Comparison of Tests for the ANOVA with Unequal Variance

Evren ÖZKİP, Berna YAZICI, Ahmet SEZER
Department of Statistics, Science Faculty, Anadolu University, Eskisehir, TURKEY.

Abstract

Most of the studies in practice concern comparison of the difference of group means. The goal of this study is testing the equality of the means of normal population when the variances are unequal. We compare the performance of the proposed tests such as the generalized *F*-test (GF), the parametric bootstrap test (PB) and test based on fiducial *p*-value (FP). Monte Carlo simulation studies are conducted to compare the empirical size and power of these tests.

Keywords: Generalized *F*-test, parametric bootstrap, fiducial *p*-value, empirical size, power

1. Introduction

Testing the equality of the means of normal populations when the variances are unknown and unequal is a fundamental problem in clinical trials and biomedical research. This problem, when only two normal means are involved, is referred to as the Behrens-Fisher problem. In the classical treatment of the problem, the homoscedasticity is usually made for convenience and mathematical tractability rather than anything else. The current literature on this problem does not provide standard statistical testing procedure. For example, the classical *F* test fail to reject the null hypothesis even for large samples when the population variances are unequal. There are numerous solutions for testing equality of means for normal data under heterogeneity, such as Welch's (1951) approximate test, James's (1951) second-order test, Brown and Forsythe's (1974) test, Weerahandi's (1995) generalized *F* test, Krishnamoorthy et al.'s (2007) parametric bootstrap test, Xu and Wang's (2008) test, Li et al.'s (2011) fudicial test and so on.

Recent advances in statistical computation have made a tremendous positive impact on fundamental sciences. In this study, we proposed some tests based on Monte Carlo simulation such as the generalized *F*-test (GF), the parametric bootstrap test (PB) and test based on fiducial *p*-value (FP). An simulation study was conducted to comparison the size and powers of these three tests.

2. Problem and Basic Solution

Assume $X_{i1}, \dots, X_{in_i}$, $i = 1, \dots, k$ are k sets of random samples independently generated from the normal population $N(\mu_i, \sigma_i^2), 1 \leq i \leq k$. Let $\bar{X}_i$ and S_i^2 be the sample mean and variance, respectively. That is,

$$\bar{X}_i = \frac{1}{n_i}\sum_{i=1}^{n_i} X_{ij} \text{ and } S_i^2 = \frac{1}{(n_i-1)}\sum_{j=1}^{n_i}\left(X_{ij} - \bar{X}_i\right)^2, i = 1, \dots, k$$

Denote $\mu = (\mu_1, \dots, \mu_k), \sigma^2 = \left(\sigma_1^2, \dots, \sigma_k^2\right)$, $\bar{\boldsymbol{X}} = (\bar{X}_1, \dots, \bar{X}_k)$ and $\boldsymbol{S^2} = \left(S_1^2, \dots, S_k^2\right)$. The problem of interest can be formulated as the following hypothesis:

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k = \mu \qquad H_1: \mu_i \neq \mu_j \text{ for some } \; i \neq j \tag{1}$$

When all σ_i^2s are known, it is well known that

$$T(\bar{\boldsymbol{X}}; \sigma^2) = \sum_{i=1}^{k} \frac{n_i}{\sigma_i^2}\left\{(\bar{X}_i - \mu_i) - \frac{\sum_{i=1}^{k} n_i(\bar{X}_i - \mu_i)/\sigma_i^2}{\sum_{i=1}^{k} n_i/\sigma_i^2}\right\}^2 \sim \chi^2_{(k-1)} \tag{2}$$

$T(\bar{\boldsymbol{X}}; \sigma^2)$ therefore can be used for testing hypothesis (1) and the corresponding *p*-value can easily be obtained.

When population variances σ_i^2s are equal unknown, a test statistic can be obtained by replacing σ_i^2 by S_i^2, $i = 1, \dots, k$, and is given by

$$T\left(\bar{\boldsymbol{X}}; \boldsymbol{S^2}\right) = \sum_{i=1}^{k} \frac{n_i}{S_i^2}\left\{(\bar{X}_i - \mu_i) - \frac{\sum_{i=1}^{k} n_i(\bar{X}_i - \mu_i)/S_i^2}{\sum_{i=1}^{k} n_i/S_i^2}\right\}^2 \tag{3}$$

which can be simplified as

$$T\left(\bar{\boldsymbol{X}}; \boldsymbol{S^2}\right) = \sum_{i=1}^{k} \frac{n_i}{S_i^2}\left\{(\bar{X}_i) - \frac{\sum_{i=1}^{k} n_i(\bar{X}_i)/S_i^2}{\sum_{i=1}^{k} n_i/S_i^2}\right\}^2 \tag{4}$$

when H_0 is true.

3. Recently Proposed Test

In this section, three solutions for the problems of testing several means have been developed to overcome some drawbacks in literature; i.e., the GF test by Weerahandi (1995), the parametric bootstrap (PB) test by Krishnamoorthy et al. (2007) and test based on fiducial *p*-value (FP) test by Li et al. (2011). We briefly review these three solutions.

3.1. The Generalized F (GF) Test

Weerahandi (1995a) proposed the following generalized test variable

$$T_{GF} = \frac{T(\bar{\boldsymbol{X}}; \sigma^2)}{T(\bar{x}; s_1^2\sigma_1^2/S_1^2, \dots, s_k^2\sigma_k^2/S_k^2)} \tag{5}$$

where

$$T\left(\bar{\boldsymbol{X}}; \boldsymbol{S^2}\right) = \sum_{i=1}^{k} \frac{n_i}{S_i^2}\left\{(\bar{X}_i) - \frac{\sum_{i=1}^{k} n_i(\bar{X}_i)/S_i^2}{\sum_{i=1}^{k} n_i/S_i^2}\right\}^2 \tag{6}$$

under the null hypothesis. Note that $U_i = (n_i - 1)S_i^2/\sigma_i^2$, are independently chi-squared variables with degrees of freedom $(n_i - 1)$ for $i = 1, \dots, k$, respectively. Then the generalized *p*-value is defined as

$$p = Pr(T_{GF} \geq t_{obs}) = Pr\left[(\bar{\boldsymbol{X}}; \sigma^2) \geq T\{\bar{x}; (n_1 - 1)s_1^2/U_1, \dots, (n_k - 1)s_k^2/U_k\}\right]$$
$$= 1 - E\left[\chi^2_{(k-1)}\left[T\{\bar{x}; (n_1 - 1)s_1^2/U_1, \dots, (n_k - 1)s_k^2/U_k\}\right] \setminus U_1, \dots, U_k\right] \quad (7)$$

where t_{obs} is the observed value of T_{GF} at $(\bar{\boldsymbol{X}}; \boldsymbol{S^2}) = (\bar{x}; s^2)$ and is actually equal to 1 and $\chi^2_{(k-1)}$ denotes the cumulative distribution function of χ^2 distribution with $k - 1$ degrees of freedom. The GF test rejects the null hypothesis in (1) whenever the generalized *p*- value in (7) is less than a given nominal level.

3.2. The Parametric Bootstrap (PB) Test

The parametric bootstrap (PB) involves sampling from the estimated models. That is, samples or sample statistics are generated from parametric models with the parameters replaced by their estimates. Recall that under $H_0: \mu_1 = \cdots = \mu_k$ all X_i's have the same mean. As the test statistic T_N in (4) is location invariant, without loss of generality, we can take this common mean to be zero. Using these facts, the parametric bootstrap *pivot variable* can be developed as follows.

Let $\bar{X}_{Bi} \sim N\left(0, \frac{S_i^2}{n_i}\right)$ and $S_{Bi}^2 \sim \frac{\chi^2_{n_i-1}}{(n_i-1)}$, $i = 1, \dots, k$. Then the PB pivot variable based on the test statistic is given by

$$\tilde{S}_b = \tilde{S}_b(\sigma_1^2, \dots, \sigma_k^2) = \sum_{i=1}^{k} \frac{n_i \bar{X}_i^2}{S_i^2} - \frac{\left(\sum_{i=1}^{k} \frac{n_i \bar{X}_i^2}{S_i^2}\right)^2}{\sum_{i=1}^{k} \frac{n_i}{S_i^2}} \quad (8)$$

Noticing the fact that X_{Bi} is distributed as $Z_i(S_i/\sqrt{n_i})$, where Z_i is a standard normal random variable, it can be easily verified that the PB pivot variable in (8) is distributed as

$$\tilde{S}_{bB}(Z_i, \chi^2_{n_i-1}; S_i^2) = \sum_{i=1}^{k} \frac{Z_i^2(n_i - 1)}{\chi^2_{n_i-1}} - \frac{\left[\sum_{i=1}^{k} \frac{\sqrt{n_i} Z_i (n_i - 1)}{S_i^2 \chi^2_{n_i-1}}\right]^2}{\sum_{i=1}^{k} \frac{n_i(n_i - 1)}{S_i^2 \chi^2_{n_i-1}}} \quad (9)$$

For a given $(s_1^2, \dots, s_k^2)$ of $(S_1^2, \dots, S_k^2)$ and level α, the PB test rejects H_0 in (1) when

$$P\{\tilde{S}_{bB}(Z_i, \chi^2_{n_i-1}; s_i^2) > \tilde{s}_b\} < \alpha \quad (10)$$

where T_{No} is an observed value of T_N (Krishnamoorthy, Lu and Mathew 2007).

3.3. The Fiducial Approach (FP) Test

Li, Wang and Liang (2011) developed new test for (1) by using the concept of fiducial and generalized *p*-value approach (FG).

Let $U_{1i}\sim N(0,1)$, $U_{2i}\sim\chi^2_{n_i-1}$, $i=1,2,\dots,k$ and be mutually independent. Note that $\bar{X}_i\sim N\left(\mu_i,\frac{\sigma_i^2}{n_i}\right)$, $(n_i-1)\,S_i^2\sim\chi^2_{n_i-1}\sigma_i^2$ for $i=1,2,\dots,k$ and these statistics are all mutually independent. We therefore can express $\bar{X}_i$ and $(n_i-1)S_i^2$ as functions of U_{1i} and U_{2i}; i.e.,

$$\bar{X}_i=\mu_i+\frac{\sigma_i^2}{n_i}U_{1i} \text{ and } (n_i-1)S_i^2=\sigma_i^2U_{2i}\,,\ i=1,2,\dots,k$$

Given an observation $\left(\bar{x}_i,s_i^2\right)$ and (u_{1i},u_{2i}), the equations of $\bar{x}_i=\mu_i+\frac{\sigma_i^2}{n_i}u_{1i}$ and $(n_i-1)\,s_i^2=\sigma_i^2u_{2i}$ have the unique solutions

$$\mu_i=\bar{x}_i-\frac{u_{1i}}{\sqrt{u_{2i}/(n_i-1)}}\sqrt{\frac{s_i^2}{n_i}}\ and\ \sigma_i^2=\frac{(n_i-1)s_i^2}{u_{2i}} \tag{11}$$

hence for given $\left(\bar{x}_i,s_i^2\right)$, the fiducial distribution of μ_i is the same as that of $T_{\mu_i}=\bar{x}_i-t_i\sqrt{s_i^2/n_i}$ $i=1,2,\dots,k$ here $t_i\sim t(n_i-1)$, $i=1,2,\dots,k$ and they are mutually independent. Then the fiducial distribution could be derived by

$$T_F(t;s^2)=\sum_{i=1}^{k}t_i^2-\frac{\left(\sum_{i=1}^{k}\frac{\sqrt{n_i}}{s_i}t_i\right)^2}{\sum_{i=1}^{k}\frac{n_i}{s_i^2}} \tag{12}$$

Here $t=(t_1,\dots,t_k)$. Because $T(x;s^2)$ is the observed value of T_F under the null hypothesis, the *p*-value for (1) is given by

$$p=Pr\{T_F\geq T(x;s^2)\} \tag{13}$$

Accordingly, we reject the null hypothesis when $p<\alpha$ for a given level α (Li, Wang and Liang 2011).

4 Simulation study

This section provides simulation studies for type I error probabilities and powers of the five methods proposed in Section 3. In this study, two configuration factors were taken into account to evaluate the performances of type I error probabilities and powers; sample size and variance.

Table 1. Type I error rates for the proposed tests.

k=3	n=(5,5,5)			n=(10,10,10,10)			n=(10,15,20,25)			n=(25,20,15,10)		
$(\sigma_1^2, \sigma_2^2, \sigma_3^2)$	GF	PB	FP	GF	PB	FP	GF	PB	FP	GF	PB	FP
(1,1,1)	0.075	0.044	0.030	0.072	0.054	0.038	0.058	0.054	0.049	0.054	0.045	0.026
(0.5,1,1.5)	0.076	0.048	0.023	0.064	0.054	0.030	0.058	0.049	0.047	0.061	0.046	0.019
(1,2,3)	0.076	0.048	0.023	0.064	0.054	0.030	0.058	0.049	0.047	0.061	0.046	0.019
(1,2,4)	0.078	0.049	0.018	0.066	0.057	0.028	0.060	0.052	0.039	0.064	0.046	0.017
$k = 4$	n = (5,5,5,5)			n = (10,10,10,10)			n = (10,15,20,25)			n = (25,20,15,10)		
$(\sigma_1^2, \dots, \sigma_4^2)$	GF	PB	FP	GF	PB	FP	GF	PB	FP	GF	PB	FP
(1,1,1,1)	0.098	0.037	0.021	0.076	0.055	0.046	0.070	0.058	0.051	0.068	0.053	0.044
(0.5,1,1.5,2)	0.097	0.043	0.029	0.074	0.050	0.043	0.068	0.049	0.044	0.078	0.056	0.048
(1,2,3,4)	0.097	0.043	0.029	0.074	0.050	0.043	0.068	0.049	0.044	0.078	0.056	0.048
(1,2,4,8)	0.090	0.046	0.031	0.076	0.050	0.043	0.064	0.051	0.047	0.080	0.055	0.054
$k = 5$	n = (5,5,5,5,5)			n = (10,10,10,10,10)			n = (10,15,20,25,30)			n = (30,25,20,15,10)		
$(\sigma_1^2, \dots, \sigma_4^2)$	GF	PB	FP	GF	PB	FP	GF	PB	FP	GF	PB	FP
(1,1,1,1,1)	0.094	0.046	0.025	0.079	0.055	0.046	0.074	0.056	0.052	0.066	0.046	0.040
(0.5,1,1.5,2,2.5)	0.104	0.053	0.030	0.077	0.044	0.037	0.067	0.053	0.048	0.066	0.049	0.045
(1,2,3,4,5)	0.104	0.053	0.030	0.077	0.043	0.037	0.067	0.053	0.047	0.070	0.049	0.045
(1,2,4,8,12)	0.102	0.054	0.031	0.072	0.043	0.036	0.067	0.052	0.047	0.069	0.049	0.044

Table 2. Powers for the proposed tests.

$k = 3$		$n = (10,10,10)$			$n = (5,10,15)$			$n = (10,20,30)$		
$(\sigma_1^2, \dots, \sigma_3^2)$	$(\mu_1, \dots \mu_3)$	GF	PB	FP	GF	PB	FP	GF	PB	FP
(1,1,1)	(0,0,0.3)	0.115	0.094	0.084	0.115	0.077	0.056	0.177	0.153	0.139
	(0,0,0.6)	0.254	0.226	0.189	0.317	0.227	0.163	0.534	0.493	0.472
	(0,0,0.9)	0.500	0.455	0.400	0.582	0.467	0.389	0.868	0.839	0.818
	(0,0,1.2)	0.726	0.690	0.653	0.816	0.731	0.659	0.991	0.986	0.982
	(0,0,1.5)	0.903	0.884	0.849	0.956	0.902	0.857	1.00	1.00	1.00
(0.2,0.4,0.6)	(0,0,0.3)	0.157	0.138	0.121	0.188	0.143	0.093	0.325	0.295	0.268
	(0,0,0.6)	0.453	0.408	0.367	0.553	0.494	0.397	0.869	0.839	0.820
	(0,0,0.9)	0.778	0.746	0.714	0.882	0.836	0.776	1.00	1.00	1.00
	(0,0,1.2)	0.958	0.946	0.932	0.99	0.984	0.975	1.00	1.00	1.00
	(0,0,1.5)	0.999	0.998	0.998	0.999	0.998	0.997	1.00	1.00	1.00
$k = 4$		$n = (10,10,10,10)$			$n = (5,10,15,20)$			$n = (10,20,30,40)$		
$(\sigma_1^2, \dots, \sigma_4^2)$	$(\mu_1, \dots \mu_4)$	GF	PB	FP	GF	PB	FP	GF	PB	FP
(1,1,1,1)	(0,0,0,0.3)	0.138	0.097	0.072	0.148	0.091	0.066	0.233	0.190	0.172
	(0,0,0,0.6)	0.278	0.221	0.185	0.402	0.259	0.208	0.710	0.644	0.620
	(0,0,0,0.9)	0.514	0.431	0.386	0.742	0.613	0.538	0.971	0.960	0.952
	(0,0,0,1.2)	0.752	0.690	0.647	0.939	0.867	0.827	0.999	0.998	0.998
	(0,0,0,1.5)	0.911	0.887	0.853	0.996	0.985	0.977	1.00	1.00	1.00
(0.2,0.4,0.6,0.8)	(0,0,0,0.3)	0.146	0.113	0.093	0.137	0.125	0.095	0.329	0.282	0.261
	(0,0,0,0.6)	0.360	0.295	0.260	0.350	0.460	0.373	0.895	0.872	0.847
	(0,0,0,0.9)	0.665	0.598	0.557	0.664	0.839	0.787	0.997	0.997	0.997
	(0,0,0,1.2)	0.884	0.846	0.809	0.889	0.99	0.978	1.00	1.00	1.00
	(0,0,0,1.5)	0.970	0.960	0.949	0.977	1.00	1.00	1.00	1.00	1.00
$k = 5$		$n = (10,10,10,10,10)$			$n = (5,10,15,20,25)$			$n = (10,20,30,40,50)$		
$(\sigma_1^2, \dots, \sigma_5^2)$	$(\mu_1, \dots \mu_5)$	GF	PB	FP	GF	PB	FP	GF	PB	FP
(1,1,1,1,1)	(0,0,0,0,0.3)	0.128	0.081	0.063	0.192	0.102	0.076	0.291	0.146	0.221
	(0,0,0,0,0.6)	0.254	0.188	0.165	0.527	0.363	0.305	0.791	0.493	0.734
	(0,0,0,0,0.9)	0.518	0.417	0.366	0.862	0.749	0.706	0.990	0.877	0.985
	(0,0,0,0,1.2)	0.745	0.660	0.613	0.980	0.958	0.937	1.00	0.985	1.00
	(0,0,0,0,1.5)	0.909	0.870	0.840	0.999	0.997	0.994	1.00	1.00	1.00
(0.2,0.4,0.6,0.8,1)	(0,0,0,0,0.3)	0.131	0.083	0.067	0.199	0.122	0.085	0.325	0.160	0.257
	(0,0,0,0,0.6)	0.285	0.204	0.181	0.595	0.455	0.390	0.863	0.548	0.818
	(0,0,0,0,0.9)	0.563	0.464	0.424	0.914	0.850	0.794	0.998	0.917	0.996
	(0,0,0,0,1.2)	0.789	0.709	0.677	0.992	0.980	0.975	1.00	1.00	1.00
	(0,0,0,0,1.5)	0.925	0.902	0.874	1.00	0.999	0.999	1.00	1.00	1.00

To obtain type I error rates and powers of the GP, PB and FP tests, we use a two-step simulation. First we generated 2500 observed vectors $(\bar{x}_1, \bar{x}_2; s_1^2, s_2^2)$, and used 5000 runs for each observed vector to estimate the *p*-value in (7), (10) and (13).

5 Discussion

We compared above proposed three methods for Behrens-Fisher problem. Monte Carlo simulation conducted to compare the empirical size and power of these tests. Simulation results show that type I error of the PB test close to nominal level than the other tests. Type I error of the GF test appears to be very liberal especially when sample sizes are large. However its power is better than the other tests. Type I error of the FP appears to be conservative especially when variances are heterogeneity.

Acknowledgment

This work was supported by the Anadolu University under Grant (Number: 1202F38).

References

[1] Brown, M.B., Forsythe, A.B. (1974). The small sample behavior of some statistics which test the equality of several means. Technometrics 16, 129-132.

[2] James, G.S. (1951). The comparison of several groups of observations when the ratios of the population variances are unknown. Biometrika 38, 324–329.

[3] Krishnamoorthy, K., Lu F., Mathew, T. (2007). 'A parametric bootstrap approach for ANOVA with unequal variances: Fixed and random models.' Computational Statistics and Data Analysis 51, 5731-5742.

[4] Li, X., Wang J., Liang H. (2011). 'Comparison of several means: A fiducial based approach.' Computational Statistics and Data Analysis 55,1993-2002

[5] Weerahandi, S. (1995). ANOVA under unequal error variances. Biometrica 38, 330-336.

[6] Welch, B.L. (1951). On the comparison of several mean values: An alternative approach. Biometrica 38, 330-336.

[7] Xu, L., Wang, S. (2007). A new generalized p-value and its upper bound for ANOVA under unequal erros variances. Communications in Statistics Theory and Methods 37, 1002-1010.