Alternatives for correcting the Tukey's statistics for unbalanced experiments

José Márcio Martins Júnior[1†], *Eric B. Ferreira*[2], *Patrícia S. Ramos*[2]

[1] *Master student in Applied Statistics and Biometrics, Federal University of Alfenas, Brazil.*
[2] *Exact Sciences Institute, Federal University of Alfenas, Brazil, CEP: 37130-000.*

Resumo: *Experimentos desbalanceados são aqueles em que os tratamentos não têm o mesmo número de repetições. Alguns testes de comparações múltiplas de médias precisam ser corrigidos para se comportarem bem em situações desbalanceadas. Para o teste de Tukey é aconselhado o uso da média harmônica dos números de repetições. O objetivo deste artigo é verificar o desempenho das médias harmônica, aritmética, ponderada, quadrática, e geométrica, o número mínimo, máximo e a mediana do número de repetições como correção para o teste de Tukey, bem como propor uma correção que melhore o desempenho da média harmônica. Foram medidas as taxas de erro tipo I e o poder dos testes, via simulação Monte Carlo, a 5% de significância. A média harmônica apresenta o melhor desempenho, mas pode ser melhorada. A proposta feita neste trabalho, conseguiu controlar o erro tipo I em todos os casos estudados.*

Palavras-chave: Tukey; testes de comparação múltipla; desbalanceamento; média harmônica; experimento.

Abstract: *Unbalanced experiments are those in which treatments do not have the same number of replications. Some multiple comparison tests need to be correct to behave well in unbalanced situations. For the Tukey test is suggested the use of the harmonic mean of the numbers of replications. The aim of this paper is to verify the performance of the harmonic, arithmetic, weighted, quadratic, and geometric means, the minimum, maximum and median number of repetitions as a correction for the Tukey test, and propose a correction that improves the performance of harmonic mean. We measured the type I error rate and power via Monte Carlo simulation, at 5% of significance. The harmonic mean performs best but could be improved. The proposal made in this work was capable of controlling the type I error in all cases studied.*

Keywords: Tukey test; multiple comparisons tests; unbalanced experiment; harmonic mean; experimentation.

Introduction

One way to produce scientific knowledge is through planning and analysis of experiments. An experiment is a controlled production of a phenomenon to be analyzed, consisting of factors, treatments and response variables (MACHADO et al., 2005). It is usually done to elect the best treatment(s) and this election is made by comparing their means.

There are three types of errors that can be made when performing inferences (RAMALHO et al., 2005). The type I error, when one rejects the true hypothesis (its probability is called α). The type II error, when one accepts a false hypothesis (its probability is called β). And the type III error, which occurs when one (some) treatment(s) are classified in the opposite way.

On the other hand, the power of a test is the probability of rejecting a false null hypothesis. It is defined as $Power = 1 - \beta$ (MORETTIN, 2000).

Most multiple comparison tests usually assume that the experiment is balanced, such as the Tukey's test does. When this characteristic is lost Tukey's statistic (Tukey, 1953) can corrected taking the harmonic mean of the number of replications, as suggested by Kramer (1956).

According to Ramalho et al. (2005), the Tukey test requires all treatment levels to have the same number of repetition and the inferences should be made for all possible pair of means.

[†]Corresponding author: jmmjunifal@gmail.com

This method uses the studentized (or standardized) range distribution (q):

$$q_{(I,\nu,\alpha)} = \frac{Max(Y_i) - Min(Y_i)}{s}$$

Let n observations Y_1, Y_2, ... ,Y_n come from a normal distribution, with mean μ and variance σ^2. Considering s^2 as the estimator for σ^2, the Minimum Significant Difference (MSD) of the test can then be defined for one level α of significance with the aim to test the hypotheses $H_0 : \mu_i - \mu_{i'} = 0$, for the expression:

$$\Delta = q_{(I,\nu,\alpha)}\sqrt{\frac{1}{2}\hat{\sigma}_{\bar{x}_1-\bar{x}_2}} = q_{(I,\nu,\alpha)}\sqrt{\frac{MSR}{J}}, \tag{1}$$

where I is the number of treatments, ν is the freedom degree of the residue, J is the number of repetition, $\hat{\sigma}_{\bar{x}_1-\bar{x}_2}$ is the standard-error from the difference between the means and MSR is the mean square of the residual.

Then, if $x_i - x_{i'}$ is greater than Δ, the $H_0 : \mu_i - \mu_{i'} = 0$ must be rejected at the nominal level α previously established.

According to Machado et al. (2005), the Tukey test controls the experimentwise type I error, but it becomes very conservative in relation to the comparisonwise error rate when the number of treatments increases.

Borges et al. (2003) state that the Tukey test for balanced experiments controls the comparisonwise and experimentwise type I error rate under normal distribution.

According to Dunnett (1980), for balanced experiments, the Tukey (1953) establishes the following confidence interval set for the quantities $\mu_i - \mu_{i'}$:

$$\bar{x}_i - \bar{x}_{i'} \pm q_{(I,\nu,\alpha)}\sqrt{\frac{MSR}{J}} \tag{2}$$

where $\bar{y}_i$ denotes the sample mean of the ith treatment; $q_{(I,\nu,\alpha)}$ is the upper quantile α of the studentized range distribution of I normal variables, and MSR is the Mean Squared of Residual from ANOVA.

In experimentation unbalanced experiments are usual, either by loss of parcels or experiments that are already planned to be unbalanced (such as unbalanced incomplete block design).

According to Dunnett (1980), Kramer in 1956 proposed a correction for multiple comparison statistics (not only for Tukey's) for unbalanced experiments. Kramer (1956) noted that the expression to the right of the sign of ± in equation (2) is equivalent to dividing the quantile for $\sqrt{2}$ and multiply by the standard error of the difference between two means. Thus, the expression is rewritten as:

$$\bar{y}_i - \bar{y}_{i'} \pm \frac{q_{(I,\nu,\alpha)}}{\sqrt{2}}\sqrt{\frac{S^2}{J_i} + \frac{S^2}{J_{i'}}} \tag{3}$$

It is easy to see that expression (3) can be rewritten as a function of the harmonic mean of the number of repetitions:

$$\bar{y}_i - \bar{y}_j \pm q_{(I,\nu,\alpha)}\sqrt{\frac{\frac{S^2}{2}}{\left(\frac{1}{J_i} + \frac{1}{J_{i'}}\right)}} \tag{4}$$

Several authors have criticize this statement, e.g., Miller (1966) cited by Dunnett (1980), which states that this modification is inaccurate and has no mathematical proof. Instead of the harmonic mean, Miller (1966) suggests using the arithmetic mean or median of the numbers of repetitions, but warning that this is for “fearless statisticians”.

On the other hand, many authors advise and use the harmonic mean as an adaptation to the number of repetitions, but is not common in the literature to find the reason to use this mean (RAMALHO et al., 2005; SAMPAIO, 2010; PIMENTEL-GOMES, 2009). Even when the reason found we are not sure it is the best best alternative.

Methodology

For this research, routines were written in R language (R CORE TEAM, 2013). These routines have the function to raffle the data, perform multiple comparisons tests and Monte Carlo simulations.

Data were simulated from a completely randomized design following the statistical model:

$$y_{ij} = \mu + \tau_i + e_{ij} \tag{5}$$

where y_{ij} is the jth replication of the ith treatment, for $i = 1, \ldots, I$, $j = 1, \ldots, J_i$, $I \in$ (2, 3, 5, 10, 15, 20, 30) and $J_i \in$ (3, 6, 9, 12, 15, 18, 21, 24); μ is a common constant (overall mean), set to zero without loss of generality; τ_i is the fixed effect of the ith treatment; and $e_{ij} \sim N(0, 1)$ is the random error associated to y_{ij}.

The unbalance level (unbalance rate) has been defined as the ratio of the maximum and the minimum number of replications:

$$\delta = \frac{\max(J_i)}{\min(J_i)}, \tag{6}$$

where J_i is the number of replications of one treatment i. Therefore $\delta > 1$.

To ensure the randomness of the experiment and to get all possible values of δ, the number of repetitions were chosen as follows:

1. A matrix was created to contain the possible numbers of replications. The unbalance rate δ was computed by dividing each pair of values (eq. 6) of the superior triangular matrix (Table 1). (As long as the balanced case does not matter, was not necessary to evaluate the main diagonal).
2. A vector of size I (treatments) was created and filled with values drawn with replacement from the set $J = [3, 6, 9, 12, 15, 18, 21, 24]$ such that a maximum fixed δ occurs. It is done by fixing the first and the last element in the sorted vector.

Table 1: Unbalance rate according to the number of replications

Replications	3	6	9	12	15	18	21	24
3	-	2.00	3.00	4.00	5.00	6.00	7.00	8.00
6	-	-	1.50	2.00	2.50	3.00	3.50	4.00
9	-	-	-	1.33	1.67	2.00	2.33	2.67
12	-	-	-	-	1.25	1.50	1.75	2.00
15	-	-	-	-	-	1.20	1.40	1.60
18	-	-	-	-	-	-	1.17	1.33
21	-	-	-	-	-	-	-	1.14
24	-	-	-	-	-	-	-	-

It is important to note that each δ occurred a different number of times since we can have more than on pair of numbers such ration gives the same. For instance, $\delta = 8$ only occurs when the minimum is 3 and the maximum is 24. On the other hand, $\delta = 2$ is the result of four ratios (Table 1)

Then, in order to present this result in 2D graphics it was computed the average of type I error rate and power for these cases with multiple estimates per δ. However, every estimate can be seen in the 3D graphics, since it displays the maximum and the minimum number of replicates (rather than the unbalance level).

Tow groups experiments were performed. Group I, with the purpose of estimating the error type I and group II, to estimate the power.

Experiments in group I were simulated under H_0. We set the effect of treatment equal to zero, i.e., $\tau_1 = \tau_2 = \ldots = \tau_I = 0$, without loss of generality.

On the other hand, in group II consecutive treatment effects were set to distance 0.5 standard error of the mean from each other.

Each group had 196,000 experiments - product of 7 possible numbers of treatments $I \in$ (2, 3, 5,10, 15, 20, 30), 28 unbalance levels (δ) that that come from the numbers of replications $J \in$ (3, 6, 9, 12, 15, 18, 21,24) and 1000 Monte Carlo runs.

It was computed only the experimentwise type I error rate, given by

$$etI = \frac{\text{Number of experiments with at least one error}}{\text{Number of experiments}} \tag{7}$$

To infer about the difference between the type I error rate and the nominal level of significance ($\alpha = 5\%$), we used the exact confidence interval for proportions with 99% of probability. Given by:

$$IC_{1-\alpha}: \left[L_I = \frac{1}{1 + \frac{(n-y+1)\, F_{\alpha/2;\nu_1=2(n-y+1),\nu_2=2y}}{y}} ; L_S = \frac{1}{1 + \frac{(n-y)}{(y+1)\, F_{\alpha/2;\nu_1=2(y+1),\nu_2=2(n-y)}}} \right] \tag{8}$$

where $F_{\alpha/2}$ is the upper quantile of an F distribution, with ν_1 and ν_2 degrees of freedom. If $y = 0$, then $L_I = 0$ and L_S is given in (8), or if $y = n$, then $L_S = 1$ and L_I is given in 8 (FERREIRA, 2005).

The 99% confidence interval for the type I error rate was $[3.3927\%, 7.0504\%]$, i.e., every observed value outside this interval can be considered different from the nominal level of significance.

It is known that when a test is liberal it tends to be more powerful (and vice versa). In order to avoid this illusion, here we compute not only the power but the *real power* of the tests.

The *real power* is a metric that allows the estimation of the power that a test would have if it controlled the type I error rate. We could not find references about real power in the literature, but it is quite intuitive. We define the real power (RP) as:

$$RP = \begin{cases} P - (etI - \alpha), & \text{if } etI > \alpha; \\ P, & \text{otherwise} \end{cases} \tag{9}$$

where P is the empirical power, etI is the type I error rate and α is the nominal level of significance.

In each experiment, 9 corrections (arithmetic mean, median, geometric mean, quadratic mean, weighted mean, the maximum and the minimum and the proposed one) for the Tukey test were evaluated plus the Tukey-Kramer (harmonic mean). In Table 2 we show most of them.

Table 2: Main functions evaluated to correct the Tukey's statistic under unbalanced condition (but the proposed one)

Arithmetic mean	Geometric Mean	Quadratic mean	Weighted mean
$J_{AM} = \frac{\sum_{i=1}^{I} J_i}{I}$	$J_{GM} = \sqrt[I]{\prod_{i=1}^{I} J_i}$	$J_{QM} = \sqrt{\frac{\sum_{i=1}^{I} J_i^2}{I}}$	$J_W = \frac{\sum_{i=1}^{I} (w_i J_i)}{\sum_{i=1}^{I} w_i}$
Harmonic mean	**Median**	**Minimum**	**Maximum**
$J_{HM} = \frac{I}{\sum_{i=1}^{I} \frac{1}{J_i}}$	$J_{Md} = md(J_i)$	$J_{Min} = \min(J_i)$	$J_{Max} = \max(J_i)$

where J_i is the number of replications of the ith treatment, I is the number of treatments and w_i is the weight for the ith treatment.

The Proposed function (J_P) is a mixture of the harmonic mean (HM) and minimum function (Min), i.e., a weighted mean, based on the number of treatments and on the unbalance level δ. Those two functions were chosen because they are limited to the area you want to achieve with a new test. HM use to be liberal and Min use to be conservative (as will be discussed later), so a mixture of them should be appropriate. Thus, J_P is defined as:

$$J_P = (1 - \hat{p})HM + \hat{p}Min$$

where $\hat{p} = \frac{a}{a+b}$; a is the distance between HM's type I error rate and the nominal level of significance (α); b is the distance between Min's type I error rate and α, as can be seen in Figure 1.

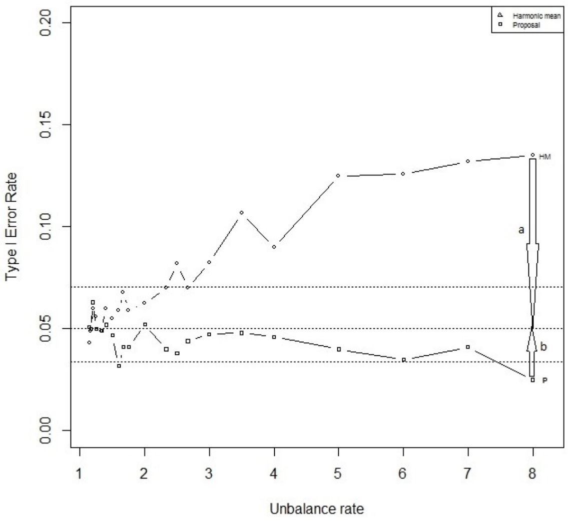


Figure 1: Illustration of a and b required for the computation of $\hat{p}$.

This is done for all possible unbalance levels δ and number of treatments I. If both are conservative or liberal use of the value that is closer to the nominal level of significance.

It is easy to see that under balanced condition, i.e. $J_1 = J_2 = \cdots = J_I = J$, all those functions (corrections) equals to J.

With the aim to estimate the functional relationship between the number of treatments (I), the unbalance level (δ), the maximum and minimum number of replication and the proportion in which the harmonic mean and minimum function should be mixed (p), a function $f()$ of all possible combinations of these quantities was investigated.

$$p = f[I, \sigma, max(J_i), min(J_i)] + \varepsilon$$

where $\varepsilon \sim N(0, \sigma_e^2)$.

The best model was selected via backward method according to the Akaike criterion (DRAPER; SMITH, 1998).

Results and discussion

In this section, we present and discuss the results of Groups I and II, i.e., Type I error rates and power for all situations studied.

First, we present the type I error rates along the unbalance levels, $\delta = 1.14$ to 8.0, for all numbers of treatments studied ($I = 2$ to 30).

In Figure 2a the results in a experiment with two treatments. Using the Min function the test is exact up to the $\delta = 2$ then becomes to be conservative. Furthermore, using the HM functions and P, the test is exact for all values of δ. It may also be noted that the function GM fails to control the type I error from $\delta = 3.5$. The functions Md and AM had an extremely similar behavior. On the other hand, QM, WM and Max fail to control the type I error for $\delta > 1.75$.

In figure 2b we can find the results for an experiment with 30 treatments. One can observe a very similar behavior, but in this case the function Min becomes conservatively at $\delta > 2.0$. Again HM and P are conservative for all values of δ. In this case, the functions GM, Md, AM, QM, WM and Max fail to control the type I error for $\delta > 2.66$. The functions AM and Md had a slightly difference between each other in this case. Md showed a slightly better performance compared to AM for strong unbalance.

The behavior of the functions for treatments between 2 and 30 can be found in Appendix.

In general, the results of the Tukey test using the Min function is the most conservative and using the Max function was the most liberal. That was expected, according to Kelseman and Rogan (1978).

The HM controlled the type I error only for some treatments numbers and values of δ. The HM was able to maintain the nominal level of significance only for experiments with a maximum of 5 treatments.

Over 5 treatments, starts controlling only small imbalances. Therefore, one can not conclude that a test is conservative, as stated by Hayter (1984).

The 3D graphics (Figure 3) also show the type I error rate, but along the maximum and the minimum number of replications. Here the behavior is more detailed since for 2D graphs we used only the mean for each time. So, it can be noticed the differences for a same δ. For those graphics we are going to describe only the behavior of the 3 more competitive functions, namely Min (bold square), HM (empty triangle) and P (empty square).

For the HM function the type I error rate tends to increase as the maximum increases in a given maximum unbalance level, e.g., let an experiment with minimum 3 and maximum 6 replications of treatments and another one with 12 and 24, then they have the same unbalance level but the second experiment tends to present a greater type I error rate. The opposite occurs for the Min function. The P function is always exact or conservative.

The Table 3 presents the range of values of δ in which the functions are conservative, liberal or exact for all numbers of treatments. This table aims to summarize the behavior of the three functions with better performance.

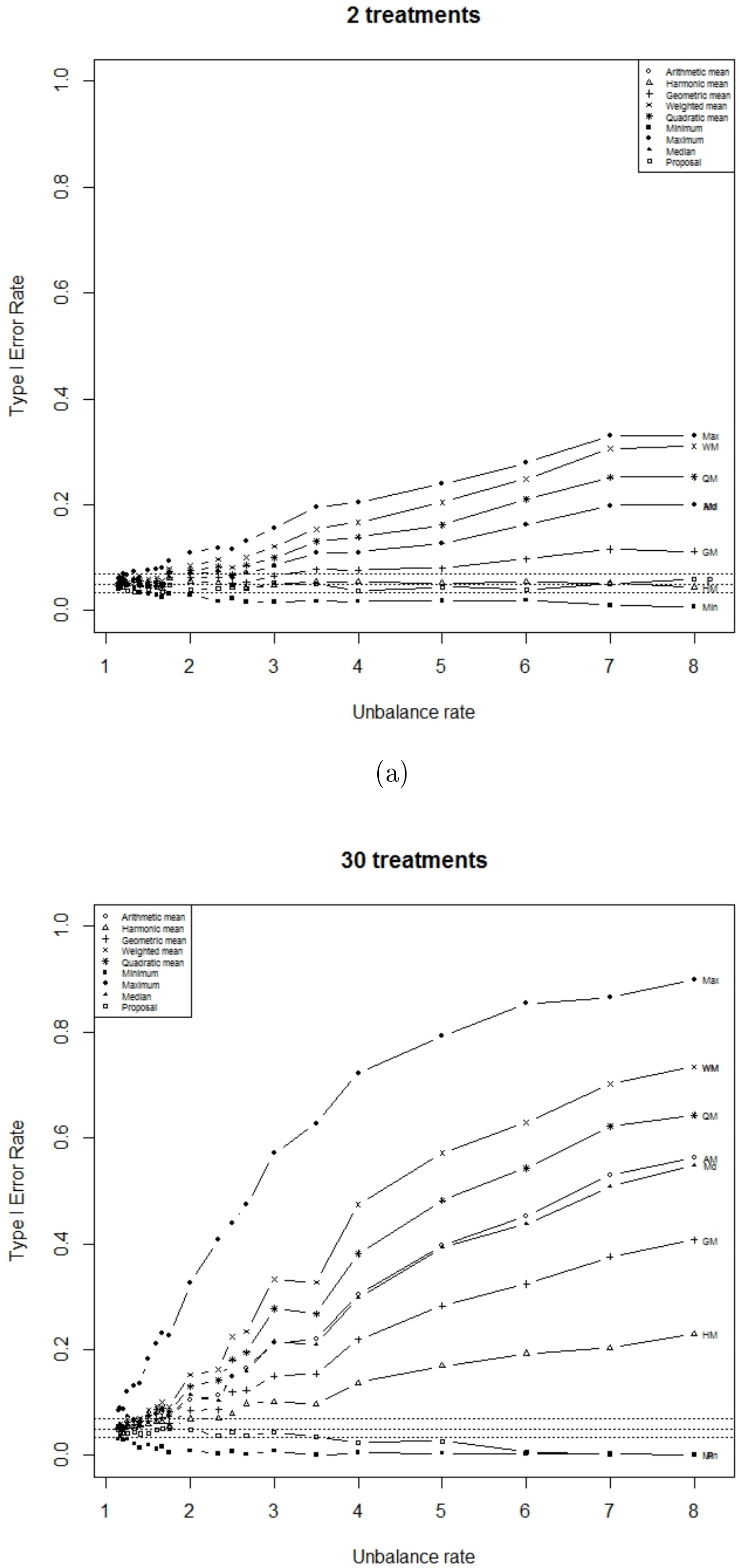
The other functions - not mentioned in Table 3 - fail to control the type I error even for small δ and few treatments. Due to their failure in controlling the type I error their powers are illusory, so we estimate the *real power* as describe previously.

Table 3: Results of the functions with the best performances analyzed

Function	Treatments	Conservator	Accurate	Liberal
HM	2	-	$\forall$	-
	3	-	$\forall$	-
	5	-	$< 8,0$	$8,0$
	10	-	$\leq 3,0$	$> 3,0$
	15	-	$\leq 2,33$	$> 2,33$
	20	-	$\leq 2,0$	$> 2,0$
	30	-	$\leq 2,33$	$> 2,33$
Min	2	$> 2,0$	$\leq 2,0$	-
	3	$> 1,5$	$\leq 1,5$	-
	5	$> 1,2$	$\leq 1,2$	-
	10	$> 1,2$	$\leq 1,2$	-
	15	$\forall$	-	-
	20	$\forall$	-	-
	30	$\forall$	-	-
P	2	-	$\forall$	-
	3	-	$\forall$	-
	5	-	$\forall$	-
	10	-	$\forall$	-
	15	$8,0$	$< 8,0$	-
	20	$> 6,0$	$\leq 6,0$	-
	30	$> 3,5$	$\leq 3,5$	-

Now the results for the Group II experiments are presented, which aimed to estimate the power of the tests. First will be presented the power in figures 4a and 4b, standing for 2 and 30 treatments, respectively. In both cases we evaluated all the possible values of δ (1.14 to 8.0), and the other numbers of treatments can be found in Appendix.

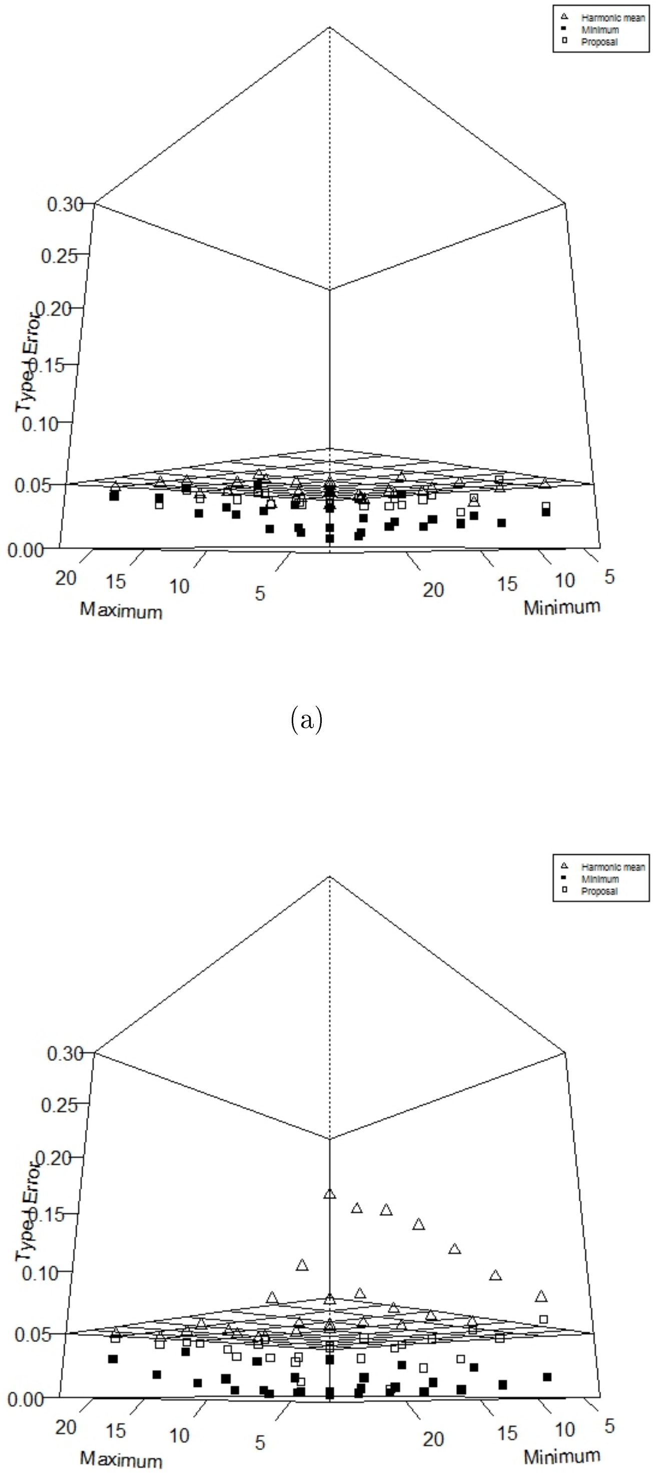
Figure 5 brings the behavior of the real power along the unbalance levels for 2 and 30 treatments. One can find the behavior of the real power for other numbers of treatments in the Appendix.



(a)

(b)

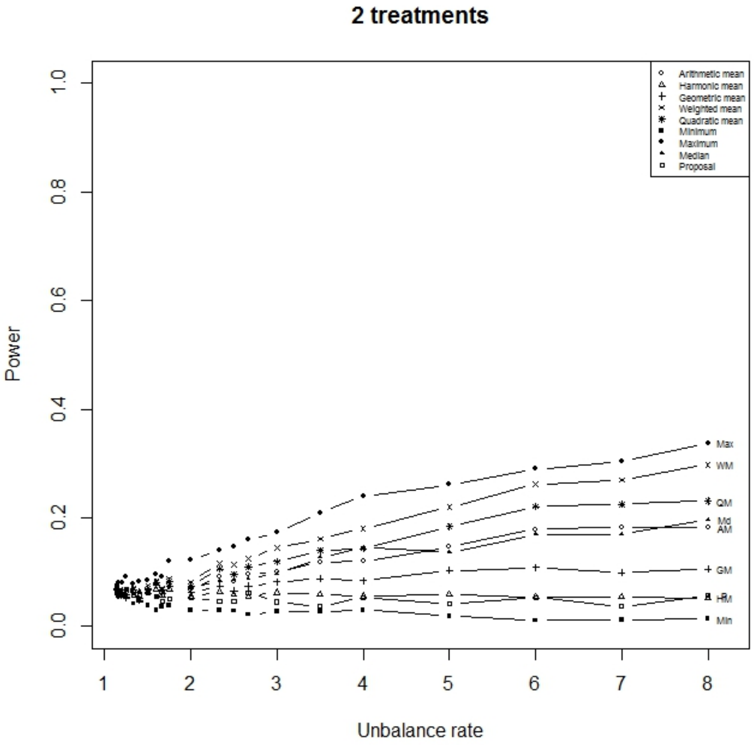
Figure 2: Type I error rate along unbalance levels (δ) for 2 (a) and 30 treatments (b).



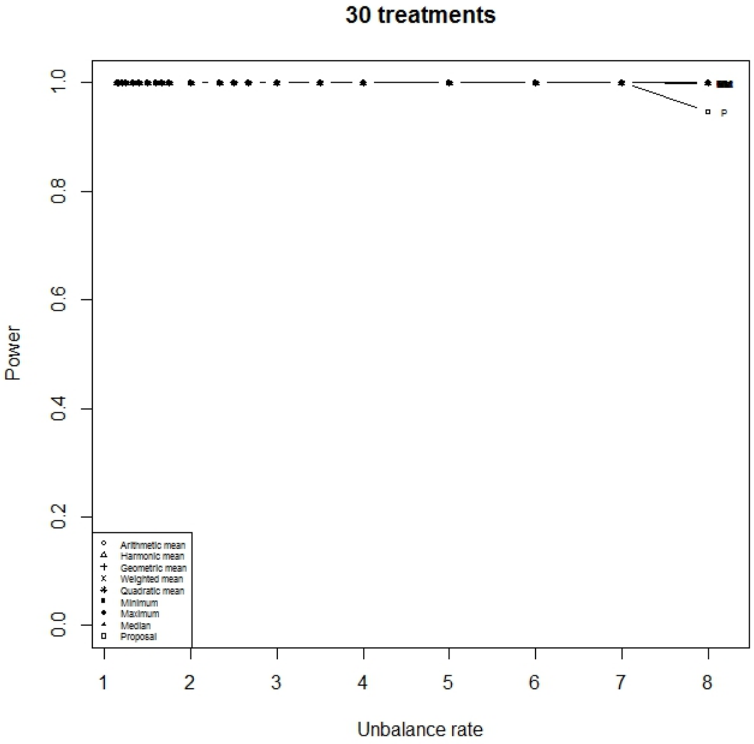
(a)

(b)

Figure 3: Type I error rate (3D) along minimum and maximum values for J_i for 2 (a) and 30 treatments (b).

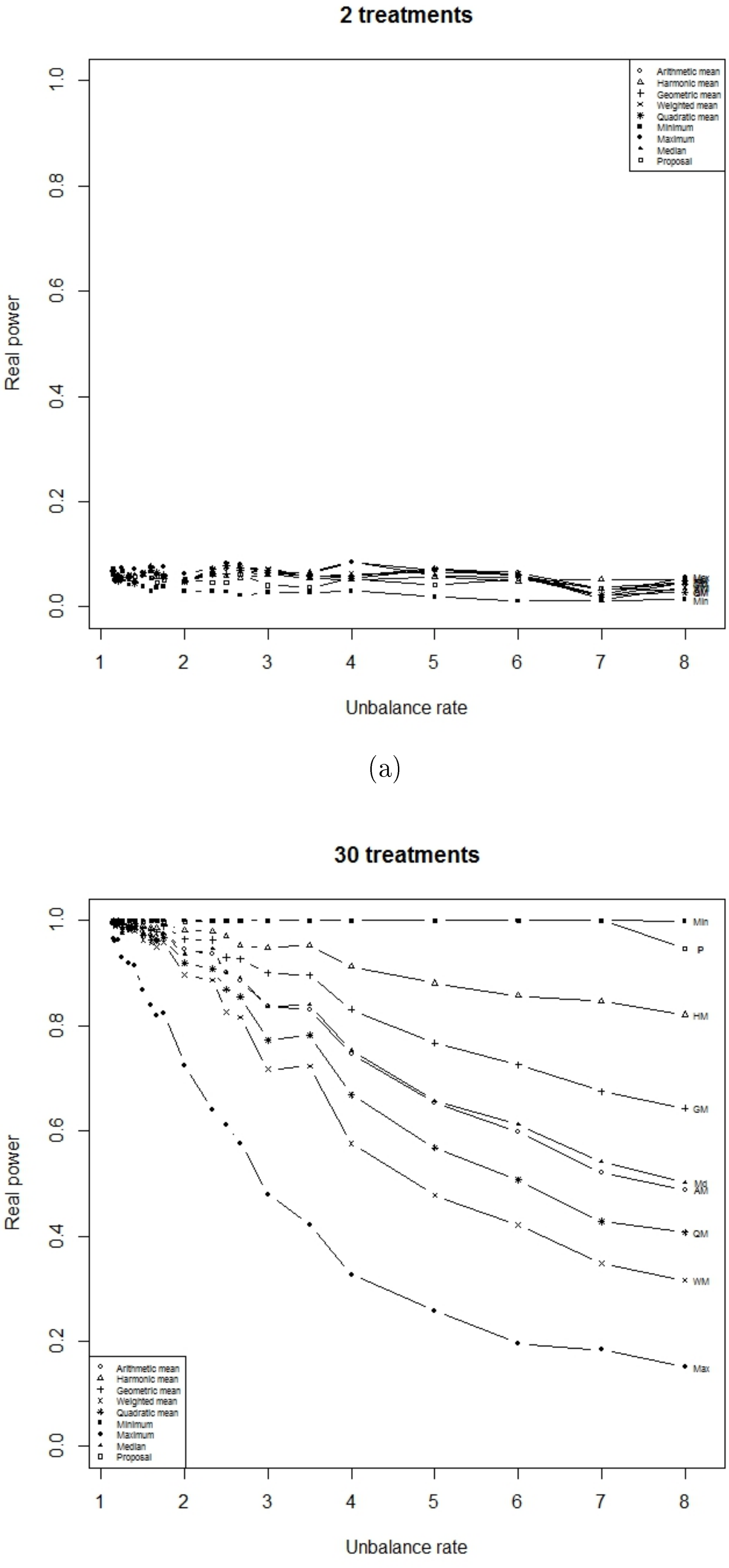


(a)



(b)

Figure 4: Power along unbalance levels (δ) for 2 (a) and 30 treatments (b).



(a)

(b)

Figure 5: Real power along unbalance levels (δ) for 2 (a) and 30 treatments (b).

For Group II of experiments (Power and Real Power) is interesting to note that the power decreases as the δ increases. It happens to all statistics that we considered reasonable, i.e., Min, P and HM, respectively. These functions are also the most affected by the unbalance level. It means that the power of these statistics greatly decreases as the unbalance is greater. On the other hand, observing statistics already considered not reasonable as Max, WM and QM, which are significantly affected by the unbalance level, it can be seen that the power slightly decreases or remains when the δ increases. This is a desired feature but illusory in statistics that practice 60% to 85% of type I error.

One possible explanation for the decrease in power with increasing δ is that some treatments' means are estimated less accurately due to the unbalance. That makes the test to accept H_0 more easily, even when the null is not true.

Possibly, the more important result of this work is the proportion to mix the harmonic mean and the minimum value of replications in a unbalanced experiment. With this information any researcher can perform a better Tukey test for his/her unbalanced experiment. The model indicated to explain the proportion for combining Min and HM was found to be

$$\hat{p} = 0.2956515 - 0.0085824I - 0.0298501\delta + 0.0055853I\delta, \tag{10}$$

where I is the number of treatments and δ is the unbalance level, i.e., the ratio between the maximum and the minimum number of replications. The determination coefficient was found to be $R^2 = 28.63\%$, what suggests that p is a function of more effects that can be considered in further studies. This is not the best model in R^2, but as the model gets more complex the gain in R^2 is not relevant, less of 4% with all the 4 initial quantities and they interactions. Hence, this model is indicated.

Conclusions

Some of the studied functions fail to control the type I error even for few treatments and weak unbalance. This is the case of the functions Max, WM, QM, AM, Md and GM. Therefore, these functions are not indicated for correcting the Tukey's statistic for unbalanced experiments.

For all corrections, the type I error seems to increase faster when the maximum number of replications is bigger. When the minimum value decreases, type I error also increases but in a slower way.

The type I error rate observed for the Min function (being even 0% in many situations) classifies it as very conservative. Consequently, the power of the Min function is growing more slowly as the number of treatments increases.

The functions Md and AM have similar behavior for both type I error and for power.

For the proposed corrections, the harmonic mean and the minimum value of replications, the power decreases and the type I error rate increases as the unbalance gets stronger, independent of the number of treatments.

The HM controlled the type I error only for some treatments numbers and values of δ. The HM was able to maintain the nominal level of significance only for experiments with a maximum of 5 treatments. Over 5 treatments, starts controlling only small imbalances.

The proposed correction (P) managed to keep the type I error rate at/under the nominal level of significance of the test, and showed a reasonable power gain in relation to HM. Therefore, the use of HM in order to calculate the number of repetitions is indicated for a few treatments (maximum 5) and small δ (maximum 2). Otherwise, we advise the use of P, which has a better performance. However, the P correction can be further improved.

As can be seen, the mixture plays its role, improving the harmonic mean correction.

References

BANZATTO, D. A.; KRONKA, S. N. *Experimentação Agrícola.* Jaboticabal: FUNEP. 2006. 237p.

BORGES, L. C.; FERREIRA, D. F.; Poder e taxas de erro tipo I dos testes Scott-Knott, Tukey e Student-Newman-Keuls sob distribuições normais e não normais dos resíduos. Revista de matemática e estatística, 21:67-83. 2003.

DRAPER, N.; SMITH, H.; Applied regression analysis. 3a ed. New York: John Wiley & Sons, 1998. 706 p.

DUNCAN, D. B. Multiple Range and Multiple F Tests. *Biometrics*, Washington, v.11, n.1, p.1-42, 1955.

DUNNETT, C. W.; Pairwise multiple comparisons in the homogeneous variance, unequal sample size case. December 1980. Journal of the American Statistical Association, vol. 75, n. 372, p. 789-795.

FERREIRA, D. F.; Estatística básica. Lavras: Editora UFLA, 2005. 676 p.

FERREIRA, E. B.; CAVALCANTI, P. P. Função em código R para analisar experimentos em DIC simples, em uma só rodada. In: 54^a Reunião da Região Brasileira da Sociedade Internatcional de Biometria, 13^o Simpósio de Estatística Aplicada à Experimentação Agronômica, 2009, São Carlos. *Programas e resumos...* São Carlos, SP: UFSCar, 2009. p. 1-5.

HAYTER, A. J.; A proof of the conjecture that the Tukey-Kramer multiple comparissions procedure is conservative. 1984. The Annalls of Statistics, 12(1), 61-75.

KESELMAN, H. J.; ROGAN, J. C.; A Comparison of the Modified-Tukey and Scheffe Methods of Multiple Comparisons for Pairwise Contrasts. March 1978. Journal of the American Statistical Association, Washington, vol. 73, n. 361, p. 47-52.

KRAMER, C. Y. Extension of multiple range tests to group means with unequal numbers of replications, *Biometrics*, 12, 307-310. 1956.

MACHADO, A. A.; DEMÉTRIO, C. G. B.; FERREIRA, D. F.; SILVA, J. G. C. da; Estatística Experimental: uma abordagem fundamentada no planejamento e no uso de recursos computacionais. Londrina: Editora da UEL, 2005.

MILLER, R.G., Jr.; Simultaneous Statistical Inference. Journal of the American Statistical Association, New York: McGraw-Hill Book Co. 1966.

MORETTIN, L. G.; Estatística básica - Volume 2 - Inferência. São Paulo: Pearson Makron Books, 2000. 182 p.

PIMENTEL-GOMES, F.; Curso de estatística experimental. 15. ed. Piracicaba: FEALQ, 2009. 451 p.

R CORE TEAM. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. 2013.

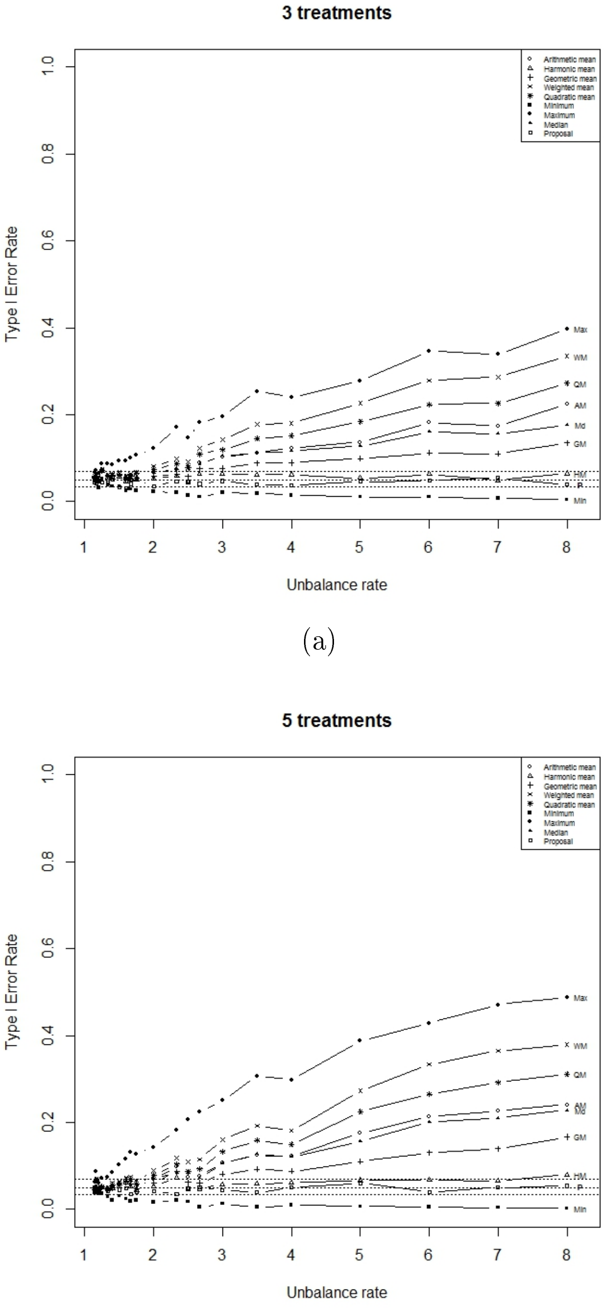
RAMALHO, M. A. P.; FERREIRA, D. F.; OLIVEIRA, A. C. de; Experimentação em genética e melhoramento de plantas. 2. ed. Lavras: UFLA, 2005. 322 p.

SAMPAIO, I. B. M.; Estatística Aplicada à Experimentação Animal. 3. ed reimpressão. Belo Horizonte: Fundação de Estudo e Pesquisa em Medicina Veterinária e Zootecnia, 2010. 264 p.

TUKEY, J.W.; The Problem of Multiple Comparisons. Un-published report, Princeton University. 1953.

Appendix: Extra graphics

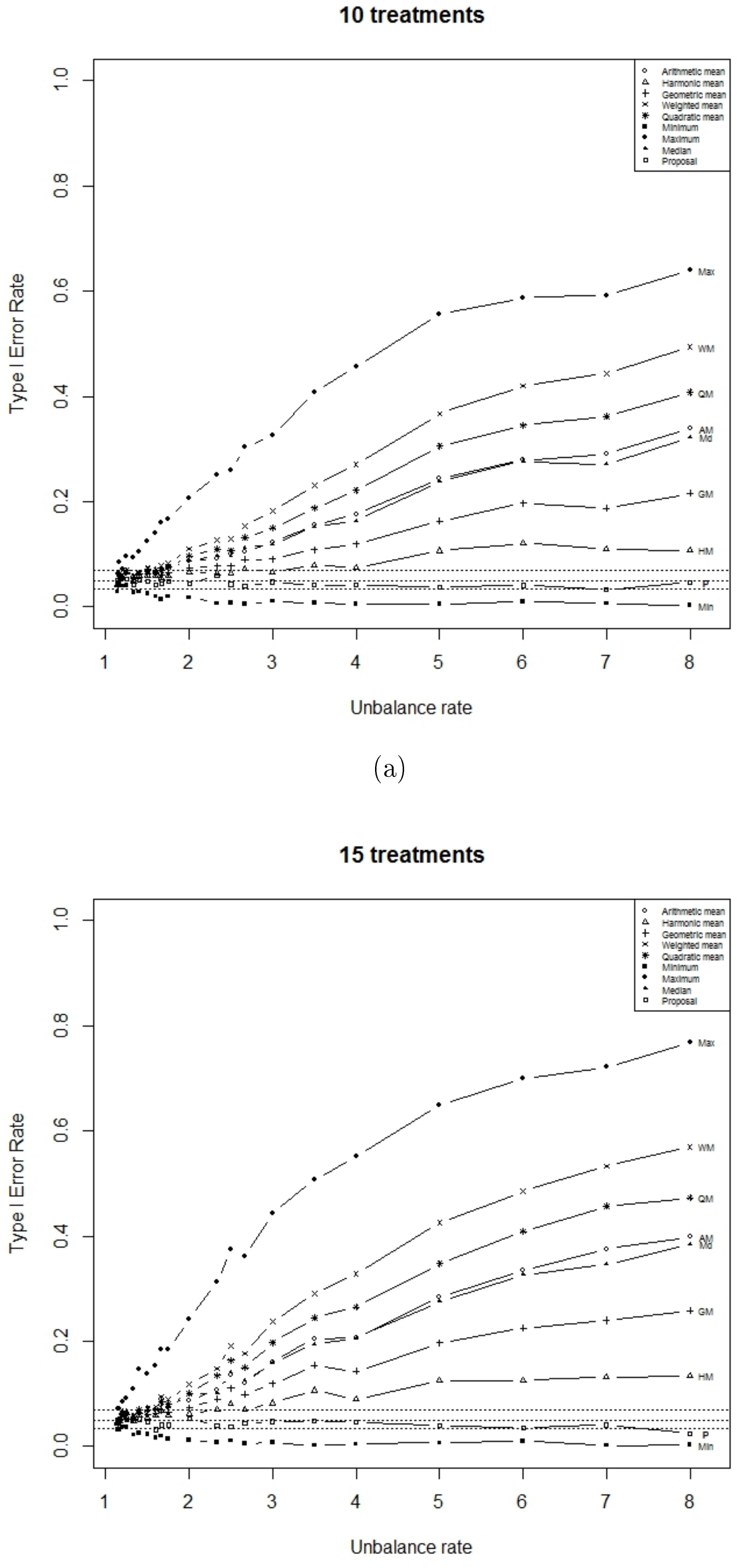
Type I error rate



(a)

(b)

Figure 6: Type I error rate along unbalance levels (δ) for 3 (a) and 5 treatments (b).



(a)

(b)

Figure 7: Type I error rate along unbalance levels (δ) for 10 (a) and 15 treatments (b).

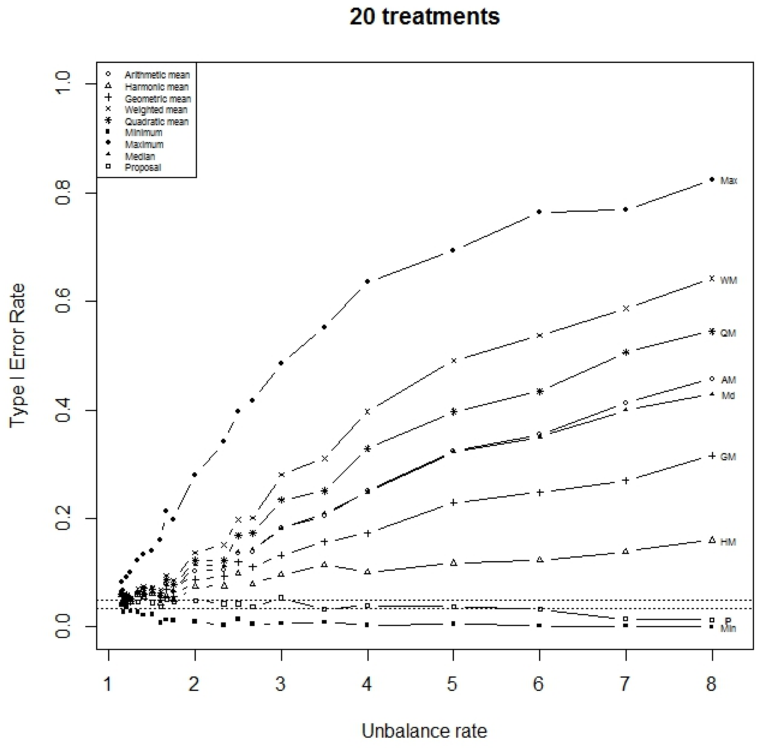


Figure 8: Type I error rate along unbalance levels (δ) for 20 treatments.

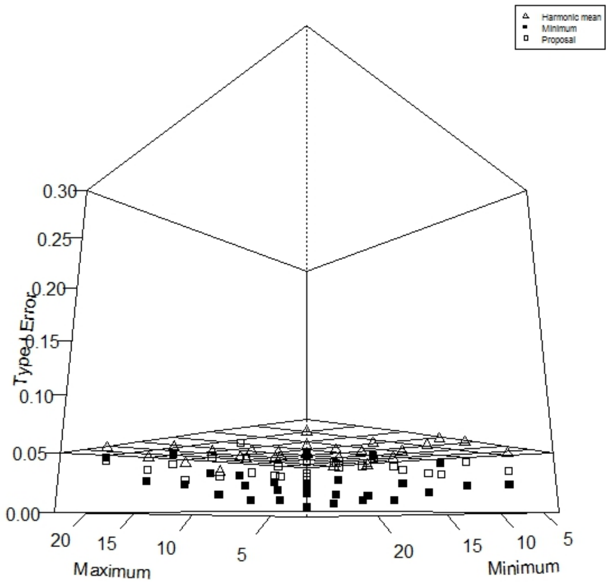
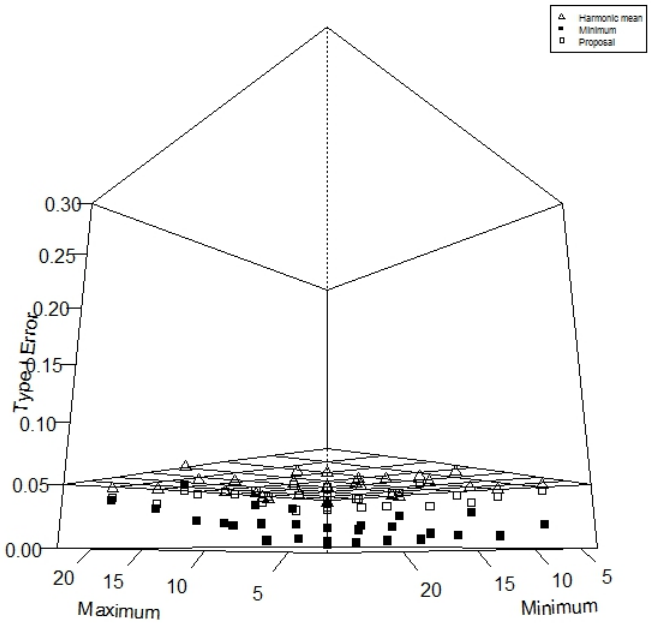
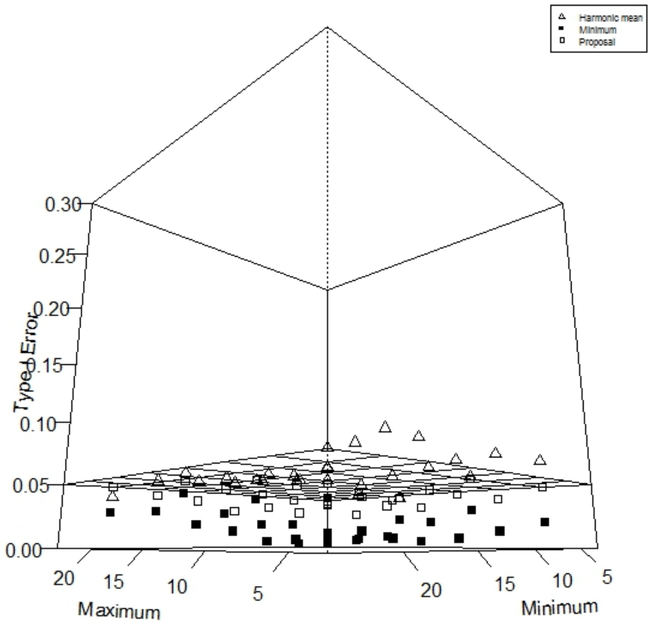


Figure 9: Type I error rate (3D) along minimum and maximum values for J_i for 3 treatments.

Type I error rate: 3D

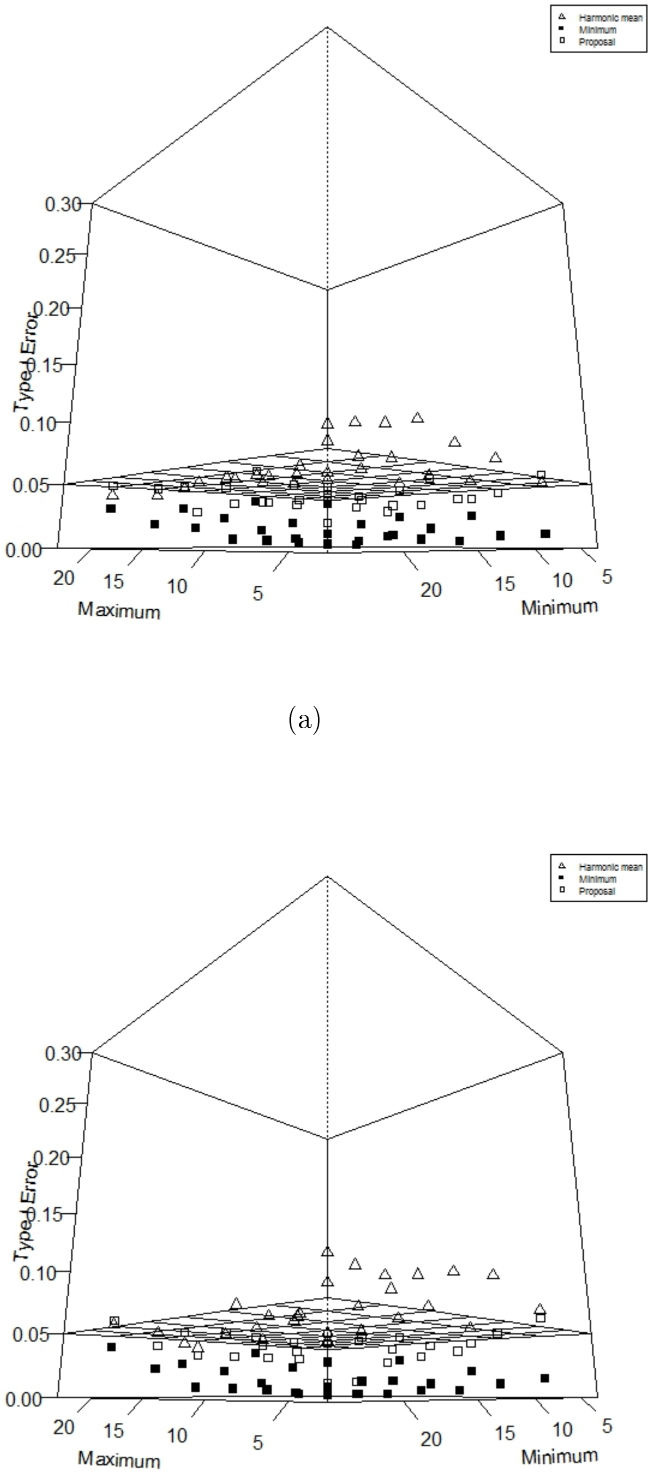


(a)



(b)

Figure 10: Type I error rate (3D) along minimum and maximum values for J_i for 5 (a) and 10 treatments (b).

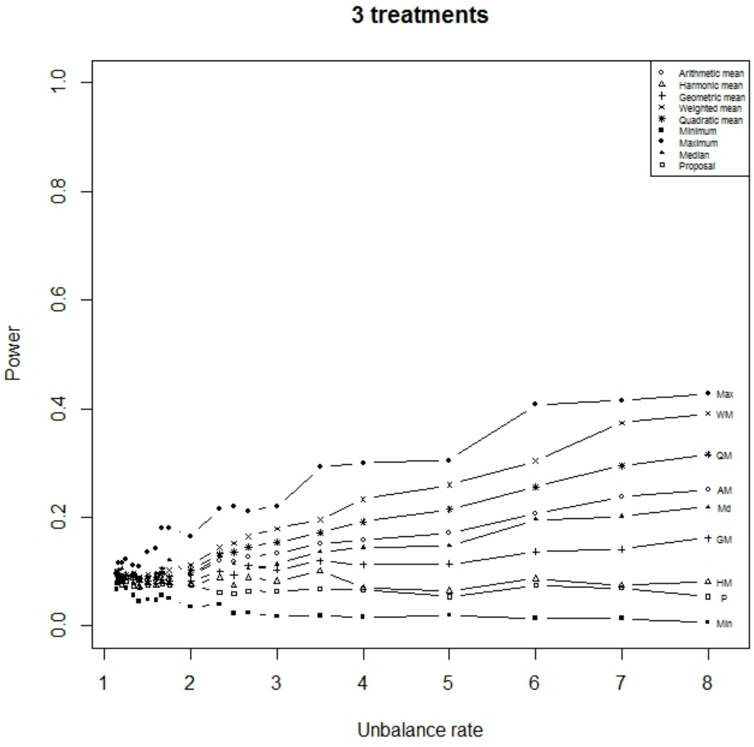


(a)

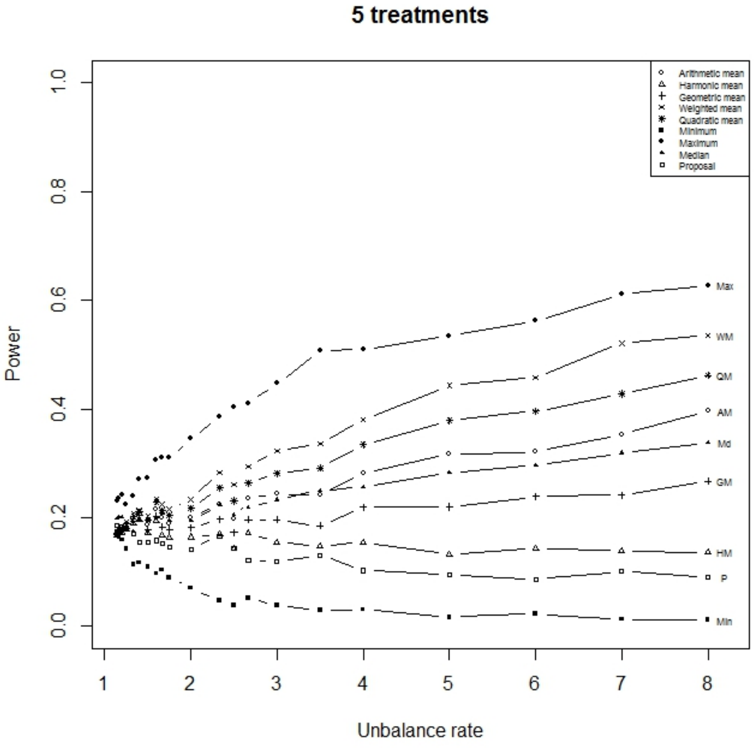
(b)

Figure 11: Type I error rate (3D) along minimum and maximum values for J_i for 15 (a) and 20 treatments (b).

Power

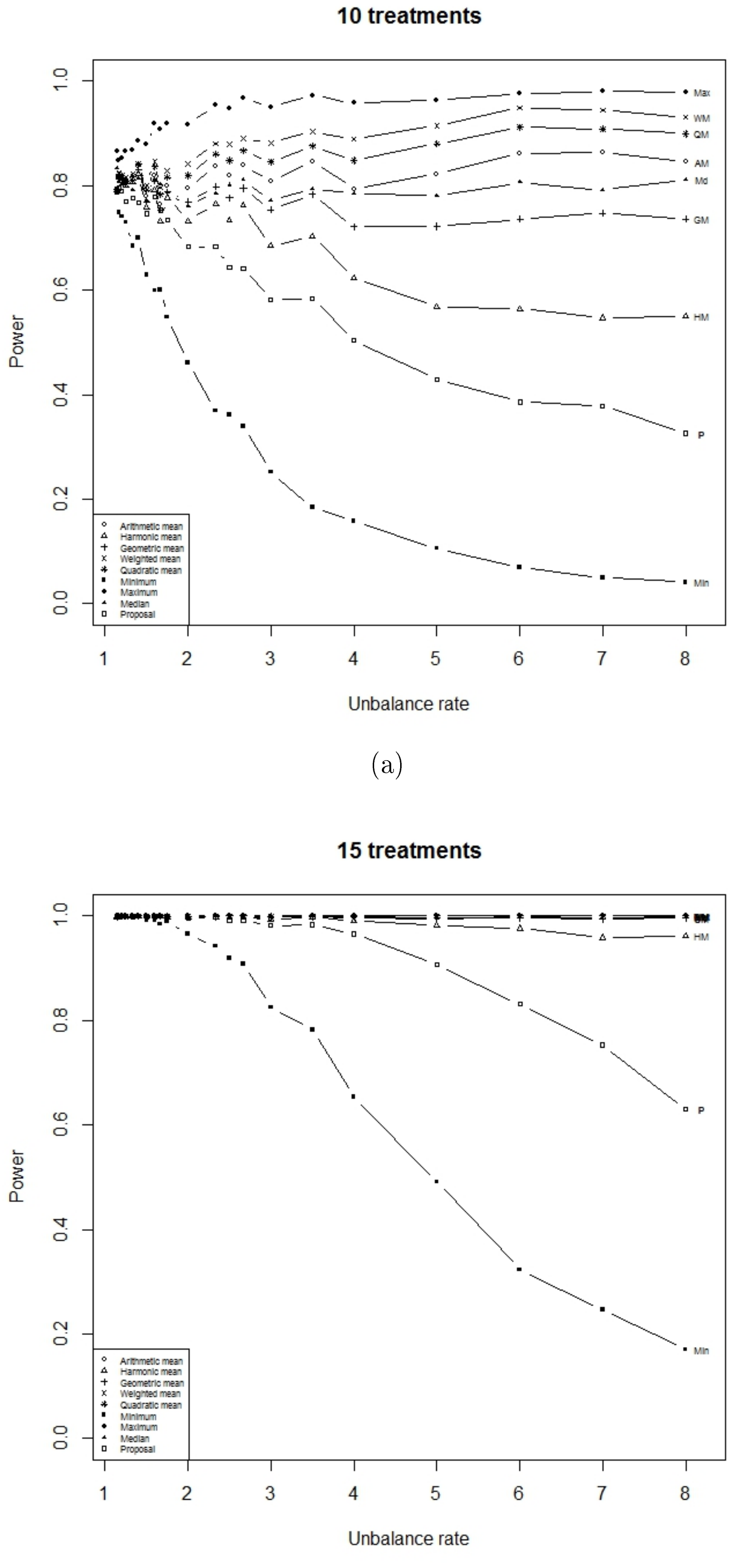


(a)



(b)

Figure 12: Power along unbalance levels (δ) for 3 (a) and 5 treatments (b).



(a)

(b)

Figure 13: Power along unbalance levels (δ) for 10 (a) and 15 treatments (b).

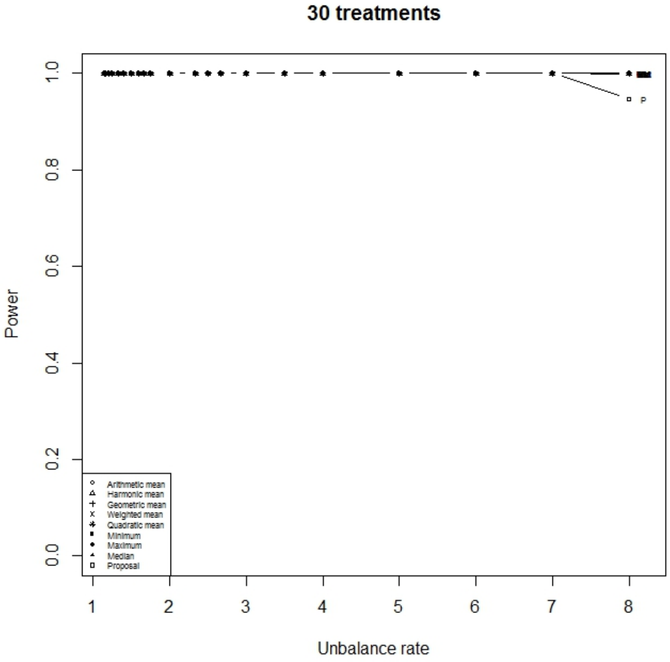


Figure 14: Power along unbalance levels (δ) for 20 treatments.

Real power

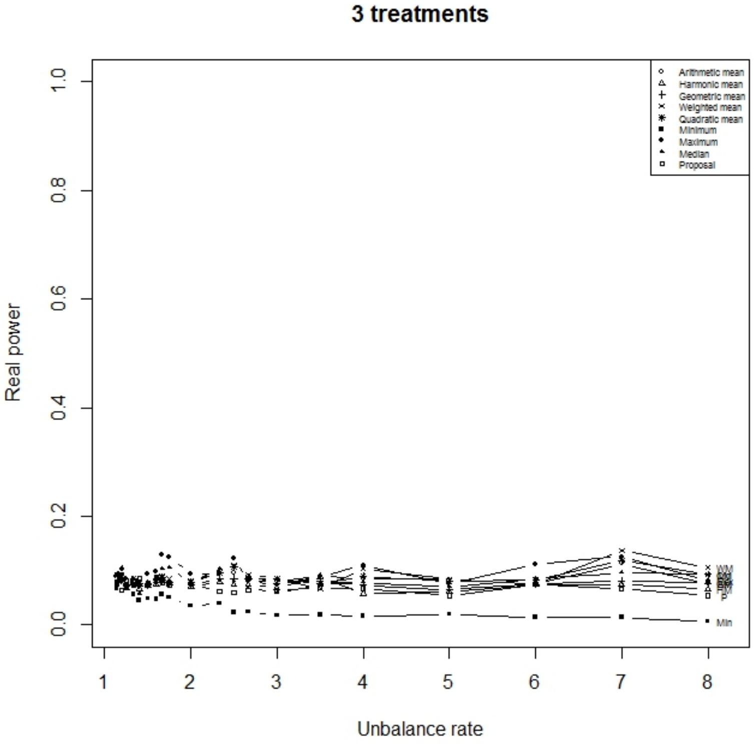
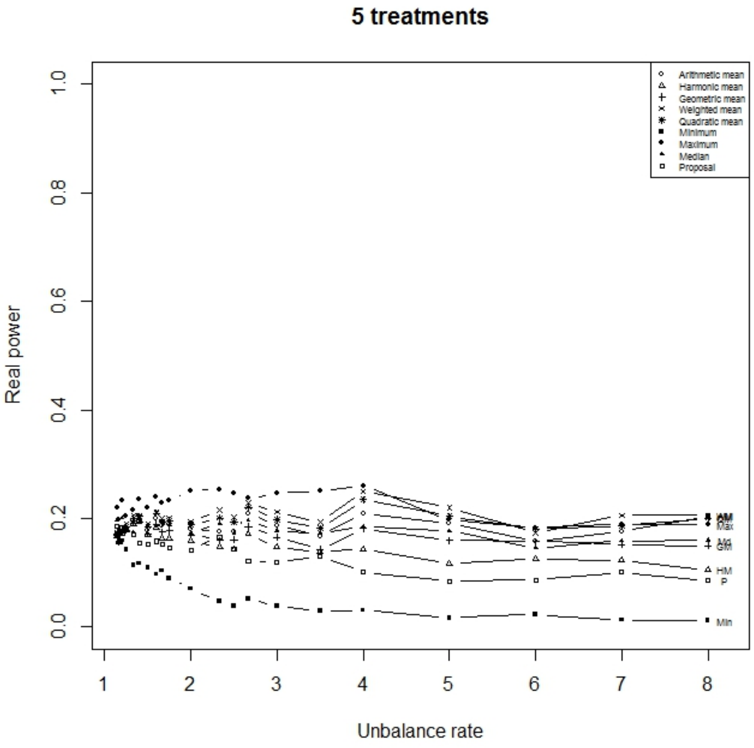
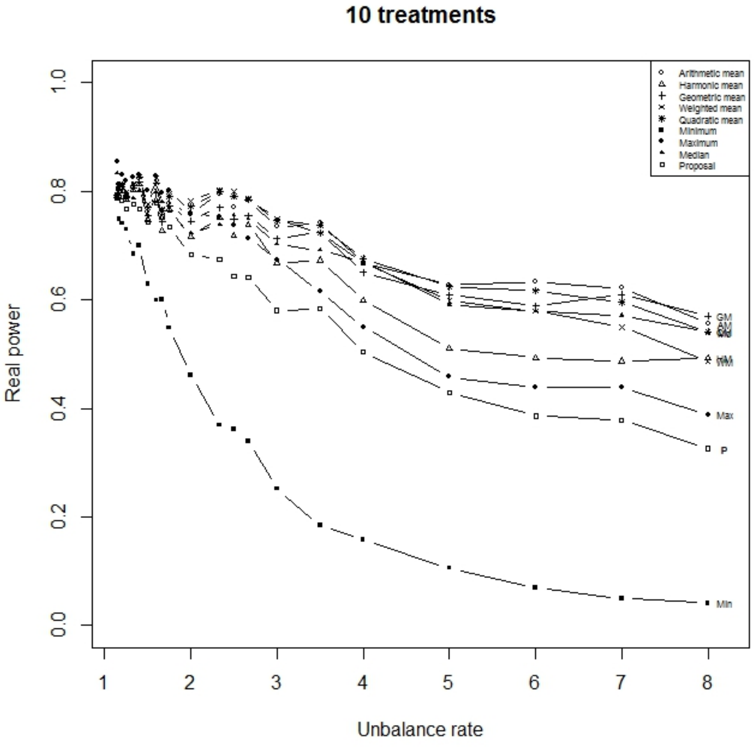


Figure 15: Real power along unbalance levels (δ) for 3 treatments.

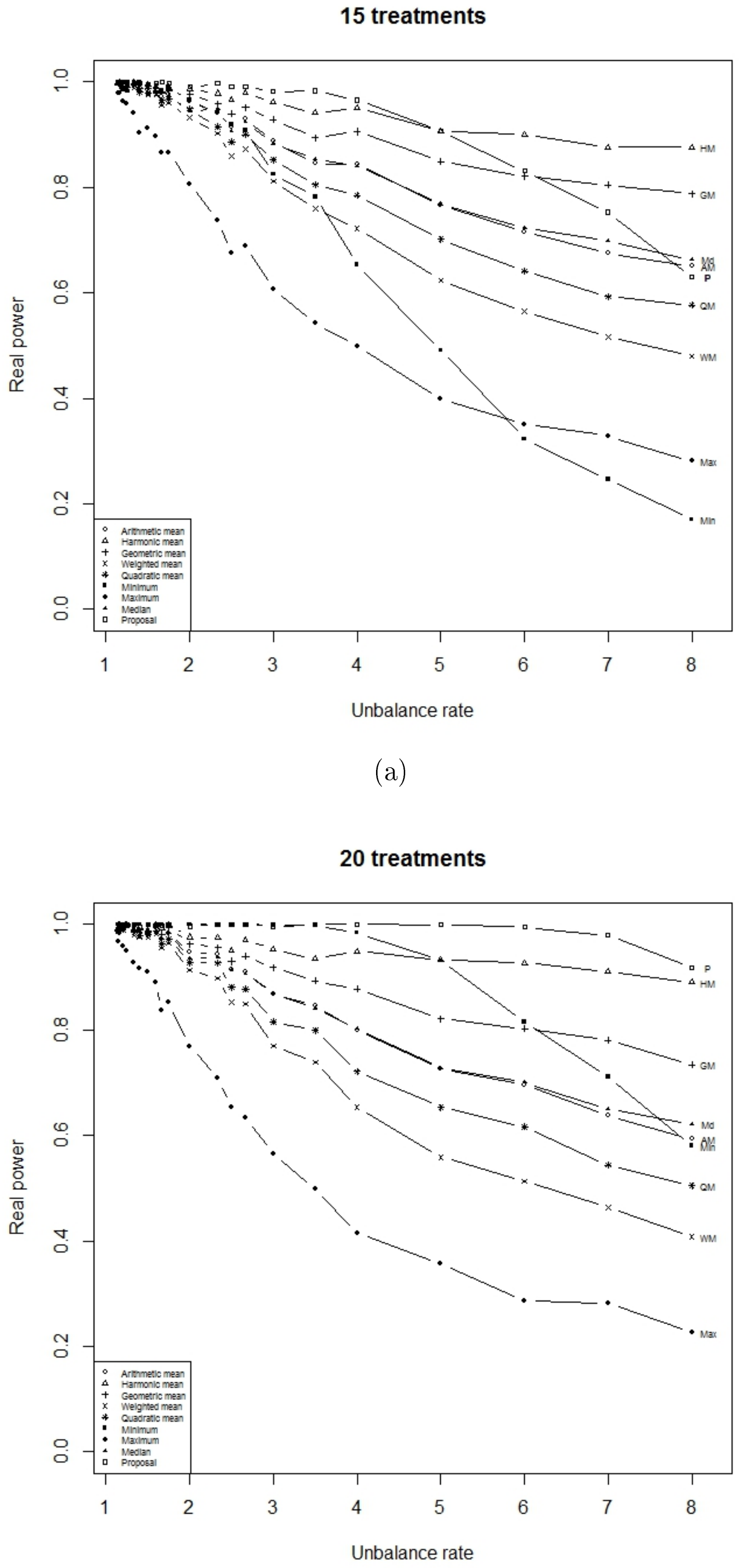


(a)



(b)

Figure 16: Real power along unbalance levels (δ) for 5 (a) and 10 treatments (b).



(a)

(b)

Figure 17: Real power along unbalance levels (δ) for 15 (a) and 20 treatments (b).