Transactions of the Nigerian Association of Mathematical Physics
***Volume 7,(March, 2018*), pp251 –256**
© Trans. of NAMP

Erratum: Comparison of Equality of two means with Equal and Unequal Variance {Transactions of the Nig. Assoc. of Maths. Phys. Vol. 5, (September and November, 2017), pp161-166}

By

Abidoye A.O.

Department Of Statistics, University of Ilorin, Ilorin, Nigeria.

The Equations and Several Symbols in this paper did not appear properly in the Vol. 5 Issue of the Transactions. The entire paper is therefore reproduced below as it ought to be.

COMPARISON OF EQUALITY OF TWO MEANS WITH EQUAL AND UNEQUAL VARIANCE

Abidoye A.O.

Department Of Statistics, University of Ilorin, Ilorin, Nigeria.

Abstract

In this paper we want to compare the conventional t –test with proposed t – test and see which one gives better results between the two when we have unequal variances. The harmonic mean of variances was proposed as an alternative to the pooled sample variance when there is heterogeneity of variances. However, when the variances are unequal, the pooled sample variance overestimates the appropriate variance and the test statistic becomes conservative and this is well known Behrens – Fisher problem.
The result shows that the proposed t – test statistic is found to be appropriate than conventional t – test statistic for the data set obtained from Kwara Agricultural Development Project (KWADP) when we have unequal variances and conventional t – test perform better when we have equal variances with the set of data obtained from Ministry of Agriculture in Kwara State .

Keyword: Equality of means, t – test statistic, harmonic mean of variances, heterogeneity of variance, constant variances.

1. INTRODUTION

The two – sample t – test is used to determine if two populations' means are equal or not equal. A common application is to test if a new process or treatment is superior to a current process or treatment. The data may either be paired or not paired. Then by paired, we mean that there is a one – to – one correspondence between the values in the two samples. That is, if $X_1, X_{2,} ...X_n$ and $Y_{1,} Y_{2,} ...Y_n$ are the two samples, then X_i corresponds to Y_i. For paired samples, the difference X_i - Y_i is usually calculated. Then, for unpaired samples, the sample sizes for the two samples may or may not be equal [1 – 5].
A test statistic is a statistic used in statistical hypothesis testing. A hypothesis test is typically specified in terms of a test statistic, considered as a numerical summary of a data set that reduces the data to one value that can be used to perform the hypothesis test. In general, a test statistic is selected or defined in such a way as to quantify, within observed data, behaviors that would distinguish the null from the alternative hypothesis where such an alternative is prescribed or that would characterize the null hypothesis if there is no explicitly stated alternative hypothesis [6 – 9] .

Correspondence Author: Abidoye A.O., E-mail: solatayo2003@yahoo.com, Tel.: +2348166010235, +2348050663983

An important property of a test statistic is that its sampling distribution under the null hypothesis must be calculable, either exactly or approximately, which allows p – values to be calculated. A test statistic shares some of the same qualities of a descriptive statistic and many statistics can be used as both test statistics and descriptive statistics. However, a test statistic is specifically intended for use in statistical testing, whereas the main quality of a descriptive statistic is that it is easily interpretable. Some informative descriptive statistics, such as the sample range do not make good test statistics since it is difficult to determine their sampling distribution. A t – test is appropriate for comparing means under relaxed conditions [3, 7, 10 – 14].

2.0 METHODOLGY

2.1 The Procedure for Proposed T -Test Statisic

We proposed a test statistic to test the hypothesis:

$H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$ ……………………………..………..….(2.1)

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

Then $X_{ij} \sim N(\mu_i, \sigma_i^2)$ and set $\sigma_1^2 \neq \sigma_2^2$. where X_{ij} are the observed values.

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.

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

$$= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \quad \text{……………………………………………………….…………(2.3)}$$

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

2.2 Distribution of Harmonic Variance

Abidoye [6] 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

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

and

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

Now p- value = $P(t = t_r > t_{cal})$ …………....……………………….……………(2.9)

where t_r is regular t – distribution and r is the appropriate degrees of freedom for the t – test .

2.3 The Procedure For Conventional T - Test Statisic

We set the conventional t – test statistic to test the hypothesis

$H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$ ……………………………………………………(2.10)

and the error term is $e_{ij} \sim N(0, \sigma^2)$ $i = 1,2.$ and $j = 1,2,...,n_i$ with constant variance and is the homogenous variances

Then $X_{ij} \sim N(\mu_i, \sigma_i^2)$ and set $\sigma_1^2 = \sigma_2^2$. where X_{ij} are the observed values.

The unbiased estimate of $(\mu_1 - \mu_2) = (\bar{X}_1 - \bar{X}_2) = Y$ ………………………………………(2.11)

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

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

$$= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \quad \ldots\ldots(2.12)$$

but $$Y = (\bar{X}_1 - \bar{X}_2) \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}) \quad \ldots\ldots(2.13)$$

2.4 DISTRIBUTION OF POOLED VARIANCES

The pooled variance is defined as S_p^2 which represents series of equal group variances. It was also shown that the sample distribution of S_p^2 is approximated by the chi – square distribution. where $S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$

$$Y = (\bar{X}_1 - \bar{X}_2) \sim N(\mu_1 - \mu_2, S_P^2(\frac{n_1 + n_2}{n_1 n_2})) \quad \ldots\ldots(2.14)$$

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

$$t = \frac{Y}{Z} \quad \ldots\ldots(2.15)$$

where

$$Y = \left(\bar{X}_1 - \bar{X}_2\right) \ldots\ldots(2.16)$$

and

$$Z = \sqrt{S_P^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \quad \ldots\ldots(2.17)$$

Now p- value = $P(t = t_r > t_{cal})$

$$\ldots\ldots(2.18)$$

where t_r is regular t – distribution and r is the appropriate degrees of freedom for the t – test .

3.0 DATA ANALYSIS

In this study, the data used were secondary data, obtained from Kwara Agriculture Development Project (KWADP), Ilorin, Kwara State. They were extracts from her Agronomic Survey Report for ten consecutive cropping seasons, covering the period 1998 – 2007. See Abidoye (2012).

Table 3.1: Showing yields of maize for ten years (1998 – 2007).

Years	1	2	3	4	5	6	7	8	9	10
Zone A	2.5	2.9	1.7	2.8	1.8	2.9	3.6	3.8	2.9	3.2
Zone B	6.8	10.2	0.3	4.5	2.1	3.8	1.2	3.3	0.9	9.3

Computation on Proposed t – test statistic

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

Zone A: $\bar{Y}_A = 2.81, S_A^2 = 0.4588, n_A = 10$

Zone B: $\bar{Y}_B = 4.24, S_B^2 = 12.147, n_B = 10$

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

$H_0 : \sigma_1^2 = \sigma_2^2$ vs $H_0 : \sigma_2^2 \neq \sigma_1^2$

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

$$F = \frac{12.147}{0.4588} = 26.4756$$

Since F_{cal} = 9.453 > $F_{9,9,(0.05)}$=3.18 we do not reject H_0 and therefore conclude that the two variances are not equal.

In the above data set, $n_i = 10$, g = 2, $n = \sum_{i=1}^{2} n_i = 20$, $S_H^2 = \left(\frac{1}{g}\sum \frac{1}{s_i^2}\right)^{-1}$, $S_H = 0.9403$

The hypothesis to be tested is

$H_0 : \mu_A = \mu_B$ Vs $H_1 : \mu_A < \mu_B; \sigma_1^2 \neq \sigma_2^2$

$$t^* = \frac{\bar{X}_A - \bar{X}_B}{S_H \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_r$$

$$= \frac{-1.430}{0.9403\sqrt{\left(\frac{1}{10} + \frac{1}{10}\right)}} = \frac{-1.430}{0.42051}$$

$$= \left|-3.4006\right| = 3.4006$$

Now p- value = $P(t^* = t_r > t_{cal}) = P(t_r > t_{cal})$

$= P(t_r > 3.4006)$

$= 0.02393 < 0.05$

This leads to the rejection of the null hypothesis, H_0 and we conclude that the mean yields of maize in the two zones are significantly different.

Computation on Conventional t – test statistic

In the above data set, $n_i = 10$, g = 2, $n = \sum_{i=1}^{2} n_i = 20$, $S_p^2 = \frac{(n_A - 1)S_A^2 + (n_B - 1)S_B^2}{n_A + n_B - 2} = 6.3031$

The hypothesis to be tested is

$H_0 : \mu_A = \mu_B$ Vs $H_1 : \mu_A < \mu_B; \sigma_1^2 = \sigma_2^2$

$$t^* = \frac{\bar{X}_A - \bar{X}_B}{S_P \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_r$$

$$= \frac{-1.430}{2.5106\sqrt{\left(\frac{1}{10} + \frac{1}{10}\right)}} = \frac{-1.430}{1.1228}$$

$$= \left|-1.2736\right| = 1.2736$$

Now p- value = $P(t^* = t_r > t_{cal}) = P(t_r > t_{cal})$

$= P(t_r > 1.2736)$

$= 0.08789 > 0.05$

This leads to the non- rejection of the null hypothesis, H_0 and we conclude that the mean yields of maize in the two zones are not significantly different.

Also in this study, we make used another data, obtained from the Ministry of Agriculture in Ilorin, Kwara State which were secondary data, covering the period 2000 – 2010.

Table 3.2: Showing yields of Sorghum for ten years (2000 – 2010).

Years	1	2	3	4	5	6	7	8	9	10
Zone A	0.9	0.7	0.8	0.6	0.9	0.9	0.7	0.8	0.8	0.9
Zone B	0.6	0.5	0.7	0.4	0.7	0.8	0.6	0.6	0.6	0.68

Computation on Proposed t – test statistic

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

Zone A: $\bar{Y}_A = 0.8, S_A^2 = 0.011, n_A = 10$

Zone B: $\bar{Y}_B = 0.62, S_B^2 = 0.013, n_B = 10$

In the above data set, $n_i = 10$, g = 2, $n = \sum_{i=1}^{2} n_i = 20$, $S_H^2 = \left(\frac{1}{g} \sum \frac{1}{s_i^2} \right)^{-1}$, $S_H = 0.1092$

The hypothesis to be tested is

$$H_0 : \mu_A = \mu_B \text{ Vs } H_1 : \mu_A < \mu_B ; \sigma_1^2 \neq \sigma_2^2 \qquad t^* = \frac{\overline{X}_A - \overline{X}_B}{S_H \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim t_r$$

$$= \frac{0.18}{0.1092 \sqrt{\left(\frac{1}{10} + \frac{1}{10} \right)}} = \frac{0.180}{0.04884}$$

$$= 3.6855$$

Now p- value = $P(t^* = t_r > t_{cal}) = P(t_r > t_{cal})$

$$= P(t_r > 3.6855)$$

= 0.02393 < 0.05

This leads to the rejection of the null hypothesis, H_0 and we conclude that the mean yields of sorghum in the two zones are significantly different.

Computation on Conventional t – test statistic

In the above data set, $n_i = 10$, g = 2, $n = \sum_{i=1}^{2} n_i = 20$, $S_p^2 = \frac{(n_A - 1) S_A^2 + (n_B - 1) S_B^2}{n_A + n_B - 2} = 0.01201$

The hypothesis to be tested is

$$H_0 : \mu_A = \mu_B \text{ Vs } H_1 : \mu_A < \mu_B ; \sigma_1^2 = \sigma_2^2 \qquad t^* = \frac{\overline{X}_A - \overline{X}_B}{S_P \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim t_r$$

$$= \frac{0.18}{0.1096 \sqrt{\left(\frac{1}{10} + \frac{1}{10} \right)}} = \frac{0.180}{0.04901}$$

$$= 3.6727$$

Now p- value = $P(t^* = t_r > t_{cal}) = P(t_r > t_{cal})$

$$= P(t_r > 1.2736)$$

= 0.03252 < 0.05

This leads to the rejection of the null hypothesis, H_0 and we conclude that the mean yields of sorghum in the two zones are significantly different.

4.0 CONCLUSION

The result of analysis shows that proposed t – test statistic is perform better when we have unequal variances in the data because the data of maize used indicate that it contribute significantly different which mean it leads to the rejection of the null hypothesis. But under conventional t – test statistic, the test statistic do not perform better because it leads to the non – rejection of the null hypothesis. In a situation when we have equal variances in the data, both test statistic perform better which indicate that both tests contribute significantly different which mean it leads to the rejection of the null hypothesis.

REFERENCES

[1] Abidoye, A.O, Jolayemi, E.T, Sanni, O.O.M and Oyejola, B. A.(2016a). On Application of Modified F – Statistic: An Example of Sales Distribution of Pharmaceutical Drug. **Journal of Science World** (www.scienceworldjournal.com) Page 23 – 26, Vol. 11, No2 .

[2] Abidoye, A.O, Jolayemi, E.T, Sanni, O.O.M and Oyejola, B. A.(2016b). Development of Testing Ordered Mean Against a Control Under Heterogeneous Variance. Journal of the Nigerian Association of Mathematical Physics. Page 125 – 128, Vol. 33.

[3] Abidoye, A.O, Jolayemi, E.T, Sanni, O.O.M and Oyejola, B. A.(2016c). The Use of Harmonic Variance: A Better Alternative for Hypothesis Testing Under Unequal Group Variances. The Journal of the Mathematical Association of Nigeria. Page 91 – 99, Vol. 43, No2

[4] Abidoye, A.O, Sanni, O.O.M, Ikoba, N. A and Job,O. (2017): On a Proposed Test Statistic For Testing Equality of Two Means Under Unequal Population Variances. Taguwa The Journal of Humanities. Faculty of Humanities Umaru Musa Yar' Adua University Katsina. Page 152 – 158, Vol. 8, No1.

[5] 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.

[6] 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.

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

[8] 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.

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

[10] Adegboye, O. S (2009): Descriptive Statistic for Students, Teachers and Practitioners. First Edition. Olad Publishing.

[11] Gupta, S.C (2004): Fundamentals of Statistics. Sixth edition. Himalaya Publishing.

[12] Adegboye, O. S (1981): "On Testing Against Restricted Alternative In Gaussian Models". Ph.D Dissertation. BGUS.

[13] Adegboye, O. S and Gupta, A. K (1986): "On Testing Against Restricted Alternative About The Mean of Gaussian Models With Common Unknown Variance".Sankhya. The Indian Journal of Statistics, series B, 48, pp 331 – 341.

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