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

1Abidoye, A. O, 2Jolayemi, E. T, 3Sanni, O. O. M and 4Oyejola, 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,

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_g = \mu$$

against non-directional alternative, $H_1$: $\mu_i \neq \mu$ for at least one $i$,

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

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
\[
H_0 : \delta_i = 0 \quad \text{vs} \quad H_1 : \delta_i \neq 0
\]
Consequently the hypothesis set is
\[
H_0 : \delta_i = 0 \quad \text{vs} \quad H_1 : \delta_i < 0 \cup \delta_i > 0 \quad \text{.................}(2.3)
\]
Simply put
\[
H_0 : \delta_i = 0 \quad \text{vs} \quad H_1 : \delta_i \neq 0
\]
Assume
\[
Y_i = \delta_i
\]
The unbiased estimate of \( Y_i \) is
\[
\hat{\delta_i} = \overline{X_i} - \overline{X} \quad \text{.................................................................}(2.5)
\]
Therefore
\[
\hat{\delta_i} \sim N[\delta_i, V(\delta_i)]
\]
where
\[
V(\hat{\delta_i}) = V(Y_i)
\]
\[
= V(\overline{X_i} - \overline{X})
\]
\[
= V(\overline{X_i}) + V(\overline{X}) - 2 \text{cov}(\overline{X_i}, \overline{X})
\]
\[
= V(\overline{X_i}) + V(\overline{X}) - 2 \text{cov}(\overline{X_i}, \frac{\sum X_i}{g})
\]
\[
= V(\overline{X_i}) + V(\overline{X}) - 2 \frac{\sum X_i}{g}
\]
\[
= \frac{\sigma_i^2}{n_i} + \frac{\sigma_H^2}{n} - \frac{2}{V(\overline{X_i})}
\]
\[
= \frac{\sigma_i^2}{n_i} + \frac{\sigma_H^2}{n} - \frac{2\sigma_i^2}{gn_i}, \text{ see Abidoye (2012)}
\]
\[
= \frac{\sigma_i^2}{n_i} + \frac{\sigma_H^2}{n} \quad \text{if} \quad n_i \ g \rightarrow \infty
\]
Therefore,
\[
V(Y_i) = \frac{\sigma_i^2}{n_i} + \frac{\sigma_H^2}{n} \quad \text{.........................................................}(2.6)
\]
\[
\sigma_H^2 \left( \frac{n_i + n}{n_i n} \right), \text{ see Abidoye (2012)}
\]
\[
\min(\hat{\delta_i}) = 0 \quad \text{and} \quad \max(\hat{\delta_i}) = 0
\]
Let
\[
Y^* = \min(\overline{X_i} - \overline{X}) \sim \lambda_2 N(\mu_i - \mu, \frac{\sigma_i^2}{n_i} + \frac{\sigma_H^2}{n}) \quad \text{..............................................}(2.7)
\]
and

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

where

\[ \lambda_1 = g(1 - \Phi(\bar{Y}_i - \bar{Y}))^{1/2}, 0 < \lambda_1 < 1 \]

and

\[ \lambda_2 = g(\Phi(\bar{Y}_i - \bar{Y}))^{1/2}, 0 < \lambda_2 < 1 \], obtained from distribution of order statistic.

### 2.1 Distribution of Harmonic Variance

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

\[ Y^* \sim \min(\bar{X}_i - \bar{X}) - \lambda_1 N(\mu_i - \mu, \sigma_H^2 (\frac{n_i + n}{n_i n})) \] .................................(2.9)

and

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

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

\[ t = \frac{Y_i}{Z} \] .................................(2.11)

where

\[ Y_i = (\bar{X}_i - \bar{X}) \] .................................(2.12)

and

\[ Z = \sqrt{\frac{S_H^2 (\frac{1}{n_i} + \frac{1}{n})}{\lambda}} \] .................................(2.13)

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

where \( \lambda \) can be \( \lambda_1 \) or \( \lambda_2 \) 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).
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

<table>
<thead>
<tr>
<th>Response</th>
<th>Levene Statistic</th>
<th>df₁</th>
<th>df₂</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10.975</td>
<td>3</td>
<td>36</td>
<td>0.000</td>
</tr>
</tbody>
</table>

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:**
\[ \bar{x}_A = 53.6, S^2_A = 250.04, n_A = 10 \]

**Zone B:**
\[ \bar{x}_B = 15.8, S^2_B = 15.7, n_B = 10 \]

**Zone C:**
\[ \bar{x}_C = 15.0, S^2_C = 13.9, n_C = 10 \]

**Zone D:**
\[ \bar{x}_D = 15.0, S^2_D = 16.7, n_D = 10 \]

**Therefore,** we consider the minimum and maximum differences of means respectively as given below:

\[ Y_1 = 53.6 - 24.9 = 28.7 \]
\[ Y_2 = 15.8 - 24.9 = -9.1 \]
\[ Y_3 = 15.0 - 24.9 = -9.9 \]
\[ Y_4 = 15.0 - 24.9 = -9.9 \]

\[ S^2_H = \left( \frac{1}{4} \sum_{i=1}^{4} \frac{1}{\delta_i^2} \right)^{-1} \]

\[ S^2_H = 20.05 \]

Then, the minimum difference of means is

\[ Y^* = \min(\bar{x}_i - \bar{x}) = (\bar{x}_C - \bar{x}) = -9.9 \]

In the above data set, \( n_i = 10 \), \( g = 4 \), \( n = \sum_{i=1}^{4} n_i = 40 \)

\[ S^2_H = \left( \frac{1}{4} \sum_{i=1}^{4} \frac{1}{\delta_i^2} \right)^{-1}, \quad S^2_H = 20.05 \]

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
\[ H_0 : \mu_i - \mu = 0 \quad \text{vs} \quad H_1 : \mu_i - \mu < 0 \]

\[ t = \frac{\min(X_i - \bar{X})}{S_H \sqrt{\left(\frac{1}{n_i} + \frac{1}{n}\right)}} \sim t_r \]

\[ = \frac{-9.9}{4.478 \sqrt{\left(\frac{1}{10} + \frac{1}{40}\right)}} = \frac{-9.9}{1.5832} = -6.25 \]

where \( r = 22.096 + 0.266(n - g) - 0.000029(n - g)^3 \)

\[ \lambda_2 = g(\Phi(\bar{Y}_i - \bar{Y}))^{1/3} = 0^+, \quad 0 < \lambda_2 < 1 \quad \text{from equation (2.1)} \]

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

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

\[ = P(t_r < -\infty) \]

\[ = 0 \]

< 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

\[ Y^* = \max(X_i - \bar{X}) = (\bar{X}_A - \bar{X}) = 28.7 \]

In the above data set, \( n_i = 10, \quad g = 4, \quad n = \sum_{i=1}^{4} n_i = 40 \)

\( s_H^2 = \left(\frac{1}{4} \sum_{i=1}^{4} \frac{1}{s_i^2}\right)^{-1} \)

\( s_H^2 = 20.05 \)

The hypothesis to be tested is
\[ H_0 : \mu_i - \mu = 0 \quad \text{against} \quad H_1 : \mu_i - \mu > 0 \]

\[ t = \frac{\max(X_i - \bar{X})}{s_H \sqrt{\left(\frac{1}{n_i} + \frac{1}{n}\right)}} \sim t_r \]
\[ r = 22.096 + 0.266(n - g) - 0.000029(n - g)^2 \]
\[ \lambda_i = g(1 - \Phi(\bar{Y}_i - \bar{Y}))^{g-1} = 4.0, \; 0 < \lambda_i < 1 \]

from equation (2.1)

Now \( p \)-value = \( P(t_r > t) = P(t_r > \frac{t_{cal}}{\lambda_i}) \)

\[ = P(t_r > \frac{18.13}{4.0}) \]
\[ = P(t_r < \frac{4.5325}{10^{-5}}) \]
\[ < 0.025 \]

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 Behrens- Fisher’s problem.

### REFERENCES


