ISSN No:-2456-2165

Alternatives Test in ANOVA with Unequal Variance and Unequal Sample Size

Júlio A. Mabuie
Zambeze University
Mozambique

Abstract:- This article presents a discussion about the methods used in experimental statistics, which considers a case study in which we have experimental units with unequal sample sizes. For this type of experimentation, non-observation of certain assumptions underlying the carrying out of the associated tests is observable, forcing the researcher to opt for alternative methods. With the aim of comparing the average pedagogical performance of students from different groups of students, a database was considered, made available in three groups of students according to the distances travelled to school. Given the observable violation of the assumptions of equality of variances and normality, for the analysis and discussion of the results, some non-parametric tests were described, for multiple comparisons and post-hoc tests, which are pairwise comparisons of means.

Keywords:- *Analysis of Variance, Unequal Sample Size, and Non-Parametric Test.*

I. INTRODUCTION

In several areas of knowledge, research focuses on experimental techniques, such as agricultural sciences, health sciences, and others. Experimental statistics, in most cases, is focused on comparing average performance in two or more groups (treatments), using the conventional test (F−statistics), assuming that some relevant assumptions are satisfied (Douglas [1]). In fact, the F−statistic, as the appropriate test for multiple comparisons, assumes that the observations in K treatments have a normal distribution and that equality of variances is observed in the K treatments. In certain fields of research, it is common for experimental units to not be balanced, and therefore, cases may be observed in which differences in treatment sizes are high (half or more), or even as small as possible. Experiments with unbalanced experimental units usually present certain problems in the analysis of variance (ANOVA) procedure (Brien [2]), since, in most cases, the underlying assumptions, such as normality and equality of variances, are not verified and it is very common that the assumptions underlying the test are influenced by these sample differences, or even caused by small deviations in variances. In variance analysis, the assumptions of normality and homoscedasticity are mandatory (Robert [3]), therefore, if the underlying assumptions have not been verified, the researcher can opt for data transformation methods (George [4]), or by using non-parametric methods, as a reasonable alternative (Ghosh [5]). Some authors propose the Kruskal and Wallis test (1952), Welch’s t-test (1947), Brown and Forsythe (1974), for multiple comparisons and the Gomes and Howell test (Roxton [6]) as a post-hoc test, as non-parametric tests. alternatives. Because it is thought that certain studies with unbalanced treatments do not adhere to rigorous statistical standards (Rosenbaum and Rubin [7]). Data transformation correction methods are difficult to implement when treatments have different sample sizes, (George [4]). Non-parametric tests, as frequently used alternatives, are powerful tools for comparing treatments, as their procedure is based on comparing medians with the general average of treatments (Fernandez [8]). With the aim of comparing the average pedagogical achievement of basic education students in rural areas, this article carries out an analysis in experimental statistics, in k groups of students, bringing a discussion about the validation of the results, taking into account certain conditions of the data.

II. ANALYSIS OF VARIANCE (ANOVA)

In this chapter, methods for designing and analyzing one-factor experiments with an arbitrary number of factor levels (treatments) are developed. For an experiment that has been completely randomized, in which we have a single factor wanting to compare different treatments, then the response observed in each of the treatments is a random variable, (Douglas [1]). Analysis of variance can be classified as unidirectional when comparing three or more categorical groups and bidirectional when comparing several groups of two factors (Anindya [5]). Normally, one-way analysis of variance in its structure presents replications, while a two-way analysis can be with or without replications For a one-way analysis of variance, give a table (1) with the observations yij that represents the j−th observation taken under treatment i, then overall there will be n observations under the i − th treatment:

Table 1 Typical Data for a Single-Factor Experiment

Trat	**1**	**2**	**…**	**n**	**Total**	**Average**
1	y_{11}	y_{12}	…	y_{1n}	$y_{1.}$	$\bar{y}_{1}$
2	y_{21}	y_{22}	…	y_{2n}	$y_{2.}$	$\bar{y}_{2.}$
…	…	…	…	…	…	…
K	y_{k1}	y_{k2}	…	y_{kn}	$y_{k.}$	$\bar{y}_{k.}$
					$y_{..}$	$\bar{y}_{..}$

A. ANOVA Model

The average model to describe the observations of an experiment is expressed as follows:

$$y_{ij} = \mu_i + \varepsilon_{ij} \begin{cases} i = 1,2,3,...,k \\ j = 1,2,3,...n \end{cases} \tag{1}$$

Where

- y_{ij}, represents the j-th observation
- μ_i , represents the average of treatment i

➢ *Random Error* ε_{ij}

The random error component includes all other sources of variability in the experiment, such as measurement errors, variability arising from uncontrollable factors, differences between experimental units (such as test material, etc.) to which treatments are subjected, and the general background noise in the process (such as variability over time, effects of environmental variables and others), (Douglas [1]). It is always convenient to think of errors as having a mean of zero, then:

$$\begin{cases} E\lfloor y_{ij} \rfloor = \mu_i \\ \varepsilon_{ij} = 0 \end{cases} \tag{2}$$

Thus the model (1) is called Model of Means and an alternative way of writing a model (1), for the data is define as:

$$y_{ij} = \mu + \tau_i \quad \text{Where} \quad i = 1,2,3,...,k \tag{3}$$

Then the model (1) can be written as follows:

$$y_{ij} = \mu_i + \tau_i + \varepsilon_{ij} \begin{cases} i = 1,2,3,...,k \\ j = 1,2,3,...n \end{cases} \tag{4}$$

By (Robert [3]), for the model (4), the parameter μ becomes a common parameter for all treatments, designated as the general average, while τ is exclusively a parameter for *ith* treatment, called the effect of the *ith* treatment, and therefore the model (4), is called Effects Model (Douglas [1]). Therefore, both the Average Model and Effects Model are linear statistical models, since the response variable y_{ij} is a linear function of the model parameters (Douglas [1]).

Although both forms are useful, the Effects Model is widely mentioned in the experimental design literature, for the intuitive reason that μ, is a constant and the treatment effects i, represent the deviations from this constant when specific treatments are applied. Both the equation (1) and the equation (4) are also called single-factor ANOVA, in which the experiment is assumed to be carried out in random order so that the environment in which the experimental units are applied, be the same, then the experimental design is a completely casual design, (Robert [3]).

The objective of the experiment will be to test hypotheses about treatment means and estimate them, assuming that errors are independent random variables and normally distributed with zero mean and constant variance within groups. Which implies that the observations have:

$$y_{ij} \rightarrow N(\mu_i + \tau_i, \sigma^2) \quad \text{Where} \quad i = 1,2,3,...,k \tag{5}$$

B. Variance Analysis with Fixed Factors

As can be seen in the table (5), in a model with a single factor, in which the total of observations under the *ith* treatments, the means of the *ith* treatments and the general average are represented. In this case, the interest of the analysis is to test the equality of the means of the k treatments, which can be described as follows:

$$E\lfloor y_{ij} \rfloor = \mu + \tau_i = \mu_i \quad \text{Where} \quad i = 1,2,3,...,k \tag{6}$$

Or by hypothesis test:

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \mu_3 = ... = \mu_k \\ H_0 : \exists! \mu_i \neq \mu_{i+1} \forall i = 1,2,3,...,k \end{cases} \tag{7}$$

In fixed effects models, the average of the *ith* treatment and two components is defined, such as in equation (6) and we generally consider μ to be the general average, by (Douglas [1]):

$$\frac{\sum_{i=1}^{k} \mu_i}{k} \rightarrow \sum_{i=1}^{k} \tau_i = 0 \tag{8}$$

The effects of the treatment or factors can be considered as deviations from the general average, therefore, the hypotheses (7) can be represented in order to test the effects of the treatments:

$$\begin{cases} H_0 : \tau_1 = \tau_2 = \tau_3 = \ldots = \tau_k = 0 \\ H_0 : \exists! \tau_i \neq 0, \forall i = 1,2,3,\ldots,k \end{cases} \tag{9}$$

C. *Decomposition of the Sum of Total Squares*

The total identification of ANOVA indicates that the total variability of the data, which is measured by the corrected total sum of squares, can be partitioned into a sum of squares of the differences between the treatment means and the general mean, plus a sum of squares of the differences in observations within treatments in relation to the treatment mean. Therefore, the difference between the observed treatment means and the overall mean is a measure of the differences between the treatment means that may be due solely to random error. Considering the table (5), we have the following decomposition:

- *Sum of Squares between Treatments*

$$SS_{Tratament} = n\sum_{i=1}^{k}\left(y_{ij} - \bar{y}_{..}\right)^2 \tag{10}$$

- *Sum of Squares within Treatments*

$$SS_{Error} = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{i.}\right)^2 \tag{11}$$

- *Sum of Total Squares*

$$SS_{Total} = SS_{Tratament} + SS_{Error} \tag{12}$$

- *Degrees of freedom*

In estimating the sum of total squares, the total number of observations, kn = N, is considered, then in estimating the sum of average total squares, we admit the degree of freedom, N − 1 in estimating the sum of squares between treatments, if we consider the total number of treatments, *k* then in estimating the sum of squares between the average treatments, we admit the degree of freedom *k − 1* and finally, there are n replicates within the *k* treatments that provide *n − 1* degrees of freedom, to estimate experimental error. Due to the existence of *k* treatments, we have the degrees of freedom to estimate the sum of squares of the mean errors:

$$k(n-1) = kn - k = N - k \tag{13}$$

- *Sum of Squares between Average Treatments*

$$MS_{Tratament} = \frac{n\sum_{i=1}^{k}\left(y_{ij} - \bar{y}_{..}\right)^2}{k-1} = \frac{SS_{Tratament}}{k-1} \tag{14}$$

For cases where the means between treatments are equal, the equation (14), estimates the population variance or the variance of all treatments σ^2 Douglas [1].

- *Sum of Mean Squares within Average Treatments*

$$MS_{Error} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{i.}\right)^2}{N-k} = \frac{SS_{Error}}{N-k} \tag{15}$$

Likewise, the equation (15), is a pooled estimate of the common variance within each of the k treatments and also, if the means within the treatments are equal to the equation (15), estimates the population variance or the variance of all treatments σ^2. Note that, if there is no difference between the treatment means, both $MS_{Tratament}$ and MS_{Error} will be an estimate of the population variance σ^2 , which implies τ = 0. Therefore, it can be noted that, for the hypothesis test (7), "if there is a difference between the means of the treatments, the expected value of the mean squares of the treatment will be greater than the population variance σ^2 ", (Robert [3]). By (Jean [9]), the hypothesis test (7) can be carried out by comparing the two estimates, which are, (14) and (15) , if there is equality between the Treatment averages these estimates will be equal. As can be seen, the identity of ANOVA provides two estimates of population variance σ^2, one based on variability between treatments and the other based on variability within treatments (Douglas [1]).

$$\exists! \mu_i \neq \mu_j \Rightarrow E\left[MS_{Tratament}\right] \succ \sigma^2 \tag{16}$$

D. *Statistical Analysis of Fixed Effects Model*

The hypothesis test presented (7), takes on that within the groups, the errors ε_{ij} are independent and have a normal distribution, with zero mean and constant variance, and the observations y_{ij} , are independent and normally distributed with mean $\mu + \tau_i$ and constant variance, then $SS_{Tratament}$, is a sum of the squares of the normally distributed random variable. In statistical analysis, it can be proven that if the null hypothesis is true, then (George [4]):

Table 2 ANOVA in Fixed Effects Models

Source	SQuar	Degree	QAverage	F-Stat
BetwenTrat	$SS_{Tratment}$	$k-1$	$MS_{Tratment}$	F_0
InTrat	SS_{Error}	$N-k$	MS_{Error}	
Total	SS_{Total}	$N-1$	$\cdots$	$\cdots$

$$F_0=\frac{MS_{Tratament}}{MS_{Errror}} \tag{17}$$

If the null hypothesis is false, then by (Robert [3]), the expected value of the numerator in (17), will be greater than the variance, and therefore we reject the null hypothesis, that is, there are significant differences between the treatment averages, and then:

$$F_0 \succ F_{\alpha,k-1,N-k} \tag{18}$$

Note that, in the ANOVA table (2), the calculations presented can be obtained using already designed computational methods.

E. *Estimation of Model Parameters*

If we have two populations with means μ_1 and μ_2, with variances σ_1^2 and σ_2^2, respectively, an estimator of the difference between μ_1 and μ_2 is provided by the statistic $\bar{x}_1-\bar{x}_2$ Therefore, to obtain a point estimate of $\mu_1-\mu_2$ and calculate the difference $\bar{x}_1-\bar{x}_2$, of the sample means. Clearly, we must consider the sampling distribution. The confidence interval for the means of two populations, considering unknown population variances, we use the sample means $\bar{x}_2$ and $\bar{x}_1$ of two independent random samples, respectively from approximately normal populations, in which, a confidence interval of 100(1 − α)% is given by:

- *For Approximately Normal Populations with Unknown but Equal Variance*

$$\bar{x}_2-\bar{x}_1-t_{\frac{\alpha}{2}}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \prec \mu_2-\mu_1 \prec \bar{x}_2-\bar{x}_1+t_{\frac{\alpha}{2},}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \tag{19}$$

- *For Approximately Normal Populations with Unknown but Different Variance*

$$\bar{x}_2-\bar{x}_1-t_{\frac{\alpha}{2}}\sqrt{\frac{S_1}{n_1}+\frac{S_2}{n_2}} \prec \mu_2-\mu_1 \prec \bar{x}_2-\bar{x}_1+t_{\frac{\alpha}{2},}\sqrt{\frac{S_1}{n_1}+\frac{S_2}{n_2}} \tag{20}$$

For an estimated model, the ANOVA model estimators, when estimating the confidence interval, note that the general mean is estimated by the general mean of the observations and that any treatment effect is just the difference between the treatment mean and the overall average (Robert [3])

$$y_{ij}=\mu_i+\tau_i+\varepsilon_{ij} \rightarrow \begin{cases}\hat{\mu}=\bar{y}_{..}\\ \hat{\tau}=\bar{y}_i-\bar{y}_{..}\end{cases} \tag{21}$$

When estimating the confidence interval, it is assumed that the errors are normally distributed with zero mean and constant variance, each treatment mean (Robert [3]), and in violation of some assumptions we resort to non-parametric methods;

$$\bar{y}_i \rightarrow ND\left(\mu_I,\frac{\sigma^2}{N}\right) \tag{22}$$

Thus, knowing the variance σ^2 , estimate the confidence interval, using the normal distribution, and in the case of unknown variance, we use MS_{Error} as an estimator of the variance with t-student distribution, to test the difference in means between two treatments.

F. *Unequal Sample Size to ANOVA*

In several studies, such as in the areas of medicine, veterinary, social impact studies and others, it is possible to observe that in most cases, the researcher has been forced to carry out certain experimental statistics on treatments with different sizes. In the example of health sciences areas, it may be necessary to carry out analysis of different types of hepatitis transmission in patients diagnosed in a given period, in order to compare some attribute associated with the patient's pathological history. If the diagnostic period is taken into account, the experimenter may have in his database differences in observations in different types of transmission. (Hepatitis A, Hepatitis B and Hepatitis C)[1] . According to (Douglas [1]). "in experiments with equal sample sizes, the power of the test is maximized", but this concept is not analytically substantiated, due to the nature of the specific experiment. In the example of comparing the average grade in four second-year classes at the Faculty of Engineering, even if there are differences in the number of students, it does not take away the merit of the experiment, and in this example, random effects are negligible, as the treatments are exposed in similar environments. For treatments with unbalanced observations, the analysis of variance described in equations (10) and (11), small modifications to the sum of squares formula, must be made.

$$N=\sum_{i=1}^{k}n_i \tag{23}$$

[1] Hepatitis is classified according to the type of transmission

$$SS_{Tratament} = \sum_{j=1}^{n} \frac{y_{ij}^2}{n_i} - \frac{y_{i.}^2}{N} \tag{24}$$

$$SS_{Total} = \sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}^2 - \frac{y_{i.}^2}{N} \tag{25}$$

- *The literature mentions two advantages in experiments with a balanced design George [4]*

- The test statistic is relatively insensitive to small deviations from the assumption of homogeneity in treatments with an equal number of replicates.
- Test power is maximized if samples are of equal size

G. Model Validation

The decomposition of variability in observations through an analysis of variance identification (12), is purely algebraic, therefore the use of the partition to formally test the absence of differences in treatment means, requires that certain assumptions be satisfied (Douglas [1]). The question of model validation would be that the observations are adequately described by the estimated model? Given the equation (4), then the errors are independent and normally distributed with zero mean and constant variance.

$$\varepsilon_{ij} \to ND(0, \sigma^2) \tag{26}$$

If these assumptions are not verified, the use of analysis of variance is not a reliable test for differences between treatment means (Johnson [10]). Violation of basic assumptions and model adequacy can simply be investigated by residual analysis (Robert [3]). The residuals are defined as observation *j* of treatment *i*, as follows:

$$\varepsilon_{ij} = y_{ij} - \hat{y}_{ij} \quad \text{Where}$$
$$\hat{y}_{ij} = \hat{\mu} + \tau_i = \bar{y}_{..} + (\bar{y}_i - \bar{y}_{..}) = \bar{y}_i \tag{27}$$

The expression (27), offers an attractive result that the estimate of any observation y_{ij} in the treatment, is just the average of the corresponding treatment (Edison [11]). Verification of the diagnosis can be easily done by graphical analysis of the residues, which can adopt other ways of treating several common anomalies, such as discrepant data (Ranghuthan [12]). For variance analysis, it is more effective to construct the graph of the residuals, in which, if the distribution of the underlying errors is normal, this graph will resemble a straight line, (Ranghuthan [12]). The use of non-parametric tests of adherence to the normal distribution, such as Pearson's Chi-Square (QQ), Kulmogorov-Smirov (KS), Shapiro Walks (SW) tests and others, is more objective, (Jean [9]). Due to the fact that F-Statistic test is slightly affected by the assumption of normality, ANOVA or other procedures related to multiple comparisons are robust to the assumption of normality (Robert [12]), and it is still necessary to implement the Levene or Barttlet test as homogeneity tests (Anindya [5]).

H. Data Transformation

A practical method to solve the problem of violating the assumptions of normality and homoscedasticity, is the transformation of Box Cox, (George [4]) which is carried out through an exponent, lambda (λ) that varies between −5 and 5, and can take any range within it. To perform the data transformation, all values of λ are considerable, and an optimal value of λ is selected, which results in the best approximation of a normal distribution curve (28).

$$y(\lambda) = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda} \to \lambda \neq 0 \\ \ln \lambda \to \lambda = 0 \end{cases} \tag{28}$$

For experimental data, the Box-Cox transformation presents certain constraints in cases where treatments are not balanced, or k treatments with differences in the number of observations. An analysis of variance, in which data transformation is not applicable as a way of solving the problem of normality and equality of variances, alternative non-parametric methods can be used (Anyndia [5]).

III. NON-PARAMETRIC TESTS IN ANOVA

A non-parametric test used to replace the t test is the Mann-Whitney test, considering a smaller number of observations within the groups and/or violating the assumption of normality (Kolmogorov-Smirnov and Shapiro-Wilk), in which homogeneity in the groups is assumed, using the Levene test (Edison [11]). The non-parametric Kruskal and Wallis test is a reasonable alternative to the F statistic in multiple comparisons, when a violation of the relevant underlying assumptions is observed, (Walpole [13]). By (Douglas [1]), Kruskal and Wallis test, it is a non-parametric test, used to test the null hypothesis that the k treatments are identical in relation to the alternative hypothesis that some treatments generate a number of observations greater than the others. Although the procedure was designed to be sensitive to the test of differences in means, it is sometimes convenient to admit the Kruskal and Wallis test as a test of equality of means in k treatments, therefore, it is a non-parametric alternative to analysis of variance usual (Jean [9]). An interesting method for estimating confidence intervals simultaneously for r pairs of treatment means (post-hoc) is the Boferroni Method, (Douglas [1]) which allows the experimenter to construct a set of r simultaneous confidence intervals for pairwise differences in treatment means, in which the confidence level is at least 100 ∗ (1−r ∗α). For a not very large r, this is a very good method that leads to reasonably short confidence intervals. In the example of having 3 intervals with a significance level of 5%, the confidence level will be 100*(1- 3*0.05)% = 85% (Douglas [1]). Other non-parametric tests for comparing pairs of means are Fisher's LSD procedure, *Tukey' HSD* test and others, which consist of paired t tests, each with a significance chosen to control the experimental error rate.

A. *Kruskal and Wallis Test Procedure (Rank−Test)*

The procedure used to implement the *Kruskal and Wallis* test starts by sorting the observations y_{ij} in ascending order and replacing each observation with its order. If there are observations with the same value, the average order is assigned to each of the linked observations, (Eduard [14]). So, let R_{ij} be, with the lowest rank it will be of the first order, where R_i is the sum of the orders in the *ith* treatment and the test statistic is:

$$H = \frac{1}{S^2}\left[\sum_{i=1}^{k}\frac{R_i^2}{n_i} - \frac{N(N+1)^2}{4}\right] \tag{29}$$

$$S^2 = \frac{1}{N-1}\left[\sum_{i=1}^{k}\sum_{j=1}^{n} R_{ij}^2 - \frac{N(N+1)^2}{4}\right] \tag{30}$$

For cases where there are no observations with the same value, we have:

$$S^2 = \frac{N(N+1)^2}{12} \tag{31}$$

$$H = \frac{12}{N(N+1)}\sum_{i=1}^{k}\frac{R_{ij}^2}{n_i} - 3(N+1) \tag{32}$$

Where:

- N – Total number of observations.
- n_i - Number of observations in the *ith* treatment
- S^2 - Variance of sorted orders
- R_i - The sum of orders classified in the *ith* treatment

For a number of observations in the *ith* treatment that is reasonably larger, the test statistic (H) follows a chi-square distribution with degrees of freedom k−1, under the null hypothesis of equality of means between the k treatments.´

$$H \succ QQ_{k+1} \tag{33}$$

For the case in which the expression (33) is true, then it rejects the null hypothesis of equality of means between the k treatments. The Kruskal and Wallis test, which consists of replacing observations with their ordinary classifications, is called Transformation by Order, and is widely useful, since applying the common F test for classifications by order and not in original data, we would have as test statistics (34) (Conover [15]):

$$F_0 = \frac{\dfrac{H}{k-1}}{\dfrac{N-1-H}{N-k}} \tag{34}$$

It can be observed that both the Kruskal-Wallis (H) increases or decreases, F0 also increases or decreases, so the Kruskal-Wallis test is equivalent to the application of the usual analysis of variance in the classifications, (Douglas [1]). Violation of the assumption of normality may be caused by the effect of outliers, (Iman [16]), therefore, it is recommended that the usual variance analysis be performed on both the original data and the classifications and (George [4]) when both procedures give similar results, the assumptions of the analysis of variance are probably satisfied reasonably well, and the standard analysis is satisfactory. When the two procedures differ, the rank transformation should be preferred because the test is less likely to be skewed by non-normality and outlier observations. In such cases, the experimenter may want to investigate the use of transformations for non-normality[2] and also, examine the data through the experimental procedure to determine whether there are outliers (Fernandez [8]). Due to the need to check which pairs of groups have significantly different means, in multiple comparisons, we must analyze post-hoc tests, using non-parametric tests to replace the t-test, given the possible violation of the assumption of normality. Considering the homogeneity in the groups, the Tukey or Bonferroni LSD test can be chosen, as they are more rigorous (Edison [11]).

B. *Non-Homogeneity Estimation*

One of the assumptions of ANOVA, in the equality of means, is the equality of variances, although it is tolerant for small variance deviations, when we have equal sample sizes in k treatments (Douglas [1]). The ANOVA test can produce very misleading results in the presence of severe heterogeneity or unequal sample size. In these cases, different analysis approaches are proposed, one of which is the Welch´s t−test (Satterthwaite [17]) test, which is slightly different from the usual test (17). The Welch test is used in one-factor Analysis of Variance, as a generalization of the Student's t test, when the assumption of homoscedasticity is violated although it presupposes approximation to normality (Welch [18]). Considering k samples for the model (4), The F statistic proposed by Welch´s t−test is given by:

$$F_w = \frac{\left[\sum_{i=1}^{k} w_i\left(\bar{y}_{i.} - \bar{y}_{..}\right)\right]/(k-1)}{\left[1 + \dfrac{2(k-2)}{t^2-1}\sum_{i=1}^{k}\dfrac{1}{f_i}\left(1 - \dfrac{w_i}{\sum_{i=1}^{k} w_i}\right)^2\right]} \tag{35}$$

[2] By (George [4]), for unbalanced treatments, transformation methods are difficult to implement

Where:

$$w_i = \frac{n_i}{S_i^2}$$

$$f_i = n_i - 1$$

$$S_i^2 = \left[\frac{\sum_{j=1}^{n_i} \left(y_i - \bar{y}_{i.} \right)^2}{n_i - 1} \right]$$

$$\bar{y}_{..} = \frac{\sum_{i=1}^{k} w_i \bar{y}_{i.}}{\sum_{i=1}^{k} w_i}$$

The null hypothesis brings an approach to comparing the test statistics with degrees of freedom f_1 and f_2, as follows:

$$F_w \rightarrow F_{f_1, f_2} \text{ Such that } f_1 = k - 1 \text{ and}$$

$$f_2 = \left[\frac{3}{k^2 - 1} \sum_{i=1}^{k} \frac{1}{f_i} \left(1 - \frac{w_i}{\sum w_i} \right)^2 \right]$$

Note that the value of f_2 is approximated by default to an integer. For *k = 2* the procedure reduces to the Student’s t test with two samples. The Brown-Forsythe test is a statistical test for equality of group variances as Levene test (Antonio [19]) based on performing an Analysis of Variance (ANOVA) on a transformation of the response variable, therefore, it is the *F* statistic resulting from an ordinary one-way analysis of variance on the absolute deviations of groups or treatment data in relation to their individual medians (Brown and Forsythe [20]), assuming an approximation to normality. The test statistic for the null hypothesis (equality of means) is given by:

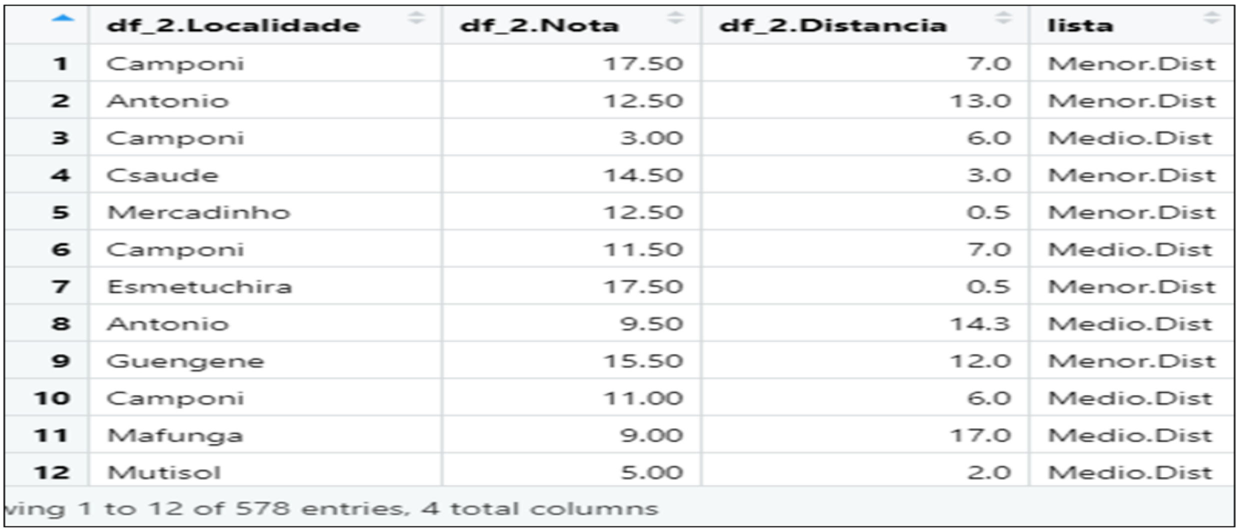
$$F_{BF} = \frac{\sum_{i=1}^{k} n_i \left(\bar{y}_{i.} - \bar{y}_{..} \right)^2}{\sum_{i=1}^{k} \left(1 - \frac{n_i}{n} \right) S_i^2} \qquad (36)$$

C. Variance Analysis and Adjacent Assumptions

- In one-way analysis of variance, if the assumption of normality is not violated in k treatments, but there is a violation of equality of variances, the usual analysis of variance test, can still be used to compare k means, resorting to non-parametric tests or even implement transformation methods to correct the data (Douglas [1])
- The non-parametric test Kruskal and Wallis, is widely used when in *k* samples, the assumption of normality is not verified, but there is similarity of distribution and variances (Anindya [5]).
- The non-parametric tests Welch’s t-test (35) and Brown and Forsyth test (36) present a reasonable alternative, considering a violation of the equality of variances (Brien [21]). Despite presenting very similar results, in the context of violating the assumption of homoscedasticity, if extremely high or low means are associated with small variations, the experimenter can use Welch’s t-test, and if extreme means are associated with variances larger, he can use Brown and Forsyth test, (George [4])
- The Games-Howell test is an improved version of the Tukey-Kramer test and is applicable in cases of violation of the assumption of equality of variances and is a t-test using Welch’s degree of freedom (Brien [2]). This method uses a strategy to control type I error for the entire comparison and is known to maintain the predefined significance level even when the sample size is different (Gearge [4]). However, the smaller the number of samples in each group, the more tolerant the type I error control is, and can be applied when the number of samples is greater than six (Brown [20]).
- For k samples with a distribution extremely different from normality and strongly heteroscedastic, the procedure for transforming the data to another distribution is recommended, which can satisfy the assumption of normality and homogeneity (George [4]). If outliers are causing non-normality and non-homogeneity, the researcher can correct the data, if he feel it has been reported incorrectly, such as using different units or missing decimals or decimals in the wrong place, can correct them or use non-parametric tests, which do not require assumptions (Conover [15]).
- One-way ANOVA is considered robust to moderate variance deviations, but unequal sample sizes affect the robustness of the homoscedasticity assumption (Keppel [22]). In fact, there is no good rule of thumb for how unequal sample sizes need to be for heterogeneity to be a problem (Rusticus [23]). A formal idea, but not generalizable, is the fact that in the existence of *K* treatments with different sizes, the possibility of reducing the different sizes based on the treatment with the smallest observations is considered (Keppel [22]). Therefore, if we have equal variances in the groups and unequal sample sizes, there is no problem, and if the variances are unequal and equal sample sizes, there is no problem (Rusticus [23]).

✓ *Example*

Data on the pedagogical performance of elementary school students, collected in a school in a rural area in the province of *Sofala− Mozambique*, were subjected to an experimental analysis, aiming to compare the average pedagogical performance of students from different locations.

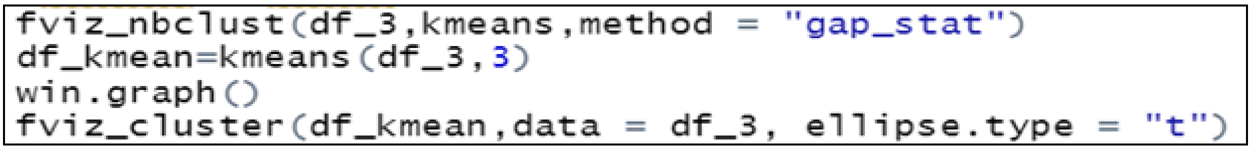


	df_2.Localidade	df_2.Nota	df_2.Distancia	lista
1	Camponi	17.50	7.0	Menor.Dist
2	Antonio	12.50	13.0	Menor.Dist
3	Camponi	3.00	6.0	Medio.Dist
4	Csaude	14.50	3.0	Menor.Dist
5	Mercadinho	12.50	0.5	Menor.Dist
6	Camponi	11.50	7.0	Medio.Dist
7	Esmetuchira	17.50	0.5	Menor.Dist
8	Antonio	9.50	14.3	Medio.Dist
9	Guengene	15.50	12.0	Menor.Dist
10	Camponi	11.00	6.0	Medio.Dist
11	Mafunga	9.00	17.0	Medio.Dist
12	Mutisol	5.00	2.0	Medio.Dist

ving 1 to 12 of 578 entries, 4 total columns

Fig 1 Data

For the analyses, a prior classification of groups was considered **list** *(figure 1)* using the cluster method by computational methods already designed for this purpose *(figure-2) (kmeans, gap−stat,..),* taking into account the distances covered, through which three heterogeneous groups were estimated (Etherington [24]).



```
fviz_nbclust(df_3,kmeans,method = "gap_stat")
df_kmean=kmeans(df_3,3)
win.graph()
fviz_cluster(df_kmean,data = df_3, ellipse.type = "t")
```

Fig 2 Grouping by Kmeans

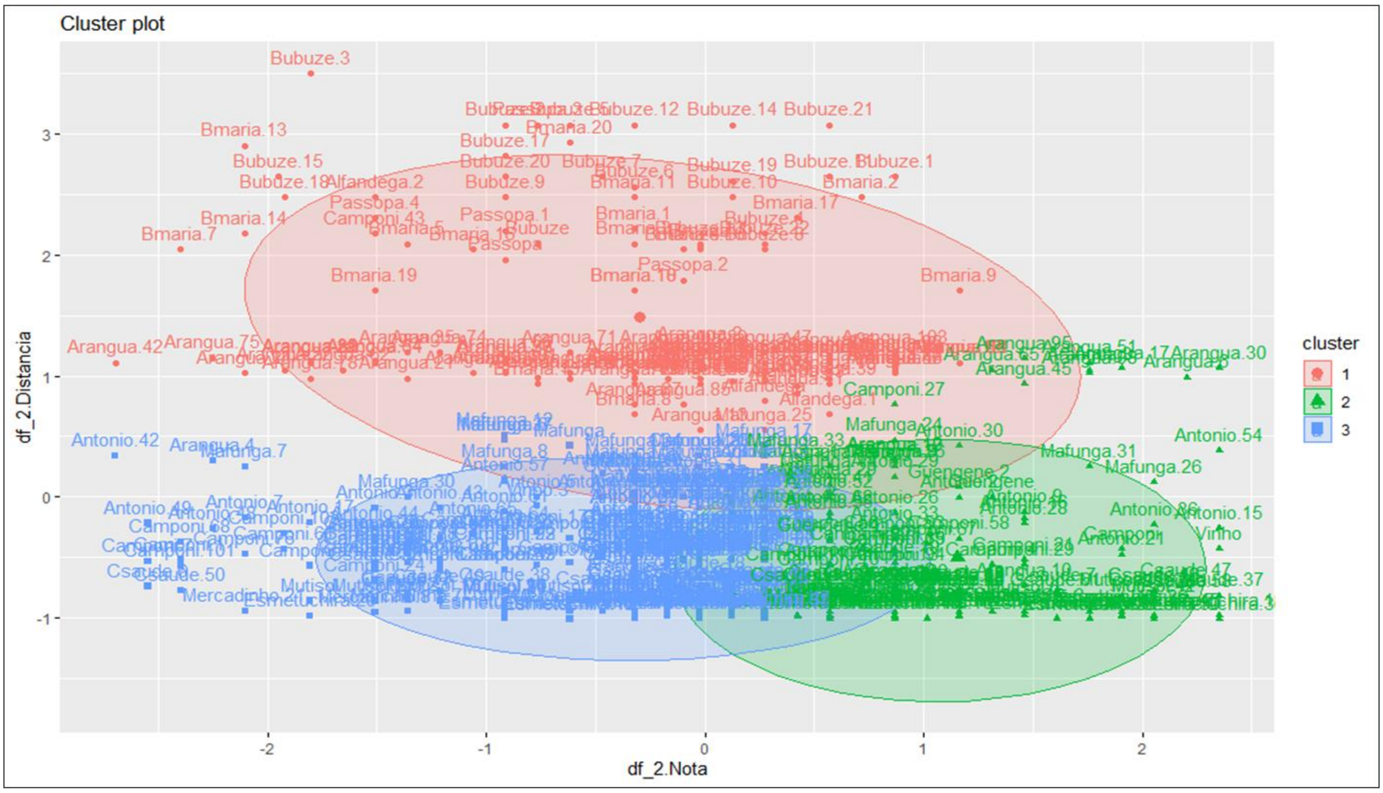
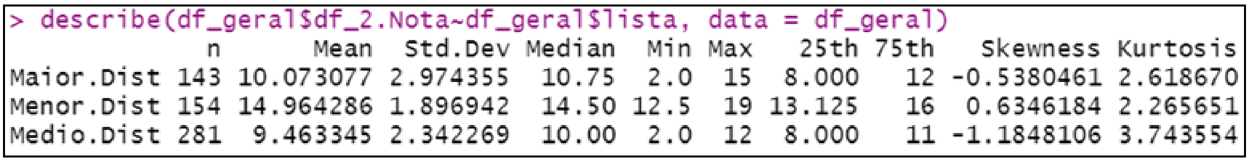


Fig 3 Treatments

To an implementation of multiple comparison methods (ANOVA), to compare the average pedagogical performance, in the three groups (treatments), the data grouping procedure used, it is inevitable that the treatments would have different replicates or not equal sample sizes. Admit useful even if it is carry out experimental statistics in more than two classes at a given level of education, different treatment sample sizes will clearly be observed.

Experiments carried out presenting different treatment sizes usually present anomalies in relation to the underlying assumptions for the analysis of variance procedure, such as equality of variances within treatments and normality. Data transformation methods, for data correction, are difficult to implement when we are dealing with experimental units with not equal sample sizes, therefore a need to use non−parametric methods for this purpose. As can be seen, the sizes of the treatments are different, with one of them having half the observations compared to the other *(figure - 4)* and the estimated groups were named as being:

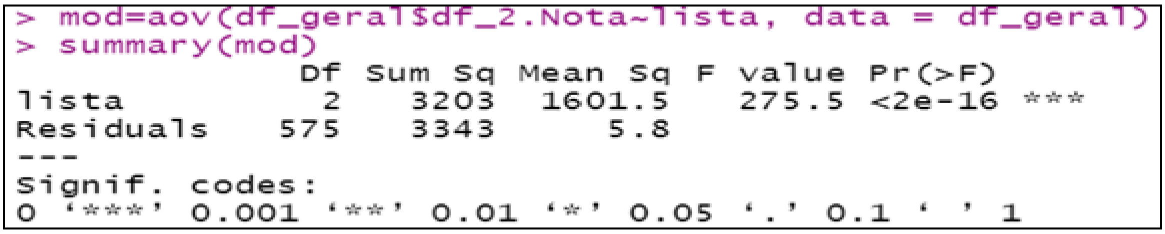
- *Shortest Distance*
- *Medium Distance*
- *Greater Distance*



```
> describe(df_geral$df_2.Nota~df_geral$lista, data = df_geral)
              n      Mean  Std.Dev Median  Min Max   25th 75th   Skewness Kurtosis
Maior.Dist 143 10.073077 2.974355  10.75  2.0  15  8.000   12 -0.5380461 2.618670
Menor.Dist 154 14.964286 1.896942  14.50 12.5  19 13.125   16  0.6346184 2.265651
Medio.Dist 281  9.463345 2.342269  10.00  2.0  12  8.000   11 -1.1848106 3.743554
```

Fig 4 Data Summary

In the ANOVA model *(figure – 5)*, the usual F Statistics test, concludes that there are differences between the treatment means, that is, there are significant differences in the average pedagogical performance between the treatments. Regarding the validation of the model, it is noted that the series of residues *(figure – 6)*, presents some deviation from normality, which may be caused by the observation of atypical data that are in some way representative of the sample, or even by the nature of the data, which have different sizes of treatments.



```
> mod=aov(df_geral$df_2.Nota~lista, data = df_geral)
> summary(mod)
             Df Sum Sq Mean Sq F value Pr(>F)    
lista         2   3203  1601.5   275.5 <2e-16 ***
Residuals   575   3343     5.8                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig 5 F-Statistic Test for ANOVA

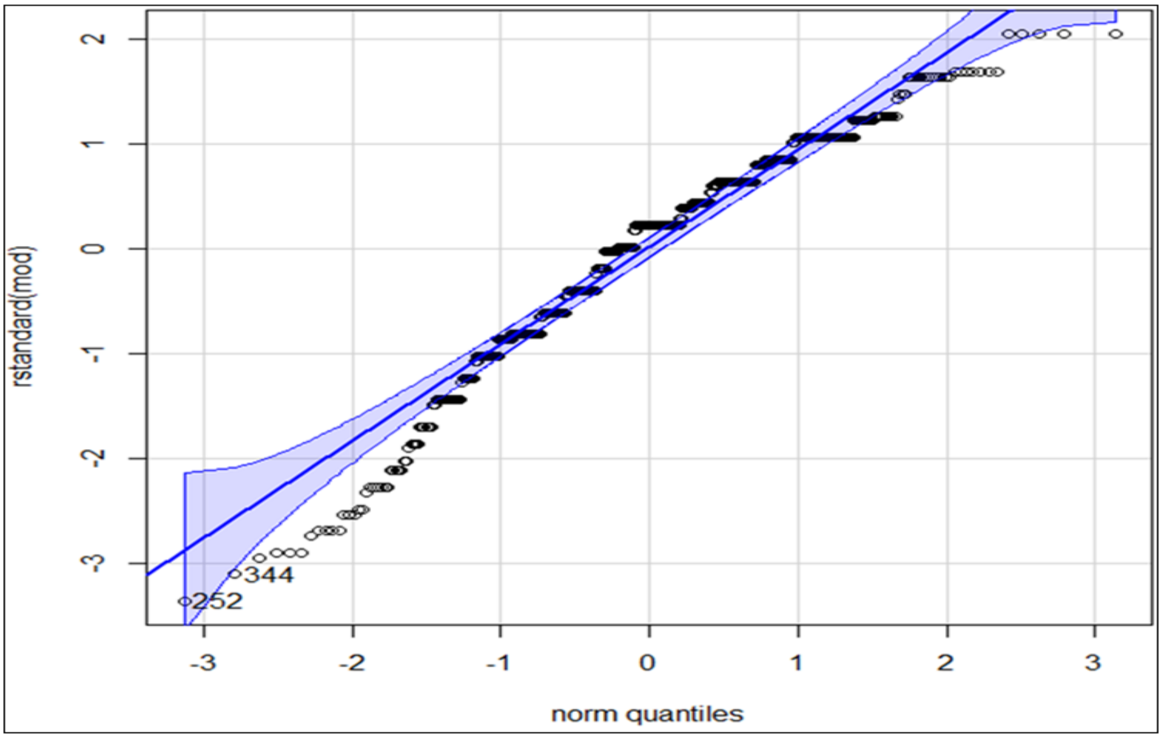


Fig 6 Small Deviations from Normality

Observing the level of asymmetry, graph *(figure – 7)* does not present serious situations of violation of the assumption of normality of the model's residuals, therefore, the idea that the samples have an approximately normal distribution is reinforced, despite the formal test *(figure – 8),* it is rejected normality.

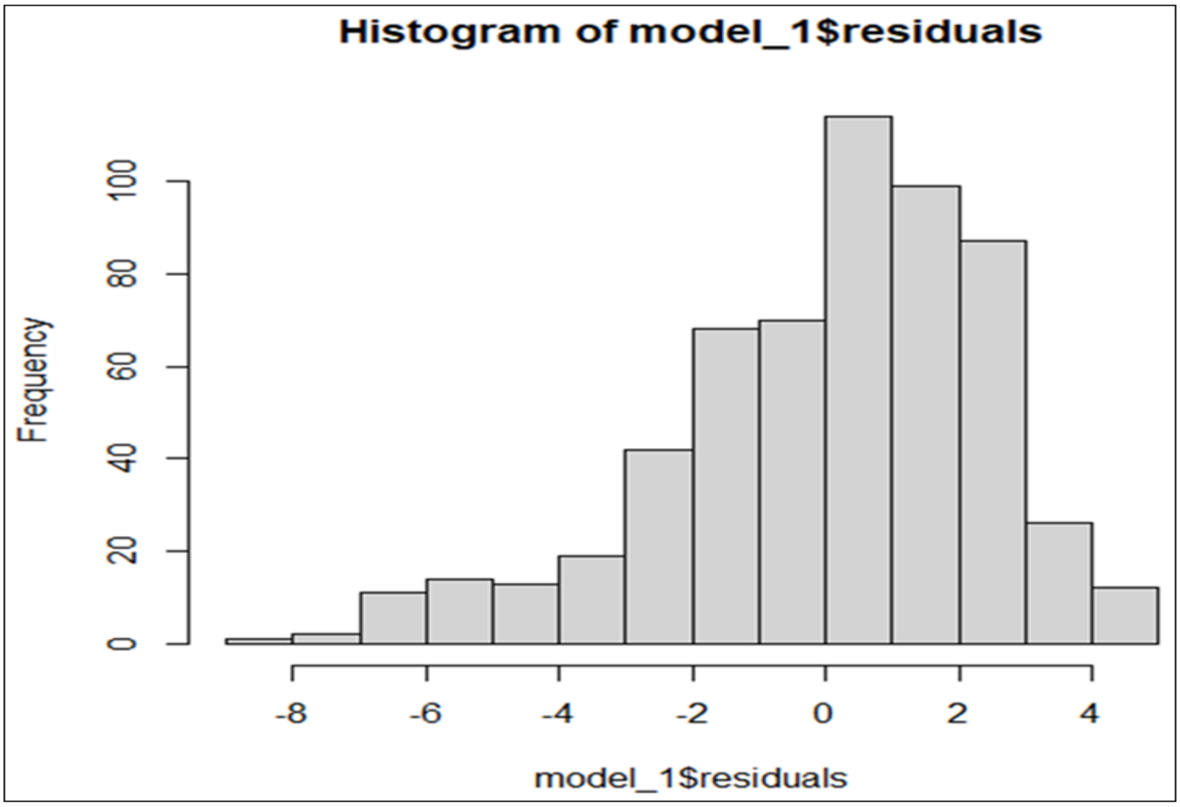
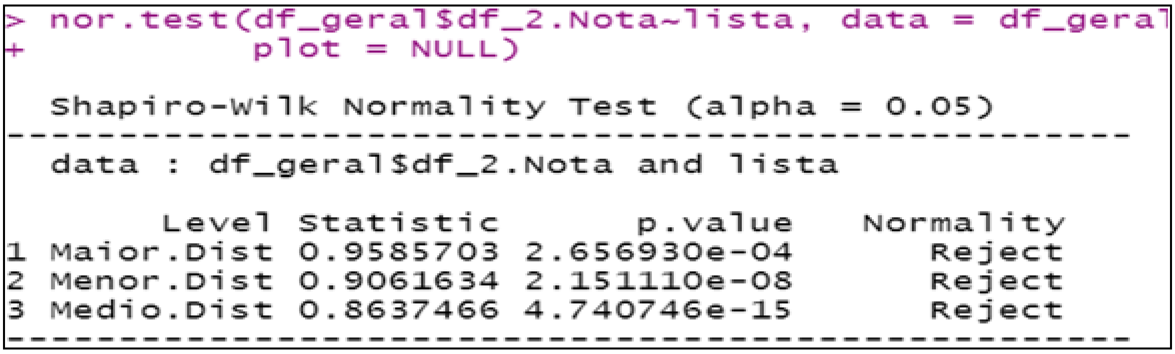


Fig 7 Histogram from Normality Residual



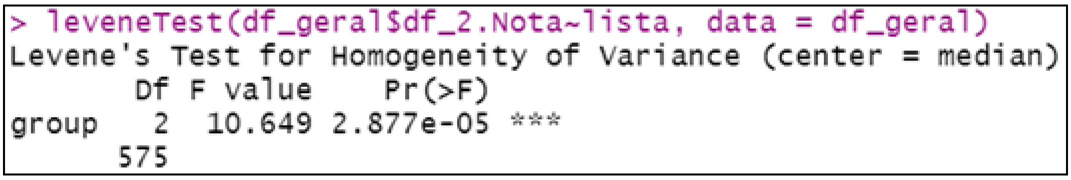
```
> nor.test(df_geral$df_2.Nota~lista, data = df_geral
+          plot = NULL)

  Shapiro-Wilk Normality Test (alpha = 0.05)
--------------------------------------------------
  data : df_geral$df_2.Nota and lista

       Level Statistic       p.value   Normality
1 Maior.Dist 0.9585703 2.656930e-04      Reject
2 Menor.Dist 0.9061634 2.151110e-08      Reject
3 Medio.Dist 0.8637466 4.740746e-15      Reject
--------------------------------------------------
```

Fig 8 Shapiro Wilk Test from Normality

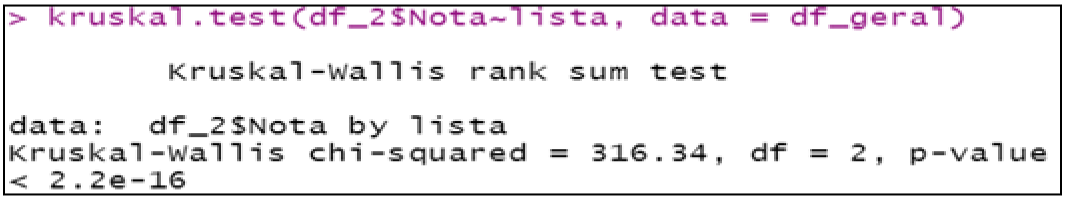
To test the equality of variations within groups and admitting doubts about the assumption of normality, both the Bartlet test and the Levene test can be used, therefore, the Levene was implemented as one of the most efficient and robust alternatives to non-normality, which considers the comparison of variability in relation to the median, in your procedure.



```
> leveneTest(df_geral$df_2.Nota~lista, data = df_geral)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value    Pr(>F)
group   2  10.649 2.877e-05 ***
      575
```

Fig 9 Levene Test for Homogeneity

The conventional F-Statistic test, obtained through one-factor analysis of variance to compare the means of independent normal populations *(figure – 5)*, presents invalid results, given the violation of the assumption of equality of variances and presents results similar to Kruskal - Wallis *(figure – 10)* and *Welch (figure – 11)* non-parametric test, both reject the hypothesis of equality of means in the three groups (p – value < 5%).

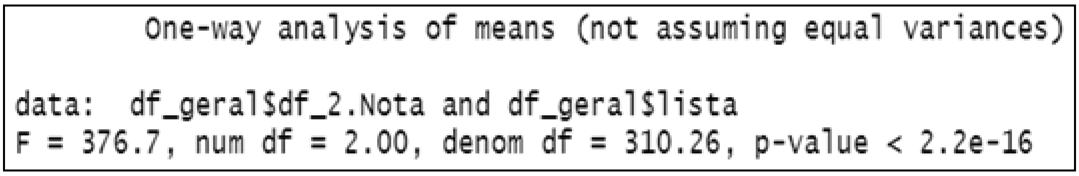


```
> kruskal.test(df_2$Nota~lista, data = df_geral)

	Kruskal-Wallis rank sum test

data:  df_2$Nota by lista
Kruskal-Wallis chi-squared = 316.34, df = 2, p-value
< 2.2e-16
```

Fig 10 Kruskal and Wallis Test



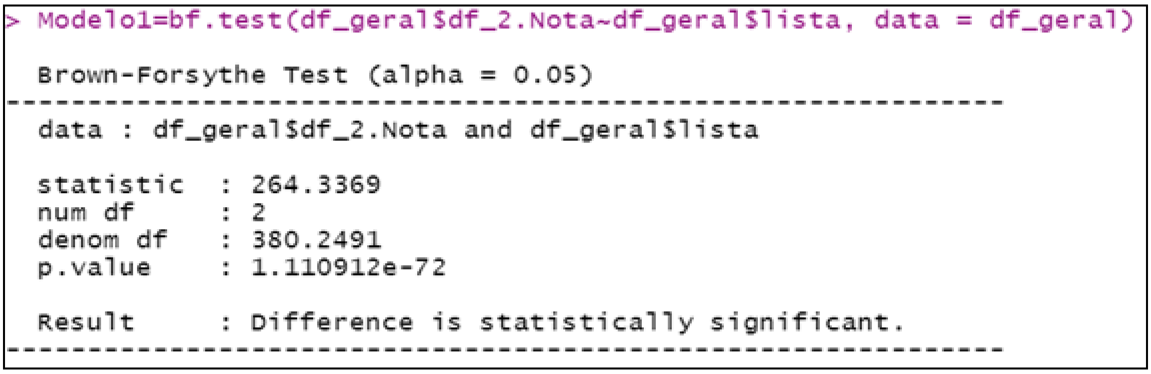
```
	One-way analysis of means (not assuming equal variances)

data:  df_geral$df_2.Nota and df_geral$lista
F = 376.7, num df = 2.00, denom df = 310.26, p-value < 2.2e-16
```

Fig 11 Welch´s Test

The Welch t-teste *(Figure – 11), F*-Statistic test *(Figure – 5),* and the Kruskal and Wallis test *(Figure – 10),* in addition to being unanimous in rejecting the hypothesis of equality of means, present similar results. This result shows that there is no serious violation of the relevant assumptions for the purpose, such as normality and equality of variances. , which can be influenced by the observation of atypical data considered representative of the sample. In general, if the assumptions of equality of variances have not been verified and assuming the approximation to normality, the researcher may choose to use non-parametric tests as a reasonable alternative for the intended inferences.

Although the tests already used tend to reject equality of means in the three groups, and although it rejects the null hypothesis, the Brown Forsythe test *(figure – 12)* may be a reasonable preference due to its robustness, taking into account counts its procedure, which is compared with Levene's test.



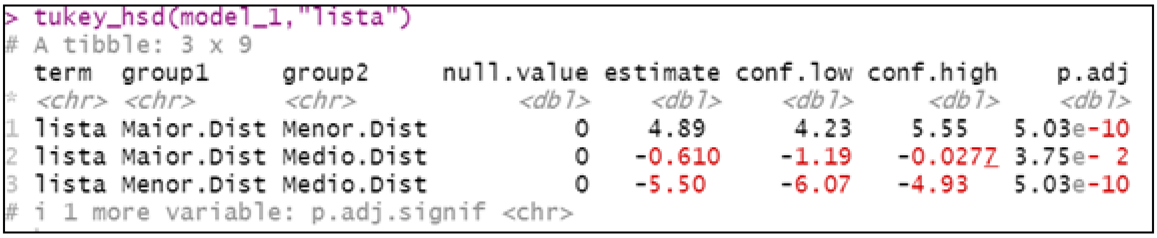
```
> Modelo1=bf.test(df_geral$df_2.Nota~df_geral$lista, data = df_geral)

  Brown-Forsythe Test (alpha = 0.05)
-------------------------------------------------------------
  data : df_geral$df_2.Nota and df_geral$lista

  statistic  : 264.3369
  num df     : 2
  denom df   : 380.2491
  p.value    : 1.110912e-72

  Result     : Difference is statistically significant.
-------------------------------------------------------------
```

Fig 12 Brown Forsythe Test



```
> tukey_hsd(model_1,"lista")
# A tibble: 3 x 9
  term  group1     group2     null.value estimate conf.low conf.high    p.adj
* <chr> <chr>      <chr>           <dbl>    <dbl>    <dbl>     <dbl>    <dbl>
1 lista Maior.Dist Menor.Dist          0    4.89      4.23    5.55   5.03e-10
2 lista Maior.Dist Medio.Dist          0   -0.610    -1.19   -0.0277 3.75e- 2
3 lista Menor.Dist Medio.Dist          0   -5.50     -6.07   -4.93   5.03e-10
# i 1 more variable: p.adj.signif <chr>
```

Fig 13 Games-Howell Test

As has already been commented on the use of post-hoc tests, admitting heterogeneity between groups, the Games-Howell test *(figure – 13),* shows that all pairs of groups are statistically different (p – value < 0.05), and therefore, the possible solutions in relation to public services, especially in the area of education, could be implemented considering the specific problems of each of the three groups

Tools Used: *RStudio, Python and Latex*

REFERENCES

[1]. C. M. Douglas, *"Design and Analysis of Experiments"*. Jhon Wiley and Suns: 8th Edition, 2013.

[2]. R. G. O'Brien, *"Robust techniques for testing heterogeneity of variance effects in factorial designs,"* 1978.

[3]. L. M. Robert, F. G. Richard, and L. H. James, *"Statistical Design and Analysis of Experiments with Applications to Engineering and Science"*. Jhon Wiley and Suns, 2th edition - 2003.

[4]. E. P. B. George, J. H. Stuart, and W. G. Hunter, *"Statistical of Experimenters"*. Jhon Wiley and Suns, 2th edition - 2005.

[5]. G. Anindya, S. Bapi, and P. Mal, *"Textile Engineering Statistical Techniques, Design of Experiments and Stochastic Modeling"*. Tylor and Francis Group, 2022.

[6]. G. Ruxton and G. Beauchamp, *"Time for some a priori thinking about post hoc testing,"* 2008.

[7]. P. R. Rosenbaum and D. Rubin, *"Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score,"* The American Statistician, 1985

[8]. G. C. J. Fernandez, *"Residual Analysis and Data Transformations as Important Tools in Statistical Analysis"*. Department of Agriculture Economics, University of Nevada-Reno, 1992.

[9]. D. G. Jean and C. Subhabrata, *"Nonparametric Statistical Inference"*. Chapman Hall/CRC, Fifth Edition - 2011.

[10]. R. Johnson, I. Miller, and J. E. Freund, *"Probability and Statistics for Engineers"*. Pearson London, 9th edition - 2017.

[11]. C. O. H. N. Edison, *"Bioestatística quantitativa aplicada"*. UFRGS-Porto Alegre, 2000.

[12]. T. E. Ranghuthan, J. Lepko Wski, H. Van, and P. Solenberger, *"A multiply imputing missing values using a sequence of regression models,"* Statistics Canada Catalogue, vol. 27, 2001.

[13]. R. E. Walpole, R. H. Myers, S. L. Myers, and K. Ye, *"Probabilidad Estadıstica para Ingenieria Ciencias"*. Naucalpan de Juarez, 9ad, 2012.

[14]. E. Chaves-Barboza, J. M. Trujillo-Torres, and J. A. Lopez-Nunez, *"Accomplishments in ˜ learning self-regulation in personal environments,"* june, 2015.

[15]. W. J. Conover, *"Practical Nonparametric Statistics*". Jhon Wiley and Suns, 3rd Edition - 1999.

[16]. F. H. Imai, N. Tsumura, and Y. Miyake, *"Perceptual color difference metric for complex images based on mahalanobis distance,"* Journal of Electronic Imaging - 2001.

[17]. F. E. Satterthwaite, *"An approximate distribution of estimates of variance components,"* International Biometric Society, vol. 2, pp. 110– 114, 1946.

[18]. B. L. Welch, "*The generalization of students problem when several different population variances are involved,"* 2019.

[19]. A. D. Alameida, N. E. Silva, and S. N. Juvêncio, "*Modifications and alternatives to the tests of levene and brown forsythe for equality of variances and means,"* 2008.

[20]. M. B. Brown and A. B. Forsythe, *"Robust tests for the equality of variances,"* 1974.

[21]. L. Breiman, *"Classification and Regression Trees"*. eBook Published, New York - 1984.

[22]. J. M. Keppel-Benson, *"Design and analysis,"* A Researcher's Handbook. Pearson., 1993

[23]. S. A. Rusticus and C. Y. Lovato, *"Impact of sample size and variability on the power and type I error rates of equivalence tests,"* A Simulation Study. Practical Assessment, Research Evaluation, vol. 19, pp. 1–10, 2014.

[24]. T. R. Etherington, *"Mahalanobis distances for ecological niche modelling and outlier detection: implications of sample size, error, and bias for selecting and parameterizing a multivariate location and scatter method",* Geography in Higher Education, 40(1), 2021.

[25]. W. H. Kruskal and W. A. Wallis, *"Use of ranks in one-criterion variance analysis,"* 1952.