.2) below. Fuller(1987) also noted that the model (2.2) was the psychometric single factor model. His development of maximum likelihood estimators follows closely that of Lawley and Maxwell(1971). Pantula and Fuller(1986) derived algorithm computing the maximum likelihood estimator and the estimated covariance matrix of the estimators of the factor model under factor vector is distributed as an normal random vector.

Maximum likelihood tests for difference among variances poses a challenge in that the likelihood can easily become unbounded. While we cannot apply the method of maximum likelihood estimate to the multivariate normal distribution of the observations to find estimates of error variances, we provide a alternative way of estimating parameters. We formulated 8 models for hypothesis tests and supplied test methods by likelihood ratio tests to see differences in means and slopes and variances.

This difficulty was overcome by finding a test statistic based on a difference in two likelihoods, both from estimates interior to parameter space. The resulting test statistic had a distribution not easily characterized and simulations were needed. Thus the major objective become to provide a useful test of variance equality.

In this paper, we consider the latent variable models for interlaboratory studies setting and derive likelihood function in Section 2 and propose the Bonferroni multiple comparison procedure for several simultaneous tests of significance in relatively nonstandard situations in Section 3. In Section 4, we also give an alternative method of formulating hypotheses in which the laboratories are segregated into subgroups defined by commonality of variances. Furthermore, we give simulations studies for the verifications and test modification in Section 5 and 6, respectively. Then concluding remarks are given in Section 7.

2. Latent Variable Model

A classical measurement error model decomposes the recorded measurement X into a true value T and a random error E, i.e.,

$$X = T + E, \tag{2.1}$$

where E $(E) = 0$ and Var(E) = σ^2, a constant.

Elaborations of the model arise from specifying the true value T to accommodate a number of materials and a number of laboratories along with allowing for non-constant variance. We consider the latent variable models for interlaboratory studies setting (ASTM, 1992).

The model that emerges is:

$$X_{ij}=\mu_i+\beta_i\tau_j+\varepsilon_{ij} \tag{2.2}$$

where X_{ij} denote observation with $i=1,\cdots,L$ indexing laboratories or tests and $j=1,\cdots,M$ indexing subjects or materials. The random ε_{ij} have zero mean and variance σ_i^2, and τ_j are taken as fixed quantities with zero mean and unit finite population variance. The μ_i and β_i may similarly be taken as fixed.

With the base model given by (2.2), eight variations on the model concerning the equality of the μ_i's, β_i's and σ_i^2's are considered.

To determine estimates of the unknown parameters, we use maximum likelihood which minimizes -log L , where

$$L(\mu_1,\cdots,\mu_L,\beta_1,\cdots,\beta_L,\tau_1,\cdots,\tau_M,\sigma_1^2,\cdots,\sigma_L^2)$$
$$=\frac{1}{(2\pi)^{\frac{LM}{2}}}\prod\frac{1}{(\sigma_i^2)^{\frac{2}{M}}}\exp\left\{-\frac{1}{2}\sum_{j=1}^{M}\frac{(X_{ij}-\mu_i-\beta_i\tau_j)^2}{\sigma_i^2}\right\}$$

(2.3)

This L is the likelihood function obtained under the assumption that the observations have a multivariate normal distribution. Anderson and Rubin (1956) showed that the likelihood function does not have a maximum. To show this fact, note that if $\mu_L=0$, $\beta_L=1$ and $\tau_j=X_{Lj}$ then $(X_{Lj}-\mu_L-\beta_L\tau_j)$ is zero so σ_L^2 can be made arbitrarily small without changing the quantity in the exponent of (2.3). As σ_L^2 approaches zero, the likelihood is unbounded. Thus the likelihood function has no maximum, and maximum likelihood estimates do not exist.

Differences among the μ_i's which might be called additive biases are widespread and can be easily explained by calibration problems. The differences by β_i's which are sometimes called scale biases also arise from relatively simple conditions. The difference among the σ_i^2's come from a complex of activities and are perhaps best considered last.

The first alternative hypothesis of interest is significance in biases μ_i, the second one is about slopes β_i and finally about error variances σ_i^2's. Proceeding in this sequence we stop whenever a failure to reject occurs.

3. Bonferroni Multiple Comparison Procedure

The Bonferroni multiple comparison procedure is customarily used when doing several simultaneous tests of significance in relatively nonstandard situations in which other methods do not apply. When one wishes to simultaneously test each of several hypotheses at a common significance level α', the generalized Type I error probability α, (i.e., the probability of rejecting at least one of those hypotheses being tested that is in fact true), is typically much in excess of α'. Thus it is usually desirable to ensure a pre-selected value for α by using a multiple comparisons testing procedure for conducting the multiple tests. Let $\chi^2_{(1)}, \cdots, \chi^2_{(R)}$ be the set of R statistics with corresponding p-values, $P_1, \cdots, P_R$ for testing hypotheses $H_{(1)}, \cdots, H_{(R)}$.

The Bonferroni multiple test procedure is performed by rejecting $H_0 = \bigcap H_i$ if any p-value is less than $\frac{\alpha}{R}$. Furthermore the specific hypothesis $H_{(i)}$ is rejected for each $P_i \leq \frac{\alpha}{R}$ $(i=1, \cdots, R)$. The Bonferroni inequality,

$$\Pr\left\{\bigcup\left(P_i \leq \frac{\alpha}{R}\right)\right\} \leq \alpha \qquad (0 \leq \alpha \leq 1),$$

ensures that the probability of rejecting at least one hypothesis when all are true is no greater than α.

A criticism of the classical Bonferroni test procedure is that it is too conservative for highly correlated test statistics. Sidak(1968, 1971) has shown that the significance level for each test $\frac{\alpha}{R}$ can be improved by using $1-(1-\alpha)^{\frac{1}{R}}$ which is less conservative. But the degree of improvement for $R<10$ and $\alpha=0.05$ is slight.

4. Alternative Method of Formulating Hypothesis

The likelihood with all variances equal is well behaved and the calculation of maximum likelihood estimates is relatively routine. The likelihood with unequal, unconstrained variances is unbounded with some variance estimates falling on the boundary of the parameter space ($\sigma_i^2 = 0$ for some i). Thus it is not possible to obtain a meaningful test of

$$H_0\colon\ \sigma_1^2 = \cdots = \sigma_L^2$$

$$H_1\colon \text{ not all } \sigma_i^2 \text{ equal.}$$

However, in interlaboratory study applications it is unlikely that the full generality inherent in H_1 is necessary. Likely departures from H_0 are those in which the laboratories are segregated into subgroups defined by commonality of variances.

The expectation that variances will likely segregate into such groups provides the rational for constructing various null and alternative hypotheses which incorporate grouping of the variances. These have the advantages of both practical relevance and boundedness of the likelihood provided all of the groups contain at least two laboratories.

We pair two laboratories and assume them to have a common variance while the other variances are free to differ. Denote these laboratories as *a1* and *a2* and their common variance as $\sigma^2_{(a1,a2)}$. Let σ^2_0 denote the common variance when the σ^2_i's are all equal. The null and alternative hypotheses are formulated as

$$H_0: \sigma^2_1 = \sigma^2_2 = \cdots = \sigma^2_L = \sigma^2_0$$

$$\text{and } H_1: \ \sigma^2_i \geq \sigma^2_{(a1,a2)}, \quad \text{for } i \neq a1, \ a2$$

where $\sigma^2_{(a1,a2)}$ is the common variance of *a1* and *a2*.

For some data observations, convergence may not be achieved if the data suggest that another laboratory has a smaller variance than the initially designated common variance $\hat{\sigma}^2_{(a1,a2)}$ of the paired laboratories. If this happens, the likelihood is unbounded and a minimum under the alternative hypothesis H_1 does not exist. Thus we would run all possible pairs $\left(R = \frac{L(L-1)}{2}\right)$ and find pairs satisfying the constraints $\sigma^2_i \geq \sigma^2_{(a1,a2)}$ for $i \neq a1, \ a2$. If more than one pair converged, we would pick the one that has the largest χ^2 or the smallest p-value. We would thus apply the Bonferroni multiple comparison test at significance level $\frac{2\alpha}{L(L-1)}$ for the alternative Method.

5. Simulation Study: Verification Phase

A simulation study was done to verify the claim of conservatism of level of significance based on assumptions about central chi-square distributed test statistics and on Bonferroni approximations.

To describe the asymptotic behavior of the likelihood ratio statistics and thereby determine the actual size of the tests for equal variances, we can simulate their behavior under the null hypothesis.

We used four sets of parameter values in (2.2):

Case 1: Equal biases and equal slopes (L=6, M=14)
Case 2: Equal biases and unequal slopes (L=6, M=14)
Case 3: Unequal biases and equal slopes (L=6, M=14)
Case 4: Unequal biases and unequal slopes (L=6, M=14).

We examined the Bekk data from Mandel and Lashof (1959). Fourteen laboratories participated in the interlaboratory study and each made 8 replicate measurements on each of 14 paper materials. Fuller (1987) analyzed the logarithms of the averages of the replication for six laboratories and we will work with these same data.

Under the null hypothesis H_0, we generated *NS*=1000, *s*=1, ⋯, *NS*, data sets for *L*=6. From the 1000 simulated data sets, likelihood ratio statistics $T_k^{(s)} = 2\{ L_k^{(s)}(\widehat{\omega}) - L_k^{(s)}(\widehat{\Omega})\}$ for s = 1, ⋯, 1000 and separately for the k = 1, ⋯, 15 combinations of pair along with its p-value, $P_{T_k^{(s)}} - value$, determined as $\left[1 - F_{\chi_1^2}(T_k^{(s)})\right]$ were calculated to study likelihood ratio statistics T_k, k = 1, ⋯, 15, have level- α tests. The actual level of significance of the test for equal variances was estimated as the proportion of replications which rejected the null hypothesis H_0 at significance level α. Define the true level of significance $P_{T_k}(\alpha) = \Pr(P_{T_k} - value < \alpha)$. The true level was estimated by

$$\widehat{P}_{T_k}(\alpha) = \frac{(P_{T_k} - value)}{15000} . \tag{5.1}$$

These estimated levels were compared with the nominal level of the test for equal variances. The test of equal variances was performed at the α = 0.10, 0.05, 0.025 and 0.01 significance levels. Table 1 shows that the estimated actual levels exceed the nominal levels.

We also calculated $T_{\max}^{(s)} = \max(T_k^{(s)},\ k = 1, \cdots, 15)$ along with its Bonferroni adjusted p-value $BAP^{(s)}$ to study whether the test statistic $T_{\max}$ is approximately level α. The $BAP^{(s)}$ were calculated by $15[1 - F_{\chi_1^2}(T_{\max}^{(s)})]$ for s=1, ⋯,1000 simulated data sets under the null hypothesis H_0. The actual level of significance of the test for equal variances was estimated as the proportion of replicates that rejected the null hypothesis H_0 at significance level α. Define the true level of significance $P_{BAP}(\alpha) = \Pr(BAP < \alpha)$. The true level was estimated by

$$\widehat{P}_{BAP}(\alpha) = \frac{(BAP^{(s)} < \alpha)}{1000} . \tag{5.2}$$

We had expected that $P_{BAP}(\alpha)$ will be less than or equal to α, but we found that estimated levels are above the true levels from Table 2. The Bonferroni correction is not conservative in this case. That is, the actual levels exceeded the Bonferroni corrected nominal levels.

Actually, the data suggest that there are bias and slope differences. Table 3 shows that the observed test statistic $T_{\max}$ of 16.718 is not significant at the 5% level. In fact, it has an empirical significance level of 12.8%. This is found by calculating $\frac{1}{1000} \sum_{s=1}^{1000} (T_{\max}^{(s)} > 16.718)$. That is, the actual levels of significance are quite far above the Bonferroni corrected nominal levels. These actual levels when slopes and biases are unequal are even more inflated than when slopes and biases are equal. This suggests that the T_k does not have a chi-square distribution under H_0.

Table 1. Proportion of rejections of H_0, $\hat{P}_{T_k}(\alpha)$, with 95% confidence limits. (L=6,NS=1000)

α	Case 1	Case 2	Case 3	Case 4
0.010	0.028 (0.025, 0.031)	0.035 (0.032, 0.038)	0.036 (0.033, 0.039)	0.047 (0.044, 0.050)
0.025	0.049 (0.046, 0.053)	0.061 (0.057, 0.065)	0.067 (0.063, 0.071)	0.079 (0.075, 0.083)
0.050	0.083 (0.079, 0.087)	0.093 (0.088, 0.098)	0.105 (0.100, 0.110)	0.114 (0.109, 0.119)
0.100	0.132 (0.127, 0.137)	0.142 (0.136, 0.148)	0.154 (0.148, 0.160)	0.164 (0.158, 0.170)

Table 2. Proportion of rejections of H_0 , $\hat{P}_{BAP}(\alpha)$, with 95% confidence limits. (L=6,NS=1000)

α	Case 1	Case 2	Case 3	Case 4
0.010	0.041 (0.029, 0.053)	0.049 (0.036, 0.062)	0.043 (0.030, 0.056)	0.084 (0.067, 0.101)
0.025	0.066 (0.051, 0.081)	0.077 (0.061, 0.094)	0.074 (0.058, 0.090)	0.112 (0.102, 0.122)
0.050	0.090 (0.072, 0.108)	0.110 (0.091, 0.129)	0.114 (0.094, 0.134)	0.157 (0.134, 0.180)
0.100	0.129 (0.125, 0.150)	0.146 (0.124, 0.168)	0.161 (0.138, 0.184)	0.202 (0.177, 0.226)

Table 3. Intended and estimated true levels for $T_{\max}$. (NS=1000)

Intended true level α	Estimated true level $\hat{P}_{BAP}(\alpha)$	$\chi^2_{1-\frac{\alpha}{15}}(4)$	Observed $T_{\max}$
0.01	0.084	19.364	16.718
0.025	0.112	17.331	
0.05	0.157	15.777	

Table 4. Estimated proportion of rejection, $\hat{P}_{Boot}(\alpha)$. (NS=200)

$P_{Boot}-value<\alpha$	$\hat{P}_{Boot}(\alpha)$	s.e.($\hat{P}_{Boot}(\alpha)$)
$P_{Boot}-value<0.01$	0.010	0.0070
$P_{Boot}-value\leq 0.01$	0.020	0.0099
$P_{Boot}-value<0.05$	0.045	0.0147
$P_{Boot}-value\leq 0.05$	0.045	0.0147
$P_{Boot}-value<0.1$	0.090	0.0202
$P_{Boot}-value\leq 0.1$	0.105	0.0217

6. Simulation Study: Test Modification Phase

The Bonferroni adjustment does not always work and thus we turn to empirical significance levels based on simulations. Replicated values of $T^{(s)}_{\max}$, $s = 1, \cdots, NS$, evaluated from the successive bootstrap samples can be used to approximate the null distribution of $T_{\max}$ and this enables an approximation of the particular empirical level of significance corresponding to the value of $T_{\max}$ evaluated from the original sample.

The approach involved the simulation of 200 sets of data (NS=200) under the null hypothesis H_0 as before by (2.2). For each simulated data set, we calculated $T^{(s)}_{\max}$ and its $P_{T_k^{(s)}}-value$ from chi-square distribution and performed q=1, ⋯, 100 bootstrap replicates (NQ=100). These 100 data sets were generated using the biases and slopes from the s^{th} generated data set and also used its pooled variance. And we calculated $T^{(s,q)}_{\max}$=max($T_k^{(s,q)}, k=1,\cdots,15$) for q=1, ⋯, 100 and for each s and $P_{Boot^{(s)}}-value$ by $\frac{1}{100}\sum_{q=1}^{100} I(T^{(s,q)}_{\max} > T^{(s)}_{\max})$. We compare the frequency distribution of 200 $P_{Boot}-value$ grouped into 10 classes, with the theoretical distribution that $P_{Boot}-value$ should follow if the error distribution is uniform. We note that the agreement between the observed (f_c) and the expected frequencies (F_c) appears good. The test criterion to apply the χ^2 test of goodness of fit is $\chi^2=\sum_{c=1}^{10}\frac{(f_c-F_c)^2}{F_c}$. The χ^2=7.4 which is less than $\chi^2_{0.05}$ =16.92 with 9 degrees of freedom, and so there is not sufficient evidence to reject at the 5% level. But $\chi^2=\sum_{c=1}^{15}\frac{(f_c-F_c)^2}{F_c}$=545.2 of the Bonferroni adjusted p-value(BAP) is rejected at the 5% level and suggest departure from a uniform distribution. The significance level based on $P_{Boot}-value$ seems close to the nominal ones but those based on BAP are excessive. This would cause too many false positives by too often finding variances to be unequal when they were not.

7. Conclusion

We fitted a multi-parameter measurement model with the material true value as a fixed latent variable by maximum likelihood estimation. We found that maximum likelihood estimation of unequal, unconstrained error variances leads to an unbounded likelihood. We formulated and studied a alternative hypothesis under which the likelihood is bounded. However, the resulting test statistics did not have central chi-square distributions. The significance levels under central chi-square assumption and with a Bonferroni adjustment for multiple testing were found by simulations to be greater than nominal levels. This result suggested that a randomization approach that furnished empirical significance levels would be better than a Bonferroni adjustment, and we verified this by further simulations.

References

[1] Anderson, T. W. and Rubin, H. (1956). Statistical inference in factor analysis. Proceedings of the Third Berkeley Symposium. vol. V. University of California Press, Berkeley.

[2] ASTM (1992). Standard practice for conducting an interlaboratory study to determine the precision of a test method, designation: E 691-92 ASTM Standards on Precision and Bias for Various Applications. (4th ed.). American Society for Testing and Materials.

[3] Fuller, W. A. (1987). Measurement Error Models. John Wiley & Sons. New York. NY. [4.] Lawley, D. N. and Maxwell, A. E.(1971). Factor Analysis as a Statistical Method, American Elsevier, New York.

[5] Mandel, J. and Lashof, T. W. (1959). The Interlaboratory Evaluation of Testing Methods, ASTM Bulletin, ASTM Headquarters, Philadelphis, PA No. 239, pp. 53-61. [6.] Pantula, S. G. and Fuller, W. A.(1986). Computational algorithm for the factor model, Comm. in statistics, Simulation, Vol.15 (1) 227-259.

[7] Proctor, C. H.(1991). Quick fitting covariance models to interlaboratory trial data. Paper given in Atlanta, 11, ASA.

[8] Sidak, Z. (1968). On Multivariate Normal Probabilities of Rectangles: Their Dependence on Correlations. Ann. Math. Statist. 39, 1425-34.

[9] Sidak, Z (1971). On Probabilities of Rectangles in Multivariate Student distributions: Their Dependence on Correlation. Ann. Math. Statist. 42, 169-75.