East Asian Journal of Multidisciplinary Research (EAJMR)
Vol. 2, No.3 2023: 1079-1084

On The Use of Analysis of Variance under Unequal group variances

Abidoye, A. O[1*], Egburonu O. D[2]

[1]Department of Statistics, University of Ilorin, Ilorin, Nigeria, [2]Centre for Critical Thinking and Research, CLAPAI Orphanage High School, Jos, Nigeria

Corresponding Author: Abidoye, A. O abidoye@unilorin.edu.ng

A R T I C L E I N F O

Keywords: Harmonic Mean of Variances, Analysis of Variance (ANOVA) ,Chi-square Distribution, Modified T–test Statistic

Received : 05, January
Revised : 10, February
Accepted: 15, March

©2023 Abidoye, Egburonu: This is an open-access article distributed under the terms of the Creative Commons Atribusi 4.0 Internasional.

A B S T R A C T

In this study, we imposed Analysis of variances test (ANOVA) which use when we have more than two treatments or different levels of a single factors that we wish to compare then we assume homogeneity of variances across the groups being compared although most of the earlier works that have addressed the problem of testing equality of mean variance overestimates the appropriate variance and the test statistic becomes conservative. This is the well-known Behrens – Fisher problem. Then we are interested in comparing several treatments means in this work, we made use the analysis of variance under unequal variances when the groups variances differ. It will be very inappropriate to use the pooled sample variance (S_P^2) as a single value for the variances, instead the sample harmonic mean of variances (S_H^2) is proposed as an alternative to the pooled sample variance when there is heterogeneity of variances. The distribution theoretically and confirmed using simulation studies and this proposed harmonic mean of variance was , examined in this work and found useful for unequal variances.
Data set from Kwara State Ministry of Health on the incidence of diabetes diseases for male patients was used to illustrate the relevance of our proposed test statistic.

INTRODUCTION

Analysis of Variance is one of the most popular models in statistics. In general, interest is in testing the homogeneity of the different group means using the classical Analysis of Variance (ANOVA). However, the standard assumption of homogeneous error variances which is crucial in ANOVA is seldomly met in statistical practice. In such a case one has to assume a model with heteroscedastic error variances. Analysis of variance assumes that the sample data sets have been drawn from populations that follow a normal distribution. One-way analysis of variance for example, evaluates the effect of a single factor on a single response variable. For example, a clinician may be interested in determining whether or not there are differences in the age distribution of patients enrolled in different study groups. To satisfy the assumptions, the patients must be selected randomly from each of the population groups, a value for age for the response variable is recorded for each sampled patient, the distribution of the response variable can then be assumed to be normally distributed. See Ott, 1984 and Abidoye (2012).

The analysis of variance uses a linear regression approach and consequently supports unequal sample sizes. This is important because designers of experiments seldom have complete control over the ultimate sample sizes in their studies. However, two-way analysis of variance for example does not support empty cells (factor levels with no sample data points). Each of the factors must have two or more levels and the factor combination must have one or more sample observations.

The conventional analysis of variance (ANOVA) is also based on the assumption of normality, independence of errors and equality of the error variances. Studies have shown that the F-test is not robust under the violation of equal error variances, especially if the sample sizes are not equal and some authors have developed an exact Analysis of variance for testing the means of g independent normal populations by using one or two stage procedures. See Jonckheere (1954), Dunnett (1964), Montgomery (1981), Dunnet and Tamhane (1997), Yahya and Jolayemi (2003).

This test procedure having a specified type one error rate of α should be powerful to determine whether the differences among the sample means are large enough to imply that the corresponding population means are different. See Abidoye et.al (2015a, 2015b).

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, $i = 1,2,...,g$ ……(2.1)

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

If we define $\delta_i = \mu_i - \mu$, ……(2.2)

then,

$H_0 : \delta_i = \mu_i - \mu = 0 \quad \forall_i \; vs \; H_1 : \delta_i \neq 0$ for at least one i

The unbiased estimate of δ_i is

$\hat{\delta}_i = \overline{X}_i - \overline{\overline{X}}$ ……(2.3)

Therefore

$\hat{\delta}_i \sim N[\delta_i, V(\delta_i)]$

where

$$V(\delta_i) = V(Y_i)$$
$$= V(\overline{X}_i - \overline{\overline{X}})$$
$$= V(\overline{X}_i) + V(\overline{\overline{X}}) - 2\,\mathrm{cov}(\overline{X}_i, \overline{\overline{X}})$$
$$= V(\overline{X}_i) + V(\overline{\overline{X}}) - 2\,\mathrm{cov}(\overline{X}_i, \frac{\sum \overline{X}_i}{g})$$
$$= V(\overline{X}_i) + V(\overline{\overline{X}}) - 2\,\mathrm{cov}(\overline{X}_i, \frac{\overline{X}_i}{g})$$
$$= \frac{\sigma_i^2}{n_i} + \frac{\sigma_H^2}{n} - \frac{2}{g}V(\overline{X}_i)$$
$$= \frac{\sigma_i^2}{n_i} + \frac{\sigma_H^2}{n} - \frac{2\sigma_i^2}{gn_i}$$, see Abidoye (2012) and Abidoye et.al (2015b)
$$= \frac{\sigma_H^2}{n} + \left(1 - \frac{2}{g}\right)\frac{\sigma_i^2}{n_i}$$ ……(2.4)

Therefore,

$$V(\delta_i) = \frac{\sigma_i^2}{n_i} + \frac{\sigma^2}{n}$$ ……..(2.5)

$$= \sigma_H^2\left(\frac{1}{n} + \frac{1}{n_i}(1 - \frac{2}{g})\right)$$ ……..(2.6)

$$\hat{\sigma}_H^2 = S_H^2$$
$$= \left[\frac{1}{g}\left(\frac{1}{s_1^2} + \frac{1}{s_2^2} + \ldots + \frac{1}{s_g^2}\right)\right]^{-1}$$

where S_i^2 is the variance of the ith group. S_H^2 has r degrees of freedom, where r = 22.096 + 0.266(n-g) – 0.000029(n-g)^2 as defined in Abidoye et. al (2015a).

S_H^2 is likened to the common variance as defined in one – way Analysis of variance ANOVA. The one – way ANOVA is presented in Table 1 below. In the table one the sum of squares between groups and adjusted sum of squares do not necessarily sum to total sum of squares, also the degrees of freedom.

RESULTS AND DISCUSSION

Table 1: One – way analysis of variance with unequal group variances

Source variation	d.f	SS	MS	F
Treatment (between groups)	t-1	$SSt = \frac{\sum Y_{i.}^2}{t} - \frac{(\sum Y_{ij})^2}{n}$	SSt/t-1 =MSSt	MSS_t / MS_H^2
Adjusted error S_H^2	r	$r \times S_H^2$		
Total	n-1	$SST = \sum Y_{ij}^2 - \frac{(\sum Y_{ij})^2}{n}$		

where $r = \Omega = [r+½]$

Small simulations show that without any loss of generality nearest integer values can be used to approximate the r degrees of freedom.

APPLICATION

The data used in this study is applicable in Medicine; the data used were 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 2: Showing the incidence 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	37	80	58	48	35	46	53	39	64	76
Zone B	14	19	12	21	23	13	15	16	11	14
Zone C	15	18	11	19	22	14	13	15	10	13
Zone D	11	19	10	18	23	12	14	16	12	15

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

$$H_0 : \sigma_A^2 = \sigma_B^2 = \sigma_C^2 = \sigma_D^2 \quad \text{vs} \quad H_1 : \sigma_i^2 \neq \sigma_j^2 \quad \textit{for atleast} \quad (i, j),$$

$$i = A, B, C, D$$

$$j = A, B, C, D$$

Table 3: Levene test for variance equality

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

Since P- value < 0.05 we therefore reject H_0 and therefore conclude that the variances are not equal.

Computation on incidence of diabetes diseases for male patients: From the data above the following summary statistics were obtained:

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

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

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

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

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

$$S_H^2 = \left(\frac{1}{4}\sum_{i=1}^{4}\frac{1}{s_i^2}\right)^{-1} \qquad S_H^2 = 20.05$$

The main hypothesis 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 $i = A, B, ..., D$

$$SST = \sum Y_{ij}^2 - \frac{(\sum Y_{ij})^2}{n}$$

$=37^2+80^2+\ldots.+12^2+15^2-7182.4$

$=31,209.6$

$$SSt = \frac{\sum Y_{i.}^2}{t} - \frac{(\sum Y_{ij})^2}{n}$$

$= 35726 - 7182.4$

$= 28543.6$

SSE = SST – SSt

= 31209.6 – 28 543.6

= 2666

ANALYSIS OF VARIANCE TABLE

Source variation	d.f	SS	MS	F
Between groups	3	28543.6	9514.5	474.54
Ajusted error	31.63	634.18	20.05	
Total	39	31209.6		

$$F_{3,31.63,\alpha} \doteq F_{3,32,\alpha} = 8.62$$

Analysis of variance shows that incidence rate of diabetes diseases in the four zones are significantly different at 5% level of significance which support the earlier results (Abidoye et. al 2015b), even though the sum of squares due to zones plus the adjusted error sum of square did not sum up to total sum of squares.

CONCLUSIONS AND RECOMMENDATIONS

In this work we have established that ANOVA test may suffice if the adjusted error (the Harmonic mean of the population variances) is known. The sum of square treatments and sum of square of adjusted error may not necessarily sum up to sum of squares; and the adjusted error degrees of freedom need not necessarily be an integer.

REFERENCES

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.

Abidoye, A.O, Jolayemi, E.T, Sanni, O.O.M and Oyejola, B. A. (2015a): Development of Test Statistic for Testing Equality of Means Under Unequal Population Variances. Ilorin Journal of Science. Accepted for publication by Faculty Of Physical Sciences, University of Ilorin a.

Abidoye, A.O, Jolayemi, E.T, Sanni, O.O.M and Oyejola, B. A. (2015b): Development of Hypothesis Testing on Type one Error and Power Function. Ilorin Journal of Science. Accepted for publication by Faculty Of Physical Sciences, University of Ilorin.

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.

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

Jonckheere, A.R (1954): "A distribution-free k-sample test against ordered alternative." Biometrical, 41, 133-145.

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

Ott, L (1984): "An Introduction To Statistical Methods and Data Analysis". Second edition. P.W.S Publisher. Boston.

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