International Journal of Mathematics and Statistics Invention (IJMSI)
E-ISSN: 2321 – 4767 P-ISSN: 2321 - 4759
www.ijmsi.org Volume 1 Issue 1 ‖ August. 2013‖ PP-38-43

Development of a Test Statistic for Testing Equality of 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.

ABSTRACT: *In this work, we propose a test statistic for testing equality of means under unequal variances. When group variances differ, It is very inappropriate to use the pooled sample variance (S_P^2) as a single value for the variances. This problem is commonly known as the Behrens – Fisher problem in the two – sample situation. Instead, the sample harmonic mean of variances (S_H^2) is proposed, examined and found useful for unequal variances. Data set from Kwara State Ministry of Health on the prevalence of diabetes diseases for male patients was used to illustrate the relevance of our proposed test statistic.*

KEYWORDS: *Harmonic mean of variances, chi- square distribution, modified t – test statistic*

I. INTRODUCTION

The conventional test statistic in ANOVA for testing equality of g population means, $H_0 : \mu_1 = \mu_2 = ... = \mu_g = \mu$ against non-directional alternative, H_1: $\mu_i \neq \mu$ for at least one i, $i = 1,2,\ldots,g$, is not appropriate under the non – homogeneity of the variances. Instead, we might be tempted to run all possible pair wise comparisons of the population means. If we assume that all the g distributions are approximately normal with means given by $\mu_1, \mu_2,..., \mu_g$ and common variance σ^2, we need to run $\binom{g}{2}$ t- test for comparing all pairs of means.

Obviously, this test procedure may be too tedious and time consuming. Besides, a more important but less apparent disadvantage of running multiple t-tests to compare means as stated above is that the probability of falsely rejecting at least one of hypothesis increases as the number of t-test increases (Ott, 1984). This was the origin of the Bonferroni multiple comparison procedure, (Neter and Wasserman, 1974).

Although, we may have the probability of a type I error fixed at α= 0.05 for each individual test, the probability of falsely rejecting at least one of those tests is larger than 0.05. In other words, the combined probability of falsely rejecting one of the $\binom{g}{2}$ hypotheses would be larger than the α value of 0.05 set for each individual test.

However, what is desirable is a single test. See Jonckheere (1954), Dunnett (1964), Montgomery (1981), Dunnet and Tamhane (1997), Yahya and Jolayemi (2003).

The interest of this work is to develop a suitable test procedure to address heterogeneity of variances, arising from this situation. See Abidoye et. al (2013b)

II. METHODOLOGY

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

$H_0 : \mu_1 = \mu_2 = ... = \mu_g = \mu$ against non-directional alternative, H_1: $\mu_i \neq \mu$, for at least one i, ………(2.1)

where the error term $e_{ij} \sim N(0,\sigma_i^2)$ $i = 1,2,\ldots,g$.

The hypothesis of equation (2.1) can be split into two cases: case I and case II which is well explained in Bonferroni test statistic, see Dunnett (1964), Guptal et. al (2006) and Abidoye et. al (2007)

Define $\delta_i = \mu_i - \mu$,……………………………………………………………………(2.2)
then, equation 2.1 can be written as
$H_0 : \delta_i = 0$ vs $H_1 : \delta_i \neq 0$

Consequently the hypothesis set is $H_0 : \delta_i = 0 vs H_1 : \delta_i < 0 \cup \delta_i > 0$ ………………(2.3)
Simply put
$H_0 : \delta_i = 0 \; vs \; H_1 :$ case I or case II……………………………………………….(2.4)
Assume
$Y_i = \delta_i$
The unbiased estimate of Y_i is
$\delta_i = \bar{X}_i - \bar{\bar{X}}$ ……………………………………………………………………………....(2.5)
Therefore
$\hat{\delta}_i \sim N[\delta_i, V(\delta_i)]$
where

$$V(\delta_i) = V(Y_i)$$
$$= V(\bar{X}_i - \bar{\bar{X}})$$
$$= V(\bar{X}_i) + V(\bar{\bar{X}}) - 2\operatorname{cov}(\bar{X}_i, \bar{\bar{X}})$$
$$= V(\bar{X}_i) + V(\bar{\bar{X}}) - 2\operatorname{cov}(\bar{X}_i, \frac{\sum \bar{X}_i}{g})$$
$$= V(\bar{X}_i) + V(\bar{\bar{X}}) - 2\operatorname{cov}(\bar{X}_i, \frac{\bar{X}_i}{g})$$
$$= \frac{\sigma_i^2}{n_i} + \frac{\sigma_H^2}{n} - \frac{2}{g} V(\bar{X}_i)$$
$$= \frac{\sigma_i^2}{n_i} + \frac{\sigma_H^2}{n} - \frac{2\sigma_i^2}{g n_i} \text{, see Abidoye (2012)}$$
$$= \frac{\sigma_i^2}{n_i} + \frac{\sigma_H^2}{n} \quad if \quad n_i g \to \infty$$

Therefore,

$$V(Y_i) = \frac{\sigma_i^2}{n_i} + \frac{\sigma^2}{n} \text{ ……………………………………………………………….............(2.6)}$$

$$\sigma_H^2 \left(\frac{n_i + n}{n_i n} \right) \text{ see Abidoye (2012)}$$

$\min(\hat{\delta}_i) = 0$ and $\max(\hat{\delta}_i) = 0$

Let

$$Y^* = \min(\bar{X}_i - \bar{X}) \sim \lambda_2 N(\mu_i - \mu, \frac{\sigma_i^2}{n_i} + \frac{\sigma^2}{n}) \text{ ………………………………………………(2.7)}$$

and

$$Y^{**} = \min(\bar{X}_i - \bar{X}) \sim \lambda_1 N(\mu_i - \mu, \frac{\sigma_i^2}{n_i} + \frac{\sigma^2}{n}) \quad \text{.....................(2.8)}$$

where

$\lambda_1 = g(1 - \Phi(\bar{Y}_i - \bar{\bar{Y}}))^{g-1}, 0 < \lambda_1 < 1$

and

$\lambda_2 = g(\Phi(\bar{Y}_i - \bar{\bar{Y}}))^{g-1}, 0 < \lambda_2 < 1$, obtained from distribution of order statistic .

2.1 Distribution of Harmonic Variance

Abidoye et al (2013a) showed that harmonic mean of group variances σ_H^2 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.

$$Y^{*} = \min(\bar{X}_i - \bar{\bar{X}}) \sim \lambda_2 N(\mu_i - \mu, \sigma_H^2(\frac{n_i + n}{n_i n})) \quad \text{.....................(2.9)}$$

and

$$Y^{**} = \max(\bar{X}_i - \bar{\bar{X}}) \sim \lambda_1 N(\mu_i - \mu, \sigma_H^2(\frac{n_i + n}{n_i n})) \quad \text{.....................(2.10)}$$

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

$$t = \frac{Y_i}{Z} \quad \text{.....................(2.11)}$$

where

$$Y_i = \left(\bar{X}_i - \bar{\bar{X}}\right) \quad \text{.....................(2.12)}$$

and

$$Z = \sqrt{S_H^2\left(\frac{1}{n_i} + \frac{1}{n}\right)} \quad \text{.....................(2.13)}$$

Now p- value = $P(t_r > t) = P(t_r^* > \frac{t}{\lambda})$ ....(2.14)

where λ can be λ_1 or λ_2 and t_r^* is regular t – distribution and r is the appropriate degrees of freedom for the t – test .

The degree of freedom r for the Harmonic mean of variances have been determined to be r = 22.096 + 0.266(n –g) – 0.000029(n-g)^2 see Abidoye et. al (2013b, 2013a) .

3 Application

The data used in this study are secondary data, collected primarily by Kwara State Ministry of Health, Ilorin, Kwara State, Nigeria. They were extracts from incidence of diabetes diseases for male patients for ten consecutive years, covering the period 2001 – 2010

Table 1: Showing the prevalence of diabetes diseases for male patients in Kwara State for ten years (2001- 2010).

Years	1	2	3	4	5	6	7	8	9	10
Zone A 1	37	80	58	48	35	46	53	39	64	76
Zone B 2	14	19	12	21	23	13	15	16	11	14
Zone C 3	15	18	11	19	22	14	13	15	10	13
Zone D 4	11	19	10	18	23	12	14	16	12	15

By the application of Levene test of equality of variances of Table 1, the variances differ from zone to zone.

Table 2: Levene test for variance equality

	Levene Statistic	df_1	df_2	P-value
Response	10.975	3	36	0.000

Hence, we can not use the conventional t –test statistic, but that which is proposed in the work. From the data in Table 1 the following summary statistic were obtained:

Zone A: $\bar{X}_A = 53.6, S_A^2 = 250.04, n_A = 10$

Zone B: $\bar{X}_B = 15.8, S_B^2 = 15.7, n_B = 10$

Zone C: $\bar{X}_C = 15.0, S_C^2 = 13.9, n_C = 10$

Zone D: $\bar{X}_D = 15.0, S_D^2 = 16.7, n_D = 10$

$\bar{X}_A$	$\bar{X}_B$	$\bar{X}_C$	$\bar{X}_D$	$\bar{\bar{X}}$
53.6	15.8	15.0	15.0	24.9

Therefore, we consider the minimumaand maximum differences of means respectively as given below:

$$Y_1 = 53.6 - 24.9 = 28.7$$
$$Y_2 = 15.8 - 24.9 = -9.1$$
$$Y_3 = 15.0 - 24.9 = -9.9$$
$$Y_4 = 15.0 - 24.9 = -9.9$$
$$S_H^2 = \left(\frac{1}{4} \sum_{i=1}^{4} \frac{1}{s_i^2} \right)^{-1}$$
$$S_H^2 = 20.05$$

Then , the minimum difference of means is

$$Y^* = \min(\bar{X}_i - \bar{\bar{X}}) = (\bar{X}_C - \bar{\bar{X}}) = -9.9$$

In the above data set, $n_i = 10$, g = 4 , $n = \sum_{i=1}^{4} n_i = 40$, $S_H^2 = \left(\frac{1}{4} \sum_{i=1}^{4} \frac{1}{s_i^2} \right)^{-1}$, $S_H^2 = 20.05$

The main hypothesis to be tested is

$H_0 : \mu_A = \mu_B = \mu_C = \mu_D = \mu$ against H_1: $\mu_i \neq \mu$, for at least one i, i.e $i = A, B, \ldots, D$

The hypothesis to be tested is

$H_0 : \mu_i - \mu = 0$ vs $H_1 : \mu_i - \mu < 0$

$$t = \frac{\min(\overline{X}_i - \overline{\overline{X}})}{S_H \sqrt{\left(\frac{1}{n_i} + \frac{1}{n}\right)}} \sim t_r$$

$$= \frac{-9.9}{4.478\sqrt{\left(\frac{1}{10} + \frac{1}{40}\right)}} = \frac{-9.9}{1.5832}$$

= -6.25

where r = 22.096 + 0.266(n –g) – 0.000029(n-g)^2

= 31.63

$\lambda_2 = g(\Phi(\overline{Y}_i - \overline{\overline{Y}}))^{g-1} = 0^+$, $0 < \lambda_2 < 1$ from equation (2.1)

Now p- value = $P(|t_r| > t) = P\left(t_r < \frac{t_{cal}}{\lambda_2}\right)$

$$= P\left(t_r > \frac{-6.25}{0^+}\right)$$

$$= P(t_r < -\infty)$$

$$= 0$$

< 0.025

In this regard, we reject H_0 and conclude that the mean of incidence for diabetes diseases in all the four zones are significantly different from the overall incidence rate at 5% level of significance. Indeed zones A could be the zone for which incidence was highest and would need a special attention.

Next we consider the maximum difference of means is

$$Y^{**} = \max(\overline{X}_i - \overline{\overline{X}}) = (\overline{X}_A - \overline{\overline{X}}) = 28.7$$

In the above data set, $n_i = 10$, g = 4 , $n = \sum_{i=1}^{4} n_i = 40$, $S_H^2 = \left(\frac{1}{4}\sum_{i=1}^{4}\frac{1}{s_i^2}\right)^{-1}$, $S_H^2 = 20.05$

The hypothesis to be tested is

$H_0 : \mu_i - \mu = 0$ against H_1: $\mu_i - \mu > 0$

$$t = \frac{\max(\overline{X}_i - \overline{\overline{X}})}{S_H \sqrt{\left(\frac{1}{n_i} + \frac{1}{n}\right)}} \sim t_r$$

$$= \frac{28.7}{4.478\sqrt{\left(\frac{1}{10}+\frac{1}{40}\right)}} = \frac{28.7}{1.5832}$$

= 18.13

where r = 22.096 + 0.266(n –g) – 0.000029(n-g)^2

= 31.63

$\lambda_1 = g(1-\Phi(\overline{Y}_i - \overline{\overline{Y}}))^{g-1} = 4.0,\ 0 < \lambda_1 < 1$ from equation (2.1)

Now p- value = $P(t_r > t) = P\left(t_r > \frac{t_{cal}}{\lambda_1}\right)$

$$= P\left(t_r > \frac{18.13}{4.0}\right)$$

$$= P(t_r < 4.5325)$$

$$= 3.845 \times 10^{-5}$$

< 0.025

Which led to the rejection of H_0 and conclude that the mean of incidence rate for diabetes diseases in all the four zones are not the same at 5% f significance?

III. CONCLUSION

In this work we have developed a test statistic for testing equality of means under unequal population 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.

REFERENCES

[1]. Abidoye, A.O, Jolayemi, E.T, Sanni, O.O.M and Oyejola, B.A (2013a): 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, Jolayemi, E.T, Sanni, O.O.M and Oyejola, B.A (2013b): Development of a TestStatistic for Testing Equality of Two Means Under Unequal Population Variances..The International Journal of Institute for science, Technology and Education. Vol. 3, No 10, page 10 -14.

[3]. 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.

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

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

[7]. 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.

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

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

[10]. Neter, J. and Wassarman, W. (1974): “Applied Linear Statiscal Model”. Richard D. Irwin. Inc. IIIinois.

[11]. Ott, L (1984): “An Introduction To Statistical Methods and Data Analysis”. Second edition. P.W.S Publisher. Boston.

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