GENERALIZED P VALUE TEST TO COMPARE SEVERAL POPULATION MEANS FOR UNEQUAL VARIANCES

Berna YAZICI, Ahmet SEZER, Evren ÖZKİP

Abstract

The size of classical F tests is fairly robust against the assumption of equal variances when the sample sizes are equal. However when the sample sizes are different, failure of the assumption of equal variances can have serious problems on the power of the F test. Classical F test fails to reject the null hypothesis even when the data actually provides strong evidence to do so. The generalized p-value method provides a promising approach to solve such problems with no adverse effect on the size of the test. In this study we demonstrate performance of the F test and generalized p value approach for different sample sizes and variances. Weather data will be used to compare temperature for different locations under some unequal variance situations.

Key Words: Generalized p-value, Unequal Variances, ANOVA, Brown-Forsythe test, Scoth-Smith test, Welch test.

INTRODUCTION

Classical ANOVA test is conducted to compare means of groups under the following hypothesis:

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k = \mu \qquad H_1: \mu_i \neq \mu_j,\ 1 \leq i \leq j \leq k, for\ each\ pair \qquad (1)$$

Independent samples from i^{th} population $X_{i1}, \ldots, X_{in}$ and $i = 1, \ldots, k$. The classical one-factor model without any restriction is as follows:

$$x_{ij} = \mu + \alpha_i + \varepsilon_{ij}\ , \qquad i = 1, \ldots, k,\ j = 1, \ldots, n_i$$

In order to test the hypothesis in Eq. 1, the following assumptions are required:

1. The samples come from normally distributed populations.
2. The variances are homogenous for the populations.
3. The samples are random and independent from each other.

Generally assumptions are satisfied for the large samples. But often we encounter situations where the second assumption is violated especially if the sample sizes are different. To overcome this problem variance-stabilizing transformation is applied and the analysis of variance is applied to the transformed populations.

The size of classical F tests is fairly robust against the assumption of equal variances when the sample sizes are equal. However when the sample sizes are different, failure of the assumption of equal variances can have serious problems on the power of the F test. Classical F test fails to reject the null hypothesis even when the data actually provides strong evidence to do so. The generalized p-value method provides a promising approach to solve such problems with no adverse effect on the size of the test. In this study we demonstrate performance of the F test and generalized p value approach for different sample sizes and variances. The results of Brown-Forsythe test, Scott-Smith test, and Welch test will be given for the presence of unequal variance. Weather data will be used to compare temperature for different locations. In the following section we introduce several tests which are used to deal with unequal variance.

1. Welch Test

Welch test is commonly used by the researchers for unequal variance situations since it is quite practical.

$$W = \frac{\sum_{i=1}^{k} w_i\left[(\bar{X}_i - \bar{X})^2/(k-1)\right]}{1+\frac{2(k-2)}{k^2-1}\sum_{i=1}^{k}\frac{1}{n_i-1}\left(1-\frac{w_i}{\sum w_j}\right)^2} \qquad (4)$$

where, $w_i = \frac{n_i}{s_i^2}$

The denominator of Eq. 4 is given by Eq. 5:

$$f = \frac{1}{\frac{3}{k^2-1}\sum_{i=1}^{k}\frac{1}{n_i-1}\left(1-\frac{w_i}{\sum w_j}\right)^2} \qquad (5)$$

The W value calculated by Eq. (4) has F distribution with *k-1* and *f* degrees of freedoms.

If $P\left(F_{k-1,f} > w\right) < \alpha$, then the null hypothesis is rejected.

2. Scott-Smith Test

Another test to compare the population means is given by Scott and Smith, 1971

$$F_s = \sum_{i=1}^{k} \frac{n_i(\bar{X}_i - \bar{X})^2}{S_i^{*2}} \qquad (6)$$

where S_i^{*2} is calculated using the following equation:

$$S_i^{*2} = \frac{n_i - 1}{n_i - 3} S_i^2 \qquad (7)$$

F_s, in Eq. 6 has the χ^2 distribution with *k* degrees of freedom.

If $P\left(F_s > f_s\right) < \alpha$ then the null hypothesis of equal means is rejected.

3. Brown-Forsythe Test

Another test for unequal variance case is to test null hypothesis of equality of means by Brown and Forsythe test 1974:

$$B = \frac{\sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2}{\sum_{i=1}^{k}\left(1 - \frac{n_i}{n}\right) S_i^2} \qquad (8)$$

B statistic has the *F* distribution with *k-1* and *v* degrees of freedoms where v is:

$$v = \frac{\left[\sum_{i=1}^{k}\left(1 - \frac{n_i}{n}\right) S_i^2\right]^2}{\sum_{i=1}^{k} \frac{\left(1 - \frac{n_i}{n}\right)^2 S_i^4}{n_i - 1}}$$

If $P\left(F_{k-1,v} > b\right) < \alpha$ holds then the null hypothesis is rejected.

4. Weerahandi's Generalized F test

While the right tail region is used to analyze the test statistic in classical F test, right tail sample space region is used for Generalized F test approach. Instead of

$$S_i^2 = \frac{1}{n_i - 1}\sum_{i=1}^{n_i}\left(X_{ij} - \bar{X}_i\right)^2$$

Weerahandi proposed the use of

$$S_i^2 = \frac{1}{n_i}\sum_{i=1}^{n_i}\left(X_{ij} - \bar{X}_i\right)^2$$

and test statistics β_j is defined as the function of $\frac{n_i S_i^2}{\sigma_i^2}$ statistic as given in Eq. 9.

$$B_j = \frac{\left(\sum_{i=1}^{j}\frac{n_i S_i^2}{\sigma_i^2}\right)}{\sum_{i=1}^{j+1}\frac{n_i S_i^2}{\sigma_i^2}} \qquad j = 1, \dots, k-1 \qquad (9)$$

Hence β_j has the Beta distribution;

$$B_j \sim beta\left[\sum_{i}^{j}\frac{(n_i - 1)}{2}, \frac{(n_{j+1} - 1)}{2}\right]$$

The random variable's value for i^{th} sample $\frac{n_i S_i^2}{\sigma_i^2}$ is obtained as using $\tilde{S}_e$ and β_j in Eq. 10

$$\frac{n_i S_i^2}{\sigma_i^2} = \tilde{S}_e(1 - B_{i-1})B_2 \dots B_{k-1}, \quad i = 2, \dots, k-1 \qquad (10)$$

The discrimination for $\frac{n_i S_i^2}{\sigma_i^2}$ is not dependent to any unknown parameter. So it is not affected by the acceptance of the

$$1 - E\left(H_{k-1,n-k}\left\{\frac{n-k}{k-1}\tilde{s}_b\left[\frac{n_1 s_1^2}{B_1 B_2 \dots B_{k-1}}, \frac{n_2 s_2^2}{(1-B_1)B_2 \dots B_{k-1}}, \dots, \frac{n_k s_k^2}{(1-B_{k-1})}\right]\right\}\right) \qquad (11)$$

Here the obtained function is the cumulative distribution function of *F* distribution with *k-1* and *n-k* degrees of freedoms. Generalized p value is the expected value of that distribution. if $p < \alpha$ then the null hypothesis is rejected.

5. Xu and Wang’s Generalized F test

The null hypothesis in Eq. 1 is defined as in Eq. 12 by Xu and Wang in 2007 .

$$H_0 : \mu_1 = \mu_k;\ \mu_2 = \mu_k; \dots ; \mu_{k-1} = \mu_k \qquad (12)$$

Or briefly, from the definitions, same hypothesis can be written as below:

$$\nu_a = (\mu_1, \mu_2, \dots, \mu_k)', \qquad \nu_b = 1_{k-1}\mu_k$$

$$H_0:\ \nu_a = \nu_b$$

$$H_1:\ \nu_a \neq \nu_b$$

Where 1_{k-1} is the vector of ones with dimension of $(k-1)\times 1$. The required equations for generalized p value developed by Xu and Wang are given below:

Let $S_i^2 = \frac{1}{n_i}\sum_{i=1}^{n_i}\left(X_{ij} - \bar{X}_i\right)^2$, $\bar{Y}_a, \bar{Y}_b, S_a, S_b$ where $\bar{Y}_a = (\bar{X}_1, \dots, \bar{X}_{k-1})'$,

$$\bar{Y}_b = 1_{k-1}\bar{X}_k \quad , S_a = diag\left(\frac{S_1^2}{n_1}, \dots, \frac{S_{k-1}^2}{n_{k-1}-1}\right)\ and \qquad S_b = \frac{S_k^2}{n_k} S_k^2 1_{k-1}1'_{k-1}$$

The value of generalized test variable *t* is obtained by Eq. 13.

$$t = (\bar{y}_a - \bar{y}_b)'(s_a + s_b)^{-1}(\bar{y}_a - \bar{y}_b) \qquad (13)$$

$$T = Y'[(s_a + s_b)^{-1/2}\left(diag\left(\frac{s_1^2}{U_1}, \dots, \frac{s_{k-1}^2}{U_{k-1}}\right) + \frac{s_k^2}{U_k} 1_{k-1}1'_{k-1}\right)(s_a + s_b)^{-1/2}]Y \qquad (14)$$

The variables $Y\ are\ U_i$ in Eq. 13 has the following distributions, $Y \sim N(0, I_{k-1})\ \ and \quad U_i \sim \chi^2_{n_i - 1}$,

If the null hypothesis is correct generalized p value is given as below

$$p = P(T \geq t | H_0)$$

APPLICATION

Global warming or temperature changes all over the world have been an attractive topic for the researchers in terms of meteorology, geography, biology, zoology, health, etc. There are numerous papers related with the results of weather changes during last decades. In this study, temperature changes has been examined for 5-year periods for the last two decades . The results for four different locations in Turkey are obtained and interpreted.

DATA SET

This study is based on twenty years (1991-2010) data of daily temperature from four different stations of Turkey. The selection of the stations based on the different population density and different areas in terms of climate. The most populated parts are north Istanbul (Göztepe), south Istanbul (Kartal), Trabzon (north part of Turkey), and Ankara (the capital, mid of Turkey). The data set is obtained from Turkish Institute of Meteorology.

Twenty years data set is divided four equal groups of five years for each working area. Firstly equality of variances are tested for all locations. While the group variances are found to be equal for Kartal and Ankara, they are not equal for Göztepe and Trabzon. ANOVA is conducted for Kartal and Ankara in order to check if the increase in temperature for the last 20 years statistically significant. The ANOVA results are interpreted for those two stations. On the other hand several tests for unequal variance situations are conducted for Göztepe and Trabzon.

RESULTS

4 different locations in Turkey are taken into account in order to check the climate change for 20 years. The data set is consisted from the maximum daily temperatures of Trabzon, Ankara, Göztepe and Kartal. Those locations are considered since they have different geographical and meteorological properties. Trabzon is located in the north east of Turkey, Ankara, the capital city, is in the middle of Turkey. Göztepe and Kartal are both in İstanbul; Kartal is located near the seaside in İstanbul and Göztepe is at the northern side of the city.

The data set of 20 years is divided four equal groups of 5-years period; 1991-1995, 1996-2000, 2001-2005 and 2006-2010. ANOVA is conducted to check if there is a difference in temperature for those time periods. Also post hoc comparison test is conducted after ANOVA in order to determine the groups with different means. The statistical analyses are conducted in SPSS 15.0.

Since the variances are equal for Ankara and Kartal, ANOVA conducted and the results are given in Table 1 and Table 2.

Table 1. ANOVA results for Ankara

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	75863.973	3	25287.991	253.169	.000
Within Groups	724770.170	7256	99.886		
Total	800634.143	7259			

Table 2. ANOVA results for Kartal

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	47085.864	3	15695.288	232.807	.000
Within Groups	429921.700	6377	67.418		
Total	477007.564	6380			

As given in Table 1 and Table 2, the null hypotheses are rejected for Ankara and Kartal which means that at least one of the time period has different mean than the other groups.

Results for Göztepe

Although we have unequal variances, ANOVA is conducted for Göztepe and it is concluded that the temperatures for four different time periods are different. The ANOVA results are given in Table 3. Other tests are also conducted for Göztepe and Trabzon because of presence of unequal variance.

Table 3. ANOVA results for Göztepe

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	4093,298	3	1364,433	19,659	.000
Within Groups	4093,298	7258	69,405		
Total	507837,174	7261			

Welch Test for Göztepe

The W statistic is obtained as W=19.349 and f is 4.032. The critical value is $F_{3;4.032}$=6.59. Since $F_{3;4.032} = 6.59 <$ W=19.349 H_0 is rejected.

Scott-Smith Test for Göztepe

Test statistic is calculated as $F_S = 57.94$ and the critical value with 4 degrees of freedom is 14.86. Since $\chi^2 = 14.86 < 57.94$ H_0 is rejected.

Brown-Forsthe Test for Göztepe

The test statistic for Brown-Forsthe is 19.7 and the critical value is $F_{3;7251.906}=2.6$.

Since $F_c=2.6 < B= 19.7$ H_0 is rejected.

Generalized F Test for Göztepe

$\tilde{S}_B = 58.207$

$B_1 \sim Beta(912.5,913)$

$B_2 \sim Beta(1825,912.5)$

$B_3 \sim Beta(2738.5,893.5)$

p= $1 - E\left(H_{3,7263}\left\{\frac{7263}{3} 58.207\left[\frac{1826(8.468)^2}{B_1 B_2 B_3}, \frac{1827(8.108)^2}{(1-B_1)B_2 B_3}, \frac{1826(8.318)^2}{(1-B_2)B_3}, \frac{1784(8.427)^2}{(1-B_3)}\right]\right\}\right)$

p= 0.00 so , H_0 is rejected

Xu and Wang's Generalized F Test for Göztepe

$$\bar{Y}_a = \begin{bmatrix} 18.160 \\ 18.439 \\ 18.914 \end{bmatrix}, \bar{Y}_b = \begin{bmatrix} 20.132 \\ 20.132 \\ 20.132 \end{bmatrix}, S_a = \begin{bmatrix} 0.039 & 0 & 0 \\ 0 & 0.036 & 0 \\ 0 & 0 & 0.038 \end{bmatrix}, S_b = \begin{bmatrix} 2.822 & 2.822 & 2.822 \\ 2.822 & 2.822 & 2.822 \\ 2.822 & 2.822 & 2.822 \end{bmatrix}$$

$$T = Y'[(s_a + s_b)^{-1/2}\left(diag\left(\frac{s_1^2}{U_1}, \dots, \frac{s_{k-1}^2}{U_{k-1}}\right) + \frac{s_k^2}{U_k} 1_{k-1} 1'_{k-1}\right)(s_a + s_b)^{-1/2}]Y$$

Since t=2.80 > T=0.49 H_0 is rejected.

Results for Trabzon

ANOVA is also conducted for Trabzon and it is concluded that the temperatures for four different time periods are different. The ANOVA results are given in Table 4.

Table 4. ANOVA results for Trabzon

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2047,130	3	682,377	13,034	.000
Within Groups	382032,516	7297	52,355		
Total	384079,646	7300			

Welch Test for Trabzon

The W statistic is obtained as W=12.652 and f is f=4055.15. Hence the critical value is $F_{3;4055.15}$= 3.12. Since $F_{3;\ 4055.15}$ = 12.652 < W=12.652 then H_0 is rejected.

Scott-Smith Test for Trabzon

Test statistic is calculated as $F_S = 39.046$ and the critical value with 4 degrees of freedom is 14.86. Since χ^2 =14.86<39.046 then H_0 is rejected.

Brown-Frosthe Test for Trabzon

The test statistic for Brown-Forsthe is calculated as B=13.353 and the critical value is $F_{3;6954.86}$ = 2.6. Since F_c=2.6< B=13.353 H_0 is rejected.

Generalized F Test for Trabzon

$\tilde{S}_B = 37.97$

$B_1 \sim Beta(912.5,913)$

$B_2 \sim Beta(1825,910.5)$

$B_3 \sim Beta(2735.5,912.5)$

$$p = 1 - E\left(H_{3,7297}\left\{\frac{7297}{3}37.97\left[\frac{1826(7.355)^2}{B_1B_2B_3},\frac{1827(7.174)^2}{(1-B_1)B_2B_3},\frac{1822(7.371)^2}{(1-B_2)B_3},\frac{1826(7.037)^2}{(1-B_3)}\right]\right\}\right)$$

p=0.00 so H_0 is rejected.

Xu and Wang’s Generalized F Test for Trabzon

$$\bar{Y}_a = \begin{bmatrix}17.910\\18.988\\19.343\end{bmatrix}, \bar{Y}_b = \begin{bmatrix}18.637\\18.637\\18.637\end{bmatrix}, S_a = \begin{bmatrix}0.03 & 0 & 0\\0 & 0.028 & 0\\0 & 0 & 0.029\end{bmatrix}, S_b = \begin{bmatrix}1.344 & 1.344 & 1.344\\1.344 & 1.344 & 1.344\\1.344 & 1.344 & 1.344\end{bmatrix}$$

$$T = Y'[(s_a + s_b)^{-1/2}\left(diag\left(\frac{s_1^2}{U_1}, \dots, \frac{s_{k-1}^2}{U_{k-1}}\right) + \frac{s_k^2}{U_k} 1_{k-1}1'_{k-1}\right)(s_a + s_b)^{-1/2}]Y$$

The result is t=0.008 > T=0.001 then H_0 is rejected.

CONCLUSIONS

In this study 4 different locations are considered. For those large data sets, the normality assumption hold, but for 2 data sets(Göztepe and Trabzon) unequal variance is observed among the groups. Ignoring the equal variance assumption , analysis of variance is applied for four data sets and null hypotheses are rejected for all locations. The other tests applied for Göztepe and Trabzon since unequal variance is observed for those locations. In the presence of unequal variance case , the null hypotheses are rejected for Göztepe and Trabzon, by the Welch test, Scott-Smith test, Brown-Forsythe test, Weerahandi's generalized p value and Xu-Wang's generalized p value.

The simulation studies in the literature reveal that, Weerahandi's and Xu and Wang's generalized p value tests give more powerful results for small sample sizes. For the large data sets results are close to each other as it is expected. The parallel results are obtained in this study.

Acknowledgment

This study is supported by the Research Foundation of Anadolu University. (Project Number: 1202F38).

REFERENCES

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

[2] Gamage, J., Mathew, T., Weerahandi, S., (2004). Generalized p-values and generalized confidence regions for the multivariate Behrens-Fisher problem and MANOVA. J. Multivariate Anal. 88, 177–189.

[3] Gamage, J. ve Weerahandi, S. (1998). Size performance of some tests in one-way ANOVA *Communications in Statistics and Simulations* 27(3), 625-640.

[4] Scott, A.J. ve Smith, T.M.F. (1971). Interval estimates for linear combinations of means. *Applied Statistics* 20(3), 276-285.

[5] Weerahandi, S. (2003). ANOVA under unequal error variances. *Biometrica* 38, 330-336.

[6] Weerahandi, S. (1995). *Exact statistical method for data analysis*. Springer-Verlag, NewYork, 2-50.

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

[8] Xu, L. and 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.

[9] Yiğit, E. and Gamgam, H. (2011). The test proposed for the one-way ANOVA under unequal variances and simulation study. Anadolu University Journal of Science and Technology-B, Theoretical Sciences, Vol.1; 57-71.