Overview of Friedman's Test and Post-hoc Analysis

Article *in* Communication in Statistics- Simulation and Computation · November 2015

DOI: 10.1080/03610918.2014.931971

CITATIONS
93

READS
13,551

3 authors, including:

D. G. Pereira
Universidade de Évora
25 PUBLICATIONS **218** CITATIONS
SEE PROFILE

Anabela Afonso
Universidade de Évora
41 PUBLICATIONS **160** CITATIONS
SEE PROFILE

Communications in Statistics - Simulation and Computation

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lssp20

Overview of Friedman's Test and Post-hoc Analysis

Dulce G. Pereira[a], Anabela Afonso[a] & Fátima Melo Medeiros[b]

[a] Department of Mathematics and Research Center of Mathematics and Applications (CIMA-UE), University of Évora, Évora, Portugal

[b] Department of Biology, University of the Azores, Azores, Portugal

Accepted author version posted online: 21 Aug 2014.Published online: 21 Aug 2015.

To cite this article: Dulce G. Pereira, Anabela Afonso & Fátima Melo Medeiros (2015) Overview of Friedman's Test and Post-hoc Analysis, Communications in Statistics - Simulation and Computation, 44:10, 2636-2653, DOI: 10.1080/03610918.2014.931971

To link to this article: http://dx.doi.org/10.1080/03610918.2014.931971

Communications in Statistics—Simulation and Computation®, 44: 2636–2653, 2015

ISSN: 0361-0918 print / 1532-4141 online
DOI: 10.1080/03610918.2014.931971

Overview of Friedman's Test and Post-hoc Analysis

DULCE G. PEREIRA,[1] ANABELA AFONSO,[1]
AND FÁTIMA MELO MEDEIROS[2]

[1]Department of Mathematics and Research Center of Mathematics and Applications (CIMA-UE), University of Évora, Évora, Portugal
[2]Department of Biology, University of the Azores, Azores, Portugal

*When the null hypothesis of Friedman's test is rejected, there is a wide variety of multiple comparisons that can be used to determine which treatments differ from each other. We will discuss the contexts where different multiple comparisons should be applied, when the population follows some discrete distributions commonly used to model count data in biological and ecological fields. Our simulation study shows that sign test is very conservative. Fisher's LSD and Tukey's HSD tests computed with ranks are the most liberal. Theoretical considerations are illustrated with data of the Azores Buzzard (*Buteo buteo rothschildi*) population from Azores, Portugal.*

Keywords Multiple comparison methods; Nonparametric tests; Related samples.

Mathematics Subject Classification 62K10; 62J15; 62K10.

1. Introduction

The parametric technique for testing whether several samples have come from identical populations is the Analysis of Variance (ANOVA) and associated F statistics.

Many of ANOVA's developments were done in agriculture research, which explains the usual terminology used in this technique. In typical applications of ANOVA, the experimenter wants to study the effect of the different levels of one or more factors, usually referred to as treatments, in a variable of interest.

There are two basic designs for comparing K treatments or groups. The first design involves K independent random samples, not necessarily of the same size; one sample from each population. For this design, the statistical tests for K independent samples should be employed. Depending on the number of factors, there is one-way ANOVA (one factor), two-way ANOVA (two factors), and so on (Chiang, 2003). In the second design, K samples of equal size are matched according to some criterion or criteria, which may affect the values of the observations. In some cases, the matching is achieved by comparing the same n individuals or cases, usually referred as blocks, under all K treatments or conditions. For such designs one of the factors is called the blocking factor, and the statistical test for K related samples should be used: the repeated measures ANOVA or the randomized-blocks ANOVA.

Received October 16, 2012; Accepted June 3, 2014
Address correspondence to Dulce G. Pereira, Department of Mathematics and Research Center of Mathematics and Applications (CIMA-UE), University of Évora, Évora, Portugal; E-mail: dgsp@uevora.pt

The assumptions associated with the statistical model underlying the ANOVA are (Sheskin, 2007):

1. scores/observations are independently drawn from normally distributed populations;
2. all populations have the same variance, i.e., homogeneity of variance;
3. means of the populations are linear combinations of levels' "effects" of each factor.

Sometimes the assumptions of ANOVA are unrealistic: observations do not meet the measurement requirements, normality of the distributions is violated or variances are not equal. There are also situations where we wish to avoid making the assumptions in order to increase the generality of the findings. In these situations, the nonparametric statistical tests would be appropriate. Nonparametric methods require less assumptions about the underlying populations. Zimmerman and Zumbo (1993) examined the effects of some ANOVA nonparametric tests with samples from Normal, Uniform, Mixed-Normal, Exponential, Laplace, and Cauchy distributions. They concluded that Friedman's test is more powerful than parametric ANOVA for very skewed (heavy tail) distributions, like Cauchy distribution. This is because usually these methods are insensitive to outliers and are often more powerful than parametric methods, if the underlying assumptions of the parametric model do not hold (Conover, 1999; Hollander and Wolfe, 1999; Sheskin, 2007).

The well-known Kruskall-Wallis test is the nonparametric alternative to one-way ANOVA with independent samples. The nonparametric Friedman's test is the alternative nonparametric procedure to the parametric ANOVA for the repeated measures one-way ANOVA or randomized-blocks one-way ANOVA. When the null hypothesis is rejected post-hoc tests can be applied to compare conditions/treatments.

Demsar (2006), García et al. (2010) and Derrac (2011) explained Friedman's test (Friedman, 1937), the Iman Davenport correction (Iman and Davenport, 1980) and some post-hoc procedures with adjusted p-values, such as Bonferroni−Dunn (Dunn, 1961), Holm (Holm, 1979), Hochberg (Hochberg, 1988), and Hommel (Hommel, 1988).

Iman and Davenport (1980) showed that the test statistic proposed by Friedman (1937) is undesirably conservative and derived a better statistic, the Iman Davenport correction.

Church and Wike (1979) argued that Wilcoxon test was the best pairwise multiple comparison procedure when compared with the Wilcoxon-Nemenyi-McDonald-Thompson and sign tests. Bonferroni procedure tests each one of the several pairs of comparisons at the same level of significance. Holm, Hochberg, and Hommel procedures are improvements of the Bonferroni procedure (Wright, 1992). These procedures uses unequal allocation to each individual hypothesis tested according to a criterion. Holm is a step-down procedure while Hochberg and Hommel are step-up procedures that sequentially test the hypotheses ordered by their significance. Demsar (2006) described a set of nonparametric tests for performing multiple comparisons and analyzed them, in contrast to well-known parametric tests, in terms of power. He concluded that the nonparametric tests are more suitable when the required conditions of the parametric tests are not fulfilled. García et al. (2010) and Derrac et al. (2011) presented a comprehensive and useful tutorial about the use of nonparametric statistical tests in computational intelligence, using tests already proposed in several papers of the literature. Garcia et al. (2010) focus on the case in which a control treatment is compared against other treatments. For a set of post-hoc procedures they performed an experimental analysis to estimate their power, with the purpose of detecting the advantages

and disadvantages of the statistical tests described. They also presented a useful guideline for their use.

In the literature, the power and relative efficiency of Friedman's test and multiple comparisons procedures are studied with continuous distributions (e.g., Church and Wike, 1979; Derrac et al., 2011; Garcia, 2010; Iman et al., 1984; O'Gorman, 2001; St. Laurent and Turk, 2013; Zimmerman and Zumbo, 1993). But Friedman's test is a nonparametric test and therefore it is not restricted to continuous distribution functions and any models with discrete observations may also be used. However, studies with discrete distributions are rare or nonexistent.

The goal of this article is to discuss the context in which the different multiple comparisons should be applied, after Friedman's test, when the population follows a Binomial, Poisson, Negative Binomial, or Uniform distribution. These discrete distributions are commonly used to model count data. Section 2 presents Friedman's test and the Iman and Davenport correction. Section 3 addresses the more general nonparametric tests for performing multiple comparisons. We have considered those that are not excessively complicated, well-known in statistics, and easily available in statistical books (Conover, 1999; Hollander, 1999; Sheskin, 2007; Siegel, 1988). Section 4 describes a simulation study that was developed to compare the power of the post-hoc procedures. The results of this simulation study and their implications are reported in Section 5. Finally, we provide an illustrative example using data of the Azores Buzzard (*Buteo buteo rothschildi*) population from Azores, Portugal.

2. Friedman's Test

For Friedman's two-way analysis of variance by ranks, the null hypothesis states that the K repeated measures or matched groups come from the same population or from populations with the same median (Siegel and Castellan, 1988). Under the null hypothesis, the test assumes that the response variable has the same underlying continuous distribution; thus it requires at least ordinal measurement of that variable. The data are casted in a two-way table having n rows and K columns. The rows represent the blocks/individuals or matched sets of individuals, and the columns represent the various conditions/treatments. The data of the test are ranks ($R_{ik}, i = 1, ..., n; k = 1, ..., K$) of the treatments by blocks ; therefore, $1 \leq R_{ik} \leq K, i = 1, .., n$. In case of ties, average ranks are used. The underlying assumptions of Friedman's test are (Conover, 1999):

1. The n K-variate random variables are mutually independent, i.e., the results within one row do not influence the results within the other rows;
2. The observations in each row can be ranked separately according to some criterion of interest.

Friedman's test determines whether the rank totals for each condition/treatment differ significantly from the values which would be expected by chance.

The test statistic suggested by Friedman's is (Siegel and Castellan, 1988)

$$T_1 = \frac{12}{nK(K+1)} \sum_{k=1}^{K} R_{.k}^2 - 3n(K+1), \tag{1}$$

where $R_{.k} = \sum_{i=1}^{n} R_{ik}$ is the sum of the ranks for treatment k over the n blocks. Under the null hypothesis, as n tends to infinity, this statistic T_1 has an asymptotic Chi-square distribution with $K-1$ degrees of freedom. At the α level of significance, the null hypothesis is rejected if $T_1 \geq \chi^2_{K-1;1-\alpha}$, where $\chi^2_{K-1;1-\alpha}$ is the $(1-\alpha)$ quantile of the Chi-square distribution with $K-1$ degrees of freedom. For small values of n and K, exact critical values have been computed and are available in Siegel and Castellan (1988).

When there are ties among observations in a given block, the test statistic T_1 may be used,

$$T_1 = \frac{4(K-1)\sum_{k=1}^{K}\left(R_{.k} - \frac{n(K+1)}{2}\right)^2}{4\sum_{i=1}^{n}\sum_{k=1}^{K} R_{ij}^2 - nK(K+1)} \overset{a}{\sim} \chi^2_{K-1}. \tag{2}$$

Iman and Davenport (1980) noted that the Chi-square approximation is sometimes poor and too conservative with high propensity to increase error Type II, i.e., has less power. These authors remarked that the "Chi-square approximation quickly falls off as K increases for fixed n" and the "F approximation improves as K increases and is liberal, but still dominates the Chi-square approximation". Therefore, they suggested the use of test statistic T_2 which has an approximate F distribution and usually is more powerful than T_1:

$$T_2 = \frac{(n-1)T_1}{n(K-1) - T_1} \overset{a}{\sim} F_{K-1;(K-1)(n-1)}. \tag{3}$$

The null hypothesis is rejected if $T_2 \geq F_{K-1;(K-1)(n-1);1-\alpha}$, where $F_{K-1;(K-1)(n-1);1-\alpha}$ is the $(1-\alpha)$ quantile of the F distribution with $K-1$ and $(K-1)(n-1)$ degrees of freedom.

Nevertheless, several software still use the Chi-square approximation.

3. Post-hoc Tests

When significant differences are detected between treatments, i.e., the null hypothesis of Friedman's test is rejected, there are several post-hoc tests that can be applied to find out which treatments differ from the others (unplanned comparisons). Therefore, the hypotheses to test are $H_0: \theta_k = \theta_j, k \neq j$ vs $H_1: \theta_k \neq \theta_j$, where θ_k represents the median of the treatment k.

There are several multiple comparison techniques available in the literature to investigate differences between pairs of population medians. However, user may be poorly informed about the apporpriate method(s) to used. Here, we focus on the tests commonly available in statistical software packages.

The Bonferroni-Dunn test (Siegel and Castellan, 1988) is flexible and can be used to test differences between two treatments, as well as among all treatments. It employs an adjustment in the critical value used to reject the null hypothesis in order to reduce the familywise Type I error rate, i.e., $1-(1-\alpha)^c$, where $c = K(K-1)/2$ is the number of comparisons and α is the per comparison Type I error rate. This adjustment insure that the overall likelihood of committing at least one Type I error in the set of comparisons will not exceed a prespecified α (Sheskin, 2007). The treatments k and j are considered significantly different if

$$|R_{.k} - R_{.j}| \geq z_1 - \frac{\alpha}{K(K-1)}\sqrt{\frac{nK(K+1)}{6}}, \tag{4}$$

where $z_{1-\frac{\alpha}{K(K-1)}}$ is the $(1-\frac{\alpha}{K(K-1)})$ quantile of the standard Normal distribution.

Fisher's LSD test can be used when test statistic is computed on the ranks R_{ik} instead of on the original data (Conover, 1999). The null hypothesis H_0 is rejected if

$$|R_{.k} - R_{.j}| \geq t_{(n-1)(K-1);1-\frac{\alpha}{2}} \sqrt{2\frac{n\sum_{i=1}^{n}\sum_{k=1}^{K} R_{ik}^2 - \sum_{k=1}^{K} R_k^2}{(n-1)(K-1)}}, \tag{5}$$

where $t_{1-\frac{\alpha}{2}}$ is the quantile of probability $(1-\frac{\alpha}{2})$ of the t distribution with $(n-1)(K-1)$ degrees of freedom. This test is computationally identical to the Bonferroni-Dunn test except for the fact that the test statistic equation does not employ any adjustment in Type I error. Thus, Fisher's LSD test does not protect the familywise Type I error rate (Skeskin, 2007).

The Wilcoxon-Nemenyi-McDonald-Thompson test (Hollander and Wolfe, 1999) controls the familywise Type I error rate so that it will not exceed the prespecified alpha value. With this test, H_0 is rejected if

$$|R_{.k} - R_{.j}| \geq r_{n;K;1-\alpha}. \tag{6}$$

The approximate values $r_{n;K;1-\alpha}$ can be found in Hollander and Wolfe (1999). When $n \to \infty$ this test is performed in a way similar to Tukey's HSD test using rank sums, and H_0 is rejected if

$$|R_{.k} - R_{.j}| \geq q_{K;n-K;1-\alpha}\sqrt{\frac{nK(K+1)}{12}}, \tag{7}$$

where $q_{K;n-K;1-\alpha}$ is the $(1-\alpha)$ quantile of the Studentized Range distribution with K and $(n-K)$ degrees of freedom. This test is in general more powerful than the Bonferroni-Dunn (Sheskin, 2007).

The Binomial Sign test (Sheskin, 2007) assumes that there is information only in the signs of the differences between paired observations. Let S be the frequency of the more occurring sign (positive or negative) in D_i, where $D_i = X_{ik} - X_{ij}$ for each block i, $i = 1, .., n$. The decision rule for the Binomial Sign test, with the Bonferroni correction, is rejected H_0 if

$$S \geq b_1 - \frac{\alpha}{K(K-1)}, \tag{8}$$

where $b_{1-\frac{\alpha}{K(K-1)}}$ is the $(1-\frac{\alpha}{K(K-1)})$ quantile of the Binomial distribution, $B(n', \frac{1}{2})$, and n' is the sample size without ties. If n' is sufficiently large, then a Normal approximation can be used, and H_0 is rejected if

$$S \geq \frac{n'}{2} - 0.5 + z_1 - \frac{\alpha}{2}\frac{\sqrt{n'}}{2}, \tag{9}$$

where $z_{1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})$ quantile of the standard Normal distribution.

The application of the Wilcoxon test (Sheskin, 2007) requires that the absolute score differences be rank-ordered, i.e., rank $|D_i|$. Let W be the total rank sum of the more occurring sign (positive or negative). The significance of individual pairs of differences

between treatments is tested by using the following inequality

$$W \geq W_{n';1} - \frac{\alpha}{K(K-1)}, \tag{10}$$

where $W_{n';1-\frac{\alpha}{K(K-1)}}$ is the $(1-\frac{\alpha}{K(K-1)})$ quantile of the Wilcoxon distribution (tabulated or software available) with the Bonferroni correction, and n' is the sample size without ties. As n' increases, the sampling distribution converges to a Normal distribution. Thus, treatment k and j are considered significantly different if

$$W \geq \frac{n'(n'+1)}{4} + z_1 - \frac{\alpha}{K(K-1)}\sqrt{\frac{n'(n'+1)(2n'+1)}{24}}, \tag{11}$$

where $z_{1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})$ quantile of the standard Normal distribution.

Both Binomial Sign and Wilcoxon test statistics are based on the score differences between treatments. The Wilcoxon test employs more information than the Binomial Sign test and, consequently, the first will provide a more powerfull test (Sheskin, 2007).

The Bonferroni procedure is very simple. However, the price of this simplicity and generality is a lack of power. To improve this situation, several modifications to the Bonferroni procedure have been proposed (Wright, 1992), such as the Holm procedure (Holm, 1979), the Hochberg procedure (Hochberg, 1988) and the Hommel procedure (Hommel, 1988). These modifications can be applied to Binomial Sign and Wilcoxon tests. Let $p_{(1)} \leq ... \leq p_{(c)}$ be the ordered p-values and $H_0^{(1)}, ..., H_0^{(c)}$ the associated null hypotheses:

- The Holm procedure (Holm, 1979) is based on a sequentially rejective algorithm. Let m be the minimal index such that $p_{(m)} > \frac{\alpha}{c-m+1}$. Reject the null hypothesis $H_0^{(1)}, ..., H_0^{(m-1)}$ and do not reject $H_0^{(m)}, ..., H_0^{(c)}$.
- The Hochberg procedure (Hochberg, 1988) proceeds similarly to the Holm procedure except for the order in which the null hypotheses are tested. Let M be the maximal index such that $p_{(M)} \leq \frac{\alpha}{c-M+1}$. Reject the null hypothesis $H_0^{(1)}, ..., H_0^{(M)}$ and do not reject $H_0^{(M+1)}, ..., H_0^{(c)}$.
- The Hommel procedure (Hommel, 1988) is more complicated to carry out than the Holm and Hochberg procedures. For this procedure, compute

$$M = \max\{a \in \{1, ..., c\} : p_{(c-a+b)} > \frac{b\alpha}{a} \text{ for } b = 1, ..., a\}. \tag{12}$$

  If the maximum does not exist, reject all $H_0^{(a)}$ $(a = 1, ..., c)$, otherwise reject all $H_0^{(a)}$ with $p_a \leq \frac{\alpha}{M}$.

Thus, Holm is step-down procedure since it starts by rejecting the hypothesis associated to the most significant test statistic and then sequentially considers the most significant of the remaining hypotheses. Hochberg e Hommel are step-up procedures where testing begins with the hypothesis associated with the least significant test statistic and continues in increasing order of significance as long no rejection occurs or there are no more hypotheses to test (Lehmann et al., 2005; Tamhane et al., 1998).

The adjusted p-values for Hochberg method are uniformly smaller than for Holm's method, and always as large as or larger than Hommel's procedure. However, both Hocheberg and Hommel procedures only preserve the familywise Type I error rate when individual test statistic are independent or positively dependent (Sarkar and Chang, 1997).

Sometimes we have more specific comparisons in mind than the set of multiple comparisons between all groups or conditions. For instance, to test a more precise hypothesis concerning the difference between one condition (a control condition) and the other conditions, then specific comparisons may be tested as well (planned comparisons). For these cases, the tests described above can be applied, but a Type I error of α/c should be used, where $c = K - 1$ is the number of comparisons. Dunnett test can also be used with rank sums and the decision rule is given by

$$|R_{.k} - R_{.j}| \geq d_{K;n-K;1-\alpha}\sqrt{\frac{nK(K+1)}{6}}, \tag{13}$$

where $d_{K;n-K;1-\alpha}$ are the tabulated values proposed by Dunnett (1964).

In the literature one can find several definitions of multiple Type II error rate and statistical power (e.g., Senn and Bretz, 2007; Westfall et al., 1999). The power of a multiple testing procedure can be defined as minimal power (probability to reject at least one false null hypothesis), complete power (probability to reject all false null hypotheses), individual power (probability to reject an individual false null hypothesis), and average power (average proportion of false null hypotheses that are rejected). Explicit formulas of minimal power and r-power (probability to reject at least r false null hypothesis) for stepwise multiple testing procedures can be found in Chen et al. (2011). In this article, we will consider the minimal power.

4. Monte Carlo Study Description

A Monte Carlo study was conducted to determine how well differences between treatments are detected by Friedman and post-hoc tests, in the analysis of discrete data. Friedman tests the null hypothesis that there are no differences between k treatments against the alternative hypothesis that there are differences between treatments. For each pair of treatments (k, j), post-hoc procedures tests that there are no differences between these two treatments against the alternative hypothesis that the treatments k and j are considered significantly different.

We assumed that the additive model holds (O'Gorman, 2001). Thus,

$$X_{ik} = \theta + \beta_i + \tau_k + \varepsilon_{ik},$$

where θ is the overall median, β_i is the unknown effect of block i, τ_k is the unknown effect of treatment k, and ε_{ik} is the random effect in block i and treatment k.

The study was designed to simulate a wide range of situations that might be encountered in the analysis of real count data. In the simulations, the following scenarios were considered for block and treatment effects:

Binomial: $\beta_i \frown B(N; p)$, $\tau_k \frown B(N; p_k)$, with $p_k = (k-1)/K$, and $\varepsilon_{ik} \frown B(N; p)$;
Poisson: $\beta_i \frown P(\lambda)$, $\tau_k \frown P(\lambda_k)$, with $\lambda_k = Nk/K$, and $\varepsilon_{ik} \frown P(\lambda)$;
Discrete Uniform: $\beta_i \frown U(0; N)$, $\tau_k \frown U(0; N_k)$, with $N_k = Nk/K$, and $\varepsilon_{ik} \frown U(0; N)$;
Negative Binomial: $\beta_i \frown NB(N; p)$, $\tau_k \frown NB(N; p_k)$, with $p_k = k/K$, and $\varepsilon_{ik} \frown NB(N; p)$.

In these scenarios, treatments have different effects, i.e., the null hypothesis in Friedman's test is false as well as the null hypotheses in the post-hoc procedures. With the increase of the number of treatments, the effect between consecutive treatments is less detectable.

For each repetition of the simulation we draw a new set of block and treatment effects from these probability distributions. In all the simulations, we considered $\theta = 0$, and $N = 8$ for the Negative Binomial distribution and $N = 15$ to others. For block effects and random errors we assumed $p = 0.5$ for Binomial and Negative Binomial distributions and $\lambda = N/2$ for the Poisson distribution. We choose these values for the parameters to provide guidance for the analysis of the Azores Buzzard data.

In biological and ecological fields, these distributions are commonly used to model count data. They provide a sufficiently wide variety of distributions that are often encountered in practice. They can be used to estimate the abundance of a certain species in the presence of uncertain detection. The Discrete Uniform distribution is the simplest and the least used since it assumes that all values of its domain are equally likely to occur. The Binomial distribution models the number of detected animals in an area with N animals and the probability of detecting an animal is p (Seber, 1982). This model assumes the independence between detections. The Poisson is related to complete spatial randomness, i.e., the density of animals is constant (homogeneous) over the study area and for a random sample of subregions of the study area, the frequency distribution of the number of animals in each region will follow a Poisson distribution (Diggle, 2003). The Negative Binomial distribution is used when animals are clumped, which implies that the variance is greater than the mean, resulting in overdispersion when compared to a Poisson distribution (White and Bennetts, 1996). When $N \to \infty$, $p \to 0$ and Np in such a way that $Np \to \lambda$, Binomial distribution converges to the Poisson distribution with mean λ. When $N \to \infty$, Negative Binomial distribution approaches to the Poisson distribution.

Several combinations of n and K were considered, namely, $n = 3, 4, 5, 6, 10, 15, 20$, and 30 blocks and $K = 3, 4, 5, 6$, and 10 treatments. For each combination of n and K, 10,000 repetitions were generated and the T_1 and T_2 test statistics and seven post-hoc procedures were computed. For each overall test statistics T_1 and T_2, we computed the proportion of repetitions that lead to rejection of the null hypothesis of Friedman's test in the total number of repetitions, and we have called it the *empirical power*. For each post-hoc procedure we counted the total number of repetitions with one or more significant comparisons when the overall Friedman's test was significant. We divided this quantity by the total number of repetitions in which the overall Friedman's test was significant, and we called it also *empirical power*. We used a significance level of $\alpha = 0.05$ for all tests in this simulation study. We considered a linearly increasing mean value for the kth treatment group. If the mean values for the treatments were separated further, all of the tests would have empirical powers near 1. On the other hand, if the mean values were closely grouped, all the tests would have powers near the α level. It is assumed that there is exactly one observation for each treatment and block combination and that the data are counts (discrete variable).

All computations were carried out using R project 2.15.1 software version (R Development Core Team, 2005). We used several functions already available in some packages for this software: friedman (agricolae package); friedman.test and pairwise.wilcox.test (stats package); friedmanmc (pgirmess package); symmetry-test (coin package); and sign.test (BSDA package).

5. Simulation Results

In our simulation study, the power of the test statistics T_1 and T_2 depends on the number of treatments per block (Figs. 1–4). Test statistic T_2, which has an approximate F distribution, did not detect the differences between treatments when $K = 3$ and $n = 3$. For this combination T_1 test statistic presented a very low power. For all other combinations of K and n,

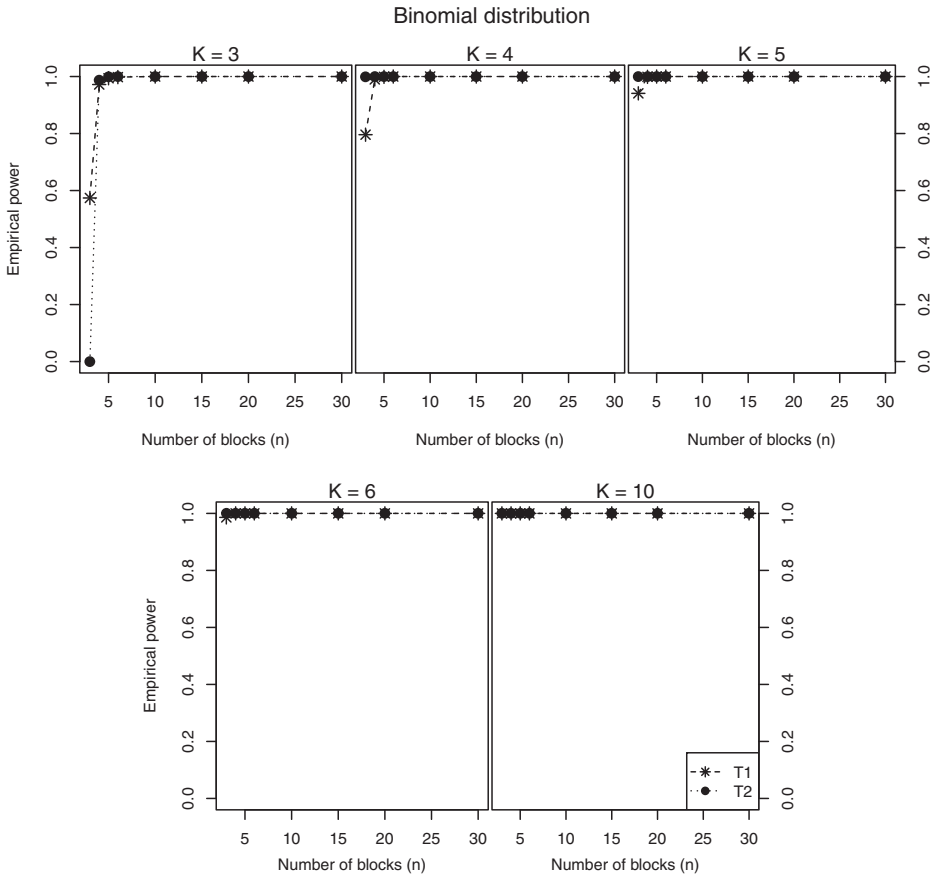

Figure 1. Block power of Friedman's test and Iman-Davenport extension with Binomial distribution, for K = 3, 4, 5, 6, and 10 treatments.

T_1 presented a more conservative behavior than test statistic T_2, especially for small values of n. When $n \geq 10$ and $K \geq 6$, T_2 test statistic presented a power near to 1. When $n \geq 25$ and $K \geq 10$ test statistics T_1 and T_2 agreed perfectly.

For the power analysis of the post-hoc tests, we carried out the procedures explained in Section 3. We computed the proportion of trials that each post-hoc rejected the null hypothesis when Friedman's test rejected its null hypothesis. The same was done for the Iman-Davenport extension. The purpose of this study was to detect and quantify the differences in power observed among the seven post-hoc procedures.

Under rejection of null hypothesis by Friedman's test or by Iman-Davenport extension, the empirical power of the several post-hoc procedures presented the same behavior. Considering the observations already made about test statistics T_1 and T_2, we present the post-hoc results under Friedman's test when $K = 3$ and $n = 3$, and for all other combinations the results are under the Iman-Davenport extension (Figs. 5–8).

For all simulations, the multiple sign test did not detect any difference between treatments, regardless the adjusted p-value considered. For this reason, we did not present the results for this test. The sign test revealed to be an extremely conservative test and therefore we do not recommend its use.

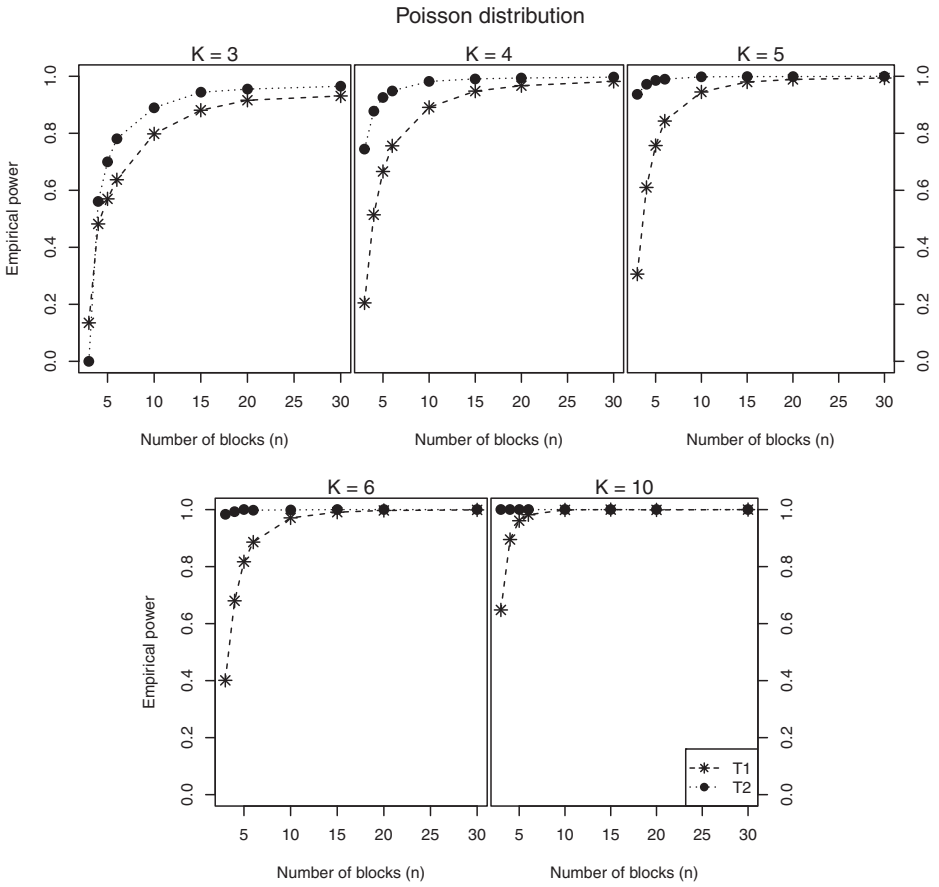

Figure 2. Block power of Friedman's test and Iman-Davenport extension with Poisson distribution, for K = 3, 4, 5, 6 and 10 treatments.

All other post-hoc procedures showed an increase in the empirical power with the increase of n when $K \geq 4$. After a certain combination of n and K the power became close to 1. For Binomial and Negative Binomial distributions all tests presented power near 1 when: $K = 3$ and $K = 4$ and $n \geq 5$; $K = 5$ and $K = 6$ and $n \geq 6$; and $K = 10$ and $n \geq 10$ (Figs. 5 and 7). For Poisson distribution, in all post-hoc procedures, the power achieved a value near 1 for higher combinations of K and n: for $K = 5, 6$ and $n \geq 30$ while for $K = 10$ combining with $n \geq 20$ (Fig. 6). For Discrete Uniform distribution we did not reach that combination (Fig. 8). The decreasing behavior observed, mainly with small values of K, is explained by the small number of overall hypotheses rejected with Friedman's test.

For all distributions, the Wilcoxon test with Holm, Hochberg, Hommel or Bonferroni corrections behaved very similarly among themselves. Holm and Bonferroni p-value adjustment presented exactly the "same empirical" power. These post-hoc tests proved to be less powerful than all other procedures considered. Fisher's LSD presented the highest power followed by Wilcoxon-Nemenyi-McDonald-Thompson test (Tukey-HSD) and Bonferroni-Dunn test.

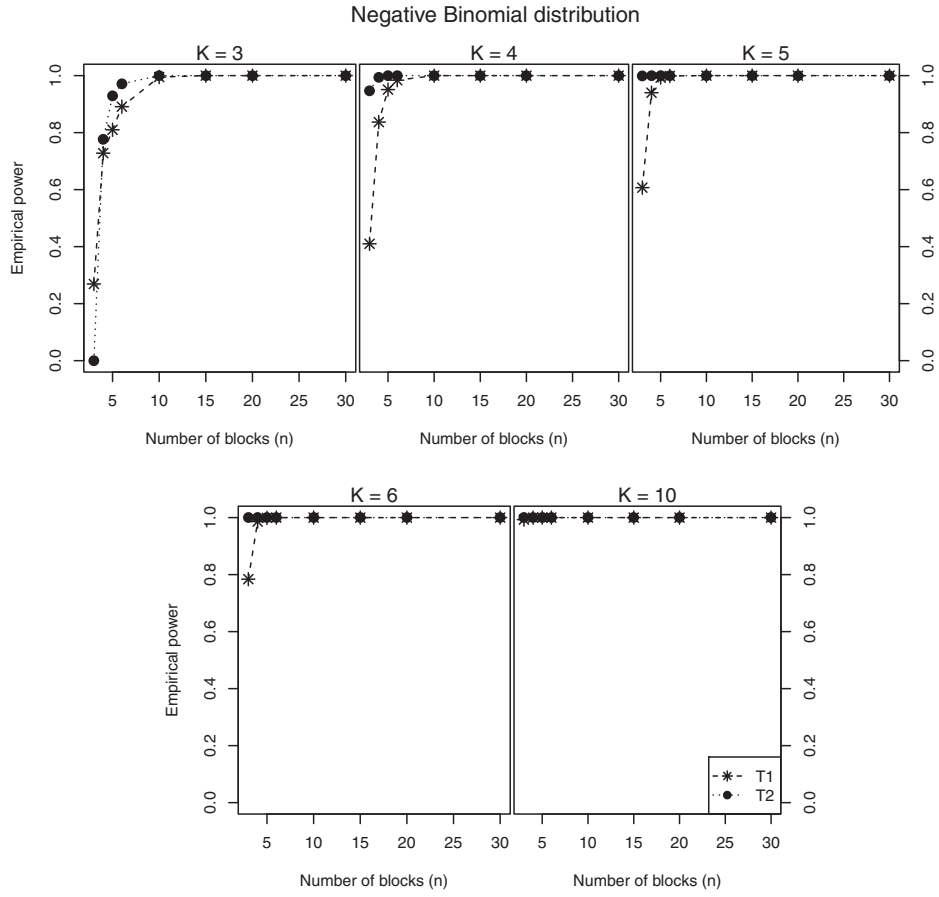

Figure 3. Block power of Friedman's test and Iman-Davenport extension with Negative Binomial distribution, for K = 3, 4, 5, 6 and 10 treatments.

These results illustrated the power of the post-hoc procedures. A large number of blocks and treatments were necessary to obtain an empirical power near 1, when the data presented a high variability within and medium variability between treatments (Uniform distribution). In the presence of medium variability within and between treatments (Poisson distribution) with 20 blocks the empirical power achieved high values. The empirical power easily achieved a value near 1, when data presented homogeneity within treatments (Binomial or Negative Binomial distributions). These was valid even if the heterogeneity between treatments decreases (Binomial distribution) or increases (Negative Binomial distribution) as the number of treatments increases. This explains the speed at which the power approaches the value 1: faster in the presence of heterogeneity between treatments and slower in the presence of heterogeneity within treatments.

6. Azores Buzzard Survey in Azores, Portugal

We illustrate the post-hoc procedures with real data from a survey that was carried out between 1996 and 1998 over a population of the Azores Buzzard (*Buteo buteo rothschildi*)

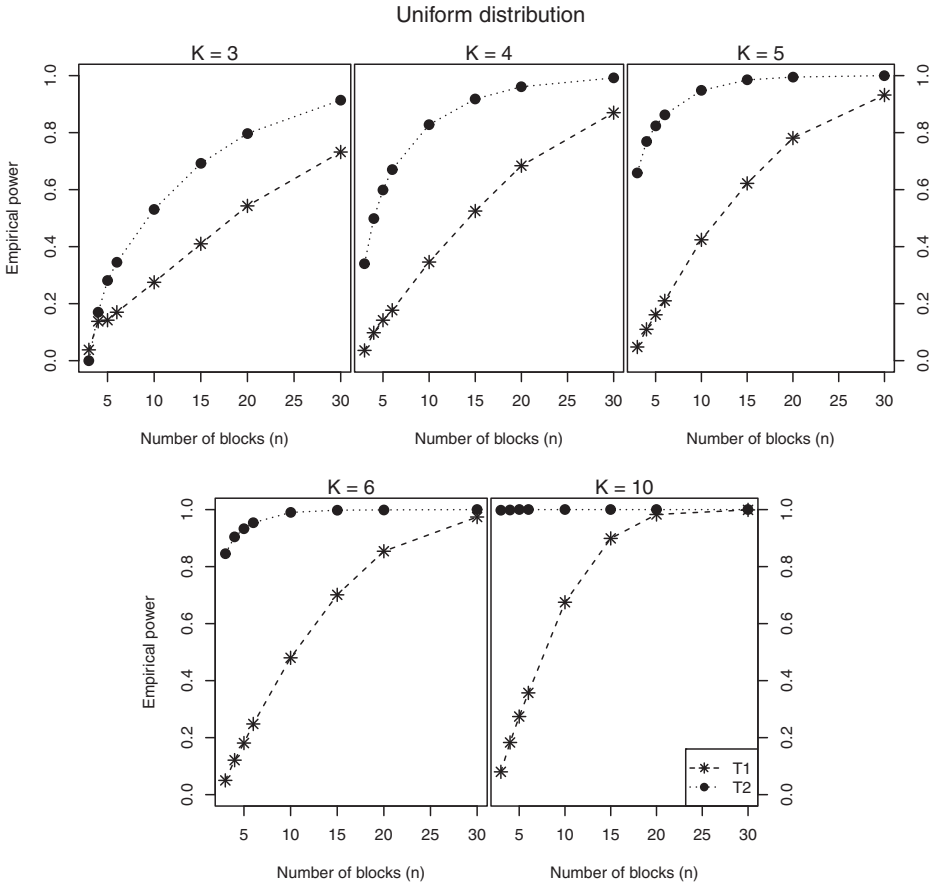

Figure 4. Block power of Friedman's test and Iman-Davenport extension with Discrete Uniform distribution, for K = 3, 4, 5, 6 and 10 treatments.

at Azores archipelago, in Portugal. In 1996, a sample of 11 units of 5×2 Km^2 were randomly chosen in S. Miguel Island and they were located at: Feteiras, Castelo Branco, Fajã de Cima, Rabo de Peixe, Relva, Pico do Carvão, Nordestinho, Sete Cidades, Furnas, Lagoa do Fogo, and Pico Bartolomeu. Each sample unit was rearranged in order to include as much habitats as possible. A line transect was drawn at each sample unit along roads and paths with the aim of covering as much area as possible. An observer crossed the transects by car at a speed of less than 40 km per hour, during the first two months of the breeding season. The observer recorded the number of detected Azores Buzzard. This survey was done only once in 1996 and three times in 1998 at the same 10 sample units. In 1998, each visit was done following the opposite direction of the previous one.

We intend to find out if there are significant differences between the number of the Azores Buzzard observed in each visit ($K = 4$ treatments) for each location ($n = 11$ blocks). We also want to evaluate if the number of detected animals is statistically different in the several locations ($K = 11$ treatments, $n = 4$ blocks).

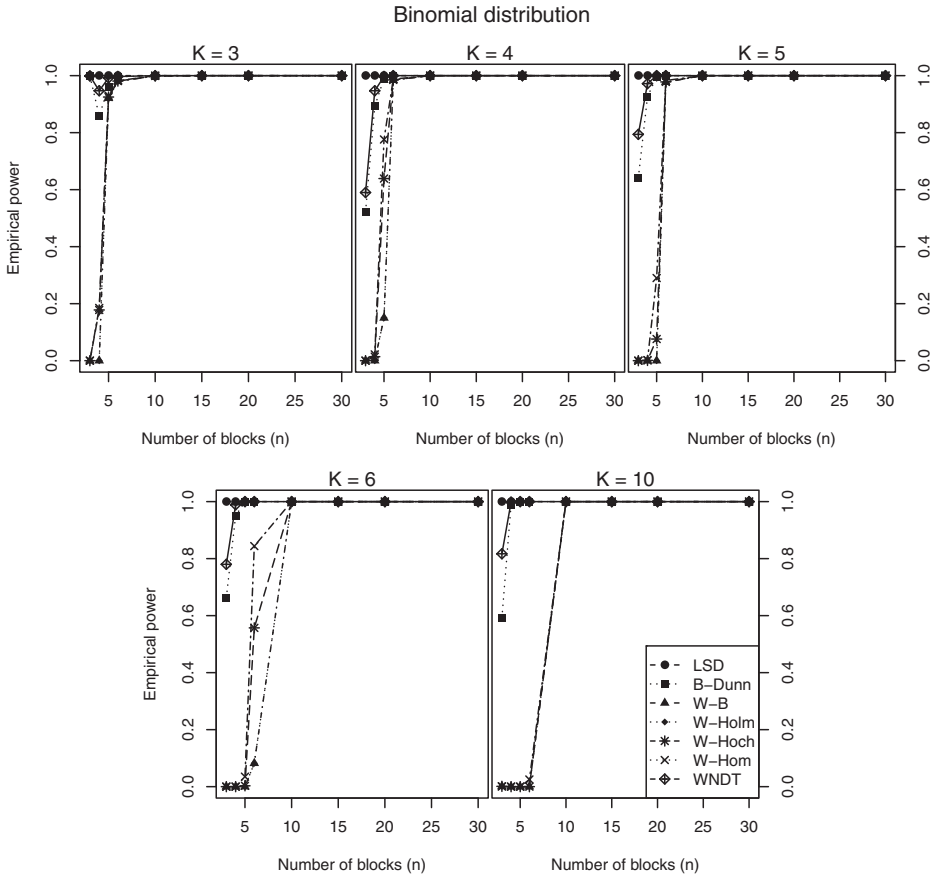

Figure 5. Comparison between all post-hoc procedures with Binomial distribution, for K = 3, 4, 5, 6 and 10 treatments.

Using Friedman's test, we conclude that there are no significant differences between the number of observed Azores Buzzard in each visit ($K = 4$, $n = 11$; $T_1 = 5.717$, $p - value_{T_1} = 0.126$; $T_2 = 2.10$, $p - value_{T_2} = 0.122$). On the other hand, there are significant differences by location ($K = 11$, $n = 4$; $T_1 = 25$, $p - value_{T_1} = 0.005$; $T_2 = 5$, $p - value_{T_2} = 0.0003$). In both cases, the $p - value$ obtained with the T_2 statistics is smaller than with T_1 statistics.

To preserve and improve habitats, species abundance is important to wildlife managers. At 0.05 significance level, with this dataset no significant differences were observed in the number of detected animals in each visit. Considering that in both years the same observational conditions (protocol, sample effort, weather conditions,...) were warranted, this may indicate that the population remained stable during that period. We also found out that the site locations were not equally preferred by this species, but most of the post-hoc test did not identify which were the preferred. Only Fisher's LSD test founded 22 pairs of locations significantly different. This finding can be explained by the fact that Fisher's LSD test is known to be a liberal test. It has the highest likelihood of committing one or more Type I errors in a set/family of comparisons (Sheskin, 2007).

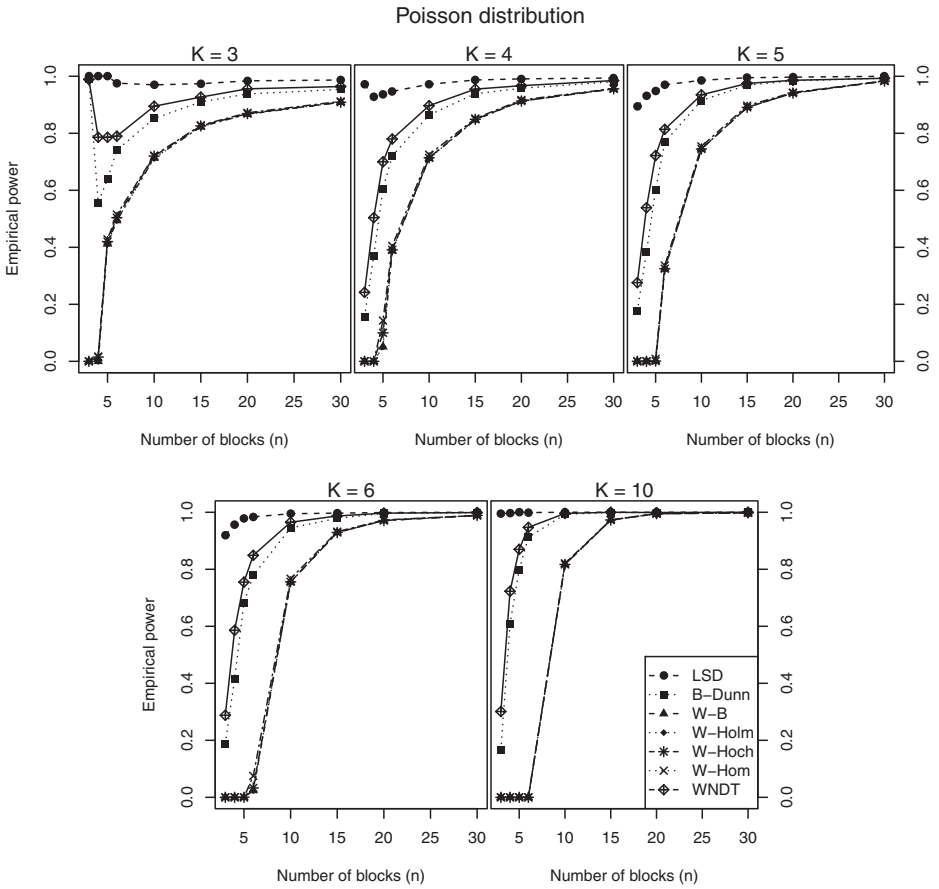

Figure 6. Comparison between all post-hoc procedures with Poisson distribution, for K = 3, 4, 5, 6 and 10 treatments.

7. Concluding Remarks

In this article, we applied Friedman's test and the Iman-Davenport extension. If one of these tests rejects, the hypothesis of equivalence of medians, the detection of specific differences among treatments can be made with the application of post-hoc statistical procedures.

We have shown in terms of empirical power that Friedman's test performs better than the Iman-Davenport extension only when $K = 3$ and $n = 3$, out of the total cases considered in this article.

The post-hoc procedures presented similar results, when we apply them after Friedman's test or the Iman-Davenport extension. In general, the number of treatments (K) used in multiple comparisons procedures should be lower than the number of blocks (n). The p-values of the post-hoc tests are lower if we increase the number of blocks used in multiple comparisons procedures; therefore, the differences among treatments are more detectable.

The multiple sign test has been described in this article. This procedure is easy to apply, but our results indicate that it has much less power with respect to more advanced techniques. This test should not be used because it is very conservative and many differences

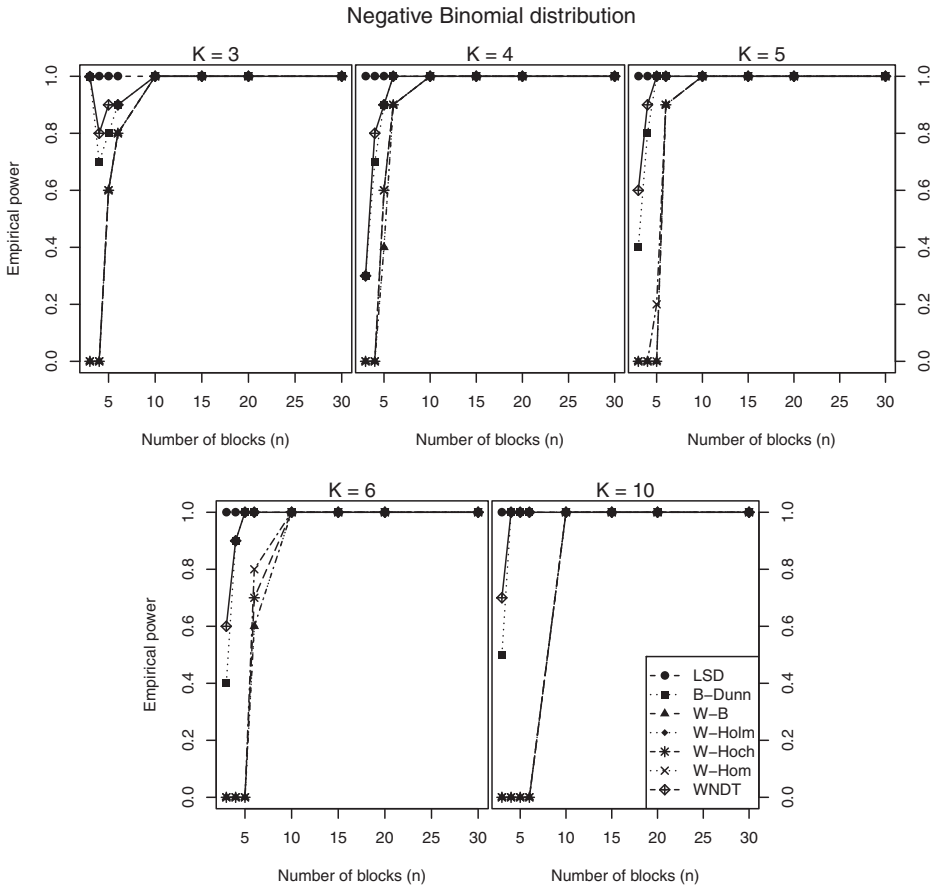

Figure 7. Comparison between all post-hoc procedures with Negative Binomial distribution, for K = 3, 4, 5, 6 and 10 treatments.

may not be detected. We only recommend its use when the differences between treatments are very clear.

One way to characterize the post-hoc tests is as liberal (powerful) or conservative (lack of power). The common post-hoc tests we focus on this article can be listed from most liberal to most conservative, although this varies somewhat depending on the situation. In this work, we have identified three groups of post-hoc procedures:

- Group 1 (extremely conservative, much lower power): Binomial Sign test.
- Group 2 (more conservative, lower power): Wilcoxon test with Holm, Hochberg, Hommel or Bonferroni correction. The differences in power between all these *p*-value adjustments are in practice rather small.
- Group 3 (more liberal, larger power): Bonferroni-Dunn test, Fisher's LSD test and Wilcoxon-Nemenyi-McDonald-Thompson test (Tukey-HSD). Fisher's LSD test and Tukey-HSD revealed to be the two most powerful procedures.

We have found that, when we consider the discrete distributions commonly used to model count data in biological and ecological fields, the Wilcoxon test with Bonferroni and

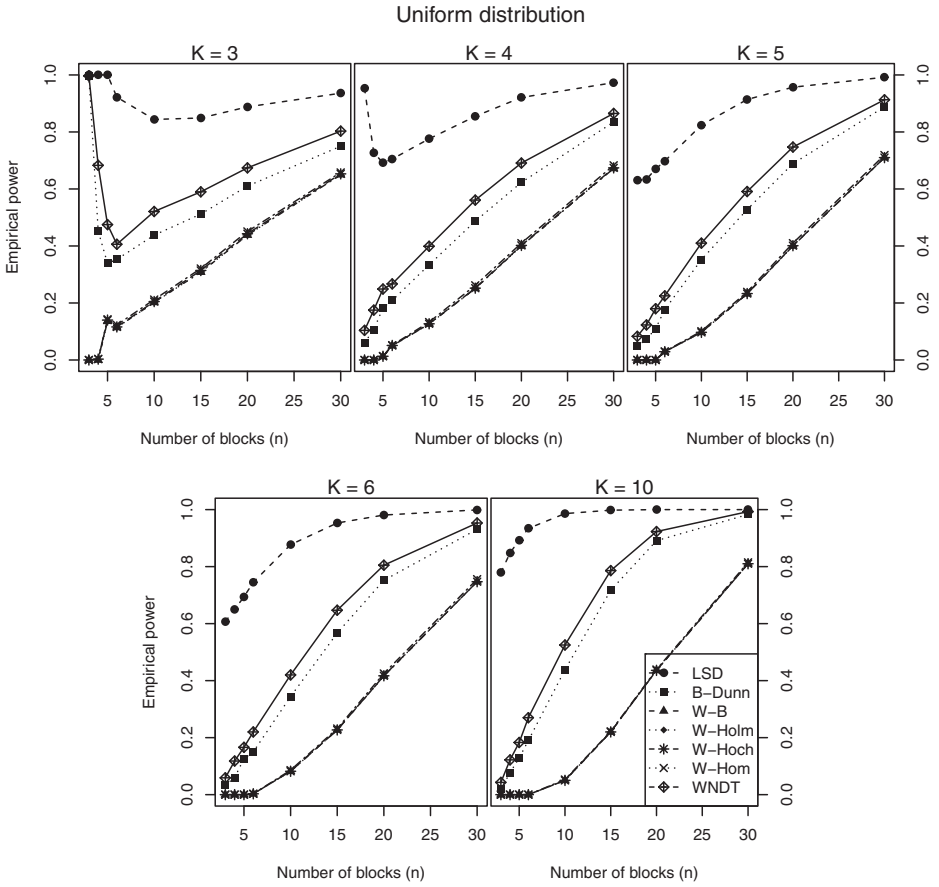

Figure 8. Comparison between all post-hoc procedures with Discrete Uniform distribution, for K = 3, 4, 5, 6 and 10 treatments.

Holm corrections gave the same results (power). In the literature (e.g. Holm, 1999), only continuous distributions are considered, and it is referred that the Bonferroni correction has less power than the modification of Holm. These findings are under Dunn's test while ours are under Wilcoxon's test. This fact will require future analysis.

We point out that the choice of any post-hoc procedure presented in this article for conducting an experimental analysis should be justified by the research. The use of the most powerful procedures does not imply that the results obtained will be better. The choice of a post-hoc procedure is ruled by a trade-off between its power and its complexity when it comes to being used or explained to non-statistical-expert readers.

Acknowledgments

A special note of thanks to our friends G. Carita, P. Filipe, A. Pinto, R. Jara, and H. Perin for their careful reading. The authors are very grateful to the Associate Editor, and two anonymous reviewers for their comments and suggestions which enriched very much our manuscript.

Funding

DGP and AA are members of the CIMA-UE, a research center financed by the Science and Technology Foundation, Portugal.

References

Chen, J., Luo, J., Liu, K., Mehrotra, D. V. (2011). On power and sample size computation for multiple testing procedures. *Computational Statistics and Data Analysis* 55:110–122.

Chiang, C. L. (2003). *Statistical Methods of Analysis*. New Jersey, NJ: World Scientific.

Church, J. D., Wike, E. L. (1979). A Monte Carlo study of nonparametric multiple-comparison tests for a two-way layout. *Bulletin of the Psychonomic Society* 14:95–98.

Conover, W. J. (1999). *Pratical Nonparametric Statistics*. 3rd ed. New York: John Wiley & Sons.

Demsar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research* 7:1–30.

Derrac, J., García, S., Molina, D., Herrera, F. (2011). A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarn and Evolutionary Computation* 1:3–18.

Diggle, P. J. (2003). *Statistical Analysis of Spatial Point Patterns*. 2nd ed. London: Arnold.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association* 56:52–64.

Dunnett, C. W. (1964). New tables for multiple comparisons with a control. *Biometrics* 20: 482–491.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association* 32:674–701.

García, S., Fernández, A., Luengo, J. Herrera, F., (2010). Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Information Sciences* 180:2044–2064.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 75:800–803.

Hollander, M., Wolfe, D. A. (1999). *Nonparametric Statistical Methods*. 2nd ed. New York: John Wiley & Sons.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6:65–70.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* 75:383–386.

Iman, R. L., Davenport, J. M. (1980). Approximations of the critical region of the Friedman statistics. *Communications in Statistics – Theory and Methods* 9:571–595.

Lehmann, E. L., Romano, J. P., Shaffer, J. P. (2005). On optimality of stepdown and stepup multiple test procedures. *The Annals of Statistics* 33:1084–1108.

O'Gorman, T. W. (2001). A comparison of the F-test, Friedman's test, and several aligned rank tests for the analysis of randomized complete blocks. *Journal of Agricultural, Biological, and Environmental Statistics* 6:367–378.

R Development Core Team (2005). *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical. Available from http://www.R-project.org.

Sarkar, S. K., Chang, C.-K. (1997). The Simes method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Association* 92:1601–1608.

Seber, G. A. F. (1982). *The Estimation of Animal Abundance and Related Parameters*. 2nd ed. London: Charles Griffin.

Senn, S., Bretz, F. (2007). Power and sample size when multiple endpoints are considered. *Pharmaceutical Statistics* 6:161–170.

Sheskin, D. J. (2007). *Handbook of Parametric and Nonparametric Statistical Procedures*. 4th ed. Boca Raton: Chapman & Hall/CRC.

Siegel, S., Castellan Jr, N. J. (1988). *Nonparametric Statistics for the Behavioral Sciences*. 2nd ed. New York: McGraw-Hill.

St. Laurent, R., Turk, P. (2013). The effects of misconceptions on the properties of Friedman's test. *Communications in Statistics – Simulation and Computation* 42:1596–1615.

Tamhane, A. C., Liu, W., Dunnett, C. W. (1998). A generalized step-up-down multiple test procedure. *The Canadian Journal of Statistics* 26:353–363.

Westfall, P. H., Tobias, R. D., Rom, D., Wolfinger, R. D., Hochberg, Y. (1999). *Multiple Comparisons and Multiple Tests Using the SAS System*. Cary, NC: SAS Institute Inc.

White, G. C., Bennetts, R. E. (1996). Analysis of frequency count data using the negative binomial distribution. *Ecology* 77:2549–2557.

Wright, S. P. (1992). Adjusted p-values for simultaneous inference. *Biometrics* 48:1005–1013.

Zimmerman, D. W., B. D. Zumbo., (1993). Relative power of the Wilcoxon test, the Friedman test, and repeated-measures ANOVA on ranks. *Journal of Experimental Education* 62:75–86.