www.iiste.org

Development of a Test Statistic for Testing Equality of Two Means under Unequal Population Variances

[1]Abidoye, A. O, [2] Jolayemi, E. T , [3]Sanni, O. O. M and [4]Oyejola, B. A
Department of Statistics, University of Ilorin, Ilorin, Kwara State, Nigeria.
*[1] abidoye@unilorin.edu.ng, [2]tejujola@unilorin.edu.ng ,[3] omsanni@unilorin.edu.ng, [4] boyejola@unilorin.edu.ng

Abstract
In this paper, we propose a test statistic for testing equality of two independent sample means for unequal variances. When group variances differ, the pooled sample variance (S_P^2) is inadequate as a single value for the variances. This problem is commonly known as the Behrens – Fisher problem. Instead, the sample harmonic mean of variances (S_H^2) is proposed, examined and found to better represent the unequal variances. The distribution of S_H^2 which is known to be generalized Beta is further approximated by the chi – square distribution with the degrees of freedom related to that of degrees of freedom of the chi – square distribution of S_P^2 . Consequently, it is used to replace the pooled sample variance in the resulting proposed t – test. An example of application is provided.
Keywords: Harmonic mean of variances, chi- square distribution, modified t – test statistic

1. INTRODUCTION

Many authors such as Jonckheere (1954), Dunnett (1964), Montgomery (1981), Dunnet and Tamhane (1997), Yahya and Jolayemi (2003), Gupta et. al (2006), worked on the conventional test statistic on equality of two population means, $H_0 : \mu_1 = \mu_2$ against non-directional alternative, H_1: $\mu_1 \neq \mu_2$. When the variances are unequal, the pooled sample variance overestimates the appropriate variance and the test statistic becomes conservative. This is the well known Behrens – Fisher problem. There is therefore the need to seek for an alternative to the pooled variance. The interest of this work is to develop a suitable test procedure to address heterogeneity of variances, see Abidoye et. al (2013)

2. METHODOLOGY

We are interested in developing a suitable test procedure to test the hypothesis:

$$H_0 : \mu_1 = \mu_2 \text{ against } H_1 : \mu_1 \neq \mu_2 \quad or \quad H_0 : \mu_1 - \mu_2 = 0 \quad vs \quad H_1 : \mu_1 - \mu_2 \neq 0 \quad \ldots\ldots(2.1)$$

when the error term $e_{ij} \sim N(0,\sigma_i^2)$ $i = 1,2.$ and $j = 1,2,\ldots,n_i$ at is, under heterogeneity of variances.

Consequently $X_{ij} \sim N(\mu_i,\sigma_i^2)$ where X_{ij} are the observed value.

The unbiased estimate of $(\mu_1 - \mu_2) = (\bar{X}_1 - \bar{X}_2) = Y$ ……………………………………(2.2)

where $\bar{X}_1, \bar{X}_2$ are the sample means for groups 1 and 2 respectively.

$$V(Y) = Var[\bar{X}_1 - \bar{X}_2]$$

$$= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \ldots\ldots\ldots(2.3)$$

but

$$Y = (\bar{X}_1 - \bar{X}_2) \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}) \ldots\ldots\ldots(2.4)$$

2.1 DISTRIBUTION OF HARMONIC VARIANCE

Abidoye et al (2007) showed that σ_H^2 harmonic mean of group variances better represents series of unequal group variances and is estimated by S_H^2. It was also shown that the sample distribution of S_H^2 is approximated by the chi – square distribution.

$$Y = (\bar{X}_1 - \bar{X}_2) \sim N(\mu_1 - \mu_2, \sigma_H^2(\frac{n_1+n_2}{n_1 n_2})) \quad \text{...(2.5)}$$

Consequently, the test statistic for the hypotheses set in equation (2.1) is

$$t = \frac{Y}{Z} \quad \text{...(2.6)}$$

where

$$Y = (\bar{X}_1 - \bar{X}_2) \quad \text{...(2.7)}$$

and

$$Z = \sqrt{S_H^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \quad \text{...(2.8)}$$

Now p- value = $P(|t_r| > t)$ ...(2.9)

where t_r is regular t – distribution and r is the appropriate degrees of freedom for the t – test , which is obtained as shown in the next section.

3. DETERMINATION OF DEGREES OF FREEDOM FOR THE DISTRIBUTION OF SAMPLE HARMONIC MEAN OF VARIANCES

3.1 SIMULATION

Let $X_{ij} \sim N(\mu_i, \sigma_i^2)$, i = 1, 2 ; j = 1,2, …, n_i $n = \sum n_i$ and set $\sigma_1^2 \neq \sigma_2^2$. For various values of n , generate X_{ij} and compute S_H^2 having estimated degree of freedom as $\hat{r} = 2\hat{\alpha}$; and α is obtained from the generalised beta with parameters λ, α, and β all of which are estimated by method of moment . The degrees of freedom (r) for the proposed test statistic in equation (2.6) was related to $\sum n_i - 2$ = n- 2 degrees of freedom for the pooled variance (S_P^2), where $S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$. The value of degree of freedom of r was obtained through the $\hat{\alpha} = Y_{(n)}(s_H^2)$. from equation (2.5), where $Y_{(n)}$ is maximum value of Y.

Table 1 gives the values of n-2 the degrees of freedom for the pooled sample variance S_p^2 against the estimated degrees of freedom $\hat{r}$ for the chi – square distribution of S_H^2 .

Table 1: Simulated values of degrees of freedom for the proposed t - test statistic

n-2	$\hat{r}=2\hat{\alpha}$	n -2	$\hat{r}=2\hat{\alpha}$	n -2	$\hat{r}=2\hat{\alpha}$
8	4.8	318	189.9	618	364.4
18	8.6	328	199.8	628	369.1
28	13.4	338	207.9	638	374.3
38	19.2	348	215.6	648	379.4
48	23.2	358	226.4	658	384.7
58	25.6	368	232.3	668	389.6
68	26.8	378	238.9	678	394.9
78	36.3	388	253.2	688	399.8
88	39.6	398	264.4	698	404.2
98	41.8	408	265.2	708	409.8
108	51.3	418	267.5	718	414.2
118	58.9	428	269.3	728	419.4
128	62.1	438	274.3	738	424.7
138	67.7	448	279.4	748	429.2
148	71.5	458	284.4	758	434.4
158	74.6	468	289.8	768	439.7
168	85.2	478	294.3	778	444.5
178	95.9	488	299.1	788	449.2
198	98.6	498	304.4	798	454.6
208	113.8	508	309.4	808	459.8
218	118.4	518	314.1	818	464.2
228	124.9	528	319.2	828	469.5
238	133.2	538	324.4	838	474.2
248	138.8	548	329.9	848	479.3
258	142.4	558	334.3	858	484.7
268	148.3	568	339.4	868	489.8
278	152.7	578	344.4	878	494.4
288	163.2	588	349.2	888	499.2
298	175.4	598	354.2	898	504.8
308	182.8	608	359.5		

3.2 RELATIONSHIP BETWEEN EMPIRICAL DEGREE OF FREEDOM (r) AND n-2

The values of empirical degree of freedom (r) for chi-square from above simulated data in Tables 1 were plotted against $\sum n_i - 2 = n - 2,$ where n is the total number of all the observations in the groups. Clearly the contending models were, simple linear, quadratic and cubic models, see Figure 1. Table 2 shows coefficient of determination R^2 and adjusted R^2

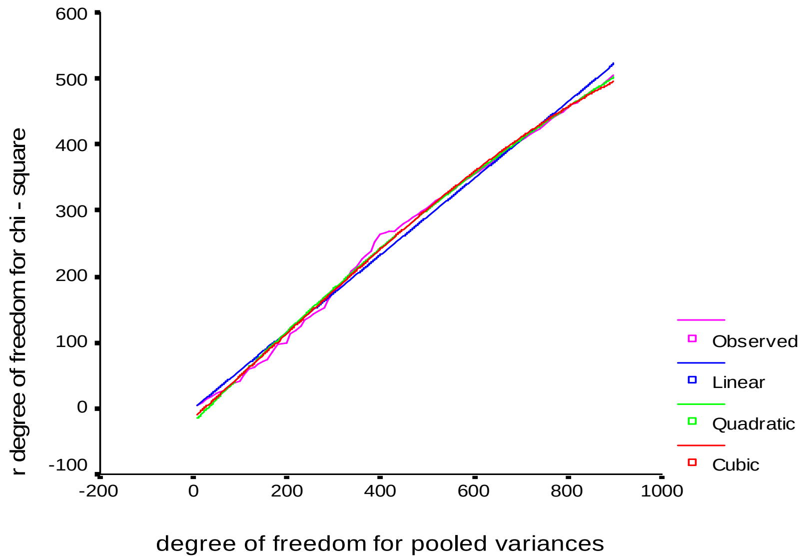

Figure 1: Plot of simulated degrees of freedom against the chi – square degrees of freedom

Table 2: Showing R^2 and Adjusted R^2 under Linear, Quadratic and Cubic

Model	**R^2**	**Adjusted R^2**	**Coefficients**
Linear	0.99337	0.99329	$a_0 = 0.607389$ $a_1 = 0.581022$
Quadratic	0.99731	0.99724	$a_0 = 20.966752$ $a_1 = 0.722882$ $a_2 = -0.000157$
Cubic	0.99760	0.99752	$a_0 = 13.856149$ $a_1 = 0.629294$ $a_2 = 0.000101$ $a_3 = -0.0000001897$

From the results, the quadratic equation (polynomial of order two) was seen to perform best in term of R^2, adjusted R^2 or the associated complexity so that the degree of freedom for the approximated chi –square is given by $r = 20.966752 + 0.722882(n-2) - 0.000157(n-2)^2$, where $n = \sum_{i=1}^{g} n_i$ is as defined earlier.

3.3 CONCLUSION

In this work we have established the degrees of freedom of the chi- square distribution proposed of the distribution of the sample harmonic mean of variances. Because the sample harmonic mean of variances has the chi – square distribution, the modified t – statistic is appropriate and eliminates the Beheren- Fisher's problem.

4. APPLICATION

In this study, the data used were secondary data, collected primarily by Kwara Agriculture Development Project (KWADP), Ilorin, Kwara State, Nigeria. They were extracts from her Agronomic survey report for ten consecutive cropping seasons, covering the period 1998 – 2007, see Abidoye (2012). The use of this test statistic is not limited to Agriculture alone, but it is equally applicable in Medicine, Social sciences and Engineering and Technology.

Table 3: Showing yields of sorghum in tons/ per acre for ten years (1998 – 2007).

Years	1	2	3	4	5	6	7	8	9	10
Zone A	1.0	1.2	1.1	1.2	1.18	0.9	1.3	0.9	1.08	1.0
Zone C	2.60	3.38	3.48	3.32	3.49	2.80	3.38	2.80	2.60	2.62

Computation on sorghum : From the data above the following summary statistics were obtained:

Zone A: $\overline{Y}_A = 1.086 \;, S_A^2 = 0.0183 \;, n_A = 10$

Zone C: $\overline{Y}_C = 3.047 \;, S_C^2 = 0.1538 \;, n_C = 10$

We need to verify the equality of the variances between these two zones. That is, testing the hypothesis;

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ vs } H_0 : \sigma_2^2 > \sigma_1^2$$

$$F = \frac{S_2^2}{S_1^2} \sim F_{(n_2-1, n_1-1, \alpha)}$$

$$F = \frac{0.1538}{0.0183} = 8.404$$

Since F_{cal} =8.404> $F_{9,9,(0.05)}$=3.18 we reject H_0 and therefore conclude that the two variances are not equal. Hench, we can not use the normal t- test statistic, thus we use our recommended modified t – test statistic.

In the above data set, $n_i = 10$, $n = \sum_{i=1}^{2} n_i = 20$, $S_H^2 = \left(\frac{1}{2} \sum_{i=1}^{2} \frac{1}{s_i^2} \right)^{-1}$, $S_H = 0.1809$

The hypothesis to be tested is

$$H_0 : \mu_C = \mu_A \text{ Vs } H_1 : \mu_C > \mu_A ; \sigma_1^2 \neq \sigma_2^2$$

$$t = \frac{\overline{X}_1 - \overline{X}_2}{S_H \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim t_r$$

$$= \frac{0.109}{0.1809 \sqrt{\left(\frac{1}{10} + \frac{1}{10} \right)}} = \frac{0.109}{0.080901}$$

= 1.34733

where r = 20.966752 + 0.722882(n –2) – 0.000157(n-2)^2
= 33.93

Now p- value = $P(|t_r| > t) = P(t_r > t)$
$= P(t_r > 1.34733)$
= 0.093393
> 0.05

This led to acceptance of H_0 and we conclude that the mean yields of sorghum in the two zones are not significantly different.

REFERENCES

[1] Abidoye, A.O, Jolayemi, E.T, Sanni, O.O.M and Oyejola, B.A (2013): On Hypothesis Testing Under Unequal Group Variances:The Use of the Harmonic variance.The International Journal of Institute for science, Technology and Education. Vol. 3, No 7, Page 11 - 15

[2] Abidoye, A. O (2012): Development of Hypothesis Testing Technique for Ordered Alternatives under heterogeneous variances. Unpublished Ph.D Thesis submitted to Dept. of Statistics, University of Ilorin, Ilorin.

[3] Abidoye, A. O, Jolayemi, E.T and Adegboye, O.S (2007): On the Distribution of a Scaled Beta Random Variable. JNSA, 19, 1 – 3.

[4] Dunnett, C. W and Tamhane, A. C (1997): Multiple testing to establish superiority / equivalence of a new treatment compare with k standard treatments. Statistics in Medicine 16, 2489 – 2506.

[5] Dunnett , C. W (1964): New tables for multiple comparison with a control. Biometric, 20, 482-491.

[6] Gupta, A.K, Solomon, W.H and Yasunori, F (2006): Asymptotics for testing hypothesis In some Multivariate variance components model Under non-normality. Journal of Multivariate Analysis archive. Vol. 97, pp 148-178.

[7] Jonckheere, A.R (1954): “A distribution-free k-sample test against ordered alternative.” Biometrical, 41, 133-145.

[8] Montgomery, D.C (1981): “Design and Analysis of Experiment.” Second Edition. John Wiley and sons Inc. New York.

[9] Yahya, W.B and Jolayemi, E.T (2003): Testing Ordered Means against a Control. JNSA,16, 40- 51.